IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 26, 2018, accepted October 29, 2018, date of publication November 9, 2018,
date of current version December 18, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2879648

Overlapping Community Detection Based on Information Dynamics

ZEJUN SUN[1,2], BIN WANG[1], JINFANG SHENG[1], ZHONGJING YU[3], AND JUNMING SHAO[3]

[1]School of Information Science and Engineering, Central South University, Changsha 410083, China
[2]Department of Network Center, Pingdingshan University, Pingdingshan 467000, China
[3]Big Data Research Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: Jinfang Sheng (jfsheng@csu.edu.cn)

This work was supported by the National Science and Technology Major Project of China under Grant 2017ZX06002005, Fundamental Research Funds for the Central Universities of Central South University under Grant 2017zzts139, National Natural Science Foundation of China under Grant 61403062, Fok Ying-Tong Education Foundation for Young Teachers in the Higher Education Institutions of China under Grant 161062, and the Science-Technology Foundation for Young Scientist of Sichuan Province under Grant 2016JQ0007.

ABSTRACT Identifying overlapping communities is essential for analyzing network structures, exploring the interactions of groups, studying network functions, and obtaining insight into the dynamics of networks. Many algorithms have been proposed for detecting overlapping communities but identifying the intrinsic communities is still a non-trivial problem because of the difficulties with parameter tuning, user bias criteria, and the lack of ground truth information. In this paper, we propose a new model called OCDID (Overlapping Community Detection based on Information Dynamics) to uncover the overlapping communities, which treats the network as a dynamical system that allows an individual to communicate and share information with its neighbors. The information flow in the network is controlled by the underlying topology structure (e.g., the community structure), and the community structure is also reflected by the information dynamics. Overlapping nodes act as bridges between multiple communities and the information from multiple communities flows through these nodes. Thus, the overlapping nodes can be identified by analyzing the information flow among communities. In addition, we use the monotone convergence theorem to confirm the convergence of our model. Experiments based on synthetic and real-world networks demonstrate that in most cases, our proposed approach is superior to other representative algorithms in terms of the quality of overlapping community detection.

INDEX TERMS Complex network, diffusion, information dynamics, overlapping community detection.

I. INTRODUCTION

Complex networks are powerful methods for representing and studying the interactions among objects in the real world. In recent decades, complex network mining has been an important research area because of its far-reaching effects in various disciplines and domains [1]–[5]. Numerous studies have demonstrated that many real networks possess community structures (groups, clusters, or modules). Detecting community structures is a core problem in the field of network computing because of its importance and practical applications, such as identifying modules, studying the interactions among objects, understanding the dynamic characteristics, and predicting the evolution of a network. In recent years, many methods have been established for disjoint community detection [6]–[8]. Most of these methods divide networks into several independent communities with no nodes shared between them. However, it is known that people can participate in diverse organizations in the real world owing to their many interests; thus, one person may belong to different communities, so most of the communities in real-world networks are overlapping. Therefore, the detection of overlapping communities has attracted increasing attention and extensive discussions. Many algorithms have been designed for identifying the overlapping communities from different perspectives, which can be divided into four categories: clique percolation-based algorithms, seed expansion-based algorithms, link partitioning-based algorithms, and dynamical-based algorithms. Among these algorithms, identifying overlapping communities using dynamic models based on the intrinsic topological structure of networks is an emerging and promising method, and it has been used widely because of its simplicity, efficiency, and

data-driven nature. The basic idea of the dynamical algorithm is to treat a network as a dynamical system before creating a dynamic model to imitate the interaction processes among the nodes. As the interactions evolve, each node will reach a steady state and the community structure will finally be revealed in an intuitive manner. SLPA [9], COPRA [10], and Synchronization of Wu *et al.* [11] are representative dynamical algorithms. However, most of the existing methods have some limitations and deficiencies. For example, the Synchronization of Kuramoto method needs to set the coupling strength parameters K_p and K_n. The COPRA algorithm also needs to set parameters for the maximum number of communities. The performance of these algorithms depends greatly on the choice of parameters. In addition, the propagation-based algorithms comprising SLPA and COPRA are disadvantageous because they produce unstable results and they detect the overlapping communities by label propagation; therefore, they do not truly model the propagation of information in real-world networks. Thus, it is important to develop mathematical models for qualitatively and formally characterizing the essential features of information propagation for overlapping community detection. However, it is difficult to model the spread of information in networks because of the complex interactions between objects as well as the diversity and vast scale of networks. Therefore, the highly accurate detection of overlapping communities is still a major problem.

In this study, we propose a new overlapping community detection algorithm based on information dynamics called OCDID. We treat the network as a closed dynamic social system and each person in the network exchanges information with others because of their common interests. The overlapping community structure is identified in a natural manner based on the information dynamics. This new approach provides a novel method for detecting overlapping communities and it has some attractive features. In the following, we first introduce the basic idea of the proposed method.

A. BASIC IDEA

In the real world, the topological structure of a network affects the spread of information. In addition, the propagation of information in networks has the power to change the network structure. For example, a person might prefer to exchange information with familiar people or people in the same circle (group or community). Similarly, these people may also form a circle because of their common interests and traits. Therefore, the communication of information between people plays an important role in the formation and development of communities. Thus, the communication of information may reflect the overlapping structure of communities. We considered a person who was connected with multiple circles or communities and found that the amount of information flowing through this person was often higher than that through his/her neighbors connected with one community. A person on the boundaries of communities can communicate with different people in distinct clusters, so the information can be spread among different communities. Thus, the dynamic features of information propagation in networks are suitable for characterizing the overlapping structure of communities. According to this assumption, we may be able to automatically identify the overlapping communities by imitating the diffusion of information among people. Thus, we developed a new method based on information dynamics to obtain insights into the partitioning of overlapping communities, where the basic idea is to treat the network as an accommodative dynamic system and study the interactions among information as it diffuses over time. In particular, in a social network, users with similar hobbies and properties are more likely to communicate with each other, and the diffusion of information between them is often more frequent. As the interactions of information spread, the people in a common community have almost the same information, whereas those located in multiple communities may receive more diverse information. Over time, the dynamics of this network reach convergence and the information in the network achieves a steady state. Thus, everyone in a common community has the same amount of information, whereas people in distinct communities have different amounts of information, and users connected with multiple communities may receive more information from different groups. Therefore, the overlapping communities can be identified by calculating the amount of information that each person possesses. The formal definition of the information dynamics model is presented in Section 3. We use a toy network as an example to better illustrate the basic idea. As shown in Fig. 1, a number of users denoted as cartoon people with different colors comprise the artificial social network. In this network, we use a corporation as an example to introduce the information dynamics process, which can be described according to the following stages: First, people possess their own knowledge as initial information because of their different occupations. For example, the initial information of v equals 0.29 (see Definition 3), as shown in Fig. 1(a). The information then diffuses through the topological structure and the interactions to communicate information between users are more frequent in a common department than among those in different departments. For example, user v exchanges information with the connected users $u_1 - u_5$ (see (5)–(9)), as shown in Fig. 1(b). The overlapping people act as bridges between communities and information from multiple communities flows through these people (u_4 is an overlapping person). Over time, the volume of information communicated between people tends to zero and the information possessed by users in a common department will be the same. Finally, the volume of information in the whole network achieves a steady state and the information dynamics in the network reach convergence (see Fig. 1(c)). Next, the communities are divided in a natural manner by calculating the different information in the network, as shown in Fig. 1(d). As a result, the overlapping nodes can be identified by considering the amount of information flowing through the bridge nodes ((10)-(12)), as shown in Fig. 1(e).

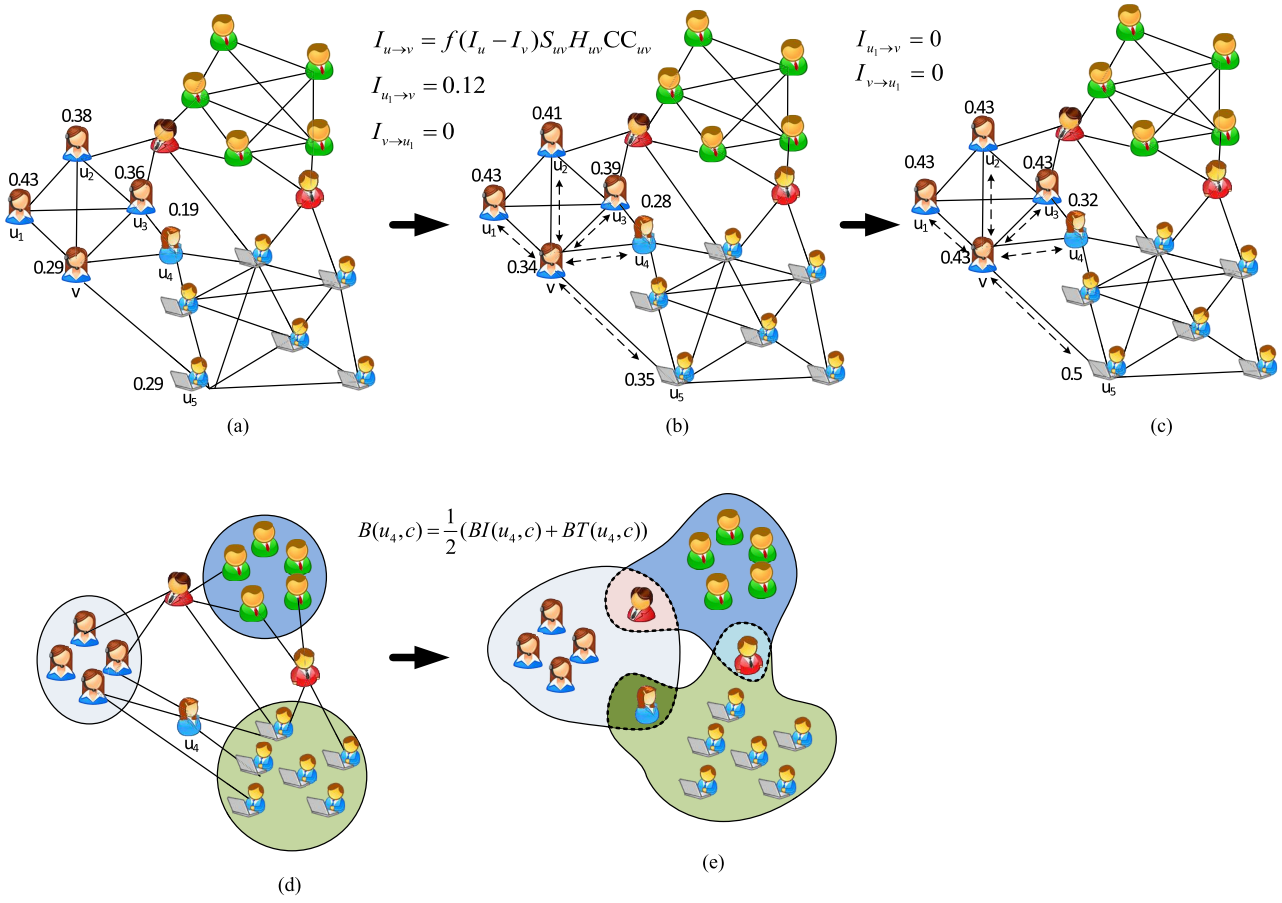


FIGURE 1. Illustration of overlapping community detection based on information dynamics. (a) An artificial social network where the lines denote the relationships between people. (b) The dashed lines with arrows represent the information communicated between people. The information possessed by every person is updated over time based on the proposed information dynamics model. (c) The information in the whole network reaches a steady state. (d) The communities are identified by calculating the different volumes of information in the network. (e) The overlapping communities are identified by computing the information flowing through the bridge nodes.

B. CONTRIBUTIONS

By imitating the information dynamics, the OCDID method exhibits several desirable properties for overlapping community detection in complex networks where the most important are as follows.

- **Intuitive and effective overlapping community detection**: The information dynamics model identifies the overlapping communities by simulating the communication of information in the real world, which accurately represents the flow of information in the network. The overlapping nodes can be identified in a natural manner by calculating the information flowing through the bridge nodes. More importantly, the OCDID reliably identifies the high-quality overlapping community, and still achieves outstanding results at low average degree networks (Fig. 3 –Fig. 4 and Table 4).
- **Parameter-free**: The OCDID method does not rely on prior knowledge and parameter adjustments, and it can automatically identify the overlapping communities based on the information dynamics obtained from the local topology.
- **Scalability**: Owing to the benefits of the proposed information dynamics model, OCDID can detect overlapping nodes based on the amount of information spread among communities, which can also be used to identify non-overlapping communities. Furthermore, OCDID can be applied to large-scale networks owing to the local interaction model and its low time complexity.

The remainder of this paper is organized as follows. In Section 2, we provide a brief survey of related research. In Section 3, we explain the information dynamics model and develop the algorithms in detail. In Section 4, we present evaluations of the performance of OCDID based on synthetic and real-world networks according to several widely used metrics. Finally, we present our conclusions in Section 5.

II. RELATED WORK

In recent decades, many algorithms have been designed for identifying overlapping communities in networks. In the following, we only provide a very brief survey of the algorithms that are relevant to the present study, which can be broadly

classified into four areas. More detailed reviews of overlapping community detection were provided by [12] and [13].

A. CLIQUE PERCOLATION-BASED METHODS

The first clique percolation-based method was CPM [14] proposed by Palla. CPM assumes that a community comprises fully connected subgraphs and it identifies the overlapping communities by searching for adjacent cliques. Therefore, CPM is more suitable for densely connected networks. In the worst case, the time complexity of CPM is exponential because it requires the computation of all the maximal cliques in the network [15], and thus it fails to identify the overlapping communities in many large social networks. Many clique percolation-based methods have been proposed recently, such as CPMw [16], SCP [17], and FCP [18]. CPMw is designed for weighted networks and it detects the overlapping communities based on a subgraph intensity threshold. SCP detects communities with a given size, where it allows multiple weight thresholds and it is faster than CPM. FCP initially obtains the maximal cliques in a similar manner to CFinder, before trying to minimize the number of overlapping clique tests that need to be conducted to obtain the k-clique communities.

B. SEED EXPANSION-BASED METHODS

Seed expansion-based methods start from a small group of nodes or a node and a community can be identified by adding neighbor nodes with a local benefit function. The benefit function characterizes the quality of the structure of the clustering. In general, seed expansion approaches comprise two steps. First, the algorithm detects seed nodes according to certain criteria. Second, the seed nodes are expanded or merged iteratively until a local quality function cannot be improved further. In the last decade, many seed expansion-based algorithms have been proposed such as iterative scan (IS) [19], LFM [20], and seed set expansion (SSE) [21]. IS starts by choosing an edge as a seed and then expands the seeds by adding or removing vertices until a density metric cannot be improved. A disadvantage of this method is that the clustering result is influenced by the choice of a random edge. LFM expands the seeds by maximizing a fitness function and the community size is controlled by the fitness function parameter. The SSE method comprises three phases: filtering, seeding, and expansion of the seed set. A drawback of SSE is that it returns different results in each run.

C. LINK PARTITIONING-BASED METHODS

Link partitioning-based methods identify overlapping community structures by partitioning links instead of nodes. They convert the original network into a *line graph* and then identify the non-overlapping link communities using disjoint community detection methods. Finally, the *line graph* is converted back into the original network, which allows nodes to be present in multiple communities. Recently, many link partitioning-based methods have been proposed such as link clustering (LC) [22], CDAEO [23], and map equation for link communities (MELC) [24]. Link partitioning allows disjoint community detection methods to be employed for identifying overlapping communities, which is a conceptually natural approach for the detection of overlapping. However, a disadvantage of these algorithms is the resolution limit problem [25].

D. DYNAMICAL-BASED METHODS

Overlapping communities can also be identified in a dynamical process and many dynamical algorithms have been proposed based on methods such as synchronization [11], [26], label propagation [9], [10], [27], spin dynamics [28], [29], and random walk [30]. For example, the synchronization dynamical-based algorithm [11] considers a node as a phase oscillator and the nodes evolve according to the designed differential equations. As the interactions among the phase oscillators proceed, the neighborhoods with common properties reach the same phase. The network converges with time and the nodes with the same phase can be partitioned into the same community. Finally, the network is divided into several communities based on their phases, where the phases of overlapping nodes are between two or more community phases. Label propagation algorithms detect overlapping communities based on the dynamics of label propagation. In a network, every vertex shares its label with neighbors. Finally, vertices with the same label are assigned to the same community. The overlapping vertices have multiple labels, and thus they are divided into several communities. COPRA [10] is a well-known label propagation-based algorithm that identifies communities based on their belonging coefficients. However, a parameter needs to be set to limit the number of communities in which a node can participate. SLPA [9] is a fast label propagation algorithm that detects overlapping communities based on speaker–listener patterns. An advantage of SLPA is that the time complexity is $O(tm)$, which is linear with the edges m, so it can handle large-scale networks. However, a disadvantage of SLPA is that the community division result is unstable.

In this study, we introduce a new dynamical-based method called OCDID for identifying the overlapping community structure. Similar to label propagation methods, this method also employs the idea of propagation and a node interacts with its immediate neighbors. However, our approach identifies overlapping communities from the perspective of information theory, which is different from label propagation methods. In addition, our method uses the amount of information spread in a network to represent the flow of information in the network, whereas label propagation algorithms propagate a label.

III. OVERLAPPING COMMUNITY DETECTION BASED ON INFORMATION DYNAMICS

In this section, we first survey preliminary concepts regarding overlapping community detection and information dynamics, before introducing the information dynamics model.

Finally, we develop the information dynamics algorithm and analyze its time complexity.

A. PRELIMINARIES

Before explaining our proposed algorithm, we formalize some basic definitions that are used in the following sections. Let $G = (V, E)$ be an undirected unweighted network with $|V|$ nodes and $|E|$ edges. The goal of overlapping community detection in network G is to find a division $D = \{C_1, C_2, \ldots, C_k\}$ where one node can belong to more than one community ($\{\exists i, j \in k, C_i \bigcap C_j \neq \emptyset, i \neq j\}$). All of the key symbols are described in Table 1.

TABLE 1. Summary of the symbols used in this study.

Symbols	Definitions
n	number of vertices in network G ($n = \|V\|$)
m	number of edges of network G ($m = \|E\|$)
D_v	degree of vertices v
D_{max}	max degree of network G
Avg_D_v	average degree of the neighbors of vertex v
$N(v)$	neighborhood set of vertex v
$Avg\_S(v)$	average similarity of the neighbors of vertex v
T_v	number of triangles for vertex v
CC_v	clustering coefficient for vertex v
CS_{vu}	contact strength of vertex u on vertex v
I_v	information about vertex v

Definition 1 (Jaccard Similarity): Given an undirected and unweighted network G = (V, E), $\Gamma(v)$ denotes the set of neighborhoods of vertex v that includes vertex v and its neighbors. The Jaccard similarity coefficient for vertices u and v is defined as follows.

$$JS_{uv} = \frac{|\ \Gamma(u) \bigcap \Gamma(v)\ |}{|\ \Gamma(u) \bigcup \Gamma(v)\ |} \tag{1}$$

In the real world, research has shown that social networks often contain strong and weak ties, which play very important roles in information propagation and community formation. We use triangles to describe the contact strength in order to clearly characterize the relationships of vertices in networks because a triangular structure can better reflect the degree of connection between vertices.

Definition 2 (Connection Strength): Given an undirected and unweighted network $G = (V, E)$, the contact strength of vertex u on v is defined as

$$CS_{vu} = \frac{|N(v) \bigcap N(u)|}{T_v} \tag{2}$$

where $N(v)$ is the number of neighbors of node v, the junction of $N(v)$ and $N(u)$ denotes the amount of triangles shared by nodes v and u, and T_v represents the quantity of triangles for node v.

In an interpersonal network, a person can obtain more information when he/she has more friends, and thus more resources. Similarly, the clustering coefficient of a node reflects the degree of aggregation with its organizations. Thus, we use the degree of a node and the clustering coefficient as its initial information.

Definition 3 (Information): Given an undirected and unweighted network $G = (V, E)$, the initial information for vertex v is defined as

$$I_v = \frac{D_v \cdot CC_v}{D_{max}} \tag{3}$$

where D_v denotes the degree of node v, D_{max} represents the maximum degree of G, and CC_v is the clustering coefficient of node v. Let I_{max} be the maximum volume of information for the nodes in a given network G. We can obtain the largest value ($I_{max} = 1$) when $D_v = D_{max}$, $CC_v = 1$.

Definition 4 (Boundary Nodes): Given an undirected and unweighted network $G = (V, E)$, let c be a community of G and the boundary node set of community c is defined as

$$BN_c = \{v \in c | \Gamma(v) \cap c \neq \emptyset, \Gamma(v) \nsubseteq c\} \tag{4}$$

where $\Gamma(v)$ is the set of neighbors of node v, including v. A boundary node is a node v that belongs to a community c and the neighbors of v do not belong to this community completely.

B. INFORMATION DYNAMICS MODEL

After defining the key notations, we now construct the information dynamics model, which has three parts: the interaction range, dynamics model, and overlapping community detection.

1) INTERACTION RANGE

The topology of the network plays an important role as the medium, which can affect the dynamics of information diffusion. Instead of observing the global interaction of the topology, we focus on the information dynamics in a local manner. In particular, every vertex interacts with its directly connected vertices. Our algorithm can handle large-scale networks because it considers local interactions.

2) DYNAMICS MODEL

After defining the interaction range for information propagation, the next critical step involves studying the diffusion model among the nodes in order to characterize the information dynamics. How should we describe the spread of information? Considering the patterns of information diffusion among people in a social network, we can see that everyone can obtain information from their neighbors as well as propagating information to them. People on the borders between organizations play coordinating roles, where they can guide relationships and the exchange of information among different organizations. Thus, the propagation of information is greatly affected by the local topology, such as the node's degree, clustering coefficient, similarity, and connection strengths. Moreover, the cost of information diffusion should be considered. Thus, in the following, we formulate the dynamics model in terms of the initial information, propagation volume, information loss, and information propagation.

1) **Initial Information.** How should we depict the initial information for one node? The degree of a node is an important indicator for characterizing its initial information. However, our experiments showed that only using the node degree as the initial information for finding overlapping communities is not ideal because many nodes have the same degree in a network, which may affect the diffusion of information. Another reason is that overlapping nodes often have larger degrees because they connect several communities. Thus, more information is transmitted so these communities may have the same amount of information and multiple communities could be assigned to one community. To solve these problems, we use the degree and clustering coefficient to represent the information for nodes ((3)). The degree of a node reflects how much the node owns a resource and the clustering coefficient denotes the closeness of one node to other nodes.
2) **Propagation Volume.** In the real world, the exchange of information is readily influenced by the surrounding environment. For example, people are more likely to choose to communicate with people with whom they share closer links and similar interests. In order to characterize the diffusion of information in a more realistic manner, we use the information difference, node similarity, connection strength, and clustering coefficient to model the amount of information propagated. Formally, we let $I_{u\to v}$ denote the information that one node u propagates to node v, which is defined as follows:

$$I_{u\to v} = f(I_u - I_v) \cdot JS_{uv} \cdot H_{uv} \cdot CC_{uv} \tag{5}$$

where $f(I_u - I_v)$ represents the information that can be propagated from node u to node v. In particular, the coupling function $f(\cdot)$ is given by

$$f(I_u - I_v) = \begin{cases} e^{(I_u - I_v)} - 1 & I_u - I_v \geq 0 \\ 0 & I_u - I_v < 0. \end{cases} \tag{6}$$

As shown above, nodes with a large amount of information are more likely to spread and influence nodes with a small amount of information. In Equation (5), JS_{uv} denotes the Jaccard similarity coefficient for node u and node v, H_{uv} represents the contact strength between nodes u and v, and CC_{uv} is the closeness of nodes u and v with their neighbors in the local topology, which is defined as

$$CC_{uv} = \frac{1}{1 + e^{-5 \cdot CC_v \cdot CC_u}} - 0.5. \tag{7}$$

3) **Information Loss.** The loss of information may occur during the information propagation process in the real world. For example, if the information disseminated is familiar or attractive to us, we can understand and spread it more easily. By contrast, owing to environmental factors, people may misunderstand, ignore, or even lose information. In order to describe the loss of information in a more realistic and accurate manner, we employ the topological features and information volume for its characterization. Formally, we define the loss of information as follows:

$$I_{(u\to v)\_cost} = \frac{Avg\_S(v)}{Avg\_D(v)} \cdot f(I_u - I_v) \cdot (1 - JS_{uv}) \tag{8}$$

where $Avg\_S(v)$ and $Avg\_D(v)$ are local topological features representing the local average similarity and local average degree, respectively. Clearly, $I_{(u\to v)\_cost}$ is positively related to $f(I_u - I_v)$ and negatively related to JS_{uv}. Thus, the information loss is greater when the amount of information propagated is higher, and the information loss is smaller when the communicating objects are more similar.
4) **Information Propagation.** To represent the information diffusion process in a network, we use iterative methods to simulate the communication of information between the nodes. Every node obtains information from its neighbors in each iteration. Based on the combined diffusion patterns, the model of the information dynamics over time is given by

$$I_v(t+1) = I_v(t) + \sum_{u \in N(v)} (I_{u\to v}(t) - I_{(u\to v)\_cost}(t)) \tag{9}$$

where $I_v(t)$ denotes the information for node v at time step t. Initially, $t = 0$ and $I_v(0)$ is the initial information for node v ((3)). We can see that the information for node v at time step $t + 1$ comprises two parts: the information at time t and the information obtained from its neighbors. In the real world, the information that we may receive cannot be negative. Therefore, $(I_{u\to v}(t) - I_{(u\to v)\_cost}(t)) \geq 0$, i.e., the information volume of every node will not decrease in each iteration.

3) OVERLAPPING COMMUNITY DETECTION

Based on the proposed information dynamics model, we can divide communities by computing different information values for the nodes in the network. The overlapping nodes are detected as follows. The overlapping nodes are connected to several communities and they can penetrate the information into the linked communities, so we can identify the overlapping nodes by calculating the information volume communicated with multiple communities. Let $B(v, c)$ denote the belonging degree of a node v that belongs to community c, which comprises two parts: the attribution coefficient based on information flow and the belonging coefficient obtained from the local topology. Formally, the definition of $B(v, c)$ is given as follows:

$$B(v, c) = \frac{1}{2}(BI(v, c) + BT(v, c)) \tag{10}$$

where

$$BI(v, c) = \frac{\sum_{u \in (N(v) \cap c)} I_{sum(u\leftrightarrow v)}}{\sum_{u \in N(v)} I_{sum(u\leftrightarrow v)}} \tag{11}$$

represents the belonging coefficient of nodes derived from the information flow in different communities, and where

$$I_{sum(u\leftrightarrow v)} = \sum_{t\in L}(I_{(u\leftrightarrow v)}(t) - I_{(u\leftrightarrow v)\_cost}(t)) \tag{12}$$

is the sum of the information volume propagated between nodes u and v. $I_{(u\leftrightarrow v)}(t)$ is the information propagated between nodes u and v at time step t, and $I_{(u\leftrightarrow v)\_cost}(t)$ indicates the cost information at time step t. $BT(v, c)$ describes the belonging coefficient of nodes based on the local topological structure, which is defined as

$$BT(v, c) = \frac{|N(v) \cap c|}{D_v}. \tag{13}$$

C. CONVERGENCE ANALYSIS

Based on the proposed information dynamics model, the information for each node in the network will reach a steady state over time. In this section, we prove the convergence of the information dynamics model. Before deriving our proof, we first present some related symbols. According to Equation (3), $I_u|_{t=i}$ denotes the information for node u at time step $t = i$ and the sequence $\{I_u|_{t=i}, i = 0, 1, 2\ldots\}$ represents the information for node u from $t = 0$ to $t = i$.

Theorem 1: Note that the maximum amount of information for one node in network G is I_{max} ($I_{max} = 1$). For each node u, the information sequence comprising $\{I_u|_{t=i}, i = 0, 1, 2\ldots\}$ is convergent.

Proof: According to Definition 3 and by using (5),(6), we can obtain:

$$I_u|_{t=i} \leq I_{max}. \tag{14}$$

Equation (14) denotes that the sequence $\{I_u|_{t=i}, i = 0, 1, 2\ldots\}$ is bounded above. Based on the information dynamics (9), for any $t = i$:

$$I_u|_{t=i} \leq I_u|_{t=i+1}. \tag{15}$$

Formula (15) shows that the sequence $\{I_u|_{t=i}, i = 0, 1, 2\ldots\}$ is non-decreasing as i increases. According to the monotone convergence theorem [31], for one vertex u, the sequence $\{I_u|_{t=i}, i = 0, 1, 2\ldots n\}$ is convergent. ■

D. OCDID ALGORITHM

In this section, we present the OCDID algorithm for identifying overlapping communities, which involves the following steps.

1) **Simulation of Information Dynamics**. Based on the proposed information dynamics models, we can simulate the information spread process. The information dynamics mainly comprise the following several steps. Initially, there is no interaction between the nodes and every node is assigned initial information according to the local topology features ((3)). Then, the information propagates in the network and the exchange of information between nodes depends on the information dynamics models ((5)–(8)). In order to find the overlapping nodes, we need to store the amount of information spread between each of the nodes ((12)). Next, we iteratively update the information volume for each node using (9)). Finally, when the amount of information propagating between each pair of nodes in the network is less than the threshold, the information of the entire network reaches a steady state. As shown in the Algorithm 1, the threshold is equal to 0.001. Because when the threshold is less than or equal to 0.001, the information dynamics in the network will reach a convergence state, and the community structure in the network is well divided. Although a higher community partitioning accuracy can be achieved when the threshold is less than 0.001, more computational time is required. Therefore, considering the performance

Algorithm 1 OCDID

Input:
Graph:$G = (V, E)$
//Initialization of information.
for each node v in V **do**
  for each node u in $N(v)$ **do**
    compute JS_{uv} using (1)
    compute CS_{uv} using (2)
  end for
  compute CC_v,$Avg\_S(v)$ and $Avg\_d(v)$
  compute the initial information I_v using (3)
end for
//Information dynamics interactions.
$Flag = TRUE$
$Threshold = 0.001$
while $Flag$ **do**
  $I_{max} = 0$
  for each node v in V **do**
    for each node u in $N(v)$ **do**
      compute the propagation volume using (5)
    end for
    // Update information over time.
    compute $I_v(t + 1)$ using (9)
    $I_{in} = I_{u\to v} - I_{(u\to v)\_cost}$
    if $I_{in} > I_{max}$ **then**
      $I_{max} = I_{in}$
    end if
  end for
  // The network reaches a balanced state.
  if $I_{max} < Threshold$ **then**
    $Flag = FALSE$
  end if
end while
// Partition communities.
$C \leftarrow$Community_detection(G, I)
// Detect overlapping communities.
$OC \leftarrow$ Ov_comm_detection(G, C)
// Return overlapping communities OC.
Output: OC

and calculation cost, we set the threshold equal to 0.001.

2) **Community Partitioning**. All of the nodes converge as the information in the network reaches equilibrium. According to the information dynamics model, the amount of information on the nodes in the same group is equal. By contrast, the information volumes on nodes in distinct groups are unequal. Therefore, we can identify the communities in a natural manner by statistically analyzing the different information volumes for the nodes in the network. As shown in Algorithm 2, when the amount of information between two neighboring nodes is less than the threshold, they are divided into the same community.

Algorithm 2 Community_detection

Input:
Graph: $G = (V, E)$, Information list: I

// Community partition.
$Threshold = 0.001$
for each node v in V **do**
  if v not in C(communities) **then**
    for each node u in $N(v)$ **do**
      if $|I_v - I_u| < Threshold$ **then**
        if u in C **then**
          if v in C **then**
            $C_u -> C_v$
          else
            $v -> C_u$
          end if
        else
          if v in C **then**
            $u -> C_v$
          else
            $u, v -> C_u$
          end if
        end if
      end if
    end for
  end if
end for
// Communities C.

Output: C

3) **Overlapping Community detection**. After detecting the communities in the network, we can identify the overlapping nodes by studying the information exchange between a node and its neighbors. Overlapping nodes have different information compared with other nodes because of the information exchange between overlapping nodes and multiple communities. Therefore, we can identify the overlapping nodes in the following steps. First, we calculate the information volumes for boundary nodes using (12). Next, we compute the belonging degree for each boundary node ((10)). Finally, the overlapping nodes are divided into multiple communities according to the belonging degree of each node. The method for detecting overlapping nodes is given in Algorithms 1–3.

Algorithm 3 Ov_comm_detection

Input:
Graph: $G = (V, E)$, Communities: C

// Identify the overlapping nodes.
$Threshold = 0.2$
for each node v in V **do**
  if v in BN_c **then**
    // NC is the community set to which node v neighbors belong.
    for c' in NC **do**
      compute the belonging degree $B(v, c')$ using (10)
      if $B(v, c') > Threshold$ and v not in c' **then**
        $c' \leftarrow v$
      end if
    end for
  end if
end for
// Overlapping communities C.

Output: C

Fig. 2(a)–(f) illustrate the overlapping detection process based on the information dynamics in a social network. Every node begins with the initial information obtained from its local structural characteristics. For example, the initial values for nodes 1, 2, 3, and 4 are 0.43, 0.29, 0.36, and 0.38, respectively, which can be calculated using formula (3) (Fig. 2(a)). Next, each node interacts with its neighbors and the nodes within the boundaries can communicate with nodes in different communities. Fig. 2(b)–(d) show the change in the amount of information for each node as the interaction process iterates. We can see that the amount of information on the nodes in the same community tends to be equal. For example, the information volumes for nodes 1, 2, 3, and 4 are 0.43, 0.42, 0.43, and 0.43, respectively (Fig. 2(d)). Each node achieves a steady state over time and the nodes in the same community have the same volume of information (Fig. 2(e)). The communities can be partitioned using the information dynamics model (Algorithm 2). Overlapping nodes are located between multiple communities, so they may obtain more information from different communities, and thus the amount of information on overlapping nodes is different from that in each community. For example, overlapping node 5 can obtain information from nodes 2, 3, 13, and 14, which belong to two different communities. The information volumes for nodes 5, 6, and 12 are 0.32, 0.14, and 0.31, respectively. Owing to the low clustering coefficient of node 6, its information has not changed but it can still be identified using formula (13). Finally, the overlapping nodes can be identified in a natural manner by calculating

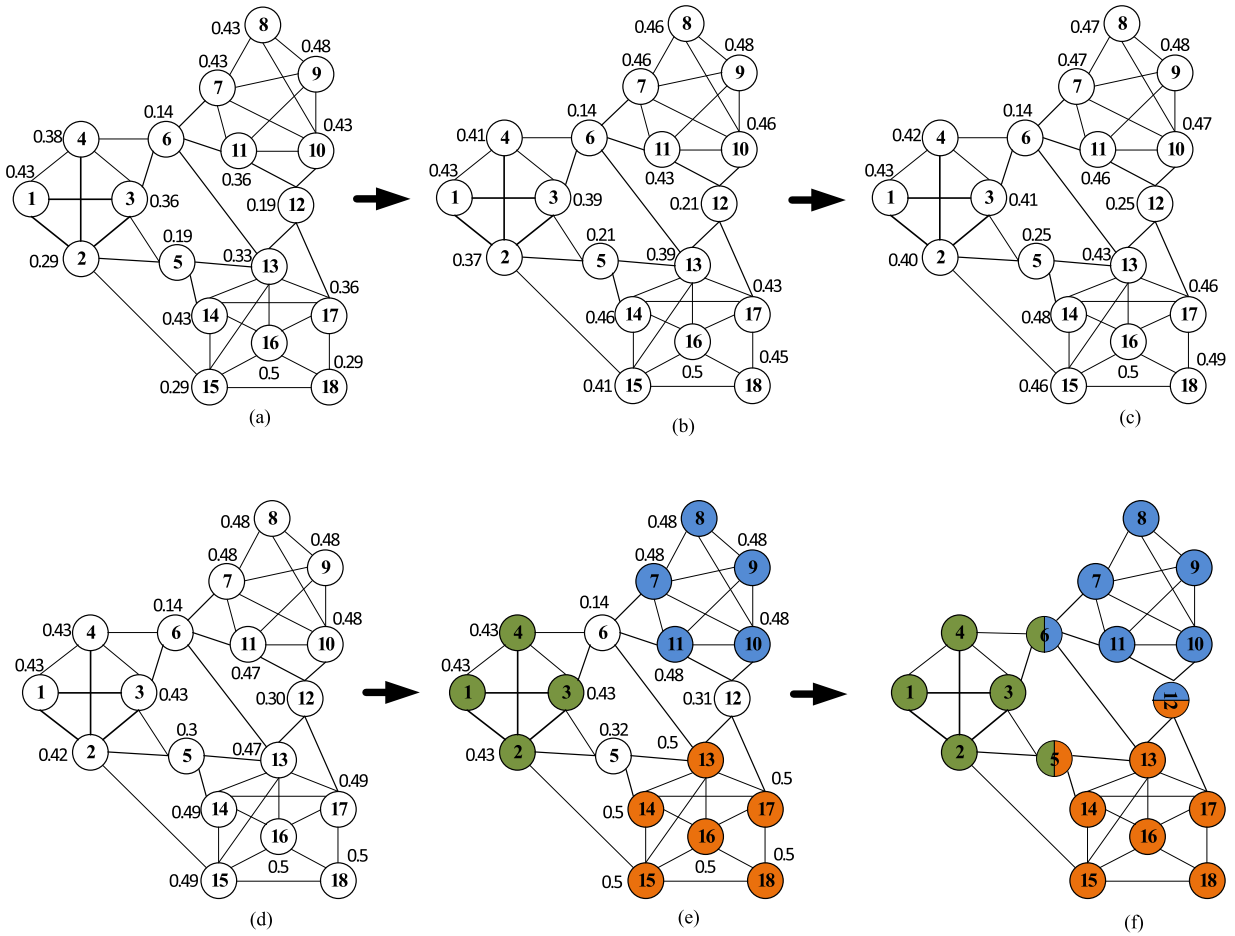


FIGURE 2. Illustration of the information dynamics.

the information volumes for the nodes in the network (formula (10)). Fig. 2(f) shows the detection of three overlapping communities.

E. COMPLEXITY ANALYSIS

The computational complexity of OCDID has three main components. In the first step, the initial information for every node is required. Moreover, for the information dynamics interaction, OCDID also needs to calculate the clustering coefficient, Jaccard similarity coefficient, and contact strength. Thus, the time complexity for initialization is $O(k \cdot n)$, where k is the average degree of the network. In the second step comprising the information dynamics process, the computational complexity is $O(L \cdot n \cdot k)$ because of the local interaction range, where L is the total number of iterations, which typically ranges between 20 and 100. In the third step, the overlapping communities are detected and the computational complexity is due to community partitioning and detecting overlapping communities. The complexity of community partitioning is $O(k \cdot n)$. The worst complexity for overlapping community identification is $O(|C| \cdot n)$, where $|C|$ is the number of communities in the network G. Hence, the computational complexity of OCDID is $O(k \cdot n + L \cdot n \cdot k + |C| \cdot n)$. We note that $k \ll n$ and $|C| \ll n$, and thus the OCDID algorithm can handle large-scale networks.

IV. EXPERIMENTS

In this section, we present the results of various experiments conducted using synthetic and real-world networks in order to demonstrate the performance of OCDID based on comparisons with several representative overlapping community detection methods. Before presenting the experimental results, we briefly introduce the algorithms used in the comparisons.

SLPA [9] is a fast overlapping community detection approach for large networks based on label propagation. SLPA spreads labels among the nodes depending on interaction rules, and it provides every node with a memory to store the received labels information.

FCP [18] is a fast clique percolation algorithm that establishes a minimal spanning forest based on the maximal cliques to reduce the need for unnecessary clique tests.

LC [22] is a well-known link clustering method that divides communities by hierarchical clustering of the

link similarity. The time complexity of LC is $O(nd_{max}^2)$, where d_{max} is the maximum degree of the nodes in the network.

COPRA [10] is a well-known overlapping community algorithm based on the label propagation algorithm [32]. First, it assigns a label to each node, where all of the belonging coefficients are set to 1. Next, every node updates its labels by computing the belonging coefficients repeatedly for its neighbors. Finally, the overlapping communities are identified by considering the label for each node.

DEMON [27] is a democratic voting approach where each node judges the community to which each of its neighbors should belong. DEMON is also based on the label propagation algorithm.

LFM [20] is a seed expansion method for detecting overlapping and hierarchical community structures. LFM starts with a random seed node and assigns a node to a community based on its fitness.

SSE [21] uses the PageRank scheme to optimize the conductance score for the community.

NECTAR [33] is an extension of the Louvain method [34] that employs greedy local search heuristic to maximize the modularity objective function.

These methods can be divided into several categories: clique percolation-based algorithms (FCP), seed expansion-based algorithms (LFM and SSE), link partitioning-based algorithms (LC), dynamical-based algorithms (SLPA, COPRA, and DEMON), and other algorithms (NECTAR). Propagation-based algorithms are used widely for overlapping community identification because of their low time complexity. However, most of the algorithms based on label propagation produce unstable results.

A. DATA DESCRIPTION

1) SYNTHETIC NETWORKS

Many complex networks exist in the real world but we rarely know the ground truth details for the overlapping communities. Therefore, we built synthetic networks with a known ground truth community structure in order to evaluate the algorithms used in the comparisons. We used the LFR benchmark to generate networks that are very similar to real-world networks [35]. This benchmark has been used widely for disjoint community and overlapping community detection. The LFR benchmark can be readily controlled to generate networks using several parameters, including the community size, average clustering coefficient, average degree, and overlap. These parameters are summarized in Table 2.

TABLE 2. Summary of the parameters for the LFR benchmarks.

Symbol	Definition
n	number of nodes ($n = \|V\|$)
μ	mixing parameter
k	average degree of nodes
k_max	max degree of the generating network
τ_1	minus exponent for the power law distributions of degree
τ_2	minus exponent for the community size distribution
C_{min}	minimum size of the community for the generating network
C_{max}	maximum size of the community for the generating network
O_n	number of overlapping nodes
O_m	number of memberships for the overlapping nodes
C	average clustering coefficient

TABLE 3. Some properties of the real-world data sets where *k* is the average degree and CC is the clustering coefficient.

Dataset	$\|V\|$	$\|E\|$	k	CC
Karate	34	78	4.588	0.571
Football	115	613	10.661	0.403
Polbooks	105	441	8.4	0.488
Jazz	198	2742	27.697	0.633
Citeseer	2,110	4,732	3.523	0.24
protein	1,870	2,277	2.435	0.171
Power grid	4,941	6,594	2.669	0.107
Amazon	334,863	925,872	5.53	0.397

2) REAL-WORLD DATA SETS

We also used several popular real-world networks with different sizes and characteristics to assess the performance of each of the algorithms used in the comparisons. The statistics for each real-world network are summarized in Table 3. All of these data sets are publicly available as network data from Newman (http://www-personal.umich.edu/m̃ejn/netdata), KONECT (http://konect.uni-koblenz.de/networks), and Stanford (http://snap.stanford.edu/data/). Next, we briefly introduce some of the real-world networks.

a) ZACHARY'S KARATE NETWORK

This is a popular network and it has been used widely for complex network mining. This network comprises a social network of friendships among 34 members of Zachary's karate club in an American university. This club was divided into two groups over time due to differences in opinions regarding leadership.

b) FOOTBALL NETWORK

This is a well-known network derived from U.S. college football games among Division IA colleges. The network includes 115 nodes, 631 edges, and 12 conferences (communities). Each node represents a team and each edge denotes a regular season game between two teams.

c) POLITICS BOOKS NETWORK

This network contains the books about politics in the United States that are sold on the Amazon.com. The network comprises 105 nodes denoting books and 441 edge representing the frequent co-purchasing of books by the same buyers.

d) CITESEER NETWORKS

This is a scientific paper citation network comprising papers with six classifications. The network comprises 2110 nodes denoting papers and 4732 edges representing the citations of papers.

e) JAZZ NETWORK

This is a collaboration network of Jazz musicians containing 198 nodes and 2472 edges. In the network, every Jazz musician denotes a node and the relationship between two musicians is represented as an edge.

f) PROTEIN NETWORK

This is a protein interactions network in yeast, which comprises 1870 nodes and 2277 edges. Each node is a protein and each edge denotes a metabolic interaction between two proteins. Studies have shown that a protein with a higher degree is more important for yeast survival than others.

g) POWER GRID NETWORK

This is a power grid network in the Western States of the USA, which contains 4941 nodes and 6594 edges. In the network, a node either denotes a generator, a transformer, or a substation, and an edge is a power supply line.

h) AMAZON NETWORK

This is an e-commerce network derived from merchandise sales data for Amazon, which comprises 334,863 nodes and 925,872 edges. Each node denotes a product and edges represent the correlations between co-purchased products.

B. EVALUATION METRICS

In the last 10 years, many evaluation criteria have been proposed for quantifying the goodness of the overlapping communities detected by different algorithms. We selected several widely used methods for evaluation, including the extended modularity (EQ) [36], extended normalized mutual information (ENMI) [20], F-score, and purity [37], [38]. The ENMI, F-score, and purity metrics require the ground-truth information for networks, but this is difficult to obtain for real-world networks. Therefore, these evaluation criteria are often used with synthetic networks. Before presenting the evaluation results, we briefly describe the evaluation indicators.

The concept of modularity was proposed by Newman and Girvan [39] as a metric for evaluating the quality of a partition. This is still one of the most widely used measures for quantifying the partitioning of a network. The modularity is defined as follows:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{d_i d_j}{2m})\delta(C_i, C_j) \tag{16}$$

where m is the number of edges in network G, A denotes the adjacency matrix of G, d_i represents the degree of vertex i, and δ is the Kronecker function, the value of which equals 1 if node i and node j belong to the same community, whereas it is 0 otherwise. To evaluate the quality of a partition based on an overlapping community, Shen *et al.* [36] proposed the EQ, which is often used when the ground truth is not known for the network. EQ is defined as follows:

$$EQ = \frac{1}{2m}\sum_{l=1}^{s}\sum_{i\in C_l, j\in C_l}\frac{1}{O_i O_j}(A_{ij} - \frac{d_i d_j}{2m}) \tag{17}$$

where O_i is the number of communities to which vertex i belongs and C_l denotes a community ($1 \le l \le s$ and s is the number of communities). The detection of the overlapping communities will be better when the value of EQ is larger. However, a disadvantage of the EQ evaluation metric is that it is not suitable for evaluating large-scale networks because of the high time complexity.

The normalized mutual information is widely used to evaluate the quality of disjoint community detection, where it originated from information theory. It is considered that if the two divisions are similar, then only a small amount of additional information is needed to infer one partition from the other. To evaluate the quality of overlapping community detection, Lancichinetti *et al.* [20] extended the normalized mutual information metric to obtain the ENMI, which is defined as follows:

$$ENMI(X|Y) = 1 - [H(X|Y) + H(Y|X)]/2 \tag{18}$$

$$H(X|Y) = 1 - \frac{1}{|C'|}\sum_{k}\frac{H(X_k|Y)}{H(X_k)} \tag{19}$$

where X and Y are random variables related to partitions C and C', respectively, and $H(X|Y)$ denotes the normalized conditional entropy for cluster X with respect to cluster Y. The ENMI value ranges between 0 and 1, were 0 denotes that the partition is completely independent of the ground truth, whereas 1 indicates a perfect match with the real partition.

The F-score measures the accuracy of overlapping community partitioning with respect to the ground truth information. The F-score is defined as:

$$F - score = \frac{2 * Accuracy * Recall}{Accuracy + Recall} \tag{20}$$

where *Accuracy* is the proportion of correctly detected overlapping nodes among all of the identified overlapping nodes and *Recall* denotes the proportion of correctly detected overlapping nodes among the total true overlapping nodes. The *Accuracy* and *Recall* are defined as

$$Accuracy = \frac{|GOC \bigcap DOC|}{|DOC|} \tag{21}$$

$$Recall = \frac{|GOC \bigcap DOC|}{|GOC|} \tag{22}$$

where *GOC* is the ground truth for the overlapping communities and *DOC* represents the detected overlapping communities. The F-score value ranges from 0 to 1 where a larger value denotes better quality.

The purity is an external evaluation metric for evaluating the quality of community detection methods and it is defined as

$$P = \frac{1}{N}\sum_{i=1}^{N} max_{1\le j\le k}\frac{|C_{ij}|}{|C_i|} \tag{23}$$

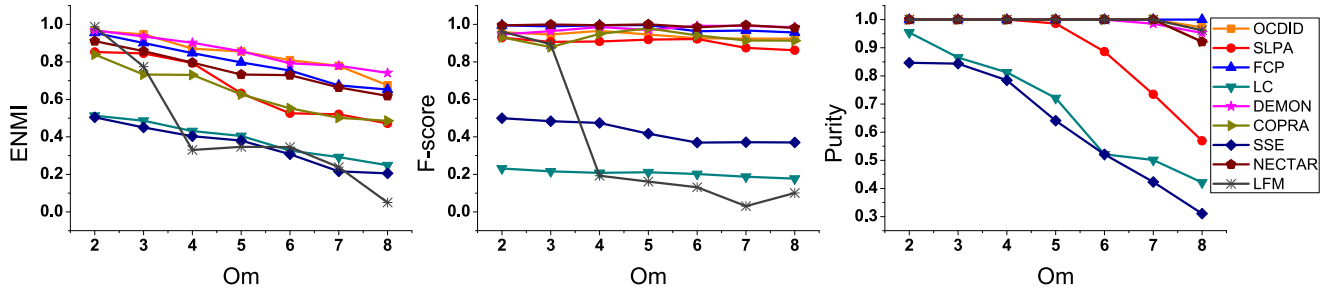


FIGURE 3. Effectiveness of different algorithms with LFR networks when the degree of overlap O_m ranged from 2 to 8.

where N denotes the number of detected communities, k is the number of different labels, $|C_i|$ is the number of nodes in community i, and $|C_{ij}|$ is the number of nodes belonging to label j and community i. A higher purity index denotes better community partitioning.

C. PERFORMANCE EVALUATION

In these experiments, the codes used for SLPA, FCP, LC, COPRA, DEMON, and LFM were obtained from GitHub (https://github.com/GraphProcessor/Community-DetectionCodes), the code for the SSE algorithm was also acquired from GitHub (https://github.com/grey-foxamine/Overlapping-Co-mmunity-Detection-Using-Seed-SetExpansion-), and the code for NECTAR was provided by the author (https://github.com/amirubin87/NECTAR). We employed the default values for all of the algorithms with tunable parameters. The average result was obtained for each method based on experiments with 20 independent runs. All of the experiments were conducted using a desktop computer with an Intel Core i5 3.3-GHz CPU and 16 GB RAM.

1) EVALUATIONS BASED ON SYNTHETIC NETWORKS

To compare the performance of the overlapping community detection algorithms, we generated several synthetic networks using the LFR benchmark with different characteristics. In particular, we set power law distributions for the node degree $\tau_1 = 2$ and the distributions of the community sizes $\tau_2 = 1$, the community size as $C_{min} = [5, 10]$ and $C_{max} = [50, 100]$, and the network size as $n = 1000$. To conduct comprehensive comparisons of the algorithms, we varied the parameters O_m, O_n, μ, and k according to the characteristics of the networks. The ground truth was known for the synthetic networks, so the ENMI, F-score, and purity metrics were used to evaluate the effects. We compared the performance of each method with different parameter settings.

1) **Effect of O_m (degree of overlap)**
In order to test the sensitivity of the different algorithms to the parameter O_m, we varied the number of memberships O_m from 2 to 8, and fixed the parameters $\mu = 0.1$, $O_n/n = 0.1$, and $k = 10$. Fig. 3 shows how the performance of each algorithm varied with the seven LFR networks according to different values of the parameter O_m. We found that as O_m increased, the performance of each method decreased. In terms of the ENMI metric, DEMON and OCDID obtained the best effects. FCP and NECTAR also achieved better results than the other algorithms. When the value of O_m increased to 8, these four algorithms still obtained good ENMI values. SLPA and COPRA also achieved acceptable results. LFM performed well when the value of O_m was low, but its effectiveness decreased significantly as O_m increased. In particular, when the value of O_m was larger than 4, the ENMI value was less than 0.4. LC and SSE did not obtain performance comparable to the other algorithms for some LFR networks. In terms of the F-score metric, OCDID, SLPA, FCP, DEMON, COPRA, and NECTAR performed very well, thereby demonstrating that these algorithms could achieve high accuracy at detecting overlapping communities. However, LC, SSE, and LFM did not obtain ideal performance. In terms of the purity metric, OCDID, FCP, DEMON, COPRA, NECTAR, and LFM achieved very high scores. However, LC and SSE obtained lower purity values as O_m increased, thereby indicating that there were many errors in the communities identified, and thus many nodes were misclassified.

2) **Effect of O_n (number of overlapping nodes)**
In order to evaluate the effectiveness of each algorithm with various values for the number of overlapping nodes O_n, we fixed the parameters $\mu = 0.1$, $O_m = 2$, and $k = 10$, and varied the overlapping density O_n/n from 0.1 to 0.8. Fig. 4 shows the performance of the different methods in terms of the ENMI, F-score, and purity metrics. The performance of each method decreased gradually as O_n increased. In terms of the ENMI metric, DEMON achieved the best performance, while OCDID, FCP, and NECTAR also obtained comparatively good results and they performed better than the other algorithms. When the number of overlapping nodes did not exceed 0.2, the LFM algorithm obtained the highest ENMI value. However, when the value of O_n/n was larger than 0.4, its effectiveness decreased significantly. LC and SSE were not ideal for some of the LFR networks. In terms of the F-score metric, FCP and NECTAR yielded the best performance, where they remained effective as the value of O_n increased.

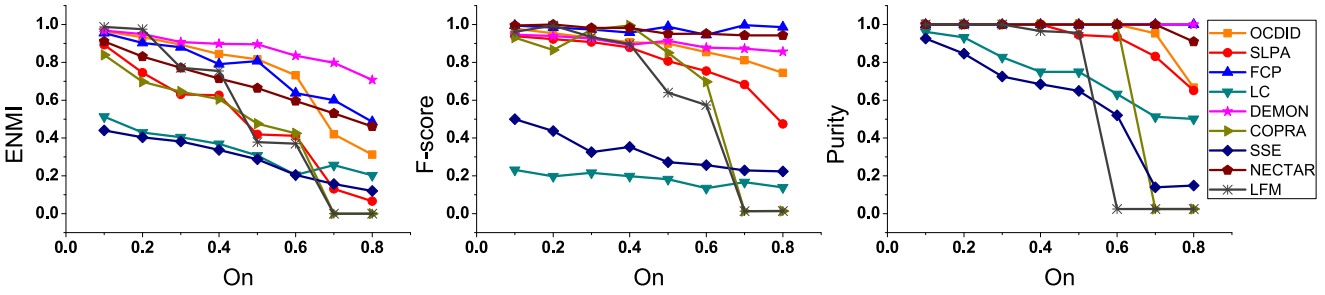


FIGURE 4. Effectiveness of different algorithms with the LFR networks when the overlapping density O_n/n varied from 0.1 to 0.8.

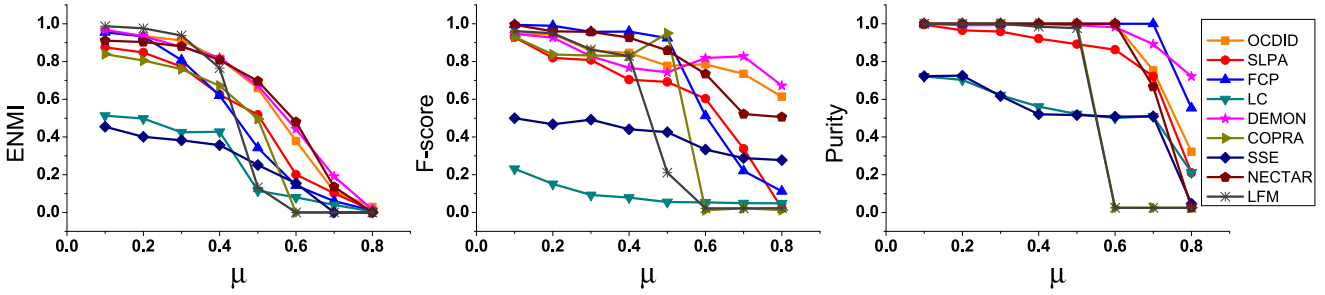


FIGURE 5. Performance of different algorithms with LFR networks when the parameter μ varied from 0.1 to 0.8.

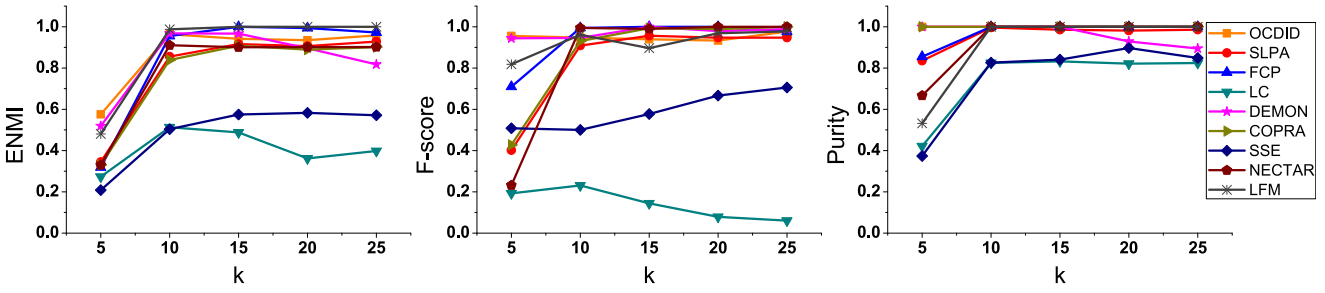


FIGURE 6. Performance of different algorithms with LFR networks when the average degree k ranged from 5 to 25.

Similar on ENMI metric, OCDID, DEMON, and SLPA still performed comparatively well. In terms of the purity metric, DEMON achieved the best performance, and NECTAR, OCDID, and SLPA also performed well.

3) **Effect of μ (community density)**
In order to further investigate the effects of the different methods on the community density, we varied the mixing parameter μ to generate networks. The mixing parameter μ represents the fraction of the edges of a node outside its community. Thus, it is often used to regulate the density of a community for detecting overlapping and disjoint communities. We varied the mixing parameter μ from 0.1 to 0.8, and fixed the other parameters as $O_m = 2$, $O_n/n = 0.1$, and $k = 10$. As shown in Fig. 5, the effectiveness of all the methods decreased gradually as the parameter μ increased. In terms of the ENMI metric, OCDID, DEMON, and NECTAR were very stable and they performed better than the other algorithms. FCP and LFM delivered impressive performance when the parameter $\mu < 0.3$, where LFM achieved the highest ENMI value. However, their effectiveness decreased significantly when the parameter μ was larger than 0.4. SLPA and COPRA also obtained comparable results. In terms of the F-score metric, OCDID, DEMON, and NECTAR yielded stable results in a similar manner to the ENMI values. FCP, COPRA, and LFM obtained very high scores when the parameter μ was less than 0.4, but their performance decreased significantly when $\mu > 0.6$. In particular, the ENMI scores for COPRA and LFM were almost equal to zero. In terms of the purity metric, OCDID, FCP, and DEMON also performed better than the other methods.

4) **Effects of k (average degree of node)**
Fig. 6 illustrates the effectiveness of each approach when the average degree k varied from 5 to 25. The other parameters were fixed at $O_m = 2$, $O_n/n = 0.1$, and $\mu = 0.1$. As shown in Fig. 6, most of the algorithms

did not perform well when the average degree of the network was relatively low. However, OCDID still produced better results than the other methods. In particular, OCDID obtained an ENMI score of 0.6 when the average degree k was as low as 5. OCDID still produced good results as the average degree k increased. In terms of the three metrics, SLPA, FCP, DEMON, COPRA, NECTAR, and LFM performed extremely well when the average degree k was larger than 10. By contrast, the LC and SSE methods did not perform well and they were not comparable to the other methods with these networks.

2) REAL-WORLD NETWORKS

In further comparisons of OCDID and the other algorithms, we tested their performance with seven popular real-world networks that exhibit distinct characteristics. These real-world data sets lacked ground truth data, so we used the EQ metric to evaluate the efficiency of the algorithms. Table 4 shows the evaluation results obtained for the different methods with real-world networks.

TABLE 4. The performances of several algorithms in real-world networks.

	Karate	Football	Polbooks	Citeseer	Jazz	Protein	Powergrid
OCDID	**0.351**	0.572	**0.436**	**0.330**	0.237	**0.570**	**0.447**
SLPA	0.316	0.402	**0.435**	0.069	0.220	0.566	**0.558**
FCP	0.130	0.464	0.351	0.221	0.003	0.109	0.162
LC	0.125	0.048	0.294	0.257	0.014	0.346	0.178
DEMON	0.229	0.554	0.198	0.024	0.060	0.115	0.087
COPRA	0.321	0.413	**0.435**	0.283	**0.420**	0.118	0.245
SSE	0.160	0.188	0.118	0.150	0.157	0.13	0.273
NECTAR	0.284	0.585	0.334	0.318	0.272	0.162	0.314
LFM	0.258	**0.699**	0.351	0.308	0.152	0.171	0.103

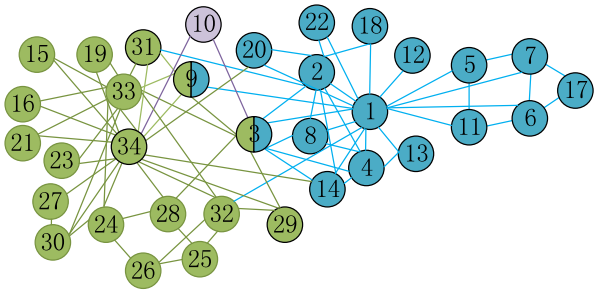


FIGURE 7. The OCDID algorithm detected two overlapping communities in the Karate network.

The OCDID method performed very well with the Zachary's Karate club network, where it obtained the highest EQ score ($EQ = 0.351$). Fig. 7 shows the overlapping community detection results produced by OCDID, which detected three communities. Nodes 3 and 9 were considered to be overlapping nodes because these nodes exchanged information among different communities. We identified the overlapping nodes by calculating the volume of information that flowed through multiple communities. The SLPA and COPRA algorithms also performed well and most of the nodes were correctly divided. However, FCP, LC, and SSE did not perform well, with low EQ scores, possibly because of the low average degree of the Karate club network. The LFM method achieved the highest EQ score with the football network. OCDID and most of the other approaches also performed well due to the higher average degree ($k = 10.66$). However, the LC and SSE algorithms incorrectly divided many of the nodes, with low EQ scores. With the Polbooks, Citeseer, and Protein networks, the OCDID algorithm still achieved the best performance and the highest scores although the average degree and clustering coefficients were low. COPRA obtained the highest EQ score with the Jazz network, and the OCDID, SLPA, and NECTAR methods also performed well, whereas FCP, LC, and DEMON failed to identify the overlapping communities. With the Power grid network, SLPA obtained the highest score ($EQ = 0.558$), and OCDID also produced excellent clustering results, where it performed better than the other algorithms. Unfortunately, the results produced by the FCP, LC, DEMON, and LFM algorithms were not ideal because the average degree and clustering coefficients were low in this network ($k = 2.669$, CC=0.107). Thus, the OCDID algorithm achieved good results with the low average degree and low clustering coefficient networks, as well as obtaining the ideal results for the high average degree networks.

Next, we employed the large-scale network derived from Amazon merchandise sales data to evaluate the performance of the algorithms. Because of the high complexity of the EQ measure, it could not be determined for this network. Thus, the ENMI, F-score, and purity metrics were used to assess the quality of overlapping community detection. Fig. 8 shows how the different algorithms performed with this network. OCDID obtained high performance, thereby indicating that the quality of overlapping community detection was high. SLPA, FCP, COPRA, and NECTAR achieved good results, but LC and DEMON did not perform well with this network. OCDID, SLPA, FCP, COPRA, and NECTAR obtained higher F-score values. The result obtained by LFM is not shown in the figure because its worst-case complexity is $O(n^2)$ and the runtime exceeded four days with this network.

In summary, DEMON and FCP performed very well with the synthetic networks, where they achieved most of the highest evaluation scores. However, the performance of these two algorithms was not ideal with real networks. By contrast, OCDID performed extremely well with both synthetic and real-world networks, where it could handle networks of different sizes and generate good partitions of the overlapping communities.

D. RUNTIME

In order to compare the computational time required for OCDID and various other algorithms at different network scales, we employed the LFR benchmark to create networks with different sizes. In particular, we fixed the parameters at $\mu = 0.1$, $O_m = 2$, $O_n/n = 0.1$, and $k = 10$, and varied the number of nodes n from 1,000 to 1,000,000. The runtimes of

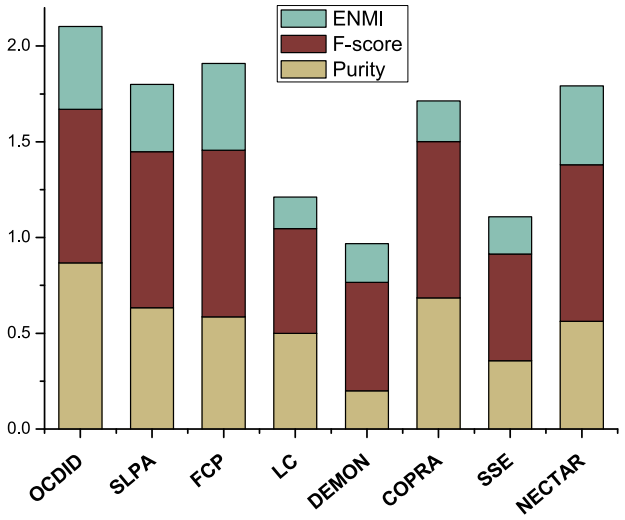


FIGURE 8. Effectiveness of different algorithms with the Amazon network.

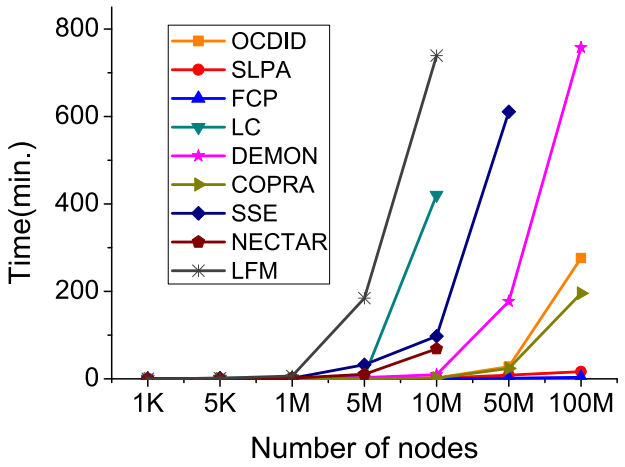


FIGURE 9. Runtimes for different algorithms with synthetic networks when the number of nodes varied from 1*K* to 100*M*.

the algorithms are compared in Fig. 9. When the number of nodes reaches 500, 000, the runtime was more than three days for the LC, SSE, and LFM algorithms, and thus the results obtained for these algorithms are not shown in the figure. Our proposed OCDID method was faster than the LC, DEMON, SSE, NECTAR, and LFM algorithms, where this advantage was particularly significant as the network size increased. This difference in the runtime required was mainly due to the low time complexity of $O(k \cdot n + L \cdot n \cdot k + |C| \cdot n)$, where $k \ll n$ and $|C| \ll n$. Therefore, the OCDID approach can handle large-scale networks, although it was slower than SLPA, FCP, and COPRA. These three algorithms required less time than OCDID but the SLPA and COPRA methods have stability problems, and FCP does not perform well with real-world networks (Table 4), especially networks with a low average degree.

V. CONCLUSIONS

In this study, we developed a new algorithm called OCDID for detecting overlapping communities in complex networks. In our algorithm, an information dynamics model is employed to simulate the communication of information between the nodes in networks. This model represents the flow of information on networks to intuitively depict the community structure and it facilitates the identification of overlapping nodes by calculating the volume of information on each node. We conducted experiments using synthetic and real-world networks to evaluate the performance of OCDID, where we compared OCDID with eight representative overlapping community detection methods. The experimental results demonstrated that OCDID performed well at identifying the overlapping communities and it was better than the representative algorithms considered in the evaluation.

REFERENCES

[1] J. Shao, F. Huang, Q. Yang, and G. Luo, "Robust prototype-based learning on data streams," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 5, pp. 978–991, May 2018.

[2] Z. Sun, B. Wang, J. Sheng, Y. Hu, Y. Wang, and J. Shao, "Identifying influential nodes in complex networks based on weighted formal concept analysis," *IEEE Access*, vol. 5, pp. 3777–3789, Mar. 2017.

[3] J. Shao, Q. Yang, Z. Zhang, J. Liu, and S. Kramer, "Graph clustering with local density-cut," in *Database Systems for Advanced Applications*. Cham, Switzerland: Springer, 2018, pp. 187–202.

[4] Z. Yu, J. Shao, Q. Yang, and Z. Sun, "ProfitLeader: Identifying leaders in networks with profit capacity," in *Proc. World Wide Web*, 2018, pp. 1–21.

[5] J. Shao, X. Wang, Q. Yang, C. Plant, and C. Böhm, "Synchronization-based scalable subspace clustering of high-dimensional data," *Knowl. Inf. Syst.*, vol. 52, no. 1, pp. 83–111, Jul. 2017.

[6] B. S. Khan and M. A. Niazi, "Network community detection: A review and visual survey," *CoRR*, vol. abs/1708.00977, Aug. 2017. [Online]. Available: http://arxiv.org/abs/1708.00977

[7] J. Shao, Z. Han, Q. Yang, and T. Zhou, "Community detection based on distance dynamics," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2015, pp. 1075–1084.

[8] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.

[9] J. Xie, B. K. Szymanski, and X. Liu, "SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *Proc. IEEE 11th Int. Conf. Data Mining Workshops*, Dec. 2011, pp. 344–349.

[10] S. Gregory, "Finding overlapping communities in networks by label propagation," *New J. Phys.*, vol. 12, no. 10, p. 103018, 2010.

[11] J. Wu *et al.*, "Overlapping community detection via network dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 85, p. 016115, Jan. 2012.

[12] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surv.*, vol. 45, no. 4, p. 43, 2013.

[13] A. Amelio and C. Pizzuti, "Overlapping community discovery methods: A survey," *Social Networks: Analysis and Case Studies*. Vienna, Austria: Springer, 2014, pp. 105–125.

[14] 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, Jun. 2005.

[15] E. Tomita, A. Tanaka, and H. Takahashi, "The worst-case time complexity for generating all maximal cliques and computational experiments," *Theor. Comput. Sci.*, vol. 363, no. 1, pp. 28–42, 2006.

[16] I. Farkas, D. Ábel, G. Palla, and T. Vicsek, "Weighted network modules," *New J. Phys.*, vol. 9, no. 6, p. 180, Jun. 2007.

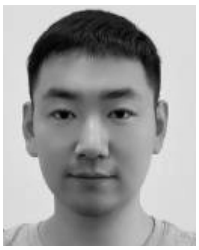
[17] J. M. Kumpula, M. Kivelä, K. Kaski, and J. Saramäki, "Sequential algorithm for fast clique percolation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, p. 026109, Aug. 2008.

[18] F. Reid, A. McDaid, and N. Hurley, "Percolation computation in complex networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2012, pp. 274–281.
[19] J. Baumes, M. Goldberg, and M. Magdon-Ismail, "Efficient identification of overlapping communities," in *Intelligence and Security Informatics*. Berlin, Germany: Springer, 2005, pp. 27–36.
[20] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New J. Phys.*, vol. 11, no. 3, p. 033015, 2009.
[21] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping community detection using seed set expansion," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA, 2013, pp. 2099–2108.
[22] Y.-Y. Ahn, J. P. Bagrow, and L. Sune, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–764, 2010.
[23] Z. Wu, Y. Lin, H. Wan, and S. Tian, "A fast and reasonable method for community detection with adjustable extent of overlapping," in *Proc. IEEE Int. Conf. Intell. Syst. Knowl. Eng.*, Nov. 2010, pp. 376–379.
[24] Y. Kim and H. Jeong, "Map equation for link communities," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 84, p. 026110, Aug. 2011.
[25] L. Chen, J. Zhang, and L.-J. Cai, "Overlapping community detection based on link graph using distance dynamics," *Int. J. Mod. Phys. B*, vol. 32, no. 3, p. 1850015, 2018.
[26] D. Li *et al.*, "Synchronization interfaces and overlapping communities in complex networks," *Phys. Rev. Lett.*, vol. 101, p. 168701, Oct. 2008.
[27] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: A local-first discovery method for overlapping communities," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2012, pp. 615–623.
[28] J. Reichardt and S. Bornholdt, "Detecting fuzzy community structures in complex networks with a potts model," *Phys. Rev. Lett.*, vol. 93, p. 218701, Nov. 2004.
[29] Q. Lu, G. Korniss, and B. K. Szymanski, "The naming game in social networks: Community formation and consensus engineering," *J. Econ. Interact. Coordination*, vol. 4, no. 2, p. 221, Nov. 2009.
[30] F. Breve, L. Zhao, and M. Quiles, "Uncovering overlap community structure in complex networks using particle competition," in *Artificial Intelligence and Computational Intelligence*. Berlin, Germany: Springer, 2009, pp. 619–628.
[31] J. Bibby, "Axiomatisations of the average and a further generalisation of monotonic sequences," *Glasgow Math. J.*, vol. 15, no. 1, pp. 63–65, 1974.
[32] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 3, p. 036106, 2007.
[33] Y. Cohen, D. Hendler, and A. Rubin, "Node-centric detection of overlapping communities in social networks," in *Proc. 3rd Int. Winter School Conf. Netw. Sci.* Cham, Switzerland: Springer, 2017, pp. 1–10.
[34] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech., Theory Exp.*, vol. 10, p. P10008, 2008.
[35] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, p. 046110, Oct. 2008.
[36] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," *Phys. A, Stat. Mech. Appl.*, vol. 388, no. 8, pp. 1706–1712, 2009.
[37] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams, and J. Fan, "Topic oriented community detection through social objects and link analysis in social networks," *Knowl.-Based Syst.*, vol. 26, pp. 164–173, Feb. 2012.
[38] B.-J. Sun, H. Shen, J. Gao, W. Ouyang, and X. Cheng, "A non-negative symmetric encoder-decoder approach for community detection," in *Proc. ACM Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA, 2017, pp. 597–606.
[39] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, p. 026113, Feb. 2004.

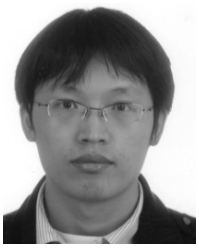
ZEJUN SUN received the B.Sc. degree in computer science from Henan Polytechnic University in 2003 and the M.Sc. degree in computer science from Xidian University, China, in 2008. He is currently pursuing the Ph.D. degree with Central South University, China. His research interests include data mining, complex network structure mining, and machine learning.

BIN WANG received the M.Sc. degree in mining engineering and the Ph.D. degree in computer science and technology from Central South University, China, in 1999 and 2003, respectively. He is currently a Professor with the School of Information Science and Engineering, Central South University. His research interests include transparent computing and software engineering.

JINFANG SHENG received the M.Sc. degree in computer science and technology and the Ph.D. degree in control theory and control engineering from Central South University, China, in 1996 and 2007, respectively. She is currently an Associate Professor with the School of Information Science and Engineering, Central South University. Her research interests include transparent computing and big data processing.



ZHONGJING YU received the B.Sc. degree in information and computing science from Shanxi Normal University in 2014. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. His current research interests include data mining and complex network analysis.



JUNMING SHAO received the Ph.D. degree (Hons.) *(summa cum laude)* from the University of Munich, Germany, in 2011. He became the Alexander von Humboldt Fellow in 2012. He not only authored papers on top-level data mining conferences, such as KDD, ICDM, and SDM (two of those papers have won the Best Paper Award), but also authored data mining-related interdisciplinary work in leading journals, including the *Brain*, the *Neurobiology of Aging*, and the *Water Research*. His research interests include data mining and neuroimaging.