HIM: DISCOVERING IMPLICIT RELATIONSHIPS IN HETEROGENEOUS SOCIAL NETWORKS

Bo Xu[1], *Jinpeng Wang*[1], *Zhehuan Zhao*[1], *Hongfei Lin*[2], *Feng Xia*[3]

[1] School of Software, [2] College of Computer Science and Technology, Dalian University of Technology, Dalian, China
[3] School of Computing Technologies, RMIT University, Melbourne, VIC 3000, Australia

ABSTRACT

To date, research on relation mining has typically focused on analyzing explicit relationships between entities, while ignoring the underlying connections between entities, known as implicit relationships. Exploring implicit relationships can reveal more about social dynamics and potential relationships in heterogeneous social networks to better explain complex social behaviors. The research presented in this paper explores implicit relationships discovery methods in the context of heterogeneous social networks. First, the creation of a novel implicit relationships dataset is described, namely HIMdata. Next, a framework for discovering implicit relationships in heterogeneous social networks is introduced. The proposed framework, HIM, innovatively integrates node attributes information and network structure information with graph convolutional networks for discovery of implicit relationships. Finally, HIM is evaluated on two different types of networks, achieving state-of-the-art performance on implicit relationship discovery tasks. The source codes are released at https://github.com/myjpgit/HIM.git.

Index Terms— Heterogeneous social networks, implicit relationship discovery, graph convolutional networks.

1. INTRODUCTION

Recently, relation mining is widely used for a broad range of applications, including user analysis [1, 2], recommendation [3,4] and semantic similarity search [5]. Current research on relation mining has primarily focused on exploring explicit relationships between entities. However, less work has explored the hidden relations that may exist beneath the explicit connections in social networks [6–8]. These non-explicit, hidden relationships are also known as implicit relationships.

Implicit relationships refer to implied relationships between entities in a social network. They can reflect the underlying reason for entities to establish explicit connections. For example, in a scientific collaboration network where collaborations between scholars are explicit relationships, the underlying reason for their collaboration may be that one scholar is the advisor of the other. In the context of the social network Weibo, the mutual following relationship between users is considered an explicit relationship. However, it may be that two users follow each other on Weibo because they are in the same city. The advisor-advisee relationship in this collaboration network and the same-city relationship in the Weibo following network could be regarded as implicit relationships. Implicit relationships are essential for understanding the potential connections between entities, revealing the formation rules of social networks, and obtaining the behavioral characteristics of entities within social networks.

Despite the importance of both explicit and implicit relationships in understanding social network dynamics [8, 9], the extraction of implicit relationships is more complicated than extracting explicit relationships, because implicit relationships may not have real links/edges. Thus, to accurately model the complex social behaviors of entities in social networks, relation mining methods must be developed that can accurately model both explicit and implicit relationships.

This paper explores methods for mining implicit relationships from social network data. First, a definition of implicit relationships is provided. Next, a novel implicit relationships dataset, known as HIMdata, is introduced to facilitate the investigation of implicit relationships from social networks. To support the discovery of implicit relationships in heterogeneous social networks, the HIM framework is proposed which consists of three major components, described as follows. In the first component, node attribute features are aggregated by graph neural networks, and the link attribute embeddings are calculated by the common neighbors (*CN*) index. In the second component, Graph Convolutional Networks (GCN) are used to model relations in the data. Finally, the third component predicts the implicit relationships for each candidate edge. Our contributions can be summarized as follows:

(1) HIMdata. We create a novel implicit relationships dataset, namely HIMdata. It consists of data from two distinct social networks: Same-City Relationship (from Weibo) and Advisor-Advisee Relationship (from Microsoft Academic Graph). HIMdata will be released publicly for research.

(2) Efficient framework. We propose a novel implicit relationships discovery framework, namely HIM. HIM aggregates node attributes and network structure information from nodes' heterogeneous neighbors to predict implicit relationships in social networks.

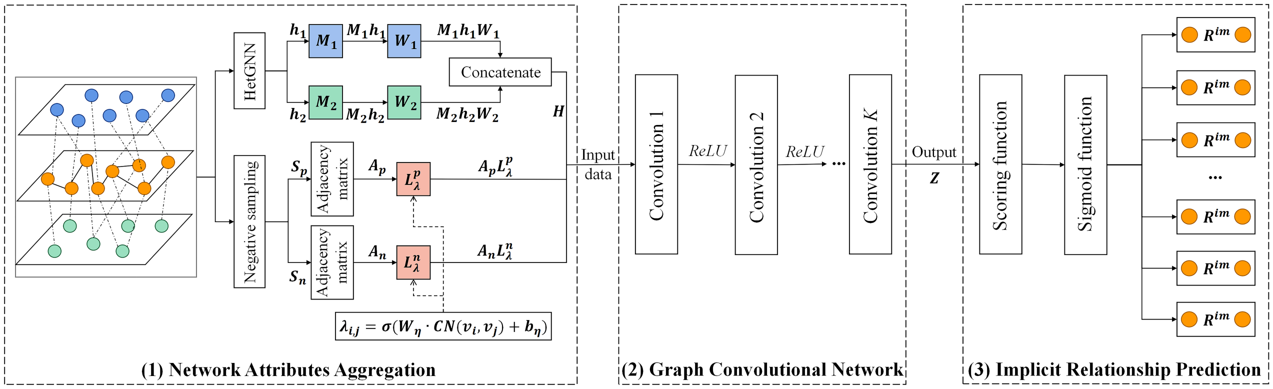

Fig. 1. The overall framework of our proposed HIM.

(3) Flexible and superior method. The ability of HIM to aggregate network attributes makes it applicable to multiple types of networks and outperforms existing SOTA methods.

2. DEFINITION OF IMPLICIT RELATIONSHIPS

Explicit relationships are represented by real edges between entities in networks. In contrast, implicit relationships refer to inconspicuous relationships between entities in networks that are not clear on the surface but reflect the underlying reasons for entities to establish explicit connections. These reasons can include shared interests, same location, work colleagues, shared purpose, and other social connections.

Some researchers have noted the implicit relationships that are hidden in social networks [10–12], but they have only considered simple implicit relationships between entities in a social network, such as relationships based on co-occurrence or common interests. However, the nature of implicit relationships is more complicated in many heterogeneous social network relationships. Thus, we propose a broader definition of implicit relationships, which is defined as:

Implicit relationships. Implicit relationships refer to the potential relationships between entities in the heterogeneous network, largely reflecting the cause of the explicit relationships. In a heterogeneous network $G = (V, E, \varphi, \psi)$, where V and E denote the sets of nodes and edges, each node $v \in V$ and each edge $e \in E$ correspond to one or more type mapping functions $\varphi(v) : v \to T$ and $\psi(e) : e \to R$, where T and R denotes the node and edge types, $|T| + |R| > 2$. The implicit relationship R^{im} can be inferred by entity features F, external knowledge K, and explicit edges E. Implicit relationship R^{im} can refer to an implicit edge type that is not explicitly displayed in heterogeneous network G.

3. THE HIM FRAMEWORK

This section elaborates on our proposed HIM framework, which is shown in Fig. 1. The following subsections provide a detailed overview of each module respectively.

3.1. Network Attributes Aggregation

To capture both node attributes and network link attributes, HIM takes advantage of the HetGNN model [13] and connected node similarity indices to fuse multiple types of information contained in different nodes.

First, inspired by previous work [14], Bi-LSTM is used as an aggregator to combine the heterogeneous neighbors' attributes of nodes under a specific relationship. We denote the t-type sampled neighbor set of $v \in V$ obtained by RWR (Random Walk with Restart) as $N_t(v)$, and use a neural network f^t to aggregate content embeddings of $v' \in N_t(v)$:

$$f^t(v) = \frac{\sum_{v' \in N_t(v)} \left[\overrightarrow{LSTM}\{x(v')\} \oplus \overleftarrow{LSTM}\{x(v')\} \right]}{|N_t(v)|} \quad (1)$$

where $t \in T$ denotes a specific node type in the heterogeneous network, $f^t(v)$ is the type-based aggregated attributes representation of node v.

The previous step generates the $|T|$ aggregated attributes representation for node v. Since different types of neighbors will make different contributions to the attributes representation of node v, the attention mechanism is used. Thus, the attributes representation of node v is defined as:

$$h_v = \alpha^{v,v} x(v) + \sum_{t=1}^{T} \alpha^{v,t} f^t(v) \quad (2)$$

where h_v is the attributes representation of node v that aggregates different types of neighbors, $\alpha^{v,*}$ denotes the importance of different attributes representation, and $*$ denotes v or t. If all nodes are considered and M_t denotes the link matrix, the final attributes matrix H of nodes can be expressed as:

$$h_t = \sum_{v=1}^{V} \left[\alpha^{v,v} x(v) + \alpha^{v,t} f^t(v) \right] \quad (3)$$

$$H = \sum_{t=1}^{T} M_t \sum_{v=1}^{V} \left[\alpha^{v,v} x(v) + \alpha^{v,t} f^t(v) \right] W_t = \sum_{t=1}^{T} M_t h_t W_t \quad (4)$$

where W_t is weight matrix and h_t denotes the attributes representation of node $v \in V$ that aggregates t-type neighbors.

Second, the similarity of two connected nodes is utilized to define the attributes representing links or connections (i.e., network structure information). The common neighbors (*CN*) approach [15] is used as the similarity index. After we convert the similarity of two connected nodes into the link attributes $\eta_{e_{v_i,v_j}}$ of edge e_{v_i,v_j}, the link attributes weight $\lambda_{i,j}$ for edge e_{v_i,v_j} is calculated as:

$$\lambda_{i,j} = \sigma(W_\eta \cdot \eta_{e_{v_i,v_j}} + b_\eta) \tag{5}$$

where W_η is the weight matrix and b_η is the bias vector. Finally, the adjacency matrices of positive samples A_p and negative samples A_n are multiplied by the link attributes weight matrix L_λ as part of the input information of the GCN model.

3.2. Graph Convolutional Network

After processing the node attributes and network link attributes, relations are modelled by using graph convolutional networks. In order to make full use of the information of neighbor nodes and thereby improve the effectiveness of information dissemination and conversion, we lead into the first-order neighbors of the target node in GCN model to extend the receptive field of a single graph convolutional layer. Taking the relation type into account, the information dissemination method of each layer can be expressed as:

$$h_i^{l+1} = ReLU\left(\sum_{r\in R} A_i^s h_i^l W_r^l + \frac{1}{|N_r^i|}\sum_{j\in N_r^i} A_j^s h_j^l W_r^l + b^l\right) \tag{6}$$

where N_r^i is the set of v_i's neighbors under the relation type $r \in R$. $A_i^s = \widetilde{D}_i^{-\frac{1}{2}}\widetilde{A}_i\widetilde{D}_i^{-\frac{1}{2}}$ is the normalized adjacency matrix of node v_j, $\widetilde{D}_i$ is the diagonal degree matrix of $\widetilde{A}_i$, $h_i^l \in \mathbb{R}^{d(l)}$ is the feature of node v_i in the l-th hidden layer.

3.3. Implicit Relationship Prediction

After the $\mathcal{K}$ layer GCN, the final representation of v_i is computed as $z_i = h_i^{\mathcal{K}}$. Next, we utilize the traditional method of link prediction to predict unknown links in G. Specifically, we use a scoring function $\mathcal{F}(v_s, R^{im}, v_o)$ to predict the score of candidate links (v_s, v_o):

$$\mathcal{F}(v_s, R^{im}, v_o) = z_s^T R z_o \tag{7}$$

Finally, the probability $\hat{\mathcal{F}}(v_s, R^{im}, v_o)$ of unlabeled triplet (v_s, R^{im}, v_o) existing under R^{im} is defined as:

$$\hat{\mathcal{F}}(v_s, R^{im}, v_o) = \phi\left(\mathcal{F}(v_s, R^{im}, v_o) + b\right) \in (0,1) \tag{8}$$

where b is a bias vector, and $\phi(x) = 1/(1+\exp(-x))$ is a sigmoid function which expresses the probability prediction of whether the new triplet (v_s, R^{im}, v_o) is correct or not.

In the training process, we use the cross-entropy function as the loss function to train the parameters of the model. Negative sampling is used to accelerate the convergence of the method. If E denotes the set of edges, the loss function $\mathcal{L}_{sum}$ for HIM can be defined as follows:

$$\mathcal{L}_{sum} = \sum_{(v_s,R^{im},v_o)\in E} \mathcal{L}(s,o) = -\frac{1}{|E|} \sum_{(v_s,R^{im},v_o)\in E}\left(\delta log\hat{\mathcal{F}}_{s,o} + \frac{1}{|S_n^o|}\sum_{v_o'\in S_n^o}(1-\delta)\, log(1-\hat{\mathcal{F}}_{s,o'})\right) \tag{9}$$

where v_o' is a random node which is not connected with v_s, $\hat{\mathcal{F}}_{s,*}$ represents $\hat{\mathcal{F}}(v_s, R^{im}, v_*)$, δ is defined as:

$$\delta = \begin{cases} 1,\ if\ (v_s, R^{im}, v_o) \in S_p \\ 0,\ if\ (v_s, R^{im}, v_o') \in S_n \end{cases} \tag{10}$$

where S_n is negative samples, which is composed of edges generated by negative sampling based on positive samples S_p.

4. EXPERIMENTS

4.1. Dataset and Baselines

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. The dataset is titled HIMdata and can be downloaded from https://github.com/myjpgit/HIMdata.git.

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 here means that the geographic locations in the personal information of two users are in the same city, which is considered as an implicit relationship. We successfully screened out 34,438 users' mutual following relationship pairs, including 10,195 users and 16,428 same-city relationship pairs.

Advisor-Advisee Relationship. Microsoft Academic Graph, which contains information about scholars and publications, was used to construct an academic co-author heterogeneous network. Here, 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. Thus, the advisor-advisee relationship is considered as an implicit relationship. We successfully obtained 8282 co-authors relationship pairs, including 7872 scholars and 2787 advisor-advisee relationship pairs.

We compare the HIM with following state-of-the-art network representation learning methods to evaluate the performance of HIM, including GATNE [16], GRCN [17], HGT

Table 1. 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 [16]	0.862	0.866	0.872	0.794	0.800	0.813
GRCN [17]	0.817	0.828	0.832	0.676	0.706	0.709
HGT [18]	0.847	0.858	0.862	0.766	0.776	0.779
GEN [19]	0.808	0.813	0.822	0.672	0.680	0.698
SLICE [20]	0.924	0.922	0.933	0.853	0.857	0.876
Shifu2 [9]	0.824	0.835	0.834	0.753	0.769	0.763
MHGCN [21]	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

Table 2. 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 [16]	0.617	0.628	0.631	0.543	0.559	0.567
GRCN [17]	0.572	0.576	0.580	0.496	0.507	0.509
HGT [18]	0.628	0.634	0.640	0.565	0.571	0.576
GEN [19]	0.648	0.653	0.652	0.582	0.593	0.589
SLICE [20]	0.686	0.691	0.695	0.611	0.618	0.622
Shifu2 [9]	0.714	0.726	0.733	0.673	0.692	0.706
MHGCN [21]	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**

[18], GEN [19], SLICE [20], Shifu2 [9] and MHGCN [21]. Furthermore, we adopt the average of the Area Under Curve (AUC) and F1 score obtained from 10 independent training cycles to evaluate the performance of different methods.

4.2. Experimental Results and Analysis

As demonstrated in Table 1 and Table 2, 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, MHGCN, and HIM, the remaining network representation learning methods have not achieved competitive performance, because methods designed for link prediction and those designed for other downstream tasks have different focuses on aggregating network information, which leads to differences in the emphasized information.

Furthermore, Shifu2, which is specially designed for mining advisor-Advisee implicit relationship, shows excellent performance in Advisor-Advisee Relationship dataset, but poor performance in Same-City Relationship dataset. These results prove that Shifu2 has strong domain pertinence, while our proposed HIM stands out by identifying multiple implicit relationships instead of focusing on a single type, demonstrating its superiority in this task.

We further compare the results of HIM and its variables, namely HIM w/o *HAtt* and HIM w/o *LAtt*. HIM w/o *HAtt* and HIM w/o *LAtt* denote the framework without heterogeneous attributes aggregation (i.e., using X_t instead of h_t in equation (4)) and without link attributes weight (i.e., without using link attributes weight $\lambda_{i,j}$). The experimental results are shown in Table 3. 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 w/o *HAtt* and HIM w/o *LAtt* on all datasets. 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.

Table 3. Performance of HIM w.r.t. AUC and F1 with Different Aggregation Attributes on Each Dataset.

Dataset	Same-City Relationship		Advisor-Advisee Relationship	
Criteria	AUC	F1 Score	AUC	F1 Score
HIM w/o *HAtt*	0.974	0.928	0.717	0.680
HIM w/o *LAtt*	0.967	0.917	**0.751**	**0.722**
HIM	**0.979**	**0.943**	0.735	0.713

We conducted additional experiments on "Terrorist Attack" dataset to demonstrate the broader applicability of HIM. Due to page limitations, these experimental results were not included in the paper. For details about the dataset, its construction methodology, and comprehensive experimental results, please refer to https://github.com/myjpgit/HIM/blob/master/Supplementary.pdf

5. CONCLUSION

In this paper, we create a large-scale implicit relationships dataset, namely HIMdata, for investigating hidden connections between entities in social networks. Further, the HIM framework is proposed, which innovatively integrates node attributes information and network structure information with Graph Convolutional Networks for implicit relationships discovery. The proposed HIM framework has important significance for many real-world applications. For example, the mining of same-city relationships could help same-city business recommendation applications. In addition, understanding the implicit relationships in social networks is very important for the analysis of the formation and development of social networks and the analysis of social users' behavior.

Acknowledgement This work was supported in part by the National Natural Science Foundation of China under Grant 62072073, Grant 61906028, Grant 62076046, 62106034 and in part by the Dalian Innovation Fund 2021JJ12GX016.

6. REFERENCES

[1] C. Mi, X. Ruan, and L. Xiao, "Microblog sentiment analysis using user similarity and interaction-based social relations," *Int. J. Web Serv. Res.*, vol. 17, no. 3, pp. 39–55, 2020.

[2] J. H. Chen and H. T. Chen, "The role of social network sites on the relationship between game users and developers: An evolutionary game analysis of virtual goods," *Inf. Technol. Manag.*, vol. 22, no. 2, pp. 67–81, 2021.

[3] C. Wang, W. Ma, M. Zhang, C. Chen, Y. Liu, and S. Ma, "Toward dynamic user intention: Temporal evolutionary effects of item rela-tions in sequential recommendation," *ACM Trans. Inf. Syst.*, vol. 39, no. 2, pp. 16:1–16:33, 2021.

[4] Z. Zhao, Y. Yang, C. Li, and L. Nie, "Guessuneed: Recommending courses via neural attention network and course prerequisite relation embeddings," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 16, no. 4, pp. 132:1–132:17, 2021.

[5] J. N. Chiang, Y. Peng, H. Lu, K. J. Holyoak, and M. M. Monti, "Distributed code for semantic relations predicts neural similarity during analogical reasoning," *J. Cogn. Neurosci.*, vol. 33, no. 3, pp. 377–389, 2021.

[6] M. R. Hamedani, I. Ali, J. Hong, and S. W. Kim, "Trustrec: An effective approach to exploit implicit trust and distrust rela-tionships along with explicitones for accurate recommendations," *Comput. Sci. Inf. Syst.*, vol. 18, no. 1, pp. 93–114, 2021.

[7] B. Xu, B. Chen, T. Zhang, J. Liu, C. Liao, and Z. Zhao, "Mining multivariate implicit relationships in academic networks," in *WWW' 22*, 2022, pp. 1148–1156.

[8] J. Liu, F. Xia, J. Ren, B. Xu, G. Pang, and L. Chi, "Mirror: Mining implicit relationships via structure-enhanced graph convolutional networks," *ACM Trans. Knowl. Discov. Data*, vol. 17, no. 4, pp. 55:1–55:24, 2023.

[9] J. Liu, F. Xia, L. Wang, B. Xu, X. Kong, H. Tong, and I. King, "Shifu2: A network representation learning based model for advisor-advisee relationship mining," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1763–1777, 2021.

[10] M. Song, N. G. Han, Y. H. Kim, Y. Ding, and T. Chambers, "Discovering implicit entity relation with the gene-citation-gene network," *PLoS ONE*, vol. 8, no. 12, pp. e84639, 2013.

[11] W. Zhou, W. Duan, and S. Piramuthu, "A social network matrix for implicit and explicit social network plates," *Decis. Support Syst.*, vol. 68, pp. 89–97, 2014.

[12] A. Spitz, S. Almasian, and M. Gertz, "Evelin: Exploration of event and entity links in implicit networks," in *WWW' 17*, 2017, pp. 273–277.

[13] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *KDD' 19*, 2019, pp. 793–803.

[14] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NIPS' 17*, 2017, pp. 1024–1034.

[15] A. L. Buchsbaum, R. Giancarlo, and J. R. Westbrook, "On finding common neighborhoods in massive graphs," *Theor. Comput. Sci.*, vol. 299, no. 1-3, pp. 707–718, 2003.

[16] Y. Cen, X. Zou, J. Zhang, H. Yang, J. Zhou, and J. Tange, "Representation learning for attributed multiplex heterogeneous network," in *KDD' 19*, 2019, pp. 1358–1368.

[17] D. Yu, R. Zhang, Z. Jiang, Y. Wu, and Y. Yang, "Graph-revised convolutional network," in *ECML/PKDD' 20*, 2020, pp. 378–393.

[18] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *WWW' 20*, 2020, pp. 2704–2710.

[19] R. Wang, S. Mou, X. Wang, W. Xiao, Q. Ju, C. Shi, and X. Xie, "Graph structure estimation neural networks," in *WWW' 21*, 2021, pp. 342–353.

[20] P. Wang, K. Agarwal, C. Ham, S. Choudhury, and C. K. Reddy, "Self-supervised learning of contextual embeddings for link prediction in heterogeneous networks," in *WWW' 21*, 2021, pp. 2946–2957.

[21] P. Yu, C. Fu, Y. Yu, C. Huang, Z. Zhao, and J. Dong, "Multiplex heterogeneous graph convolutional network," in *KDD' 22*, 2022, pp. 2377–2387.