Data & Knowledge Engineering 117 (2018) 183–194

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

An overlapping community detection algorithm in complex networks based on information theory

HongFang Zhou[a,b,*], Yao Zhang[a,b], Jin Li[a,b]

[a] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, 710048, China
[b] Shaanxi Key Laboratory of Network Computing and Security Technology, Xi'an, 710048, China

ARTICLE INFO

Keywords:
Overlapping community detection
Complex networks
Clustering
Data mining

ABSTRACT

In this paper, a new algorithm for overlapping community detection is proposed. First, we propose a node importance evaluation matrix to calculate the important degree for each node; second, we put forward the difference function to detect overlapping points in complex networks; finally, we use triangle principle to detect communities in complex networks. We adopt two measures of Normalized Mutual Information and Modularity to evaluate the algorithm. The experimental results show that our algorithm has a good performance on detecting overlapping community.

1. Introduction

In real life, many real-world networks exist in the form of complex networks, such as private relationship network in social systems, a food chain network in a biological system and World Wide Web etc. These complex networks present a community structure, i.e. vertices groups that have a higher density of edges within them and a lower density of edges between them. Community detection is useful for understanding the properties of network structure and predicting the behaviors of networks [1].

The problem of community structure detection in complex networks has attracted the attention of researchers. Researchers have proposed many community detection algorithms. These methods have been successfully applied in some real complex networks. However, many of the research work mainly focus on hard partition of complex networks (a node can only be divided into a community). But in fact, one node may belong to multiple communities in the real networks. For example, in social networks, one person is usually involved in several social groups such as family, colleagues and friend [2]. Therefore, Kelley et al. [3] pointed out that overlap is indeed an important feature of many real-world networks. Palla et al. [4] introduced the concept of overlapping community and proposed the clique percolation method.

Overlapping community detection algorithms mainly include clique percolation method [2,5,6], block model [7,8], edge clustering algorithm [9,10] and label propagation algorithm [11,12]. The specifications of the different applications are mainly based on the overlapping ratio between different communities. In some approaches, it is required that nodes belonging to multiple communities are strictly restricted. And in other methods, it prefers highly overlapping community structures. In this paper, a new overlapping community detection algorithm based on information theory is proposed. The important node is taken as the cluster center. It is closer to the real community structure. The contributions of the paper are as follows.

(1) The node importance contribution matrix is proposed to evaluate the node importance.
(2) The difference function is proposed to detect the overlapping points.

[*] Corresponding author. School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, 710048, China.
E-mail addresses: zhouhf@xaut.edu.cn (H. Zhou), zhangyao_1108@163.com (Y. Zhang), 286374431@qq.com (J. Li).

https://doi.org/10.1016/j.datak.2018.07.009
Received 1 March 2017; Received in revised form 6 July 2018; Accepted 24 July 2018
Available online 27 July 2018

In this paper, the community detection algorithm we proposed uses the triangle principle for community partition. The algorithm is simple and it is easy to be understood and realized. The remaining parts of the paper are organized as follows. Section 2 introduces the related work of the paper. Section 3 introduces our algorithm in details. Section 4 is the experimental analysis. Finally, the fifth part is the conclusion.

2. Related works

Some research on complex networks has attracted the attention of researchers. One of the most important tasks in complex network analysis is the community detection. In the past decade, scholars proposed many effective community detection algorithms. Among them, the representative methods include clustering-based method [13,14], modularity-based method [15], spectral algorithm [16–18], dynamic algorithm [17–22], statistical inference-based method [18] and matrix factorization method [19].

The communities of complex network can be divided into overlapping and non-overlapping ones. Most existing community detection algorithms are designed to get the hard partition of the network. That is, a node can only belong to a community. Radicchi et al. [20] proposed a local algorithm. It uses edge clustering coefficient and removes the edges which own the lower clustering coefficient. Pons and Latapy [21] use the method based on random walks to compute the nodes' similarity. This method belongs to the hierarchical clustering algorithm. In 2002, Newman and Girvan [22] proposed the Modularity to evaluate community quality. Some methods [23,24] adopt the optimized modularity measure to detect community structures. And it is useful for us to understand the community structures in the networks. Fortunato et al. [25] proposed a variant one on method presented by Girvan and Newman. It uses the information centrality. Kernighan-Lin [26] proposed a heuristic algorithm. It partitions the networks using a greedy algorithm based on the known community number. The method [27] proposed by Raghavan et al. uses the measure of label propagation. First, each node in the network is given an independent label. After each step of the algorithm is executed, each node is labeled as the label which the majority of its neighbors own.

Although the algorithms mentioned above have been applied in the real-world network, they are unable to detect overlapping communities. However, a node may belong to a community. The overlapping communities do exist in some real networks. For example, a member of a social network can belong to both family and hobby groups and a protein may interact with multiple protein complexes. In recent years, some methods have been proposed for overlapping community detection. They may contain five categories which are link partitioning, clique percolation, fuzzy detection, local optimization and agent-based algorithm. CPM (Clique Percolation Methods) [4] assumes that edges in the same community are able to form a clique, but edges in the different communities are unable to form clique. Wang et al. [28] proposed ACC algorithm to detect overlapping communities structures by using two new concepts of the maximum subgraph and the clustering coefficient.

The link partition methods use links instead of nodes to partition community. If node links are partitioned into multiple groups, a node is overlapping. Ahn et al. [9] proposed a link partitioning method to detect overlapping community in a network. Lancichinetti et al. [15] proposed a method based on a local benefit function to find overlapping communities. The fuzzy detection methods use a membership vector to measure the case in which a node belonging to multiple communities. Zhang et al. [29] provided NMF (Nonnegative Matrix Factorization) to detect overlapping communities. Zarei et al. [30] also proposed an algorithm based on NMF to detect overlapping communities. This method uses Laplacian matrix in a given network. Besides, NMF can also be used to detect communities on large-scale networks. Xie et al. [31] proposed an extended LPA method, which relies on the dynamic interaction process of speaker-listener to discover overlapping communities. Besides, there exist some other overlapping community detection algorithms, such as RaRe-IS algorithm [32]. This algorithm consists of two stages which are initialization and improvement. In the initialization step, the algorithm applies RaRe to build a set of seed clusters. In the improvement step, the algorithm applies a IS method. The IS method updates each seed cluster by adding or removing one node at a time. Its disadvantage is that the values of several parameters need to be adjusted. Chao et al. [33] proposed NLA method. The method adopts the measure of node location analysis and evaluates the node mass using Page Rank algorithm. The community affiliation of the node is determined by the location of the peak-valley in its network topology.

3. Algorithm

In this section, we propose an overlapping community discovery algorithm based on a triangle principle. Two adjacent nodes and their shared neighbor nodes can form a triangle to determine whether the two nodes belong to the same community. In the community detection algorithm, the node importance contribution matrix is proposed.

3.1. Node importance

Compared with other nodes in the network, the important nodes can affect greatly the structure and function of the networks. We can measure the importance of a node by the number and the importance of its neighbor nodes. The more neighbor nodes a node has, the more important it is and the more important the neighbor node is, the more important it is.

Definition 1. Given a network $G = (V, E)$, $V = \{v_i | i = 1, \ldots, n\}$ represents a collection of nodes, n represents the number of nodes,

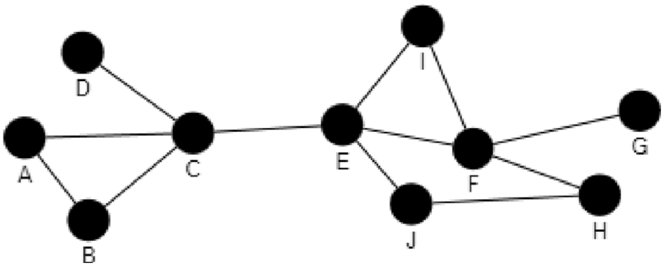


Fig. 1. Relationship of social networks.

$E = \{v_i, v_j | v_i, v_j \in V\}$ represents a collection of edges, $m =$ |E| represents the number of edges, $d(v_i)$ represents the degree of the node v_i, then the direct contribution of v_i to v_j is defined as follows.

$$S(v_i, v_j) = J(v_i, v_j) + F(v_i, v_j) \tag{1}$$

Here, $J(v_i, v_j)$ quotes Jaccard correlation and indicates the effect of the node v_i on the node v_j. That is, $J(v_i, v_j) = \frac{|d(v_i) \cap d(v_j)|}{|d(v_i) \cup d(v_j)|}$; $J(v_i, v_j)$ represents that the common adjacent points of the nodes v_i and v_j accounted for the proportion of all their adjacent points, the more the number of common adjacent points a node has, the greater the influence of a node is. Besides $J(v_i, v_j) \in [0,\ 1]$; and $F(v_i, v_j) = \frac{1}{d(v_i)}$ must be satisfied.

In real life, the influence of the node v_i on the node v_j is different from the influence of the node v_j on the node v_i. Let's take an example. In a real social network, v_i and v_j have known each other. But the social circle of v_i is very large, and that of v_j is very small. Then the probability of the impact of v_i on v_j is greater than the probability of the impact of v_j on v_i. Therefore, we can conclude that the impact of v_i on v_j is greater than the impact of v_j on v_i, if the degree of node v_i is much higher than that of the node v_j in complex networks. In addition, there may exist a bridge between v_i and v_j. As shown in Fig. 1, the graph represents a social network. For nodes C and E, we can know that $J(C, E)$ equals 0 in accordance with the above formula. And it is obviously unreasonable. Therefore, we add an additional node p in the network so that the node p is connected with every node in the network. Any two connected nodes have at least one common node. The improved method avoids the situation described in Fig. 1. As shown in Fig. 2, the number of common nodes between any pair of nodes in the network is increased by 1.

In summary, we modified the $J(v_i, v_j)$ in Eq. (1). And the improved $J(v_i, v_j)$ is shown in below.

$$J(v_i, v_j) = \frac{|d(v_i) \cap d(v_j)\ | + 1}{|d(v_i) \cup d(v_j)\ | + 2} \tag{2}$$

In Eq. (2), the direct contribution of the interconnected nodes only takes into account the direct impact of adjacent nodes, but it did not consider the characteristics of the entire network. Based on this, we use the mutual information theory to think over the influence of indirect relationship. We model the complex network as a graph $G = (V, E)$. $V = \{v_i | i = 1, \ldots, n\}$ represents a set of nodes and n represents the number of nodes. $E = \{v_i, v_j | v_i, v_j \in V\}$ represents the collection of edges and $m =$ |E| represents the number of edges. In the information theory, the information is defined as the movement of things or the uncertainty description, as the message is sent to destination from the source over the channel, the uncertainty can be eliminated and further information can be obtained.

For any node, the complex network can be viewed as a complex communication system model, if we consider its m edges as m data stream, we can use the information theory to evaluate the node importance. This kind of evaluation method based on mutual

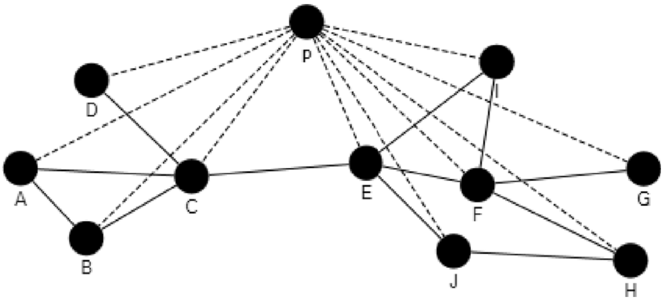


Fig. 2. Relationship of social network after adding additional the node p.

information takes full account of the global network. The evaluation method is used to evaluate the importance of nodes by information amount. The information amount in each node is determined by its edges.

Definition 2. Given an unweighted, undirected network $G = (V, E)$, the mutual information $I(v_i, v_j)$ of the node v_i to the node v_j is defined as follows.

$$I(v_i, v_j) = \begin{cases} \log d(v_i) - \log d(v_j), & v_i \textit{ and } v_j \textit{ are directly connected.} \\ 0, & \textit{otherwise} \end{cases} \tag{3}$$

The information amount $I(v_i)$ of the node v_i is the sum of the mutual information between the node v_i and other nodes. And $I(v_i)$ is defined as follows.

$$I(v_i) = \sum_{j=0}^{n} I(v_i, v_j) \tag{4}$$

The node information amount represents the contribution of the nodes in the whole network. We give the definitions of contribution degrees of the direct neighbors and the information amount of a node. Considering the importance contribution of a node to other nodes, we construct the node importance contribution matrix by referring to the adjacency matrix, denoted as CM. The contribution degree and the node information quantity are combined. The contribution matrix of node importance is given as follows.

$$CM = \begin{bmatrix} I_1 & \delta_{12}S_{12}I_2 & \cdots & \delta_{1n}S_{1n}I_n \\ \delta_{21}S_{21}I_1 & I_2 & \cdots & \delta_{2n}S_{2n}I_n \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1}S_{n1}I_1 & \delta_{n2}S_{n2}I_2 & \cdots & I_n \end{bmatrix} \tag{5}$$

In Eq. (5), $\delta_{ij} = 1$ if nodes i and j are directly connected, otherwise $\delta_{ij} = 0$. S_{ij} represents the direct contribution of one node i to another node j. The node importance contribution matrix CM represents the extent of contribution of each node to others. Then the importance degree M_i of the node i is defined as follows.

$$M_i = \sum_{j=1}^{n} CM_{ij} = I_i + \sum_{j=1, j\neq i}^{n} \delta_{ij}S_{ij}I_j \tag{6}$$

It can be seen that the importance of the node v_i is composed of the information of itself and the sum of the contributions of all adjacent nodes to the node v_i.

3.2. Community detection method

We apply the node importance to the community detection. The higher the node importance is, the more likely it is to be the clustering center.

In the community detection process, nodes that are not assigned to two or more communities are extensible, and these are referred to as central nodes or extended nodes. Nodes that are assigned to two or more communities are called as non-extensible nodes. In general, if a node is more important, it often has more links with other nodes. Node importance is a direct reflection of a nodes' influence. The greater the node importance is, the stronger the node's influence is. Therefore, the node importance can be measured by the concrete importance value of the node. In this study, we select the node with maximum importance as the central node (extensible node). The central node of the expansion method is shown in the rules.

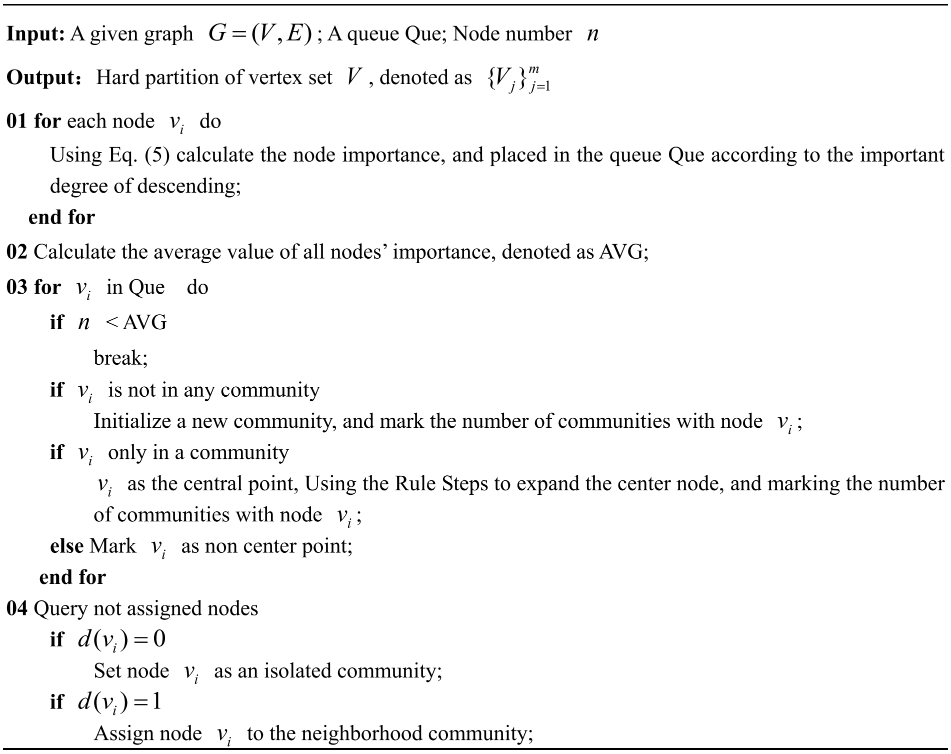
Rule Steps:

(1) Select a center node v_i, which has maximum importance. If the node v_i is not allocated, then we initialize the community C and add the node v_i into the C. If the node v_i is assigned to a community, then we set the community containing the v_i as the initial community C;
(2) If the node v_i and its neighbor node A have shared neighbor node B and the three nodes constitute a triangle, the nodes A and B should join the community C, otherwise the node A does not join the community C;
(3) Test another neighbor node, repeat step (2), until all neighbor nodes are tested.

Community structure in a complex network has the characteristics of high connection density. A dense community structure usually contains complete subgraphs. Triangles are the basic elements of complete subgraphs. Therefore, we can judge whether the vertices belong to the community according to whether the vertices form a triangle. The neighbor node and the center node are considered to be triangular if the neighbor node and the center node have their common neighbor nodes. We consider that the neighbor node and the central node belong to the same community. This is done in accordance with Eq. (5) to calculate the importance degrees of all the nodes. The nodes are ranked in descending order referring to their importance, and the node with the highest importance degree is selected each time.

If the node is not assigned to any community, a new community is initialized with the node, and then the community created by the node is expanded. If the node and its neighbor node A have a common neighbor node B and the three nodes can form a triangle, the nodes A and B join the community. The algorithm does not terminate until the importance of nodes in the queue is less than the average of all node importance. The concrete steps of the algorithm are as follows.

Algorithm 1. community detection method.



```
Input: A given graph $G=(V,E)$; A queue Que; Node number $n$
Output：Hard partition of vertex set $V$, denoted as $\{V_j\}_{j=1}^{m}$
01 for each node $v_i$ do
        Using Eq. (5) calculate the node importance, and placed in the queue Que according to the important
        degree of descending;
   end for
02 Calculate the average value of all nodes' importance, denoted as AVG;
03 for $v_i$ in Que  do
        if $n$ <AVG
            break;
        if $v_i$ is not in any community
            Initialize a new community, and mark the number of communities with node $v_i$;
        if $v_i$ only in a community
            $v_i$ as the central point, Using the Rule Steps to expand the center node, and marking the number
            of communities with node $v_i$;
        else Mark $v_i$ as non center point;
      end for
04 Query not assigned nodes
        if $d(v_i)=0$
            Set node $v_i$ as an isolated community;
        if $d(v_i)=1$
            Assign node $v_i$ to the neighborhood community;
```

3.3. Overlapping node detection

If the node v_i is an overlapping node, there must be several center nodes for v_i to choose which center nodes for their own center nodes are not very different. Therefore, we use the difference function to express the difference between the node v_i and each center point to judge whether v_i is a overlapping node. The difference function is defined as Eq. (7).

$$\mathit{difference}\,(v_i,\ C_i,\ C_k) = |J\,(v_i,\ C_i) - J\,(v_i,\ C_k)| \tag{7}$$

Where C_i represents the center point of the vertex v_i determined by this algorithm. C_k is the center point of other community. $J\,(v_i, C_i)$ expresses the influence degree of vertex v_i and the center point C_i. The difference function $\mathit{difference}(v_i, C_i, C_k)$ means the difference between C_i as the center point and C_k as the center point for the node v_i. At this point, we can give the characterization of overlapping points. It is considered that when the condition $\mathit{difference}(v_i, C_i, C_k) \leq \theta$ (θ is the threshold) is satisfied, and $C_i \neq C_k$, the vertex v_i is the overlapping vertex of the community with C_k as the center point. Here, the parameter θ as a threshold is a non-negative real number. It controls the scale of overlapping points. With the increase of θ, there are more and more overlapping points. The overlapping point detection algorithm is shown in Algorithm 2.

Algorithm 2. Overlapping points detection algorithm.

```
Input: The hard partition of vertex set $V$, denoted by $\{V_j\}_{j=1}^{m}$, where $m$ is the number of communities
    detected by Algorithm 1; threshold $\theta$.
Output: Overlapping point set $OverlapPointSet_j$
(Use the previously mentioned Algorithm 1 for hard partitioning of the vertex set $V$)
01 Run the Algorithm 1 to get the hard partition $\{V_j\}_{j=1}^{m}$ of the vertex set $V$, and store the center points
   in the set C;
02 for each community $V_j \in \{V_j\}_{j=1}^{m}$ do
     Initializes the overlapping point collection $OverlapPointSet_j = \Phi$;
   end for
03 for each vertex $v_i \in V$ do
       for each center point $j \in C$ do
           if $difference(v_i, C_i, j) \le \theta$, and $j \ne C_i$
               $OverlapPointSet_j = OverlapPointSet_j \cup \{v_i\}$
           end if
       end for
   end for
04 Returns a collection of overlapping points $\{OverlapPointSet_j\}_{j=1}^{m}$
```

3.4. Complexity analysis

After a description of the algorithm, given a graph G contains n vertices, m edges. The average value of all nodes' importance is e, the calculation of time complexity for the node importance is $O(n^3)$. The time complexity of node sorting is $O(n \log n)$. For an extended node, it needs to find its neighbor nodes and their common neighbor. The time complexity is $O(e^2)$, and the time complexity of extending n nodes is a $O(ne^2)$. For overlapped nodes, it needs to calculate the number of communities to which it belongs, and the computational complexity is $O(ne)$. The time complexity of finding unassigned nodes is $O(ne)$. Therefore, the total computational complexity of our algorithm is approximately $O(n^3)$.

4. Experiment

We use the experiment to evaluate the proposed algorithm. All experiments are performed on a PC with Windows XP, an i3 CPU (2.16 GHz) and 1 GB main memory. The programming environment is JDK 1.7.

4.1. Datasets

In this paper, the following three datasets are selected to complete the experiment.

(1) Zachary's karate club dataset: Zachary's karate club network contains 34 nodes and 78 edges. A node represents a club member and edges represent the interrelationships among members. Due to the contradictory whether increased club member fees between club managers and coaches. Finally, the club network is split into two smaller clubs.
(2) Dolphin network: the dolphin network is derived from Lusseau's study about the behavior of 62 bottlenose dolphins living in New Zealand's Doubtful Sound. The dolphin network includes 62 nodes and 159 edges. Nodes represent the 62 dolphins and each edge represents the interaction between the two dolphins.
(3) College football network: College football network consists of 115 nodes and 613 edges. Each node represents the college team of the United States in the 2000 football season, and each edge represents that the corresponding two teams have had at least one game.

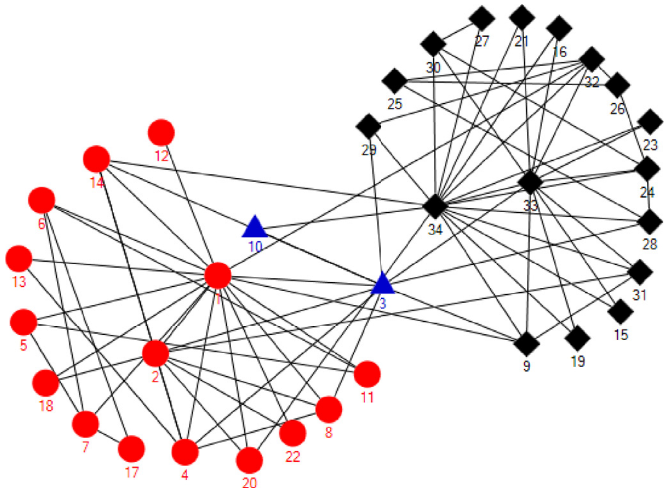


Fig. 3. Results of network community detection on Zarchary's Karate Club.

4.2. Evaluation measures

We select two widely used measures, NMI and Modularity, to evaluate the community quality.

(1) NMI(Normalized Mutual Information): NMI is the evaluation criterion of the result of network partition. NMI can evaluate the similarity between the detected community and the real community. Given two partitions A and B of a network, N is defined as a confusion matrix, where the rows correspond to the “real” communities, and the columns correspond to the “found” communities. N_{ij} is the number of nodes in the real community i that appear in the found community j. Normalized Mutual Information $I(A, B)$ is defined as follows:

$$I(A,B)=\frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}N_{ij}\log\left(\frac{N_{ij}*N}{N_{i.}*N_{.j}}\right)}{\sum_{i=1}^{C_A}N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{C_B}N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \tag{8}$$

Where the number of real communities is denoted C_A and the number of found communities is denoted by C_B, the sum over row i of matrix N_{ij} is denoted by $N_{i.}$, and the sum over column j is denoted by $N_{.j}$. If the found partitions are identical to the real communities, then $I(A, B)$ takes a value of 1. If the partition found by the algorithm is totally independent of the real partition, then $I(A, B)$ takes a value of 0.

(2) Modularity: Modularity function is a commonly used to measure the quality of community. The higher the value of modularity is, the more accurate the result of community detection is. Shen et al. gives the evaluation criteria of overlapping communities. Overlapping community modularity EQ is defined as follows.

$$\mathrm{EQ}=\frac{1}{2m}\sum_{i}\sum_{v\in c_i,w\in c_j}\frac{1}{Q_vQ_w}\left[A_{vw}-\frac{k_vk_w}{2m}\right] \tag{9}$$

Where k_v and k_w are the degrees of node v and node w. A_{vw} is the adjacency matrix of the network. If the node v and node w are adjacent, m is 1; otherwise m is 0; c_i represents the ith community. Q_v indicates how many communities the node v belongs to. m is the total number of edges in the network.

4.3. Experimental results and analysis

First, we analyze the results of the proposed algorithm on community detection on three real datasets.

As shown in Fig. 3, the experimental dataset is the Zachary's Karate Club network in which we set $\theta = 0.07$. The proposed

Table 1
Community partition results on Zachary's Karate Club.

Community number	Internal nodes	Overlapping nodes
1	1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22	
2	9, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34	3, 10

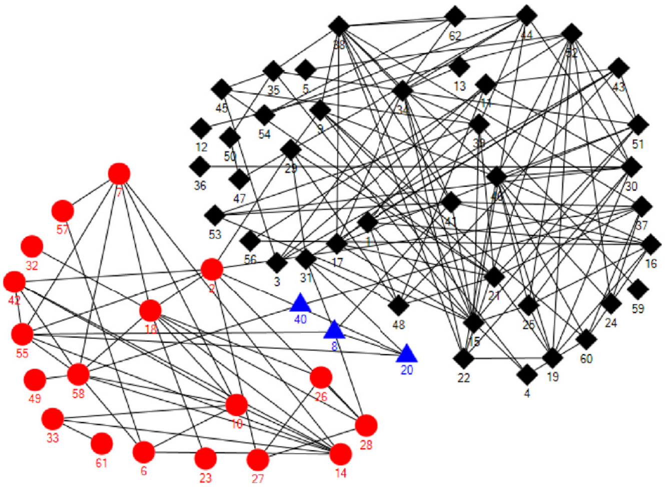


Fig. 4. Results of network community detection on Dolphin network.

Table 2
Community partition results on Dolphin network.

Community number	Internal nodes	Overlapping nodes
1	2, 6, 7, 10, 14, 18, 23, 26, 27, 28, 32, 33, 42, 49, 55, 57, 58, 61	
2	1, 3, 4, 5, 9, 11, 12, 13, 15, 16, 17, 19, 21, 22, 24, 25, 29, 30, 31, 34, 35, 36, 37, 38, 39, 41, 43, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 56, 59, 60, 62	8, 20, 40

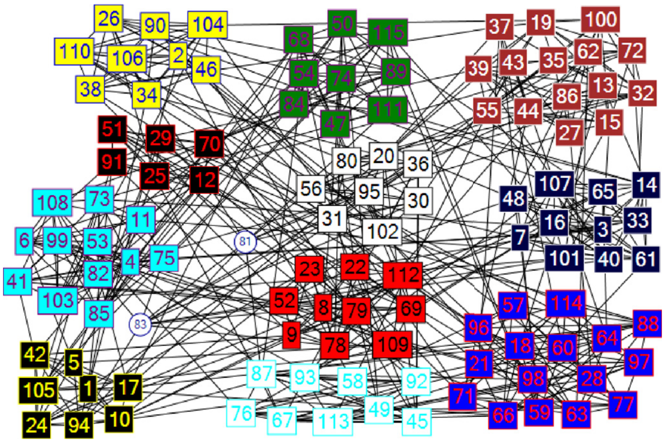


Fig. 5. Results of network community detection on College football network.

algorithm divides the network into two communities 'C_1' and 'C_2' which take node '1' and '34' as the centers, respectively. The set of overlapping points is {3, 10}. As shown in Table 1, two partitioned communities are displayed in details. Node '34' has the biggest importance, and its value is 31.3584. Similarly, the importance of node '1' is 27.2718. We can find two nodes (node '34' and node '1') are the core nodes. We can also see that nodes '2', '3' and '33' are also important shown in Fig. 3. Node '33' is the combined representative node in community 'C_2'. Node '3' is the overlapping node and it is almost the same as the connections between the two communities, so it can be the hub node which contacts with two communities.

As is shown in Fig. 4, the experimental dataset is the Dolphin network in which we set $\theta = 0.07$. The proposed algorithm divides the network into two communities which take node '15' and node '18' as the centers respectively. The set of overlapping points is {8, 20, 40}. As shown in Table 2, two partitioned communities are displayed in details. Besides, node '15' has the biggest importance, and its value is 10.5381. So, it is a central point. Node '18' is the second central node and its importance is 9.2203. So, it is second only to node '15'. Besides, nodes '46', '52' and '58' are potential ones. Their importances are 7.1089, 8.1378 and 7.4982 respectively. The neighboring set of node '55' is {2, 8, 20, 42, 58}, nodes '2', '42' and '58' are internal nodes in community 'C_1'. Node '58' is the vice-representative node in community 'C_1'. Therefore, it can be more reasonable to partition node '55' as an internal node in community 'C_1' instead of an overlapping node between community 'C_1' and 'C_2'.

Table 3
Community partition results on College Football network.

Community number	Internal nodes	Overlapping nodes
1	12, 25, 29, 51, 70, 91	
2	8, 9, 22, 23, 52, 69, 78, 79, 109, 112	
3	2, 26, 34, 38, 46, 90, 104, 106, 110	
4	18, 21, 28, 57, 59,60, 63, 64, 66, 71, 77, 88, 96, 97, 98, 114	
5	47, 50, 54, 68, 74, 84, 89, 111, 115	81, 83
6	3, 7, 14, 16, 33, 40, 48, 61, 65, 101, 107	
7	1, 5, 10, 17, 24, 42, 94, 105	
8	4, 6, 11, 41, 53, 73, 75, 82, 85, 99, 103, 108	
9	13, 15, 19, 27, 32, 35, 37, 39, 43, 44, 55, 62, 72, 86, 100	
10	45, 49, 58, 67, 76, 87, 92, 93, 113	
11	20, 30, 31, 36, 56, 80, 95, 102	

As is shown in Fig. 5, the experimental dataset is the College Football network. There are five independent teams. It is node '24', node '37', node '43', node '81' and node '83' respectively. In the experiment, we set $\theta = 0.07$. The proposed algorithm divides the network into 11 groups. The central nodes of eleven communities are {2, 89, 44, 70, 30, 7, 6, 8, 1, 67, 77} and their importances are {2.0178, 1.4068, 1.5728, 1.0866, 0.7978, 1.4455, 2.3729, 1.4471, 1.7598, 0.7376, 1.0266}. We can find that almost all teams are correctly detected. The set of overlapping points is {81, 83}. As shown in Table 3, the eleven partitioned communities are displayed in details. Nodes '81' and '83' are exactly an independent team.

Secondly, we analyze the effect of θ on the number of overlapping points on three datasets. We show the effect of these values on overlapping points in Table 4.

As it is revealed in Table 4, the overlapping point number is very sensitive to the value of θ. The reason is the difference function.

Table 4
Effect of θ on the number of overlapping points on three datasets.

Datasets	θ	Number of overlapping points	Overlapping point set
Zarchary	$0 < \theta < 0.04$	1	{3}
	$0.04 \le \theta < 1$	2	{3,10}
dolphin	$0 < \theta < 0.06$	2	{8,40}
	$0.06 \le \theta < 1$	3	{8,20,40}
college football	$0 < \theta < 0.05$	1	{81}
	$0.05 \le \theta < 1$	2	{81,83}

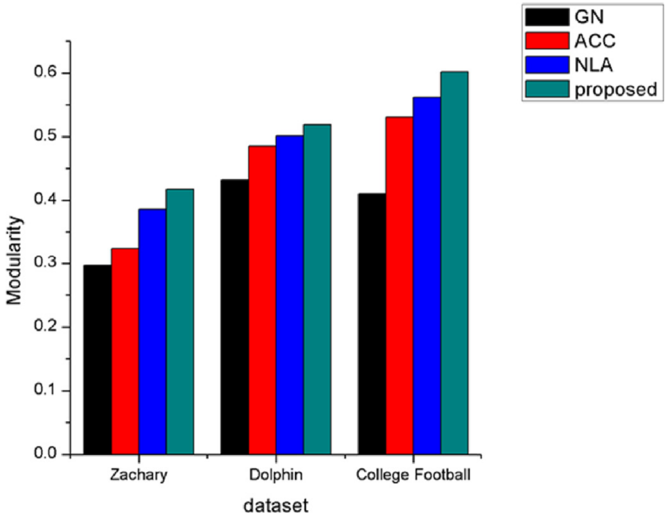


Fig. 6. Comparison of Modularity on three real datasets.

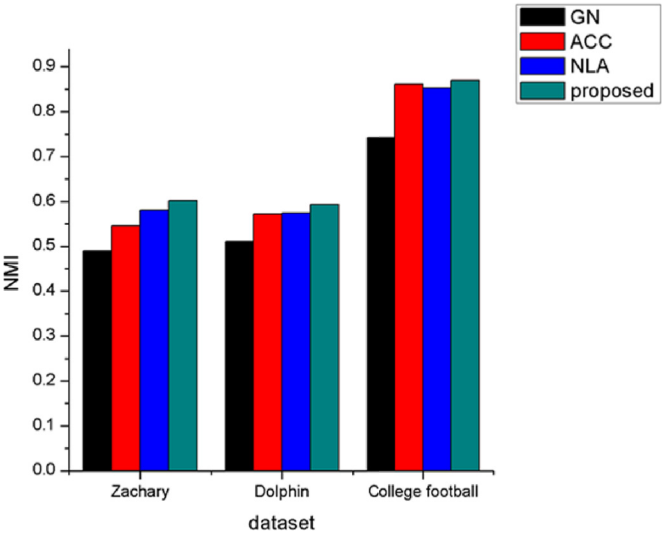


Fig. 7. Comparison of NMI on three real datasets.

With the increase of θ, the number of overlapping points increases but θ must be greater than threshold. Besides, for different network, θ is also different. For Zachary's Karate Club, the threshold of θ is 0.04. The thresholds of Dolphin and College Football are 0.06 and 0.05 respectively.

Finally, we compare the proposed algorithm with three contrast algorithms including ACC algorithm [28], GN algorithm [22] and NLA algorithm [32] on three real datasets. In the experiment, the parameter θ is set to be 0.07. We use modularity and NMI as evaluation criteria. We perform the four algorithms 10 times, and compare the average values of ten results. Figs. 5 and 6 are the results of modularity and NMI respectively.

In Fig. 6, the modularity comparisons are shown about four algorithms on the three real datasets. The higher the modularity is, the more accurate of the community partition is. In the Zachary's Karate Club network, our algorithm outperforms the other three ones. It is shown that our proposed method can preferably find central points of cluster and obtain better community structure. The modularity of GN and ACC is extremely low. It shows that the partitioning results of GN and ACC are poor. On the Dolphin dataset, GN has the worst result. ACC, NLA and our proposed algorithms have slight difference. On the College Football dataset, our algorithm is also superior to other three ones. Its modularity is 0.6. The modularity of NLA is about 0.55, and the result of community detection is not bad. NLA algorithm can almost predict the topological position of most nodes precisely. ACC algorithm is slightly inferior to NLA, Therefore, we can know that ACC algorithm cannot effectively merge two communities into a new larger community. On the whole, GN algorithm performs worst on three datasets, and our proposed algorithm outperforms ACC and NLA.

On the Zachary as shown in Fig. 7, NMI of our proposed method is 0.6, which is slightly higher than that of ACC and NLA. However, GN algorithm is the worst and NMI is less than 0.5. On the Dolphin dataset, the NMI values of ACC and NLA are basically the same. The proposed algorithm is slightly higher than ACC and NLA. The NMI of GN algorithm is worst. Overall, the differences of ACC, NLA and our proposed one for community detection results are very small. On the College Football dataset, NMI of our method reaches 0.9. It shows that the results of our proposed algorithm partitioning the network closer to the real one. ACC is slightly higher than NLA. NMI of ACC reached about 0.85, NMI of GN algorithm is more than 0.7. On the whole, on the three datasets, NMI of GN is the lowest and the quality of the community detection is the worst, our proposed algorithm is the largest, and the community detection quality is better. But from the experimental results, the quality of the community are good when use ACC and NLA to detect.

5. Conclusion

In this paper, a new algorithm for overlapping community detection is proposed. First, we propose a node importance contribution matrix to calculate the similarity between each pair of nodes. Second, the difference function is proposed to detect the overlapping points. Finally, we use triangle principle for detecting communities in complex networks. Our proposed algorithm can effectively detect overlapping communities in real network datasets. Next, we will perform overlapping community detection in weighted and directed networks.

Acknowledgments

The corresponding author would like to thank the support from the National Natural Science Foundation of China under the Grant of 61402363, the Education Department of Shaanxi Province Key Laboratory Project under the Grant of 15JS079, Xi'an Science Program Project under the Grant of 2017080CG/RC043(XALG017), the Ministry of Education of Shannxi Province Research Project under the Grant of 17JK0534, and Beilin district of Xi'an Science and Technology Project under the Grant of GX1625.

References

[1] Z.X. Wang, Z.T. Chen, Y. Zhao, S.D. Chen, A community detection algorithm based on topology potential and spectral clustering, Sci. World J. 9 (2014) 329325.

[2] Z.G. Luo, F. Ding, X.Z. Jiang, J.L. Shi, New progress on community detection in complex networks, J. Natl. Univ. Def. Technol. 33 (1) (2011) 47–52.

[3] S. Kelly, M. Goldberg, M. Magdon-Ismail, Defining and discovering communities in social networks, Handbook of Optimization in Complex Networks, Springer, US, 2011, pp. 139–168.

[4] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–881.

[5] J.M. Kumpula, M. Kivela, K. Kaski, J. Saramaki, Sequential algorithm for fast clique percolation, Phys. Rev. E 78 (2) (2008) 026109.

[6] C. Lee, F. Reed, A. McDaid, N. Hurley, Detecting Highly Overlapping Community Structure by Greedy Clique Expansion, (2010) Preprint.

[7] A. Decelle, F. Krzakla, C. Moore, L. Zdeborova, Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications, Phys. Rev. E 84 (6) (2011) 066106.

[8] P.K. Gopalan, D.M. Blei, Efficient discovery of overlapping communities in massive networks, Proc. Natl. Acad. Sci. Unit. States Am. 110 (36) (2013) 14534–14539.

[9] Y.Y. Ahn, J.P. Bagrow, S. Lehman, Link communities reveal multiscale complexity in networks, Nature 466 (7307) (2010) 761–764.

[10] T. Evans, R. Lambiotte, Line graphs, link partitions and overlapping communities, Phys. Rev. E 80 (2) (2009) 016105.

[11] A. Lancichinetti, F. Radicchi, J.J. Ramasco, S. Fortunato, Finding statistically significant communities in networks, PLoS One 6 (4) (2011) e18961.

[12] T. Népusz, A. Petróczi, L. Négyessy, F. Bazsó, Fuzzy communities and the concept of bridgeness in complex networks, Phys. Rev. E 77 (1) (2008) 016107.

[13] A. Lancichinetti, S. Fortunato, Consensus clustering in complex networks, Sci. Rep. 2 (2012).

[14] J. Kim, T. Wilhelm, Spanning tree separation reveals community structure in networks, Phys. Rev. E 87 (3) (2013) 032816.

[15] A. Lancichinetti, S. Fortunato, J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, N. J. Phys. 11 (3) (2009) 033015.

[16] J.Q. Jiang, A.W.M. Dress, G. Yang, A spectral clustering-based framework for detecting community structures in complex networks, Appl. Math. Lett. 22 (9) (2009) 1479–1482.

[17] D. Lai, C. Nardini, H. Lu, Partitioning networks into communities by message passing, Phys. Rev. E 83 (1) (2011) 016115.

[18] P.K. Gopalan, D.M. Blei, Efficient discovery of overlapping communities in massive networks, Proc. Natl. Acad. Sci. Unit. States Am. 110 (36) (2013) 14534–14539.

[19] F. Wang, T. Li, X. Wang, S. Zhu, C. Ding, Community discovery using nonnegative matrix factorization, Data Min. Knowl. Discov. 22 (3) (2011) 493–521.

[20] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. U.S.A. 101 (9) (2004) 2658–2663.

[21] P. Pons, M. Latapy, Computing communities in large networks using random walks, Lect. Notes Comput. Sci. 3733 (2) (2005) 284–293.

[22] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. U.S.A. 99 (6) (2002) 7821–7826.

[23] Z.H. Wu, Y.F. Lin, H.Y. Wan, S.F. Tian, K.Y. Hu, Efficient overlapping community detection in huge real-world networks, Physica A 391 (7) (2012) 2475–2490.

[24] Y.P. Li, Y.M. Ye, E.K. Wang, Fast computation of modularity in agglomerative clustering methods for community discovery, Int. J. Adv. Comput. Technol. 3 (4) (2011) 153–164.

[25] S. Fortunato, V. Latora, M. Marchiori, Method to find community structures based on information centrality, Phys. Rev. E 70 (5) (2004) 056104.

[26] B.W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, Bell Syst. Tech. J. 49 (2) (1970) 291–307.

[27] U. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (2007) 036106.

[28] Y.Z. Cui, X.Y. Wang, J.Q. Li, Detecting overlapping communities in networks using the maximal sub-graph and the clustering coefficient, Physica A 405 (2014) 85–91.

[29] S. Zhang, R.S. Wang, X.S. Zhang, Uncovering fuzzy community structure in complex networks, Phys. Rev. E 76 (4) (2007) 046103.

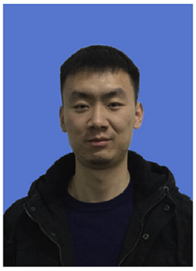
[30] M. Zarei, D. lzadi, K.A. Samani, Detecting overlapping community structure of networks based on vertex-vertex correlations, J. Stat. Mech. Theor. Exp. 2009 (11) (2009) 11013.

[31] J. Xie, B.K. Szymanski, X. Liu, Slpa: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process, 2011 IEEE 11th International Conference on Paper Presented at: Data Mining Workshops, ICDMW, 2011.

[32] J. Baumes, M. Goldberg, M. Krishnamoorty, M. Magdon-Ismail, N. Preston, Finding communities by clustering a graph into overlapping subgraphs, Proceedings of IADIS Applied Computing, 2005, pp. 97–104.

[33] Z.X. Wang, Z.C. Li, X.F. Ding, J.-H. Tang, Overlapping community detection based on node location analysis, Knowl.-Based Syst. 105 (2016) 225–235.

HongFang Zhou received her B.S. and M.S. degrees from Xi'an University of Technology in 1999 and 2002 respectively. And she received her Ph.D. degree from Xi'an Jiaotong University in 2006. She is now an associate professor in Xi'an University of Technology, China. Her research interests include artificial computing, software and theory and heterogeneous information network.



Yao Zhang received the B.S. degree in Computer science and technology from Xi'an University of Posts & Telecommunications in 2015. Now he is studying for Master degree in the School of Computer Science and Engineering, Xi'an University of Technology. His research interests are focus on statistical machine learning and data mining.

Jin Li received his B.S. Degree from Xi'an University of Technology, in 2014. He is a postgraduate of Xi'an University of Technology. His research interests include artificial computing and Data Mining.