Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-022-03774-4

ORIGINAL RESEARCH

NOCD: a new overlapping community detection algorithm based on improved KNN

Shi Dong[1] · **Mudar Sarem**[2,3]

Received: 23 March 2021 / Accepted: 9 February 2022

Abstract
In social networks, the community detection algorithm is very important for understanding the structures and the functions of these networks. A lot of researches have been done on the overlapping community detection algorithms as the overlapping is a significant feature of such networks. However, though many algorithms have been introduced to detect overlapping communities, the detection of the overlapping community is still a challenging task. In fact, the traditional static methods which partitioned the network structure could not efficiently obtain the latest community structure. The problems of high computational complexity and low identification accuracy need to be solved. To address these issues, in this paper, we propose a New Overlapping Community Detection algorithm based on improved KNN (called NOCD), which can timely adjust the community structure based on different network changes, and ultimately obtains the results of the community partitions with a high degree of Q module. To deal with the weighted social networks, NOCD adopts similarity instead of distance to evaluate the network. The experimental results show that the proposed NOCD algorithm compared with the COPRA, the CPM, the DeCom, the PLPA, and the AI-LPA algorithms can effectively improve the detection accuracy, the efficiency of parallel computing, and reduce the time complexity.

Keywords Social networks · Overlapping community · Community detection · Q module

1 Introduction

The social networks in the real world tend to be sparse and self-organized. Within the community interaction density, these networks are usually far greater than the inter community interaction density in the community structure. In the current social network analysis, there is an important trend and need to use detection methods for complex networks especially with heterogeneous relationship and dynamic social network (where the node or the edge varies with time change) to mine the implicit community structure. The existing community detection algorithms are mostly based on static social networks, and mainly contain algorithms based on hierarchical clustering and algorithms based on graph partitioning. Many researchers have made an improved in-depth study based on the traditional algorithms, so a lot of optimization algorithms including algorithms based on information entropy and classification are proposed. The Top Leaders Community Detection method based on K-means which introduced in Khorasgani et al. (2010) was a typical algorithm for community detection in information networks. This algorithm depended on the number of the input communities, and under the premise of accurate number of the input communities, it could achieve better community divisions. This kind of algorithm strongly depends on the number of the input communities. If there is a big difference between the input number of the communities and the actual division number of the communities, this may lead to a meaningless community division; so there is a need to use the algorithm without relying on prior information to get the preliminary divisions. The division number of the communities is considered as the input, and this will undoubtedly lead to a great time overhead. The algorithm without the number of input communities usually adopts the ideas of greedy algorithm

✉ Shi Dong
dongshi@zknu.edu.cn

1 School of Computer Science and Technology, Zhoukou Normal University, Zhoukou 466001, China

2 School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

3 General Organization of Remote Sensing, 12586 Damascus, Syria

Published online: 26 February 2022

and heuristic algorithm, so the optimal community structure is obtained on the purpose of a certain optimization objective function. Although this kind of algorithm does not need to re-enter the number of the communities, it often requires introducing additional parameters and adjusting the values of these parameters according to the actual situation of the network, which is a drawback of such algorithms. In addition, this kind of algorithms often has high time complexity, and the community division structure is not accurate enough. Thus, it is a very challenging task to reduce the dependence on a priori information and realize the continuous optimization of the running time and accuracy of the algorithm. The research on community mining and detecting communities in networks can provide a clear vision to understand the functional structures of the networks.

In this work, we propose an overlapping community detection algorithm based on improved KNN called a New Overlapping Community Detection algorithm (NOCD). The main idea of this algorithm is given as follows: first, in order to find out the high quality of the community, the network distance will be replaced by the similarity in the improved KNN proposed algorithm. Secondly, in the node community, if the node belongs to different communities, it has naturally become an overlapping node for connecting different communities.

The main contributions of this work are the following:

An overlapping community detection algorithm based on improved KNN called a New Overlapping Community Detection algorithm (NOCD) is proposed.
We have used a simulation method and real networks to verify the performance of our proposed NOCD algorithm. The experimental results show that the proposed NOCD algorithm not only has a very good time complexity but also has excellent performance compared with several traditional algorithms.

The rest of this paper is organized as follows. The related works and researches are introduced in Sect. 2. Section 3 discusses the overlapping community detection process and our proposed NOCD algorithm. In Sect. 4, we give out the performance evaluation of the NOCD algorithm compared with some selected traditional algorithms. And finally, in Sect. 5 , we present our conclusion of this paper.

2 Related work

A traditional community detection algorithm divides a network into several disconnected communities (i. e., association, cluster, group, etc.), and each node must belong to only one community. There are many typical algorithms proposed in the last two decades for detecting community structures, such as the optimization algorithm based on module (Newman et al. 2004; Lee et al. 2012; Shang et al. 2013), the spectral clustering algorithm (Shen et al. 2010; Jiang et al. 2009), the hierarchical clustering algorithm (Girvan et al. 2002; Blondel et al. 2008), the label propagation algorithm (Raghavan et al. 2007; Subelj et al. 2011), and the algorithm based on information theory (Rosvall et al. 2008). However, in many real social networks, the network is usually not isolated between communities, but it may overlap with other networks. That is to say, in the social network, some nodes belong to only one community and there are some other nodes that belong to multiple communities at the same time. In the social network, for example, any person can belong to several different communities (e.g., school, family, friends, etc.) according to different classification methods. Therefore, there is a significant need to find the overlapping communities in the social network structure. To further describe the concept of overlapping community, we have taken the network presented in Fig. 1 as an example. As we can see in Fig. 1, the network consists of three communities C1, C2, and C3 and one overlapping vertex 7 which should be regarded as member of the three communities. Therefore, the communities C1, C2 and C3 are called as the overlapping communities. At present, the research on overlapping community detection has attracted more and more attention, and some representative algorithms have been proposed. For example, the Clique Percolation Method (CPM) presented in Palla et al. (2005) was based on the concept that the internal links in a community are likely to form cliques due to their high densities. The main idea of this method is to move a clique on a graph, in some way, so it would probably be trapped inside its original community because it could not cross the bottleneck formed by the inter-community links.

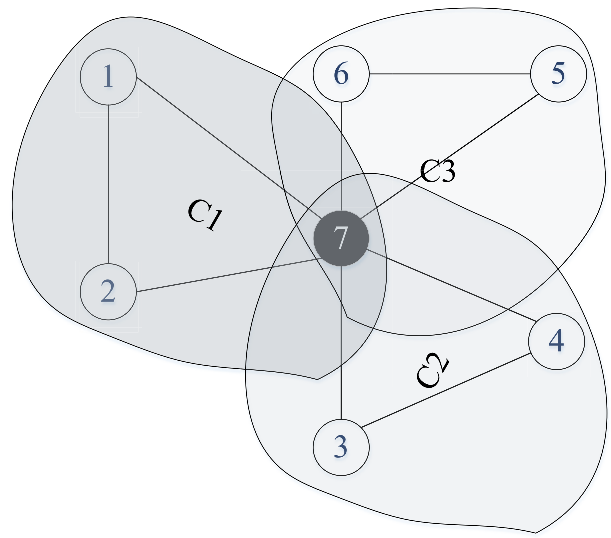

Fig. 1 A network with three communities and one overlappling vertex

Similar algorithms are presented in the literatures (Kumpula et al. 2008; Farkas et al. 2007). The optimization and the extended algorithms based on the local community respectively start from the seeds of different nodes according to a set of optimized functions to explore the local community structure together with the seeds for each local community form. In addition, some algorithms such as LFM (Lancichinetti et al (2009)), DEMON (Coscia et al (2012)), and OSLOM (Lancichinetti et al (2011)) were introduced for the overall overlapping community structure. In the algorithms based on label propagation such as COPRA (Gregory (2010)), the BMLPA (Wu et al (2012)), and SLPA (Xie and Szymanski (2012)), only the initial label was assigned to each node, and then the label and the membership degree were updated according to the neighbor nodes of each node. Finally, the same label node was given to the same community, and the nodes with multiple tags were overlapped for connecting different communities. The algorithm based on LINK clustering such as LINK (Ahn et al (2010)), LINK Maximum Likelihood (Ball et al (2011)), and LINK-Comm (Kim and Jeong (2011)), is hard to be partitioned in the network edge set, therefore the result of the edge is divided into community structures of the corresponding node. Bhatia et.al. (2019) propose DeCom algorithm to discover overlapping communites. DeCom adopts autoencoder based layered approach to initialize seed nodes and to decide the number of communities via the network structure. It can linearly deal with large graphs. However, its computational cost is huge. Xu et al. (2019) present an extended adaptive density peaks clustering for overlapping community detection, called EADP which introduces the idea of weights and a novel distance function based on common nodes in this paper. Although its detection accuracy for overlapping communities has been improved, when facing large-scale social networks, running time efficiency of the algorithm needs to be further improved. Van Lierde et al. (2019) put forward to a method based on an extension of the notion of normalized cut and introduce a hierarchical version of the algorithm to automatically detect the number of communities. Nevertheless, when facing large-scale complex network, it is difficult to detect overlapping communities. Liu et al. (2019) propose the CDCLM algorithm which adopts the triangle-based coarsening strategy to reduce the network scale. However, this method is only limited to the detect overlapping communities for static social networks and not applicable to dynamic social networks. Li et al. (2018) raise sparse symmetric non-negative matrix factorization (ssNMF) to detect the overlapping community. The technology of non-negative matrix factorization and sparse coding is used in this paper. Wang et al. (2021) propose a scalable and efficient approach for overlapping community detection(SIMGT) which can associate each node with a utility function. However, the strategic choices available of the method need further be updated. Gao et al. (2021) present the constrained personalized PageRank diffusion with a dynamic transition matrix. It can reduce the problem of redundant diffusion. Ramesh et al. (2021) propose a merged-maximal clique representation scheme which can reduce the number of maximal cliques. Sathyakala et al. (2021) put forward a weak clique based multi objective evolutionary algorithm to detect the overlapping communities. Experimental results show that the algorithm can reduce the time complexity of the algorithm and improve the performance of the algorithm. To sum up, the overlapping community detection had made some achievements. However, due to the complexity increasing of the network structure in the practical applications, the difficulty of the community detection is also increased. In fact, how to more accurately and effectively identify the network within the overlapping community structure is still a challenging task for the researcher. Therefore, finding out a new, more efficient, and robust method for finding the overlapping communities worth a further in-depth discussion and research.

3 A proposed new overlapping community detection algorithm

3.1 Problem statement

As the size of the social networks increases continuously, the community detection algorithms should be fast and accurate. Currently, though the research on community detection has shown advances such as in detecting the overlapping networks, but it still remains an open research area. In this paper, our proposed overlapping community detection algorithm can maintains higher classification accuracy and in the same time lower time complexity than the overlapping community detection algorithms proposed previously. Also, the accuracy can be improved by using similarity index for partitioning the overlapping nodes.

3.2 Basic concepts of overlapping community detection

In this sub-section, we formalize some concepts which are needed to realize the overlapping community detection process.

We have Previously expressed the real-world social community construction as a collection of nodes and edges: where either the dense or the sparsely connected nodes represent the communities. However, there are some nodes in an overlapping community belong to multiple communities as demonstrated in Fig. 2. For example, in Fig. 2, we can notice that node number 4 belongs to both communities C1 and C5 with a relationship to other objects, hence it is similar between them. Again, we can see that node number 7, 9,

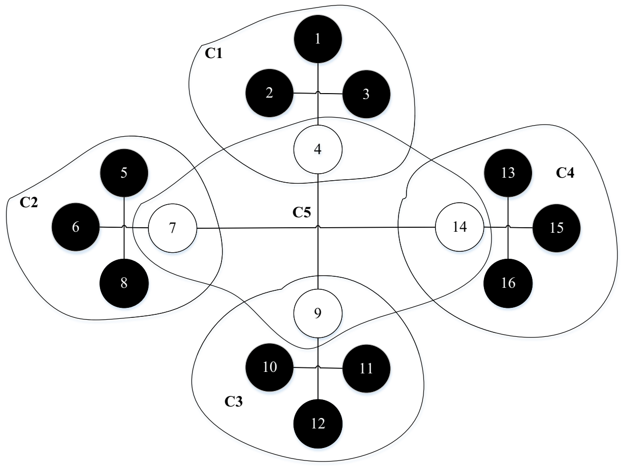

Fig. 2 Detection of five communities in graph G

and 14 belong to C2, C3, and C4 communities respectively, and these nodes also belong to community C5. For a given network *G (V, E)*, *V* represents the collection of nodes and *E* is the collection of edges. For each node *v,* the equation *I(v)={u|(u, v) E}* is true for all the neighbor nodes connected to the node *v*, where |*I(v)*| denotes the number of *I(v)* and $I_i(v)$ is the *ith* of *I(v)*. Most KNN classification algorithms use Euclidean distance which is defined as follows.

$$dist(x, y) = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^2} \tag{1}$$

where x_i and y_i represent different nodes.

However, this paper adopts similarity as the distance functions. Suppose that the similarity between the nodes *q* and *n* is defined as *S*∈*(q,n)* [0,1] and the SimRank (Jeh and Widom 2002) is defined as a way of recurrence, if *q*=*n*, then *S(q,n)*=1, otherwise, the *S(q,n)* is defined as follows according to the similarity of its neighbor nodes:

$$S(q, n) = \frac{A}{|I(q)||I(n)|} \sum_{i=1}^{|I(q)|} \sum_{j=1}^{|I(n)|} S(I_i(q), I_j(n)) \tag{2}$$

where *A* is an attenuation factor and it is a constant between 0 and 1. However, the node *q* or *n* may not have any neighbor nodes, in this case, there is no way to infer the degree of the similarity between the nodes *q* and *n*. So the set *S(q,n)*=0, that is to say when *I(q)*=*ϕ* or *I(n)*=*ϕ*, then *S(q,n)=0.* Thus the Eq. (2) is further expressed as follows:

$$S(q, n) = \begin{cases} 1; & if\ q = n \\ \frac{A}{|I(q)||I(n)|} \sum_{i=1}^{|I(q)|} \sum_{j=1}^{|I(n)|} S(I_i(q), I_j(n)); & if\ q \neq n\ and\ I(q) \neq \phi\ or\ I(n) \neq \phi \\ 0; & if\ I(q) = \phi\ or\ I(n) = \phi \end{cases} \tag{3}$$

As it is shown in Algorithm 1, by using the given formulas for the parameters, the algorithm first goes through some initializations (i.e. a_i), where a_i is the fraction of the edges which connect vertices in the community *i*. A more detailed explanation for initializing the formulas is presented in Clauset (2005).

3.3 Overlapping community detection based on improved KNN algorithm

In this paper, we use the above similarity concept to replace the distance between the nodes in our proposed overlapping community detection algorithm. In this proposed algorithm firstly, the setting of K value adopts the cross validation in probability theory. When a training dataset is given, it is useful for selecting a proper K value. In this paper, the K value is the number of nodes to be identified. *K* value is initialized as 1, then each node refers to a community and is given a community label *LK*, and then *K* is gradually iterated. With the increasing of *K* value, the number of the communities is also increasing, and the process number of the nodes which are belonging to different communities are calculated and such nodes are labeled as overlapping nodes. The overlapping of two communities is considered as an overlapping community. These steps are repeated until *K* value reaches the maximum. The proposed algorithm is divided into three algorithms. In the first algorithm which is called the initialization, the nodes and the edges of the social networks are represented by the matrix, and then the corresponding data such as $S(q, n)$ is initialized. In the second algorithm, we detect and obtain the overlapping communities by applying an improved KNN algorithm which uses a similarity measure method to acquire the node distance, and then selects the higher similarity of the neighbor nodes with the core node. Then, the nodes are labeled in the same community as the core nodes. The next step in this algorithm is to compute the nodes which belong to different communities and determine them as the overlapping nodes, so the overlapping communities are obtained. The third algorithm which is called New Overlapping Community Detection algorithm (NOCD) presents the performance evaluation of the overlapping community. In this algorithm, we calculate the extension of modularity *EQ*, the fitness function *Qov*, and the Normalized Mutual Information (NMI) respectively, and evaluate the performance of the proposed overlapping community algorithm according to these three different ways. The structure of the Algorithm is shown in Fig. 3. The following three algorithms present the detailed pseudo code of the proposed algorithm.

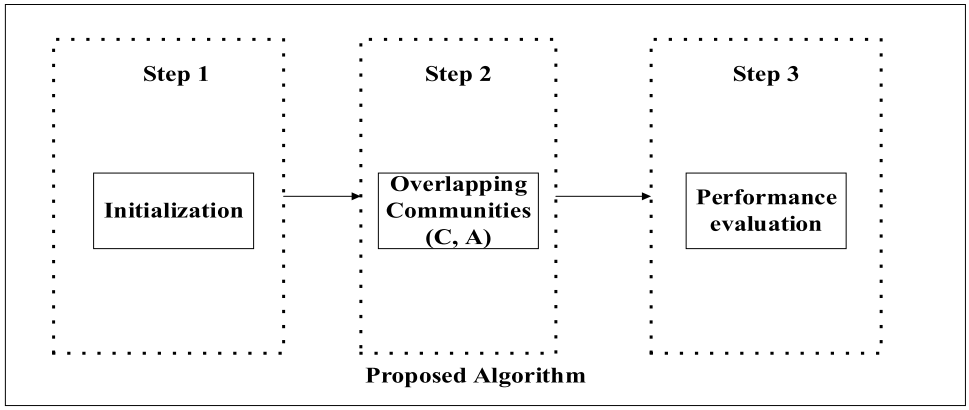

Fig. 3 Our proposed algorithm

Algorithm 1: Initialization

Input: The nodes n, q and the edges m in a network and the adjacent matrix A.

Output: The initial matrix ΔQ.

$\boldsymbol{a_i}$ **<-0**
  for i:=1 to s do
    for j:=1 to t do
      $S(q,n)=\frac{1}{q}\sum_{i=1}^{|I(q)|}\sum_{j=1}^{|I(n)|} s(I_i(q), I_j(n))$
      Updates every node until completing;
end for
end for

Algorithm 2: Gain overlapping communities (C, A)

Input: The nodes n, q and the edges m in a network and the adjacent matrix A.

Output: The set of overlapping communities C'

num is the number of communities in C;
for $\boldsymbol{i}$**:=1 to** ***num*** **do**
for $\boldsymbol{j}$**:=1 to** $\boldsymbol{n}$ **do**
for k:=1 to *num* do
  generating communities
  if s_{ij}>0.5 then
add node j to community C_k
  if node j ∈community C and ∈community C_k
node j is labeled as overlapping node
  generating community C'
end if
end for

Algorithm 3: NOCD algorithm

Initialize all nodes (Initialization)
Algorithm 2: Gain overlapping communities (C, A)-> generating community C';
num is the number of communities in C';
for $\boldsymbol{k}$**:=1 to** ***num*** **do**
  generating EQ
  $EQ=\frac{1}{2|m|}\sum_{n}\sum_{i\in C_n, j\in C_n}\frac{1}{o_i o_j}\left[A_{ij}-\frac{k_i k_j}{2|m|}\right]$
  generating Q_{OV}
  $Q_{ov}=\frac{1}{|m|}\sum_{c=1}^{n_c}\sum_{ij}\left(r_{ijc}A_{ij}-s_{ijc}\frac{k_{i,c}{}^{out}k_{j,c}{}^{in}}{|m|}\right)$
  generating NMI
  $H(X|Y)=\frac{1}{|c|}\sum_{k}\frac{H(x_k|Y)}{H(x_k)}$
end for

3.4 Time complexity analysis

Theorem 1. *The total time complexity of the proposed NOCD algorithm is O(m).*

Proof. Suppose that the network G contains n nodes and m edges. Essentially, in the first phase of the NOCD algorithm, we traverse each edge and find the similarity of the edges. Since we only need to examine every node and its neighbors, so the time complexity is $O(m)$. In the second stage of the NOCD algorithm, we apply this procedure for each edge and the connected nodes are assigned to the corresponding community. Since the time complexity in this stage is also $O(m)$, so the total time complexity of the proposed NOCD algorithm is $O(m)$.

4 Performance evaluation metric and experimental analysis

In this section, firstly we give out the performance evaluation metric. Then the experimental analysis is done via comparing with existing algorithms in a different network environment (synthetic networks and real networks).

4.1 Performance evaluation

For evaluating the efficiency of the proposed NOCD algorithm, in this paper, we have compared the NOCD algorithm with other community detection methods. The first method to be compared with our algorithm is the method proposed by Shen et al. (2009). They proposed a relatively simple method that can be used to evaluate the overlapping community discovery module for extended function *EQ*, which is defined as follows:

$$EQ = \frac{1}{2|m|}\sum_{n}\sum_{i\in C_n, j\in C_n}\frac{1}{o_i o_j}\left[A_{ij} - \frac{k_i k_j}{2|m|}\right] \tag{4}$$

where $|m|$ is the total number of the edges in the network, A_{ij} is the element of the adjacent matrix of the network. o_i and o_j are the number of communities to which vertex i and j belong respectively, and k_i and k_j are the degrees of vertex i and j respectively.

The second method to be compared with our NOCD algorithm is the method proposed by Nicosia et al. (2009). They thought that the stochastic model of the edge is not the same as the common network edge of the modularity function. So, based on the graph theory, they gave out a new module for the function model. The new modularity is defined as follows:

$$Q_{ov} = \frac{1}{|m|}\sum_{c=1}^{n_c}\sum_{ij}(r_{ijc}A_{ij} - s_{ijc}\frac{k_{i,c}^{out}k_{j,c}^{in}}{|m|}) \tag{5}$$

where r_{ijc} and s_{ijc} are the contributed portions of the modularity given by community c due to the link $l(i, j)$. $K_{i,c}{}^{out}$ is the out-degree of node i, (i. e., the number of links going out of i), and $K_{i,c}{}^{in}$ is the in-degree of node j, (i. e., the number of links coming into j).

We have adopted the extended Normalized Mutual Information (NMI) (Lancichinetti et al 2009) as the third method to identify the accuracy of three methods chosen for comparison (i. e., the CPM. the COPRA, and our proposed NOCD). The NMI value range between 0 and 1and it can measure the similarity between the detected partition and the true partition. The larger the NMI value is, the better the partition result is. Figure 2 shows the best NMI values of the three algorithms (i. e., the CPM. the COPRA, and our proposed NOCD) on six synthetic datasets. The main formulas of overlapping NMI for partition c are given by Eq. (6) and Eq. (7) as follows:

$$\mathrm{H(X|Y)} = \frac{1}{|c|}\sum_{k}\frac{H(x_k|Y)}{H(x_k)} \tag{6}$$

$$\mathrm{NMI} = 1 - \frac{1}{2(H(X|Y) + H(Y|X))} \tag{7}$$

where *X(Y)* is the random variable associated with the partition *C* and *C′*. H(X|Y) is the normalized conditional entropy of X to Y, and H(Y|X) is the normalized conditional entropy of Y concerning X.

In addition, we have adopted the precision, the recall and the *F1*-measure to further evaluate the performance. As it was presented in Bu et al. (2015), let $C_r(\partial)=\{v_i|P_{i,r}>\partial\}$, where C_r represents the *rth* overlapping community, ∂ is the membership threshold. $\partial\in[0,1]$, and $P_{i,r}$ denotes the membership degree of the node i that belongs to community r. It can control the scale of the overlapping community. The precision P(∂), the recall R(∂), and the F1-measure F1 are expressed by Eqs. (8), (9), and (10) as follows:

$$\mathrm{P}(\partial) = \frac{\sum_{r=1,2\ldots,n}\sum_{v_i\in C_r(\partial)}\frac{|C_r(\partial)\cap T_i|}{|C_r(\partial)|}}{\sum_{r=1,2\ldots,n}\sum_{v_i\in C_r(\partial)} 1} \tag{8}$$

$$\mathrm{R}(\partial) = \frac{\sum_{r=1,2\ldots,n}\sum_{v_i\in C_r(\partial)}\frac{|C_r(\partial)\cap T_i|}{|T_i|}}{\sum_{r=1,2\ldots,n}\sum_{v_i\in C_r(\partial)} 1} \tag{9}$$

$$F1 = \frac{2P(\partial)R(\partial)}{P(\partial)+R(\partial)} \tag{10}$$

where T_i is the ground truth community including the node v_i.

4.2 The experimental results

In order to study the performance of our proposed algorithm and compare with other algorithms, we select some dataset including in synthetic networks and real networks.

4.2.1 Synthetic networks

The most widely used synthetic benchmark for comparison of community detection algorithms is the LFR (Lancichinetti-Fortunato-Radicchi) model which was introduced in Lancichinetti et al. (2008). Therefore, we have selected six LFR benchmark network data sets. The synthetic network

Table 1 Synthetic networks parameters

Datasets	Node	μ	Community size range	O_n	O_m
LFR1	1000	0.1	10,50	100	2,3,4,5,6
LFR2	1000	0.3	10,50	100	2,3,4,5,6
LFR3	5000	0.1	10,50	500	2,3,4,5,6
LFR4	5000	0.3	10,50	500	2,3,4,5,6
LFR5	5000	0.1	20,100	500	2,3,4,5,6
LFR6	5000	0.3	20,100	500	2,3,4,5,6

information is shown in Table 1. In the experiments, we have tested these six widely-used real networks (i. e., LFR1 to LFR6). The *EQ* and *Qov* are used as performance metrics to evaluate our proposed NOCD algorithm compared with the CPM, COPRA, DeCom Bhatia et al. (2019), PLPA Sheng et al. (2019), and NI-LPA El Kouni et al. (2020) methods. From Table 2, it is obvious that the *EQ* and the *Qov* for our proposed NOCD algorithm have obtained the maximum values among the six algorithms.

The reason is that the community construction affects the formation of the node community for our NOCD algorithm. In order to get better results for community detection, our NOCD algorithm can ignore the isolated edges in the node community. So, the values of the *EQ* and the *Qov* for our NOCD algorithm are significantly higher than that of the other four algorithms (i. e., the CPM, COPRA, DeCom, PLPA, and NI-LPA).

Figure 4 shows that the NMI values of the community founded by our NOCD algorithm are greater than that of the other four algorithms in any benchmark network. By comparing LFR1 with LFR2, LFR3 with LFR4, and LFR5 with LFR6 respectively, we can notice that the network topology becomes much fuzzier and the NMI values of each algorithm will be better when $\mu = 0.1$ than that when $\mu = 0.3$. For instance, our NOCD, the NI-LPA, the PLPA, the DeCom, the COPRA, and the CPM algorithms can nearly uncover 83%, 82.5%, 82%, 81%, 80%, and 79% of accurate communities respectively on LFR1 with $O_m = 2$. However, the accuracies of the detection for the six algorithms decrease to 78%, 76%, 72%, 71.5%, 70% and 69% respectively on LFR2 with $Om = 2$. In general, there is a common variation tendency for the curve in each graph. As the number of O_m varying from 2 to 6, the curve is declining. The reason might be that a bigger O_m denotes a node could belong to more communities. This will make it harder to identify overlapping communities. In other words, when the partition is much fuzzier, which means that more nodes belong to multi communities, it is difficult to get the true partition. Therefore, the NMI values will be low. However, our NOCD algorithm will get higher NMI values than the other algorithms on each dataset anyway.

Table 3 shows the average precision, the recall, and the F1-measure values under different thresholds ∂. As we can see from Table 3, when $\partial = 0.8$, F1 can obtain the maximum values in the six datasets. In dataset LFR3, the value of F1 is equal to 0.529 which is the largest in all the datasets. So, when the threshold ∂ is equal to 0.8, our NOCD algorithm can have the best performance.

To verify algorithms' running time, we adopt LFR to create a group of dataset. The vital parameters are set. Where $\mu = 0.1$ and $O_m = 2$, the number of nodes varies from 1000 to 20,000. In Fig. 5, we can see that with the increase of the number of nodes, the running time of all algorithms is increasing. Meanwhile, in the six algorithms, our proposed algorithm is more efficient on running time. The reason why is that proposed algorithm only adopts similarity to replace the traditional distance function and the the total time complex degree of the proposed NOCD algorithm is *O(m).*

4.2.2 Real networks

We have adopted five real networks datasets including Karate (Zachary (1977)), Dolphin network (Lusseau et al. (2003)), NCAA college football network (Girvan et al. 2002), Jazz (Gleiser et al. 2003), email (Guimera et al. 2003). The datasets in detail are depicted in Table 4.

Karate: a weighted Zachary's interaction network between 34 members of a Karate club.
Dolphin network: a social network including 62 dolphins, which are frequent associations with each other.

Table 2 Community detection results

Datasets	CPM		COPRA		DeCom		PLPA		NI-LPA		NOCD	
	EQ	Q_{ov}	EQ	Q_{ov}	EQ	Qov	EQ	Q_{ov}	EQ	Q_{ov}	EQ	Q_{ov}
LFR1	0.437	0.498	0.478	0.523	0.487	0.514	0.493	0.531	0.512	0.523	0.579	0.538
LFR2	0.476	0.512	0.483	0.536	0.512	0.524	0.562	0.543	0.587	0.558	0.623	0.579
LFR3	0.497	0.523	0.523	0.547	0.545	0.557	0.582	0.561	0.596	0.573	0.654	0.593
LFR4	0.471	0.547	0.543	0.578	0.556	0.589	0.597	0.583	0.625	0.603	0.667	0.621
LFR5	0.492	0.579	0.558	0.585	0.572	0.593	0.604	0.598	0.634	0.612	0.676	0.634
LFR6	0.498	0.589	0.663	0.592	0.671	0.607	0.672	0.629	0.678	0.636	0.689	0.658

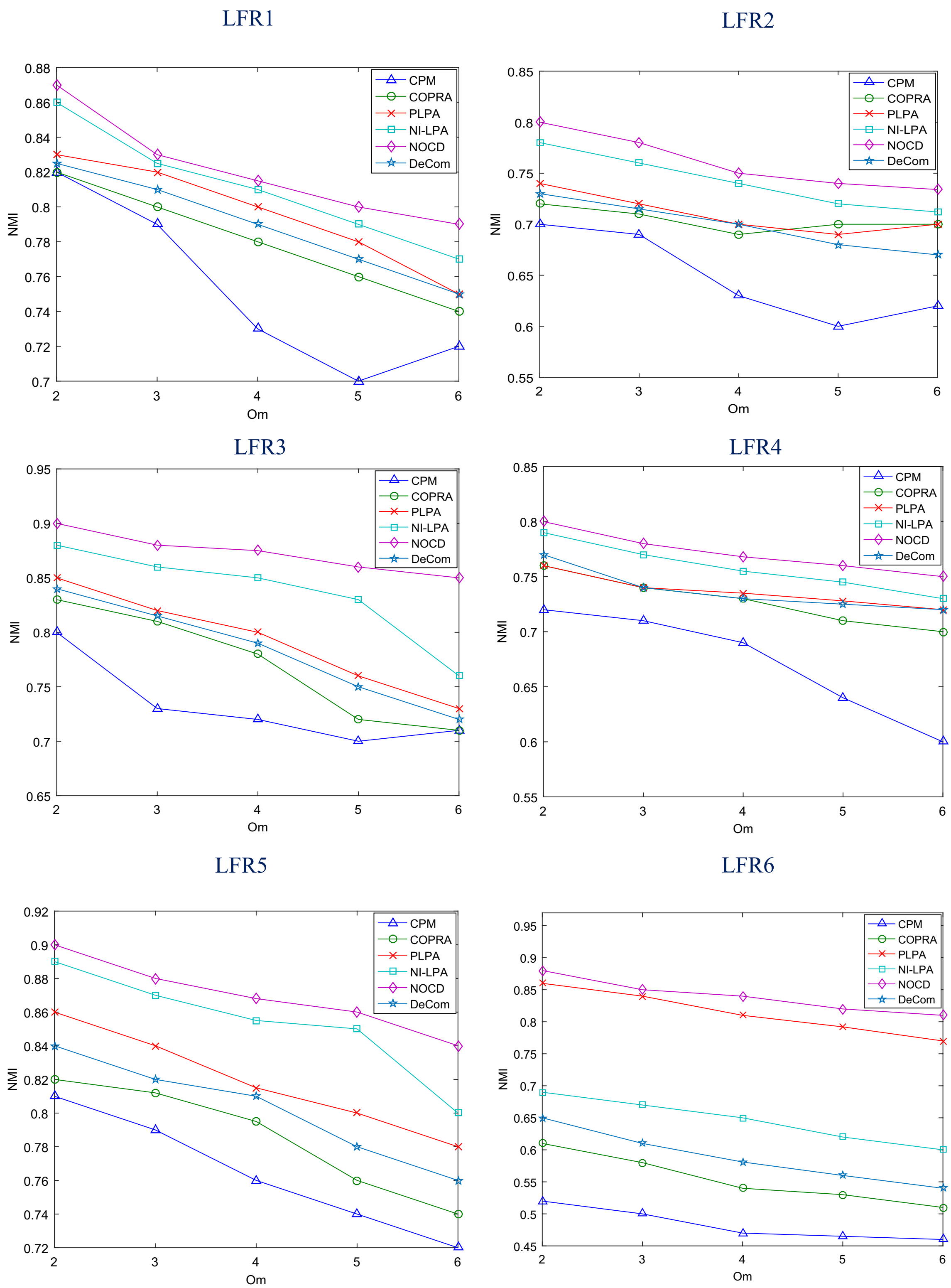

Fig. 4 Comparative NMI values of six algorithms on six datasets

Table 3 Average performance of our proposed algorithm under different thresholds ∂

Datasets	$\partial=0.2$			$\partial=0.5$			$\partial=0.6$			$\partial=0.8$		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LFR1	0.483	0.31	0.378	0.497	0.325	0.347	0.514	0.337	0.407	0.562	0.419	0.48
LFR2	0.49	0.38	0.428	0.51	0.392	0.443	0.527	0.413	0.463	0.573	0.428	0.49
LFR3	0.52	0.41	0.458	0.56	0.43	0.486	0.571	0.441	0.498	0.61	0.467	0.529
LFR4	0.487	0.36	0.414	0.495	0.378	0.429	0.508	0.385	0.413	0.548	0.402	0.464
LFR5	0.51	0.43	0.467	0.542	0.453	0.494	0.563	0.467	0.511	0.594	0.47	0.525
LFR6	0.502	0.42	0.457	0.537	0.438	0.482	0.552	0.453	0.498	0.586	0.464	0.518

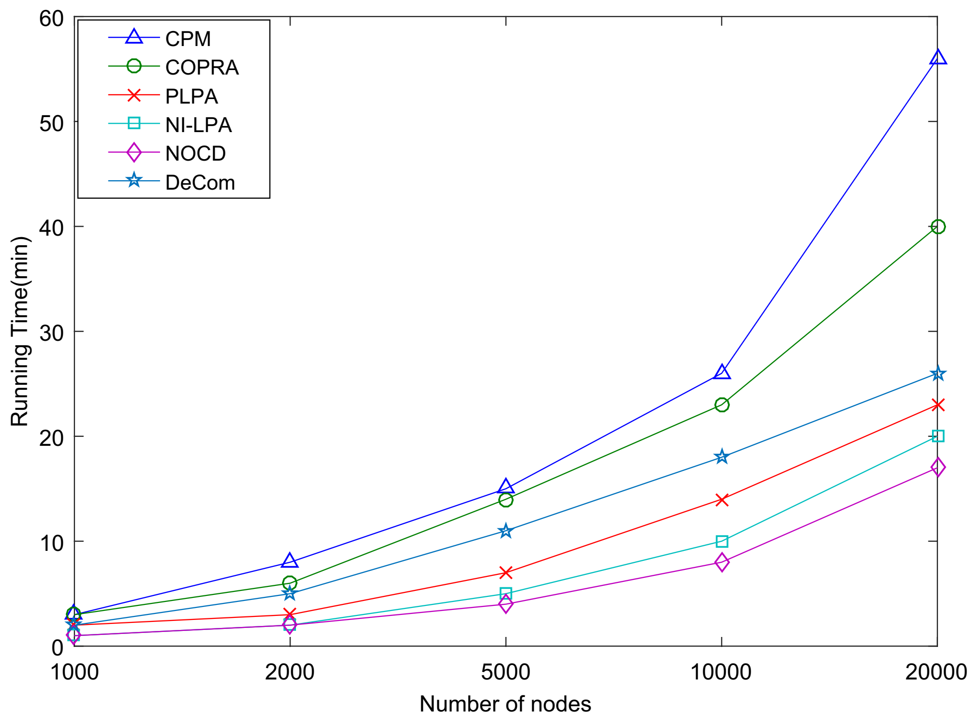

Fig. 5 Comparative running time of six algorithms on the different number of nodes

Table 4 Real networks

Dataset	Node	Edge	Community	Average degree
Karate	34	78	3	4.59
Dolphin	62	159	2	5.13
Football	115	613	12	10.66
Jazz	198	2742	4	27.7
Email	1133	5451	11	9.62

NCAA college football network: a social network consisting of 115 college football teams.
Jazz: it is list of edges of the network of Jazz musicians.
Email: This is a network data which consists of 1133 nodes and 5451 edges from a research group within the University of Rovira ivirgili used to analyze individual social relationships within the research group.

Table 5 The EQ and Q_{ov} performance of six algorithms on five real networks

Datasets	CPM		COPRA		DeCom		PLPA		NI-LPA		NOCD	
	EQ	Q_{ov}	EQ	Q_{ov}	EQ	Qov	EQ	Q_{ov}	EQ	Q_{ov}	EQ	Q_{ov}
Karate	0.132	0.167	0.224	0.221	0.238	0.229	0.263	0.242	0.278	0.257	0.289	0.274
Dolphin	0.3	0.316	0.187	0.152	0.276	0.291	0.3	0.323	0.32	0.341	0.332	0.349
Football	0.44	0.432	0.398	0.363	0.407	0.398	0.44	0.452	0.46	0.468	0.479	0.493
Jazz	0.224	0.234	0.202	0.236	0.218	0.231	0.224	0.227	0.236	0.238	0.273	0.244
Email	0.292	0.297	0.048	0.058	0.296	0.283	0.29	0.279	0.301	0.297	0.316	0.394

Table 5 shows the results (EQ and Q_{ov} values) of the six algorithms on the five real networks datasets. It is obvious that our method (NOCD) has well overlapping community detection accuracy compared with other algorithms in the real networks. The traditional CPM and COPRA have low EQ and Q_{ov} values. Our proposed NOCD is overall slightly better than the NI-LPA Algorithm. However, on some data sets the advantage is obvious such as Jazz, EQ value of NOCD is 0.273 which is significantly larger than the value of the NI-LPA Algorithm.

5 Conclusion

In this paper, in order to detect the overlapping communities in the large-scale networks starting from a high quality partition, we have proposed a new overlapping community detection algorithm (called NOCD). Our proposed NOCD algorithm can identify the overlapping nodes from the boundary and the inner node set in turn based on the deduced conditions for overlapping nodes. Also, our NOCD algorithm can always give better results on the aspect of quality than the other traditional algorithms used for comparison in this paper. Further more, the proposed NOCD method performs very well on the aspect of speed, especially for huge real-world networks. The main advantages of our proposed NOCD algorithm are: firstly, it makes extending the weighted networks easy by replacing the distance with the similarity. Secondly, since discovering the overlapping nodes among the communities of different pairs is completely independent in our algorithm, this means that the proposed NOCD algorithm is highly amenable to parallel implementation. The experimental results show that the overlapping community detection algorithm based on improved KNN (i.e., NOCD) compared with the COPRA, the DeCom, the CPM, the PLPA, and the AI-LPA algorithms can effectively improve the detection accuracy and reduce the time complexity.

Acknowledgements The authors would like to thank the anonymous reviewers for their comments which helped them in improving the quality of the paper. This paper is supported by the Key Scientific and Technological Research Projects in Henan Province (Grand No. 192102210125) and Open Foundation of State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) (SKLNST-2020-2-01) .

Funding The funding has been received from Key Scientific and Technological Research Projects in Henan Province with Grant no. 192102210125; Open Foundation of State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) with Grant no. SKLNST-2020-2-01.

Data availability statement The LFR (Lancichinetti-Fortunato-Radicchi) model introduced in [36] is the most widely used synthetic benchmark for the comparison of community detection algorithms.

Declarations

Conflict of interest The authors of the paper certify that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466(7307):761–764
Ball B, Karrer B, Newman ME (2011) Efficient and principled method for detecting communities in networks. Phys Rev E 84(3):036103
Bhatia V, Rani R (2019) A distributed overlapping community detection model for large graphs using autoencoder. Fut Gen Comput Syst 94:16–26
Blondel VD, Guillaume JL, Lambiotte R et al (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
Bu Z, Wu Z, Cao J et al (2015) Local community mining on distributed and dynamic networks from a multiagent perspective. IEEE Trans Cybern 46(4):986–999
Clauset A (2005) Finding local community structure in networks. Phys Rev E 72(2):026132
Coscia M, Rossetti G, Giannotti F, et al (2012) Demon: a local-first discovery method for overlapping communities. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 615–623
El Kouni IB, Karoui W, Romdhane LB (2020) Node importance based label propagation algorithm for overlapping community detection in networks. Expert Syst Appl 162(113):020
Farkas I, Abel D, Palla G et al (2007) Weighted ´ network modules. New J Phys 9(6):180
Gao Y, Yu X, Zhang H (2021) Overlapping community detection by constrained personalized pagerank. Expert Syst Appl 173(114):682
Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826
Gleiser PM, Danon L (2003) Community structure in jazz. Adv Complex Syst 6(04):565–573
Gregory S (2010) Finding overlapping communities in networks by label propagation. New J Phy 12(10):103018
Guimera R, Danon L, Diaz-Guilera A et al (2003) Self-similar community structure in a network of human interactions. Phys Rev E 68(6):065103
Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 538–543
Jiang JQ, Dress AW, Yang G (2009) A spectral clustering-based framework for detecting community structures in complex networks. Appl Math Lett 22(9):1479–1482
Khorasgani RR, Chen J, Zaiane OR (2010) Top leaders community detection approach in information networks. In: 4th SNA-KDD workshop on social network mining and analysis, Citeseer
Kim Y, Jeong H (2011) Map equation for link communities. Phys Rev E 84(2):026110
Kumpula JM, Kivela M, Kaski K et al (2008) Sequential algorithm for fast clique percolation. Phys Rev E 78(2):026109
Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78(4):046110
Lancichinetti A, Fortunato S, Kertesz J (2009) Detecting the overlapping and hierarchical community structure in complex networks. New J Phys 11(3):033015

Lancichinetti A, Radicchi F, Ramasco JJ et al (2011) Finding statistically significant communities in networks. PloS One 6(4):e18961
Lee J, Gross SP, Lee J (2012) Modularity optimization by conformational space annealing. Phys Rev E 85(5):056702
Li X, Hu Z, Wang H (2018) Combining nonnegative matrix factorization and sparse coding for functional brain overlapping community detection. Cogn Comput 10(6):991–1005
Liu Z, Xiang B, Guo W et al (2019) Overlapping community detection algorithm based on coarsening and local overlapping modularity. IEEE Access 7:57943–57955
Lusseau D, Schneider K, Boisseau OJ et al (2003) The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behav Ecol Sociobiol 54(4):396–405
Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
Nicosia V, Mangioni G, Carchiolo V et al (2009) Extending the definition of modularity to directed graphs with overlapping communities. J Stat Mech Theory Exp 2009(03):P03024
Palla G, Derenyi I, Farkas I et al (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818
Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106
Ramesh A, Srivatsun G (2021) Evolutionary algorithm for overlapping community detection using a merged maximal cliques representation scheme. Appl Soft Comput 112(107):746
Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105(4):1118–1123
Sathyakala M, Sangeetha M (2021) A weak clique based multi objective genetic algorithm for overlapping community detection in complex networks. J Ambient Intell Humaniz Comput 12(6):6761–6771
Shang R, Bai J, Jiao L et al (2013) Community detection based on modularity and an improved genetic algorithm. Phys A 392(5):1215–1231
Shen HW, Cheng XQ (2010) Spectral methods for the detection of network community structure: a comparative analysis. J Stat Mech Theory Exp 10:P10020
Shen H, Cheng X, Cai K et al (2009) Detect overlapping and hierarchical community structure in networks. Phys A 388(8):1706–1712
Sheng J, Wang K, Sun Z et al (2019) Overlapping community detection via preferential learning model. Phys A 527(121):265
Subelj L, Bajec M (2011) Unfolding communities in large complex networks: combining defensive and offensive label propagation for core extraction. Phys Rev E 83(3):036103
Van Lierde H, Chow TW, Chen G (2019) Scalable spectral clustering for overlapping community detection in large-scale networks. IEEE Trans Knowl Data Eng 32(4):754–767
Wang Y, Bu Z, Yang H et al (2021) An effective and scalable overlapping community detection approach: integrating social identity model and game theory. Appl Math Comput 390(125):601
Wu ZH, Lin YF, Gregory S et al (2012) Balanced multi-label propagation for overlapping community detection in social networks. J Comput Sci Technol 27(3):468–479
Xie J, Szymanski BK (2012) Towards linear time overlapping community detection in social networks. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, pp 25–36
Xu M, Li Y, Li R et al (2019) Eadp: an extended adaptive density peaks clustering for overlapping community detection in social networks. Neurocomputing 337:287–302
Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33(4):452–473