Physica A 447 (2016) 508–519

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A general method of community detection by identifying community centers with affinity propagation

Wei-Feng Guo, Shao-Wu Zhang *

Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an, 710072, China

HIGHLIGHTS

- A general method suitable for unweighted, weighted, undirected, directed and signed network.
- Construct the dissimilarity distance matrix with different measures.
- Extract a candidate center set of community with AP algorithm.
- Determine the community by selecting the center subset to maximum the modularity.

ARTICLE INFO

Article history:
Received 28 July 2015
Received in revised form 29 October 2015
Available online 29 December 2015

Keywords:
Dissimilarity distance matrix
Centers
Community
Modularity
Affinity propagation

ABSTRACT

Detection of community structures is beneficial to analyzing the structures and properties of networks. It is of theoretical interest and practical significance in modern science. So far, a large number of algorithms have been proposed to detect community structures in complex networks, but most of them are suitable for a specific network structure. In this paper, a novel method (called CDMIC) is proposed to detect the communities in un-weighted, weighted, un-directed, directed and signed networks by constructing a dissimilarity distance matrix of network and identifying community centers with maximizing modularity. For a given network, we first estimate the distance between all pairs of nodes for constructing the dissimilarity distance matrix of the network. Then, this distance matrix is input to the affinity propagation (AP) algorithm to extract a candidate center set of community. Thirdly, we rank these centers in descending order according to the sum of their availability and responsibility. Finally, we determine the community structure by selecting the center subset from the candidate center set in an incremental manner to make the modularity maximization. On three real-world networks and some synthetic networks, experimental results show that our CDMIC method has higher performance in terms of classification accuracy and normalized mutual information (NMI), and ability to tolerate the resolution limitation.

1. Introduction

Many complex systems in nature and society can be commonly modeled as complex networks or graphs with nodes representing individuals or organizations and edges representing the interactions among these nodes [1]. It has been shown that many real world networks (e.g. social networks, biological networks and the Internet) have a structure of modules or

* Corresponding author.
E-mail address: zhangsw@nwpu.edu.cn (S.-W. Zhang).

http://dx.doi.org/10.1016/j.physa.2015.12.037
0378-4371/© 2015 Elsevier B.V. All rights reserved.

communities which are characterized by the sub-graphs of densely connected nodes and sparsely connected with other parts of the networks. Detecting the community structures is beneficial to analyzing the structures and properties of networks [2].

So far, many approaches have been introduced to detect the communities of complex networks, which can be roughly categorized into positive network community detection and signed network community detection according to the characteristics of network structure. Positive network communities are defined as the groups of nodes in which the links in intra-group are dense but the links in inter-groups are sparse [2]. Signed network communities are defined as the groups of nodes in which not only the positive links in intra-group but also the negative links inter-groups are dense [3].

For the methods of detecting the positive network communities, there are two classical algorithms, i.e., spectral bisection [4] and Kernighan–Lin [5]. Spectral bisection algorithm uses the basis of eigenvectors of Laplacian matrix of a graph to find the optimal cuts of the networks, which has been shown to be a NP-complete problem [6]. Kernighan–Lin algorithm divides the networks according to the optimization of the number of intra- and inter-communities edges using a greedy algorithm, which is sensitive to the initial partition. The change in the order of an initial partition may significantly alter the detecting results. To address the issues of the two classical methods [7–11], several algorithms based on optimizing the modularity (represented by Q function) are developed to detect the community structures of complex networks, especially for weighted networks and directed networks [12–15]. However, the existing methods based on modularity maximization suffer from two major limitations. One is that the maximization of Q is an NP-hard problem, though a few of the optimization techniques such as simulated annealing [16] and extreme optimization [10] are introduced to obtain the suboptimal solutions. Another is resolution limit so-called that it cannot detect the communities whose node number is smaller than a predefined threshold [17]. Thus, several algorithms have been proposed to alleviate the resolution limit by redefining the modularity function or adding weights to the edges [18,19].

The methods of detecting positive network community cannot be suitable for detecting the community structure of signed networks with positive links and negative links. Thus, various approaches have been proposed to detect the communities of signed networks by designing the improved modularity function, adopting an agent-based heuristic strategy and other strategies [3,15,20]. Although above-mentioned algorithms are suitable for the undirected networks, they cannot be extended to directed networks.

In this paper, we will introduce a novel method (called CDMIC) to detect the community structures of weighted, un-weighted, directed, un-directed and signed networks. The key strategy of our CDMIC method is to identify the community centers such that the modularity for the network partition is maximization. By using different similarity measures, CDMIC estimates the similarity between pair of nodes in the network, transforming it into dissimilarity by a decreasing function to obtain a distance matrix of the network. Then we use affinity propagation (AP) algorithm to extract the candidate center set of community, ranking these centers in descending order according to the sum of their availability and responsibility. By selecting the center subset from the candidate center set in an incremental manner to partition the network, and calculating their corresponding modularity, we choose the partition of modularity maximization as the final result of community detection. On three real-world networks and some synthetic networks, our CDMIC method shows higher performance in terms of classification accuracy and normalized mutual information (NMI), and also has strong robustness and ability to tolerate the resolution limitation.

This paper is organized as follows. In Section 2, we explain some crucial concepts, and briefly introduce the affinity propagation algorithm. In Section 3, we describe our novel CDICM method for detecting the community structures in detail. Experimental results of some synthetic networks and three real-world networks are shown in Section 4. In Section 5, we discuss the effects of similarity measures, different modularity measures, and dissimilarity distance matrix. Finally, our conclusions are presented in Section 6.

2. Basic concepts

2.1. Similarity measure

Similarity measures play an important role in the research of complex networks. Employing appropriate similarity measure to grasp as much network structure information as possible can help to accurately detect the community structures of a network. In Table 1, we list the ten similarity measures [21], which are often used to measure the similarity between any two nodes or two links in different networks. The measures of Common Neighbors (CN), Salton, Jaccard, Sørenson, Hub Promoted (HP), Hub Depressed (HD), Leicht–Holm–Newman (LHN), Preferential Attachment (PA) and Adamic–Adar (AA) are based on the local structural information (i.e. neighborhood information). In addition, the first seven measures, from CN to LHN, only differ in the denominator. If the investigated network simultaneously has large clustering coefficient and large degree heterogeneity, there are significant differences among those seven measures [21]. PA is a proximity measure and often used to quantify the functional significance of edges subject to various network-based dynamics, which does not require information on the neighborhood of each node [21]. AA refines the simple counting of common neighbors by assigning the less connected neighbors more weight [21,22]. Assuming that each transmitter has a unit of resource, and equally distribute it between all its neighbors, then resource allocation (RA) index can be defined as the amount of resource v_j received from v_i, which works well on the networks with large clustering coefficient, high degree heterogeneity and absence of a strongly assortative linking pattern [21].

Table 1
Definition of ten similarity measures.

Measures	Definition	Measures	Definition
CN	$s_{ij} = \left\|\Gamma(v_i) \cap \Gamma(v_j)\right\|$	HD	$s_{ij} = \frac{\left\|\Gamma(v_i)\cap\Gamma(v_j)\right\|}{\max\{k(v_i),k(v_j)\}}$
Salton	$s_{ij} = \frac{\left\|\Gamma(v_i)\cap\Gamma(v_j)\right\|}{\sqrt{k(v_i)\times k(v_j)}}$	LHN	$s_{ij} = \frac{\left\|\Gamma(v_i)\cap\Gamma(v_j)\right\|}{k(v_i)\times k(v_j)}$
Jaccard	$s_{ij} = \frac{\left\|\Gamma(v_i)\cap\Gamma(v_j)\right\|}{\left\|\Gamma(v_i)\cup\Gamma(v_j)\right\|}$	PA	$s_{ij} = k(v_i) \times k(v_j)$
Sørenson	$s_{ij} = \frac{2\left\|\Gamma(v_i)\cap\Gamma(v_j)\right\|}{k(v_i)+k(v_j)}$	AA	$s_{ij} = \sum_{z\in\Gamma(v_i)\cap\Gamma(v_j)} \frac{1}{\log k(z)}$
HP	$s_{ij} = \frac{\left\|\Gamma(v_i)\cap\Gamma(v_j)\right\|}{\min\{k(v_i),k(v_j)\}}$	RA	$s_{ij} = \sum_{z\in\Gamma(v_i)\cap\Gamma(v_j)} \frac{1}{k(z)}$

2.2. Modularity

In order to evaluate the community structure of a network, Newman and Girvan [7] proposed the following modularity Q_u as a measure of network partition.

$$Q_u = \sum_i (e_{ii} - a_i^2) \tag{1}$$

where e_{ii} is the fraction of edges belonging to community i in the all edges of the network, and a_i is the fraction of edges that connect to nodes in community i. However above modularity measure can only tackle the community structure of unweighted networks. Thus, Newman [8] proposed an extended modularity Q_e for measuring weighted networks, which is defined as

$$Q_e = \frac{1}{2m}\sum_{ij}\left(S_{ij} - \frac{s_i s_j}{2m}\right)\delta(C_i, C_j) \tag{2}$$

where S_{ij} is the edge weight between node i and j, $s_i = \sum_j S_{ij}$ is the weight sum of edges attached to node i, C_i is the community to which node i is assigned, the δ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise, and $m = \frac{1}{2}\sum_{ij} S_{ij}$ is the weight sum of edges in network G.

In the case of directed networks, Leicht [12] proposed the following modularity Q_d as a measure of network partition.

$$Q_d = \frac{1}{m}\sum_{ij}\left(S_{ij} - \frac{s_i^{in} s_j^{out}}{m}\right)\delta(C_i, C_j) \tag{3}$$

where s_i^{in} and s_j^{out} are the in- and out-weight sum of the node, C_i is the community to which node i is assigned.

In the case of signed networks, Gómez [20] proposed the following modularity Q_s as a measure of network partition.

$$Q_s = \frac{1}{2m^+ + 2m^-}\sum_{ij}\left(S_{ij} - \left(\frac{s_i^+ s_j^+}{2m^+} - \frac{s_i^- s_j^-}{2m^-}\right)\right)\delta(C_i, C_j) \tag{4}$$

where S_{ij} is the weight connecting node i and node j for the signed network; $s_i^+ = \sum_j \max\{0, s_{ij}\}$ denotes the weight sum of positive edges of node i, and $s_i^- = \sum_j \max\{0, -s_{ij}\}$ denotes the weight sum of negative edges of node i; $2m^+$ and $2m^-$ are weight sum of the positive and negative edges attached to nodes, respectively.

2.3. Affinity propagation algorithm

Affinity Propagation (AP) [23] algorithm takes as input a collection of real-valued distance between data points, and exchange the real-valued information between data points until a high-quality set of centers and corresponding clusters gradually emerges. Let $V = \{v_1, v_2, \ldots, v_N\}$ be a set of data points and also let $\sigma(v_i)$ is the index of the nearest community center associated to data point v_i, then the goal is to find the mapping σ maximizing the functional $E(\sigma)$ defined as [24]

$$E[\sigma] = \sum_{i=1}^{N} s(v_i, v_{\sigma(v_i)}) \tag{5}$$

where $s(v_i, v_j)$ reflect the similarity between the nodes v_i and v_j. In this paper, we set $s(v_i, v_j)$ to be the negative value of the distance between nodes v_i and v_j. The initialized values of $s(v_i, v_i)$ is set equally each other as the median of the input similarities.

The resolution of the optimization problem defined by Eq. (5) is achieved by a message passing algorithm, considering two types of messages: the“responsibility” $r(v_i, v_k)$ and the “availability” $a(v_i, v_k)$. $r(v_i, v_k)$ reflects the accumulated evidence for how well-suited v_k is to serve as a community center for node v_i, taking into account other potential centers for node v_i. $a(v_i, v_k)$ reflects the accumulated evidence for how appropriate it would be for node v_i to choose node v_k as its center, taking into account the support from other nodes that node v_k should be a center. At any node during affinity propagation, availabilities and responsibilities can be combined to identify centers.

To begin with, the availabilities and responsibilities are initialized as $a(v_i, v_k) = 0,\ r(v_i, v_k) = 0$. The responsibility and availability are computed iteratively by using the following rules:

$$r(v_i, v_k) \leftarrow s(v_i, v_k) - \max_{k' \neq k}\{a(v_i, v_{k'}) + s(v_i, v_{k'})\} \tag{6}$$

$$a(v_i, v_k) \leftarrow \min\left\{0, r(v_k, v_k) + \sum_{i' \notin \{i,k\}} \max\{0, r(v_{r'}, v_k)\}\right\} \tag{7}$$

$$a(v_k, v_k) \leftarrow \sum_{i' \neq k} \max\{0, r(v_{i'}, v_k)\}. \tag{8}$$

Index $\sigma(v_i)$ of community center associated to v_i is finally defined as [24]

$$\sigma(v_i) = \arg\max_k\{a(v_i, v_k) + r(v_i, v_k),\ k = 1, 2, \ldots, N\}. \tag{9}$$

The algorithm is stopped after a maximal number of iterations or when the centers did not change for a given number of iterations.

3. CDMIC method

In this section, we will describe our CDMIC algorithm in detail. CDMIC algorithm consists of three main steps: (i) Constructing the dissimilarity distance matrix for a given network, (ii) Identifying the candidate center set of the community with AP algorithm, (iii) Computing the modularity of the communities corresponding candidate centers to obtain the final community partition with modularity maximization. Fig. 1 is the schematic diagram of our CDMIC algorithm which is described in detail as follows.

3.1. Constructing the dissimilarity distance matrix

For weighted networks, the weight among edges is looked as the similarity. For un-weighted network including the directed and un-directed network, signed network, the similarity can be calculated by using one of the similarity measures in Table 1. Next, we compute the dissimilarity ρ_{ij} of edge e_{ij} linking node v_i and v_j by the following formula.

$$\rho_{ij} = f(s_{ij}) \tag{10}$$

where s_{ij} is the similarity or strength linking nodes v_i and v_j, and $f(x)$ is a decreasing function, e.g., $f(x) = \exp(-\alpha x)$ or $f(x) = \beta/x$.

Suppose that the distance d_{ij} between nodes v_i and v_j is defined as the sum of dissimilarity along the shortest path L_{ij} between nodes v_i and v_j, then, we can construct the dissimilarity distance matrix $D = [d_{ij}]_{n\times n}$ for an n-node network G. In the directed network, if there is no shortest path between two nodes, the distance is set to infinite. The dissimilarity distance matrix is used as the input of AP algorithm to identify the candidate center set of community.

3.2. Identifying the candidate center set of community

The distance matrix D is taken as the input of AP algorithm to obtain the initial center set $\{v_{\sigma(v_1)}, v_{\sigma(v_2)}, \ldots, v_{\sigma(v_m)}\}$ of a network. According to the sum of the availability and responsibility of these initial centers, we rank the centers in descending order to form a candidate center set $Z = \{z_1, z_2, \ldots, z_\lambda, \ldots, z_m\}$.

3.3. Obtaining the final community partition with modularity maximization

We select the center subset from the candidate center set Z in an incremental manner to partition the network, and calculate their corresponding modularity. That is, first select z_1 as the community center forming a community C_1, and also calculate its modularity Q_1 value; then, select z_1, z_2 as the community centers and assign each member to its closest center forming two communities C'_1, C'_2, i.e., $\{v \in C'_1 \text{ if } D(v, z_1) \leq D(v, z_2)\}$, where D is the dissimilarity distance matrix, and calculate the modularity Q_2 value of this network partition. In the situation of that the distances of one node with two or more community centers are equal, we assign this node to the community center with higher value of the sum of the availability and responsibility. Repeat this process until the network is divided into m community and also calculate its modularity Q_m. Comparing these modularity values, we choose the partition corresponding to maximum Q_λ value as the final result of community detection, that is, if Q_λ is maximum among all the modularity values, and then the final community set is $C^* = \{C^*_1, C^*_2, \ldots, C^*_\lambda\}$.

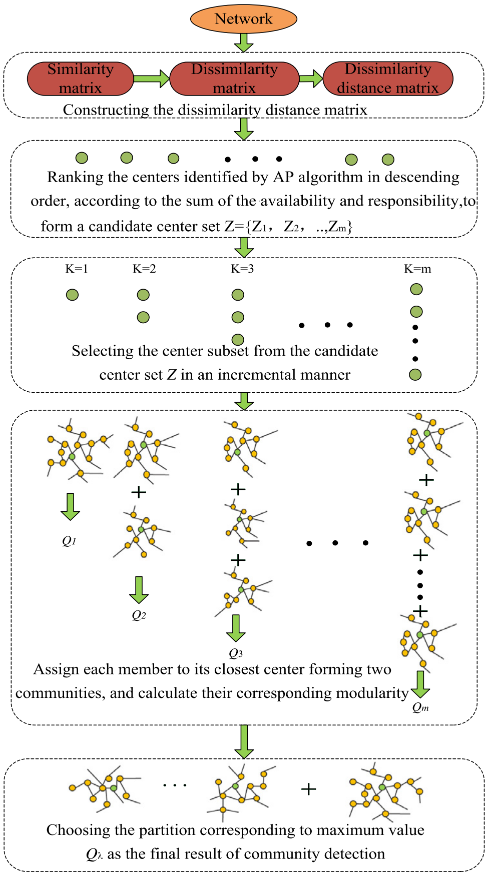

Network
Similarity matrix
Dissimilarity matrix
Dissimilarity distance matrix
Constructing the dissimilarity distance matrix
Ranking the centers identified by AP algorithm in descending order, according to the sum of the availability and responsibility,to form a candidate center set Z={Z1, Z2, ..,Zm}
K=1
K=2
K=3
K=m
Selecting the center subset from the candidate center set Z in an incremental manner
Q1
Q2
Q3
Qm
Assign each member to its closest center forming two communities, and calculate their corresponding modularity
Choosing the partition corresponding to maximum value Qλ as the final result of community detection

Fig. 1. Schematic diagram of CDMIC algorithm showing the process of detecting the network communities.

3.4. Computational complexity

The computational complexity of our CDMIC method contains four parts. For the phase of transforming the un-weight network into the weighted network, it needs to scan all the edges by using the similarity measures. Thus the maximum complexity is in the order of $O(E)$, where E is the total number of edges. In the phase of constructing the distance matrix by using the shortest path algorithm, $O(N^3)$ times should be investigated in the worst case, where N is the total number of nodes. In the phase of forming the initial centers, using the AP algorithm will result in a complexity $O(N^2)$. In the phase of computing the modularity, we need to repeat $O(m)$ times $(m < N)$ for obtaining the maximum modularity. Therefore, the overall complexity of our CDMIC approach is $O(N^3) + O(E) + O(N^2) + O(m)$, which could be considered as $O(N^3)$.

4. Experimental results

The detecting community algorithm proposed in this paper is implemented using MATLAB programming language running on a PC with 2.67 GHz processor, 4 GB memory and Win7 operating system. The computer-generated networks, i.e., synthetic networks [2,25] and signed networks [3] and three classic real-world networks, i.e., glossary network [26], football network [2] and collaboration network [27], are used to evaluate the performance of our CDMIC method in the following text. In supplementary material, we give the results of Zachary network [28], Dolphin network [29] with CDMIC method (see Appendix A).

4.1. Synthetic network

To test the performance of our CDMIC method on networks with varying degrees of community structure, we generate a set of un-weighted synthetic networks with known community structure by using the planted l-partition model [25], which partitions a network with $n = g * l$ nodes in l communities with g nodes each. Nodes of intra-community are linked with a probability p_{in}, whereas nodes of inter-community are linked with a probability p_{out}. Because the probabilities p_{in} and p_{out} are not independent, the parameters $z_{in} = p_{in}(g-1)$ and $z_{out} = p_{out}g(l-1)$ are commonly used to generate the linked edges of intra- and inter-communities, respectively. In this paper, we fix $l = 4$, $g = 32$ and the average total degree $\langle k \rangle = 16$ to

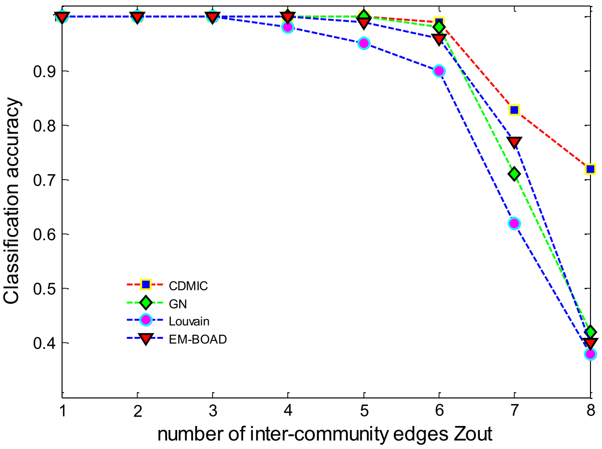


Fig. 2. The classification accuracy of four algorithms on the computer-generated networks at different z_{out}. The accuracy of each point is an average over 100 realization of the networks. Use Q_u to evaluate the quality of network partition. For $z_{out} = 1, 2, \dots, 6$, use CN index to measure the similarity, for $z_{out} = 7, 8$, Jaccard index.

generate a set of synthetic networks, which were originally designed by Girvan and Newman [2] and have been considered as the benchmark to test the performance of community detection algorithms.

The classification accuracy defined as the fraction of nodes that are classified into their correct community is used to evaluate the performance of different community detection algorithms. The average results of 100 networks for each z_{out} are shown in Fig. 2, from which we can see that the performance of our CDMIC is the best, especially, when $z_{out} > 6$, the accuracy of our CDMIC is far higher than that of GN [7], Louvain [30] and EM-BOAD algorithms [1].

4.2. Signed network

Signed network is a special network (e.g. social network) which consists of both positive and negative links. To evaluate the performance of our CDMIC method, we compare it with the existing state-of-the-art algorithms (i.e., EA_{HC}-SN and CSA_{HC}-SN) on the signed networks used in EA_{HC}-SN and CSA_{HC}-SN algorithms [15]. A randomly generated signed network is labeled as $SN(C, [n_1, n_2, \dots, n_C], P_N, P_-, P_+)$, where C is the number of communities, $[n_1, n_2, \dots, n_C]$ is the number of nodes in each community, P_N is the ratio of the number of negative edges in inter-community to that of positive edges in intra-community, P_- is the ratio of the number of negative edges in intra-community to that of positive edges in intra-community, and P_+ is the ratio of the number of positive edges in inter-community to that of negative edges in inter-community. These parameters are used to control community structure.

In the process of computing the similarity between two nodes, the similarity value is negative if the link between two nodes is negative. We select the exponential function to obtain the dissimilarity, and set parameter $\alpha = 100$. We also adopt the following normalized mutual information (NMI) index [31] to estimate the similarity between the true partitions and the detected ones.

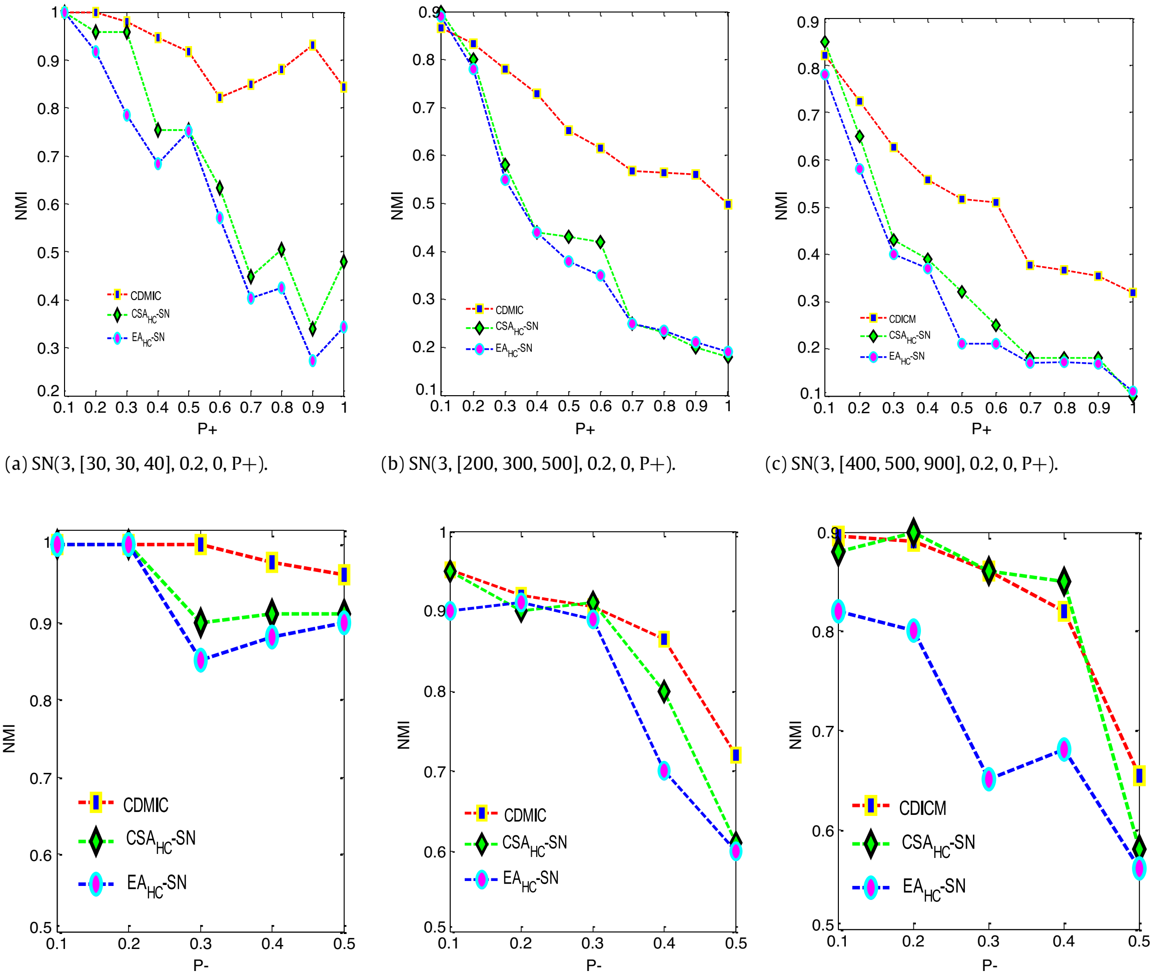
$$NMI(A, B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{C_A} N_{i.}\log\left(\frac{N_{i.}}{N}\right) + \sum_{j=1}^{C_B} N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \tag{11}$$

where A is a true partition, B is the partition detected with one algorithm, C_A is the number of real communities, C_B is the number of found communities, N is the total number of nodes of the network, N_{ij} is the number of nodes shared in common by the real community i in partition A and by the found community j in partition B, $N_{i.}$ is the sum over row i of matrix N_{ij}, and $N_{.j}$ is the sum over column j.

The results of CDMIC, EA_{HC}-SN and CSA_{HC}-SN methods on the signed networks generated with different P_+ and P_- are shown in Fig. 3. For each network, 10 independent runs are conducted for each algorithm. From Fig. 3, we can see that the performance of our CDMIC method is better than that of the EA_{HC}-SN and CSA_{HC}-SN for different structure and scale networks. Especially, when $P_+ \geq 0.4$, the *NMI* values of CDMIC are far higher than that of EA_{HC}-SN and CSA_{HC}-SN, which means that the communities obtained by CDMIC are more accurate than those obtained by EA_{HC}-SN and CSA_{HC}-SN.

4.3. Glossary network

Glossary network is a directed word network, which describes the connections among a set of technical terms, such as “tree” and “digraph”, contained in a glossary of network jargon [32,33]. It consists of 67 nodes denoting the technical terms



(a) SN(3, [30, 30, 40], 0.2, 0, P+).

(b) SN(3, [200, 300, 500], 0.2, 0, P+).

(c) SN(3, [400, 500, 900], 0.2, 0, P+).

(d) SN(3, [30, 30, 40], 0.2, P−, 0.1).

(e) SN(3, [200, 300, 500], 0.2, P−, 0.1).

(f) SN(3, [400, 500, 900], 0.2, P−, 0.1).

Fig. 3. The NMI obtained by EA_{HC}-SN, CSA_{HC}-SN and CDMIC algorithms on the signed networks with different structure and scale. Jaccard index was used to measure the similarity among nodes, and Q_s to evaluate the quality of network partition.

and 118 directed edges representing the links from one term to another term if the second term is used to describe the meaning of the first term.

Table 2 shows the results of functional modules detected by Newman [32], Rosvall [34], Palla [35], Wang [26] and our CDMIC methods, from which we can see that our CDMIC method can detect more functional modules than other four methods, and the isolated points are smaller than that of Palla, Rosvall and Wang methods. Fig. 4 shows the structures of functional modules detected by CDMIC method in the directed glossary network, which appear to correspond to the meaningful groups in understanding the relations among glossary terms. For example, module 1 (M1) deals with words that describe the tree structure. The term “Decision Tree” can be explained by its downstream terms, i.e., “Binary Search Tree”, “m-aryTree” and “Offspring”, and the term “Offspring” is the basic foundation for the formation of other upstream terms. The glossary terms of module 2 (M2) represent the real definition of the fundamental term “Isomorphic”. Module 3 (M3) provides an overview of the term “Walk” and contains the glossary terms derived from the fundamental term “Walk”. All other modules represent not only groups of term with similar meanings, but also the etymology of the network jargon. Above results indicate that CDMIC method is more effective than that of other four methods for detecting the functional modules in the directed glossary network.

4.4. Football network

Football network describes an American college football game between Division I colleges for the 2000 season, which consists of 115 nodes denoting the football teams and 616 edges representing the regular season games [2]. It is an unweighted network and can be divided into 12 communities according to athletic conference. Each community contains 8–12 teams.

Table 2
Results of Newman, Rosvall, Palla, Wang and CDMIC methods on the glossary network.

Method	Number of modules	Isolated points
Newman [32]	9	0
Palla [34]	5	37
Rosvall [35]	8	6
Wang [26]	12	3
CDMIC	15	0

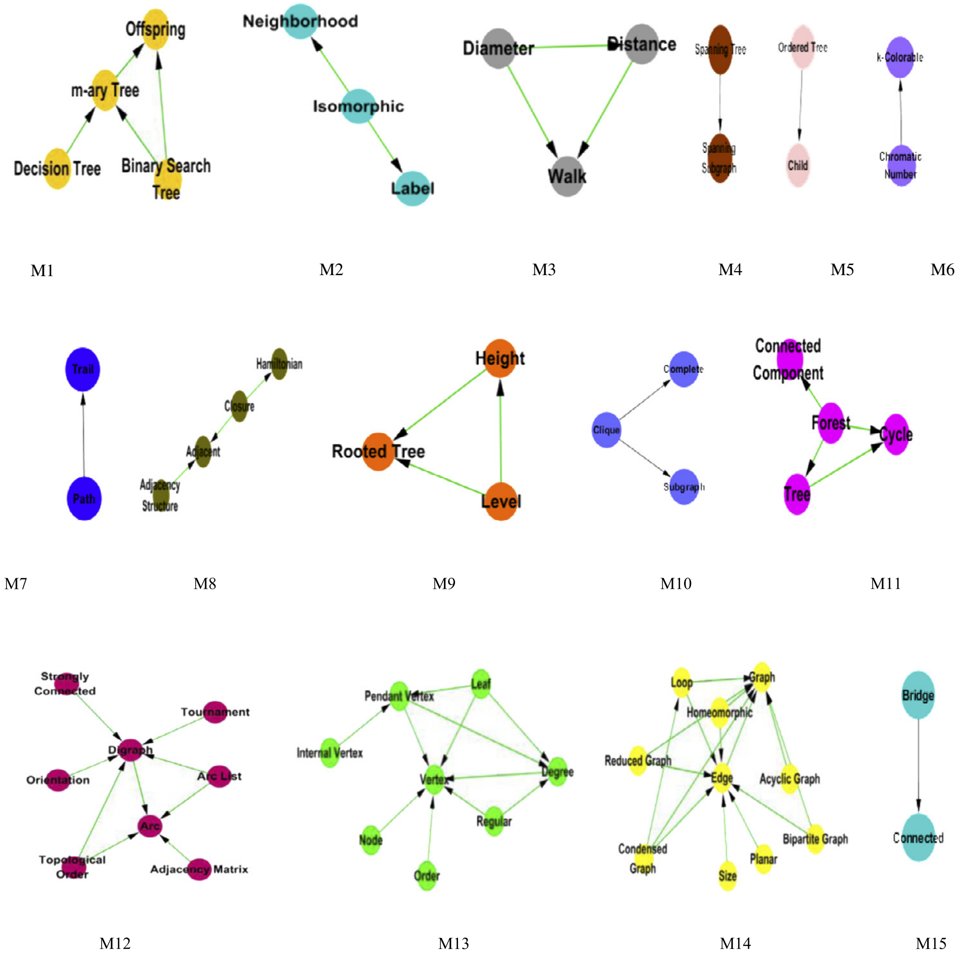


Fig. 4. Structures of functional modules detected by our CDMIC method using Sorenson (or Jaccard) index.

The results of GN [7], Chen [14], EM-BOAD [1] and our CDMIC algorithms on the Football network are shown in Table 3 and Fig. 5. From Table 3, we can see that the performance of CDMIC is the best in terms of *NMI* and modularity. The *NMI* value of CDMIC arrives at 0.9242, which is 0.2313, 0.0639 and 0.0222 higher than that of GN, Chen, EM-BOAD algorithms. CDMIC divides 115 teams into 12 communities and almost all the teams are correctly grouped with other teams in their conference. GN algorithm detected 6 communities with the modularity of 0.546, which under-estimates the communities. Chen's algorithm detected 13 communities with the modularity of 0.5868, and EM-BOAD algorithm detected 14 communities with the modularity of 0.562, both of which over-estimate the communities. Fig. 4 shows that 6 communities are correctly detected by our CDMIC approach, and just 9 teams (Central Florida, Louisiana Tech, Minnesota, Navy, Notre Dame, Texas Christian, Connecticut, Boise State, Utah State) in other 6 communities are wrongly divided. By analyzing the topological properties of these misclassified teams, we found that these teams have sparse connections intra-communities and density links inter-communities.

Table 3
Results of GN, EM-BOAD, Chen and CDMIC algorithms on the football network.

Method	N_{C_0}	N_{C_d}	NMI	Q
GN	12	6	0.6929	0.5460
Chen	12	13	0.8630	0.5868
EM-BOAD	12	14	0.9020	0.5620
CDMIC	12	12	0.9242	0.6277

Note: N_{C_0} and N_{C_d} denote the number of original and detected communities, respectively.

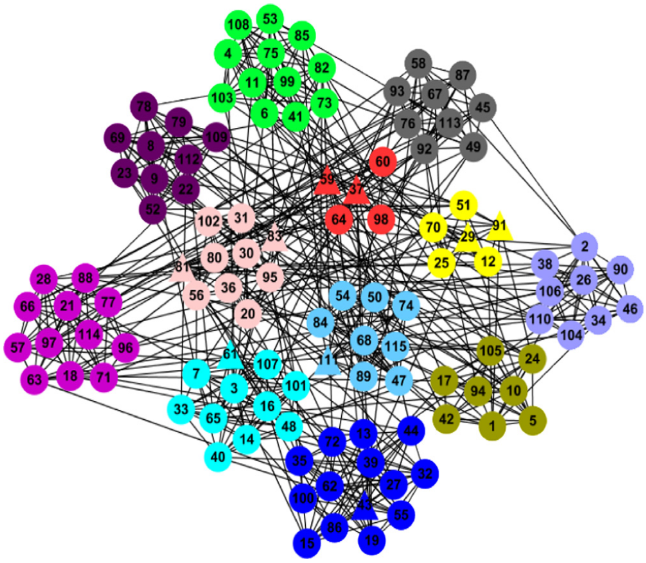


Fig. 5. Twelve communities detected by our CDMIC method on the college football network. The circle nodes are correctly divided into their communities, and the triangle nodes are wrongly divided into their communities.

4.5. *Collaboration network*

Collaboration network is a weighted network of scientists who are publishing on the topic of networks [27,36]. The 1589 nodes represent scientists and 2742 edges denote the co-authorships among the scientists if they coauthored one or more articles during the same time period. The weights are derived from the number of joint publications. In this paper, we just take the largest connected component with 379 scientists and 914 links as the test network. Since the actual communities are unknown in this collaboration network, we cannot compute its *NMI* and compare with other existing methods for highlighting the performance of our CDMIC method. Here, we focus on the problem of resolution limit in community detection. Ref. [17] reported that most of modularity maximization algorithms may fail to resolve communities with few than $\sqrt{L/2}$ edges, where L is the number of edges in entire network.

For this collaboration network, our CDMIC method detected 23 communities with modularity 0.8517. Among these communities, there are 14 large communities, and 9 small communities whose edge numbers are less than $\sqrt{914/2} \approx 21$. Table 4 shows the topology parameters of these small communities, where V_{size}, E_{in}, E_{out} and D_c are the number of nodes and edges in this intra-community; edges leave from this community and modularity density. From Table 4, we can conclude that there are 13.89% edges in the intra-communities and 2.84% edges in the inter-communities, which indicate that these small community structures detected basically meet the definition of community, and our CDMIC method can tolerate the resolution limit problem by discovering communities smaller than $\sqrt{L/2}$.

5. Discussion

5.1. *Effects of different similarity measures*

In order to evaluate the effect of different similarity measures to the community detection, we used the signed network ($SN(3, [30, 30, 40], 0.2, 0.5, 0.7)$), synthetic network ($Z_{out} = 5$) and football network as the test networks, and selected ten similarity measures to investigate the effect in terms of *NMI*. Table 5 shows the results of our CDMIC by using ten similarity measures on the three networks, from which we can see that except PA and LHN, other eight similarity measures can obtain better results of community detection, but there are little difference among them on different networks, e.g., Sørenson and

Table 4
Parameters of 9 small communities whose edge numbers are small than 21.

Community number	V_{size}	E_{in}	E_{out}	D_c
1	8	16	4	12.3810
2	5	6	2	6.3750
3	10	13	7	4.5000
4	8	15	4	10.8571
5	9	18	2	11.6518
6	9	12	2	5.0089
7	6	13	1	17.6000
8	10	17	2	7.9444
9	9	17	2	10.3661
Sum	74(0.1953)	127(0.1389)	26 (0.0284)	86.6843(0.3221)

Note: V_{size}, E_{in}, E_{out} and D_c are the number of nodes and edges in this intra-community; edges leave from this community and modularity density..

Table 5
Results (in terms of *NMI*) of ten similarity measures with CDMIC on three networks.

Measure	Signed network	Synthetic network	Football
PA	0.4039	0.6692	0.7215
AA	0.9017	0.8768	0.8983
CN	0.9017	0.8190	0.8983
HD	0.8962	0.9425	0.9111
HP	0.9139	0.8442	0.8732
RA	0.9582	0.8902	0.8401
LHN	0.5640	0.8907	0.9111
Salton	0.8562	0.8900	0.9111
Jaccard	0.8526	0.8937	0.9242
Sørenson	0.9582	0.9418	0.9111

Table 6
Results (in terms of NMI) of modularity measures with CDMIC on the signed network (SN(3, [30, 30, 40], 0.2, 0.5, 0.7)), football network and Zachary network.

Modularity	Signed	Football	Zachary
Q_u	0.9183	0.9242	0.7055
Q_e	0.9426	0.9242	0.8394
Q_s	0.9790	–	–

RA obtain the best detection results on the signed network, Sørenson and HD arrive at the best detection results on the synthetic network, while Jaccard gives the best results on the football network. LHN is not suitable to measure the signed network for community detection. PA performs the worst on these three networks, because it is often used to quantify the functional significance of edges subject to various network-based dynamics. Maybe it is suitable to dynamical networks.

Besides above ten similarity measures, a number of similarity measures based on global structural information, such as the average commutation time of a random walk, the pseudo inverse of the Laplacian matrix, Katz index and its variant, the transferring similarity, and the PageRank index [37–42], have been proposed to grasp the structure information of networks. These measures may give better results.

5.2. Effects of different modularity measures

We firstly select three modularity measures (i.e., Q_u, Q_e, Q_s) to investigate the effects on the signed network (*SN*(3, [30, 30, 40], 0.2, 0.5, 0.7)), football network and Zachary network. Football network is an un-weighted network, and Zachary network is a weighted network. The results are listed in Table 6, from which we can see that Q_s obtains higher *NMI* and is suitable for signed network. Q_e is suitable for weighted network. The *NMI* values of Q_u and Q_e are same in the football network, because un-weighted network is a special network (i.e., equal weight network) of weighted network, meaning that both of Q_u and Q_e can be used to detect communities in un-weighted networks.

We also use Q_u, Q_e and Q_d to detect communities in the glossary network which is a directed network, and find that both of Q_u and Q_e detect 7 communities, while Q_d detects 15 communities. In addition, some communities detected by Q_u and Q_e contain a mix of terms that fail to describe the connections among these terms.

Above results show that modularity measures can affect the detecting results of final community. It is better for selecting the modularity measure according to the network type.

Table 7
NMI of adjacency matrix, Jaccard similarity matrix and dissimilarity distance matrix used as the input of AP.

Method	Signed network	Synthetic network	Football
Adjacency matrix [43]	0.4836	0.5028	0.6618
Jaccard similarity matrix [44]	0.8021	0.6672	0.8860
Dissimilarity distance matrix	0.8562	0.8059	0.9111
CDMIC	0.9582	0.9425	0.9242

5.3. Effect of dissimilarity distance matrix to the AP algorithm

For the un-weighted networks, Refs. [43,44] took the adjacency matrix and Jaccard similarity matrix as the input of AP algorithm for detecting the communities, here we construct the dissimilarity distance matrix as the input of AP. Adjacency matrix reflects the link relationship between any two nodes. Jaccard similarity matrix measures the similarity between any two nodes with Jaccard index. Our dissimilarity distance matrix represents the dissimilarity strength between any two nodes by calculating the sum of dissimilarity along the shortest path between two nodes. In order to show the advantage of our dissimilarity distance matrix, we took these three matrices as the input of AP on the signed network ($SN(3, [30, 30, 40], 0.2, 0.5, 0.7)$), synthetic network ($Z_{out} = 5$) and football network. The comparing results are listed on the Table 7, from which we can see that using our dissimilarity distance matrix as the input of AP can obtain higher *NMI* values than that of other two matrices. In addition, our strategy of maximizing the modularity to select the community centers can effectively partition the network and obtain the rational community structures.

6. Conclusions

In this paper, we proposed a novel analysis framework, called CDMIC, for detecting the communities of un-weighted, weighted, un-directed, directed and signed networks. The two key aspects of our CDMIC method are constructing a dissimilarity distance matrix of network as the input of AP and identifying community centers by maximizing modularity. Three real-world networks and some synthetic networks are used to evaluate the performance of our CDMIC method. For the synthetic networks, signed networks and football network, our CDMIC method has higher performance in terms of classification accuracy and *NMI*. For Glossary network, CDMIC detects 15 functional modules and also there are no isolated data points. For collaboration network, CDMIC detects 14 large communities and 9 small communities whose edge numbers are less than $\sqrt{L/2}$ (L is the number of edges in entire network). These experimental results demonstrate that our CDMIC method can effectively detect the community structure of un-weighted, weighted, undirected, directed and signed networks, and it also has the strong ability to tolerate the resolution limit problem.

Acknowledgment

This paper was supported by the National Natural Science Foundation of China (91430111, 61473232 and 61170134).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.physa.2015.12.037.

References

[1] J. Li, X. Wang, J. Eustace, Detecting overlapping communities by seed community in weighted complex networks, Physica A 392 (2013) 6125–6134.
[2] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (2002) 7821–7826.
[3] B. Yang, W.K. Cheung, J. Liu, Community mining from signed social networks, IEEE Trans. Knowl. Data Eng. 19 (2007) 1333–1348.
[4] M. Fiedler, Algebraic connectivity of graphs, Czechoslovak Math. J. 23 (1973) 298–305.
[5] B.W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, Bell Syst. Tech. J. 49 (1970) 291–307.
[6] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, D. Wagner, On finding graph clusterings with maximum modularity, WG, vol. 7, 2007, pp. 121–132.
[7] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004) 026113.
[8] M.E. Newman, Analysis of weighted networks, Phys. Rev. E 70 (2004) 056131.
[9] J. Ruan, W. Zhang, Identifying network communities with a high resolution, Phys. Rev. E 77 (2008) 016104.
[10] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2005) 027104.
[11] X. Wang, G. Chen, H. Lu, A very fast algorithm for detecting community structures in complex networks, Physica A 384 (2007) 667–674.
[12] E.A. Leicht, M.E. Newman, Community structure in directed networks, Phys. Rev. Lett. 100 (2008) 118703.
[13] V. Nicosia, G. Mangioni, V. Carchiolo, M. Malgeri, Extending the definition of modularity to directed graphs with overlapping communities, J. Stat. Mech. Theory Exp. 2009 (2009) P03024.
[14] D. Chen, M. Shang, Z. Lv, Y. Fu, Detecting overlapping communities of weighted networks via a local algorithm, Physica A 389 (2010) 4177–4187.
[15] Y. Li, J. Liu, C. Liu, A comparative analysis of evolutionary and memetic algorithms for community detection from signed social networks, Soft Comput. 18 (2014) 329–348.
[16] R. Guimera, S. Mossa, A. Turtschi, L.N. Amaral, The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles, Proc. Natl. Acad. Sci. 102 (2005) 7794–7799.

[17] S. Fortunato, M. Barthélemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. 104 (2007) 36–41.
[18] J.W. Berry, B. Hendrickson, R.A. LaViolette, C.A. Phillips, Tolerating the community detection resolution limit with edge weighting, Phys. Rev. E 83 (2011) 056119.
[19] Z. Li, S. Zhang, R.-S. Wang, X.-S. Zhang, L. Chen, Quantitative function for community detection, Phys. Rev. E 77 (2008) 036109.
[20] S. Gómez, P. Jensen, A. Arenas, Analysis of community structure in networks of correlated data, Phys. Rev. E 80 (2009) 016114.
[21] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (2009) 623–630.
[22] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Networks 25 (2003) 211–230.
[23] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.
[24] X. Zhang, C. Furtlehner, M. Sebag, Data streaming with affinity propagation, in: Machine Learning and Knowledge Discovery in Databases, Springer, 2008, pp. 628–643.
[25] A. Condon, R.M. Karp, Algorithms for graph partitioning on the planted partition model, Random Struct. Algorithms 18 (2001) 116–140.
[26] B. Wang, L. Gao, Y. Gao, Control range: a controllability-based index for node significance in directed networks, J. Stat. Mech. Theory Exp. 2012 (2012) P04011.
[27] M.E. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[28] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. (1977) 452–473.
[29] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (2003) 396–405.
[30] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (2008) P10008.
[31] L. Danon, A. Diaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, J. Stat. Mech. Theory Exp. 2005 (2005) P09008.
[32] M.E. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167–256.
[33] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. 424 (2006) 175–308.
[34] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. 105 (2008) 1118–1123.
[35] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814–818.
[36] X. Ma, L. Gao, X. Yong, L. Fu, Semi-supervised clustering algorithm for community structure detection in complex networks, Physica A 389 (2010) 187–197.
[37] F. Göbel, A. Jagers, Random walks on graphs, Stochastic Process. Appl. 2 (1974) 311–336.
[38] L. Yen, M. Saerens, F. Fouss, A link analysis extension of correspondence analysis for mining relational databases, IEEE Trans. Knowl. Data Eng. 23 (2011) 481–495.
[39] L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1953) 39–43.
[40] E. Leicht, P. Holme, M.E. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2006) 026120.
[41] D. Sun, T. Zhou, J.-G. Liu, R.-R. Liu, C.-X. Jia, B.-H. Wang, Information filtering based on transferring similarity, Phys. Rev. E 80 (2009) 017101.
[42] D. Horowitz, S.D. Kamvar, The anatomy of a large-scale social search engine, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 431–440.
[43] J. Vlasblom, S.J. Wodak, Markov clustering versus affinity propagation for the partitioning of protein interaction graphs, BMC Bioinformatics 10 (2009) 99.
[44] Y. Wang, L. Gao, Detecting protein complexes by an improved affinity propagation algorithm in protein–protein interaction networks, J. Comput. 7 (2012) 1761–1768.