Iterative mining algorithm based on overlapping communities in social networks

Jiayin Feng, Jinling Song, Dongyan Jia, and Limin Shen

***Abstract*—**Social networks with complex structure and large scale emerged with the development of social network sites. Various network communities gradually form complex structural pattern in the production and living of people. The competitive advantages and community distribution in networks can be obtained through analyzing the structure of community. Therefore the research key point of the current data mining field is how to find out the potential structure of such large-scale social networks. Currently, most of real networks have overlapping communities. All the users can be allocated to different communities according to different allocation rules. But the complex structure of network and mass node information are difficulties for mining large-scale social network communities. Based on the discussion of relevant theories of complex network and mining algorithm, this study summarized and analyzed several algorithms for mining overlapping communities and put forward a high-efficient and effective overlapping community iterative mining algorithms. Moreover, experiments were carried out to verify its effectiveness and high efficiency. This work provides an improvement direction of relevant technologies for researchers who engage in network community data mining.

***Keywords*—**complex network, community mining, mining algorithm, overlapping community

I. Introduction

MANY structures in the real world can be represented by simple diagrams, for example, geographic information networks, and the internal connections can be easily discovered by combining different simple diagrams. Such a large structure with complex functions is often called complex network. Complex network is a traditional subject has extremely high research values. Many subjects including but not limited to physics and informatics have studied complex network at the beginning of the last century. But limited by hardware technology in the last century, mass data obtained from complex network are difficult to be efficiently counted and recorded, let alone analysis. Study on complex network was confined within graphic surface. For example, Euler's seven bridges problem is the origin of graph theory analysis [16].

Human as a social creature has the attribute of sociality which has a great influence on the civilization of human society. Each person stands for one node and the connection between people is expressed by side; such as a combination pattern forms social network [20]. With the development of computer technology and the popularity of the Internet, communication between people has become easier. People can communicate with each other without face to face. Under the support of the Internet, new communication tools such as QQ and WeChat have emerged, which not only facilitates communication, but also makes the gathering of people easier. Moreover some people established different social networking sites after discovering the commercial opportunity underlying communication. It is particularly important for the managers of these websites to understand the potential interests of the users. Recommendation system is an algorithmic tool which can offer proposals for the potential interests of users.

Community structure has become an important technical research subject in complex network. Although many community detection algorithms have been proposed, most of them concentrate on independent community structures. However, communities often overlap each other in many real network structures. Therefore, it is necessary to develop overlapping community discovery algorithms [18].

With the emergence of social network sites, social network has become more complex and larger. Social network is featured by community structure, i.e. different points are closely connected to from group or colony through certain relation. The academic world defines it as community. In recent years, Fortunato [1] and Coscia [2] investigated independent and overlapping communities and compared mining algorithms. The difference was that Fortunato studied based on the principles of algorithms and Coscia studied different definitions of community. Xie et al. [3] divided algorithms into five categories, i.e. local expansion and optimization algorithm, side segmentation algorithm, clique percolation algorithm, fuzzy

This study was supported by the National Natural Science Foundation of China under Grant No. 61772450 (Research on software behavior analysis and judgement based on process algebra and coordination trust for mobile computing), Hebei Province Natural Science Foundation of China under Grant No. F2017203307 (Research on method and technology of software monitoring based on behavior for mobile computing protection) and the Science and Technology Plan of Hebei Province under Grant No. 17210701D (Transparent monitoring and protection technology of data and development of cloud service platform).

J. Y. Feng, J. L. Song and D. Y. Jia are with Hebei Normal University of Science & Technology, Hebei 066004, China (e-mail: feng_ada2001@163.com)

L. M. Shen is with School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei, 066004, China.

detection algorithm and agency and dynamics based algorithm [4]. They made experimental analysis on 14 algorithms. Wang and Zhang [5] analyzed four overlapping community mining algorithms which were frequently used in recent years, i.e. SVINET, UEOC, SLPA and TopGC, explored the principles of the algorithms, summed up their application characteristics and scope, and moreover applied them in multiple large-scale social network.

II. Overlapping community mining algorithm

Complex network [6] refers to network with part of or all properties including self-organization, self-similarity, attractor, worldlet and scale free. In short, it refers to network with high complexity, i.e. structural complexity, network evolution, connection diversity, dynamical complexity, node diversity and multiple complexities.

As studies on the properties of complex network went deeper, an aggregation phenomenon has been discovered in most real networks, which is called community structure. Community structure in complex network has been extensively concerned by researchers for its modularity and heterogeneity [19]. Community is the set of individuals with the same characteristics. Complex network communities are overlapped. Nodes which belong to multiple communities in network are called overlapping nodes [7]. Communities with overlapping nodes are called overlapping communities. Overlapping communities usually plays a large and even a key role in complex network structure.

A. Multi-label propagation

Label propagation algorithm, the basis of all label based algorithms [8, 9], is simple and effective. The only label corresponding to each node in network was iterated to make them the most carried label; then label nodes are divided after stabilization to obtain communities containing same label nodes. But the algorithm is too simple that it can only mine non-overlapping community. Therefore the algorithm was improved and some new algorithms such as COPRA and SLPA were obtained.

The improved multi-label propagation algorithm can mine overlapping community. The nodes of the algorithm can own multiple labels. Multiple labels which each node is corresponding to in network are iterated to be the label of all neighbors to obtain evaluation on degree of community overlapping [10]. But the algorithm has a defect, i.e. uncertainty. It is easy to cause label assimilation and result in many same communities.

B. Cique percolation

Cique percolation method as the earliest overlapping community mining algorithm was developed based on cique percolation theory. The algorithm considers that community is the set of fully connected subgraph sharing nodes [11]. Cique percolation can be used to identify community structure in network, search for all complete subgraphs containing k nodes, and construct new graph taking k-clique as nodes. If two nodes had (k-1) public nodes, then a side will be established between the nodes; each connection subgroup is a community. The algorithm also has defects. As the selection of k value can produce large impacts on mining results, the application of the algorithm had large limitation in reality.

Cique percolation method also has corresponding improved algorithms such as weighed Cique percolation method and speedy cique percolation method. But Cique percolation method is more suitable for solving search problems in complex network rather than mining community.

C. Local optimization

The basic idea of local optimization based mining algorithm is to divide communities based on the local information of network. Such an algorithm is the most common and has achieved great achievements. Initial nodes which need optimization are confirmed; then nodes with the largest measurement increment near initial nodes are merged according to the measurement standards of community structure. The expansion of different communities is independent; different communities realize mining of community structure through absorbing and merging the same node.

III. Design of overlapping community mining algorithm

A. Putting forward an algorithm

Previous studies concerning community mining were carried out under the premise that communities were not overlapped. But with the constant advancement of community mining technologies, many evidences suggest that different communities are intersected in real network. Real network is a complex.

Real network is a complex network with a large scale, real-time dynamic updating and a large number of points of intersection. Traditional community mining technology which takes non-intersected communities as the precise will need a large amount of resources and cause heavy calculation when collecting global topological information. Hence a local information based overlapping community mining algorithm is proposed. Real network can be regarded as overlapped local community structure. The algorithm has extremely high operation algorithm and can carry large calculated quantity when mining overlapping community structure.

Local community mining is realized by obtaining the most suitable community structure through merging initial nodes and neighbour nodes and covering all the nodes in the whole network through local community mining. In this way, the local community of a node can be obtained based on the network information of local network structure around initial nodes. Hence the method has high operation efficiency. Algorithms based on the method include LFM algorithm and GCE algorithm.

B. Design of algorithm

1. Design philosophy

Traditional algorithms select initial seed nodes randomly. As a result, the accuracy of the final community mining results relies heavily on the selection of initial seed nodes. To improve the quality and stability of mining, local central nodes can be selected as the initial seed nodes. But the maximum node degree does not mean the largest impact. Mining taking points with the largest impact as the center can obtain community structure more accurately. Fitness function used in mining algorithm can play a key role; however traditional local fitness function does not take the features of communities into account, i.e. whether node degree is equivalent to impact.

To solve the above problems, an improved algorithm, the core node based overlapping community mining algorithm (COCMA), was proposed in this study. The algorithm only needs to do fuzzy clustering according to node similarity rather than specify undeterminable value of k like the traditional community mining algorithm. The specific algorithm flow is shown in Figure 1.

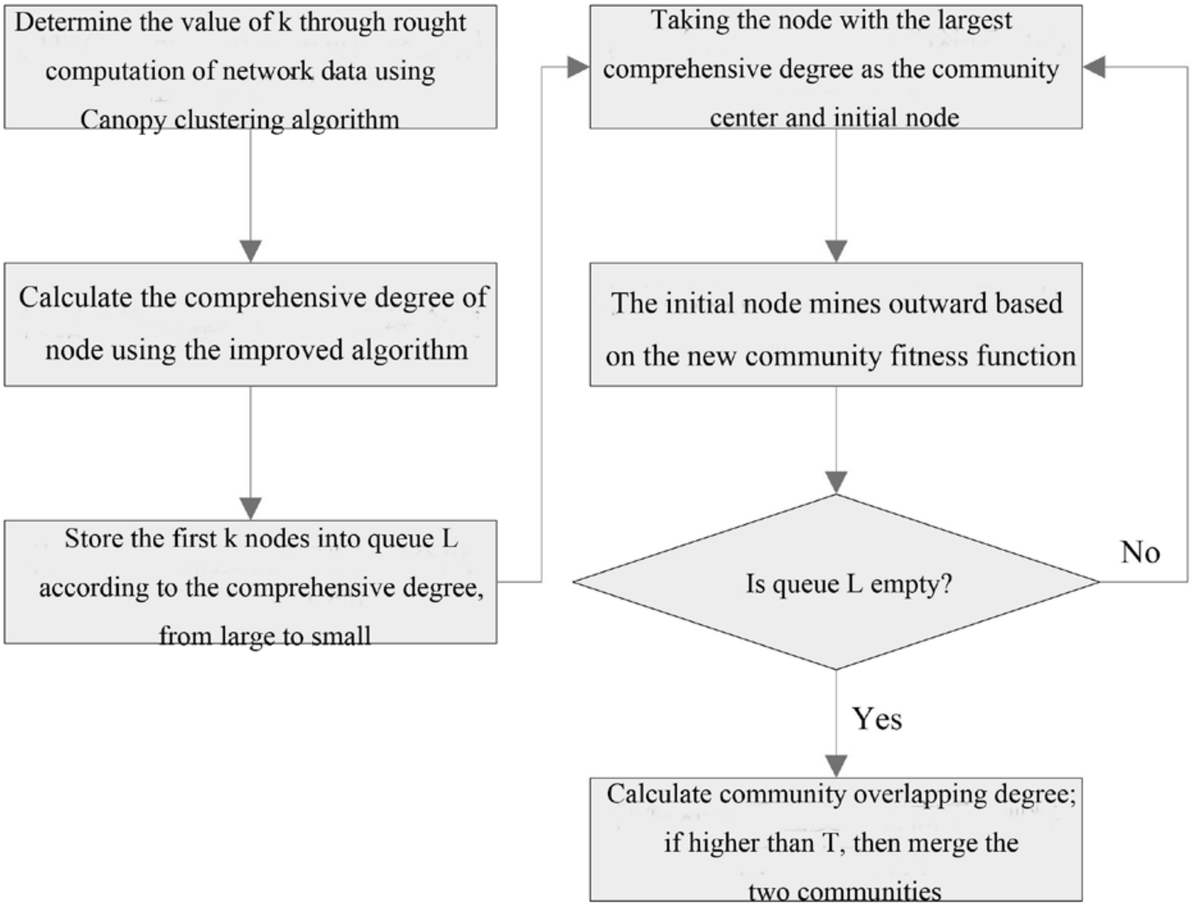
Determine the value of k through rought computation of network data using Canopy clustering algorithm
Calculate the comprehensive degree of node using the improved algorithm
Store the first k nodes into queue L according to the comprehensive degree, from large to small
Taking the node with the largest comprehensive degree as the community center and initial node
The initial node mines outward based on the new community fitness function
Is queue L empty?
No
Yes
Calculate community overlapping degree; if higher than T, then merge the two communities

Fig. 1 The specific flow of the core node based overlapping community mining algorithm

2. Selection of core nodes

Nodes in complex network are individuals. Communities in network usually take several core nodes with great influence as the center. Communities in complex network are the set of core nodes and surrounding neighbour nodes. Therefore community mining can be rapid and accurate from the aspect of community core nodes; otherwise it will cause huge calculated quantity and non-ideal results. Therefore finding out core nodes is the key.

Core nodes can be found through measuring the influence of nodes including betweenness centrality, closeness centrality [13] and degree centrality [14].

The calculation of betweenness centrality and closeness centrality needs the shortest route between two nodes which is difficult to obtain; hence degree centrality was selected. The influence of a node is correlated to the size of node degree. Therefore it was defined as:

$$C_D(v) = \sum_{w\in neighbors(v)}^{N} s_{vw} = p_v^{out}$$

where s_{vw} stands for the element in network adjacent matrix and p_v^{out} stands for the out degree of node v.

In real network, nodes usually do not have enough degree and influence; hence they cannot become core nodes. But its surrounding neighbor nodes have great influence. There are also nodes with large degree and great influence, but the influence of its surrounding neighbor nodes is quite small. But such nodes are not core nodes. Therefore the common influence of nodes and surrounding nodes should be considered together.

Comprehensive scale of nodes was defined as following through improving degree centrality.

$$C_{DD}(v) = C'_{DD}(v) + C''_{DD}(v) = p_v^{out} + \sum_{w \in neighbors(v)} p_w^{out}$$

where $C'_{DD}(v)$ stands for the influence of node v and $C''_{DD}(v)$ stands for the comprehensive influence.

3. Mining local community

Community fitness stands for the tightness of community. Therefore fitness function needs to be defined when initial nodes are determined as community core for overlapping community mining. Though common fitness function is simple and direct, the influence of community structure density. Therefore it cannot successfully mine community with all structures, which is easy to cause omission. Therefore it was defined as following through improving fitness function.

$$F_C = \alpha \frac{p_{in}^C}{N_C(N_C - 1)} + (1-\alpha)\frac{p_{in}^C}{p_{in}^C + p_{out}^C}$$

where N_C stands for the number of nodes in community and α stands for controllable parameter (0 < α ≤ 0.5). The value of α was 0.2 in this study.

After community core nodes were selected using the aforementioned algorithm, the above method was used to screen and absorb the neighbor nodes of core nodes to realize mining of local community.

4. Merging overlapping community

As nodes in complex community were overlapped, mining local community may cause severe or complete overlapping of communities [15], as shown in Fig. 2. Therefore the highly overlapped communities should be merged.

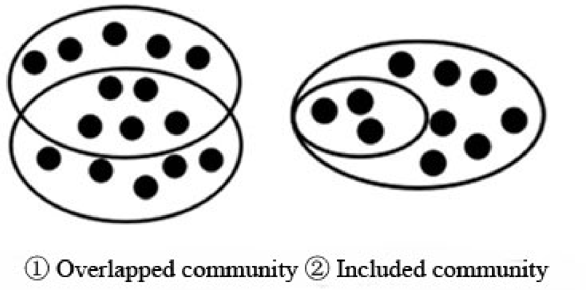

Fig. 2 The overlapped and included communities

As communities had different sizes, the degree of community overlapping could not be defined based on the proportion of the overlapped nodes. Therefore degree of overlapping should be defined using the following method.

$$Overlap(C_v, C_w) = \frac{|C_v \cap C_w|}{MIN(C_v, C_w)}$$

where C_v and C_w stand for two overlapped communities, numerator $|C_v \cap C_w|$ stands for the number of nodes included by the overlapped part, and denominator $MIN(C_v, C_w)$ is the number of nodes included in the smaller community among two communities.

IV. Experiments and results

In real network, data set is usually massive, complicated and changeable. Therefore requirements on the efficiency and accuracy of algorithm are high. This study selected three real networks to test the practicability of the designed mining algorithm. The scale of data sets in real network is shown in Table 1.

The designed algorithm was compared with LFM, GCE and CPM, and the results are shown in Table 2.

Table 2 shows that COCMA algorithm had a stable performance and favorable mining effect; LFM had general and instable mining performance and was easy to be affected by the errors of initial node selection; GCE had better and more stable performance compared to LFM algorithm, but it was not as good as COCMA; CPM depended on the selection of k value which could directly determine mining results.

Table 1 Explanation for data sets in real networks

Name	Type	Nodes	Edges	Communitie s	Description
email-Eu-core	Directed,	1, 005	25, 571	42	E-mail network

	Communities				
com-Amazon	Undirected, Communities	334,863	925,872	75,149	Amazon product network
com-DBLP	Undirected, Communities	317,080	1,049,866	13,477	DBLP collaboration network

Table 2 Comparison of mining results

	Modularity extension function				
Network	COCMA	LFM	GCE	CPM	
				k=3	k=5
email-Eu-core	0.5167	0.2467	0.4065	0.1687	0.3774
com-Amazon	0.5873	0.3821	0.4647	0.3648	0.4691
com-DBLP	0.5291	0.3159	0.4184	0.4038	0.4952

Through comparison, it was found that COCMA had excellent comprehensive performance and was a good overlapping community iterative mining algorithm.

V. Conclusion

With the development of society, the progress of computer technology and the popularization of the Internet, connections between people have become more and more frequent. In the past, social networks were not complex because of the limitation of communication devices, and the overlapping of different social networks was not obvious. But social networks become complicated because of the emergence of the Internet based social networking sites. Moreover researchers gradually discover the overlapping between community networks as community mining algorithms develop.

The traditional non-overlapping community mining algorithms have been not able to accurately and efficiently mine community network because of its complexity and overlapping. In recent years, various overlapping community mining algorithms have been put forward. This study focused on the improvement of algorithm accuracy and computation speed. The development history and characteristics of complex network were introduced briefly, and the problems existing in the traditional mining algorithms were analyzed. Moreover the role of nodes in networks was described, and a new algorithm called COCMA was proposed.

In this paper, three mining algorithms, COCMA, LFM and GCE were analyzed and compared. COCMA is an improved algorithm. On the basis of local optimization mining algorithm, the algorithm of degree centrality was improved. It mined overlapping communities in social networks through mining local communities and merging overlapping communities. The experiment showed that the performance of COCMA was quite stable in three real networks, and the mining effect of the algorithm was relatively good; the mining effect of LFM was general and not stable, and it was easy to result in the difference in the mining effect because of the error of the initial node selection; the GCE algorithm was more stable and better than the LFM, but there was a gap with COCMA; CPM algorithm depended on the selection of k value, and whether the K value was proper or not directly determined the result of mining. The above findings revealed that the improved algorithm, COCMA, with excellent performance in all aspects was a good overlapping community iterative mining algorithm.

Through comparison and analysis, it was found that COCMA had a favorable mining effect; however the algorithm had some deficiencies. Firstly the algorithm focused on un-weighed and undirected networks and made no discrimination on all the nodes. Notes in real networks are treated differently. Secondly the distributed storage of data in complex network will be conflict with the current algorithm.

Studies concerning overlapping communities in social network remain to be further explored and improved. Besides improving the current exploration mode, we can also apply new methods in studies on overlapping community, which has great practical significance.

References

[1] S. Fortunato, “Community detection in graphs,” Physics Reports, vol. 486, no. 3–5, pp. 75-174, 2010.

[2] M. Coscia, F. Giannotti, D. Pedreschi, “A classification for community discovery methods in complex networks,” Statistical Analysis & Data Mining the Asa Data Science Journal, vol. 4, no. 5, pp.512-546, 2011.

[3] J. Xie, S. Kelley, B.K. Szymanski, “Overlapping community detection in networks: The state-of-the-art and comparative study,” Acm Computing Surveys, vol.45, no.4, pp.1-35, 2011.

[4] M. Herrlich, B. Waltherfranks, R. Schröderkroll, J Holthusen, R Malaka, “Proxy-Based Selection for Occluded and Dynamic Objects,” International Conference on Smart Graphics. Springer-Verlag, pp.142-145, 2011.

[5] L.D. Wang, Y. Zhang, “Overlapping Community Detection in Large-scale Social Networks,” Journal of Hangzhou Normal University (Natural Science), vol.15, no.3, pp.331-336, 2016.

[6] Y. Mi, L. Zhang, X. Huang, Y. Qian, G. Hu, X. H. Liao, “Complex networks with large numbers of labelable attractors,” Epl, vol.95, no.5, pp.58001, 2011.

[7] D. He, D. Jin, Z. Chen, W. Zhang “Identification of hybrid node and link communities in complex networks,” Scientific Reports, vol.5, 2013.

[8] Z. Lin, X. Zheng, N. Xin, D. Chen, "CK-LPA: Efficient community detection algorithm based on label propagation with community kernel," Physica A Statistical Mechanics & Its Applications, vol.416, no.C, pp.386-399, 2014.

[9] L. Xia, L. Zhang, L. Guo, Y. S. Zhang, J. P. Zhang, J. Yang, "Applied research of node similarity label propagation in social networks," Computer Engineering & Applications, vol.50, no.14, pp.103-109, 2014.

[10] L. Wang, "An overlapping community mining algorithm based on the multi-label propagation," Journal of Yunnan Minzu University (Natural Science Edition), vol.24, no.3, pp.252-256, 2015.

[11] A. Choudhary, "Fast Algorithms for the Maximum Clique Problem on Massive Graphs with Applications to Overlapping Community Detection," Internet Mathematics, vol.11, no.4-5, pp.421-448, 2015.

[12] D. Prountzos, K. Pingali, "Betweenness centrality," ACM Sigplan Notices, vol.48, no.8, 2013.

[13] K. Wehmuth, A. Ziviani, "Distributed assessment of the closeness centrality ranking in complex networks," Computer Networks, vol. 57, no.13, pp.2536-2548, 2013.

[14] P. Bródka, K. Skibicki, P. Kazienko, K. Musiał, "A degree centrality in multi-layered social network. International Conference on Computational Aspects of Social Networks," IEEE, pp.237-242, 2012.

[15] S. Choi, J.Y. Park, H.W. Park, "Using social media data to explore communication processes within South Korean online innovation communities, " Scientometrics, vol. 90, no.1, pp.43-56, 2012.

[16] S. Zhang, R. S. Wang, X. S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy c c mathContainer Loading Mathjax -means clustering," Physica A Statistical Mechanics & Its Applications, vol. 374, no. 1, pp. 483-490, 2007.

[17] H. Li, D. Wu, W. Tang, N. Mamoulis, "Overlapping Community Regularization for Rating Prediction in Social Recommender Systems," ACM Conference on Recommender Systems, pp. 27-34, 2015.

[18] X. Wen, W. N. Chen, Y. Lin, T. L. Gu, H. X. Zhang, Y. Li, Y. L. Yin, J. Zhang, "A Maximal Clique Based Multiobjective Evolutionary Algorithm for Overlapping Community Detection," IEEE Transactions on Evolutionary Computation, vol. PP, no. 99, pp. 1-1, 2016.

[19] R. Wang, H. Yang, H. Wang, D. Wu, "Social overlapping community-aware neighbor discovery for D2D communications," IEEE Wireless Communications, vol. 23, no. 4, pp. 28-34, 2016.

[20] Z X Wang, Z C Li, X F Ding, J. H. Tang, "Overlapping community detection based on node location analysis," Knowledge-Based Systems, vol. 105, pp. 225-235, 2016.