ISSN 1674-1056
http://cpb.iphy.ac.cn
http://iopscience.iop.org/cpb
中国物理 B
Chinese Physics
中国物理
Chinese Physics B

Detecting overlapping communities based on vital nodes in complex networks

Xingyuan Wang(王兴元), Yu Wang(王宇), Xiaomeng Qin(秦小蒙), Rui Li(李睿), Justine Eustace

Citation: Chin. Phys. B . 2018, 27(10): 100504. **doi:** 10.1088/1674-1056/27/10/100504

Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Empirical topological investigation of practical supply chains based on complex networks

Hao Liao(廖好), Jing Shen(沈婧), Xing-Tong Wu(吴兴桐), Bo-Kui Chen(陈博奎), Mingyang Zhou(周明洋)

Chin. Phys. B . 2017, 26(11): 110505. **doi:** 10.1088/1674-1056/26/11/110505

Cluster synchronization of uncertain complex networks with desynchronizing impulse

Cai Guo-Liang, Jiang Sheng-Qin, Cai Shui-Ming, Tian Li-Xin

Chin. Phys. B . 2014, 23(12): 120505. **doi:** 10.1088/1674-1056/23/12/120505

Adaptive function projective synchronization of uncertain complex dynamical networks with disturbance

Wang Shu-Guo, Zheng Song

Chin. Phys. B . 2013, 22(7): 070503. **doi:** 10.1088/1674-1056/22/7/070503

Adaptive generalized matrix projective lag synchronization between two different complex networks with non-identical nodes and different dimensions

Dai Hao, Jia Li-Xin, Zhang Yan-Bin

Chin. Phys. B . 2012, 21(12): 120508. **doi:** 10.1088/1674-1056/21/12/120508

Generalized projective synchronization of fractional-order complex networks with nonidentical nodes

Liu Jin-Gui

Chin. Phys. B . 2012, 21(12): 120506. **doi:** 10.1088/1674-1056/21/12/120506

Detecting overlapping communities based on vital nodes in complex networks*

Xingyuan Wang(王兴元)[1,2,†], Yu Wang(王宇)[2], Xiaomeng Qin(秦小蒙)[2], Rui Li(李睿)[3], and Justine Eustace[2]

[1] School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
[2] Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China
[3] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China

(Received 21 March 2018; revised manuscript received 9 August 2018; published online 27 September 2018)

Detection of community structures in the complex networks is significant to understand the network structures and analyze the network properties. However, it is still a problem on how to select initial seeds as well as to determine the number of communities. In this paper, we proposed the detecting overlapping communities based on vital nodes algorithm (DOCBVA), an algorithm based on vital nodes and initial seeds to detect overlapping communities. First, through some screening method, we find the vital nodes and then the seed communities through the pretreatment of vital nodes. This process differs from most existing methods, and the speed is faster. Then the seeds will be extended. We also adopt a new parameter of attribution degree to extend the seeds and find the overlapping communities. Finally, the remaining nodes that have not been processed in the first two steps will be reprocessed. The number of communities is likely to change until the end of algorithm. The experimental results using some real-world network data and artificial network data are satisfactory and can prove the superiority of the DOCBVA algorithm.

Keywords: complex networks, overlapping communities, vital nodes, seed communities

PACS: 05.10.–a, 02.10.Ox **DOI:** 10.1088/1674-1056/27/10/100504

1. Introduction

There are many kinds of complex systems in nature and society. Most of them can be described as networks, where the nodes (or vertexes) represent entities and the edges represent the relationship between the nodes.[1–4] The community structure is the most important feature of complex networks, and it is very important to study the community structure in the network from the perspective of theoretical research and practical application. The main reasons are four points:

(i) The research community can better understand the function and the overall structure between the nodes and sides from the local network.

(ii) We can understand the dynamics of the whole network better from the perspective of mass organizations.

(iii) We can explain the underlying causes of the network from a macro perspective and help to understand the real evolution process and mechanism of the network.

(iv) It provides a clustering method for real world networks, and does not rely on other attributes of data.

It is widely believed that communities are widespread in large networks,[5,6] where the nodes have high density of inner-group edges and low density of inter-group edges.[7–11] Recently, a large number of community detection algorithms have been proposed. Early researches focused on non-overlapping communities where a node belongs to a single community in Ref. [12], such as Kernighan-Lin algorithm in Ref. [13], and the spectral bisection algorithm on the basis of the eigenvectors of the Laplacian matrix of a network in Ref. [14]. The parameters in these methods seem to be too strict to show the link of reality social networks because a node may belong to more than one community. In fact, communities in social networks usually overlap with each other since many active users can possibly participate in multiple groups at the same time.[15] For example, a student may participate in two community activities, and the proportion between the two communities is the same. Thus, more and more researchers begin to concentrate on detecting overlapping communities.

Hierarchical clustering is a common approach which detects community based on the similarity or intensity between vertices.[16] According to this approach, the community structure is considered to be hierarchical by

adding or removing edges or nodes.[16,17] At the same time, the core vertex is also the key to detect the communities.[16] In general, the nodes with high degree are likely to be vital nodes in complex networks, and they are key points in detecting communities. Through the vital nodes, we can quickly find the center of a community. The method of detecting overlapping communities based on vital nodes algorithm (DOCBVA) uses clustering method to achieve the detection of overlapping networks. Because of the screening of vital nodes, we can find the seed communities quickly and cor-

*Project supported by the National Natural Science Foundation of China (Grant Nos. 61672124, 61370145, 61173183, and 61503375) and the Password Theory Project of the 13th Five-Year Plan National Cryptography Development Fund, China (Grant No. MMJJ20170203).

†Corresponding author. E-mail: wangxy@dlut.edu.cn

© 2018 Chinese Physical Society and IOP Publishing Ltd

rectly. At the same time, during the research, it is shown that most setting ranges of the parameters related to the intimacy between nodes and communities are so strict that many common nodes are ignored. Therefore, we use a new parameter to measure the intimacy between nodes and communities. In addition, in many algorithms,[7,13] the number of communities is known in advance or constant once they are identified. To solve this problem, we select the main seeds of club firstly, then the number of communities is constantly updated during the next steps. By detecting the data of the real-world networks and artificial networks, our method can get the community structures quickly and accurately.

The rest of this paper is organized as follows. First, we give the crucial concepts of our new overlapping community detection algorithm. Next, we describe the detailed realization of the algorithm. Finally, we use DOCBVA to analyze real-world networks and artificial networks and show the results and conclusions. It is worth mentioning that, in this paper, we just consider the unweighted networks.

2. Basic concepts

2.1. Basic notations

Let $G=(V,E)$ be a graph representing a network, where V is the set of N nodes and E is the set of M connections. M is the total number of edges in the network. We denote the network community structure as $C=\{C_1,C_2,\ldots,C_k\}$. k_i is the degree of vertex v_i, $k_{\text{in}}^{C_i}$ is the number of edges the vertex v_i connecting to the community of C_i, and $k_{\text{out}}^{C_i}$ is the total number of edges with the vertex v_i connecting to other communities. We allow $C_i \cap C_j \neq \phi$ so that the network communities can overlap with each other. A neighborhood set $N(v_i)$ is the collection of neighboring nodes of vertex i (v_i). $N(C)$ is the collection of neighboring nodes of community C.

Definition 1 A neighborhood set $I(v_i)$, $I(v_i) \in V$, is a set of all vertices adjacent to vertex v_i including itself, $I(v_i) = N(v_i) \cup v_i$.

2.2. Objective function

Our objective is to find the seed communities of networks which are more distant from each other and extend the seeds. In networks, the connection crossings between communities should be less than those inside them.[13] Unlike the case of disjoint community structures, overlapping communities allow a node to belong to more than one community. In this paper, we use three functions to realize the process of community detection.

First, in order to save the running time, we will find all nodes whose degrees are greater than the average of the network, and organize them into a set S in descending order. By doing this, we can filter out some important nodes. Then we select the $v_i (v_i \in S)$ which has the maximum degree and has not been assigned to the seeds of communities as the seed of absorbing community. Because S is ordered, v_i is the first node in S. If the vertex v_j meets the following condition, traverse S and absorb the vertex $v_j \in S$ until no more nodes can be absorbed.

Definition 2 Let $v_i, v_j \in S$, v_i and v_j are in the same community when $I(v_i) \cap I(v_j) \geq I(v_j)/2 (i < j)$.

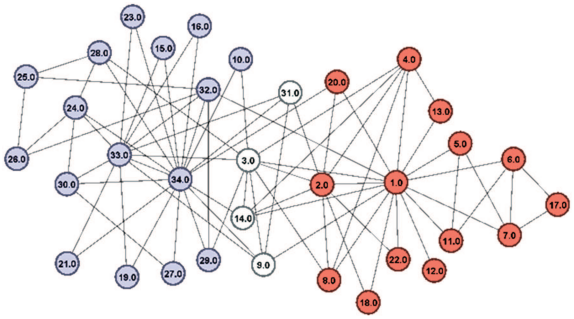


Fig. 1. (color online) Two communities and four overlapping vertices in friendship network from Zachary's karate club study.

Definition 2 suggests that there are enough public neighbors between v_i and v_j. Then we can make a judgment that they have a much closer relationship with each other than with others. Let us assume that v_i and v_j are not in the same community. Because of the condition $I(v_i) \cap I(v_j) \geq I(v_j)/2(i < j)$, the vertex v_i is closer with another community which contains v_j, which is consistent with the definition of community.

In order to explain the process of selecting the core vertex and the seed of absorbing community, Zachary's network of karate club members is considered in Fig. 1, from which we can see the average degree of the network is 5 (the accurate value is 4.6, here we round it up to an integer) and the S is {34, 1, 33, 3, 2, 4, 32, 9, 14, 24}. Then we choose the node $v_i = \{34\}$ as the seed of the absorbing community C_1, $C_1 = C_1 \cup v_i$. Now the set of $S = \{1,33,3,2,4,32,9,14,24\}$. Table 1 is the analysis process of finding the seed communities of Zachary's network. Through the pretreatment of vital nodes, we can get two seed communities {34, 33, 3, 32, 9, 24} and {1, 2, 4, 14}.

However, the community structure is not always strongly related with nodes of large degree. For example, the community in power grid can be composed of nodes with homogeneous degrees. For this type of network, our approach is still applicable. For example, figure 2 is a simple network, where nodes have relatively homogeneous degrees. In this case, the set of S will include most of the nodes in the network, $S = \{2,4,3,5,6,7,8,9\}$. We can still obtain some seeds of communities, {2, 4, 3, 5, 6} and {7, 8, 9}. Meanwhile, because we have dealt with sufficient nodes, the following processing time will be greatly shortened. There are still some defects when we deal with tree like and household networks.

Table 1. Analysis of the process of finding the seed communities of Zachary's network.

	v_i	v_j		S	Seed communities
Round 1	34			{1, 33, 3, 2, 4, 32, 9, 14, 24}	$C_1 = \{34\}$
	34	1	$I(v_i) \cap I(v_j) < I(v_j)/2$	{1, 33, 3, 2, 4, 32, 9, 14, 24}	$C_1 = \{34\}$
	34	33	$I(v_i) \cap I(v_j) \geq I(v_j)/2$	{1, 3, 2, 4, 32, 9, 14, 24}	$C_1 = \{34,33\}$
	34	3	$I(v_i) \cap I(v_j) \geq I(v_j)/2$	{1, 2, 4, 32, 9, 14, 24}	$C_1 = \{34,33,3\}$
	34	2	$I(v_i) \cap I(v_j) < I(v_j)/2$	{1, 2, 4, 32, 9, 14, 24}	$C_1 = \{34,33,3\}$
	34	4	$I(v_i) \cap I(v_j) < I(v_j)/2$	{1, 2, 4, 32, 9, 14, 24}	$C_1 = \{34,33,3\}$
	34	32	$I(v_i) \cap I(v_j) \geq I(v_j)/2$	{1, 2, 4, 9, 14, 24}	$C_1 = \{34,33,3,32\}$
	34	9	$I(v_i) \cap I(v_j) \geq I(v_j)/2$	{1, 2, 4, 14, 24}	$C_1 = \{34,33,3,32,9\}$
	34	14	$I(v_i) \cap I(v_j) < I(v_j)/2$	{1, 2, 4, 14, 24}	$C_1 = \{34,33,3,32,9\}$
	34	24	$I(v_i) \cap I(v_j) \geq I(v_j)/2$	{1, 2, 4, 14}	$C_1 = \{34,33,3,32,9,24\}$
Round 2	1			{2, 4, 14}	$C_2 = \{1\}$
	1	2	$I(v_i) \cap I(v_j) \geq I(v_j)/2$	{4, 14}	$C_2 = \{1,2\}$
	1	4	$I(v_i) \cap I(v_j) \geq I(v_j)/2$	{ 14}	$C_2 = \{1,2,4\}$
	1	14	$I(v_i) \cap I(v_j) \geq I(v_j)/2$	ϕ	$C_2 = \{1,2,4,14\}$

Second, we will extend the seed communities. It is shown that most setting ranges of parameters related to the intimacy between nodes and communities are too strict in the searching process. As shown in Fig. 3, by calculating the absorbing degree $A(C_i,\mu)$ in Ref. [7] and the fitness function $f_{C_i}^{\mu}$ in Ref. [18], we can get $A(C_1,\mu) > A(C_2,\mu)$ and the fitness $f_{C_1}^{\mu} > 0$ and $f_{C_2}^{\mu} < 0$. Therefore, the node μ is categorized into the community of C_1. But in reality, especially when the difference of two link weights between μ and two communities is very small, we will take the node as a common node. To solve this problem, we put forward a new parameter to measure the intimacy between the node and the community. For a community C_i and a node μ, the attribution degree $\alpha(C_i,\mu)$ between C_i and μ is defined as

$$\alpha(C_i,\mu) = \frac{\left|k_{\text{in}}^{C_i} - k_{\text{out}}^{C_i}\right|}{k_{\text{in}}^{C_i}} (k_{\text{in}}^{C_i} \neq 0). \tag{1}$$

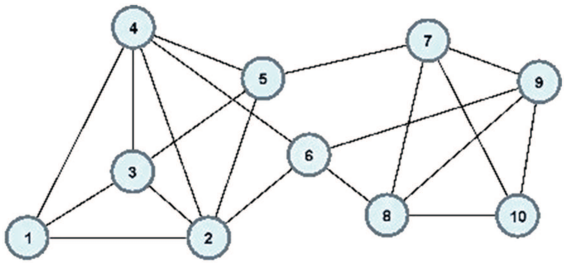


Fig. 2. (color online) A simple network where nodes have relatively homogeneous degrees.

It reflects how tight the node μ is with community C_i. $k_{\text{in}}^{C_i}$ is the number of edges between μ and community C_i, and $k_{\text{out}}^{C_i}$ is the number of edges between μ and others. The attribution degree $\alpha(C_i,\mu)$ reflects the ratio of the difference between the inner and outer edges of μ and C_i to the total degrees of μ. If $\alpha(C_i,\mu) < 1$, we will consider that the node μ belongs to the community C_i. Then, in Fig. 3, we can easily get $\alpha(C_1,\mu) < 1$ and $\alpha(C_2,\mu) < 1$. The node μ is the common node. Therefore, in this case, the attribution degree $\alpha(C_i,\mu)$ is more reasonable.

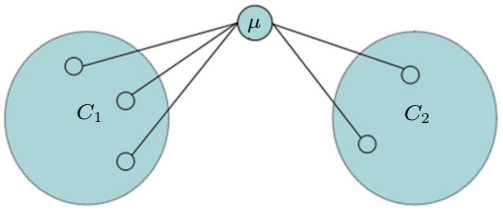


Fig. 3. (color online) A simple example discussing the question of belongingness.

For the unassigned nodes, we will consider the fitness function[18]

$$F = \frac{k_{\text{in}}^{C_i}}{k_i}. \tag{2}$$

Each node belongs to the community that has the maximum value of F.

2.3. Modularity

In order to quantify the community structure of a network, Newman and Girvan[19] proposed the modularity Q as a measure of network partition

$$Q = \sum_i (e_{ii} - a_{ii}^2) = \text{Tr}\, e - \left\| e^2 \right\|, \tag{3}$$

where e_{ii} is the fraction of weights of edges belonging to community i in the total weights of all edges, and a_{ii} is the fraction of edges which have one or both vertices inside i.[20] If the network is unweighted, the weight of each edge is equal to 1. But it faces several problems, for example, it suffers a resolution limit problem,[21] and cannot tackle overlapping community structure. Therefore, Newman proposed an extended modularity defined as[22]

$$Q_0 = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{4}$$

where A_{ij} is the weight of the link between vertex i and vertex j, k_i is the degree of vertex i, c_i is the community to which the vertex i belongs, and m is the sum of the edges of complex network

$$m = \frac{1}{2} \sum_{i,j} A_{ij}. \tag{5}$$

And if $c_i = c_j$, the function $\delta(c_i, c_j) = 1$, which means i and j belong to the same community, otherwise $\delta(c_i, c_j) = 0$. In order to accurately measure the quality of the overlapping community structure, many researchers have made an extension of modularity, such as EQ, which was proposed by Shen in Ref. [23]. Here we will not go into details.

3. Algorithm

3.1. Detecting the seeds of communities

The detection of the seeds of communities is the first and also the most important phase of our method. Generally speaking, the seeds of communities provide us an overview of the network as well as the critical information of communities. First, we will get the set of vital nodes S through the pretreatment of V where the degree of every node is greater than the average degree of network, which can effectively reduce the number of nodes that need to be treated. And S is an ordered collection where nodes are sorted in descending order by degree. It is worth mentioning that a node in S belongs to only one club. For example, consider Zachary's network of karate club members in Fig. 1, we can get $S = \{34, 1, 33, 3, 2, 4, 32, 9, 14, 24\}$, and two seeds of communities {34, 33, 3, 32, 9, 24} and {1, 2, 4, 14} will be found through Algorithm 1.

The details are presented in Algorithm 1.

Algorithm 1

Input: The set of S

Output: The seeds of communities $C' = \{C'_1, C'_2, \ldots, C'_{k_1}\}$

Begin

While $S \neq \phi$

$C' = S[0] \cup C' / *S[0]$ is the node with maximum degree*/

For $u \in S$ do

If $I(S[0]) \cap I(u) \geq I(u)/2$ or there exists a fully connected graph constituted by $S[0]$, $I(S[0])$, and u, then

$C' = C' \cup u$

$S = S - u$

Update $N(C')$

End if

End for

End while

End

During the experiment, we find that the condition of $I(S[0]) \cap I(u) \geq I(u)/2$ is so severe that too many nodes are divided into a club alone. So by relaxing the condition, we add another condition that u is in the same community with $S[0]$ when they have a complete subgraph between u, $S[0]$, and at least two common neighbor nodes of them. The experimental results show that additional condition does not have any effect on previous condition, nor does it obtain better experimental results.

3.2. Extending the seed communities

As soon as the first procedure finishes, the raw network community structure can be pictured as a collection of dense parts of the network together with outliers. Because of the condition setting in step 1, there is a low overlapping score between the two seeds. Then we just need to extend the seed communities.

Algorithm 2

Input: The seeds of communities $C' = \{C'_1, C'_2, \ldots, C'_{k_1}\}$

Output: Extended community structure $C = \{C_1, C_2, \ldots, C_{k_1}\}$

Begin

$C = C'$

For $C_i \in C$ do

For $u \in N(C_i)$ do

If $\alpha(C_i, \mu) < 1$ or $k_{\text{out}}^{C_i} = 0$ then

$C_i = C_i \cup u$

Update $N(C_i)$

End if

End for

End for

End

For Zachary's network, because two seeds of communities {34, 33, 3, 32, 9, 24} and {1, 2, 4, 14} have been found through Algorithm 1, we can get $N(C_1) = \{1, 2, 4, 8, 10, 14, 15, 16, 19, 20, 21, 23, 25, 26, 27, 28, 29, 30, 31\}$ and $N(C_2) = \{5, 6, 7, 8, 9, 11, 12, 13, 18, 20, 22, 31, 32, 34\}$. Then we will get the community structure like Fig. 1. We also analyze the network in Fig. 2 and get two communities {1, 2, 3, 5, 6} and {6, 7, 8, 9, 10}. Vertex 6 is the common node. All nodes of these two networks are divided into the community they should belong to. As a result, the Algorithm 3 will be skipped.

3.3. Revisiting unassigned nodes

Even when the above two procedures are executed, there will still be remaining nodes due to their less attraction to the rest of the network. Therefore, we need to revisit these nodes. Here, we use the fitness function to quantify the similarity between a node u and a neighbor community C. Fitness function is described as

$$F = \frac{k_{\text{in}}^{C_i}}{k_i}. \tag{6}$$

We can use the value of F to group the remaining nodes into appropriate communities or classify them as outliers based on their connectivity structures. The node will be regarded as an independent community when the node does not meet Eqs. (1) and (6), and will attend the next round.

The detailed description is presented in Algorithm 3.

Algorithm 3

Input: Extended community structure $C = \{C_1, C_2, \ldots, C_{k_1}\}$

Output: The final community structure $C = \{C_1, C_2, \ldots, C_k\}$

Begin
  For $u \in V$ and u does not belong to any community do
    For $C_i \in C$ do
      If F_{C_i} is maximum
        $C_i = C_i \cup u$
      Else C_i is independent community
      End if
    End for
  End for
End

3.4. Complexity

First, we should take at least $O(n\log^n)$ to sort the vertices by degree. Then in the average case, the number of nodes we choose in S is the half of V. So, the average time complexity of Algorithm 1 is $O(n^2/8)$, and the total complexity of Algorithm 2 and Algorithm 3 is about $O(m)$. At the worst, it costs $O(n\log^n + n^2/8 + m)$ to find all communities.

4. Applications

The overlapping community detecting algorithm proposed in this paper is implemented by Java programming language running on a PC with a 2.2-GHz processor, 3.0-GB memory, and Windows 7 operating system. Some artificial network data of Lancichinetti–Fortunato–Radicchi (LFR) benchmark graphs[24] and real-world networks are used to evaluate the proposed algorithm. The real-world networks include the friendship network from Zachary's karate club study, dolphin"s associations, college football, etc. These networks can be downloaded from Ref. [25].

4.1. Friendship network

The well-known karate club study of Zachary[26] is widely used as a test case for new methods in complex networks. It consists of 34 members of a karate club as vertices and 78 edges representing friendships among members of the club. By chance, a dispute arose between the club's administrator and the club's instructor, and as a result, the club eventually split into two smaller clubs, centered on the administrator and the instructor.[20] The network and its fission are depicted in Fig. 4. The administrator and the instructor are represented by node 1 and node 33, respectively.

As shown in Fig. 4, two overlapping communities with four nodes 3, 14, 9, and 31 in common are detected by our method, with $Q_0 = 0.332$. In Ref. [27], there are two communities and five common nodes, which include nodes 3, 9, 10, 14, and 31. This network is also discussed in other methods, and further details are provided in Table 2. Consistent with the real split, our algorithm finds two communities. Although we have found four common nodes which are different from the reality, it is reasonable to regard them as common nodes of the two communities.

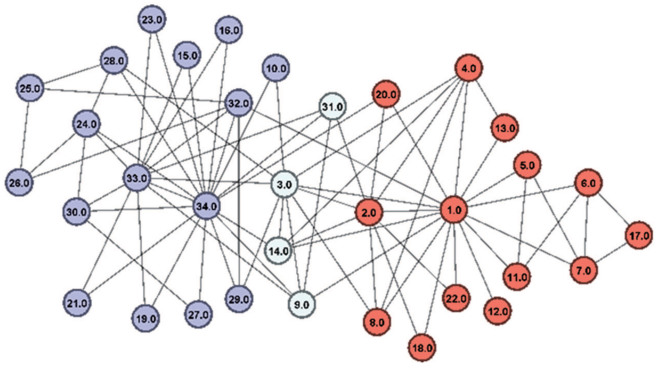


Fig. 4. (color online) Two communities and four overlapping vertices in friendship network from Zachary's karate club study.

4.2. Dolphin's associations

It describes the associations between 62 dolphins living in Doubtful Sound, New Zealand, with ties between dolphin pairs being the statistically significant frequent association.[20] This network can be split naturally into two groups. But in this paper, we find four communities by our algorithm, which are represented by different colors in Fig. 5. The modularity of dolphin's network calculated by our method is 0.511. In the EM-BOAD algorithm of Ref. [7] and the GN algorithm of Ref. [16], this network splits into three communities and two large communities with modularity Q_0 of 0.457 and modularity Q of 0.38, respectively.

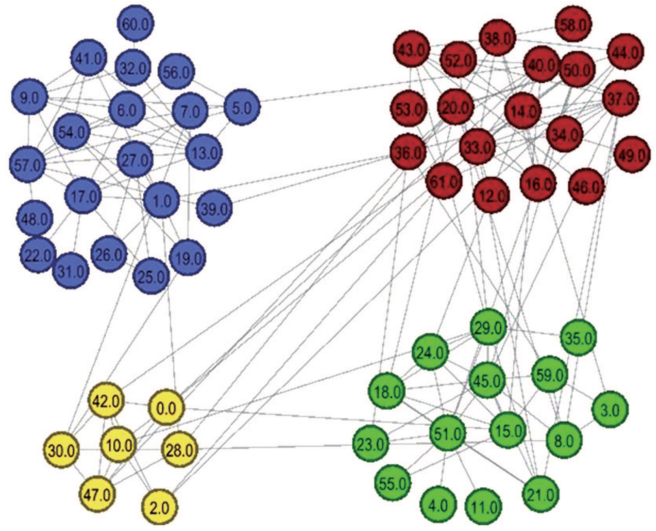


Fig. 5. (color online) Communities in dolphin's living network (no common node is found in this network).

4.3. College football

The third network instance is the college football games in division IA in USA in the fall of 2000,[8] which is a network representing the schedule of games between American college football teams. It includes 115 teams from 12 conferences, and 616 edges represent regular season games. The

Chen algorithm[27] splits the network into 13 non-overlapping communities. 10 non-overlapping communities are detected by our algorithm, as shown in Fig. 6. From Fig. 6, there are two teams that do not belong to any conference, and these tend to be grouped with the team with which they are most closely associated. Thus, they are divided into the other ten teams. At the same time, several members are divided into other teams falsely. Although there are some differences with the real alliance, from the angle of the links of reality games, the experimental result is also reasonable.

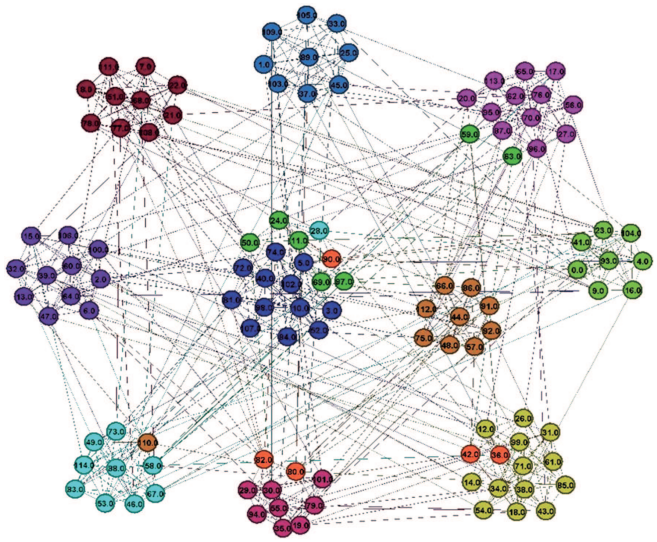


Fig. 6. (color online) Ten non-overlapping communities of college football network.

In order to further verify the performance of our algorithm, we have applied it to a number of test-case networks that are commonly used for efficiency comparisons (see Table 2). This Table gives the performances of other algorithms and our algorithm for detecting community structures in networks of various sizes. For each algorithm/network, the Table displays the modularity that is achieved and the computation time. EM-BOAD[7] is a method of detecting overlapping communities by the seed community in weighted networks. In order to make a comparison with our method effectively, we set the weight of every edge as 1. The local optimization fitness method LFM[18] is evaluated and analyzed by EQ.[23] A method of label propagation algorithm SLPA[28] is also used in this paper. For the same division result, there is little difference between these two values of modularity. As we can see from Table 2, our algorithm has better efficiency in either modularity or running time.

Table 2. Execution time and modularity on some real-world networks.

	Karate	Dolphin	Football	Power grid	Cond-mat-2003[a]
Nodes/Edges	36/78	62/159	116/615	4951/6594	31163/120029
EM-BOAD	0.331/0 s	0.493/0 s	0.578/0 s	0.796/15 s	0.46/180 s
GN	0.36/0.4	0.457/2.8 s	0.585/95 s	–	0.428/1305 s
LFM	0.559/0.1 s	0.875/0.2 s	0.423/0.4 s	0.755/93 s	0.802/11632 s
SLPA	0.359/0 s	0.417/0 s	0.538/0.2 s	0.789/3.5 s	0.65/55.6 s
DOCBVA	0.332/0 s	0.511/0 s	0.596/0 s	0.815/4 s	0.645/130 s

[a]In order to compare with our algorithms, cond-mat-2003 used here is unweighted.

4.4. LFR benchmark graphs

To further verify the performance of the algorithm, we also apply the DOCBVA to the artificial networks. Figure 7 shows the results of the experiment. 10 kinds of networks are randomly generated, each of which is executed 10 times to obtain 100 values. Each value in Fig. 7 is the mean time of the 100 values. From Fig. 7, we can get the results that the run time of our algorithm is less than other algorithms and the performance is higher. With the increase of the mixing parameter μ, the running time has a remarkable increase, but it is still in the acceptable range.

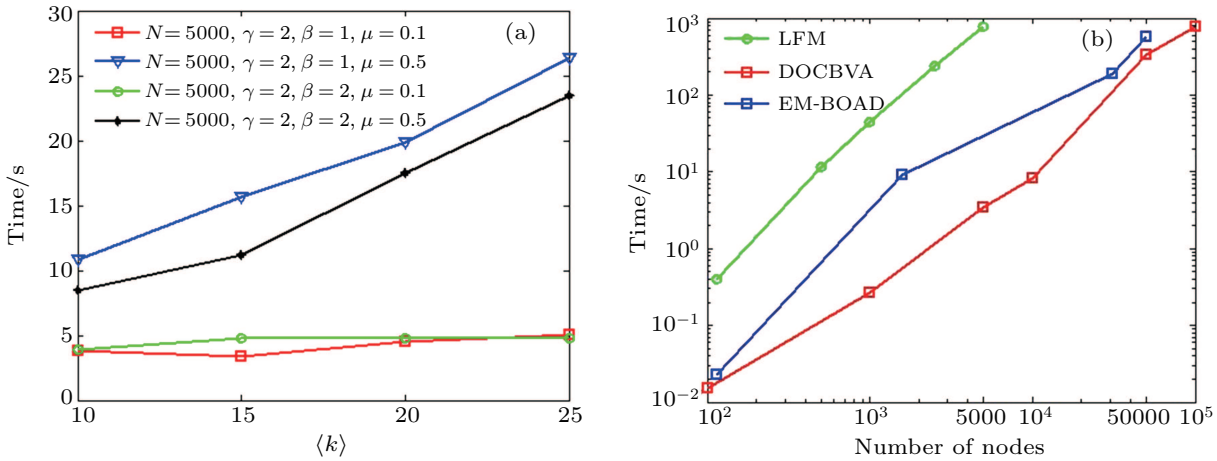


Fig. 7. (color online) The execution time of our method on artificial networks. (a) N is the number of nodes, γ is the minus exponent for the degree sequence, β is the minus exponent for the community size distribution, μ is the mixing parameter, and $\langle k \rangle$ is the average degree of network. (b) The used networks are LFR benchmark graphs[23] with mixing parameter of 0.1, average degree of 15, maximum degree of 50, minimum degree of 20, exponent for the degree distribution of 2, and exponent for the community size distribution of 1.

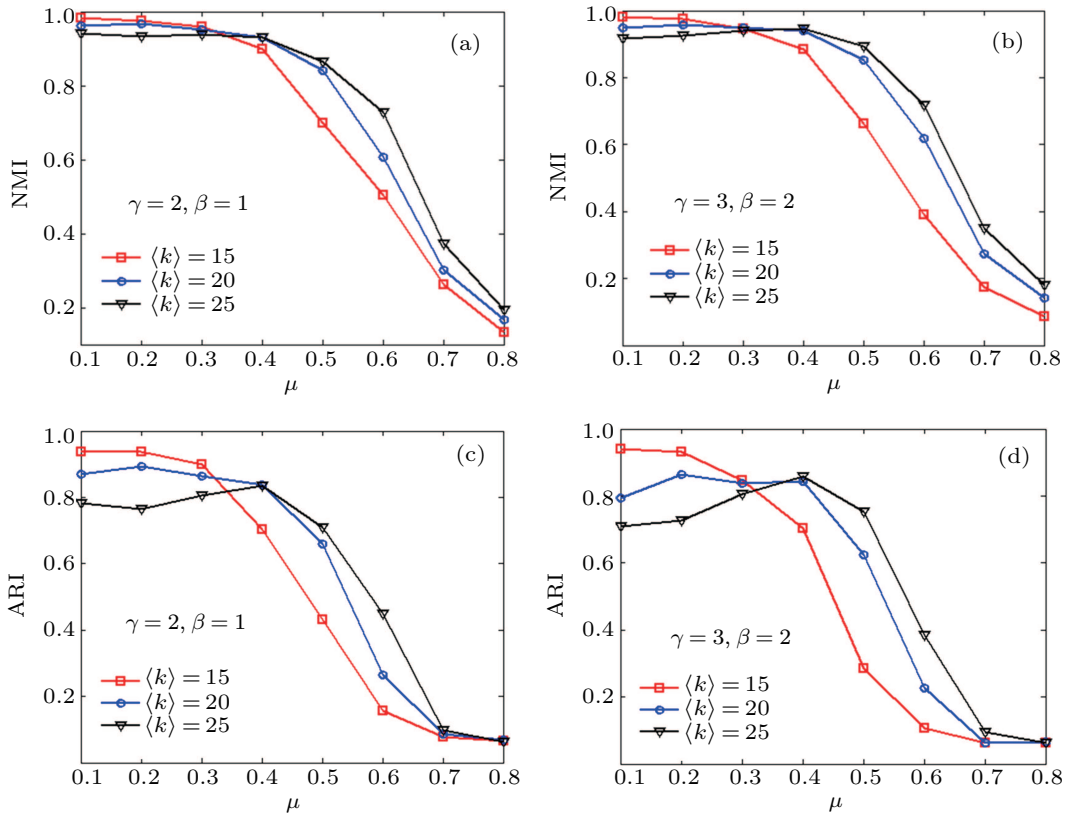


Fig. 8. (color online) Test of our algorithm on the LFR benchmark. The number of nodes is $N = 1000$, γ is the exponent for the degree distribution, β is the exponent for the community size distribution, and $\langle k \rangle$ is the average degree. Although not shown in the figure, the average number of overlapping nodes is 10 in all networks. We use two standards NMI (top) and ARI (bottom) simultaneously to measure the detected performance.

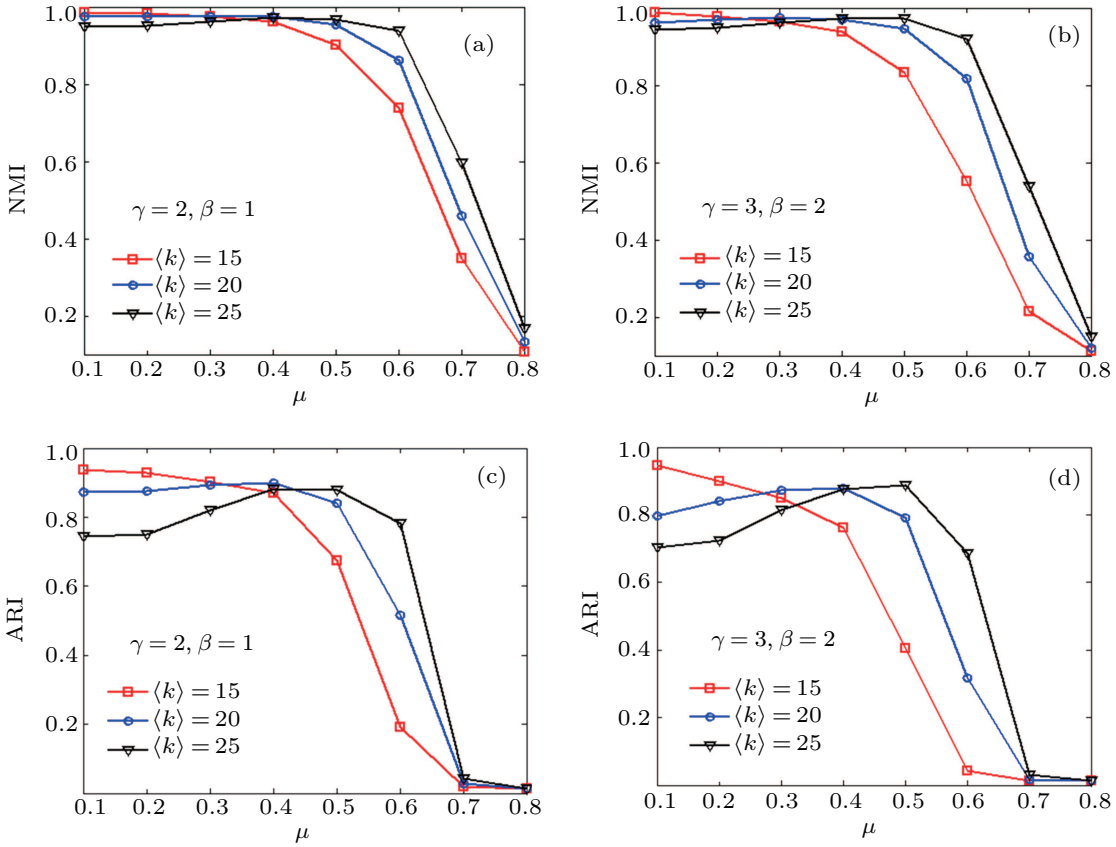


Fig. 9. (color online) Test of our algorithm on the LFR benchmark. The number of nodes is $N = 5000$. Two standards NMI (top) and ARI (bottom) are used simultaneously to measure the detected performance.

To measure the similarity of ground truth communities and found communities, we use normalized mutual information (NMI), an information theoretic similarity measure. In order to measure the performance of overlapping communities, we use the extensional NMI realized by Eustace *et al.*[29,32] The adjusted rand index (ARI)[30–32] is also used in our method. Similar to NMI, the value of ARI is also in the range between 0 and 1. Two partitions are mutually independent when ARI is close to 0 and vice versa.[33–35] The results are shown in Figs. 8 and 9. From the figures, we can draw a conclusion that the proposed method gives good results, especially for large size systems.

Table 3. The overlap coverage of our method test on the LFR benchmark.

n	on	μ					
		0.1	0.2	0.3	0.4	c0.5	0.6
1000	10	11/0.7	9/4.7	2/0.2	1/0.1	1/0.1	1/0.1
5000	50	20/11.3	31/18.4	9/7.5	1/0.1	1/0.1	0/0

Table 3 gives the overlap coverage of our method. In the generated networks, the numbers of nodes (n) are set as 1000 and 5000. The mixing parameters μ ranges from 0.1 to 0.6. The numbers of the overlapping nodes are set as 10 and 50, respectively. The maximum and average numbers of overlapping nodes are obtained, respectively. As show in Table 3, our algorithm can detect the overlapping nodes effectively.

5. Conclusions

A new DOCBVA algorithm has been proposed to detect overlapping community structures with high efficiency. The most significant improvement of our method is the decrease of the running time. Our algorithm needs to find the seed communities and then extend them. Compared with previous algorithms, our method adopts a new parameter to estimate the intimacy between nodes and communities. The absorbing parameter α is so loose that it can accurately describe the relationship of nodes and communities. We use DOCBVA to analyze real-world networks and artificial networks. From the results we can conclude that the proposed method gives good results.

References

[1] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[2] Wang X Y and Zhao T F 2017 *Commun. Nonlinear Sci. Numer. Simul.* **48** 63
[3] Newman M E J 2003 *SIAM. Rev.* **45** 167
[4] Cui Y Z and Wang X Y 2014 *Physica A* **407** 7
[5] Shen X J, Wang Z F and Wu L N 2006 *Int. J. Mod. Phys. C* **17** 1055
[6] Chen Z, Jia M Y and Yang B 2015 *Comput. Sci. Inf. Syst.* **12** 843
[7] Li J Q, Wang X Y and Eustace J 2013 *Physica A* **392** 6125
[8] Girvan M and Newman M E J 2002 *Proc. Natl. Acad. Sci. USA* **99** 7821
[9] Fortunato S 2010 *Phys. Rep.* **486** 75
[10] Palla G Derényi I Farkas I and Vicsek T 2005 *Nature* **435** 814
[11] Zhang Z W and Wang Z Y 2015 *Physica A* **421** 25
[12] Zhang X W, You H B and Zhu W 2015 *Physica A* **421** 233
[13] Kernighan B W and Lin S 1970 *Bell. Syst. Tech. J.* **49** 291
[14] Fiedler M 1973 *Czech Math. J.* **23** 298
[15] Nguyen N P, Dinh T N, Nguyen D T and Thai M T 2012 *IEEE Third Int. Conf. Soc. Computing*, January 3, 2012, Boston, USA, pp. 201, 35
[16] Wang X Y and Li J Q 2013 *Physica A* **392** 2555
[17] Lu J G 2005 *Chin. Phys. B* **14** 703
[18] Lancichinetti A, Fortunato S and Jnos K 2009 *New J. Phys.* **11** 033015
[19] Newman M E J and Girvan M 2004 *Phys. Rev. E* **69** 026113
[20] Chen D B and Shang M S 2010 *Physica A* **389** 4177
[21] Fortunato S and Barthélemy M 2007 *Proc. Natl. Acad. Sci. USA* **104** 36
[22] Newman M E J 2004 *Phys. Rev. E* **70** 056131
[23] Shen H, Cheng X and Cai K 2009 *Physica A* **388** 1706
[24] Lancichinetti A, Fortunato S and Radicchi F 2008 *Phys. Rev. E* **78** 046110
[25] http://www-personal.umich.edu/~mejn/netdata/
[26] Zachary W W 1977 *J. Anthropol. Res.* **33** 452
[27] Wang Y W, Liu M and Liu Z W 2014 *Appl. Math. Comput.* **246** 572
[28] Yang W, Wang Y W and Zeng Z G 2015 *Neurocomputing* **164** 252
[29] Eustace J, Wang X and Cui Y 2015 *Physica A* **421** 510
[30] Li Z C and Tang J H 2015 *IEEE Trans. Multimedia* **17** 1989
[31] Li Z C and Tang J H 2017 *IEEE Trans. Image Process.* **26** 276
[32] Yang W, Wang Y W and Xiao J W 2017 *Int. J. Control* **90** 1
[33] Wirth-Lima A J, Silva M G and Sombra A S B 2018 *Chin. Phys. B* **27** 023201
[34] Liao H, Shen Q, Wu X T, Chen B K and Zhou M Y 2017 *Chin. Phys. B* **26** 110505
[35] Zhang X Y, Meng Y Y and Zhang H 2011 *Chin. Phys. Lett.* **28** 12070