Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 4988601, 11 pages
https://doi.org/10.1155/2023/4988601

Research Article

Overlapping Community Detection Algorithm Based on High-Quality Subgraph Extension in Local Core Regions of Network

Yang Zhao,[1] **Kun Deng**,[2,3] **Xingyan Liu**,[3] **and Jiqiang Yao**[1]

[1]*College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua 321004, China*
[2]*Key Laboratory of Medical Electronics and Digital Health of Zhejiang Province, Jiaxing University, Jiaxing 314001, China*
[3]*College of Information Science and Engineering, Jiaxing University, Jiaxing 314001, China*

Correspondence should be addressed to Kun Deng; dengkun@hrbeu.edu.cn

Received 17 October 2022; Revised 5 March 2023; Accepted 11 March 2023; Published 15 April 2023

Academic Editor: Petros Nicopolitidis

Community structure is an important feature of complex networks. Detecting overlapping communities in complex networks is a hot research topic in data mining and graph theory, aiming at the shortcomings of community detection algorithm based on seed expansion, such as the instability of community detection results caused by randomly selecting seeds, the similarity of selected seeds leading to similar communities after different seed expansion, and the increase of calculation caused by deleting nodes in the process of seed expansion. This paper proposes an overlapping community detection algorithm based on high-quality subgraph extension in local core regions of the network (OLCRE). First, a novel seed community selection method is designed. By analyzing the sum of node degrees of the subgraph formed by a node and its neighbor nodes in the local core region of the network and the tightness of the internal and external connections of the subgraph, a seed community selection function is proposed. According to this function, high-quality subgraphs are selected from all the local core regions of the network as seed communities. Then, taking the definition of community as the guideline, a new community expansion strategy is proposed. Considering the influence of the neighbor node on the inner and outer connection tightness of the seed community comprehensively, it is determined whether the neighbor node can join the seed community. Finally, after the completion of all seed community expansion, overlapping nodes and possible missing nodes should be simplified and redetected to further improve the quality of community detection. The proposed algorithm is tested on the artificial and real-world networks and compared with several overlapping community detection algorithms. The experimental results verify the effectiveness and feasibility of the proposed algorithm.

1. Introduction

Many complex systems in the real world can be shown in the form of complex networks through their connection modes. Components in the system can be regarded as nodes in the network, and the connection relations between different components can be regarded as edges in the network, for example, the social network [1] that is interconnected among people and the metabolic network [2] that is connected through chemical reactions. With the further study of complex networks, it is found that community structure is the basic statistical characteristic among them. A community in a complex network can be understood as a collection of nodes with similar characteristics, which is usually represented by close connections between nodes within the same community, while sparse connections between nodes in different communities. The purpose of community detection studied in this paper is to reveal the real community structure in complex networks, which has important theoretical significance and practical value for the topological structure analysis and functional analysis of complex networks [3]. At present, the research achievements of community detection have been widely applied in the public opinion analysis and control [4], search engines [5], personalized interest recommendation [6], and other fields. In addition, in view of the actual needs of epidemic transmission prevention and control [7], the community structure, which is between the macro- and micronetwork characteristics, is taken as the entry point, and community detection of social networks is combined with the epidemic transmission, so as to provide important information about the transmission risk class of persons involved in the epidemic for epidemic transmission prevention and control.

Up to now, many classical complex network community detection algorithms have been proposed, which can be divided into two categories according to their community detection results: nonoverlapping community detection algorithms and overlapping community detection algorithms. The nonoverlapping community detection algorithm divides the complex network into multiple disjoint communities. However, in real-world networks, there exist overlapping communities; that is, a node can belong to multiple communities at the same time. In a social network, for example, a person may belong to multiple social circles (family circle, friend circle, and colleague circle). Therefore, the detection of overlapping communities in complex networks has more practical value. According to the different research perspectives, the overlapping community detection algorithms are mainly divided into algorithms based on label propagation, algorithms based on cliques, algorithms based on local extension, algorithms based on edge, algorithms based on nonnegative matrix factorization, and algorithms based on spectral clustering.

Algorithms based on label propagation, for example, the SLPA algorithm [8], firstly initialize labels for nodes in the network and then carry out label propagation. The storage space of each node will save all labels received in the process of label propagation. In order to prevent too many overlapping nodes, the label control threshold is set to determine which labels will be saved in the storage space of nodes. After label propagation stops, nodes with the same label are divided into the same community, and nodes with multiple labels are considered overlapping nodes. OMKLP algorithm [9] proposed a new core node evaluation model by analyzing the node degree and local coverage density of the subgraph formed by this node and its neighbor nodes and assigned the same label to the core node and its neighbor nodes to achieve fast convergence of the algorithm. In the process of label propagation, each node adopts an asynchronous update to receive the community label corresponding to the maximum belonging coefficient of its neighbor nodes. After label propagation stops, nodes with the same label are divided into the same community, while nodes with multiple community labels are overlapping nodes.

Algorithms based on cliques, for example, the CPM algorithm [10], start from the complete subgraph and detect the community through the penetration of the complete subgraph. The nodes belonging to multiple disconnected cliques are overlapping nodes. LOC algorithm [11] firstly finds out all the cliques in the network and selects the local maximum density node as the initial community. Then, the clique participated by the node whose fitness function value is positive among the neighbor nodes of the initial community is added to the community. If the node does not participate in the formation of the clique, only the node is added to the community. Because a node can belong to multiple cliques and be added into different communities, overlapping community structures can be detected.

Algorithms based on local extension, for example, the LFM algorithm [12], start from different seed nodes and expand the community by constantly optimizing the fitness function value of the community. The nodes that are extended into multiple communities are overlapping nodes. The ECES algorithm [13] weights the network graph according to the similarity between nodes and then selects the node with the highest centrality value as the core node and expands it. This process is repeated in the remaining set of nodes until there are no nodes left.

Algorithms based on edge, for example, the LC algorithm [14], use the Jaccard function to calculate edge similarity, construct a hierarchical tree of edge community combined with the clustering method, and then truncate the hierarchical tree to obtain edge community by using partition density function. Since a node can connect multiple edges, overlapping nodes appear naturally when the community to which the edge belongs is determined. Finally, the edge community is transformed into a node community to obtain the structure of the overlapping community. LCDEL algorithm [15] firstly transforms the node graph into the line graph, constructs the adjacency matrix of the line graph, calculates the node distance matrix of the line graph using the NDML metric, and obtains the feature matrix of the node distance matrix by principal component analysis. Finally, clustering on the feature matrix by k -means clustering algorithm combined with ensemble learning is performed to obtain the overlapping community structure.

Algorithms based on nonnegative matrix factorization, for example, the DNMF algorithm [16], directly find the discrete community membership matrix, which can assign explicit community memberships to nodes without postprocessing. In addition, the pseudosupervision module is added to DNMF to utilize the identification information in an unsupervised way, which further enhances its robustness. The AGNMF-AN algorithm [17] uses an augment attributed graph to combine both the topological structure and attributed nodes of the network and introduces an effective framework to update the affinity matrix, in which the weight of the affinity matrix in each iteration is modified adaptively instead of using a fixed affinity matrix. In addition, the $l_{2,1}$ -norm is also used to reduce the impact of random noise and outliers on the community quality, which greatly improves the effectiveness of this algorithm.

Algorithms based on spectral clustering, for example, the SPOC algorithm [18], can extract prior information such as the likelihood of each node belonging to multiple communities from available metadata and node centrality measure, and a hierarchical algorithm is introduced to automatically detect communities. The ASC algorithm [19] constructs a new affinity matrix based on both the network structure and attribute information and does not need to define control parameters to combine structure and attribute. In addition, extra nodes and edges are not added to the original network which makes the algorithm suitable for application to large-scale networks.

In recent years, local community detection algorithms based on seed extension can detect communities without the complete structural information of complex networks and have high efficiency [20–22] and validity [23–25], so it

Table 1: The notations used in this paper.

Notations	Meaning
G	An undirected and unweighted graph G
V	$V=(v_1, v_2, \cdots, v_n)$ is a nonempty finite set of nodes
E	$E=(e_1, e_2, \cdots, e_m)$ is a nonempty finite set of edges
k_v	The degree of node v
$\lvert e_{v1,v2} \rvert$	$\lvert e_{v1,v2} \rvert = 1$ if there is an edge connection between nodes v_1 and v_2. Otherwise, $\lvert e_{v1,v2} \rvert = 0$
SG	A subgraph formed by a node and its neighbor nodes
$\mathrm{adj}(i)$	The set of neighbor nodes of node i
$\mathrm{CNE}(e_{i,j})$	The common neighbor edge of $e_{i,j}$
$\mathrm{CT}(e_{i,j})$	The cluster triangle in which edge $e_{i,j}$ participates
n	The number of nodes in community C
bv	The boundary nodes of community C
$\lvert C_{\mathrm{bv}} \rvert$	The number of boundary nodes of community C

is widely used in the field of community detection. However, in terms of overlapping community detection, there are still shortcomings in the quality and stability of community detection, which are manifested as the instability of community detection results caused by randomly selecting seeds, the similarity of selected seeds leading to similar communities after different seed expansion, and the increase of calculation caused by deleting nodes in the process of seed expansion. In view of the above shortcomings, this paper proposes an overlapping community detection algorithm based on high-quality subgraph extension in local core regions of the network (OLCRE). The major contributions of this paper are as follows:

(1) A new method of seed community selection is proposed; that is, the subgraphs with tight internal connections and sparse external connections in the local core regions of the network are selected as seed communities, which conforms to the definition of community and ensures the high quality of selected seed communities. Moreover, the selected seed communities by this method are determined, which can avoid the wobble of the community detection results

(2) A new seed community expansion strategy is proposed, which takes the definition of community as the guideline. Considering the influence of the neighbor node on the tightness of the internal and external connections of the seed community comprehensively, it is to decide whether the neighbor node can join the seed community, so that the seed community would expand towards the direction of tight internal connections and sparse external connections and finally obtain high-quality community structure

(3) The OLCRE algorithm proposed in this paper does not need to set any parameters. It can be applied to networks of different scales and types and has universal applicability. The experimental results show that the OLCRE algorithm is effective and feasible, which is tested on artificial networks and real-world networks and compared with several overlapping community detection algorithms

2. Basic Concepts and Definitions

A complex network can be modeled as an undirected and unweighted graph $G=(V, E)$, where $V=(v_1, v_2, \cdots, v_n)$ is a nonempty finite set of nodes and $E=(e_1, e_2, \cdots, e_m)$ is a nonempty finite set of edges. Table 1 lists the notations used in this paper and gives a brief explanation. The basic concepts and definitions used in this paper are described below.

Definition 1 (Seed community selection function). The seed community selection function, denoted by $\mathrm{SCS}(i)$, is defined as follows:

$$\mathrm{SCS}(i)=\sum_{v\in \mathrm{SG}} k_v \cdot \frac{\sum_{v_1,v_2\in \mathrm{SG}} \lvert e_{v_1,v_2} \rvert}{\sum_{v\in \mathrm{SG},u\notin \mathrm{SG}} \lvert e_{v,u} \rvert}, \tag{1}$$

where SG represents the subgraph formed by node i and its neighbor nodes and k_v represents the degree of node v. $\lvert e_{v1,v2} \rvert = 1$ if there is an edge connection between nodes v_1 and v_2. Otherwise, $\lvert e_{v1,v2} \rvert = 0$. Likewise, $\lvert e_{v,u} \rvert = 1$ if there is an edge connection between nodes v and u. Otherwise, $\lvert e_{v,u} \rvert = 0$.

The larger the value of $\mathrm{SCS}(i)$ corresponding to the subgraph SG formed by node i and its neighbor nodes, the more located the subgraph is in the local core region, and the more tightly connected the subgraph is internally and sparsely connected to the external region.

Definition 2 (Common neighbor edge). The common neighbor edge of edge $e_{i,j}$, denoted by $\mathrm{CNE}(e_{i,j})$, is defined as follows:

$$\mathrm{CNE}(e_{i,j})=\{e_{i,u}, e_{j,u} \in E \mid u \in \mathrm{adj}(i) \cap \mathrm{adj}(j)\}, \tag{2}$$

where $\mathrm{adj}(i)$ is the set of neighbor nodes of node i and $\mathrm{adj}(j)$ is the set of neighbor nodes of node j.

Definition 3 (Cluster triangle). The cluster triangle in which edge $e_{i,j}$ participates, denoted by $\mathrm{CT}(e_{i,j})$, is defined as follows:

$$\mathrm{CT}(e_{i,j}) = \{e_{i,j}\} \cup \mathrm{CNE}(e_{i,j}), \tag{3}$$

where $\mathrm{CT}(e_{i,j})$ represents the set of cluster triangles in which edge $e_{i,j}$ participates.

The more cluster triangles an edge participates in, the tighter the edge is connected to its neighbor edges. The more cluster triangles exist in the community, the tighter the connection within the community.

Definition 4 (Node to the community interior influence function). The node to the community interior influence function, denoted by I, is defined as follows:

$$I = \frac{\sum_{e_{i,j}\in E_{c'}} \left|\mathrm{CT}(e_{i,j})\right|}{3(n_c+1)} - \frac{\sum_{e_{i,j}\in E_c} \left|\mathrm{CT}(e_{i,j})\right|}{3n_c}, \tag{4}$$

where E_c is the edge set of community C and, likewise, $E_{c'}$ is the edge set of community C' formed when a neighbor node joins community C. $|\mathrm{CT}(e_{i,j})|$ represents the number of cluster triangles in which edge $e_{i,j}$ participates, and n_c represents the number of nodes in community C.

If the corresponding I value is greater than 0 after a node joins community C, it indicates that the node joining community C can improve its internal connection tightness.

Definition 5 (Community boundary nodes). The boundary nodes of community C, denoted by bv, are defined as follows:

$$\mathrm{bv} = \left\{\mathrm{bv} \in V_c, u \notin V_c \middle| \exists e_{\mathrm{bv},u} \in E\right\}, \tag{5}$$

where V_c is the node set of the community C.

Definition 6 (Node to the community exterior influence function). The node to the community exterior influence function, denoted by E, is defined as follows:

$$E = \frac{\sum_{\mathrm{bv}\in V_{c'}, u\notin V_{c'}} \left|e_{\mathrm{bv},u}\right|}{\left|C'_{\mathrm{bv}}\right|} - \frac{\sum_{\mathrm{bv}\in V_c, u\notin V_c} \left|e_{\mathrm{bv},u}\right|}{\left|C_{\mathrm{bv}}\right|}, \tag{6}$$

where V_c is the node set of the community C and, likewise, $V_{c'}$ is the node set of the community C' formed when a neighbor node joins community C. $|C'_{\mathrm{bv}}|$ represents the number of boundary nodes of community C', and $|C_{\mathrm{bv}}|$ represents the number of boundary nodes of community C. $|e_{\mathrm{bv},u}| = 1$ if there is an edge connection between boundary node bv and node u. Otherwise, $|e_{\mathrm{bv},u}| = 0$.

If the corresponding E value is less than 0 after a node joins community C, it indicates that the node joins community C to make its connections with the outside more sparse.

Definition 7 (Community quality optimization function). The community quality optimization function, denoted by M, is defined as follows:

$$M = \frac{\sum_{v_1\in V_{c'}, v_2\in V_{c'}} \left|e_{v_1,v_2}\right|}{\sum_{\mathrm{bv}\in V_{c'}, u\notin V_{c'}} \left|e_{\mathrm{bv},u}\right|} - \frac{\sum_{v_1\in V_c, v_2\in V_c} \left|e_{v_1,v_2}\right|}{\sum_{\mathrm{bv}\in V_c, u\notin V_c} \left|e_{\mathrm{bv},u}\right|}. \tag{7}$$

The community quality optimization function is used to simplify overlapping nodes and redetect possible missing nodes so as to further improve the quality of community detection results.

3. The OLCRE Algorithm

3.1. General Description of the OLCRE Algorithm. As shown in Algorithm 1, the OLCRE algorithm firstly traverses the global network and, according to the seed community selection function SCS, selects the subgraphs with close internal connections and sparse external connections from the local core regions of the network as seed communities. In the seed community expansion stage, the influence of the neighbor node on the inner and outer connection tightness of the seed community is comprehensively considered to determine whether the neighbor node could join the seed community. When the corresponding I value and E value of a neighbor node of the seed community meet the requirements of $I > 0$ and $E < 0$, the neighbor node can join the seed community. Otherwise, it cannot join the seed community. When all neighbor nodes of a seed community do not meet the expansion strategy, the seed community stops expanding and continues to expand the rest of the seed communities until all the seed communities complete expansion. After the expansion of all seed communities is completed, overlapping nodes and possible missing nodes are simplified and redetected according to the proposed community quality optimization function, so as to further improve the quality of community detection. Finally, the output is the overlapping community structure C. Through the above steps, the overlapping community detection of complex networks is completed.

3.2. Seed Community Selection. Seed selection is a key step of overlapping community detection algorithm based on seed expansion, which has an important impact on the results of community detection. In this paper, a novel seed community selection method is proposed. According to the seed community selection function SCS, subgraphs with close internal connections and sparse external connections are selected from local core areas of the network as seed communities, see Algorithm 2 for the specific process.

The seed community selection algorithm first starts from any node i in the network and calculates the respective SCS values of node i and its neighbor nodes, respectively. If the *SCS* value of node i is not the largest, the search will continue along the direction of the maximum SCS value. After the

```
Input : Graph $G = (V, E)$
Output : Overlapping community structure $C$
1: $C = \varnothing$;
2: According to seed community selection algorithm (Algorithm 2), seed community
   set, denoted by Seeds, are selected from network $G$;
3: Select any seed community, denoted by $s$, and go to Step 4 if Seeds $\neq \varnothing$. Otherwise,
   go to Step 5;
4: Remove $s$ from Seeds, and then expand it into a community structure $C_s$ according to
   the seed community extension algorithm (Algorithm 3), and add $C_s$ to $C$, returning to
   Step 3;
5: Simplify and re-detect overlapping nodes and possible missing nodes;
6: Output overlapping community structure $C$;
```

Algorithm 1: The OLCRE algorithm.

```
Input : Graph $G = (V, E)$
Output : Seed community set $Seeds$
$Seeds = \varnothing$;
for each $i \in V$ do
  if node $i$ has been accessed then
    continue;
  else
    mark node $i$ as visited;
  end if
  $max \leftarrow$ the $SCS$ value of node $i$ is calculated;
  while true do
    $value \leftarrow$ $SCS$ values of all neighbor nodes of node $i$ are calculated, and all
    neighbor nodes are marked as visited. The node with the maximum $SCS$ value is
    selected. If the node with the maximum $SCS$ value is not unique, a node $j$ with the
    maximum $SCS$ value is randomly selected;
    if $max >= value$ then
      $Seeds \leftarrow$ the subgraph formed by node $i$ and its neighbor nodes serves as a seed
      community;
      break;
    else
      $max = value$;
      $i = j$;
    end if
  end while
end for
```

Algorithm 2: Seed community selection algorithm.

node with the maximum SCS value is found in a region, the subgraph formed by this node and its neighbor nodes is regarded as a seed community. If the node with the maximum SCS value is not unique, a node is randomly selected. Then, the search for seed communities continues in unvisited areas of the network until all nodes in the network have been traversed. Finally, the subgraphs with tight internal connections and sparse external connections in all local core regions of the network have been searched and used as seed communities.

3.3. Seed Community Expansion. In the stage of seed community expansion, a novel seed community expansion strategy is designed according to the proposed node to the community interior influence function I and node to the community exterior influence function E, see Algorithm 3 for the specific process.

The new seed community expansion strategy is as follows: select any neighbor node i of the seed community and calculate the corresponding I value and E value of the node. If $I(i) > 0$ and $E(i) < 0$ are satisfied, the node will be added to the seed community; otherwise, it cannot be added to the seed community. When all neighbor nodes of the seed community do not meet the expansion strategy, the seed community stops expanding and then continues to expand the rest of the seed communities until all the seed communities have completed the expansion.

3.4. Dealing with Overlapping Nodes and Missing Nodes. After the expansion of all seed communities is completed,

```
Input : Graph $G = (V, E)$, Seed community set $Seeds$
Output : Overlapping community set $C$
$C = \varnothing$;
for each $s \in Seeds$ do
    $C_s = s$;
    While true do
        select any neighbor node $i$ of the seed community $C_s$;
        if $I(i) > 0$ and $E(i) < 0$ then
            $C_s = C_s \cup i$;
        end if
        if all neighbor nodes of seed community $C_s$ do not satisfy $I > 0$ and $E < 0$ then
            break;
        end if
    end while
    $C = C \cup C_s$;
end for
```

Algorithm 3: Seed community expansion algorithm.

if any node is not added to the community, the missing node will be added to the community with its corresponding maximum M value according to the community quality optimization function M. In addition, in order to prevent the excessive overlapping phenomenon from affecting the quality of community detection, it is necessary to simplify the detected overlapping nodes. The M value of the overlapping node corresponding to the community where it is located is calculated, respectively. If the M value is positive, the overlapping node is kept in the community where it is located; if the M value is negative, the overlapping node is removed from the community where it is located. When the M values of the overlapping node corresponding to the communities where it is located are all negative, it will be added to the community with the corresponding largest M value.

3.5. Time Complexity Analysis. Assume that the number of nodes in network G is n and the average degree of nodes is k. The number of seed communities, the number of overlapping nodes, and the number of missing nodes detected by the OLCRE algorithm are r, o, and l, respectively. Firstly, high-quality subgraphs are selected from local core areas of the network as seed communities, whose time complexity is $O(k^2 n)$. After that, the time complexity for all seed communities to complete the extension is $O(k^2 r + k^2 n)$. Finally, the time complexity of simplifying and redetecting overlapping nodes and missing nodes is $O(\mathrm{kor} + \mathrm{klr})$. To sum up, the time complexity of the OLCRE algorithm is $O(2k^2 n + k^2 r + \mathrm{kor} + \mathrm{klr})$. Since r, k, l, and o are far less than n, the time complexity of the OLCRE algorithm is about $O(jk^2 n)$, where j is a constant.

4. Experimental Results and Analysis

4.1. Experimental Data Sets

4.1.1. Artificial Networks. Since the LFR benchmark network [26] is very similar to the real-world complex network in the statistical characteristics of node degree and community size distribution, this paper uses this benchmark network as the test data set for the proposed algorithm and other comparison algorithms. The parameters of the LFR benchmark network are shown in Table 2.

Table 2: Parameters of LFR benchmark network.

Parameters	Meaning
n	The number of nodes in the network
k	The average degree of nodes
$k_{\max}$	The maximum degree of nodes
μ	The mixing parameter
$C_{\min}$	The number of nodes in the smallest community
$C_{\max}$	The number of nodes in the biggest community
O_n	The number of overlapping nodes
O_m	The number of memberships of the overlapping nodes

In order to objectively reflect the performance of each algorithm, four groups of different types of artificial networks (see Table 3) are generated by changing the mixing parameter μ, the number of overlapping nodes O_n, the number of memberships of the overlapping nodes O_m, and the number of nodes in the network n by using the LFR toolkit. They, respectively, are artificial network group N1 with a gradually fuzzy community structure, artificial network group N2 with a gradually increasing number of overlapping nodes, artificial network group N3 with a gradually increasing number of communities to which overlapping nodes belong, and artificial network group N4 with a gradually increasing number of nodes.

4.1.2. Real-World Networks. In order to compare the performance of each algorithm in detecting network community structure, seven real-world network data sets of different sizes and types are used in this paper. They, respectively, are the Zachary karate club network (Karate for short) [27], bottlenose dolphin network (Dolphins for short) [27], books about US politics network (Polbooks for short) [28],

Table 3: Parameter settings of the LFR benchmark networks.

Network	n	k	$k_{\max}$	$C_{\min}$	$C_{\max}$	O_n	O_m	μ
N1	2000	20	40	20	100	200	2	0.1~0.4
N2	2000	20	40	20	100	400~1000	2	0.1
N3	2000	20	40	20	100	50	3~6	0.1
N4	500~100000	20	40	20	100	$n/10$	2	0.1

US election blog network (Polblogs for short) [27], author collaboration network (Netscience for short) [27], trust network (PGP for short) [29], and friendship network (HR for short) [30]. The details of the seven real-world networks are listed in Table 4.

Table 4: The information of the seven real-world networks.

Network	Number of nodes	Number of edges	Average degree
Karate	34	78	4.59
Dolphins	62	159	5.13
Polbooks	105	441	8.4
Polblogs	1490	19022	25.53
Netscience	1588	2742	3.45
PGP	10680	24316	4.55
HR	54573	498202	18.26

4.2. Evaluation Metrics and Experimental Settings

4.2.1. Evaluation Metrics. Since the community structure of the artificial network is known, normalized mutual information (NMI for short) [12] is used as the evaluation metric of artificial network community detection results. NMI is used to measure the similarity between the community structure detected by the algorithm and the real community structure, and its value range is [0,1]. The more accurate the community structure detected by the algorithm, the larger the corresponding NMI value. The NMI is defined as follows:

$$\mathrm{NMI} = \frac{-2\sum_{x=1}^{C_N}\sum_{y=1}^{C_D} M_{\mathrm{xy}} \log\left(\left(M_{\mathrm{xy}} \times M\right)/\left(M_{x\cdot} \times M_{\cdot y}\right)\right)}{\sum_{x=1}^{C_N} M_{x\cdot} \log\left(M_{i\cdot}/M\right) + \sum_{y=1}^{C_D} M_{\cdot y} \log\left(M_{\cdot j}/M\right)}, \tag{8}$$

where C_N is the number of real communities in the artificial network and C_D is the number of communities detected by the algorithm on the artificial network. The rows of matrix M correspond to the real community results of the artificial network, and the columns of matrix M correspond to the community results detected by the algorithm on the artificial network. M_{xy} is the number of overlapping nodes between the real community x and the community y detected by the algorithm. $M_{x}\cdot$ is the sum of elements of M in row x and $M_{\cdot y}$ is the sum of elements of M in column y.

Since the community structure of the real-world network is unknown, the extend modularity (EQ for short) [31] is adopted as the evaluation metric of the community detection results of the real-world network. EQ is used to measure the tightness of community connection, and its value range is [0,1]. A higher EQ value means that the community quality detected by the algorithm is better. The EQ is defined as follows:

$$\mathrm{EQ} = \frac{1}{2m}\sum_{z=1}^{c}\sum_{i,j\in C_z}\frac{1}{O_iO_j}\left[A_{\mathrm{ij}} - \frac{k_ik_j}{2m}\right], \tag{9}$$

where m is the number of edges in the network. c is the number of communities detected by the algorithm in the real-world network. O_i is the number of communities to which node i belongs, and k_i is the degree of node i. A_{ij} is an adjacency matrix element of the network. $A_{\mathrm{ij}} = 1$ if there is an edge connection between nodes i and j. Otherwise, $A_{\mathrm{ij}} = 0$.

4.2.2. Experimental Settings. The OLCRE algorithm is tested on artificial network data sets and real-world network data sets and compared with overlapping community detection algorithms DNMF [16], CoEuS [32], MULTICOM [33], and APAL [34] to verify the effectiveness and feasibility of the OLCRE algorithm. The experimental running environment is a computer equipped with an Intel Core i9-11900K 3.50 GHz processor, 32 GB memory, and Windows 10 operating system. The algorithm proposed in this paper is programmed by MATLAB R2021a, and the source code has been publicly shared and is available at https://github.com/GitZhaoY/OLCRE.git.

Table 5 lists the year, programming language, and time complexity of each comparison algorithm, where m represents the number of edges in the network, n represents the number of nodes in the network, s represents the number of seeds, h represents the number of nodes within the seed community, c represents the number of communities, and t represents the number of iterations. From the data listed in Table 5, it can be seen that both the CoEuS algorithm and the MULTICOM algorithm have linear time complexity, which is on the same order of magnitude as the time complexity of the OLCRE algorithm proposed in this paper. The time complexity of the DNMF algorithm is $O(n^2)$ order of magnitude, which is significantly higher than that of the OLCRE algorithm. The time complexity of the APAL algorithm is $O(m^3/n^2)$, which indicates that it has good operating efficiency on sparse networks and is not suitable for dense networks.

4.3. Experimental Results on Artificial Networks. Figures 1–3 and Table 6, respectively, show the comparison results of the evaluation metric NMI obtained by each algorithm running

Table 5: The introduction of the comparison algorithms.

Algorithm	Year	Language	Time complexity
CoEuS	2017	Java	$O(\mathrm{shm})$
MULTICOM	2018	Python	$O(\mathrm{cn})$
DNMF	2019	MATLAB	$O\left(\mathrm{tcn}^2 + \mathrm{tc}^2 n\right)$
APAL	2021	Python	$O\left(m^3/n^2\right)$

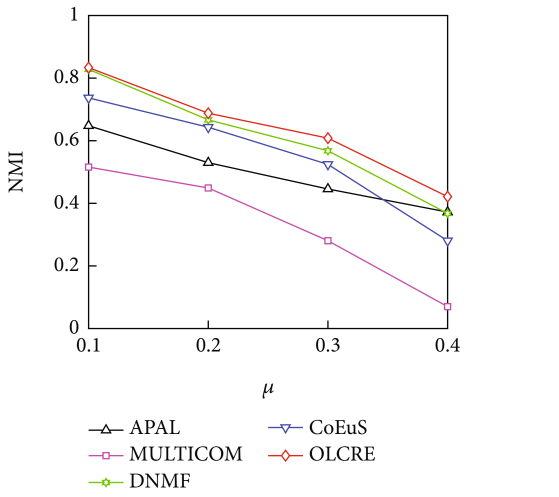

Figure 1: The comparison results of NMI values obtained by each algorithm on the network group N1 with a gradually fuzzy community structure.

on four groups of different types of artificial networks. In the network group N1, with the increase of μ value, that is, the network community structure is gradually blurred, the community detection accuracy of each algorithm decreases, but the community detection accuracy of the OLCRE algorithm is better than that of each comparison algorithm under different μ values. In the network group N2 with a gradually increasing number of overlapping nodes and the network group N3 with a gradually increasing number of communities to which overlapping nodes belong, the community detection accuracy of the OLCRE algorithm is better than that of each comparison algorithm. From the experimental data listed in Table 6 ("\\" means that the algorithm failed to detect communities in this experimental running environment), it can be seen that the community detection accuracy of the OLCRE algorithm is relatively stable and better than that of each comparison algorithm in the network group N4 with gradually increasing number of nodes.

According to the above experimental results, it is shown that the seed community selection method and community expansion strategy of the OLCRE algorithm proposed in this paper are effective and can be applied to networks of different scales and types.

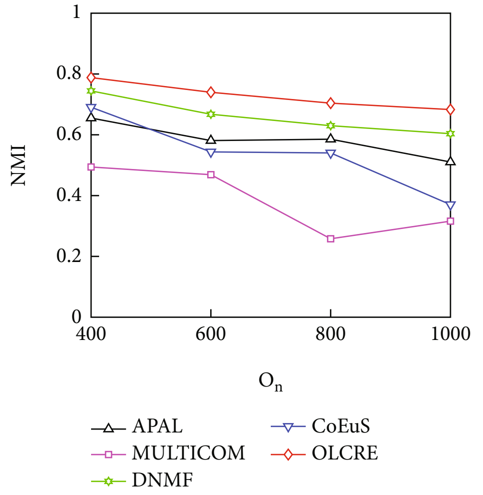

Figure 2: The comparison results of NMI values obtained by each algorithm on the artificial network group N2 with a gradually increasing number of overlapping nodes.

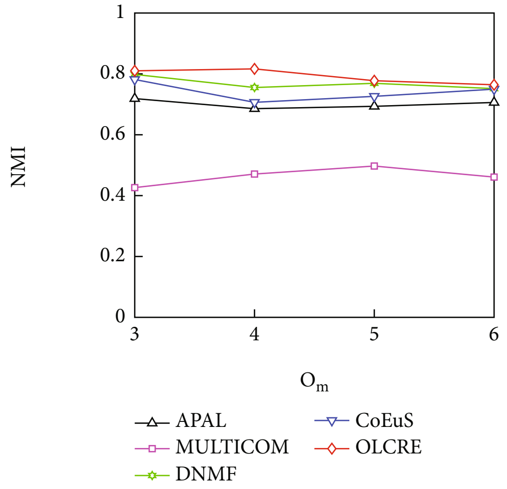

Figure 3: The comparison results of NMI values obtained by each algorithm on the artificial network group N3 with a gradually increasing number of communities to which overlapping nodes belong.

4.4. Experimental Results on Real-World Networks. Table 7 lists the results of EQ values obtained by the OLCRE algorithm and other four overlapping community detection algorithms running on seven real-world network data sets ("\\" means that the algorithm failed to detect communities in this experimental running environment). As can be seen from the experimental results listed in Table 7, the EQ values obtained by the OLCRE algorithm on the Dolphins network, Polbooks network, Polblogs network, Netscience network, PGP network, and HR network are all higher than those obtained by each comparison algorithm. The EQ value obtained by the OLCRE algorithm only on the Karate network is slightly lower than that obtained by the DNMF algorithm.

The reason why the OLCRE algorithm does not obtain the maximum EQ value on the Karate network is analyzed below. Figures 4 and 5, respectively, show the community detection results of the OLCRE algorithm and the DNMF algorithm on the Karate network. It can be seen from the comparative analysis of Figures 4 and 5 that the DNMF algorithm does not detect overlapping nodes in the Karate

TABLE 6: The comparison results of NMI values obtained by each algorithm on the artificial network group N4 with a gradually increasing number of nodes.

NMI	$n = 500$	$n = 5000$	$n = 50000$	$n = 100000$
OLCRE	0.8234	0.8142	0.8238	0.8266
CoEuS	0.6672	0.7487	0.7147	0.0080
MULTICOM	0.5618	0.4384	0.4278	0.4680
DNMF	0.7376	0.7700	\\	\\
APAL	0.5871	0.7126	0.6535	0.6524

TABLE 7: The comparison results of EQ values obtained by each algorithm on the real-world networks.

EQ	OLCRE	CoEuS	MULTICOM	DNMF	APAL
Karate	0.3678	0.2275	0.1038	0.3715	0.2698
Dolphins	0.4905	0.3249	0.3252	0.4804	0.2938
Polbooks	0.4569	0.3686	0.4304	0.4451	0.3395
Polblogs	0.4206	0.0707	0.0795	0.3438	0.0050
Netscience	0.9177	0.1506	0.1480	0.8092	0.7905
PGP	0.6466	0.5077	0.1432	0.5664	0.3307
HR	0.4078	0.0020	0.0060	\\	0.0340

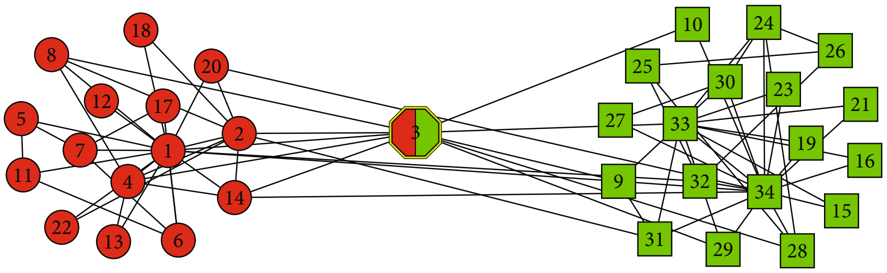

FIGURE 4: The community detection result of the OLCRE algorithm on the Karate network.

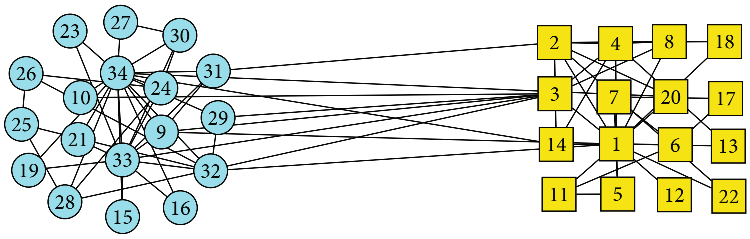

FIGURE 5: The community detection result of the DNMF algorithm on the Karate network.

network, while the OLCRE algorithm detects node 3 as the overlapping node, which loses some connection tightness. Therefore, the EQ value obtained by the OLCRE algorithm is slightly lower than that obtained by the DNMF algorithm.

5. Conclusions

The OLCRE algorithm proposed in this paper firstly selects high-quality subgraphs from all local core regions of the network as seed communities according to the proposed seed community selection function. Then, the seed communities are expanded in turn according to the proposed expansion strategy. Finally, after the completion of all seed community expansion, overlapping nodes and possible missing nodes should be simplified and redetected to further improve the quality of community detection. In this paper, four groups of artificial networks with different types and scales are designed and compared with several overlapping community algorithms. The community detection accuracy of the OLCRE algorithm on these four groups of artificial networks is better than that of each comparison algorithm. In the experiments on seven real-world networks, the OLCRE algorithm only fails to obtain the maximum value of EQ on the Karate network, and the results on the other six real-world networks are all higher than those of the comparison algorithms. In conclusion, the experimental results verify that the OLCRE algorithm is effective and feasible. In addition, the OLCRE algorithm does not need to set any parameters and only needs to master the basic network information (nodes and edges) to complete the detection of overlapping communities. It can be applied to networks of different scales and types and has universal application.

Data Availability

The data sets used to support this study are obtained from http://snap.stanford.edu/data/ and http://www-personal.umich.edu/~mejn/netdata/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61370083), the Humanity and Social Science Research Project of Ministry of Education of China (no. 22JDSZ3023), the ZheJiang Provincial Natural Science Foundation of China (no. LY15F020040), the ZheJiang Provincial Education Science Planning Project of China (no. 2020SCG046), and the Ministry of Education's Industry-University Collaboration Education Project (nos. 220603372015422 and 220604029012441).

References

[1] M. Eirinaki, J. Gao, I. Varlamis, and K. Tserpes, "Recommender systems for large-scale social networks: a review of challenges and solutions," *Future Generation Computer Systems*, vol. 78, pp. 413–418, 2018.

[2] M. Shimobayashi and M. N. Hall, "Making new contacts: the mTOR network in metabolism and signalling crosstalk," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 3, pp. 155–162, 2014.

[3] D. Jin, D. Y. Liu, B. Yang et al., "Fast complex network clustering algorithm using local detection," *Acta Electonica Sinica*, vol. 39, no. 11, pp. 2540–2546, 2011.

[4] C. Qian, J. Cao, J. Lu, and J. Kurths, "Adaptive bridge control strategy for opinion evolution on social networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 21, no. 2, article 025116, 2011.

[5] A. Sidiropoulos, G. Pallis, D. Katsaros, K. Stamos, A. Vakali, and Y. Manolopoulos, "Prefetching in content distribution networks via web communities identification and outsourcing," *World Wide Web*, vol. 11, no. 1, pp. 39–70, 2008.

[6] Z. Y. Yu, J. J. Chen, K. Guo, Y. Z. Chen, and Q. Xu, "Overlapping community detection based on influence and seeds extension," *Acta Electonica Sinica*, vol. 47, no. 1, pp. 153–160, 2019.

[7] G. Ren and X. Wang, "Epidemic spreading in time-varying community networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 24, no. 2, article 023116, 2014.

[8] J. Xie, B. K. Szymanski, and X. Liu, "Slpa: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *IEEE 11th international conference on data mining workshops*, pp. 344–349, Vancouver, Canada, 2011.

[9] K. Deng, W. P. Li, F. H. Yu et al., "Overlapping community detection in complex networks based on multi kernel label propagation," *Journal on Communications*, vol. 38, no. 2, article 5366, 2017.

[10] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[11] J. Ma and J. Fan, "Local optimization for clique-based overlapping community detection in complex networks," *IEEE Access*, vol. 8, pp. 5091–5103, 2019.

[12] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, article 033015, 2009.

[13] K. Berahmand, A. Bouyer, and M. Vasighi, "Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 1021–1033, 2018.

[14] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[15] Z. Liu, H. Wang, G. Wang, and Y. Zhou, "Link community detection based on ensemble learning," *Modern Physics Letters B*, vol. 34, no. 27, p. 2050293, 2020.

[16] F. Ye, C. Chen, Z. Zheng, R. H. Li, and J. X. Yu, "Discrete overlapping community detection with pseudo supervision," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 708–717, New York, USA, 2019.

[17] K. Berahmand, M. Mohammadi, F. Saberi-Movahed, Y. Li, and Y. Xu, "Graph regularized nonnegative matrix factorization for community detection in attributed networks," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 1, pp. 372–385, 2023.

[18] H. V. Lierde, T. W. S. Chow, and G. Chen, "Scalable spectral clustering for overlapping community detection in large-scale networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 4, pp. 754–767, 2020.

[19] K. Berahmand, M. Mohammadi, A. Faroughi, and R. P. Mohammadiani, "A novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrix," *Cluster Computing*, vol. 25, no. 2, pp. 869–888, 2022.

[20] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping community detection using neighborhood-inflated seed expansion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 5, pp. 1272–1284, 2016.

[21] L. Bai, X. Cheng, J. Liang, and Y. Guo, "Fast graph clustering with a new description model for community detection," *Information Sciences*, vol. 388-389, pp. 37–47, 2017.

[22] H. J. Li, Z. Bu, A. Li, Z. Liu, and Y. Shi, "Fast and accurate mining the community structure: integrating center locating and membership optimization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2349–2362, 2016.

[23] C. Lee, F. Reid, A. McDaid, and N. Hurley, "Detecting highly overlapping community structure by greedy clique expansion," https://arxiv.org/abs/1002.1827.

[24] P. Liakos, A. Ntoulas, and A. Delis, “Scalable link community detection: a local dispersion-aware approach,” in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 716–725, Washington DC, USA, 2016.

[25] R. Kanawati, “Empirical evaluation of applying ensemble methods to ego-centred community identification in complex networks,” *Neurocomputing*, vol. 150, pp. 417–427, 2015.

[26] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms,” *Physical Review E*, vol. 78, no. 4, 2008.

[27] M. E. J. Newman, “Network data from Mark Newman's home page,” 2012, http://www-personal.umich.edu/~mejn/netdata/.

[28] M. E. J. Newman, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[29] M. Boguná, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, “Models of social networks based on social distance attachment,” *Physical Review E*, vol. 70, no. 5, 2004.

[30] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, “Gemsec: graph embedding with self clustering,” in *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pp. 65–72, Vancouver, Canada, 2019.

[31] H. W. Shen, X. Q. Cheng, and J. F. Guo, “Quantifying and identifying the overlapping community structure in networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 7, article 07042, 2009.

[32] P. Liakos, A. Ntoulas, and A. Delis, “COEUS: community detection via seed-set expansion on graph streams,” in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 676–685, Boston, MA, USA, 2017.

[33] A. Hollocou, T. Bonald, and M. Lelarge, “Multiple local community detection,” *ACM SIGMETRICS Performance Evaluation Review*, vol. 45, no. 3, pp. 76–83, 2018.

[34] O. Doluca and K. Oğuz, “APAL: adjacency propagation algorithm for overlapping community detection in biological networks,” *Information Sciences*, vol. 579, pp. 574–590, 2021.