Fuzzy overlapping community detection based on local random walk and multidimensional scaling

Wenjun Wang\textsuperscript{a,b}, Dong Liu\textsuperscript{a,b,c,*}, Xiao Liu\textsuperscript{a,b}, Lin Pan\textsuperscript{a,b}

\textsuperscript{a} School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
\textsuperscript{b} Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300072, China
\textsuperscript{c} School of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China

**Highlights**

- A novel fuzzy overlapping community detection algorithm is proposed based on data analysis perspective.
- The variant of SRW index is used for measuring the distance between the pairs of nodes for community detection process.
- The network is projected into low-dimensional space by multidimensional scaling method.
- The proposed algorithm is efficient and effective to discover the fuzzy overlapping community structure in networks.

**Abstract**

A fuzzy overlapping community is an important kind of overlapping community in which each node belongs to each community to different extents. It exists in many real networks but how to identify a fuzzy overlapping community is still a challenging task. In this work, the concept of local random walk and a new distance metric are introduced. Based on the new distance measurement, the dissimilarity index between each node of a network is calculated firstly. Then in order to keep the original node distance as much as possible, the network structure is mapped into low-dimensional space by the multidimensional scaling (MDS). Finally, the fuzzy c-means clustering is employed to find fuzzy communities in a network. The experimental results show that the proposed algorithm is effective and efficient to identify the fuzzy overlapping communities in both artificial networks and real-world networks.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, community detection in complex networks has become a very interesting research area. It has been attracting considerable attention crossing many areas from physics, biology, and economics to sociology [1]. The community structure in real networks usually has a specific function such as cycles or pathways in metabolic networks and collections of pages on the same or related topics on the web community [2]. In order to comprehensively understand the function of different networks, enormous efforts have been devoted to develop methods that can extract community structures from complex networks.

Many of the community detection algorithms focus on identifying disjoint communities [3], however, which ignore an important fact that each node may appear in more than one community in real complex networks. In reality, a complex
network is usually composed of several communities with nodes that overlap with each other. For example, a person usually connects to several social groups including family, friends, and colleagues; a researcher may be active in several areas.

The overlapping community can be categorized into two distinct types: crisp and fuzzy [4]. The crisp overlapping community means that each node fully belongs to its associated communities. For example, in social networks, a person is very likely to belong to many communities of different types: colleagues, friends, relatives, etc. But the fuzzy overlapping community refers to those nodes belonging to the communities with different belonging coefficients. For instance, in collaboration networks, a researcher associates with several communities but with different levels of commitment regarding time and resources. In this work, we focus on identifying a fuzzy overlapping community in complex networks.

In the last decades, a few fuzzy overlapping community detection methods have been proposed. Zhang [5] proposed a method based on combining the spectral mapping, fuzzy clustering and optimization of a quality function. The network is mapped into Euclidean space by eigendecomposing a related matrix firstly and then the fuzzy \( c \)-means algorithm is adopted to detect communities. Nepusz [6] modeled the task as a non-linear constrained optimization problem. Consequently, the similarities are optimized via gradient-based constrained optimization methods to make such connected vertices similar and disconnected ones dissimilar. Psorakis [7] designed a probabilistic method based on Bayesian non-negative matrix factorization (NMF), which produces the point estimation for the model parameters through a maximum a posteriori (MAP) scheme. FOG [8] is a stochastic framework. In the FOG framework, an algorithm is introduced for dictating relationships among individuals, groups, and observable interactions as a generative model for link data. In addition, Wu [9] proposed a method to detect a fuzzy overlapping community via the clustering phenomenon of network oscillators. In this method, the phase of an overlapping node quantifies the membership degree by which communities it belongs to. Although some methods for discovering a fuzzy overlapping community have been presented recently, there is still space for improving their performance and universality.

Random walk is a Markov chain which describes the sequence of nodes visited by a random walker, and has been successfully used in community detection [10–14]. Particularly, random walk is used to measure the dissimilarity between two nodes in order to identify community [15–17]. Unfortunately, they are not designed for identifying fuzzy overlapping communities.

In this paper, a fuzzy overlapping community detection method is developed based on a data analysis perspective. The distance matrix describing the distance between each node in a network is established according to the strategy of local random walk. The nodes in the network are mapped into low-dimensional spaces with the help of multidimensional scaling to keep the distance between a pair of nodes as much as possible. It is observed that if the distance between two nodes is sufficiently small, they should be in the same community with a high belonging coefficient. As a result, based on the above observation, the nodes are clustered in low-dimensional space using the fuzzy \( c \)-means. Fig. 1 provides an example to illustrate our proposed method.

The rest of the paper is organized as follows. Section 2 contains the computation of a distance matrix by local random walk, the basic ideas of multidimensional scaling and fuzzy \( c \)-means, and the algorithm with the corresponding strategies. The parameters of the algorithm are described in Section 3. The experimental results are presented in Section 4 to demonstrate the efficient and effective way of the proposed method. Section 5 summarizes the conclusions.
2. Method

Let \( N = \{V, E\} \) denote an unweighted and undirected network, where \( V \) is the set of \( n \) nodes and \( E \) is the set of \( m \) edges. The network structure is determined by \( n \times n \) adjacency matrix \( A \). Each element \( A_{ij} \) of \( A \) is equal to 1 if there is an edge existing between the nodes \( i \) and \( j \), otherwise 0. If there are \( k \) communities, a corresponding \( n \times k \) belonging coefficient matrix \( U \) is defined, in which \( u_{ic} \) represents a measure of the strength of association between node \( i \) and community \( c \). For \( u_{ic} \), the following constraints are enforced to be satisfied, \( 0 \leq u_{ic} \leq 1 \) and \( \sum_{c=1}^{k} u_{ic} = 1 \). Based on the above definitions, the adjacency matrix \( A \) and belonging coefficient matrix \( U \) are the input and output of our method, respectively.

In this method, identifying a fuzzy overlapping community can be divided into three stages: (i) computing a distance matrix of the network by local random walk; (ii) mapping the nodes in the network into low-dimensional space by MDS; and (iii) clustering the nodes into communities by using fuzzy \( c \)-means. The detailed description of this method is given in the sequence.

2.1. Local random walk

Random walk is a Markov chain describing the sequence of nodes visited by a random walker, and used to measure the similarity between pairs of nodes in networks. There are some random-walk-based similarity indices, such as average commute time [18] (ACT) and mean first passage time [15] (MFPT). Saerens et al. [19–21] used the average commute time as a (dis)similarity measure and Zhou [15,16] used the average first passage time instead. However, ACT and SFPT are the methods based on global information with computational complexity \( O(n^2) \), so that they cannot be applied in the networks with a large number of nodes.

Ref. [22] presented a node similarity metric based on local random walk, which has lower computational complexity compared with other random-walk-based similarity indices. Given a random walker starting from node \( x \), \( \pi_{xy}(t) \) denotes the probability that this walker locates at node \( y \) after \( t \) steps. This system evolution equation can be formulated as \( \pi_{xy}(t) = P^t \pi_{xy}(0) \) and its initial system status is \( \pi_{xy}(0) = \delta_{xy} \). The local random walk(LRW) index at time step \( t \) is thus defined as

\[
s_{xy}^{LRW}(t) = \frac{k_x}{2|E|} \cdot \pi_{xy}(t) + \frac{k_x}{2|E|} \cdot \pi_{yx}(t),
\]

where \( k_x \) denotes the degree of node \( x \), and \( |E| \) is the number of links of the networks.

Ref. [22] proposed the SRW index, where the random walker is continuously released from the starting point, resulting in a higher similarity between the target node and the nodes nearby. The equation is defined as:

\[
s_{xy}^{SRW}(t) = \sum_{l=1}^{t} s_{xy}^{LRW}(l).
\]

In this paper, a similar idea of SRW is employed to measure the distances between each pair of nodes in the network. The more similar to the nodes, the smaller the distance will be. Note that the distance between the same node is zero. The distance between node \( x \) and node \( y \) is defined as follows:

\[
d_{xy}(t) = \left\{ \begin{array}{ll} 1 - s_{xy}^{SRW}(t), & x \neq y; \\ 0, & x = y. \end{array} \right.
\]

So the distance matrix \( D = \{d_{ij}\} \) of the network can be calculated based on the above formula. It is obvious that the matrix \( D \) is symmetric. Note that our distance function is a metric if the steps of the random walk are limited. It is easy to prove that \( d_{xy}(t) + d_{xz}(t) \geq d_{xz}(t) \) if \( t \leq |E|/2 \) (see the Appendix for details).

2.2. Multidimensional scaling

Here, in order to facilitate efficient community detection, the multidimensional scaling (MDS) method [23,24] is introduced to map each node in the network into lower-dimensional space. Multidimensional scaling is a general approach which can achieve a lower-dimensional representation of data, while trying to preserve the distances between the data points [25].

Given a distance matrix \( D \in \mathbb{R}^{n \times n} \) as the input of MDS, every entry \( d_{ij} \) denotes the distance between a pair of nodes \( i \) and \( j \) in the network. Let \( H \) be the idempotent centering matrix, such that

\[
H = I - \frac{1}{n} \vec{1} \vec{1}^T,
\]

where \( I \) is the identity matrix, and \( \vec{1} \) is a \( n \)-dimensional column vector with each entry being 1. The inner product matrix obtained from \( D \) can be computed using

\[
\tilde{D} = -\frac{1}{2} H(D \odot D)H.
\]
where $\odot$ denotes the Hadamard product of the matrix. According to the singular value decomposition (SVD) theorem, $\tilde{D} = B^{-1}AB$, where diagonal matrix $A = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Choosing the top $p$ eigenvectors of $\tilde{D}$ as axes of a low-dimensional space and projecting $D$ on them, approximative representations of nodes in $p$-dimensional space can be obtained. Note that $p$ is usually much less than $n$ (from our experiments, $p$ is 2, by which we can obtain reasonable results).

2.3. Fuzzy c-means

To implement the task of fuzzy overlapping community detection, we use fuzzy c-means (FCM) [26,27] to cluster the nodes in $p$-dimensional space. Fuzzy c-means is an unsupervised clustering algorithm in which each point has a certain membership degree of belonging to different clusters. These membership degrees are in the range $[0, 1]$ and indicate the strengths of the association between the nodes and the particular community.

The FCM minimizes an objective function $J_f$,

$$J_f(\tilde{U}, \tilde{C}) = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^f \|x_i - c_j\|^2,$$

where $u_{ij}$ is the membership degree of the $i$-th node to the $j$-th cluster and $d_{ij} = \|x_i - c_j\|$ is the distance between the $i$-th node and the center of the $j$-th cluster. During optimizing $J_f$, the constraint $\sum_{j=1}^{k} u_{ij} = 1$ must be satisfied. The parameter $f$ controls the fuzziness of the algorithm. As $f$ turns out to be larger, the process is more fuzzy. The $c_j$ can be calculated by the following equation:

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^f x_i}{\sum_{i=1}^{n} u_{ij}^f}.$$

The $u_{ij}$ can be calculated via the following equation:

$$u_{ij} = \frac{1}{\sum_{i=1}^{k} (d_{ij}/d_{ij})^{2/(f-1)}}.$$

The $J_f$ can be minimized by iterative optimization with the update of membership degree $u_{ij}$ and the cluster center $c_j$.

2.4. The algorithm and complexity

In this section, the implementation of the proposed method is detailed, and its computational complexity is analyzed. Our method is summarized in Algorithm 1.

The complexity of the algorithm depends on the highest complexity of the three parts involved in the algorithm. In the SRW-Step, the complexity of $t$-step SRW is approximately $O(n\langle k \rangle^t)$ [22], where $\langle k \rangle$ is the average degree of the network and $\langle k \rangle$ is generally much smaller than $n$. In the next step, MDS can be done by Chalmer’s Linear Iteration Algorithm [28] and the complexity is $O(n^2)$. The last step can be achieved by the method proposed by Kolen [29] and its complexity is $O(nkp)$, where $k$ is the number of clusters, and $p$ is the dimension of data points. Therefore, the time complexity of our method is $O(n^2)$.

3. Parameterization of the algorithm

In this section, several important parameters of the algorithm are analyzed, including the number of communities $k$, the number of random walk steps $t$, and the embedding dimension $p$.

3.1. Determining the number of communities

The most important parameter of the method is $k$ which defines the number of communities that the algorithm tries to discover in the network. In other words, the number of communities needs to be known in advance. In practice, however, the number of communities is difficult to obtain due to a lack of ground truth of the network. In order to avoid human intervention for choosing $k$, we adopt the strategy of maximizing the value of the fuzzified modularity [6] to determine the optimal number of communities.

The fuzzified modularity is a variant of Newman Girvan modularity [30], and it can be defined as follows:

$$Q_f = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{k_i k_j}{2m} \right) s_{ij},$$

where $s_{ij}$ is the strength of the link between nodes $i$ and $j$.
**Algorithm 1**: Fuzzy overlapping community detection

**input**: adjacency matrix $A$, the number of random walk steps $t$, the number of community $k$, embedding dimension $p$

**output**: belonging coefficient matrix $U$

**SRW-Step:**

Initialize $\pi$ at time 0

for $i = 1; i \leq t; i++$ do

- Compute $s_{LRW}(i)$ in Eq. (1)

$S^{LRW}(t) \leftarrow S^{LRW}(t) + S^{LRW}(i)$

Compute $D$ in Eq. (3)

**MDS-Step:**

Compute $\tilde{D}$ in Eq. (4) and Eq. (5)

Compute $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ according to the SVD, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$

$\tilde{U} \leftarrow \{V_1, \ldots, V_p\}$, where $V_i$ is the eigenvector corresponding to eigenvalue $\lambda_i$

**FCM-Step:**

Set fuzzifier $f$ and initialize $U$ randomly

while $\Delta > \epsilon$, where $\epsilon$ is a predefined small positive constant do

- Compute each cluster center in Eq. (7)

- Compute all possible distances $d_{ij}$, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$

- Update fuzzy partition matrix in Eq. (8)

Compute $\Delta = \|U_q - U_{q-1}\|$, where $q$ denotes $q$th iteration

return $U$

---

**Fig. 2.** The number of random walk steps in the LFR benchmark. In the LFR benchmark in which fraction of overlapping vertices is 0.5 (see 4.1 for details), the acceptably stable results (Fuzzy Rand Index) can be achieved although the number of random walk steps is not greater than the diameter of the generated network.

where $s_{ij} = \sum_{c=1}^{k} u_{ic}u_{jc}$. Initially, the fuzzified modularity $Q_f$ is computed by setting $k = 2$. Then keep on increasing the value of $k$ until community structure cannot improve the $Q_f$ function. The optimal number of communities $k$ is the one corresponding to the greatest value of fuzzified modularity $Q_f$. It is noteworthy that the time complexity of our algorithm will become $O(kt^2)$ without prior knowledge about the number of communities.

---

**3.2. Determining the number of random walk steps**

Another question remaining to be considered is “how many random walk steps do we need to ensure the reasonableness of our algorithm?”. So far, we do not have a mathematical conclusion. The experiments on the LFR benchmark and five real-world networks used in this work suggest that the acceptable and reasonable results can be obtained although the number of random walk steps is smaller than the diameter of the network. So, we set the diameter of the network as the number of random walk steps $t$ in the following experiments (see Figs. 2 and 3).

---

**3.3. Determining the embedding dimension**

The embedding dimension $p$ is another important parameter which defines the dimension of space embedding the network. It may seem difficult to select the appropriate value for this parameter. However, in this work, an interesting phenomenon is that the parameter is not crucial to the final result of the method. For reducing the computational complexity of the FCM-step of the algorithm, the parameter is taken as 2 for the following experiments.
Fig. 3. The number of random walk steps in real-world networks. In the five real-world networks (see 4.2 for details), the acceptably stable results (modularity) can be obtained although the number of random walk steps is smaller than the diameter of the network.

4. Experiments

To evaluate the performance of our algorithm on discovering fuzzy overlapping community structures, for artificial networks, the fuzzy rand index measure [31] is used to quantitatively compare the known fuzzy partition with the fuzzy partition found by the state-of-the-art algorithms. A higher value of this measure indicates that the quality of a detected partition is better. As to real-world networks, since we do not know the ground truth of community structures, the modularity measure [30] is employed to assess the quality of a partition. In the following experiments, it is assumed that the number of communities is unknown in advance.

4.1. Artificial networks

The LFR benchmark network [32,33] is an artificial network for community detection, which is claimed to possess some basic statistical properties found in real networks, such as heterogeneous distributions of degree and community size. It also allows communities to overlap. Many parameters are involved to specify properties of generated networks in this benchmark: $N$ (number of nodes), $\langle k \rangle$ (average degree), $k_{\text{max}}$ (maximum degree), $c_{\text{min}}$ and $c_{\text{max}}$ (minimum and maximum community size), $\tau_1$ (exponent of power-law distribution of nodes degree), $\tau_2$ (exponent of power-law distribution of community sizes), $\mu$ (mixing parameter), $o_m$ (number of communities each overlapping node belongs to), and $o_n$ (number of overlapping nodes).

However, the LFR benchmark cannot be directly applied to fuzzy overlapping community detection because it is not able to generate fuzzy overlapping community structures. According to the fuzzy generator proposed by Gregory [4], we convert crisp overlapping communities to fuzzy form by adding a random belonging coefficient (to each occurrence) of each item.

To evaluate the effectiveness of our method, we compare it to other well-known fuzzy overlapping community detection algorithms, including Fuzzyclust [6] and NMF [34]. The Fuzzyclust method can determine the number of communities using a fuzzified variant of the modularity function, just like the strategy for determining the number of communities in this work. The NMF algorithm can also detect community structure without the prior knowledge of the number of communities. As far as we know, two referenced algorithms have the time complexity $O(kdn^2)$ and $O(kn^2)$ respectively, where $d$ is the number of iterations leading to convergence. Therefore, the time complexity of our algorithm is equal to or lower than that of the two algorithms mentioned above.

In Fig. 4, we generate benchmark networks with two values of the mixing parameter ($\mu \in \{0.1, 0.03\}$) and the average degree ($\langle k \rangle \in \{12, 24\}$), and the fraction of overlapping nodes ($o_m/N$) ranging from 0.02 to 1. The rest of the parameters are set by $N = 500$, $k_{\text{max}} = \langle k \rangle \times 2.5$, $c_{\text{min}} = \langle k \rangle \times 2/3$, $c_{\text{max}} = c_{\text{min}} \times 5$, $\tau_1 = 2$, $\tau_2 = 1$, and $o_m = 2$. It is worth noting that determining the parameters of our algorithm is the key problem. In this part, we set $p = 2$, $t$ as the approximate diameter of the generated network.

From Fig. 4, it can be concluded that our algorithm by and large performs quite well and its accuracy is better than that of the other two algorithms. However, it is slightly worse than that of Fuzzyclust when the fraction of overlapping nodes is relatively small. As the overlapping increases, the performance of our algorithm is significantly enhanced.

4.2. Real-world networks

The popular modularity is to measure the quality of a disjoint partition of the network, but not to be directly applied to measure the fuzzy partition. For the purpose of the experiment, we get rid of the fuzzy overlapping by assigning a node to a single community, which has the maximum degree of membership. To evaluate the effectiveness of our method, we compared our method against the extremal optimization (EO) [35] and Louvain method [36] by testing them on a variety of popular real network datasets (see Table 1).
Fig. 4. Results (Fuzzy Rand Index) of fuzzy algorithms on the networks with fuzzy overlapping. The algorithms include Fuzzyclust, NMF and our method. Each point in the graph represents the average over 100 runs on randomly generated LFR benchmark networks with the given parameters.

### Table 1

<table>
<thead>
<tr>
<th>Network</th>
<th>n</th>
<th>m</th>
<th>(k)</th>
<th>Diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karate [37]</td>
<td>34</td>
<td>78</td>
<td>2.3</td>
<td>5</td>
</tr>
<tr>
<td>Dolphins [38]</td>
<td>62</td>
<td>159</td>
<td>5.1</td>
<td>8</td>
</tr>
<tr>
<td>Lesmis [39]</td>
<td>77</td>
<td>254</td>
<td>6.6</td>
<td>5</td>
</tr>
<tr>
<td>Football [40]</td>
<td>115</td>
<td>613</td>
<td>10.6</td>
<td>4</td>
</tr>
<tr>
<td>PGP [41]</td>
<td>10680</td>
<td>24316</td>
<td>4.5</td>
<td>24</td>
</tr>
</tbody>
</table>

### Table 2

Maximum modularity of EO, Louvain and our method for different real-world networks. Another metric is the number of communities found at the configuration with maximum modularity.

<table>
<thead>
<tr>
<th>Network</th>
<th>Q_EO</th>
<th>C_EO</th>
<th>Q_Louvain</th>
<th>C_Louvain</th>
<th>Q_our method</th>
<th>C_our method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karate</td>
<td>0.41</td>
<td>4</td>
<td>0.42</td>
<td>4</td>
<td>0.38</td>
<td>4</td>
</tr>
<tr>
<td>Dolphins</td>
<td>0.51</td>
<td>4</td>
<td>0.52</td>
<td>5</td>
<td>0.48</td>
<td>4</td>
</tr>
<tr>
<td>Lesmis</td>
<td>0.53</td>
<td>5</td>
<td>0.57</td>
<td>6</td>
<td>0.55</td>
<td>6</td>
</tr>
<tr>
<td>Football</td>
<td>0.58</td>
<td>8</td>
<td>0.60</td>
<td>10</td>
<td>0.54</td>
<td>11</td>
</tr>
<tr>
<td>PGP</td>
<td>0.85</td>
<td>365</td>
<td>0.87</td>
<td>372</td>
<td>0.85</td>
<td>343</td>
</tr>
</tbody>
</table>

The results from five real networks are shown in Table 2. As the data indicated, our method performs competitively, even though our consideration of maximizing modularity is absent, compared with the EO and Louvain method. In addition, it has the advantage of providing fuzzy partitioning solutions, i.e. the node belonging coefficient of each community.

### 5. Conclusions

In this paper, a fuzzy overlapping community detection framework is presented based on a data analysis perspective. Based on the distance calculated by local random walk, it is assumed that a pair of nodes with smaller distance are more likely to be grouped into the same community. Because the nodes belonging to the same community can be mapped into the low-dimensional space with close positions by MDS, fuzzy \( c \)-means is used to partition the network structure. As a result, every node is allowed to belong to multiple communities with different belonging coefficients.

There are many similarity (dissimilarity) measures to represent the proximity between two nodes [42]. These similarity measures can be divided into two categories: local measures and global measures. The local measures are the node-dependent measures that require only the information about node degree and the nearest neighborhood. It is easy to
calculate the proximity between two nodes but less accurate. The global measures belong to the path-dependent measures that require global knowledge of the network topology. Comparing with the local similarity measures, global measures can provide much more accurate results but with higher time complexity. To balance work load and improve the accuracy, the variant of SRW is employed as a dissimilarity measure in this work, which needs a little bit more information than local measures but can provide a competitively accurate result compared with the global measures.

In addition, in contrast with previous works about discovering a fuzzy overlapping community (see Section 1 for details), our algorithm has three advantages: (i) the time complexity of the algorithm is competitive with other fuzzy overlapping community detection algorithms mentioned above, (ii) the algorithm can obtain reasonable results, and (iii) the method is given based on a new idea about data analysis.

In conclusion, this method has been tested using a variety of networks, both real-world and artificial. The experimental results show that it can efficiently discover the fuzzy overlapping community structure of complex networks. Moreover, the modules of MDS-Step and FCM-Step can be easily replaced if more effective dimensionality reduction and fuzzy clustering methods are developed with the advance of technologies in the fields of pattern recognition and machine learning.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments and suggestions. This work was supported by the National Science-technology Support Plan Project of China (No. 2013BAK02B06, and No. 2012BAK03B06), the National Natural Science Foundation of China (No. 91224009) and the National Science Plan Project of the Education Department, Henan Province (No. 2010B520014).

Appendix

The aim of this section is to prove that the distance function \( d_{xy}(t) \) is a metric. Note that here we only focus on the few-step random walk (i.e., \( t \ll m \)) avoiding the stationary state.

To prove that our distance function \( d_{xy}(t) \) described in Eq. (3) is metric with limited steps of random walk, given that \( d_{xy}(t) \geq 0 \) and \( d_{xy}(t) = d_{yx}(t) \) hold, we only need to show the following triangle inequality

\[
d_{xy}(t) + d_{xz}(t) \geq d_{xz}(t), \quad (A.1)
\]

if \( t < \frac{|E|}{2} \), where \( t \) denotes the steps of random walk and \( |E| \) is the number of edges of the networks. If \( x = y \) or \( y = z \), it is obvious that our distance function satisfies Eq. (A.1). If \( x \neq y \) and \( y \neq z \), according to Eqs. (2) and (1), the above equation can be written as

\[
1 - s_{SRW}^{xy}(t) + 1 - s_{SRW}^{yz}(t) \geq 1 - s_{SRW}^{xz}(t)
\]

\[
1 + s_{SRW}^{xz}(t) \geq s_{SRW}^{xy}(t) + s_{SRW}^{yz}(t)
\]

\[
1 + \sum_{l=1}^{t} s_{SRW}^{xz}(l) \geq \sum_{l=1}^{t} s_{SRW}^{xy}(l) + \sum_{l=1}^{t} s_{SRW}^{yz}(l).
\]

Then the equation is equivalent to

\[
1 + \frac{k_x}{2|E|} \sum_{l=1}^{t} \pi_{xz}(l) + \frac{k_z}{2|E|} \sum_{l=1}^{t} \pi_{zx}(l) \geq \frac{k_x}{2|E|} \sum_{l=1}^{t} \pi_{xy}(l) + \frac{k_y}{2|E|} \sum_{l=1}^{t} \pi_{yx}(l) + \frac{k_y}{2|E|} \sum_{l=1}^{t} \pi_{yz}(l) + \frac{k_y}{2|E|} \sum_{l=1}^{t} \pi_{zy}(l) + \frac{k_z}{2|E|} \sum_{l=1}^{t} \pi_{yz}(l). \quad (A.2)
\]

From Eq. (A.2), it can be seen that \( (A.2) \) can be satisfied if the value of its right part is less than or equal to 1. Without loss of generality, the first term on the right can be estimated as follows:

\[
\frac{k_x}{2|E|} \sum_{l=1}^{t} \pi_{xy}(l) < \frac{k_x}{2|E|} \cdot t \cdot \max_{r \in \{1, 2, \ldots, t\}} \{ \pi_{xy}(l) \}
\]

\[
< \frac{k_x}{2|E|} \cdot t \cdot \frac{1}{k_x}
\]

\[
< \frac{t}{2|E|}.
\]

It can be clearly seen that the first term will be less than \( \frac{1}{4} \) if \( t < |E|/2 \). Hence, it can be obtained that the distance function \( d_{xy}(t) \) is a metric if \( t < |E|/2 \).
References


