A Fast Method of Detecting Overlapping Community in Network Based on LFM

Yanan Li*, Zhengyu Zhu
College of Computer Science, Chongqing University ,Chongqing, China.

* Corresponding author. Tel.: +8615683859795; email: Yanan_lee@foxmail.com.
Manuscript submitted February 13, 2015; accepted June 10, 2015.
doi: 10.17706/jsw.10.7.825-834

Abstract: Detect overlapping communities efficiently and effectively in various social networks has been more and more important. Aiming at the high complexity of expanding strategy and the defect of generating many homeless nodes in LFM, we propose a quick algorithm based on local optimization of a fitness function(QLFM). The proposed algorithm firstly select a node as seed randomly .With a local fitness function ,the algorithm then will expand from inside to outside of the seed according to the Breadth-First-Search in graph. As different seeds will expand to different communities independently ,and these communities have same nodes ,thus our method can detect overlapping nodes quickly and efficiently. An empirical evaluation of the method using real and synthetic datasets shows that the method give better result not only in time efficiency, but also in quality aspect than other methods at the overlapping community detection.

Key words: Overlapping communities, social networks, detecting communities, community structure.

1. Introduction

In real world, networks are usually used to represent many complex systems. The nodes in networks represent the individuality of systems, and edges in networks represent relationship between different individualities in system. The study of networks shows that these networks always have some common properties, such as power law degree distribution, clustering and community structures. Community structures indicate that networks are made of many groups. Vertices in same group are much more connected to each other than those vertices in different groups. Community exits in many networks in real world, such as scientific collaboration network, organizations in social networks, protein complexes in biological networks. Knowledge of community structure can be useful to reveal functional organization in networks and to know more about feature of networks.

2. Relate Works

Many traditional community detecting methods hold that each node can only belong to one community, such as Modularity optimization[1], [2], Hierarchical clustering [3], [4], Spectral Algorithms [5], [6], label propagation algorithm[7], [8], Methods based on statistical inference [9]. However in some real networks, communities are not independent , nodes can belong to more than one community ,which will lead to overlapping communities . For example, a researcher may belong to more than one research group, or a protein may exist in multiple complex systems .Therefore, the identification of overlapping communities is of central importance.

Recently ,Many overlapping community detection algorithms have been developed .Clique percolation

method [10] (CPM)holds that community are form of a set of connected sub-graphs which own shared vertices.it use clique percolation to detect overlapping communities . Similar methods include[11], [12]; Methods based on label propagation [13] firstly allocated a single label to each node ,And then update the label and membership degree of each node according to its neighbor nodes .finally , nodes owning the same label will be divided to the same community ,and those nodes who has more than one labels will be overlapping nodes .Methods based on edges [14] build line graph in which the node comes from the edge in original graph .then it use a non-overlapping community detection method to process the line graph ,as one node can belong to several edges ,thus ,it detect overlapping communities successfully. Methods based using the fitness function (such as LFM [15], DEMON [16], OSLM[17]) assume that the communities are local structures, which comprise of the nodes of the modules themselves and the extension to the nodes in its neighborhood. This method identifies communities as sub graphs obtained by maximaization of a fitness measure.

Although detecting these overlapping communities has gained a lot of attention ,it is still a challenging task for researchers to study how to detect overlapping communities in such complex networks using a more quick and efficient way.

To aim at the high complexity of expanding strategy and the defect of gaining many homeless nodes in LFM[15] , this paper propose a quick local fitness optimization method(QLFM).

In this paper ,we firstly use a local fitness function (or benefit function)which is also used in LFM as a criterion to decide whether one node is expanded .Then ,We choose a random node as a seed(started node),Our method will expand from inside to outside of the seed according to the Breadth-First-Search in graph .As different seeds will expand to different communities independently ,and these communities have same nodes ,thus our method can detect overlapping nodes quickly and efficiently . We test our method on Benchmark networks and real networks , result shows that our method has lower time-complexity and can gain higher quality communities.

3. Method

As usual , social network can be described as a graph $P=(V,E)$,in which $V=\{v_1,v_2,\cdots,v_n\}$ represents a set of nodes ,and $E=\{e_1,e_2,\cdots,e_m\}$ represents a set of edges , n and m are the numbers of node and edge. $e=(u,v)$ represents an edge whose nodes are u,v .Being Same with LFM We define fitness function as:

$$f_G=\frac{k_{in}^G}{\left(k_{in}^G+k_{out}^G\right)^\alpha} \tag{1}$$

where k_{in}^G , k_{out}^G are the total internal and external degrees of the nodes of module G , α is positive real-valued parameter, controlling the size of the communities. The internal degree of a module equals the double of the number of internal links of the module. The external degree is the number of links joining each member of the module with the rest of the graph .for any node i ,we define its fitness function as a variation of the fitness of sub-graph G with ({ $G+i$ })and without node ($\{G-i\}$).the expression is :

$$f_G^i=f_{G+i}-f_{G-i} \tag{2}$$

In which f_{G+i} is the fitness value of graph $\{G+i\}$, f_{G-i} is the fitness value of graph $\{G-i\}$.

If $f_G^i>0$, then the node i will increase the value of fitness of graph G ,add i to G .if $f_G^i\le 0$, then the node i isn't useful for increasing the value of fitness of graph G , So ignore i .

Before we introduce our expansion strategy ,we should see how it expand in LFM

1) Choose a node A as seed randomly and initialize a new community C ;

2) Choose all neighbors of all nodes in C ,calculate the fitness value for each neighbor;

3) Add the node who has maximum fitness value to C ,thus generate a new community C' ;

4) Caculate fitness value of each node in C' ;

5) if a node turns out to have negative fitness, delete it from C' , yielding C'' , goto step 4);

6) Otherwise, go to step 2) ;

When all neighbors of all nodes in community are negative, the community reach the best structure ,thus the method will stop expand and select next seed. In order to improve the efficiency, LFM doesn't expand from every node. it just choose a node which has never be included by communities as seed randomly.

However, in LFM, it will calculate the fitness value of all neighbors of all nodes in community in each expanding iteration.it is nothing in small network, but for those large community the neighbors set will became larger, so it is usually low efficient.

When community has expanded one node successfully or remove one node, LFM will have a backtracking to recalculate the fitness value of all nodes in new community .this step not only waste much time but yield many homeless nodes. for example, suppose we select node A as seed and begin to expand from it yielding a community C .if fitness value of A is negative in backtracking step, then node A will be removed from C . As we know, most of nodes in C are the neighbors of A and they will have no chance to be as seed, thus, node A will never be detected by other communities and become homeless node at last. Our tests also prove that the LFM will generate many homeless nodes especially in large networks. Obviously, homeless node will decrease the quality of LFM .Based on the above defects, We give a new expansion strategy .the algorithm description is:

Input : $P=(V,E)$, α

Output :Community Set CS

1) select a node v_0 from V randomly.

2) If $Vistited[v_0]$!=0 turn into step 1);

$Q.Enqueue(v_0)$;//enqueueing

$Vistited[v_0]=1$;

$C=\{v_0\}$;

$CS=CS\cup C$;

while($Q.IsEmpty()=false$)

{

$node=Q.Dequeue()$; //delete

Foreach $n\in Neighbour(node)\cap n\notin C$

If $f_c^n>0$ then

$Q.Enqueue(v_0)$;

$Vistited[n]=1$;

$C=C\cup n$;

EndIf

EndFor

}

End

The algorithm stops when all nodes have been assigned to at least one group .otherwise it will select a new node who have never been assigned to any community as seed. We use a queue Q in method .with the help of property fisrt-in and first-out of Q, we can easily detect the community layer by layer . $Visitred[v_0]$ is used to record whether node v_0 is visited before. $Neighbour(node)$ is a set of all neighbors of $node$ .In our method ,When a node is selected as seed ,it will be kernel of community and never be removed. For each iteration ,we only explore one neighbor node .the fitness value decide whether this node is added to community .our method is a greedy expansion strategy .In other words ,if the fitness value of neighbor node is positive we think it can help to optimize the structure of the community ,so we add it to community without selecting the best node from all neighbors .of course if the fitness value turns out to be negative ,we will ignore it simply and explore the next neighbor. Without backtracking after one node is explored ,our method can avoid generating homeless node and save much compute time. Experiments will prove that our expansion strategy can give better result.

Complexity Analysis : Suppose the number of community in graph is n_c, and the average community size is m ,then the whole time complexity of QLFM is $O(n_c * m)$,while the LFM reach $O(n_c * m^2)$ and $O(n^2)$ in worst-case. As we know, $n_c * m$ is the equal to $n_o + n$ where n_o is the number of overlapping nodes .So it can also be expressed as $O(n_o + n)$, Thus our method has linear time with the size of network.

4. Experiment

In this section, we will compare our method(QLFM) with other famous methods (CPM[10], COPRA[13], LFM[15])on synthetic and real-world networks, respectively. All implementations of these algorithms are supported by their authors. As for the parameters, CPM, where we set k = 4, which returned the best results overall; COPRA, where we set v = 3; about LFM and QLFM ,we set the same value $\alpha = 0.9$ .As COPRA, LFM, QLFM are not stable algorithms, And all results shown in experiments are the best of ten independently runs .

We firstly compare the time efficiency in artificial networks , then we compare the quality of result in both artificial networks and real networks; finally, we do experiments to explore how the parameter α in QLFM impact the result.

4.1. Experiments Environment

The hardware environment is Inter Pentium Dual 2.20GHz, 2G RAM, 320G hard disk; operating system is 32-bit win7; the development tool is MyEclipse .

4.2. Experiment Data

We choose four real networks (see Table 1) and LFR benchmarks as the experiment dataset .

Table 1. Information of Real Networks

Name	Nodes	Edges	Ref
Karate	34	78	[18]
Football	115	613	[3]
Dolphins	62	159	[19]
Email	1133	5451	[20]

LFR[21] benchmark graphs are datasets used widely.it has two advantages ,the first is that it has scale-free degree and community size distributions as well as overlapping communities. besides, its

known community structure can help to evaluate the quality of detected community.

To construct LFR synthetic networks, ten parameters should be given. N is the number of vertices ; k is average degree; $k\max$ is the maximum degree of node ; $\min c$ is the size of the smallest community; $\max c$ is the size of the largest community; on is the number of overlapping vertices; om is the number of communities that each overlapping vertex belongs to; mu is the mixing parameter which is the probability of node connect to the other nodes out of community. We will introduce the detailed parameter configuration of each LFR network in the next few sections.

4.3. Evaluation Criteria

1) For real-world networks, we use Modularity which is popular quality function to quantify communities, and larger values indicate better community structures .here we use Q_{ov} [22]. The value of overlap modularity depends on the number of communities to which each vertex belongs and the strength of its membership to each community. We assume that each vertex belongs equally to all of the communities of which it is a member . To calculate it We assume the membership of node i and j to community c is $\alpha_{i,c}$ and $\alpha_{j,c}$ .We firstly define a function $F\left(\alpha_{i,c},\alpha_{j,c}\right)$:

$$F\left(\alpha_{i,c},\alpha_{j,c}\right)=\frac{1}{(1+e^{-f(\alpha_{i,c})})\left(1+e^{-f\left(\alpha_{j,c}\right)}\right)} \tag{3}$$

where $f\left(x\right)=60x-30$ .according to function F , Q_{ov} is calculated as :

$$\beta_{I(i,j),c}=F\left(\alpha_{i,c},\alpha_{j,c}\right) \tag{4}$$

$$\beta^{out}_{I(i,j),c}=\frac{\sum_{j\in V}F\left(\alpha_{i,c},\alpha_{j,c}\right)}{|\mathrm{V}|} \tag{5}$$

$$\beta^{in}_{I(i,j),c}=\frac{\sum_{i\in V}F\left(\alpha_{i,c},\alpha_{j,c}\right)}{|\mathrm{V}|} \tag{6}$$

$$Q_{ov}=\frac{1}{m}\sum_{c}\sum_{i,j\in V}\left(\beta_{I(i,j),c}A_{i,j}-\beta^{out}_{I(i,j),c}\beta^{in}_{I(i,j),c}\frac{k_i^{out}k_j^{in}}{m}\right) \tag{7}$$

2) For the LFR networks, as we know the community structure before. Normalized Mutual Information (the detailed definition can be found in paper [15]) can be used to measure the similarity of the ground truth and found communities. NMI = 1 means that the found communities perfectly match the real communities. Smaller values of NMI indicate worse detection results.

4.4. Time Comparison

To compare the time efficient with LFM, we generate 20 LFR networks .the parameter N is set as 1000~10000 and 10000~100000. Other parameters are same (k =10, $k\max$ =50, $\min c$ =10, $\max c$ =50; mu =0.1). We can see the results in Fig. 1. from the Fig. 1, we can easily find that both QLFM and LFM have near linear running time with the growth of size of network in the left figure .this is because both methods detect the overlapping community by scanning the each nodes once or more in graph. However, We can find the QLFM is far lower than LFM. When N is over 80000, the time of LFM will increase rapidly. this is

because QLFM only consider one neighbor when expanding while LFM will calculate all the neighbors. Besides, QLFM has no back tracking while LFM will recalculate all nodes in current community. when the network is large, These steps is obviously a waste of time. So QLFM can get better time efficiency than LFM.

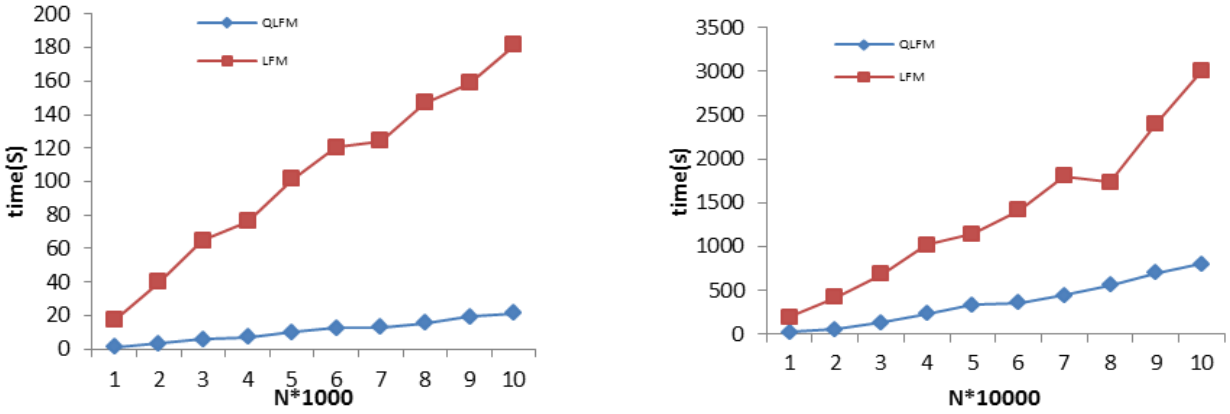


Fig. 1. Time efficiency comparison in different size of networks.

4.5. Synthetic Networks

In this experiment ,we will produce 64 networks with different properties including network size, mix parameter mu , om . Table 2 is the detail information .We adjust parameter N , $\min c$, $\max c$ to generate different size networks including small network with small communities (SS), small network with large communities (SL), large network with small communities (LS), and large network with large communities (LL).For each size network ,we make mu increase from 0.1 to 0.5 by setting step length as 0.05 to gain more complex networks(S1~S4). Also, we make om increase from 2 to 7 by setting step length as 1 to create networks with different degrees of overlapping (S4~S8).

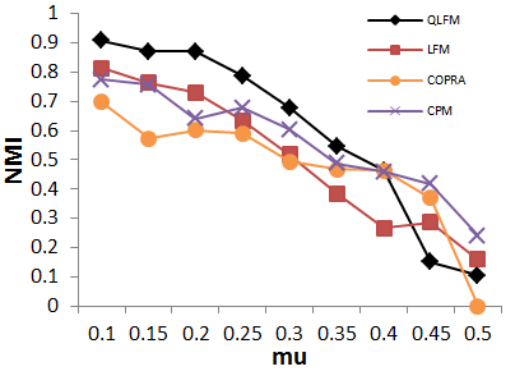
Fig. 2 shows the experimental results. the X-axis is value of mu in Fig. (a)-(d). In Fig. (e)-(h), X-axis represents om . the Y-axis is the final NMI . From results shown in Fig. 2, we can conclude:

1) QLFM can work well even when adapted to different size of networks with different size of communities. Like most of other several algorithms, QLFM is sensitive to mixing parameter, the NMI will decrease quickly with the growth of mu which can be found from Fig. (a)-(d). When mu >0.4, QLFM even can't detect meaningful communities. In Fig. (e)-(h), although The increase of om also has a bad impact on final NMI, this impact is very weak. We can see that with the increasing of om, the NMI decrease more slowly.
2) Comparing with LFM, we can find in all cases QLFM get better performance than LFM, this is because that the LFM will generate many homeless nodes while QLFM nearly has no homeless node. these homeless nodes will influence the result significantly.
3) To compare with COPRA , we can find in most cases QLFM give better result than COPRA exception in Fig. (b). It is worth noting that the results of COPRA jump widely such as in Fig.(c), Fig. (d), Fig. (f), Fig. (h). Especially in Fig. (d), with the growth of mixing parameter ,the result of other methods tend to give worse result, while result of COPRA appears locally better abnormally. this suggests that CORPA is not as stable as QLFM and LFM.
4) Besides , we can also find that QLFM get better results than CPM in most of cases.

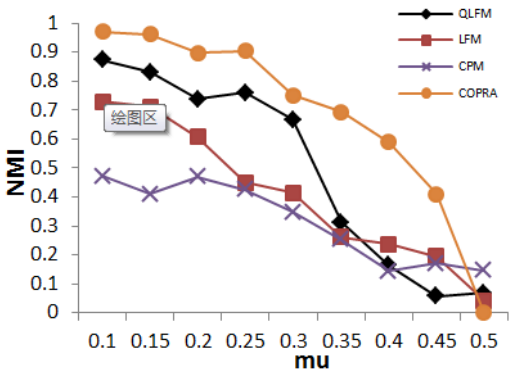
In summary QLFM will produce high quality overlapping communities than other algorithms on synthetic network.

Table 2. Information of LRF Networks

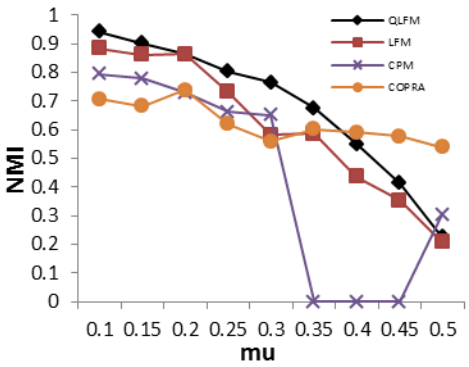
Id	N	k	max k	min c	max c	on	om	mu
S1	1000	10	50	10	50	100	2	0.1~0.5
S2	1000	10	50	20	100	100	2	0.1~0.5
S3	5000	10	50	10	50	100	2	0.1~0.5
S4	5000	10	50	20	100	100	2	0.1~0.5
S5	1000	10	50	10	50	100	2~8	0.1
S6	1000	10	50	20	100	100	2~8	0.1
S7	5000	10	50	10	50	100	2~8	0.1
S8	5000	10	50	20	100	100	2~8	0.1



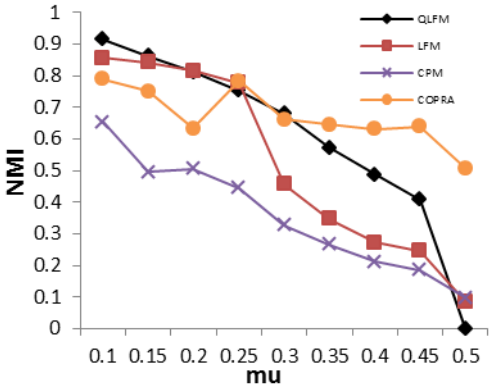
(a). Small G Small C.



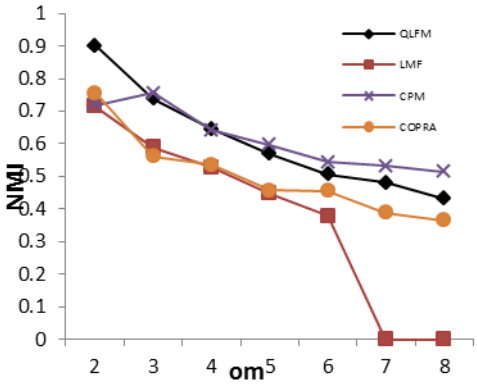
(b). Small G Large C.



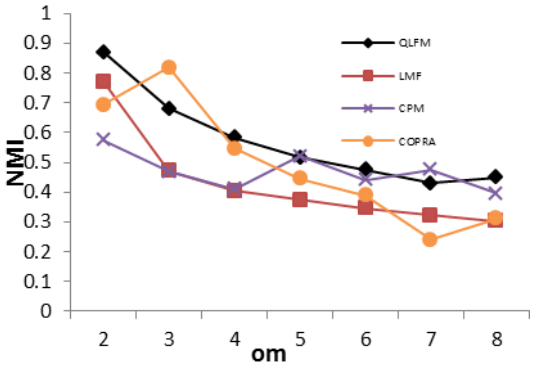
(c). Large G Small C.



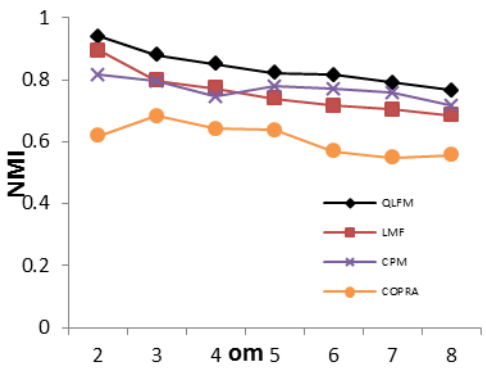
(d). Large G Large C.



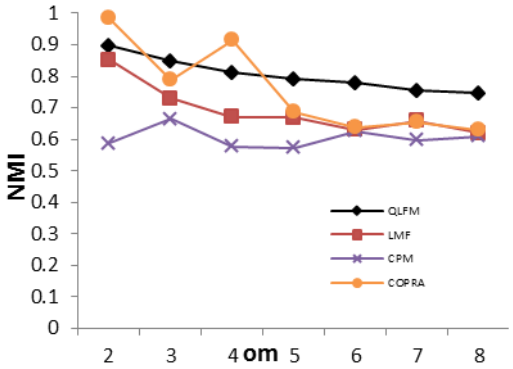
(e). Small G Small C.



(f). Small G Large C.



(g). Large G Small C.



(h). Large G Large C.

Fig. 2. Community results on synthetic network.

4.6. Real Word Networks

In real networks, we use Q_{ov} to evaluate the final community structure. Table 3 shows the results. From the table, we can find that QLFM get the biggest Q_{ov} on Karate and Dolphin network .Although QLFM doesn't give the best result in Football and Email network ,it still get the second best results, especially on Football network, Q_{ov} of QLFM is nearly equal to that of LFM. So, QLFM can find better quality overlapping community structure in real networks as a whole.

Table 3. Result on Real Network

	QLFM	LFM	CPM	COPRA
Karate	0.721	0.431	0.259	0.476
Dolphin	0.730	0.388	0.269	0.639
Football	0.655	0.661	0.641	0.653
Email	0.434	0.206	0.352	0.737

4.7. Parameter Choice

In order to study how α influences the final result of QLFM, We adjust α in three real networks(Karate ,Dolphin ,Football). Fig. 3 shows relationship between the average Q_{ov} of 10 times of test with α .

From the Fig. 3, we can find that when α <0.6 QLFM give the worst performance (Q_{ov} =0) in all the three networks. When α >1.7, Q_{ov} of three networks begin to descend .QLFM can find the best average value when 0.7< α <1.6 although the best average Q_{ov} is given in different value of α to three networks . So we advise to choose α among [0.7, 1.6].

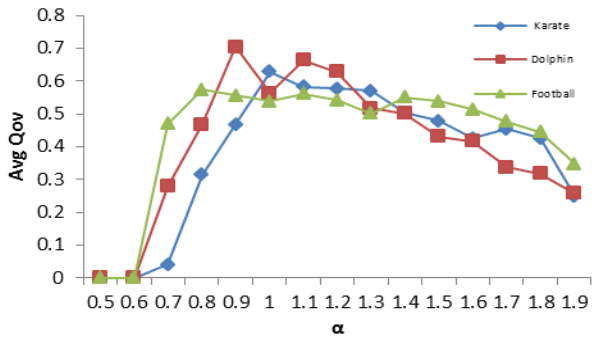


Fig. 3. Relationship Between Average Q_{ov} with α.

5. Conclusion

In this paper, we have introduced an efficient expansion strategy to detect overlapping community structure based on one local fitness function. We compare our method with LFM，CPM and COPRA on both synthetic Networks and real networks. The results show that the proposed algorithm perform better than other methods in both quality and run-time. In addition, we also analyze the impact of experimental parameters α on the result.

References

[1] Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*.

[2] Shang, R. H., *et al*. (2013). Community detection based on modularity and an improved genetic algorithm. *Physica a-Statistical Mechanics And Its Applications*, *392(5)*, 1215-1231.

[3] Girvan, M., & Newman, M. E. (2001). Community structure in social and biological networks.

[4] Blondel, V. D., *et al*. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics-Theory and Experiment*.

[5] Shen, H. W., & Cheng, X. Q. (2010). Spectral methods for the detection of network community structure: A comparative analysis. *Journal of Statistical Mechanics-Theory and Experiment*.

[6] Jiang, J. Q., Dress, A. W. M., & Yang, G. K. (2009). A spectral clustering-based framework for detecting community structures in complex networks. *Applied Mathematics Letters*, *22(9)*, 1479-1482.

[7] Raghavan, U. N., R. Albert, & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev E Stat Nonlin Soft Matter Phys*.

[8] Subelj, L., & Bajec, M. (2011). Unfolding communities in large complex networks: combining defensive and offensive label propagation for core extraction. *Phys Rev E Stat Nonlin Soft Matter Phys*.

[9] Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Natl Acad Sci*, *105(4)*, 1118-1123.

[10] Palla, G., *et al*. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 814-818.

[11] Kumpula, J. M., *et al*. Sequential algorithm for fast clique percolation. *Phys Rev E Stat Nonlin Soft Matter Phys*.

[12] Farkas, I., *et al.* Weighted network modules. *New Journal of Physics*, *9(6)*, 180.

[13] Gregory, S., (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics*.

[14] Ahn, Y. Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*.

[15] Lancichinetti, A., Fortunato, S., & Kertesz, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*.

[16] Coscia, M., *et al*. (2012). DEMON: A local-first discovery method for overlapping communities. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 615-623).

[17] Lancichinetti, A., *et al.* (2011). Finding Statistically Significant Communities in Networks. *Plos One*, *6(4)*.

[18] Zachary, W. W., (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 452-473.

[19] Network Data. Retrieved, from http://www.cs.bris.ac.uk/~steve/networks/congapaper/.

[20] Gregory, S. Network Research. Retrieved, from http://www.cs.bris.ac.uk/~steve/networks/copra/.

[21] Lancichinetti, A., Fortunato, S., & Radicchi, F. (2008). Benchmark graphs for testing community

detection algorithms. *Physical Review E*, *78(4)*.

[22] Nicosia, V., *et al*. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics-Theory and Experiment*.

Yanan Li was born in China, in 1991. He recevied his master degree from College of Computer Science, Chongqing University, Chongqing, China. His research interest is detecting overlapping community in social network.

Zhengyu Zhu was born in China. He is a Ph.D., professor and a doctoral supervisor, a senior member of CCF (E200009348S). His research interests are web intelligent search, e-commerce applications, data mining technology.