Hindawi
Complexity
Volume 2021, Article ID 7428927, 11 pages
https://doi.org/10.1155/2021/7428927

Research Article

A Local Extended Algorithm Combined with Degree and Clustering Coefficient to Optimize Overlapping Community Detection

Jing Liu,[1] **Junfang Guo**,[2] **and Qi Li**[1]

[1]*Department of Computer Science and Engineering, Shaoxing University, Shaoxing, Zhejiang, China*
[2]*Department of Computer Science and Engineering, Wenzhou University, Wenzhou, Zhejiang, China*

Correspondence should be addressed to Junfang Guo; gjf@wzu.edu.cn and Qi Li; liqi0713@foxmail.com

Received 27 August 2021; Revised 29 November 2021; Accepted 9 December 2021; Published 29 December 2021

Academic Editor: Siew Ann Cheong

Community structure is one of the most important characteristics of complex networks, which has important applications in sociology, biology, and computer science. The community detection method based on local expansion is one of the most adaptable overlapping community detection algorithms. However, due to the lack of effective seed selection and community optimization methods, the algorithm often gets community results with lower accuracy. In order to solve these problems, we propose a seed selection algorithm of fusion degree and clustering coefficient. The method calculates the weight value corresponding to degree and clustering coefficient by entropy weight method and then calculates the weight factor of nodes as the seed node selection order. Based on the seed selection algorithm, we design a local expansion strategy, which uses the strategy of optimizing adaptive function to expand the community. Finally, community merging and isolated node adjustment strategies are adopted to obtain the final community. Experimental results show that the proposed algorithm can achieve better community partitioning results than other state-of-the-art algorithms.

1. Introduction

Complex networks are ubiquitous in the real world, such as social networks, academic cooperation networks, world wide networks, and biological networks [1]. They are generally composed of nodes (individuals) and edges (relationships between individuals). For example, in social networks, nodes represent people and edges represent relationships between people. Although these networks belong to different fields, they follow the same laws. (1) Small world effect: complex networks have small average paths and large aggregation coefficients. (2) Scale-free: the degree of nodes in the network obeys the power-law distribution. (3) Community structure: the network can be divided into multiple groups with relatively close internal edges, and the connections between groups are relatively sparse.

Community structure is one of the most important structural features of complex networks [2]. It is ubiquitous in various complex networks in the real world. For example, in social networks, individuals with common interests have closer relationships and form communities of common interests. The community detection technology can predict the hobbies of new network users. In the protein network [3], proteins with the same or similar functions constitute each community. Community detection technology can identify the group of unfamiliar proteins and thus discover the protein function. In the academic cooperation network, scholars who have similar research directions or have participated in similar projects constitute a community. Community detection technology can help cross-research projects between statistics departments.

According to the different characteristics of community structure, the community structure is divided into two categories: nonoverlapping community [4] and overlapping community [5, 6]. Nonoverlapping community means that the nodes in the network belong to only one community;

each community exists independently. In the real world, network communities are often not independent. Different communities usually share some nodes, which is called overlapping community, and the nodes shared among communities are called overlapping nodes. Figure 1 is a nonoverlapping community network. It can be seen from Figure 1 that there are no public nodes between the two communities. Figure 2 shows the overlapping community structure, and the black circle represents an overlapping node.

Overlapping community detection in complex networks has attracted the attention of many scholars and achieved many results [7], such as LFM [8], COPRA [9], and LINK [10]. However, most overlapping community algorithms lack effective seed selection and community optimization methods, and these algorithms often get community results with low accuracy [11]. To solve these problems, this paper proposes an overlapping community detection algorithm based on information fusion. The main contributions of this paper are as follows:

(i) We propose an overlapping community detection algorithm based on information fusion, which improves the quality of community detection through an effective seed selection method and community optimization methods.

(ii) We propose a seed selection method of fusion degree and clustering coefficient. The method combines the weight factor, degree, and clustering coefficient to calculate the node importance, which ensures that the seed has a large total node influence and there is a high degree of similarity between nodes.

(iii) Finally, we verify the algorithm's performance on synthetic networks and real networks. The experimental results show that compared with the state-of-the-art algorithms, our proposed method can find more accurate community structures.

The remainder of this paper is organized as follows: in Section 2, the related work of locally extended overlapping community detection method is introduced; Section 3 describes the implementation of overlapping community detection algorithm based on information fusion; Section 4 gives the specific experimental results; and finally, the research work of this paper is summarized in Section 5.

2. Related Work

Overlapping community detection method based on local expansion is one of the most important methods to deal with the problem of overlapping community detection in large-scale networks [12], which includes two steps: firstly, some nodes or some node sets in the network are selected as the seed of each community and continue to expand outward through the fitness function (optimization function) until a certain termination condition is met to form a community. Finally, the fitness function reaches the local optimal value as the termination condition for the end of community

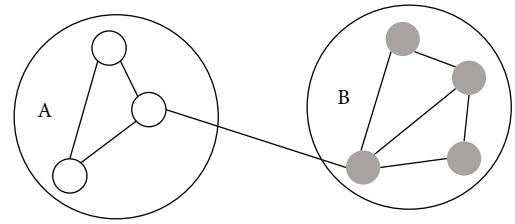

Figure 1: An example of nonoverlapping community structure.

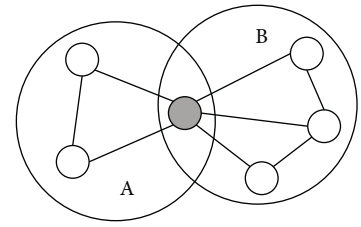

Figure 2: An example of overlapping community structure.

expansion. Formally speaking, given an undirected network $G(V, E)$, a set of seed nodes S, and a fitness function $f(C)$, the goal of the community expansion process is to find a subgraph C and make $f(C)$ reach the local optimal value [13] ($S \subseteq C$). The subgraph C is the result of the community expanded by S. At present, there are a large number of community quality evaluation functions that can be used as the fitness function of community expansion, such as modularity [14], subgraph density [15], centrality [16], conductivity [17], and edge-surplus [18].

Since the expansion process of each seed is independent, an overlapping community structure is formed when the two seeds are expanded to the same node. Meanwhile, the expansion process of the method generally only needs the local network information, which has high efficiency and is extremely suitable for large-scale networks. Lancichinetti et al. [19] proposed the LFM algorithm, which is a typical representative of the local expansion method. A subgraph G of the network is defined as follows:

$$f_G = \frac{k_{\text{in}}^G}{\left(k_{\text{in}}^G + k_{\text{out}}^G\right)^{\alpha}}, \tag{1}$$

where k_{in}^G represents the total internal degree of all nodes in the subgraph and k_{out}^G is the total externality. The symbol α is an adjustable parameter. Meanwhile, the fitness of a node A relative to a subgraph G is defined as follows:

$$f_G^A = f_{G+\{A\}} - f_{G-\{A\}}, \tag{2}$$

where $G + A$ and $G - A$ indicate that the node A belongs to G and does not belong to G, respectively. LFM algorithm first randomly selects a node in the network as an initial subgraph, continuously adds nodes around the subgraph, and removes nodes inside the subgraph according to formula (1) to increase the fitness of the subgraph until the fitness remains unchanged. Then, it selects a node that does not belong to any community and repeats the above expansion process until all nodes in the network belong to at least one community.

Lee et al. [20] proposed the GCE algorithm, which selects the maximum cliques with the number of nodes not less than K as the seeds and uses the same fitness function as the LFM algorithm to expand each community. Baumes et al. [21] proposed two overlapping community detection algorithms based on local expansion: IS (Iterative Scan) and RARE (Rank Eemoval). These two algorithms are mainly oriented to directed networks. The RARE algorithm takes some nonadjacent subgraphs of the network as seeds, removes the moderately large nodes in the network, and then adds the removed nodes to each community that can increase the community density function. Andersen et al. [22] proposed a community expansion method based on PageRank. The algorithm takes an initial node in the network as an extension object. Firstly, the approximate PageRank vector p starting from node u is calculated, and then a sweep technique is used to select the node set with the best conductivity around node u. The nodes in the network are sorted according to $p(v_i)/d(v_i)$ from large to small (where d (v_i) represents the degree of node v_i), and a node sequence is generated. The first k nodes in the sequence are selected to form a set. Different k values correspond to different sets, and the set with the lowest conductivity is the community expanded by node u.

Based on the above work, Silistre et al. [23] proposed two selection seed selection strategies: GRACLUS CENTERS and SPREAD HUBS. Both methods select a single node as a seed and use the community expansion method proposed by Andersen et al. [22]. Zhang et al. proposed [24] the CFCD algorithm and defined the core similarity. On this basis, they defined the core centrality of nodes and the core fitness of communities. The algorithm selects the node with the largest centrality that is not in the core of any existing community as a seed and takes the set composed of the node and its adjacent nodes as the initial community. However, due to the lack of fast and effective seed selection and community optimization methods, these algorithms often get community results with lower accuracy. In order to solve these problems, we propose a seed selection method based on the importance of nodes based on the degree of fusion and clustering coefficient, which ensures that the seeds have a large total node influence and also ensures that the internal nodes of the seeds have a high degree of similarity.

3. Proposed Method

In this section, we introduce the implementation process of the algorithm in detail. The main steps of overlapping community detection algorithm based on information fusion (OCDIF) include seed nodes selection (Section 3.1), local community expansion (Section 3.2), community merging (Section 3.3), and isolated nodes adjustment (Section 3.4). The principle and process of each step are described in detail as follows.

3.1. Seed Node Selection. In the existing overlapping community detection methods based on local expansion, the selection of seed nodes is random. However, it can not obtain a better community structure, and the result of community detection is unstable. In this paper, we take the node influence value as the node importance index and the node influence value as the selection order of seed nodes. The influence value of the node is large, which indicates that the node occupies an important position in the network. Then, the community structure can be guaranteed to have a certain reference value by expanding the community from this node. Moreover, due to the fixed selection order of seed nodes, stable community detection results can be obtained. It overcomes the defect of poor stability of existing overlapping community detection results based on local expansion.

The more neighbors of a node in the network, the greater the influence of this node, so the propagation ability of a node mainly depends on the sum of its direct neighbor degrees. However, considering the local properties of network nodes, different nodes with the same sum of direct neighbor degrees may have different propagation capabilities. Therefore, in addition to the nodes degree, other nodes attributes need to be considered. At the same time, the topological connection between the nodes and its neighbors also has an impact on the propagation ability. The greater the clustering coefficient, the greater the importance of the node.

Combining the degree and clustering coefficient of nodes and entropy weight, we propose a seed selection method of fusion degree and clustering coefficient. Firstly, the concepts of degree and clustering are introduced. The degree value of a node in an undirected network is defined as the number of nodes directly connected to the node. Given an undirected network $G = (V, E)$, the corresponding adjacency matrix $A = \{a_{ij}\}_{n*n}$, D_i indicates the ability of a node to communicate directly with other nodes. The larger the $D(i)$ value, the more important the node is:

$$D(i) = \sum_{j=1}^{N} a_{ij}, \tag{3}$$

where k_i represents the total number of neighbors of the node i and e_i represents the actual number of undirected edges between k_i neighbors.

Firstly, the degree and clustering coefficient of nodes are normalized to dimensionless. The matrix R is created according to the normalized value. n represents the number of nodes in the network, r_{1j} represents the normalized value of the degree of the node j, and r_{2j} represents the normalized value of the clustering coefficient of the node j, as shown in

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \end{bmatrix}. \tag{4}$$

Secondly, calculate the entropy value E_i of the node i, as shown in

$$E_i = -\ln\ (n)^{-1} \sum_{j=1}^{n} r_{ij} \ln\left(r_{ij}\right). \tag{5}$$

The weight w_i of degree and clustering coefficient is obtained according to entropy E_i, as shown in

$$w_i = \frac{1 - E_i}{2 - \sum_{i=1}^{2} E_i}, \quad i = 1, 2. \tag{6}$$

The weight factor $V_w(i)$ of the node i is calculated according to the weight calculated by the above formula. The specific calculation formula is as follows:

$$V_w(i) = w_1 \times D(i) + w_2 \times CC(i), \tag{7}$$

where $V_w(i)$ represents the weight factor of the node i. $D(i)$ and $CC(i)$ represent the degree value and the processed clustering coefficient of node i, respectively. w_1 and w_2 represent the contribution weight of the degree value and the clustering coefficient, respectively.

Finally, formula (8) is the calculation process of the node importance CLC. Arrange the nodes in descending order and store them in the vector X, select node A that is not allocated to any community from X in order, and treat node A as a local community C. The pseudocode of the nodes' importance evaluation is shown in Algorithm 1.

$$\mathrm{CLC}(i) == (i) \times V_w(i) + \sum_{j \in \Gamma(i)} V_w(j). \tag{8}$$

3.2. Local Community Expansion. This section mainly introduces the process of local community expansion. The condition for starting and stopping is that the adaptation function reaches the local maximum. The adaptation function is shown in

$$f(c) = \frac{K_C^{\mathrm{in}}}{\left(K_C^{\mathrm{in}} + K_C^{\mathrm{out}}\right)^{\alpha}}, \tag{9}$$

where K_C^{in} and K_C^{out} represent the total number of internal and external degrees of local community, respectively. α is a parameter greater than 0, which is used to adjust the community scale. Lanchinetti pointed out that the result of community detection is the best when α is 0.9, so all experiments in this paper set the value of parameter α to 0.9. Each time the algorithm expands the local community, the neighbors of the local community are added with a marker bit, which indicates that the node has joined the community.

If a node is only connected to the current node, it is considered that the node must have the same label as the current node. According to this idea, we introduce a concept "similar," as shown in Figure 3. All edges of an existing local community A and node 2 are connected not only to the nodes in local community a, but also the outside local community A. Therefore, node 2 cannot be directly added to A. It is necessary to calculate the fitness function value of node 2 before deciding whether they can be added to A. All edges connected with node 1 are in A, node 1 is directly added to local community A and the fitness function value will not be calculated. This greatly reduces the time of the algorithm and improves the quality of community detection.

3.3. Community Merging. After the expansion of local communities, there are many small-scale local communities in the divided community results because it also follows the trend of "birds of a feather flock together" in the network, forming a large-scale community structure rather than scattered small communities. In order to obtain the ideal result of community detection, these small communities need to be merged. Silistre et al. [23] gave the concept of community overlap, which is used to judge whether two communities can be merged into one community. The greater the degree of overlap is, the more reasonable it is to merge the two communities into one. The calculation of community overlap is shown in

$$\mathrm{OS}\left(C_i, C_j\right) = \frac{\left|V_i \cap V_j\right|}{\min\left(\left|V_i\right|, \left|V_j\right|\right)} + \frac{\left|E_i^{\mathrm{in}} \cap E_j^{\mathrm{in}}\right|}{\min\left(\left|E_i^{\mathrm{in}}\right|, \left|E_j^{\mathrm{in}}\right|\right)}, \tag{10}$$

where V_i and V_j represent the set of nodes in C_i and C_j and E_i^{in} and E_j^{in} represent the set of inner edges of C_i and C_j.

According to the degree of community overlap, it can be judged whether two communities can be merged into one community and calculate the average value of community overlap $\overline{\mathrm{OS}}$, which is shown in

$$\overline{\mathrm{OS}} = \frac{\sum_{C_i, C_j \in C} \mathrm{OS}\left(C_i, C_j\right)}{|C|}. \tag{11}$$

The algorithm first judges whether the overlap degree of any two communities is greater than $\overline{\mathrm{OS}}$. When the overlap degree $\mathrm{OS}(c_i, C_j) > \overline{\mathrm{OS}}$, the two communities are merged.

3.4. Isolated Nodes Adjustment. After the implementation of community merging, there are still isolated nodes. It is necessary to judge whether the isolated nodes can become a community. The judgment of an isolated node is mainly divided into two points. As shown in Figure 4, node 1 is an isolated node and is not connected to other nodes. At this time, the node can exist as an independent community.

Another situation is that, as shown in Figure 5, node 1 is an isolated node, but node 1 is connected to nodes 3 and 2. According to formula (12), there is node similarity S_{vw} between the isolated node and its neighbors:where k_v and k_w represent the degree of node v and node w, respectively. Then, calculate the average similarity between the node and all neighbor nodes $\overline{s}$ according to

$$S_{vw} = \frac{|\Gamma(v) \cap \Gamma(w)|}{\sqrt{k_v k_w}}, \tag{12}$$

$$\overline{S} = \frac{\sum_{w \in N} |\Gamma(v) \cap \Gamma(w)| / \sqrt{k_v k_w}}{|\Gamma(v)|}. \tag{13}$$

If $S_{vw} \geq \overline{S}$, it will be allocated to the neighbors community. The isolated node may be allocated to multiple communities. It is also in line with the requirements of overlapping communities. The pseudocode of community merging and isolated nodes adjustment are shown in Algorithm 2.

4. Experimental Results and Analysis

This article uses Python language to implement the OCDIF algorithm. The seed selection algorithm proposed in this

Input: a network $G(V, E)$, the number of nodes in the network n
Output: the importance of each node
(1) *Initialize* $D = \varnothing$, $CC = \varnothing$
(2) **for** i in n **do**
(3) Calculate $D(i)$ using formula (4) during D decomposition
(4) **end for**
(5) **for** i in n **do**
(6) Calculate $CC(i)$ using formula (5)
(7) **end for**
(8) Create matrix R using formula (6)
(9) **for** i in 2 **do**
(10) Calculate E_i using formula (7)
(11) **end for**
(12) **for** i in n **do**
(13) Calculate $CLC(i)$ using formula (10)
(14) **end for**

Algorithm 1: Node importance evaluation algorithm.

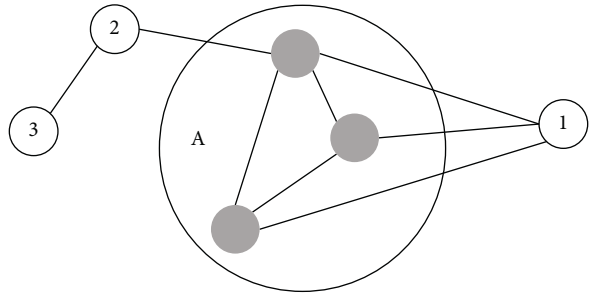

Figure 3: A special case.

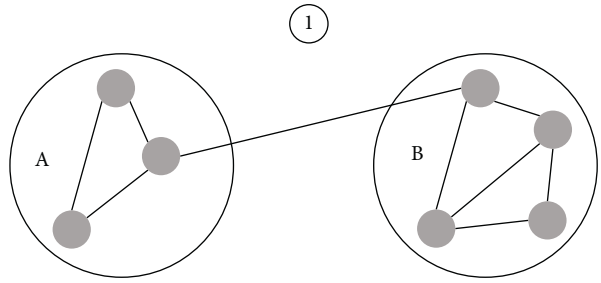

Figure 4: Isolated nodes without edges.

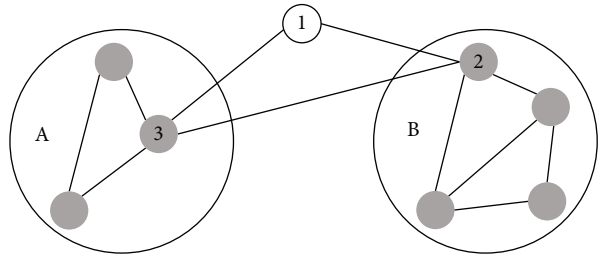

Figure 5: Isolated node with connected edges.

```
Input: local subgraph LC
Output: community detection results OC
(1)  ♯ Community merging
(2)  OS = calculateAvgOS(LC)
(3)  OC = []
(4)  for i in LC do
(5)    j = i + 1
(6)    for j in LC do
(7)      if OS_ij > OS then
(8)        OC.append(i ∪ j)
(9)      end if
(10)   end for
(11) end for
(12) ♯ Isolated nodes adjustment
(13) for i in OC do
(14)   if len(i) == 0 and otherSide(i[0]) then
(15)     v = i[0]
(16)     neighbors[v] = findeNeighbors(v)
(17)     S = calculateAvgS(neighbors[v], v)
(18)     for w in neighbors[v] do
(19)       if S_vw > S then
(20)       addIsolateNode(v, w,OC)
(21)       end if
(22)     end for
(23)   end if
(24) end for
```

Algorithm 2: Community merging and isolated nodes adjustment.

paper is compared with similar methods on the real network. At the same time, the OCDIF algorithm is compared with other overlapping community detection algorithms on real large-scale networks. The experimental environment is Core(TM) i5-4590, 3.3 GHz CPU, 16 GB memory.

4.1. Selection Effect of Seed Node. This section mainly verifies the effect and accuracy of the node importance ranking algorithm. The experimental data set uses the social friendly network (So-Colgate) [25] and the power grid network (PowerGrid) [25]. Among them, the socially friendly network consists of 3,482 nodes and 14,241 edges. The grid network consists of 4,940 nodes and 6,595 edges. For these three networks of different sizes, the node importance is calculated separately, and the ranking results and the ranking results calculated by the SIR propagation model are combined to determine each method (ControlRank [26], MBA [27], NIBNA [28], ODEF [28], CRRank [28]) accuracy.

Table 1 shows the experimental results. It can be found that the accuracy of the seed algorithm proposed in this paper in the So-Colgate network and PowerGrid network is slightly better than other algorithms within a given part of the propagation probability range. In terms of propagation probability, the MBA algorithm is also better, but overall it is lower than the algorithm proposed in this article. In terms of its average accuracy, the advantages of this article are even more obvious. On the network So-Colgate, compared to the ControlRank, MBA, NIBNA, ODEF, and CRRank algorithms, the algorithm proposed in this paper has an average increase of 18.4%, 34.5%, 37.1%, 31.5%, and 20.3%. On the network PowerGrid, compared to the ControlRank, MBA, NIBNA, ODEF, and CRRank algorithms, the algorithm proposed in this article has increased by an average of 10.4%, 4.8%, 24.4%, 22%, and 2.1%.

4.2. Community Detection Results. This section tests the performance of OCDIF on synthetic networks and real networks. We select the overlapping community detection algorithms based on the global structure and local structure of the static networks as the comparison objects of OCDIF (CLPA [29], GREESE [30], ILPA [31], LMD [32], McFFMM [33], MCMOEA [34], MPEA [35], and SSLPA [36]). We use the following two common indicators to evaluate the quality of the community detection: (1) *F1-Score* (average F1 value) and (2) *NMI* (normalized mutual information).

(1) *F1-Score*: this standard measures the accuracy of algorithm community detection by quantifying the degree of correspondence between the algorithm detection community and the real community. Given two community structures of a network $P_1 = C_1, \ldots, C_2$ and $P_2 = C_1', \ldots, C_n'$, the average F1 value is defined as follows:

$$\frac{1}{2}\left(\frac{1}{|P_1|}\sum_{C_i} F1\left(C_i, C_{f(i)}'\right) + \frac{1}{|P_2|}\sum_{C_i' \in P_2} F1\left(C_{f(i)}', C_i'\right),\right. \tag{14}$$

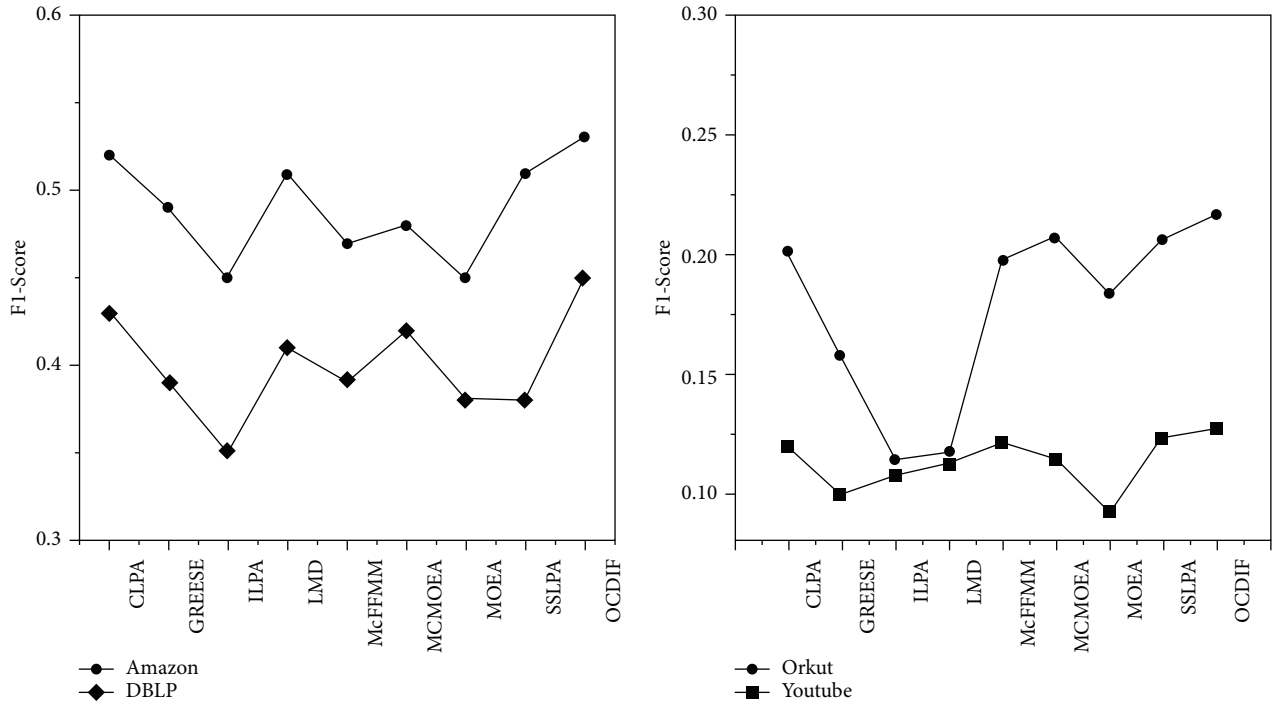

Figure 6: Average F1 value on networks.

where $F1(C_i, C'_j)$ is the harmonic average of the Precision and Recall between the two communities:

$$\text{precision}\left(C_i, C'_j\right) = \frac{\left|C_i \cap C'_j\right|}{\left|C_i\right|},$$

$$\text{recall}\left(C_i, C'_j\right) = \frac{\left|C_i \cap C'_j\right|}{\left|C'_j\right|}, \tag{15}$$

$$f(i) = \arg\max F1\left(C_i, C'_j\right),$$

$$f'(i) = \text{argmax}_j F1\left(C_j, C'_i\right).$$

(2) Generalized standard mutual information NMI (normalized mutual information): this standard is proposed to measure the accuracy of overlapping community detection algorithms. NMI evaluates the algorithm's accuracy by quantifying the similarity between the community discovered by the algorithm and the real community. The value range of NMI is [0, 1]. The larger the value, the higher the quality of the detection community.

4.3. Real Network. This paper uses the benchmark data set provided by SNAP [37] to conduct experiments. The network provided by this data set contains a real community structure, which is convenient for testing the algorithm. Table 2 shows the data of the four large-scale networks selected in this article. Next, we will briefly introduce these four networks:

DBLP: it is a collaborative network of authors. Each node in the network represents an author. If two authors have published at least one article together, then there is an edge connection between them. A journal or conference represents a community, and the community is composed of authors who have published articles in the journal or conference.

Amazon: it is a commodity network. Each node in the network represents a commodity. If two commodities are frequently purchased at the same time, there is an undirected edge between them. Each product category provided by Amazon corresponds to a real community.

YouTube: it is a YouTube social network. Each node represents a user of the network. If two users establish a friendship, then there is an edge connection between them. In this network, a community refers to a group created by users, and a community is a collection of users who join the group.

Orkut: It is a Orkut social network. Similar to the YouTube network, the nodes in this network represent users, and the edges represent the friendship between users. A community also refers to a group created by a user, and a community is a collection of users who have joined the group.

Figures 6 and 7 show the average F1 value and NMI value of the test algorithm on networks. Experimental results show that, in terms of the accuracy of community detection, the OCDIF algorithm outperforms all overlapping community detection algorithms based on global information and local information. On both networks, the OCDIF algorithm obtained the highest average F1 value and NMI value. Our

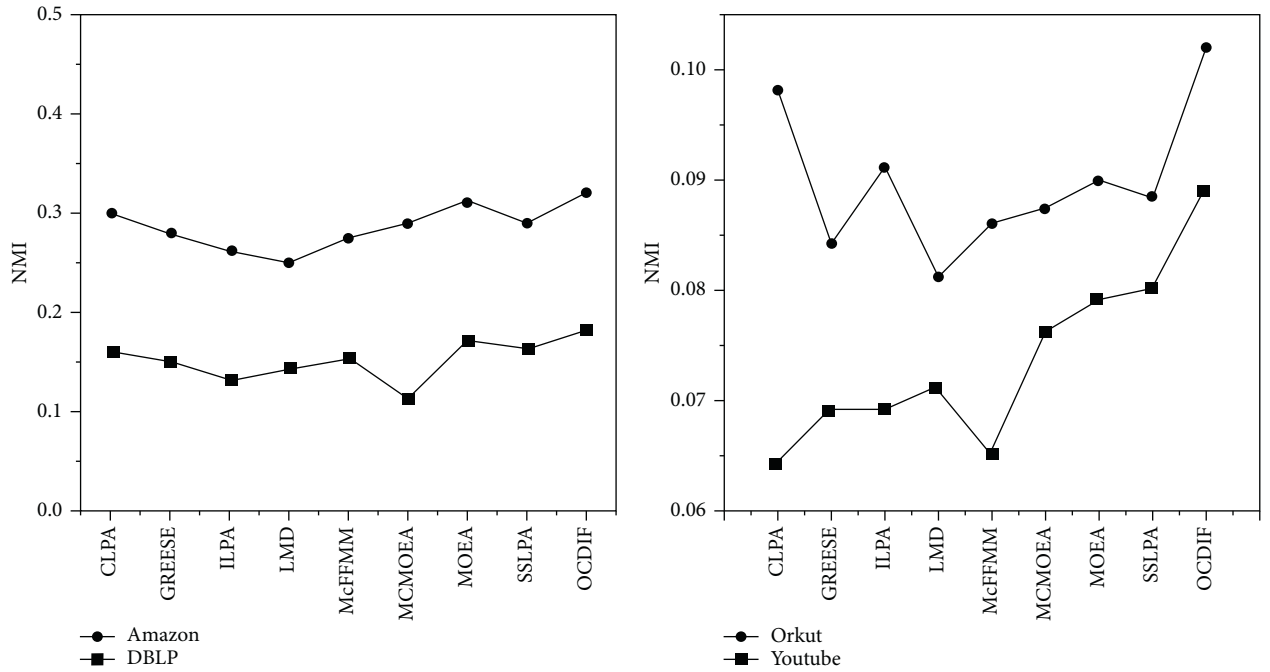

Figure 7: NMI value on networks.

Table 1: Comparison of the algorithm's accuracy in So-Colgate network and PowerGrid network.

Network	Node	Edge	OCDIF	ControlRank	MBA	NIBNA	ODEF	CRRank
So-Colgate	3482	14 241	**0.812**	0.797	0.784	0.511	0.556	0.647
PowerGrid	4940	6595	**0.704**	0.631	0.67	0.532	0.549	0.689

The significance for bold values in Table 1 is that the results of OCDIF are better than those of other algorithms.

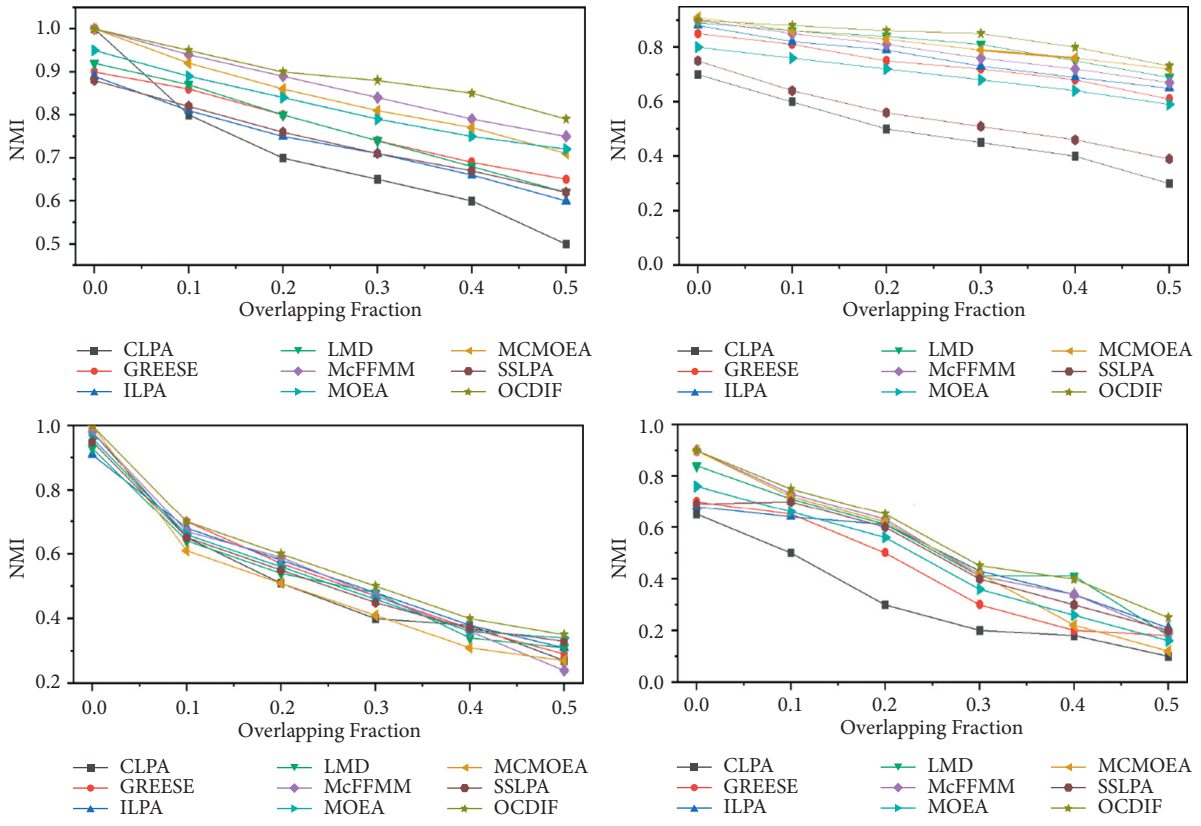

Figure 8: Experimental results on the synthetic network (N1, N2, N3, N4).

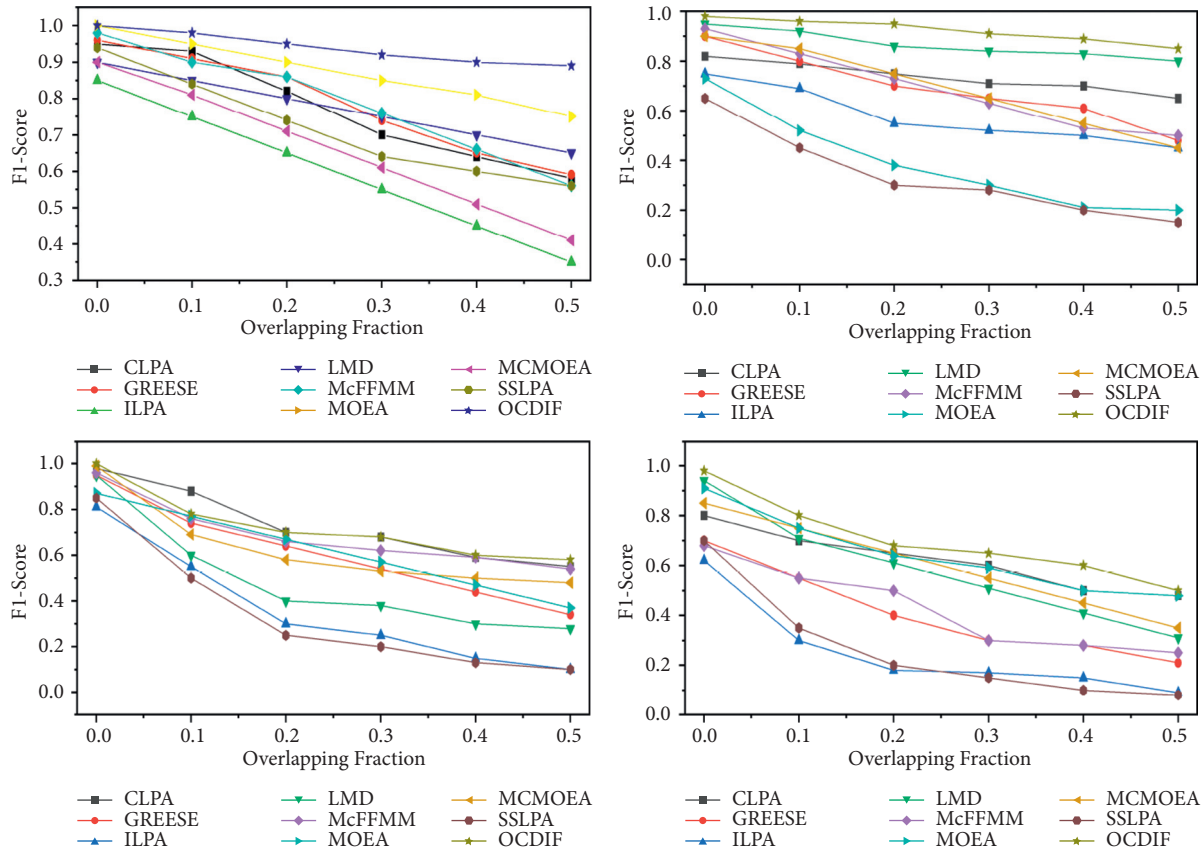

Figure 9: Experimental results on the synthetic network (N1, N2, N3, N4).

Table 2: Characteristics of the test network.

Network	Node	Edge	Community
DBLP	317,080	1,049,866	13,477
Amazon	334,863	925,872	75,149
YouTube	1,134,890	2,987,624	16,386
Orkut	3,072,441	117,185,083	15,301,901

Table 3: Data information of LFR network.

Network	o_m	m_u	o_n
N1	2	0.1	0–5000
N2	2	0.3	0–5000
N3	4	0.1	0–5000
N4	4	0.3	0–5000

algorithm is also the only local algorithm that exceeds all global algorithms in the test algorithm. The ILPA algorithm performed the worst, and the CLPA algorithm performed close to the OCDIF algorithm. On the DBLP network, the average F1 value of the OCDIF algorithm is 22.2% higher than ILPA, 20% higher than MOEA, and 6.6% higher than CLPA. On the Amazon network, the average F1 value of the OCDIF algorithm is 18.2% higher than ILPA, 17.3% higher than MOEA, and 6.4% higher than CLPA. On the DBLP network, the NMI value of the OCDIF algorithm is 27.8% higher than ILPA, 5.6% higher than MOEA, and 11.1% higher than CLPA. On the Amazon network, the NMI value of the OCDIF algorithm is 24.2% higher than ILPA, 6.1% higher than MOEA, and 9.1% higher than CLPA. Therefore, compared to these overlapping community detection algorithms for large-scale networks, the OCDIF algorithm can obtain a more accurate community structure in real networks.MCMOEA and SSLPA algorithms are relatively good, close to the algorithm's accuracy proposed in this article. The worst performing is GREESE and MOEA algorithms. On the YouTube network, the average F1 value of the OCDIF algorithm is 30.8% higher than MOEA, 7.7% higher than MCMOEA, and 15.4% higher than SSLPA. On the Orkut network, the average F1 value of the OCDIF algorithm is

20.8% higher than MOEA, 12.5% higher than MCMOEA, and 6.25% higher than SSLPA. On the YouTube network, the average NMI value of the OCDIF algorithm is 11.1% higher than MOEA, 16.7% higher than MCMOEA, and 10% higher than SSLPA. On the Orkut network, the average NMI value of the OCDIF algorithm is 18.2% higher than MOEA, 19.1% higher than MCMOEA, and 26.4% higher than SSLPA. Therefore, compared to these overlapping community detection algorithms for large-scale networks, the OCDIF algorithm can obtain a more accurate community structure in real networks. The experimental results show that, compared to the current mainstream overlapping community detection algorithms, the OCDIF algorithm can quickly and with high quality complete large-scale network overlapping community detection.

4.4. Artificial Synthetic Network. In this section, the LFR overlapping benchmarks proposed by Lancichinetti and Fortunato are selected to generate the experimental network. This overlap benchmark is widely used to evaluate the performance of overlapping community detection algorithms. The degree of generated network nodes and community size conform to a power-law distribution. In the previous chapter, we have introduced the parameters included in the LFR benchmark network. This chapter uses this overlapping benchmark to generate 4 groups of networks, which have the same parameter values as follows: $N = 10\,000$, $k = 15$, $\max_k = 50$, $\min_c = 10$, $\max_c = 50$, and other parameter values are shown in Table 3. Each group of networks contains 6 types of networks, in which the value range of on is [0, 0.5 N]; o_m is set to 2 and 4, respectively; m_u is set to 0.1 and 0.3, respectively, representing a low-mix network and a high-mix network.

This section also chooses the average F1 value and the generalized standard mutual information NMI (normalized mutual information), two evaluation indicators, to analyze the accuracy of the OCDIF algorithm community detection.

Figures 8 and 9 show the results of the NMI value and the average F1 value on a given artificial synthesis network. With the increase of overlapping nodes between communities, the community structure becomes more ambiguous, and the difficulty of finding the community increases. On the four groups of networks, each test algorithm has a different degree of reduction in the accuracy of finding the community. It can be seen from the data that the seed selection method proposed in this chapter is more stable when dealing with networks with fuzzy community structures. In addition, in the comparison algorithm, it can be seen that the seed selection method is superior to other algorithms in terms of the accuracy of expanding the community and the stability of dealing with networks with fuzzy community structure. The seed selection method proposed in this paper can get a more precise community structure.

5. Conclusion

We propose an information fusion overlapping community detection algorithm. The method is divided into four steps: seed node selection, local community expansion, community merging, and isolated nodes adjustment. Considering the local nature of network nodes, different nodes with the same direct neighbor degree may have different influences. We propose a seed selection method based on the degree of fusion and clustering coefficient, which ensures that the seeds have a large total node influence and also ensures that the internal nodes of the seeds have a high degree of similarity. The experimental results show that the algorithm greatly improves the efficiency of community detection and obtains more accurate results.

Most networks in real life are not static and will change over time, such as the removal and increase of edges between nodes in the network. As the nodes and edges in the network change, the community structure in the network will change accordingly. However, most of the existing community detection algorithms study static networks, and the research on dynamic networks is necessary and has great practical significance.

Data Availability

Data will be available at http://snap.stanford.edu/data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] X. Liu, Y. Du, M. Jiang, and X. Zeng, "Multiobjective particle swarm optimization based on network embedding for complex network community detection," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 437–449, 2020.

[2] M. Al-Andoli, S. C. Tan, and W. P. Cheah, "Parallel stacked autoencoder with particle swarm optimization for community detection in complex networks," *Applied Intelligence*, pp. 1–21, 2021.

[3] B. Škrlj, J. Kralj, and N. Lavrač, "Embedding-based silhouette community detection," *Machine Learning*, vol. 109, no. 11, pp. 2161–2193, 2020.

[4] Z. Ghalmane, M. El Hassouni, and H. Cherifi, "Immunization of networks with non-overlapping community structure," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–22, 2019.

[5] X. Teng, J. Liu, and M. Li, "Overlapping community detection in directed and undirected attributed networks using a multiobjective evolutionary algorithm," *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 138–150, 2019.

[6] Y. Li, J. He, Y. Wu, and R. Lv, "Overlapping community discovery method based on two expansions of seeds," *Symmetry*, vol. 13, no. 1, p. 18, 2021.

[7] A. Kumar, D. Barman, R. Sarkar, and N. Chowdhury, "Overlapping community detection using multiobjective genetic algorithm," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 3, pp. 802–817, 2020.

[8] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 70, no. 6, Article ID 066111, 2004.

[9] P. Chunaev, "Community detection in node-attributed social networks: a survey," *Computer Science Review*, vol. 37, Article ID 100286, 2020.

[10] H. Jiang, Z. Liu, C. Liu, Y. Su, and X. Zhang, "Community detection in complex networks with an ambiguous structure using central node based link prediction," *Knowledge-Based Systems*, vol. 195, Article ID 105626, 2020.

[11] T. He, L. Bai, and Y.-S. Ong, "Vicinal vertex allocation for matrix factorization in networks," *IEEE Transactions on Cybernetics*, vol. 2021, Article ID 3051606, 2021.

[12] Y. Zhou, G. Sun, Y. Xing, R. Zhou, and Z. Wang, "Local community detection algorithm based on minimal cluster," *Applied Computational Intelligence and Soft Computing*, vol. 2016, Article ID 3217612, 2016.

[13] Y. Fang, Y. Yang, W. Zhang, X. Lin, and X. Cao, "Effective and efficient community search over large heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 13, no. 6, pp. 854–867, 2020.

[14] J.-X. Yang and X.-D. Zhang, "Finding overlapping communities using seed set," *Physica A: Statistical Mechanics and Its Applications*, vol. 467, pp. 96–106, 2017.

[15] X. Wang, G. Liu, J. Li, and J. P. Nees, "Locating structural centers: a density-based clustering method for community detection," *PLoS One*, vol. 12, no. 1, Article ID e0169355, 2017.

[16] E. Nathan, A. Zakrzewska, J. Riedy, and D. Bader, "Local community detection in dynamic graphs using personalized centrality," *Algorithms*, vol. 10, no. 3, p. 102, 2017.

[17] 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.

[18] J. Cao, S. Wang, and H. Wang, "Detecting communities on topic of transportation with sparse crowd annotations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 1017–1022, 2016.

[19] 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 ID 033015, 2009.

[20] C. Lee, F. Reid, A. McDaid, and N. Hurley, "Detecting highly overlapping community structure by greedy clique expansion," arXiv preprint arXiv:1002.1827.

[21] J. Baumes, M. K. Goldberg, M. S. Krishnamoorthy, M. Magdon-Ismail, and N. Preston, "Finding communities by clustering a graph into overlapping subgraphs," *IADIS AC*, vol. 5, pp. 97–104, 2005.

[22] R. Andersen, F. Chung, and K. Lang, "Local graph partitioning using pagerank vectors," in *Proceedings of the 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pp. 475–486, IEEE, Berkeley, CA, USA, October 2006.

[23] A. Silistre, O. Kilincceker, F. Belli, M. Challenger, and G. Kardas, "Community detection in model-based testing to address scalability: study design," in *Proceedings of the 2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pp. 657–660, IEEE, Sofia, Bulgaria, September 2020.

[24] J. Zhang, X. Ding, and J. Yang, "Revealing the role of node similarity and community merging in community detection," *Knowledge-Based Systems*, vol. 165, pp. 407–419, 2019.

[25] I. M. Kloumann and J. M. Kleinberg, "Community membership identification from small seed sets," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1366–1375, New York, NY, USA, August 2014.

[26] J. Zhou, X. Yu, and J.-A. Lu, "Node importance in controlled complex networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 3, pp. 437–441, 2018.

[27] J. Zeng, C. Shao, X. Wang, and F. Miao, "Evaluation method for node importance based on attraction between nodes," *International Journal of Modern Physics C*, vol. 29, no. 12, Article ID 1850125, 2018.

[28] M. Xu, J. Wu, M. Liu, Y. Xiao, H. Wang, and D. Hu, "Discovery of critical nodes in road networks through mining from vehicle trajectories," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 2, pp. 583–593, 2018.

[29] J. H. Chin and K. Ratnavelu, "Detecting community structure by using a constrained label propagation algorithm," *PLoS One*, vol. 11, no. 5, Article ID e0155320, 2016.

[30] K. Asmi, D. Lotfi, and A. Abarda, "The greedy coupled-seeds expansion method for the overlapping community detection in social networks," *Computing*, pp. 1–19, 2021.

[31] S. Dong, "Improved label propagation algorithm for overlapping community detection," *Computing*, vol. 102, no. 10, pp. 2185–2198, 2020.

[32] Q. Chen, T.-T. Wu, and M. Fang, "Detecting local community structures in complex networks based on local degree central nodes," *Physica A: Statistical Mechanics and Its Applications*, vol. 392, no. 3, pp. 529–537, 2013.

[33] S. Yazdanparast, T. C. Havens, and M. Jamalabdollahi, "Soft overlapping community detection in large-scale networks via fast fuzzy modularity maximization," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 6, pp. 1533–1543, 2020.

[34] X. Wen, W.-N. Chen, Y. Lin et al., "A maximal clique based multiobjective evolutionary algorithm for overlapping community detection," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 3, pp. 363–377, 2016.

[35] E. Osaba, J. Del Ser, D. Camacho, M. N. Bilbao, and X.-S. Yang, "Community detection in networks using bio-inspired optimization: latest developments, new results and perspectives with a selection of recent meta-heuristics," *Applied Soft Computing*, vol. 87, Article ID 106010, 2020.

[36] F. Cheng, C. Wang, X. Zhang, and Y. Yang, "A local-neighborhood information based overlapping community detection algorithm for large-scale complex networks," *IEEE/ACM Transactions on Networking*, vol. 29, no. 2, pp. 543–556, 2020.

[37] Y. Su, K. Zhou, X. Zhang, R. Cheng, and C. Zheng, "A parallel multi-objective evolutionary algorithm for community detection in large-scale complex networks," *Information Sciences*, vol. 576, pp. 374–392, 2021.