IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 10, 2018, accepted May 10, 2018, date of publication May 21, 2018, date of current version July 25, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2838568

Semi-Supervised Community Detection Based on Distance Dynamics

LILIN FAN, SHENGLI XU, DONG LIU, AND YAN RU
School of Computer Science and Technology, Henan Normal University, Xinxiang 453007, China

Corresponding author: Dong Liu (liudonghtu@gmail.com)

This work was supported in part by the joint funds of the National Natural Science Foundation of China under Grant U1404604, in part by the Science and Technology Project of Henan Province under Grant 152102310313, and in part by the Ph.D. Research Startup Foundation of Henan Normal University under Grant QD14135.

ABSTRACT Community detection methods that are based entirely on the topology of the network do not always achieve higher accuracy. This implies that the topological information alone is insufficient to accurately uncover the community structures of networks. Recently, some methods were proposed that used prior information to improve the performance and accuracy of community detection. However, most of these methods have high time consumption and are not suitable for dealing with large-scale networks. In this paper, we propose a fast semi-supervised community detection method called SemiAttractor that integrates the prior information into the distance dynamics model. Experimental results from both artificial and real-world networks show that the proposed method can effectively improve the accuracy of community detection and reduce the time costs.

INDEX TERMS Community detection, semi-supervised, distance dynamics, graphs.

I. INTRODUCTION

Numerous studies have shown that community structures often exist in the real-world [1], [2]. For example, there are communities in social networks that are formed by groups of common interests or similar social backgrounds [3], [4]. There are communities exposed by functional units in biochemical networks that help them move in a better direction [5], [6]. The in-depth study of community structures helps us better understand the systematic structures and functional characteristics of real-world networks. Therefore, community detection has become increasingly important in the field of complex networks.

In recent years, scholars have increasingly devoted themselves to the field of community detection and proposed a large number of community detection algorithms. The classic community detection algorithms include the KL method [7], the Metis method [8], the maximum stream/minimum cut method [9], [10], the GN algorithm [11], [12], the modularity [4], the Louvain Spectral clustering [13], the MCL [14], the Infomap [25], the Multi-level Learning based Memetic Algorithm [39], the nature-inspired algorithms [42], [43] and the label propagation(LPA) [22]. All of the above methods belong to unsupervised schemes, which means that they can only use the network's topological information.

In real applications, we can obtain some useful prior information such as individual labels and pairwise constraints [15]. Individual labels are the labels which are placed on individual nodes in the networks. The pairwise constraints indicate that two nodes belong to the same community (a ‘‘must-link’’ constraint)or different communities (a ‘‘cannot-link’’ constraint). However, how to effectively combine the prior information with the topology of the network to guide the process of community detection is still a challenge.

Recently, scholars have proposed useful semi-supervised community detection methods, such as the NMF-LSE [16], the SNMF-SS [17], the SS-masterl [18], the SpinGlassSS [19] and Discrete potential theory [20]. By incorporating the prior information, these semi-supervised community detection methods effectively improve the accuracy of community detection. However, most of these methods have high time complexity and create the difficulty in dealing with large-scale networks.

Lately, *Shao et al.* [21] proposed a community detection method based on distance dynamics. This method uses the dynamic distance model to describe the distance links in the network and defines three kinds of intuitive interaction modes. It dynamically discovers the community structure by simulating the changes of the distance between nodes in the network. This algorithm is capable of revealing high-quality communities and can handle large-scale networks. However, this method exclusively considers the network topology and ignores any type of prior information for community detection performance.

How to integrate the prior information into distance dynamics is an important challenge. To solve this problem, we propose a novel semi-supervised community detection algorithm based on distance dynamics. We combine the pairwise constraints with the topology of the network to modify the distances between the nodes. This method reduces the time consumption and makes full use of the prior information. The comparisons performed on both artificial and real-world networks show that our proposed SemiAttractor algorithm can significantly improve the community detection performance.

The structure of this paper is as follows. Section II briefly reviews the previous works related to semi-supervised community detection. Section III details the semi-supervised community detection algorithm based on distance dynamics and the complexity analysis of this algorithm. Section IV compares the experimental results of our algorithm with other related algorithms on real and artificial networks. Section V summarizes this article.

II. RELATED WORK

In recent years, scholars have proposed many semi-supervised community detection methods [16]–[20], [23], [26], [27], [35]. These methods take advantages of the network topology and also utilize prior information to improve the accuracy of community detection.

Recently, many semi-supervised community detection methods based on the NMF model [38] have been proposed. Zhang [18] proposed a semi-supervised community detection framework. Under this framework, many forms of NMF-based algorithms, such as the NMF-LSE, SNMF and BNMF, can all be employed for optimization. To effectively utilize the available prior information, *Zhang et al.* considered that the pairwise constraints between nodes can be used for logical derivation between nodes. Therefore, *Zhang et al.* [24] proposed an enhanced semi-supervised community detection method. They found that the "must-link" constraint gives better community detection results than the "cannot-link". *Zhang et al.* [16] proposed a semi-supervised symmetric non-negative matrix factorization method for community detection. It incorporates pairwise constraints into the adjacency matrix of the original network to guide the clustering of nodes. This method theoretically proved the equivalence between the objective function and modularity. *Liu et al.* [26] presented a semi-supervised non-negative matrix factorization model based on the graph regularization and "must-link" constraints. They refined this model by introducing a set of parameters to adjust the degree of each node, which leads to a new semi-supervised NMF model, called PSSNMF that considers the node popularity. Unfortunately, PSSNMF can only use the "must-link". However, the number of communities for most NMF-based community detection methods needs to be fixed in advance.

Other researchers proposed some semi-supervised community detection methods based on the physics model. Eaton and Mansbach [19] proposed a semi-supervised Spin-Glass model [37] and proved that the method of minimizing the Hamiltonian energy equation by using the potts model is mathematically equivalent to the modularity Q. They considered prior information for both individual community labels and pairwise constraints, construct the rewards and punishments $U(C)$, construct a new Hamiltonian energy equation $H'(C)$, and finally obtain the best energy equation $H'(C)$ for community detection. In addition, *Eaton et al.* conducted experiments on noise effects and found that the semi-supervised Spin-Glass model can effectively counteract the effects of noise. To use individual labels as prior information, *Liu et al.* [20] designed a semi-supervised community detection algorithm. Based on the discrete potential theory, *Liu et al.* considered each node to be an elementary charge, and transformed community detection into potential transfer theory based on the heat transfer process. However, most of the semi-supervised community detection methods based on the physics model have high time complexity and limited abilities to deal with large-scale networks.

Recently, Shao *et al.* [21] proposed the distance dynamics model to detect the community structure of complex networks based on the links. Each link is associated with an initial Jaccard distance [36], which indicates the similarity between two neighbor nodes. The distance between links in a network is affected by three aspects. The direct link between two nodes can reduce the distance between links. The common neighbors of two nodes can also reduce the distance between links. Exclusive neighbors will increase the distance of the link. With the evolution of time and drive of network topology, the distance of the links between those nodes with the highest similarity first synchronizes and then rapidly decreases to 0. Meanwhile, the distance of the links between those nodes with the highest dissimilarity rapidly increases to 1. Then, in a sequential process, more nodes synchronize together, and the distance between them gradually decreases or increases. Finally, the intrinsic communities of different sizes in the network are searched by disconnecting the internal links of the network. However, the distance dynamics model cannot use any type of prior information for community detection.

The motivation of this paper is to integrate the prior information into the community detection method based on the distance dynamics model. We combine the original network topology with the pairwise constraints to modify the distance between the nodes. The initial kinetic distance of the links in the network has changed. By doing so, we can reduce the time step to accelerate the convergence of the distance between nodes and improve the accuracy of our algorithm.

III. SEMI-SUPERVISED COMMUNITY DETECTION BASED ON DISTANCE DYNAMICS

In this section, we propose a novel semi-supervised community detection method based on distance dynamics. Each link is associated with an initial distance d. The distance dynamics are based on three interaction models. Finally, as time passes, due to the influence of its neighbors, the initial distance d of each link dynamically increases or decreases. Driven by prior information and network topology, all distances d will converge, and the communities can be obtained by dislodging the links with maximal distances. At the end of this section, we give an analysis of the complexity of this method.

A. PRELIMINARIES

To clearly illustrate our approach, some definitions are introduced. Let $G = (V, E, W)$ be an undirected graph, where V is the set of all nodes, E is the set of all links and W is the weight of the corresponding links. Each link $e = \{u, v\} \in E$ represents the link between two nodes u and v, and $\omega(u, v) \in W$ indicates the weight of the corresponding link. $\forall e = \{u, v\} \in E$ and $\omega(u, v) = 1$ in the case of an unweighted graph. For an undirected graph $G = (V, E, W)$ with n nodes, the corresponding adjacency matrix A is:

$$A_{uv} = \begin{cases} 1 & if\ e = \{u, v\} \in E, \\ 0 & otherwise. \end{cases} \tag{1}$$

where $A_{uv} = 1$ denotes the link between nodes u and v if the link exists. Otherwise, $A_{uv} = 0$. The adjacency matrix A represents a symmetric matrix of $n \times n$.

$ngb(u)$ denotes all the neighbors of node u, but it does not include itself.

$$ngb(u) = \{v|Auv = 1, v = 1, 2, \ldots, n\} \tag{2}$$

$deg(u)$ indicates the degree of node u.

$$deg(u) = \sum A_u \tag{3}$$

For the two nodes u and v that are connected by a link $e = \{u, v\} \in E$, their common neighbors set $CN(e)$ is defined as:

$$CN(e) = ngb(u) \cap ngb(v) \tag{4}$$

The formula of the Jaccard distance [36] of the two nodes u and v is defined as follows:

$$d(u, v, 0) = 1 - \frac{card(CN(e)) + 2}{card(ngb(u) \cup ngb(v))} \tag{5}$$

where $card(CN(e))$ denotes the number of common neighbors of nodes u and v but not includes nodes u and v. $d(u, v, 0)$ indicates the initial distance between two nodes.

For the weighted graph, the Jaccard distance of the two nodes u and v is defined as follows:

$$d(u, v, 0) = 1 - \frac{\sum\limits_{x \in ngb(u) \cap ngb(v)} (w(u, x) + w(v, x))}{\sum\limits_{\{x,y\} \in E; x,y \in ngb(u) \cup ngb(v)} w(x, y)} \tag{6}$$

In real applications, there are some prior information that can be used to guide the community detection. More concretely, pairwise constraints specify the relative community membership for pairs of nodes. They can serve to identify pairs of nodes that belong to nodes in the same community (a "must-link" constraint) or different communities (a "cannot-link" constraint). We denote the set of "must-link" and "cannot-link" constraints as C_{ml} and C_{cl} respectively.

We introduce a new matrix A^* to add pairwise constraints. If nodes u and v belong to the same community, $A^*_{uv} = \alpha < -1$. If not, $A^*_{uv} = \beta > 1$, and otherwise $A^*_{uv} = 0$. Here we integrate the prior information into the adjacency matrix A such that $A^* = A^* + A$. A^* contains the following:

$$A^*_{uv} = \begin{cases} 1 & if\ u \dot{\sim} v, \\ 1 + \alpha & if\ u \tilde{=} v, \\ 1 + \beta & if\ u \tilde{\neq} v, \\ \alpha & if\ u \doteq v, \\ \beta & if\ u \neq v, \\ 0 & otherwise. \end{cases} \tag{7}$$

where $u \dot{\sim} v$ means that there is a link between nodes u and v but it is not sure whether they are in the same community. $u \tilde{=} v$ means that there is a link between nodes u and v and they are in the same community. $u \tilde{\neq} v$ indicates that there is a link between nodes u and v but they belong to different communities, $u \doteq v$ means that there is no link between nodes u and v but they are in the same community, $u \neq v$ means that there is no link between nodes u and v and they belong to different communities.

B. SEMI-SUPERVISED INTERACTION MODEL

The intrinsic links of real-world networks give a natural way to model the interaction. For each node, it naturally interacts with its adjacent nodes. Let $e = \{u, v\} \in E$ be a link between two adjacent nodes u and v. There are three distinct scenarios that influence the distance $d(u, v, 0)$ and rely on its local topological structure. (1) DI indicates the influence from the interactions of direct link nodes. (2) CI indicates the influence of common neighbors. (3)EI indicates the influence of exclusive neighbors.

To incorporate the prior information into the community detection method based on the distance dynamics model, we modify the initial distance $d(u, v, 0)$, $DI(u, v, 0)$, $CI(u, v, 0)$ and $EI(u, v, 0)$ as follows. We modify the value of equation(5) or (6) using the matrix A^*_{uv} that contains the prior information and the network topology. If $A^*_{uv} = 1 + \alpha$, then $d(u, v, 0) = 0$. If $A^*_{uv} = 1 + \beta$, then $d(u, v, 0) = 1$. Otherwise, $d(u, v, 0)$ is still the initial value.

$$d(u, v, 0) = \begin{cases} 0, & if\ \{u, v\} \in C_{ml}, \\ 1, & if\ \{u, v\} \in C_{cl}, \\ d(u, v, 0), & otherwise. \end{cases} \tag{8}$$

The distance $d(u, v, 0)$ between nodes u and v is obviously influenced by two directly linked nodes u and v, as one node attracts another towards itself. Thus, the distance $d(u, v, 0)$ decreases. We use $DI(u, v, 0)$ to describe this distance change:

$$DI(u,v,0)=\begin{cases}0, & \mathit{if}\ \{u,v\}\in c_{ml},\\ 1, & \mathit{if}\ \{u,v\}\in c_{cl},\\ -(\dfrac{f(1-d(u,v,0))}{\deg(u)}+\dfrac{f(1-d(u,v,0))}{\deg(v)}), & \\ \quad \mathit{otherwise}. & \end{cases} \tag{9}$$

where $f(\cdot)$ is a coupling function. We use $\sin(\cdot)$ in this study.

The influence from the common neighbors of nodes u and v attract the two nodes. Thus,it changes the smaller distance. We use $CI(u,v,0)$ to denote the influence of the common neighbors of two nodes:

$$CI(u,v,0) = \begin{cases}0, & \mathit{if}\ \{u,v\}\in c_{ml},\\ 1, & \mathit{if}\ \{u,v\}\in c_{cl},\\ -\sum\limits_{x\in CN(e)}(\dfrac{f(1-d(x,u,0))(1-d(x,v,0))}{\deg(u)} + \dfrac{f(1-d(x,v,0))(1-d(x,u,0))}{\deg(v)}), & \\ \mathit{otherwise}. & \end{cases} \tag{10}$$

In addition, for two nodes connected by the link $e=\{u,v\}$, there will also exist some neighbors that exclusively belong to node u or v. We denote them as $EN(u)=\{ngb(u)-CN(e)\}-\{v\}$, and similarly, $EN(v)=\{ngb(v)-CN(e)\}-\{u\}$. These exclusive neighbors also affect the distance between nodes. We use $EI(u,v,0)$ to express this influence:

$$EI(u,v,0) = \begin{cases}0, & \mathit{if}\ \{u,v\}\in c_{ml},\\ 1, & \mathit{if}\ \{u,v\}\in c_{cl},\\ \left(\begin{array}{l} -\sum\limits_{x\in EN(u)}(\dfrac{f(1-d(x,u,0)\rho(x,u,0))}{\deg(u)})\\ -\sum\limits_{y\in EN(v)}(\dfrac{f(1-d(y,v,0)\rho(y,v,0))}{\deg(v)})\end{array}\right), & \\ \mathit{otherwise}. & \end{cases} \tag{11}$$

Here, we use ρ to denote the degree of positive or negative influence on distance $d(u,v,0)$, and λ is the cohesion parameter.

$$\rho(x,u,0)=\begin{cases}1-d(x,v,0), & 1-d(x,v,0)\geq\lambda,\\ 1-d(x,u,0)-\lambda, & \mathit{otherwise}.\end{cases} \tag{12}$$

Finally, the dynamic distance can be updated as:

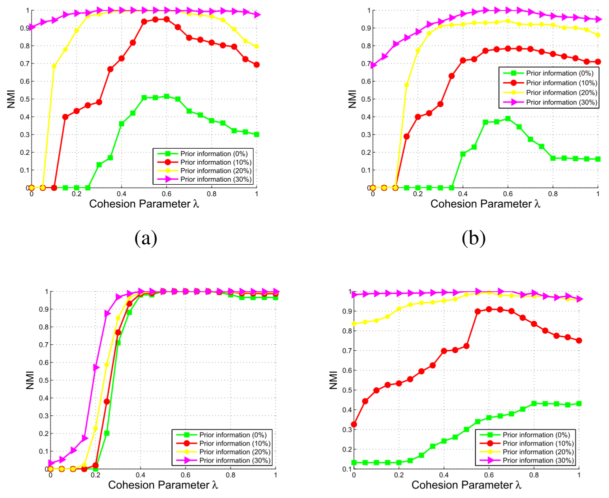
$$d(e,t+1)=d(e,t)+DI(e,t)+CI(e,t)+EI(e,t) \tag{13}$$

Where t is the number of time steps, $d(e,t+1)$ is the distance at the last time stamp t. Finally, when $d(e,t)\leq 0$ or $d(e,t)\geq 1$ for all links, the iteration process is stopped. After the iterations, the links where $d(e,t)\geq 1$ are removed from the original network, and the final communities are obtained.

C. THE SEMIATTRACTOR ALGORITHM

In this section, we specifically describe the SemiAttractor algorithm. The convergence of the distance of links is accelerated by modifying the initial distance with the network topology and pairwise constraints. Finally, the distances of all links converge, and the community structure of the network is naturally detected by dislodging the links with long distances.

The cohesion parameter λ is used to determine the positive or negative interaction influence on the distances from exclusive neighbors (see equation(12)). Figure 1 plots the NMI with different values of λ, ranging from 0 to 1 on the synthetic and real-world networks. Through the analysis of the NMI, we can see that the SemiAttractor allows for the emerging optimum partitioning with the parameter λ in a stable range $[0.5, 0.7]$. The clustering results with respect to distinct parameters are further illustrated in Figures 1(a)-(d). Extensive experiments further demonstrate that SemiAttractor usually produces an optimum result within the range $\lambda\in[0.5,0.7]$. Finally, the pseudocode of the SemiAttractor is given in Algorithm 1. We set the cohesion parameter $\lambda=0.6$ in our algorithm SemiAttractor.



(a)

(b)

(c)

(d)

FIGURE 1. The accuracy of cohesion parameter λ on the artificial benchmark and real-world networks. No prior information (0%) is converted into the Attractor algorithm. (a) λ on LFR $\mu=0.7$ network. (b) λ on LFR $\mu=0.8$ network. (c) λ on Football network. (d) λ on Political Events network.

Time Complexity Analysis: To study the SemiAttractor, we need to get the initial distance of the two nodes under one link in the network, where the time is calculated as $O(|E|)$. It is also necessary to calculate the corresponding Jaccard distance of the private neighbor with time complexity $O(k|E|)$, where k is the approximate average number of individual neighbors of the two linked nodes. During each iteration, we only need to recall the distance at the previous time stamp. Therefore, the time complexity of this step is $O(t|E|)$,

Algorithm 1 Semi-Supervised Community Detection Based on Distance Dynamics (SemiAttractor)

```
Input: undirected graph G=( V,E,W), cohesion
       parameter $\lambda$ the constraints matrix $A^*$
Output: the community label $c_i$ of each node $u_i$
Generate adjacency matrix A by E
for each link $e = (u, v) \in E$ do
  Compute the initial distance $d(e, 0)$ by equation(5)
    for each node $x \in EN(u)$ do
      Compute the distance $d(u, x, 0)$ by equation(6)
    end for
    for each node $y \in EN(v)$ do
      Compute the distance $d(v, y, 0)$ by equation(6)
    end for
end for
$A^* = A + A^*$
Modified the all distance $d(e, 0)$ by the constraint
  matrix $A^*$ using equation(8)
while (any($d(e, t) > 0 \& d(e, t) < 1$)
   $d(e, t + 1) = d(e, t) + DI(e, t) + CI(e, t) + EI(e, t)$
    by equation(13)
end while
for each link $e = (u, v) \in E$ do
  Modified the constraint matrix $A^*$ by the distance
   $d(e, t + 1)$
end for
Find the community by BFS
if $d(e, t + 1) = 1$ then
  Remove the link $e$ from the network
end if
Find the resulting components (communities) C
return C;
```

where t is the number of time steps. Hence, the total time complexity is $O(|E|+k|E|+t|E|)$. During the operation of the SemiAttractor, the more prior information we use, the smaller the value of t is. Section IV presents the change of t in more detail.

IV. EXPERIMENTS

In this section, we first detail the datasets used in the experiments. Second, we introduce four other benchmark algorithms to compare with the SemiAttractor. Then, we specifically describe how to use the prior information. Subsequently, we present the evaluation standard of the experimental results. Finally, we conduct a specific analysis of the experimental results. All experiments were run on Intel (R) Core i7-7820 CPU 2.9GHz, processor with 32.0GB RAM.

A. DATASETS

In the experiments, we use two types of artificial benchmarks and several classic real-world networks. The networks are some detail described as follows:

Synthetic LFR networks [28]: The LFR networks were designed by *Lancichinetti et al.* These networks have practical features, such as power law distribution and community size, and they are most commonly used as simulation datasets in current community detection research. They are also the artificial networks that are closest to real-world networks. The main parameters of LFR networks are as follows. N represents the number of nodes, k represents the average degree of nodes in the network, $maxk$ represents the node's maximum degree, and γ represents the parameter of node degree distribution. β represents the community size distribution. $minc$ represents the number of nodes included in the smallest community. $maxc$ represents the number of nodes included in the largest community. μ is the mixed parameter, which indicates the probability that a node is connected to the external community. The larger the value of μ is, the more difficult the community will be found. By setting different parameters, different types of simulated networks can be generated. LFR networks have known community structures, so they can be used to evaluate the quality of communities that were found by our algorithm. In our experiments, all parameters of LFR networks were set as follows: $N = 1000$, $k = 10$, $maxk = 50$, $minc = 10$, and $maxc = 100$.

Real-world Datasets:

Football Network [5] *Newman* created a complex social network based on the 2000 American College Football League. This network consists of 115 nodes and 613 links. The nodes in the network represent the football teams. The links connected by the two nodes represent the competition between the two teams. In this network, the 115 football teams are divided into 12 conferences.

Polbooks Network [29] *Valdis Krebs* built the Polbooks network based on the online sales of American political books on Amazon. In this network, the nodes represent political books sold by online booksellers, and the links indicate that the books were bought by the same buyers. This network consists of 105 nodes and 441 links. The nodes are divided into the three categories of "liberal", "neutral", and "conservative".

Polblogs Network [30] The Polblogs network was compiled by *Lada Adamic* in 2005. It represents the political inclination of people who posted blogs, and these data are displayed by the blog directory. Based on the incoming and outgoing links and posts during the 2004 presidential election period, the network contains 1490 nodes and 9518 links. Each node in this dataset has a corresponding attribute description (0 or 1) in which 0 indicates left or liberal and 1 indicates right or conservative.

Political Events Network The political events network represents the cooperation and hostility between 196 countries. It was created through the determination of 336,555 different political events between countries (or representatives of those countries) reported in the press from January 1, 2005, to December 4, 2010. Each event has a corresponding *Goldenstein* score [31]. We can form 988 links based on these scores, and then get the entire network.

When we choose a threshold of 3 for our experiment, we can determine whether the relationship between states is hostile or cooperative. The goal of the community division is to restore six geopolitical areas that represent 6 communities.

Enron Network [32] (http://www.cs.cmu.edu/ enron/) The Enron e-mail dataset was generated from the Enron network, which consists of a set of emails representing the e-mail communications between 156 people. There are 673 links in this network. Each link indicates that at least two people connected by this link have sent or received at least one e-mail. According to the job titles provided by *Shetty* and *Adibi*, this network can be divided into two communities.

Cora Network [33] The Cora dataset is composed of machine learning essays. It covers 2,708 scientific publications and 5,429 links. The publications are grouped into seven categories, each of which is described by a vector of 0/1 values.

BlogCatalog Network This is the data set that Tang and Liu [40] crawled from the social network BlogCatalog (http://www.blogcatalog.com). BlogCatalog is a social blog directory website. We can grab the group membership of the friendship network by gaming. The data set contains 10,312 nodes, 333,983 links are divided into 39 communities.

Northeastern Network Traud *et al.* [41] used Facebook data from Northeastern University in September 2005 to analyze the social structure of the Facebook friendship network from both micro and macro perspectives. The data set contains 13,882 nodes and 381,936 links, which are divided into 7 communities.

B. BASELINE

To show the effectiveness of our method, our experiments are compared with the number of time steps and the accuracy of the Attractor algorithm. Then, our experiment selects three representative semi-supervised community detection algorithms to compare the accuracy with of our algorithm.

Attractor [21] is an unsupervised community detection method based on distance dynamics. A detailed description of this method was previously mentioned in Section II.

PSSNMF [26] is the NMF framework based on the graph regularization and "must-link" constraints. A set is introduced to reconstruct the parameters of the NMF model to adjust the degree of each node and utilize the influence of the node's degree to make full use of the prior information.

NMF-LSE [18] is a semi-supervised community detection method based on the Least Squares Error under the NMF framework. This method directly encodes the prior information into the adjacency matrix and modifies the topology of the network to clarify the network's community structure.

Spin-GlassSS [19] is a semi-supervised community detection method based on the spin-glass model. The prior information is added by introducing a penalty function into the Hamiltonian energy equation and then minimizing the Hamiltonian energy equation using the potts model. Finally, the new module function is optimized by the simulated annealing algorithm to get the community division in the network.

In our experiments, the PSSNMF, NMF-LSE and Spin-GlassSS give the actual number of communities in advance. The other parameters are the same as those used in the original papers.

C. HOW TO USE THE PRIOR INFORMATION

There are many kinds of prior information, such as pairwise constraints and individual labels. Given an undirected graph G that has n nodes and k communities, there are $N = \frac{n(n-1)}{2}$ pairs of pairwise constraints. The constraints are divided into two types: "must-link" and "cannot-link". The total number of "must-link" is $N_{ml} = \sum_{c=1}^{k} \frac{n_c(n_c-1)}{2}$, where n_c represents the number of nodes included in the c-th community. The total number of "cannot-link" is $N_{cl} = N - N_{ml}$. When we choose pairwise constraints, we randomly select two nodes from set V. However, there may be no link between the two nodes. In other words, the existence of this prior information does not affect the distance. To make full use of the prior information, we use the logical reasoning techniques proposed by Zhang et al. [24] to enhance the prior information. Nodes i and j are in the same community, and the nodes j and k are in the same community. Therefore, nodes i and k are in the same community, and the matrix A^*_{ik} can be modified as α. However, if nodes j and k are in different communities, then nodes i and k are in different communities, and the matrix A^*_{ik} can be modified as β.

In our experiments, we consider that the PSSNMF can only use the "must-link" in the pairwise constraints. Similarly, the other three semi-supervised algorithms randomly select the same number of constraints from the total constraints. Since the acquisition of the prior information is random, all experiments must run 50 times to get the average value with the same amount of prior information. More details of the real-world networks are shown in Table 1.

D. EVALUATION STANDARDS

In our experiments, the authoritative evaluation methodology of the normalized mutual information(NMI) [34] is selected. It is defined as follows:

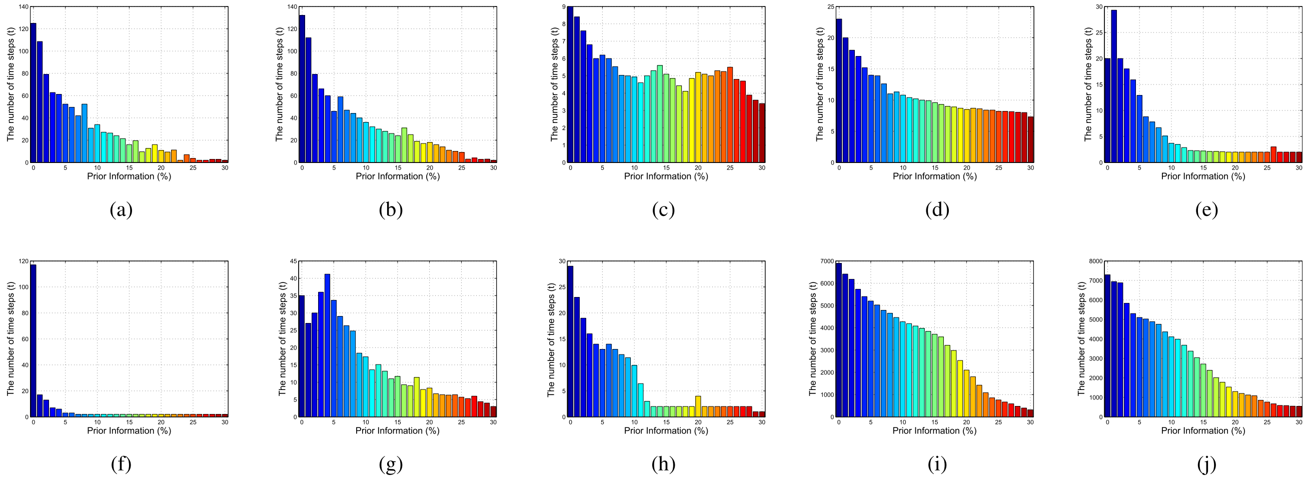
$$NMI(C_1, C_2) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} c_{ij} \log \frac{c_{ij}c}{c_i^{(1)} c_j^{(2)}}}{\sqrt{(\sum_{i=1}^{k} c_i^{(1)} \log \frac{c_i^{(1)}}{c})(\sum_{j=1}^{k} c_j^{(2)} \log \frac{c_j^{(2)}}{c})}}$$

where C_1 is the ground-truth community label and C_2 is the computed community label. k and c are the numbers of communities and nodes, respectively. c_{ij} is the number of nodes in the ground-truth community i that are assigned to the computed community j. $c_i^{(1)}$ is the number of nodes in the ground-truth cluster i, and $c_j^{(2)}$ is the number of nodes in the computed cluster j, where $NMI \in [0, 1]$.

TABLE 1. Statistics of real-world datasets used.

Datasets	$\lvert V\rvert$	$\lvert E\rvert$	$\#Class$	AD	CC	N_{ml}	N
Football	115	613	12	10.661	0.403	529	6555
Polbooks	105	441	3	8.400	0.488	2157	5460
Enron	156	673	2	9.179	0.461	7450	12090
Polblogs	1490	19090	2	12.768	0.172	554449	1109305
Political Events	196	988	6	5.041	0.210	3921	19110
Cora	2708	5429	7	4.010	0.293	657055	3665278
BlogCatalog	10312	333983	39	64.776	0.463	3019350	53163516
Northeastern	13882	381936	7	55.026	0.457	4595406	96348021

Statistics of real-world datasets, where $|V|$: all nodes of the network; $|E|$: all links of the network; $\#Class$: the number of real communities in the network; AD: average degree; CC: clustering coefficient; N_{ml}: "must-link" constraints of the network; N: all constraints of the network



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

FIGURE 2. The number of time steps (t). 0% prior information represents the Attractor algorithm. (a) LFR $\mu = 0.7$. (b) LFR $\mu = 0.8$. (c) Football network. (d) Polbooks network. (e) Enron network. (f) Polblogs network. (g) Political events network. (h) Cora network. (i) BlogCatalog network. (j) Northeastern network.

E. RESULTS ANALYSIS

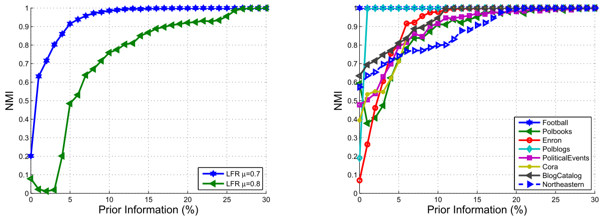
1) COMPARISON OF THE SEMIATTRACTOR AND ATTRACTOR IN THE NUMBER OF TIME STEPS AND ACCURACY OF COMMUNITY DETECTION

We first assess the number of time steps (t) of the SemiAttractor and Attractor algorithms on artificial and real-world networks, as shown in FIGURES 2(a)-(j). It is easy to determine that the t required for all networks to complete the community detection consistently and rapidly decreases the used prior information increases. Notice that no prior information (0%) is converted into the Attractor algorithm. In FIGURES 2(a)(b), in LFR networks that $\mu = 0.7$ and $\mu = 0.8$, the respective t required for the Attractor algorithm reaches 125 and 132. As the prior information increases, t decreases. Furthermore, when the prior information reaches 30%, the number of time steps becomes very few. In particular, in FIGURE 2(f), we can see that when the prior information is 0%, the t required by the Attractor algorithm reaches 117. Once we add a small amount of the prior information to the network, the number of time steps of in the SemiAttractor algorithm will decrease significantly.

However, there are some unsatisfactory results. In FIGURE 2(e), when 1% prior information is used by the SemiAttractor algorithm, its t to complete the Enron network community detection is greater than that of the Attractor algorithm. In FIGURE 2(g), when $4 \sim 5\%$ prior information is added by the SemiAttractor algorithm, the number of time steps to complete the Political Events network community detection is greater than the Attractor algorithm. Take the Enron network as an example. When we integrate 1% prior information into network topology, the number of time steps t is equal to 24 in the SemiAttractor algorithm at nodes 54 and 72, while the Attractor algorithm requires the number of time steps $t = 20$. t equals to 29 in the SemiAttractor algorithm at nodes 72 and 75, but the Attractor algorithm only needs the number of time steps $t = 20$. The reason for the relatively higher t of the SemiAttractor may be those nodes connect multiple communities with tight internal connections and sparse external connections. The experiments reveal that the increase of t is caused by a few inter-community links. In addition, most of the distances of links in the

SemiAttractor algorithm converge faster than in the Attractor algorithm.

From FIGURES 2(a)-(j), when the prior information is increased to 10%, the t is reduced to half of the Attractor algorithm. These experimental results all prove that our algorithm can effectively reduce the number of time steps t and the time costs.



(a)

(b)

FIGURE 3. **The performance 0% to 30% prior information on the LFR benchmark and Real-world networks. 0% prior information represents the Attractor algorithm. (a) LFR networks. (b) Real-world networks.**

FIGURES 3(a)(b) depict the performance of SemiAttractor algorithm on the artificial and real-world networks, with no prior information (0%) is converted into the Attractor algorithm. In the case of an increase in prior information, the NMI of most networks increases rapidly. In addition, we can find that when the prior information is increased to 15% in both artificial and real-world networks, the NMI values of our algorithm all reach 0.9 or more. From FIGURE 3 (a), when μ is set to 0.7, the NMI value of the SemiAttractor continues to increase with the addition of prior information. When μ is set to 0.8, the NMI value of the SemiAttractor maintains an upward trend. Moreover, when the amount of prior information is 30%, the SemiAttractor's NMI value almost reaches 1, which means that the algorithm achieved the best community division. From FIGURE 3 (b), in particular, the NMI value of the Polblogs network is 0.1903, which is calculated by the Attractor algorithm. However, the NMI value rapidly increases to 0.9972 after adding 1% prior information, which is almost perfect.

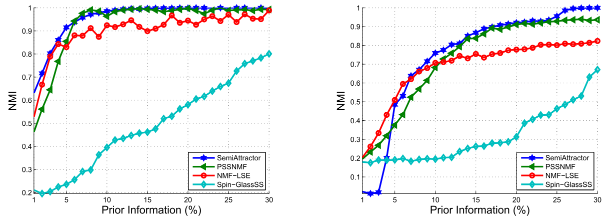
There are also unsatisfactory results with the Polbooks network. For example, the NMI value of the Polbooks network by the Attractor algorithm is 0.5917, but the NMI value of the Polbooks network with the SemiAttractor algorithm is lower than that with the Attractor algorithm after 1% prior information is added. This situation may occur because a small amount of prior information may not improve the accuracy of community detection under the topology of particular networks. Nonetheless, this situation does not affect the overall effect because when the prior information continues to increase, the NMI value will continue to rise.

Through the comparison of t between the Attractor and SemiAttractor algorithms with different numbers of prior information in FIGURE 2 and the change of NMI values in FIGURE 3, it is easy to discover that the SemiAttractor algorithm can quickly reduce time consumption and improve the accuracy with a small amount of prior information.

2) SEMIATTRACTOR ALGORITHM IS CONTRASTED WITH THE OTHER THREE SEMI-SUPERVISED COMMUNITY DETECTION ALGORITHMS

To evaluate the effectiveness of our proposed SemiAttractor algorithm, we show the comparative performances of our algorithm with the PSSNMF, NMF-LSE and Spin-GlassSS on both artificial and real-world networks.

As seen in FIGURE 4(a), in LFR networks, when μ is equal to 0.7, it is easy to see that our SemiAttractor algorithm performs better than the other three semi-supervised algorithms. When using our algorithm for community detection, its NMI values are the highest. In particular, when the prior information is increased above 10%, the NMI values obtained by the SemiAttractor algorithm reach almost 1, which means that we achieve the best community division. Of course, the accuracy of community division results obtained by the PSSNMF algorithm is also very high. From FIGURE 4(b), when μ is equal to 0.8, which is the most difficult case for community detection, we can find that the SemiAttractor algorithm is inferior to other algorithms only when the prior information is less than 7%, when the proportion of prior information is higher than 7%, with the increase of prior information, the NMI values obtained by our algorithm gradually increase. Moreover, when the proportion of prior information reaches $26 \sim 30\%$, only the NMI values corresponding to our algorithm reach 1, which is not achieved by the other three semi-supervised algorithms.



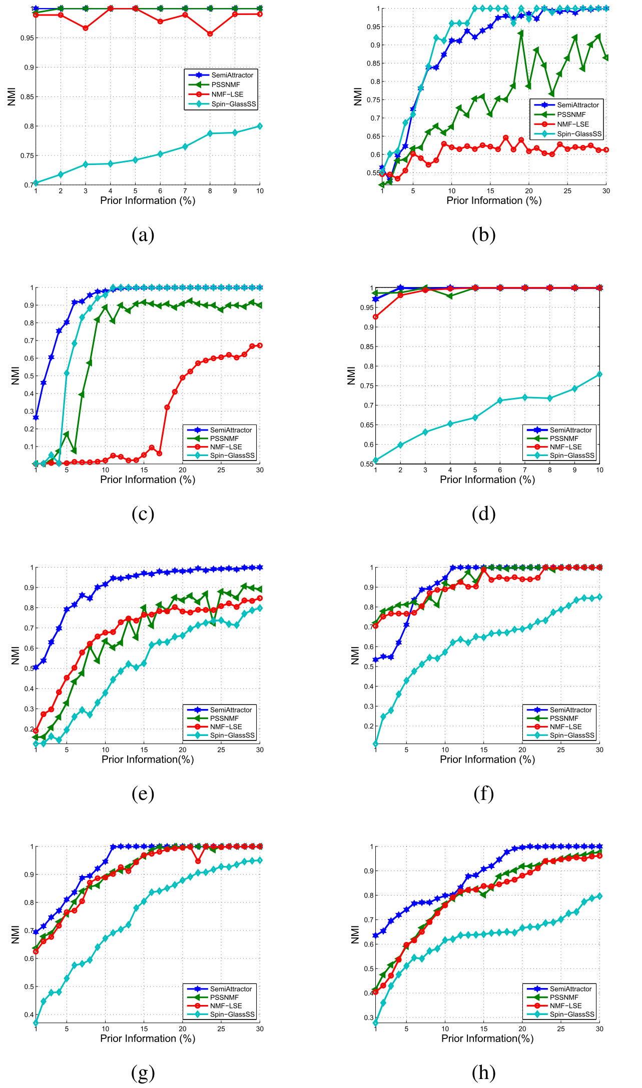
(a)

(b)

FIGURE 4. **The experimental results of the SemiAttractor, PSSNMF, NMF-LSE and Spin-GlassSS on the LFR benchmark networks. (a) LFR $\mu = 0.7$. (b) LFR $\mu = 0.8$.**

To further verify the validity of our algorithm, we conducted experiments on the eight real-world networks listed in TABLE 1. We compare our SemiAttractor algorithm with other three semi-supervised algorithms in their performances on real-world networks, as depicted in FIGURE 5.

From FIGURES 5(a)(d), we see that our algorithm achieve the best community detection results in the Football and Polblogs networks. When the ratio of prior information is 1%, the corresponding NMI values reach 1. From FIGURES 5(c)(e)(g)(h), we see that our algorithm all outperforms the other algorithms on the Enron, Political Events, BlogCatalog and Northeastern networks, regardless of the proportion of prior information. Moreover, when the prior



(a) (b) (c) (d) (e) (f) (g) (h)

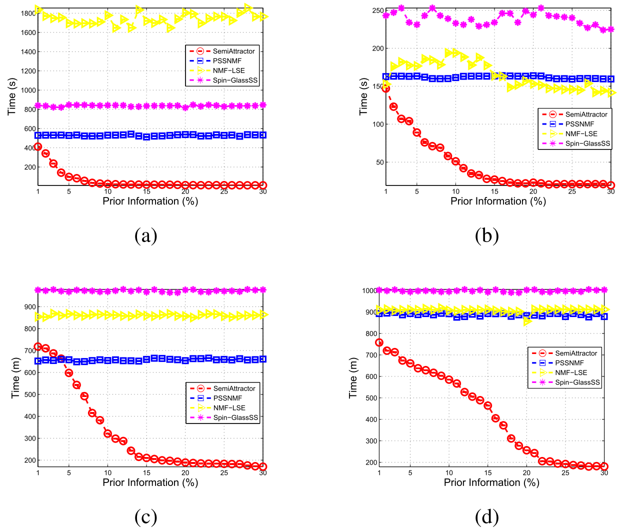
FIGURE 5. The experimental results of SemiAttractor, PSSNMF NMF-LSE and Spin-GlassSS on real-world networks. (a) Football network. (b) Polbooks network. (c) Enron network. (d) Polblogs network. (e) Political events network. (f) Cora network. (g) BlogCatalog network. (h) Northeastern network.

information exceeds 25%, the NMI values reached 1. From FIGURE 5(f), the performance of the SemiAttractor outperforms the other algorithms when the prior information exceeds 6%, when the ratio of prior information is 11%, the corresponding NMI values reach 1. However, from FIGURE 5(b), we see that due to the particularity of the Polbooks network, the effects of the four algorithms appear to be non-monotonic situations with the increase of the proportion of prior information. Nonetheless, on the whole, the NMI values are still gradually increasing, although our SemiAttractor algorithm is inferior to the other algorithms when the prior information is less than 4%. When the proportion of prior information exceeds 26%, the corresponding NMI of our algorithm reaches almost 1, which shows the superiority of the SemiAttractor algorithm over the NMF-LSE and PSSNMF algorithms.

Overall, our SemiAttractor algorithm shows better performance on both artificial and real-world networks. In addition, the number of real communities is given in the PSSNMF, NMF-LSE and Spin-GlassSS algorithms in our experiments. However, our SemiAttractor algorithm does not require the real number of communities. This further illustrates that the SemiAttractor algorithm is very effective.

3) RUNTIME

In order to evaluate the ability of the SemiAttractor algorithm to handle large-scale networks, we chose four real-world networks to compare the running time of the algorithm, Polblogs, Cora, BlogCatlog and Northeastern. FIGURES 6(a)-(d) show the running time of different semi-supervised algorithms on real networks. We can observe that the running time of the SemiAttractor algorithm decreases with the prior information increasing. The number of edges has a significant influence on the SemiAttractor algorithm. The time consumption increases with the increase of edge numbers. Because the time complexity of SemiAttractor algorithm is linear with the number of edges. The time step is reduced through the drive of prior information and network topology, the time consumption is then reduced. While the other three semi-supervised algorithms are not sensitive to the prior information, the relationship with the number of nodes is more obvious. In other words, the amount of priori information cannot effectively reduce the time consumption of the algorithm.



(a) (b) (c) (d)

FIGURE 6. The runtime of the different algorithms. (a) Cora network. (b) Polblogs network. (c) BlogCatalog network. (d) Northeastern network.

V. CONCLUSION

In this paper, we add the prior information into the community detection method based on distance dynamics to generate a semi-supervised community detection algorithm. Specifically, we directly encode the prior information into the adjacency matrix of the network. It is driven by the network topology and prior information. The distance of the links between those nodes with highest similarity first synchronizes and then rapidly decreases to 0. Meanwhile, the distances of the links between those nodes with the highest dissimilarity rapidly increases to 1. The SemiAttractor algorithm shows better performance through testing on both artificial and real-world networks. As the prior information increases, the results of community detection are closer to the real communities.

However, our proposed SemiAttractor algorithm has some drawbacks. First, since the SemiAttractor method relies on the accuracy of the prior information, its ability to discover accurate community structures rapidly degrades as the prior information is perturbed by noise. This will be an interesting topic to study in future work. Second, the SemiAttractor method is a disjointed community detection algorithm, and it cannot find the overlapping structure of the network. How to make our SemiAttractor algorithm applicable for the detection of overlapping communities with prior information is the next question to be considered.

ACKNOWLEDGMENT

The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropol. Res.*, vol. 33, no. 4, pp. 452–473, 1977.

[2] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, p. 845, 2000.

[3] N. F. Johnson *et al.*, "New online ecology of adversarial aggregates: ISIS and beyond," *Science*, vol. 352, no. 6292, pp. 1459–1463, 2016.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, p. 10008, Oct. 2008.

[5] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002.

[6] Z. Dezső and A. L. Barabási, "Halting viruses in scale-free networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 65, no. 5, p. 055103(R), 2002.

[7] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell Syst. Tech. J.*, vol. 49, no. 2, pp. 291–307, 1970.

[8] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 359–392, 1999.

[9] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of Web communities," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2000, pp. 150–160.

[10] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-organization and identification of Web communities," *Computer*, vol. 35, no. 3, pp. 66–70, Mar. 2002.

[11] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, p. 066133, 2003.

[12] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, p. 066111, 2004.

[13] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM J. Res. Develop.*, vol. 17, no. 5, pp. 420–425, 1973.

[14] S. Dongen, "A cluster algorithm for graphs," *Inf. Syst.*, vol. 10, pp. 1–40, 2000.

[15] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained K-means clustering with background knowledge," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 577–584.

[16] S. Zhang, R.-S. Wang, and X.-S. Zhang, "Uncovering fuzzy community structure in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 2, p. 046103, 2007.

[17] X. Ma, L. Gao, X. Yong, and L. Fu, "Semi-supervised clustering algorithm for community structure detection in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 389, no. 1, pp. 187–197, 2010.

[18] Z.-Y. Zhang, "Community structure detection in complex networks with partial background information," *EPL (EuroPhys. Lett.)*, vol. 101, no. 4, p. 48005, 2013.

[19] E. Eaton and R. Mansbach, "A spin-glass model for semi-supervised community detection," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 900–906.

[20] D. Liu, X. Liu, W. Wang, and H. Bai, "Semi-supervised community detection based on discrete potential theory," *Phys. A, Stat. Mech. Appl.*, vol. 416, pp. 173–182, Dec. 2014.

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

[22] 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.

[23] X. Deng, Y. Wen, and Y. Chen, "Highly efficient epidemic spreading model based LPA threshold community detection method," *Neurocomputing*, vol. 210, pp. 3–12, Oct. 2016.

[24] Z.-Y. Zhang, K.-D. Sun, and S.-Q. Wang, "Enhanced community structure detection in complex networks with partial background information," *Sci. Rep.*, vol. 3, Nov. 2013, Art. no. 3241.

[25] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 2, pp. 1118–1123, 2008.

[26] X. Liu, W. Wang, D. He, P. Jiao, D. Jin, and C. V. Cannistraci, "Semi-supervised community detection based on non-negative matrix factorization with node popularity," *Inf. Sci.*, vol. 381, pp. 304–321, Mar. 2017.

[27] X. Shi, H. Lu, Y. He, and S. He, "Community detection in social network with pairwisely constrained symmetric non-negative matrix factorization," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 541–546.

[28] 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, no. 4, p. 046110, 2008.

[29] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, 2006.

[30] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in *Proc. Int. Workshop Link Discovery*, 2005, pp. 36–43.

[31] J. S. Goldstein, "A conflict-cooperation scale for WEIS events data," *J. Conflict Resolution*, vol. 36, no. 2, pp. 369–385, 1992.

[32] J. Shetty and J. Adibi, "The Enron email dataset database schema and brief statistical report," Inf. Sci. Inst., Univ. Sothern California, Los Angeles, CA, USA, Tech. Rep., 2005, vol. 4, no. 1. [Online]. Available:ftp://ftp.isi.edu/sims/philpot/data/enron-mysqldump.sql.gz

[33] G. Forman, "A pitfall and solution in multi-class feature selection for text classification," in *Proc. 21st Int. Conf. Mach. Learn.*, Banff, AB, Canada, 2004, p. 38.

[34] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *J. Stat. Mech., Theory Exp.*, vol. 2005, p. 09008, Sep. 2005.

[35] L. Yang, X. Cao, D. Jin, X. Wang, and D. Meng, "A unified semi-supervised community detection framework using latent space graph regularization," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2585–2598, Nov. 2015.

[36] C. Hennig and B. Hausdorf, *Design of Dissimilarity Measures: A New Dissimilarity Between Species Distribution Areas*. Berlin, Germany: Springer, 2006, pp. 29–37.

[37] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 1, p. 016110, 2006.

[38] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

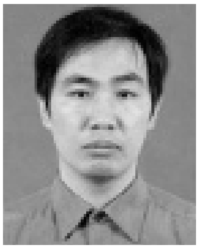
[39] L. Ma, M. Gong, J. Liu, Q. Cai, and L. Jiao, "Multi-level learning based memetic algorithm for community detection," *Appl. Soft Comput.*, vol. 19, pp. 121–133, Jun. 2014.

[40] L. Tang and H. Liu, "Relational learning via latent social dimensions," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 817–826.

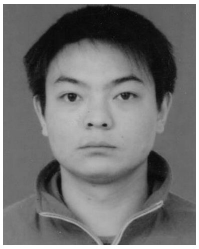
[41] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of Facebook networks," *Phys. A, Stat. Mech. Appl.*, vol. 391, no. 16, pp. 4165–4180, 2011.

[42] S. Chand and S. Mehta, Ą"Community detection using nature inspired algorithm," in *Hybrid Intelligence for Social Networks*, H. Banati, S. Bhattacharyya, A. Mani, and M. Koppen, Eds. Ą Springer, 2017, pp. 47–76.

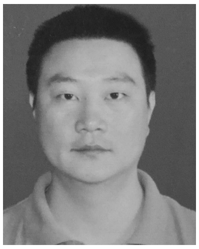
[43] R. Babers, A. E. Hassanien, and N. I. Ghali, "A nature-inspired metaheuristic Lion Optimization Algorithm for community detection," in *Proc. 11th Int. Comput. Eng. Conf. (ICENCO)*, Dec. 2015, pp. 217–222.



LILIN FAN received the B.S. degree in computer science from Henan Normal University in 1994 and the M.S. and Ph.D. degrees in computer science from Southwest Jiaotong University in 2009. He is currently an Associate Professor of computer science with Henan Normal University. His research areas include communication networks and network architecture.



SHENGLI XU received the B.S. degree in computer science and technology from Henan Normal University, Xinxiang, China, in 2015, where he is currently pursuing the M.S. degree in computer technology. His current research interests include machine learning and complex network.



DONG LIU received the B.S. and M.S. degrees in computer science from Zhengzhou University in 2004 and the Ph.D. degree in computer science from Tianjin University in 2013. He is currently an Associate Professor of computer science with Henan Normal University. His research interests include complex network analysis and machine learning.

YAN RU received the B.S. degree in computer science and technology from Henan Normal University, Xinxiang, China, in 2017, where she is currently pursuing the M.S. degree in computer technology. Her current research directions include the community detection in complex network and machine learning.