Computing and Informatics, Vol. 38, 2019, 1091–1110, doi: 10.31577/cai_2019_5_1091

OVERLAPPING COMMUNITY DETECTION EXTENDED FROM DISJOINT COMMUNITY STRUCTURE

Yan Xing

School of Computer Science and Technology
Civil Aviation University of China
Tianjin, 300300, China
&
School of Computer Science and Technology
China University of Mining and Technology
Xuzhou, Jiangsu, 221116, China
e-mail: yxing425@163.com

Fanrong Meng, Yong Zhou, Guibin Sun, Zhixiao Wang*

School of Computer Science and Technology
China University of Mining and Technology
Xuzhou, Jiangsu, 221116, China
e-mail: {mengfr, yzhou}@cumt.edu.cn, sunguibinbest@qq.com, zhxwang@cumt.edu.cn

Abstract. Community detection is a hot issue in the study of complex networks. Many community detection algorithms have been put forward in different fields. But most of the existing community detection algorithms are used to find disjoint community structure. In order to make full use of the disjoint community detection algorithms to adapt to the new demand of overlapping community detection, this paper proposes an overlapping community detection algorithm extended from disjoint community structure by selecting overlapping nodes (ONS-OCD). In the algorithm, disjoint community structure with high qualities is firstly taken as input, then, potential members of each community are identified. Overlapping nodes are determined according to the node contribution to the community. Finally, adding

* Corresponding author

overlapping nodes to all communities they belong to and get the final overlapping community structure. ONS-OCD algorithm reduces the computation of judging overlapping nodes by narrowing the scope of the potential member nodes of each community. Experimental results both on synthetic and real networks show that the community detection quality of ONS-OCD algorithm is better than several other representative overlapping community detection algorithms.

Keywords: Disjoint community detection, overlapping community detection, potential member, overlapping node

Mathematics Subject Classification 2010: 68-Q87

1 INTRODUCTION

Complex network is a relatively stable relation system which is formed by the interaction between individual members. Many real-world complex systems can be described by the form of complex networks, such as social networks, scientists cooperation networks, web networks, protein interaction networks, etc. [1]. Extensive studies have shown that complex networks not only have the properties of small world [2] and scale-free [3], but they also have the characteristic of community (module or cluster) structure. A community in a network is a group of nodes with dense connections within the group and only sparse connections between them [4]. Research on community detection of complex networks has important theoretical significance and wide application prospect. Community detection in complex networks can help to explore the structure and function of the network, find the hidden laws and predict their behavior [1]. Therefore, community detection is the basis and key of network analysis.

Traditional community detection algorithms divide the network into a number of disjoint communities. Each node can only belong to one community. Representative methods include modularity optimization algorithms [5, 6, 7], spectral clustering algorithms [8, 9], hierarchical partition algorithms [10, 11], label propagation based algorithms [12, 13], information theory based algorithms [14], and so forth. However, in many real complex networks, communities are usually not isolated from each other, but overlap and cross each other. Some nodes may belong to many communities at the same time. For example, a researcher may belong to different research groups. Therefore, finding overlapping community structure in complex networks has more practical significance.

Currently, the research on overlapping community detection has attracted more and more attention. After the development of the past few years, there have been a number of algorithms to detect overlapping communities. For example, the clique percolation method (CPM) [15], algorithms based on local community optimization and expansion (LFM [16], OSLOM [17], DEMON [18], etc.), multi label propaga-

tion algorithms (COPRA [19], BMLPA [20], SLPA [21], etc.), algorithms based on link clustering (LINK [22], LinkComm [23], LGPSO [24], LLCM [25], LBLP [26], GaoCD [27], etc.). But the computational complexity of these algorithms is generally very high while the accuracy and stability is low. The research on disjoint community detection has reached a higher level in the past decades, and some high quality algorithms in terms of both computational complexity and accuracy have been developed. By contrast, the development of overlapping community detection is not enough.

In most cases, disjoint community structures with high qualities already contain the basic and major community structure in the network, except the overlapping part [28]. On this basis, we only need to further identify the overlapping nodes in the community. Overlapping node detection can help us to understand the characteristics of nodes more comprehensively and plays a key role in community evolution. Literature [29] proposed a new algorithm based on disjoint community detection results. Firstly, the border nodes of each community are detected according to the results of disjoint communities. Then the impact of these border nodes on the corresponding community is analyzed. If the impact value is greater than 0, the border node is added into this new community and remains in the original communities. Otherwise, the border node is removed from this community. Finally, the overlapping community structure is obtained. OCDBIDC [30] is also based on the results of disjoint community detection. But it only adds boundary nodes which increase the boundary sharpness of a community into the community.

Inspired by these, this paper proposes an overlapping community detection algorithm extended from disjoint community by selecting overlapping nodes, named ONS-OCD. Firstly, ONS-OCD determines potential members of each community based on the given disjoint community structure. According to the optimization theory, if the quality of the division is already high, then the addition of a new node will not obviously change the intensity of the community. Thus, we can get two conditions to judge whether a node is a potential member of a community. One is that it should be the external fringe node of the community, namely that there are edges between the node and the internal nodes of this community. Another is that the similarity between the node and the community should be larger than the given threshold. Then, ONS-OCD detects the overlapping nodes to get the overlapping community structure. It analyses every single potential node of each community. If the influence of the potential node on the community is larger than zero, we add this node to the community and mark it as an overlapping node.

The main idea behind ONS-OCD and its contributions are presented below:

1. ONS-OCD firstly finds the potential members of each community to reduce the detection scope of overlapping nodes;
2. ONS-OCD uses the node similarity based on the heuristic DFS encoding which is more precise to measure the relationship between nodes.

We verify the performance of the proposed algorithm on synthetic and real networks. Extensive experimental studies confirm that ONS-OCD can detect overlapping community structures more effectively compared with some other state-of-the-art algorithms.

The rest of this paper is organized as follows: Section 2 introduces the basic theories related to this paper. In Section 3, we describe the main idea of the proposed algorithm. The experimental results on both synthetic and real networks in Section 4 demonstrate the effectiveness of the proposed algorithm. The conclusion is given in Section 5.

2 BACKGROUND AND RELATED WORKS

2.1 Representation of Complex Network

A complex network can be modeled as a graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes and $E = \{e_1, e_2, \ldots, e_m\}$ is the set of edges, n and m are the number of nodes and edges in the network. $N(v_u)$ represents the neighbor set of node v_u and $Com(v_u)$ represents the community set which node v_u belongs to. $C = \{C_1, C_2, \ldots, C_k\}(1 < k < n)$ is the set of community structures, where $C_i \in C$ is a nonempty subset of V and the union of all communities are the union of all nodes in the network, $\bigcup_{i=1}^{k} C_k = V$.

Disjoint community detection algorithms divide the nodes of the network into some non-overlapping subsets. That is to say each node must belong to only one community and the intersection of any two communities is empty, $C_i \bigcap C_j = \Phi$, $i, j = 1, 2, \ldots, k$ and $i \neq j$. While overlapping community detection algorithms allow nodes to belong to one or more communities.

2.2 Node Structural Similarity

Structural similarity is a commonly used method for measuring the node similarity in complex networks. There are many methods to compute the structural similarity and these methods determine node similarity based solely on the structure of the network. Since structural equivalence is too restrictive for practical use, some simplified similarity measures can be used [31]. Here we introduce the cosine similarity. If the node v_u and node v_w are connected, the structural similarity of the node v_u and node v_w is represented as $\mathrm{Scosine}(v_u, v_w)$ and calculated by Equation (1).

$$\mathrm{Scosine}(v_u, v_w) = \frac{|N(v_u) \bigcap N(v_w)|}{\sqrt{|N(v_u)||N(v_w)|}}. \tag{1}$$

The structural similarity between two nodes represents the degree of their shared neighbors.

2.3 DFS Encoding of Nodes

Depth first search (DFS) encoding [32] is a repeated random process based on the DFS for the graph. In each process, the DFS is started from a randomly selected node and each node is re-encoding by DFS traversal order. The coding of node v_u is marked as $DFS(v_u)$. Thus, for any two nodes v_u and v_w, the absolute difference between their coding indicates the distance of these two nodes, denoted as $dis(v_u, v_w) = |DFS(v_u) - DFS(v_w)|$. The similarity between nodes v_u and v_w is represented by the reciprocal of the distance between them, $s(v_u, v_w) = 1/dis(v_u, v_w)$. Repeat the random process many times, and average these similarities between node v_u and v_w as the final node similarity $S_{DFS}(v_u, v_w)$.

2.4 Node Similarity Based on Heuristic DFS Encoding

DFS encoding is a depth first search process starting from a random node and encoding each node based on the traversal order. In this paper, the heuristic rules of heuristic DFS (HDFS) encoding guides the DFS process to traverse the nodes in the same community firstly. That is to say, in the traversal process, the node which has the maximum structural similarity with the current expansion node is always firstly chosen to be traversed. For any two nodes, if their values of HDFS encoding are close, the similarity between them is large.

Since HDFS encoding has some randomness, the node similarity is calculated by using the average value of multiple HDFS encoding. For any two nodes v_u and v_w, the similarity based on HDFS encoding is denoted as $S_{HDFS}(v_u, v_w)$. In this paper, the execution number of HDFS encoding is set to be the number of communities in the network, and in each process, the node with the largest node degree of each community is selected as the initial expanding node.

3 OVERLAPPING NODE SELECTION METHOD

We propose an overlapping node selection method based on the disjoint community structure. In order to better understand the algorithm model, we first introduce a few definitions, and then detailedly introduce the process of the algorithm proposed in this paper.

3.1 Related Definitions

Definition 1 (Similarity between node and community)**.** The maximum similarity between the node and the community members is the similarity between the node and the community. The similarity between node v_u and the community C_i is denoted as $\mathrm{SNC}(v_u, C_i)$ and calculated by Equation (2).

$$\mathrm{SNC}(v_u, C_i) = \max_{v_w \in C_i} S(v_u, v_w) \tag{2}$$

where $S(v_u, v_w)$ is the similarity between node v_u and node v_w and any kind of node similarity in complex network can be used in this formula. In this paper, we choose the node similarity based on HDFS coding to measure the similarity between any two nodes. Here $S(v_u, v_w) = S_{HDFS}(v_u, v_w)$.

Definition 2 (The potential member of a community)**.** For a community, the nodes in the network can be divided into three classifications: external nodes, internal nodes and fringe nodes. External node means a node outside of the community; internal node is the node which is within the community and is not connected with the external node; fringe node is within the community and connected with the external node. In Figure 1, the area of I is the internal node of community C, B is the fringe node of the community C, U is the external node of community C. The external nodes can be further divided into true external nodes and external fringe nodes. The external fringe node is the node which is outside the community and is connected with the fringe node, represented by UB.

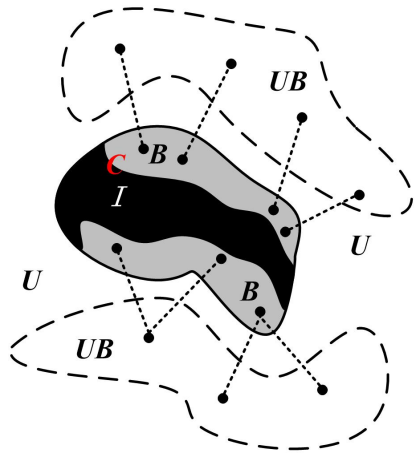


Figure 1. Node classification in complex networks

The potential member of the community needs to meet two conditions at the same time. It must be the external fringe node of the community and the similarity between the node and the community is greater than a given threshold (the threshold value can be set to the similarity between the node and the current community it belongs to).

Definition 3 (Community strength)**.** Based on the theory that the similarity between the nodes in the same community should be as large as possible, and the nodes in different communities should be as different as possible, we define the community strength as the ratio of the sum similarity between internal nodes of the community and their adjacent nodes within the community to the sum similarity between

internal nodes of the community and all their adjacent nodes in the networks. The larger the ratio is, the more obvious the community structure is and the greater the community strength is. Community strength calculation formula is shown as Equation (3).

$$R(C_i) = \frac{\sum_{v_u \in C_i} \sum_{v_w \in C_i, v_w \in N(v_u)} S_{HDFS}(v_u, v_w)}{\sum_{v_u \in C_i} \sum_{v_w \in N(v_u)} S_{HDFS}(v_u, v_w)} \tag{3}$$

where $\sum_{v_w \in C_i, v_w \in N(v_u)} S_{DHFS}(v_u, v_w)$ is the sum of similarity between node v_u and all its neighbor nodes within the community C_i and $\sum_{v_w \in N(v_u)} S_{DHFS}(v_u, v_w)$ is the sum of similarity between node v_u and all its neighbor nodes in the networks.

Definition 4 (The influence of node on community)**.** The variation of the community strength before and after the node joins the community is the influence of the node on the community. The calculation of the influence of the node v_u on the community C_i, denoted as $F(C_i, v_u)$, is shown as Equation (4).

$$F(C_i, v_u) = R(C_i \bigcup \{v_u\}) - R(C_i \backslash \{v_u\}). \tag{4}$$

Definition 5 (Overlapping node)**.** If a node belongs to more than one community at the same time, it is an overlapping node. That is to say, if $|Com(v_u)| > 1$, node v_u is an overlapping node.

3.2 Pseudo Code of the Algorithm

ONS-OCD contains two stages. The first stage is to find the potential members of each community and construct the potential node set (PNS) of each community. The second stage is to analyze the potential member nodes and get the set of overlapping nodes (ONS). In the first stage, ONS-OCD selects the external fringe node of the community. Then, it determines whether the node is a potential member of the community according to the similarity between the node and the community, and obtains the potential members of the node set PNS (line 7–9). In the second stage, ONS-OCD traverses PNS set of each community and calculates the influence of every node on the community. If the influence of node v_u on the community is positive, node v_u is added to the community and becomes an overlapping node (line 15–24).

3.3 Time Complexity Analysis

Assuming that the network G contains n nodes and m edges, the time complexity analysis of the improved algorithm proposed in this paper is as follows:

1. Compute the HDFS similarity: The time complexity of HDFS is $O(m)$, and it is repeated k times, where k is the number of communities in the network and $k << n$. So the time complexity is $O(km)$;

Algorithm 1 Overlapping community detection extended from disjoint community structure (ONS-OCD)

Input: $G = (V, E)$, disjoint community structure $DC = \{DC_1, DC_2, \ldots, DC_k\}$
Output: overlapping community structure $OC = \{OC_1, OC_2, \ldots, OC_k\}$

// The first stage, find the potential members
for each $DC_i \in DC$ **do**
  $PNS[i] \leftarrow \Phi$
  for each $v_u \in DC_i$ **do**
    for each $v_w \in N(v_u)$ **do**
      if $v_w \notin DC_i$ and $v_w \in DC_j$ **then**
        if $\mathrm{SNC}(v_w, DC_i) > \mathrm{SNC}(v_w, DC_j)$ **then**
          $PNS[i] \leftarrow PNS[i] \bigcup \{v_w\}$
        end if
      end if
    end for
  end for
end for
// The second stage, find the overlapping nodes
for each $PNS[i] \in PNS$ **do**
  $ONS_i \leftarrow \Phi$
  $OC_i \leftarrow DC_i$
  for each $v_u \in PNS[i]$ **do**
    if $F(DC_i, v_u) > 0$ **then**
      $OC_i \leftarrow OC_i \bigcup \{v_u\}$
      $ONS_i \leftarrow ONS_i \bigcup \{v_u\}$
    end if
  end for
end for
return OC

2. Judge the community potential node: $O(nd)$, where d is the average degree of nodes in the network;
3. Judge the overlapping node: $O(n'd)$, where n' is the number of potential nodes and $n' << n$.

The time complexity is $O(km) + O(nd) + O(n'd)$, taking into account that in many real networks k, n' and d are much less than n, m has the linear relationship with n, therefore, the overall time complexity of ONS-OCD is $O(m)$ or $O(n)$.

4 EXPERIMENTS

This section compares the performance of ONS-OCD with COPRA [19], LFM [16], CFinder (The implementation version of CPM algorithm) [33] and OCDBIDC [30],

where COPRA, LFM and CFinder are representative algorithms which detect overlapping communities directly and they are all widely accepted. So we compare these algorithms with the algorithm proposed in this paper. OCDBIDC and ONS-OCD belong to the same kind of algorithms which detect overlapping communities based on the disjoint community structure. We design this group contrast experiment to verify the performance of the proposed algorithm in this kind of algorithm.

All the simulations are carried out in a desktop PC with Intel® Core™ i5-2400 3.1 GHz processor and 4 GB memory under Windows 7 OS. We implement LFM, ONS-OCD and OCDBIDC in Microsoft Visual Studio 2010 environment using C++. Other algorithms are realized with Java language.

4.1 Experimental Data

1) LFR Benchmark Networks. LFR benchmark networks [34, 35] are currently the most commonly used synthetic networks in community detection, including the following parameters. N is the number of nodes; $avgk$ is the average degree of nodes in the network; $maxk$ is the maximum degree of nodes; $minc$ is the number of nodes that the minimum community contains; $maxc$ is the number of nodes that the biggest community contains; mu is a mixed parameter, which is the probability of nodes connected with nodes of external community. The greater mu is, the more difficult it is to detect the community structure; om is the number of memberships of the overlapping nodes and on represents the number of overlapping nodes. We can generate different types of networks by setting different values of these parameters.

2) Real Networks. We also make experiments on eight well known real networks, including Zachary's karate club networks (Karate), Dolphins social networks (Dolphins), American political books networks (Polbooks), American College Football networks (Football), and so on. The detailed information of each network is shown in Table 1.

Network ID	Network Name	Number of Nodes	Number of Edges	References
R1	Karate	34	78	[36]
R2	Dolphins	62	159	[36]
R3	Political Books	105	441	[36]
R4	Football	115	613	[36]
R5	Email	1 133	5 451	[37]
R6	Political Blogs	1 490	19 090	[36]
R7	Netscience	1 589	2 742	[38]
R8	PGP	10 680	24 316	[37]

Table 1. The information of real networks

4.2 Evaluation Criteria

1) Normalized Mutual Information (NMI). For LFR benchmark network, we use normalized mutual information (NMI) [18] as the evaluation criteria to compare results of different algorithms, since the groundtruth of the community structure has already been known.

Assuming the true community collection of the network is C, the membership of node i can be considered as a binary array of $|C|$ entries. If node i is present in the k^{th} community, $(x_i)_k = 1$, otherwise $(x_i)_k = 0$. We can regard the k^{th} entry of this array as the realization of a random variable X_k, whose probability distribution is $P(X_k = 1) = N_k/N$, $P(X_k = 0) = 1 - N_k/N$, where N_k is the number of nodes in the k^{th} community and N is the number of nodes in the networks. The same holds for random variable Y_l associated to the l^{th} community of the community detection result C'. We can define the conditional entropy to infer X_k given a certain Y_l, $H(X_k|Y_l) = H(X_k, Y_l) - H(Y_l)$. In particular, we can define the conditional entropy of X_k with respect to all the components of Y.

$$H(X_k|Y) = \min_{l \in \{1,2,\ldots,|C'|\}} H(X_k|Y_l). \tag{5}$$

The definition of the normalized conditional entropy of X with respect to Y is in Equation (6).

$$H(X|Y) = \frac{1}{|C|} \sum_k \frac{H(X_k|Y)}{H(X_k)}. \tag{6}$$

The expression for $H(Y|X)$ can be determined in the same way. So, the normalized mutual information (NMI) is finally defined as Equation (7).

$$NMI(X|Y) = 1 - [H(X|Y) + H(Y|X)]/2. \tag{7}$$

The large NMI value indicates that the community detection result is good, and vice versa.

2) F-Measure. For overlapping community detection algorithms, the ability of identifying overlapping nodes in the network is an important aspect to measure the performance of these algorithms. F-Measure [21] is one of the most important criteria which are widely used to measure the accuracy of algorithms in the field of machine learning. So we use F-Measure to compare the overlapping nodes detecting ability of ONS-OCD and OCDBIDC. The calculation formula of F-Measure is shown as Equation (8).

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where Precision indicates the ratio of the correct number of detected overlapping nodes to the total number of detected overlapping nodes (on^d). Recall is

calculated by dividing the correct number of detected overlapping nodes by the total number of overlapping nodes (on) in the network.

The larger F-Measure value is the better the detected overlapping nodes are [39].

3) Overlapping modularity. As the true community structure of most real networks is unknown, we use the overlapping modularity (EQ) [7] as the evaluation criteria. It is calculated by Equation (9).

$$EQ = \frac{1}{2m} \sum_{i=1}^{K} \sum_{u,v \in C_i} \frac{1}{O_v O_u} \left(A_{uv} - \frac{d_u d_v}{2m} \right) \tag{9}$$

where m represents the number of edges in the network; A is the adjacency matrix of the network; if node u and node v are directly connected, $A_{uv} = 1$, otherwise, $A_{uv} = 0$; d_u and d_v respectively denote the degree of node u and node v. O_u and O_v respectively denote the number of communities which node u and node v belong to.

The larger EQ value is the better the result of community detection is [40].

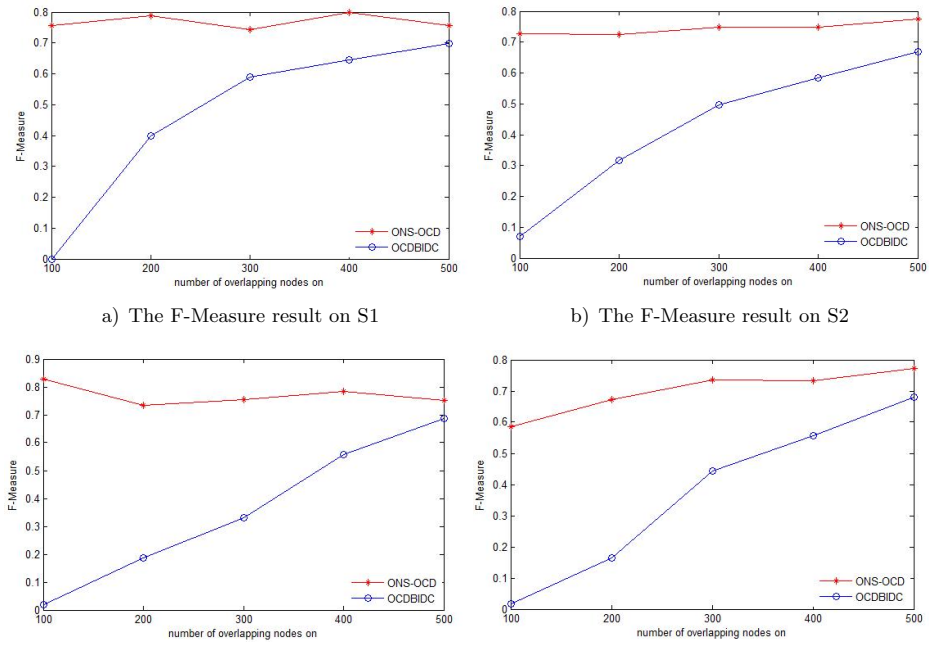
4.3 Experimental Comparison on Synthetic Networks

We use four groups of synthetic networks to evaluate the effectiveness of ONS-OCD. The details of these networks are shown in Table 2. All the networks share the common parameters of $N = 1\,000$, $avgk = 15$, $maxk = 50$ and $om = 2$. Each group contains six networks with on ranging from 0 to 500 and they also share parameters $minc$, $maxc$ and mu. The community size $minc$, $maxc$ are set to 10, 50 and 20, 100, respectively, implying small community networks and large community networks; mu is set to 0.1 and 0.3, respectively representing low and high hybrid network.

Network ID	N	$avgk$	$maxk$	$minc$	$maxc$	mu	om	on
S1	1 000	15	50	10	50	0.1	5	0-500
S2	1 000	15	50	10	50	0.3	5	0-500
S3	1 000	15	50	20	100	0.1	5	0-500
S4	1 000	15	50	20	100	0.3	5	0-500

Table 2. The information of four groups of LFR networks

1) The Comparison of Overlapping Nodes Detection. First, in the case of the ideal high quality input, we compare the overlapping nodes detection ability of ONS-OCD and OCDBIDC. We do experiments on the four groups of LFR networks ($S1 \sim S4$) and choose one of the real labels of all the nodes as the input. Figure 2 depicts the results of ONS-OCD and OCDBIDC on four groups of LFR benchmark networks. The abscissa represents the number of overlapping nodes from 100 to 500, and the ordinate is the F-Measure of the results.



a) The F-Measure result on S1

b) The F-Measure result on S2

c) The F-Measure result on S3

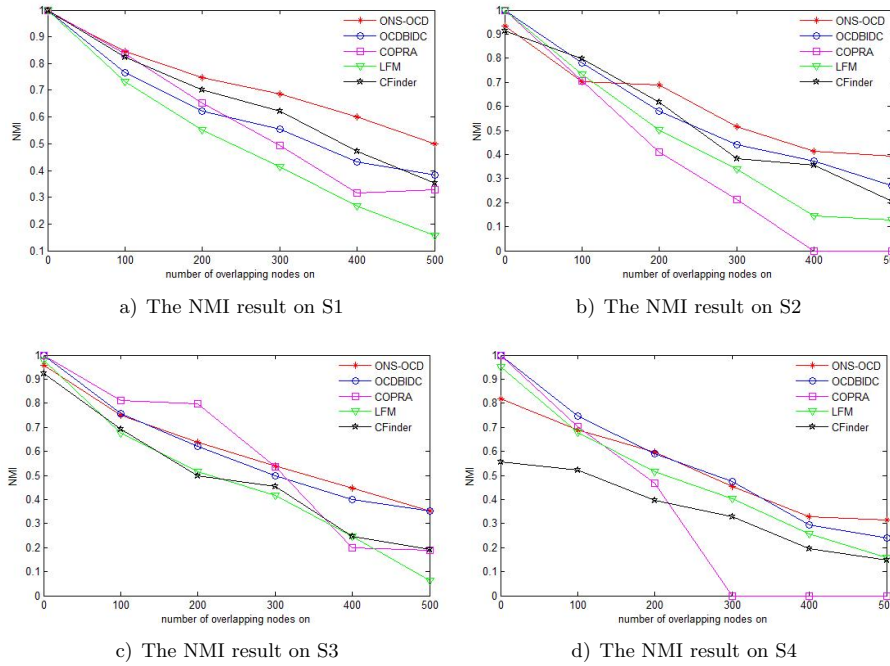
d) The F-Measure result on S4

Figure 2. The overlapping nodes detecting results of the two algorithms on LFR benchmark networks

From Figure 2, it is observed that the results of the proposed algorithm are better than OCDBIDC in all these four group networks. And in the networks with different number of overlapping nodes, the overlapping nodes selection ability of ONS-OCD is basically unchanged. The overlapping nodes in the network can be well detected by ONS-OCD. In the contrast, OCDBIDC has very poor ability to detect overlapping nodes in the network with small number of overlapping nodes. When there are 100 overlapping nodes in the network, the F-Measure value of the result detected by OCDBIDC is less than 0.1. The ability of OCDBIDC to detect the overlapping nodes improves with the increase of the number of overlapping nodes in the network, but it is still worse than ONS-OCD.

2) The Comparison of Overlapping Community Detection. Label propagation algorithm (LPA) [12] is one of the fastest community detection algorithms, with nearly linear time complexity. The algorithm is simple and does not need any parameter, thus receiving quite a lot of attention from numerous scholars. So we use the community detection result of LPA as the input information to ONS-OCD and OCDBIDC. Three classical overlapping community detection algorithms (COPRA, LFM and CFinder) are added in this comparison. The parameters of the algorithms are set as follows: in COPRA v is varied from 2

to 10 with a step size of 1; in LFM is set from 0.8 to 1.6 with a step size of 0.1; the parameter k in CFinder is initially set to 3 and increased by a step size of 1 up to 8; the parameter r of OCDBIDC is ranging from 0.1 to 1 and increased by a step size of 0.1. Each algorithm obtains different results under different parameters, and the best results of NMI are selected as the final result.



a) The NMI result on S1

b) The NMI result on S2

c) The NMI result on S3

d) The NMI result on S4

Figure 3. The overlapping community detecting results of the five algorithms on LFR benchmark networks

From these four group experimental results in Figure 3 it can be seen that in most cases, the overlapping community detection results obtained by the algorithm proposed in this paper are similar to other traditional overlapping community detection algorithms. The NMI of experimental results of all five algorithms on these four group networks decreases with the increasing number of overlapping nodes. Some traditional overlapping community detection algorithms failed to detect the overlapping community structure of the networks with too many overlapping nodes. Such as the NMI of COPRA in the second and the fourth group of networks is zero when *on* is larger than 400 and 300, respectively, while the results obtained by the proposed algorithm are optimal in most of these networks. From the experimental results on the third group of networks, it can be seen that COPRA is better than ONS-OCD and OCDBIDC when the number of overlapping nodes is less than 300. This is because the results of LPA on these networks are not satisfactory, which

shows that the initial input of disjoint community structure has great influence on these two algorithms. In summary, the experimental results on these four groups of networks show that ONS-OCD can get good results of overlapping community structure in most cases, but it is affected by the initial disjoint community input.

4.4 Experimental Comparison on Real Networks

We still use the result of LPA as the input of ONS-OCD and OCDBIDC. The parameters of the algorithms are set as follows: in COPRA v is varied from 2 to 10 with a step size of 1; in LFM is set from 0.8 to 1.6 with a step size of 0.1; the parameter k in CFinder is initially set to 3 and increased by a step size of 1 up to 8; the parameter r of OCDBIDC is ranging from 0.1 to 1 and increased by a step size of 0.1. For the five algorithms, the maximum EQ from each result under different parameters is selected as the final result. Table 3 shows the experimental results on the eight real networks, and for every instance, the best EQ and efficiency are presented in boldface.

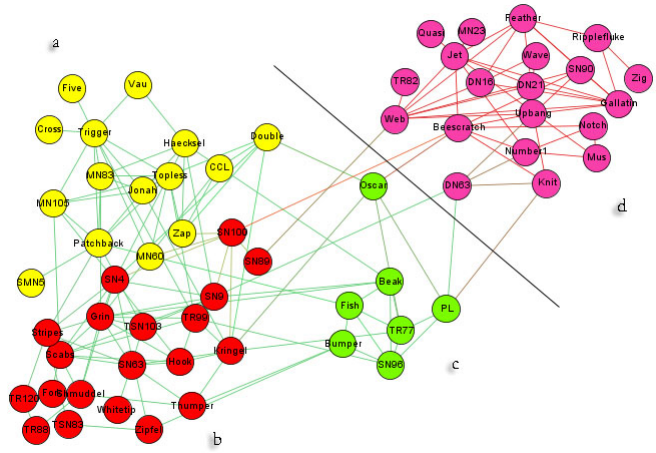
Network ID	*EQ*				
	COPRA	LFM	CFinder	ONS-OCD	OCDBIDC
R1	0.370	0.374	0.186	**0.733**	0.581
R2	0.204	0.436	0.361	0.730	**0.732**
R3	0.444	0.494	0.437	**0.826**	0.821
R4	0.583	0.566	0.548	**0.633**	0.620
R5	0.519	0.309	0.265	**0.650**	0.632
R6	0.765	0.748	0.758	**0.809**	0.804
R7	0.426	0.188	–	**0.913**	0.905
R8	0.780	0.622	0.389	0.811	**0.818**

Table 3. The comparison of results on real networks

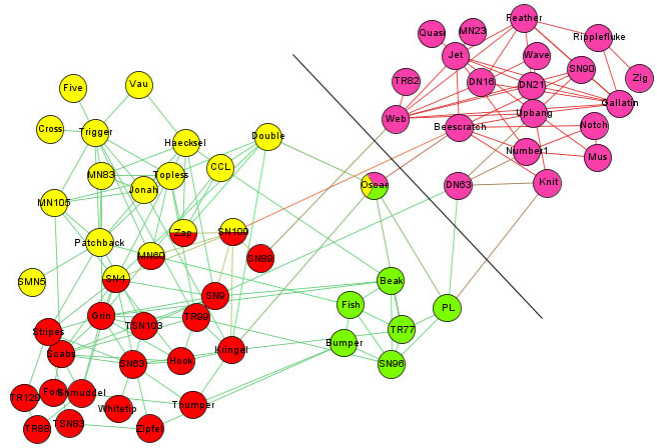
It can be seen from Table 3 that in the all real networks besides R2 (Dolphins) and R8 (PGP), the overlapping modularity of ONS-OCD is higher than those of the other four algorithms. The results of ONS-OCD on R2 (Dolphins) and R8 (PGP) are only second to OCDBIDC algorithm. Overall, the quality of the overlapping communities detected by ONS-OCD on the real networks is superior to several other algorithms.

4.5 Instance Analysis

The nodes of Dolphins are divided into two regions by a straight line in Figure 4, which represents the real division of the network. Figure 4 a) shows the community structure of Dolphins detected by LPA and Figure 4 b) is the overlapping community detection result of ONS-OCD on Dolphins. LPA algorithm divides the Dolphins data set into four communities (marked as community a, b, c and d) with different colors. It divides one of the real Dolphins communities into three. In Figure 4 b), five



a) The result of LPA on Dolphins



b) The overlapping community detection result of ONS-OCD

Figure 4. The community detection result on Dolphins

overlapping nodes (SN4, MN60, SN100, Zap and Oscar) with two or three kinds of colors are found on the basis of the result in Figure 4 a) by ONS-OCD. As can be seen in Figure 4, these nodes are closely connected between two or three communities. Node SN4 has seven adjacent nodes in the community b, which is obviously more than that in the community a which SN4 belongs to in Figure 4 a). So ONS-OCD identifies SN4 as the overlapping nodes. In summary, we can see that ONS-OCD can well identify overlapping nodes closely connected between communities.

5 CONCLUSION

In this paper, we have presented a novel overlapping node selection method to extend disjoint community structure to overlapping communities (ONS-OCD). This algorithm takes the high quality disjoint community structure as the input. Firstly, it uses the node similarity based on the heuristic DFS encoding to get the potential members of each community. Then the potential members of every community are analyzed, and the influence of the nodes on the community is calculated. Finally, the final overlapping nodes are obtained based on the node influence on communities. Since it does not need to analyze all the nodes in the network and further reduces the detection scope of overlapping nodes by the selection of potential members, it can improve the efficiency of the algorithm.

Through experiments on various synthetic networks and real networks, ONS-OCD is compared with three representative overlapping community detection algorithms (COPRA, CFinder and LFM) and OCDBIDC which also detects overlapping communities based on disjoint community structure. The results show that ONS-OCD has some advantages in the quality of community detection on the synthetic networks and real networks. In summary, ONS-OCD can identify overlapping nodes very well to get the high quality of the overlapping community structure.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities of Civil Aviation University of China (Grant No. 3122018C020 and 3122018C021), the Scientific Research Foundation of Civil Aviation University of China (Grant No. 600/600001050115 and 600/600001050117), the National Natural Science Foundation of China (Grant No. 61572505 and 61876186).

REFERENCES

[1] Yang, B.—Liu, D.-Y.—Liu, J.-M.—Jin, D.—Ma, H.-B.: Complex Network Clustering Algorithms. Ruan Jian Xue Bao/Journal of Software, Vol. 20, 2009, No. 1, pp. 54–66, doi: 10.3724/SP.J.1001.2009.00054.

[2] Watts, D. J.—Strogatz, S. H.: Collective Dynamics of 'Small-World' Networks. Nature, Vol. 393, 1998, No. 6638, pp. 440–442, doi: 10.1038/30918.

[3] Barabási, A.-L.—Albert, R.: Emergence of Scaling in Random Networks. Science, Vol. 286, 1999, No. 5439, pp. 509–512, doi: 10.1126/science.286.5439.509.

[4] Newman, M.—Barabási, A.-L.—Watts D. J.: The Structure and Dynamics of Networks. Princeton University Press, 2006.

[5] Newman, M. E. J.—Girvan, M.: Finding and Evaluating Community Structure in Networks. Physical Review E, Vol. 69, 2004, No. 2, Art. No. 26113, doi: 10.1103/PhysRevE.69.026113.

[6] Lee, J.—Gross, S. P.—Lee, J.: Modularity Optimization by Conformational Space Annealing. Physical Review E, Vol. 85, 2012, No. 5, Art. No. 056702, doi: 10.1103/PhysRevE.85.056702.

[7] Shen, H.—Cheng, X.—Cai, K.—Hu, M.-B.: Detect Overlapping and Hierarchical Community Structure in Networks. Physica A: Statistical Mechanics and Its Applications, Vol. 388, 2009, No. 8, pp. 1706–1712, doi: 10.1016/j.physa.2008.12.021.

[8] Shen, H.-W.—Cheng, X.-Q.: Spectral Methods for the Detection of Network Community Structure: A Comparative Analysis. Journal of Statistical Mechanics: Theory and Experiment, Vol. 2010, 2010, No. 10, Art. No. P10020, doi: 10.1088/1742-5468/2010/10/P10020.

[9] Huang, L.—Li, R.—Chen, H.—Gu, X.—Wen, K.—Li, Y.: Detecting Network Communities Using Regularized Spectral Clustering Algorithm. Artificial Intelligence Review, Vol. 41, 2014, No. 4, pp. 579–594, doi: 10.1007/s10462-012-9325-3.

[10] Girvan, M.—Newman, M. E. J.: Community Structure in Social and Biological Networks. Proceedings of the National Academy of Sciences of the United States of America (PNAS), Vol. 99, 2002, No. 12, pp. 7821–7826, doi: 10.1073/pnas.122653799.

[11] Blondel, V. D.—Guillaume, J.-L.—Lambiotte, R.—Lefebvre, E.: Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics: Theory and Experiment, Vol. 2008, 2008, No. 10, Art. No. P10008, doi: 10.1088/1742-5468/2008/10/P10008.

[12] Raghavan, U. N.—Albert, R.—Kumara, S.: Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks. Physical Review E, Vol. 76, 2007, No. 3, Art. No. 036106, doi: 10.1103/PhysRevE.76.036106.

[13] Šubelj, L.—Bajec, M.: Unfolding Communities in Large Complex Networks: Combining Defensive and Offensive Label Propagation for Core Extraction. Physical Review E, Vol. 83, 2011, No. 3, Art. No. 036103, doi: 10.1103/PhysRevE.83.036103.

[14] Rosvall, M.—Bergstrom, C. T.: Maps of Random Walks on Complex Networks Reveal Community Structure. Proceedings of the National Academy of Sciences of the United States of America (PNAS), Vol. 105, 2008, No. 4, pp. 1118–1123, doi: 10.1073/pnas.0706851105.

[15] Palla, G.—Derényi, I.—Farkas, I.—Vicsek, T.: Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature, Vol. 435, 2005, No. 7043, pp. 814–818, doi: 10.1038/nature03607.

[16] Lancichinetti, A.—Fortunato, S.—Kertész, J.: Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. New Journal of Physics, Vol. 11, 2009, No. 3, Art. No. 033015, doi: 10.1088/1367-2630/11/3/033015.

[17] Lancichinetti, A.—Radicchi, F.—Ramasco, J. J.—Fortunato, S.: Finding Statistically Significant Communities in Networks. PLoS ONE, Vol. 6, 2011, No. 4, Art. No. e18961, doi: 10.1371/journal.pone.0018961.

[18] Coscia, M.—Rossetti, G.—Giannotti, F.—Pedreschi, D.: DEMON: A Local-First Discovery Method for Overlapping Communities. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12), Beijing, 2012, pp. 615–623, doi: 10.1145/2339530.2339630.

[19] Gregory, S.: Finding Overlapping Communities in Networks by Label Propagation. New Journal of Physics, Vol. 12, 2010, No. 10, Art. No. 103018, doi: 10.1088/1367-2630/12/10/103018.

[20] Wu, Z.-H.—Lin, Y.-F.—Gregory, S.—Wan, H.-Y.—Tian, S.-F.: Balanced Multi-Label Propagation for Overlapping Community Detection in Social Networks. Journal of Computer Science and Technology, Vol. 27, 2012, No. 3, pp. 468–479, doi: 10.1007/s11390-012-1236-x.

[21] Xie, J.—Szymanski, B. K.—Liu, X.: SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process. Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW '11), Vancouver, 2011, pp. 344–349, doi: 10.1109/ICDMW.2011.154.

[22] Ahn, Y.-Y.—Bagrow, J. P.—Lehmann, S.: Link Communities Reveal Multiscale Complexity in Networks. Nature, Vol. 466, 2010, No. 7307, pp. 761–764, doi: 10.1038/nature09182.

[23] Kim, Y.—Jeong, H.: Map Equation for Link Communities. Physical Review E, Vol. 84, 2011, No. 2, Art. No. 026110, doi: 10.1103/PhysRevE.84.026110.

[24] Huang, F.-L.—Xiao, N.-F.: Discovering Overlapping Communities Based on Line Graph and PSO. Zidonghua Xuebao/Acta Automatica Sinica, Vol. 37, 2011, No. 9, pp. 1140–1144.

[25] Pan, L.—Jin, J.—Wang, C.-J.—Xie, J.-Y.: Detecting Link Communities Based on Local Information in Social Networks. Tien Tzu Hsueh Pao/Acta Electronica Sinica, Vol. 40, 2012, No. 11, pp. 2255–2263.

[26] Yu, L.—Wu, B.—Wang, B.: LBLP: Link-Clustering-Based Approach for Overlapping Community Detection. Tsinghua Science and Technology, Vol. 18, 2013, No. 4, pp. 387–397, doi: 10.1109/TST.2013.6574677.

[27] Shi, C.—Cai, Y.—Fu, D.—Dong, Y.—Wu, B.: A Link Clustering Based Overlapping Community Detection Algorithm. Data and Knowledge Engineering, Vol. 87, 2013, pp. 394–404, doi: 10.1016/j.datak.2013.05.004.

[28] Chakraborty, T.: Leveraging Disjoint Communities for Detecting Overlapping Community Structure. Journal of Statistical Mechanics: Theory and Experiment, Vol. 2015, 2015, Art. No. P05017, doi: 10.1088/1742-5468/2015/05/P05017.

[29] Wang, X.—Jiao, L.—Wu, J.: Adjusting from Disjoint to Overlapping Community Detection of Complex Networks. Physica A: Statistical Mechanics and Its Applications, Vol. 388, 2009, No. 24, pp. 5045–5056, doi: 10.1016/j.physa.2009.08.032.

[30] Li, Y.—Liu, G.—Lao, S.-Y.: Overlapping Community Detection in Complex Networks Based on the Boundary Information of Disjoint Community. 2013 25th Chinese Control and Decision Conference (CCDC), Guigang, 2013, pp. 125–130, doi: 10.1109/CCDC.2013.6560906.

[31] Tang, L.—Liu, H.: Community Detection and Mining in Social Media. Morgan and Claypool Publishers, 2010, doi: 10.2200/S00298ED1V01Y201009DMK003.

[32] Yu, T.—Xiao, Y. H.—He, Z. Y.—Wu, W. T.: Fast Randomized Algorithm for Community Detection in Large Networks. Journal of Computer Research and Development, Vol. 46, 2009, pp. 406–412.

[33] `http://cfinder.org/`.

[34] Lancichinetti, A.—Fourtunato, S.—Radicchi, F.: Benchmark Graphs for Testing Community Detection Algorithms. Physical Review E, Vol. 78, 2008, No. 4, Art. No. 046110, doi: 10.1103/PhysRevE.78.046110.

[35] Lancichinetti, A.—Fortunato, S.: Benchmarks for Testing Community Detection Algorithms on Directed and Weighted Graphs with Overlapping Communities. Physical Review E, Vol. 80, 2009, No. 1, Art. No. 016118, doi: 10.1103/PhysRevE.80.016118.

[36] http://www-personal.umich.edu/~mejn/netdata/.

[37] http://www.cs.bris.ac.uk/~steve/networks/copra/.

[38] Newman, M. E. J.: Finding Community Structure in Networks Using the Eigenvectors of Matrices. Physical Review E, Vol. 74, 2006, No. 3, Art. No. 036104, doi: 10.1103/PhysRevE.74.036104.

[39] Zhou, X.—Liu, Y.—Zhang, J.—Liu, T.—Zhang, D.: An Ant Colony Based Algorithm for Overlapping Community Detection in Complex Networks. Physica A: Statistical Mechanics and Its Applications, Vol. 427, 2015, pp. 289–301, doi: 10.1016/j.physa.2015.02.020.

[40] Zhou, X.—Liu, Y.—Wang, J.—Li, C.: A Density Based Link Clustering Algorithm for Overlapping Community Detection in Networks. Physica A: Statistical Mechanics and Its Applications, Vol. 486, 2017, pp. 65–78, doi: 10.1016/j.physa.2017.05.032.

Yan Xing is currently Lecturer in School of Computer Science and Technology, Civil Aviation University of China. Her research interests include data mining, complex network and community detection.

Fanrong Meng is Professor at the School of Computer Science and Technology, China University of Mining and Technology. Her research interests include database technology, data mining and knowledge discovery.

Yong Zhou is Professor at the School of Computer Science and Technology, China University of Mining and Technology. His research interests include data mining, genetic algorithm, artificial intelligence and wireless sensor network.

Guibin Sun is a master student at the School of Computer Science and Technology, China University of Mining and Technology. His research interests include data mining, complex network and community detection.

Zhixiao Wang is Professor at the School of Computer Science and Technology, China University of Mining Technology. His research interests include field theory application, and social network analysis.