Finding Multi-granularity Community Structures in Social Networks Based on Significance of Community Partition

Chenxu Gong
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China
chx.gong@foxmail.com

Guoyin Wang*
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China
wanggy@ieee.org

Jun Hu
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China
hujun@cqupt.edu.cn

Ming Liu
CISDI Information Technology Co.,Ltd
Chongqing, China
ming.liu@cisdi.com.cn

Li Liu
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China
liliu@cqupt.edu.cn

Zihe Yang
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China
zihe.yang@outlook.com

***Abstract*—Community structure detection is an important and valuable task in social network studies as it is the base for many social network applications such as link prediction, recommendation, etc. Most social networks have an inherent multi-granular structure, which leads to different community structures at different granularities. However, few studies pay attention to such multi-granular characteristics of social networks. In this paper, a method called MGCD (Multi-Granularity Community Detection) is proposed for finding multi-granularity community structures of social networks. At first, a network embedding method is used to obtain the low-dimensional vector representation for each node. Then, an effective embedding-based strategy for weakening the detected community structures is proposed. Finally, a joint learning framework, which combines network embedding and community structure weakening is developed for identifying the multi-granularity community structures of social networks. Experimental results on real-world networks show that MGCD outperforms the state-of-the-art benchmark methods on finding multi-granularity community structure tasks.**

Keywords—multiple granularity, social network analysis, community detection, network embedding, granular computing

I. Introduction

Community structure is an important feature of networks. The prevalence of network analysis has drawn many research interests on community detection, and many community detection methods have been proposed. The current researches focused primarily on the most significant community structure in networks. That is, they only find the most significant community structure, instead of the community structure on other weak granularities. However, the community structures on weak granularities are often of high value in practical applications. For example, in social network of duty crime, the significant community structure often represents the relationship between relatives or colleagues, but the criminal groups are represented by some weak community structures. The community structure in networks should be analyzed from a multi-granular perspective [18], which can provide a more comprehensive way for understanding the relationships among communities in complex social networks. To solve this problem, a meta-approach named HICODE (Hidden Community Detection) [1] has been proposed, which gives three community weakening methods (remove edge, reduce edge, and reduce weight) for identifying the hidden community structure by using the existing community detection algorithms such as OSLOM [11], Link Community [12], Infomap [13], and Louvain method [14] iteratively. HICODE represents a network by an adjacency matrix. For large networks, such as networks with millions of nodes, this network representation method has disadvantages such as high computing complexity, poor parallel processing ability and unable to apply machine learning methods when analyzing large networks.

In this paper, a multi-granularity community detection method (MGCD) is proposed for identifying the multi-granularity community structures of social networks. As shown in **Figure 1**, given a network, its multi-granularity community structure can be found by using MGCD. Different granularities indicate the community structure with different significances. To overcome the shortages of traditional network representation methods, each node is represented as a low-dimensional vector by network embedding. Then, a simple and effective strategy to weaken the community structure is proposed based on Gaussian mixture model. Finally, the multi-granularity community structure can be obtained by using these methods iteratively.

The main contributions of this paper include:

- The concept of multi-granularity community structure is defined and the significance of a community partition in a social network is given.

This work was supported by the National Key Research and Development Program of China under Grant 2017YFC0804002, the National Natural Science Foundation of China under Grant 61572091, and Grant 61772096. Chongqing Research Program of Basic Research and Frontier Technology under Grant cstc2017jcyjAX0406 and cstc2017jcyjAX0325.

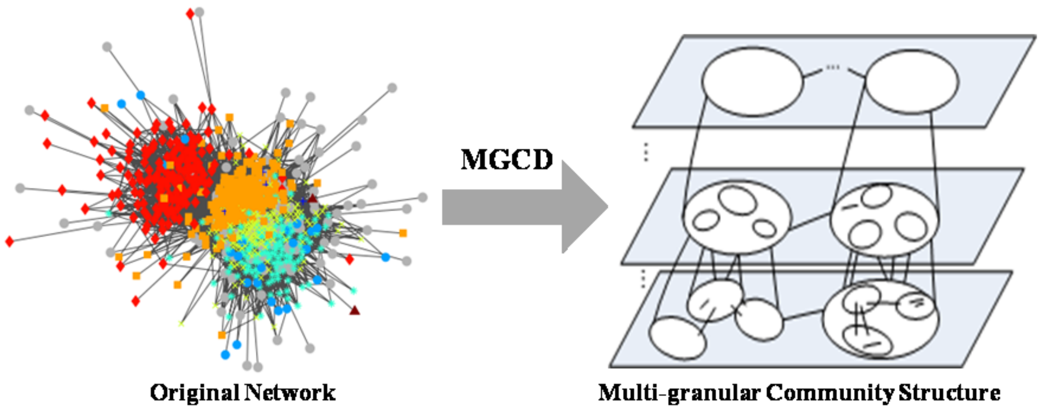

Fig. 1. **Multi-granularity community Detection Framework.**

- Network embedding is used to represent each node as a low-dimensional vector as it can overcome the disadvantages of traditional network representation methods.
- A novel community detection method is developed. It can obtain the multi-granularity community structures, which provides a more comprehensive perspective for understanding the relationships among communities in social networks.

The rest of this paper is organized as follows. Section II summarizes the related works. Section III formally defines the problem of multi-granularity community detection. Section IV introduces the MGCD in detail. Section V presents the experimental results. Finally, we conclude this paper in section VI.

II. Related Works

Our work is related to community detection and network embedding. A large number of studies on community detection have been done in the past [9]. Most community detection methods are based on network structure, such as graph based community detection algorithm [20], edge clustering algorithm [12], seed diffusion method [21], random walk [20], hierarchical clustering [22] and modularity [19].

Network embedding is the state-of-the-art network representation framework. Network embedding aims to represents each node as a low-dimensional vector according to the original network structure, such that two nodes "close" on the network have similar vector representations in a low-dimensional space. Network embedding can overcome the shortages of traditional network representation methods. Many network embedding methods have been proposed recently [8]. For example, DeepWalk [2] finds the distribution of nodes appearing in short random walks is similar to the distribution of words in natural language, so it uses Skip-Gram [10] to learn the representation of nodes. Similar to DeepWalk, Node2Vec [5] improves the random walk strategy, combining local information and global information together to get better result. LINE [6] first learns the representations of nodes that preserve the first-order and second-order proximities. ComE [15] investigates that community structure is an important attribute of the network, considering that the embedding vectors are all generated by the same Gaussian Mixture Model. On the basis of maintaining the first-order and second-order proximities, the high-order proximity (community structure) is added when training network embedding. SDNE [4] first proposes a semi-supervised deep model, which can capture nonlinear network structure through multi-layer nonlinear functions. Network embedding has been shown very successful in preserving the network structure and it has been widely used in node classification [2,23,24], graph visualization [4,25], link prediction [7,26], etc.

III. Preliminaries

In this section, the problem of finding multi-granularity community structures is formulated and the definitions of some concepts used in this paper is given.

Problem Formulation. *As input, giving a graph* ***G=(V,E)*** *to represent a network, where* ***V*** *is the node set and* ***E*** *is the edge set, each node* $\boldsymbol{v_i} \in \boldsymbol{V}$ *representing a data object and each edge* $\boldsymbol{e_{ij}} = (\boldsymbol{v_i}, \boldsymbol{v_j})$ *representing the relationship of two nodes,* $\boldsymbol{w_{ij}} > 0$ *is the weight of edge* $\boldsymbol{e_{ij}}$*. In practice, social networks can be both directed/undirected and weighted/unweighted. In this paper we only discuss undirected and weighted/unweighted graphs. As output, we aim to get the multi-granularity community structures of the network, which shows the community structures with different significance.*

Definition 1. (Multi-granularity community Structure) *Multi-granularity community structure refers to the community detection under different significance. Assume that the network has* ***n*** *granularities and the* ***i****-th granularity can be divided into* $\boldsymbol{k_i}$ *communities.* $\{\boldsymbol{G_1}, \dots, \boldsymbol{G_n}\}$ *represents the multi-granularity community structure. For the* ***i****-th granularity* $\boldsymbol{G_i} = \{\boldsymbol{C_{i1}}, \dots, \boldsymbol{C_{ik_i}}\}$, $\boldsymbol{C_{ik_i}}$ *is the* $\boldsymbol{k_i}$*-th community under the* ***i****-th granularity.*

Generally, the number of edges in the same community is larger than the number of edges among different communities. If there are many edges in a community, the relationship between nodes in this community is close, and the community is easy to be found. According to this observation, a concept called significance of community partition is put forward to evaluate the difficulty of finding current community structure. The higher the significance is, the easier the community to be found. The definition is given as follow.

Definition 2. (Significance of Community Partition) *Giving a graph* ***G=(V, E)****, where* ***V*** *is the node set,* ***E*** *is the edge set and* $\boldsymbol{w_{ij}}$ *is the weight of edge* $\boldsymbol{e_{ij}}$*. Suppose the* *network is divided into K communities, denote by* $\boldsymbol{C} = \{\boldsymbol{C_1}, \dots, \boldsymbol{C_K}\}$*. The significance of* $\boldsymbol{C}$ *is:*

$$S = \frac{\sum_{k=1}^{K} \sum_{(v_i, v_j \in C_k)} w_{ij}}{\sum_{k=1}^{K} \sum_{(v_i \in C_k, v_j \notin C_k)} w_{ij} + \sum_{k=1}^{K} \sum_{(v_i, v_j \in C_k)} w_{ij}} \tag{1}$$

In particular, we set $\boldsymbol{w_{ij} = 1}$ to calculate the significance when $\boldsymbol{G}$ is an unweighted graph.

IV. Multi-granularity community Detection

In this section, a novel approach named MGCD is introduced to identify the multi-granularity community structure in social networks. MGCD uses network embedding to represent a network. Additionally, a simple and effective approach is proposed to weaken the detected community structure. Finally, the multi-granularity community structure can be obtained by using these methods iteratively.

A. Learning Embedding

LINE [6] defines the first-order proximity in a network as the local pairwise proximity between two nodes. The first-order proximity between a pair of nodes (v_i, v_j) in a network is the similarity between their neighborhood network structures. In other words, if two nodes share the same neighbor nodes, then their embedding should be similar. LINE trains the first-order proximity embedding by minimizing the objective function

$$O_1 = -\sum_{(v_i, v_j) \in E} w_{ij} \, log\, \sigma(\varphi_j^T \varphi_i) \tag{2}$$

where $\varphi_i \in \mathbb{R}^d$ is the embedding of node v_i, and $\sigma(x) = 1/(1 + exp(-x))$ is a sigmoid function.

To preserve the second-order proximity [6], LINE and DeepWalk introduce a new role called "context" for each node in the network. In this "context", each node not only acts as the node itself, but also as the context of other nodes. LINE uses negative sampling to characterize the second-order proximity of nodes, and trains the second-order proximity embedding by minimizing the objective function

$$O_2 = -\sum_{v_i \in V} \sum_{v_j \in C_i} [\log \sigma\left({\varphi'}_j^T \varphi_i\right) + \sum_{t=1}^{m} \mathbb{E}_{v_n \sim P_{neg}(v_n)} [log\, \sigma(-{\varphi'}_n^T \varphi_i)]] \tag{3}$$

where $\varphi'_j \in \mathbb{R}^d$ is the context embedding of node v_j, C_i is the context (a set of neighbors) of node v_j , and $v_n \sim P_{neg}(v_n)$ indicates that the probability of sampling to node v_n should be $P_{neg}(v_n)$ when sampling a node v_n from negative table. We set $P_{neg}(v_n) \propto {d_{v_n}}^{3/4}$, which is proposed in [10] and d_{v_n} is the degree of node v_n.

However, traditional network embedding focuses on preserving first-order and second-order proximities. It ignores the high-order proximity in the network. For example, community structure is a significant feature of network, so the node embedding should be similar when they are in the same community. Cavallari et. al raise a network embedding algorithm named ComE [15], which takes community structure into consideration when training embedding. Motivated by Gaussian mixture formulation [16], ComE defines community embedding as a multivariate Gaussian distribution in a low-dimensional space. Each node's embedding is generated by the same GMM (Gaussian Mixture Model). For example, if a network has K communities, then each node's embedding φ_i is generated by a Gaussian mixture model $\sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mu_k, \Sigma_k)$, where $\mathcal{N}(\mu_k, \Sigma_k)$ is a multivariate Gaussian distribution , π_k is the mixture coefficient with $\sum_{k=1}^{K} \pi_k = 1$, $\mu_k \in \mathbb{R}^d$ is a mean vector, and $\Sigma_k \in \mathbb{R}^{d \times d}$ is a covariance matrix. Then, an objective function for preserving the high-order proximity is defined as

$$O_3 = -\frac{\alpha}{K} \sum_{k=1}^{K} \pi_{ik} \, log\, \mathcal{N}(\varphi_i | \mu_k, \Sigma_k) \tag{4}$$

where $\alpha \geq 0$ is a trade-off parameter, π_{ik} is the probability of node v_i belonging to community k (with $k \in \{1, \dots, K\}$).

In this paper, both LINE and ComE are used to train the embedding, which can preserve the first-order, second-order and high-order proximities of the network. Then, it turns to a problem of minimizing the objective function

$$\mathcal{L} = O_1 + O_2 + O_3 \tag{5}$$

Then, the objective function $\mathcal{L}$ can be minimized by asynchronous stochastic gradient algorithm (ASGD) [17]. Gradient the objective function to φ_i, we have

Algorithm 1 Multi-granularity community Detection

Input: a network $\boldsymbol{G=(V, E)}$, the embedding dimension $\boldsymbol{d}$, the negative table size $\boldsymbol{m}$, the offset value $\boldsymbol{\lambda}$, the number of sampling $\boldsymbol{S}$, and the number of communities $\{\boldsymbol{n_1}, \dots, \boldsymbol{n_k}\}$
Output: multi-granularity community structure $\{\boldsymbol{G_1}, \dots, \boldsymbol{G_n}\}$

Initialize alias table;
Initialize negative table;
for 1:$\boldsymbol{S}$ do
  alias sampling an edge $(v_i, v_j) \in \boldsymbol{E}$;
  Update φ_i and φ_j by (6);
end for
for 1:$\boldsymbol{S}$ do
  alias sampling an edge $(v_i, v_j) \in \boldsymbol{E}$;
  Update φ_i by (7);
  Update φ'_j by (9);
end for
for all $\boldsymbol{n_i}$ in $\{\boldsymbol{n_1}, \dots, \boldsymbol{n_k}\}$ **do**
  fit GMM by EM with current embedding;
  for 1:$\boldsymbol{S}$ do
    random sample a node $v_i \in \boldsymbol{V}$;
    Update φ_i by (8);
  end for
  calculate community partition $\boldsymbol{G_i}$ by (10) ;
  for all C_k in G_i **do**
    for all v_i in C_k **do**
      Update φ_i by (11);
    end for
  end for
end for
return community detection result $\{\boldsymbol{G_1}, \dots, \boldsymbol{G_n}\}$;

TABLE I. TABLE TYPE STYLES

Source	Dataset	|V|	|E|	Significance		
				Granularity 1	*Granularity 2*	*Granularity 3*
Facebook	0.edges	333	2,519	0.442	0.346	0.283
	107.edges	1,034	26,749	0.688	0.184	0.478
	348.edges	224	3,192	0.686	0.094	0.760
	414.edges	150	1,693	0.937	0.108	0.691
	686.edges	168	1,656	0.971	0.260	0.885
	1684.edges	786	14,024	0.828	0.264	0.560
	1912.edges	747	30,025	0.706	0.057	0.457
	3437.edges	534	4,813	0.725	0.334	0.649

$$\frac{\partial O_1}{\partial \varphi_i} = -\sum_{(v_i,v_j)\in E} w_{ij}\, \sigma\left(\varphi_j^T \varphi_i\right)\varphi_j \quad (6)$$

$$\frac{\partial O_2}{\partial \varphi_i} = -\sum_{v_j \in C_i}[\sigma\left({\varphi'_j}^T \varphi_i\right)\varphi'_j$$
$$+\sum_{t=1}^{m} \mathbb{E}_{v_n\sim P_{neg}(v_n)}[\sigma(-{\varphi'_n}^T \varphi_i)(-\varphi'_n)]] \quad (7)$$

$$\frac{\partial O_3}{\partial \varphi_i} = \frac{\alpha}{K}\sum_{K=1}^{K} \pi_{ik}\Sigma_k^{-1}(\varphi_i - \mu_k) \quad (8)$$

Gradient the objective function to φ'_j, we have

$$\frac{\partial O_2}{\partial \varphi'_j} = -\sum_{v_i\in V}[\delta(v_j \in C_i)\sigma({\varphi'_j}^T \varphi_i)\varphi_i$$
$$+\sum_{t=1}^{m} \mathbb{E}_{v_n\sim P_{neg}(v_n)}[\delta\left(v_n = v_j\right)\sigma\left(-{\varphi'_n}^T \varphi_i\right)\left(-\varphi_j\right)]] \quad (9)$$

Denote that $\Phi = \{\varphi_i\}$ and $\Phi' = \{\varphi'_i\}$, finally, the node embedding Φ and context embedding Φ' are obtained by ASGD.

B. Community Detection

After training network embedding, all the nodes are represented as low-dimensional vectors. Then, next step is to consider how to find communities based on the embedding result. When training the embedding to preserve the high-order proximity, the embedding is fitted into a Gaussian mixture model, while Gaussian mixture model is just a soft clustering method. Assume that there are K communities, then node v_i belongs to community C_k who has the maximum probability among the Gaussian mixture model. So the community detection result can be gotten by (10) at the same time.

$$v_i \in C_k, C_k = \arg\max_k \mathcal{N}(\varphi_i|\mu_k\,, \Sigma_k) \quad (10)$$

where $\mathcal{N}(\varphi_i|\mu_k\,, \Sigma_k)$ is a multivariate Gaussian distribution with mean vector μ_k and covariance matrix Σ_k.

C. Weakening Community Structure

According to the previous hypothesis, the embedding of all nodes is generated by a Gaussian mixture model. By calculating the nodes' probability of belonging to different communities, the nodes are divided into the community with the largest probability. In view of this, a straightforward way to weaken community structures is to reduce the probability of each node belonging to its original community (that is, increase the distance between the node's embedding vector and the mean vector of its original community). The formula for weakening community structure is given as follow

$$\varphi_i := \varphi_i + \lambda(\varphi_i - \mu_{k|\varphi_i\in C_k}) \quad (11)$$

where λ is the offset value, it indicates the step of φ_i far away from community C_k. By reducing the probability of each node belonging to its original community, the probability of nodes belonging to other undiscovered communities can be increased. In this way, the detected community structures are weakened, and the community structures with low significance can be found.

D. Summary

The inference algorithm of MGCD is summarized in **Algorithm 1**. In lines 1-2, the alias table and negative table are initialized for alias sampling and negative sampling. In lines 3-5, the node embedding is updated by first-order proximity. In lines 6-9, the node embedding and context embedding are updated by second-order proximity. The multi-granularity community structures are found in lines 10-22. In line 11, a Gaussian mixture model is fitted for community detection and embedding. In lines 12-14, the node embedding is updated by high-order proximity. In line 15, the community structure at current granularity is obtained. In lines 16-21, the detected community structure is weakened. All above is the details of our algorithm.

V. EXPERIMENTS

In this section, MGCD is tested on several datasets with different offset values and the experimental results are analyzed. The code used during the experiments is provided at the follow link[1].

A. Datasets

The Facebook datasets on SNAP[2] are selected to evaluate our model. Facebook is an online social networks, edges represent interactions between people. Each node has many labels, indicating the features of nodes.

The nodes are divided into three granularities according to different attributes (Granularity 1: employer, people in the same community has the same employer. Granularity 2: school, people in the same community used to study in the same school. Granularity 3: work start date, people join the company on the same day.), and each granularity has different significance. The details are shown in **Table 1.**

[1] https://github.com/chxgong/MGCD.git
[2] http://snap.stanford.edu/

TABLE II. MULTI-GRANULARITY COMMUNITY DETECTION RESULTS.

	offset value	λ = 0.5			λ = 1.5			λ = 2.5			λ = 3.5			λ = 4.5		
	granularity	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
0.edges	**NMI (%)**	22.8	**32.3**	**38.2**	21.8	30.2	35.3	23.9	29.1	35.1	21.0	29.5	36.3	**24.6**	30.0	34.9
	Micro-F1(%)	20.4	**17.7**	**19.2**	21.3	16.2	20.7	22.5	14.4	22.5	25.5	18.6	24.6	**24.9**	16.8	25.2
107.edges	**NMI (%)**	25.6	**57.3**	**33.8**	**27.3**	56.9	33.6	26.2	57.1	33.4	26.9	57.1	30.8	26.7	56.8	31.2
	Micro-F1(%)	8.6	**14.7**	**9.5**	**8.4**	14.0	8.4	8.2	13.5	14.0	7.9	14.5	19.9	8.0	13.3	20.3
348.edges	**NMI (%)**	14.2	51.4	12.0	13.5	52.9	12.3	13.7	52.1	11.8	**14.8**	52.5	6.8	13.4	**53.6**	**12.8**
	Micro-F1(%)	38.4	15.2	38.4	42.9	20.1	29.9	34.8	13.8	23.2	**45.1**	20.5	45.1	35.3	**16.1**	**45.1**
414.edges	**NMI (%)**	4.1	58.7	5.0	**4.57**	57.2	8.5	3.4	**59.8**	7.6	3.6	59.6	8.4	4.6	59.4	**12.1**
	Micro-F1(%)	72.7	21.3	42.7	**80.7**	21.3	45.3	68.3	**22.7**	40.0	76.7	23.3	52.0	72.0	18.7	**61.3**
686.edges	**NMI (%)**	**7.13**	25.0	1.75	2.0	24.4	0.2	2.4	**27.5**	**3.5**	3.1	25.6	2.0	3.4	22.9	0.3
	Micro-F1(%)	**70.8**	29.2	86.9	82.1	31.6	86.9	84.5	**32.1**	**89.9**	87.5	27.8	85.1	85.7	25.6	88.1
1684.edges	**NMI (%)**	**9.1**	33.4	25.0	9.0	33.4	**26.3**	8.3	32.7	25.3	8.4	33.7	25.0	9.2	**33.9**	25.8
	Micro-F1(%)	**21.8**	10.1	9.9	18.6	10.4	**10.4**	16.2	9.4	8.9	16.3	9.7	10.1	17.7	**10.9**	13.6
1912.edges	**NMI (%)**	**23.9**	67.5	**34.0**	22.7	**67.8**	32.9	23.4	67.4	31.8	22.9	67.5	29.5	23.4	67.1	30.0
	Micro-F1(%)	**7.6**	16.3	**9.1**	7.9	**17.3**	9.1	7.5	17.3	12.2	10.2	15.9	25.3	9.0	14.9	24.0
3437.edges	**NMI (%)**	16.0	33.2	13.5	16.8	**33.2**	**15.6**	**17.2**	32.8	12.7	16.6	32.9	13.8	15.6	33.0	14.1
	Micro-F1(%)	19.7	13.5	19.7	18.9	**14.0**	**15.9**	**18.4**	13.7	21.7	15.2	13.3	23.0	19.3	13.7	26.4

TABLE III. MULTI-GRANULARITY COMMUNITY DETECTION RESULTS

	Dataset	**0.edges (λ = 0.5)**			**107.edges (λ = 0.5)**			**348.edges (λ = 4.5)**			**414.edges λ =4.5**		
	Granularity	***1***	***2***	***3***	***1***	***2***	***3***	***1***	***2***	***3***	***1***	***2***	***3***
HC:MOD	**NMI (%)**	20.4	32.2	25.7	21.5	39.2	17.6	**17.5**	34.8	12.3	**5.2**	38.8	10.1
	Micro-F1(%)	12.8	9.5	10.7	**29.8**	6.9	**25.7**	25.0	9.8	23.7	48.3	17.5	46.9
MGCD	**NMI (%)**	**22.8**	**32.3**	**38.2**	**25.6**	**57.3**	**33.8**	13.4	**53.6**	**12.8**	4.6	**59.4**	**12.1**
	Micro-F1(%)	**20.4**	**17.7**	**19.2**	8.6	**14.7**	9.5	**35.3**	**16.1**	**45.1**	**72.0**	**18.7**	**61.3**
	Dataset	**686.edges (λ = 2.5)**			**1684.edges (λ = 1.5)**			**1912.edges (λ = 0.5)**			**3437.edges (λ = 1.5)**		
HC:MOD	**NMI (%)**	**11.9**	21.8	2.3	**9.9**	28.9	13.9	13.8	40.5	14.7	12.3	28.4	11.7
	Micro-F1(%)	28.1	16.8	28.7	**23.9**	**18.3**	**26.4**	**25.3**	7.6	**24.2**	16.0	9.4	18.2
MGCD	**NMI (%)**	2.4	**27.5**	**3.5**	9.0	**33.4**	**26.3**	**23.9**	**67.5**	**34.0**	**16.8**	**33.2**	**15.6**
	Micro-F1(%)	**84.5**	**32.1**	**89.9**	18.6	10.4	10.4	7.6	**16.3**	9.1	**18.9**	**14.0**	**15.9**

B. *Evaluation Metrics*

In this paper, we use both *normalized mutual information* (NMI) [3] and *micro-f1* [2] to evaluate our model. NMI is used to measure the closeness between predicted communities and ground truth communities. Micro-f1 is used to evaluate the accuracy of predicted communities.

C. *Baselines*

HICDOE is the state-of-art method for finding hidden community structure in social networks. HICODE proposed three community weakening methods: remove edge, reduce edge, and reduce weight for identifying the hidden community structure by using the existing community detection algorithm iteratively. HICODE are used as the benchmark method with reduce edge as community weakening strategy and Louvain method as basic community detection method, which is proved to be the best method for finding hidden community structures by now.

D. *Parameters and Environment*

In the experiment, the negative sampling number $\boldsymbol{m}$ is 100,000,000, the embedding dimension $\boldsymbol{d}$ is 28, the number

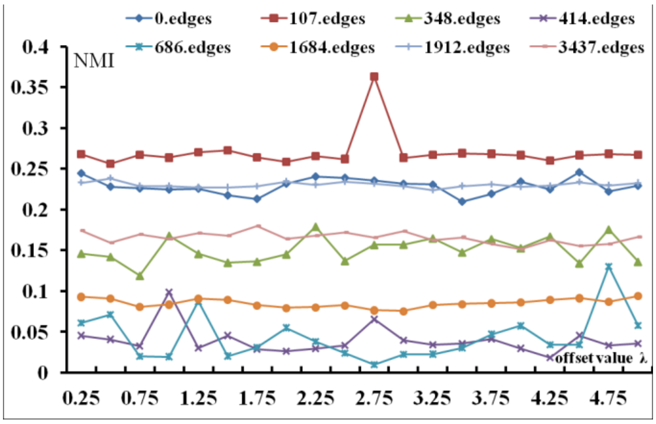

Fig. 2. **NMI results with different λ on granularity 1.**

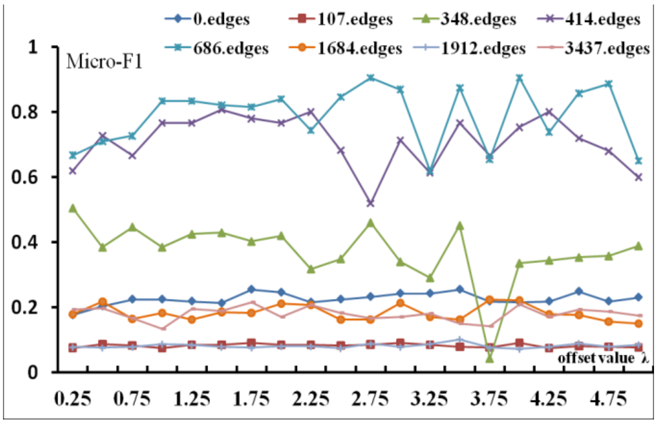

Fig. 3. **Micro-F1 results with different λ on granularity 1.**

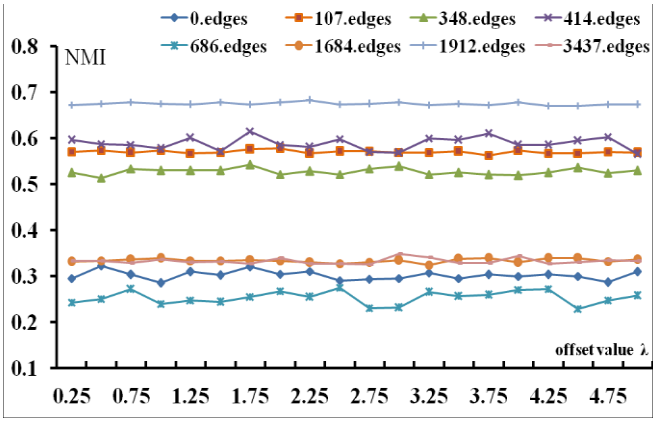

Fig. 4. **NMI results with different λ on granularity2.**

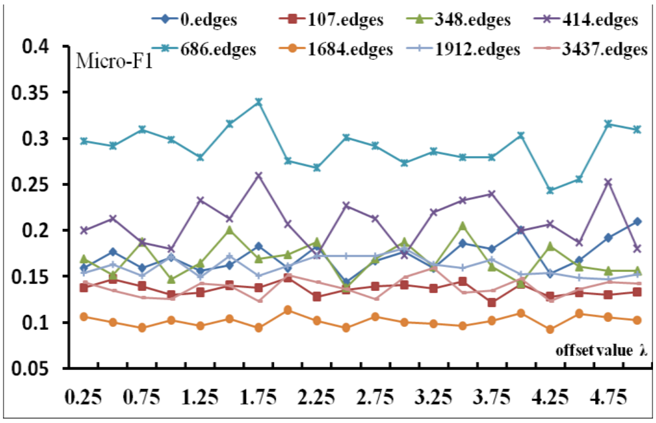

Fig. 5. **Micro-F1 results with different λ on granularity2.**

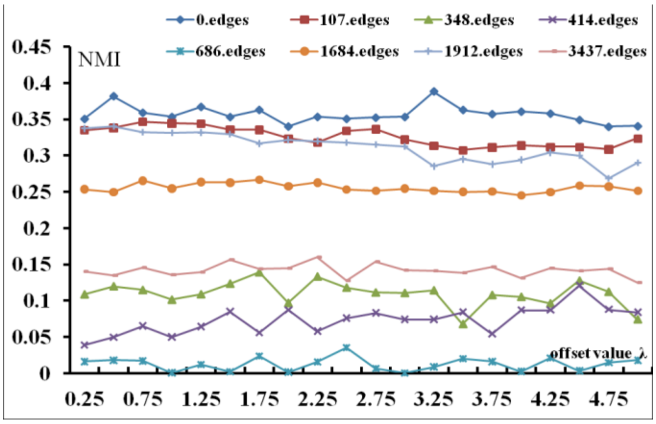

Fig. 6. **NMI results with different λ on granularity 3.**

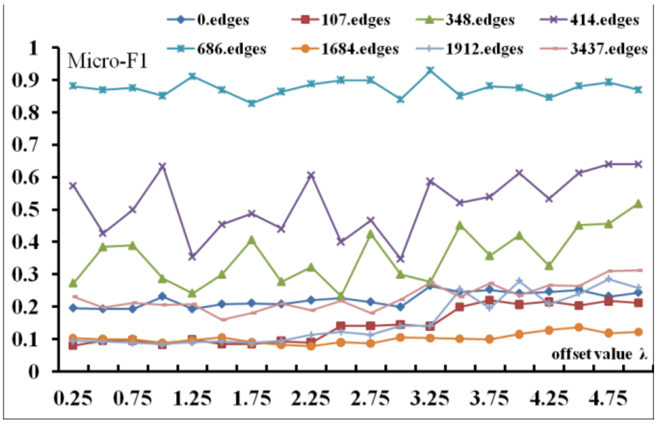

Fig. 7. **Micro-F1 results with different λ on granularity 3.**

of sampling $\boldsymbol{S}$ is 100,000. For offset value $\boldsymbol{\lambda}$, in order to compare the impact of different offset values on the experimental results, we set $\boldsymbol{\lambda}$ from 0.25 to 5, 0.25 per step. All the experiments are conducted on Windows with Intel(R) Core(TM) i5-4590 3.30GHz CPU and 16GB memory.

E. Result Analysis

Because all records are occupied too much space, the experimental results of λ from 0.5 to 4.5, 1.0 per step are recorded in **Table 2**. Additionally, the NMI and Micro-f1 results (with λ from 0.25 to 5.0, 0.25 per step) are plotted in **Figures 2-7**. The results of contrast experiment with HICODE are recorded in **Table 3**.

For granularity 1, because of colleague relationship is the more significant relationship in social network, people serve in the same company are divided into the same community. Theoretically, the value of λ should have no impact on the result of community partition on this granularity. That is, NMI and micro-F1 should be equal with different λ. However, the results show that the NMI and micro-f1 are not equal with different λ. The reason is that the EM algorithm is used to fit a GMM when detecting community structures. Nevertheless, there will be errors at each time of fitting the GMM, resulting in the fluctuation of the final result. For granularity 2, people who once studied in the same school are divided into the same community. For dataset 1912.edges with λ=1.5, the NMI is as high as 0.681, which is a favorable result. For granularity 3, people who join the company on the same day are divided into the same community. It is hard to find this kind of community structure by traditional community detection methods. Our model can obtain this kind of community structure. For example, for dataset 686.edges with λ=2.5, the micro-f1 is as high as 0.899. In **Table 3**, on the majority of datasets, our method has positive performance both in NMI and micro-f1 compared with HICODE.

According to **Figure 2** and **Figure 3**, it can be found that the values of λ are different when NMI and micro-F1 get the maximum value. The reason is that different datasets have different network structures. For example, if there is a large distance between two different communities in a dataset, the smaller λ cannot get a better weakening effect. So we cannot give an optimal λ selection suggestion, one need to select an appropriate λ based on the dataset.

VI. Conclusions

In this paper, the multi-granularity community structure and the significance of a community partition are defined. Additionally, a novel approach named MGCD is proposed for finding the multi-granularity community structure of social networks. MGCD provides a more comprehensive view for understanding the relationships among communities in complex social networks. Experimental results on real-world networks show that MGCD outperforms the state-of-the-art benchmark methods on finding multi-granularity community structure.

This paper uses the network embedding method as the network representation method. It can overcome the shortcomings of the traditional representation method, but may not be able to maintain all the characteristics of the original network. In addition, the community structure weakening strategy may result in errors while fitting Gaussian mixture model, which may lead to poor results. In the future, we plan to use a new network embedding method to maintain more information in the original network, and explore new community weakening strategy in order to obtain better results.

References

[1] K. He, Y. Li, S. Soundarajan, and J. E. Hopcroft, “Hidden community detection in social networks,” *Information Sciences*, vol. 425, pp. 92–106, 2018.

[2] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.

[3] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu, “Learning deep representations for graph clustering.” in *Proceedings of the 28th AAAI Conference on Artificial Intelligence,* 2014, pp. 1293–1299.

[4] D. Wang, P. Cui, and W. Zhu, “Structural deep network embedding,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 1225–1234.

[5] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.

[6] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, “Line: Large-scale information network embedding,” in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1067–1077.

[7] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, “Asymmetric transitivity preserving graph embedding,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 1105–1114.

[8] P. Cui, X. Wang, J. Pei, and W. Zhu, “A survey on network embedding,” *IEEE Transactions on Knowledge & Data Engineering*, 2017, pp.1-1.

[9] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” *Physics Reports*, vol. 659, pp. 1–44, 2016.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[11] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, “Finding statistically significant communities in networks,” *PloS one*, vol. 6, no. 4, pp. e18961, 2011.

[12] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, “Link communities reveal multiscale complexity in networks,” *nature*, vol. 466, no. 7307, pp. 761, 2010.

[13] M. Rosvall and C. T. Bergstrom, “Maps of random walks on complex networks reveal community structure,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 4, pp. 1118–1123, 2008.

[14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, pp. P10008, 2008.

[15] S. Cavallari, V. W. Zheng, H. Cai, K. C.-C. Chang, and E. Cambria, “Learning community embedding with community detection and node embedding on graphs,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 377–386.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[17] B. Recht, C. Re, S. Wright, and F. Niu, “Hogwild: A lock-free approach to parallelizing stochastic gradient descent,” in *Advances in neural information processing systems*, 2011, pp. 693–701.

[18] G. Wang, “DGCC: data-driven granular cognitive computing,” *Granular Computing*, vol. 2, no. 4, pp. 343–355, 2017.

[19] M. E. Newman, “Modularity and community structure in networks,” *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[20] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” *Nature*, vol. 435, no. 7043, pp. 814, 2005.

[21] J. J. Whang, D. F. Gleich, and I. S. Dhillon, “Overlapping community detection using seed set expansion,” in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 2099–2108.

[22] M. E. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Physical review E*, vol. 69, no. 2, pp. 026113, 2004.

[23] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. EliassiRad, “Collective classification in network data,” *AI magazine*, vol. 29, no. 3, pp, 2008.

[24] X. Wang, P. Cui, J. Wang, J. Pei, W. Zhu, and S. Yang, “Community preserving network embedding.” *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, 2017, pp. 203–209.

[25] I. Herman, G. Melanc¸on, and M. S. Marshall, “Graph visualization and navigation in information visualization: A survey,” *IEEE Transactions on visualization and computer graphics*, vol. 6, no. 1, pp. 24–43, 2000.

[26] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.