Contents lists available at ScienceDirect

Physica A

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

Detecting overlapping and hierarchical communities in complex network using interaction-based edge clustering

Paul Kim, Sangwook Kim [*]

School of Computer Science and Engineering, Kyungpook National University, 401 IT-4, 80 Daehakro, Bukgu, 702-701 Daegu, Republic of Korea

HIGHLIGHTS

- Most community detection methods use network topology and edge density.
- These methods decompose nodes connected by high weights into different communities, even when they intuitively belong to a single community.
- We propose a method of detecting overlapping and hierarchical communities in complex networks using interaction-based edge clustering.
- We find that the community quality and the overlap quality for our method surpass the results of the other methods.

ARTICLE INFO

Article history:
Received 6 January 2014
Received in revised form 10 September 2014
Available online 20 September 2014

Keywords:
Community detection
Edge clustering
Overlapping communities
Hierarchical structure
Complex network

ABSTRACT

Most community detection methods use network topology and edge density to identify optimal communities. However, in these methods, several objects that are connected by high weights may be decomposed into different communities, even when they intuitively belong to a single community. In this case, it is more effective to classify the objects into the same community because they perform important roles in controlling and understanding the network. To achieve this goal, in this paper, we propose a method of detecting optimal community structures in a complex network using interaction-based edge clustering. Our approach is to consider network topology as well as interaction density when identifying overlapping and hierarchical communities. Additionally, we measure the differences between the quantity and quality of intra- and inter-community interactions to evaluate the quality of the community structure. We test our method on several benchmark networks with known community structures. Additionally, after applying our method to several real-world complex networks, we evaluate our method through comparison with other methods. We find that the community quality and the overlap quality for our method surpass the results of the other methods.

1. Introduction

Networks that describe complex systems or concepts can be decomposed into communities or groups. Communities are usually subgraphs: the density of edges within the community is greater than the density of edges between communities [1]. The detection of community structures can be easy to understand, and a network can often be analyzed efficiently by dividing it into several groups [2]. Such communities often exist in social networks, biological networks and infrastructure

[*] Corresponding author. Tel.: +82 53 940 8881.
E-mail address: kimsw@knu.ac.kr (S. Kim).

http://dx.doi.org/10.1016/j.physa.2014.09.035

networks. There are many methods available to detect communities in complex networks. However, because most methods and algorithms identify communities based on the network topology, the community structure is strongly influenced by the edge density. Specifically, the identification of real communities with different structures and edge densities is difficult. Additionally, although most complex networks are directed and weighted graphs, because each member interacts with other members, most conventional methods do not consider the edge direction and weight simultaneously [3]. Therefore, we propose a community detection method based on the interactions of the members. This method evaluates and maximizes the difference between the quantity and quality of intra- and inter-community interactions to identify optimal communities. We refer to the communities identified by our method as interaction communities (IC).

Interaction communities are a specific type of community structure that maximizes the internal interaction within the community and minimizes the external interaction between communities. This concept differs in several ways from the conventional community definition. The candidate community structures identified using our method are different from the structures identified using modularity or edge density. When the weights of the edges in a network differ, the difference between interaction communities and the ideal community structure is greater. The primary reason for this difference is that we use the interaction density instead of the network topology to determine the community structure. This means that our quality function for evaluating the community structure in the clustering process returns a high value if the weights of the intra-community edges are maximized. However, when the weights of all edges in the network are the same, our method identifies a community structure that is consistent with a structure based on the edge density, such as modularity. Additionally, because we determine the clustering order in terms of edge directions and weights, a structure based on interaction communities is fundamentally different from the results of other clustering methods.

Various techniques, including hierarchical clustering, modularity optimization, detection of dense subgraphs, and statistical inference, among many others, have been used to detect community structures. The Girvan–Newman Algorithm (GN) is a well-known method [4]. The GN algorithm, which is based on divisive hierarchical clustering, probes the community structure by removing high levels of space between edges. The optimal communities identified by this method consist of hierarchical structures selected by means of modularity measurements [5,6]. This method detects non-overlapping communities.

However, an important property of communities in the real world is that a node can belong to several communities [7]. Methods for detecting overlapping communities have been the subject of extensive study. These methods attempt to allow a node to be shared among several groups. The cluster-overlap Newman–Girvan algorithm (CONGA), which has been used to extend the GN algorithm, is a divisive hierarchical algorithm that clusters undirected and unweighted networks [8,9]. In this method, an overlapping node is divided into several nodes, and the overlapping communities are evaluated using Nicosia's modularity [10]. The clique percolation method (CPM) is a popular method of detecting overlapping communities, but it has a non-hierarchical structure [11,12]. CPM combines two communities that share $k - 1$ nodes after identifying the maximal k-cliques in the network. Agglomerative hierarchical clustering based on maximal cliques (EAGLE) can identify both overlapping communities and hierarchical structures [13]. In this algorithm, overlapping communities are evaluated using a quality function that extends Girvan–Newman modularity. A link-community detection method (LC) based on agglomerative hierarchical clustering has been proposed. This method identifies overlapping communities and hierarchical structures by grouping two links that share one node [14].

Edges in complex networks can have directions and weights because the members interact directly with a measurable frequency and duration. Additionally, collaboration or communication events between the same members in a social network can be repeated, and a higher frequency of collaboration or communication usually indicates a closer relationship [12]. In this context, CPMd (CPM with directions) and CPMw (CPM with weights) can account for the directions and weights of edges, respectively, in the detection of overlapping communities [15,16]. CPMdw (CPM with directions and weights) mixes CPMd and CPMw and can detect communities in a weighted and directed network. Likewise, LC can be extended to LCd (LC with directions), LCw (LC with weights) and LCdw (LC with directions and weights). These methods can detect only dense communities while accounting for edge directions and weights.

To find communities in real-world complex networks, it is important to consider several approaches to community detection. These approaches include the detection of overlapping communities and hierarchical structure as well as consideration for directed and weighted edges. However, most methods for community detection do not use these approaches simultaneously. Moreover, most methods decompose nodes connected by high weights into different communities, even when they intuitively belong to a single community. The primary reason for this behavior is that these methods treat the edges connecting nodes as inter-community edges if they are bridges. In this case, it is more effective to classify the objects into the same community because they perform important roles in controlling and understanding the network. Therefore, in this paper, we consider network topology as well as interaction density in determining edge weights for the identification of overlapping and hierarchical communities. To achieve this goal, we propose a method of detecting optimal community structures in a complex network using interaction-based edge clustering. This method is based on single-linkage hierarchical clustering and searches for overlapping community structures in a weighted and directed network. Fig. 1 illustrates the differences between our method and previously proposed methods.

2. Methods

Our objective is to identify optimal communities while minimizing the influence of edge density and maximizing the quantity and quality of internal interactions. Consequently, we propose a method consisting of two processes. First, we

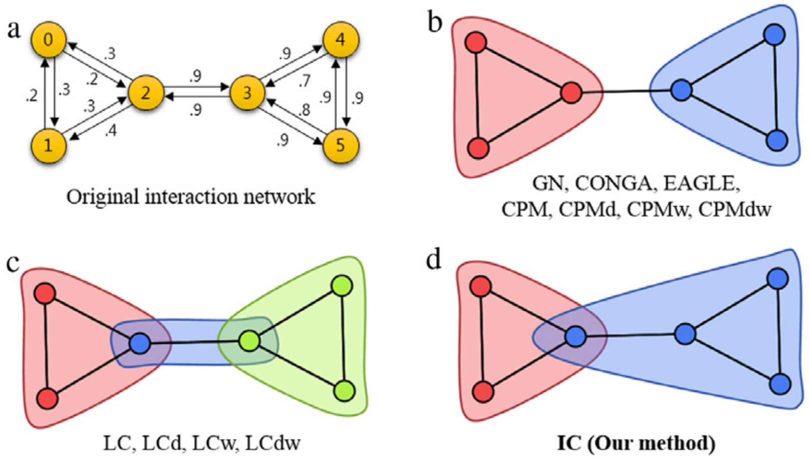


Fig. 1. Differences between our method and other methods. Given a network such as (a), (b)–(d) illustrate various derived community structures. In (b), existing methods are depicted that can detect an overlapping community but do not detect overlapping nodes. Methods based on LC, as demonstrated in (c), can identify community structures with high edge density values, although these evaluations include both edge directions and weights.

identify community structures using edge clustering based on single-linkage hierarchical clustering. Second, we evaluate the quality of the community structure using a quality function (Q). Finally, we select the optimal community structure with the highest Q.

We assume that the target network is a directed and weighted graph ($G = \{N, E\}$), where N is the set of members in the network and E is the set of directed and weighted edges. Each edge e_{ij} is a connection between nodes i and j. If $e_{ij} \in E$ and $e_{ij} \neq e_{ji}$, then nodes i and j are the sender and the recipient, respectively. Additionally, each edge has an interaction weight (w_{ij}) that describes the quality or frequency of the interaction between nodes i and j. This weight falls within the range $0 \leq w_{ij} \leq 1$. Graph G can consist of a set of communities ($C = \{C_0, \ldots, C_K\}$). Each community ($C_k = \{N_k, E_k\}$) contains a set of nodes (N_k) and a set of edges $\left(E_k = \left\{e_{ij} \mid i, j \in N_k, w_{ij} \neq 0\right\}\right)$. The sets of nodes and edges in graph G can be rewritten as $N = \bigcup_k N_k$ and $E = \bigcup_k E_k$, respectively.

2.1. Interaction-based edge clustering

We detect a set of optimal communities (C) via edge clustering based on single-linkage hierarchical clustering in a directed and weighted graph G. The algorithm is as follows:

1. Generate the set of edge pairs $\left(P = \left\{\left(e_{ik}, e_{kj}\right), \ldots\right\}\right)$.
2. Calculate the similarity $\left(S\left(e_{ik}, e_{kj}\right)\right)$ and distance $\left(d\left(e_{ik}, e_{kj}\right)\right)$ for all pairs of edges.
3. Initialize a community $C_k = \left\{N_k = \{i, j\}, E_k = \left\{e_{ij}\right\}\right\}$ using each edge e_{ij}. Afterward, generate the initial set of communities $C = \left\{C_0, \ldots, C_{|E|}\right\}$.
4. For the pair of edges $\left(e_{ik}, e_{kj}\right)$ with the smallest distance, merge two communities C_v and C_u when e_{ik} belongs to C_u and e_{kj} belongs to C_v. Afterward, remove $\left(e_{ik}, e_{kj}\right)$ from P.
5. Evaluate the set of communities C using a quality function Q.
6. Repeat steps 4 through 5 until $|P| = 0$.

The condition required to establish a pair of edges e_{ik} and e_{kj} is that node k should be a neighbor of nodes i and j. Therefore, two edges share a node (k) simultaneously, as depicted in Fig. 2. The type of edge pair varies depending on the directions of nodes i and j. In Fig. 2(a), node i is an out-neighbor of node k. However, in Fig. 2(c), node i is an in-neighbor of node k. In the directed graph, if the 3-clique is fully connected, there are 12 edge pairs. The similarity of a pair of edges $\left(e_{ik}, e_{kj}\right)$ is used to compare the directions and weights of the edges related to nodes i and j. To assess the similarity between all possible edge pairs in a target network, we assume that the network is in a weighted space and that all edges in the network are vector components. Additionally, we represent node i by a weighting vector $\left(a_i^+ = \left\{\tilde{A}_{i0}, \ldots, \tilde{A}_{i|N|}\right\} \text{ or } a_i^- = \left\{\tilde{A}_{0i}, \ldots, \tilde{A}_{|N|i}\right\}\right)$, as shown in Fig. 2(d) and (e) [14]. In Fig. 2(b), node i can be represented by a_i^+ because node k is an out-neighbor of node i. However, in a situation similar to that illustrated in Fig. 2(c), node i can be represented by a_i^- because node k is an in-neighbor of node i. Each weighting vector a_i^+ includes a set of vector components $\left\{\tilde{A}_{ij}\right\}$, where $\{j\}$ is the set of out-neighbors of node i. The vector component $\tilde{A}_{ij}$ is

$$\tilde{A}_{ij} = \frac{1}{|n^{\text{out}}(i)|} \sum_{i' \in n^{\text{out}}(i)} w_{ii'} \delta_{ij} + w_{ij} \tag{1}$$

where w_{ij} is the weight on edge e_{ij}, $n^{\text{out}}(i) = \left\{j \mid w_{ij} > 0\right\}$ is the set of out-neighbors for node i, $\delta_{ij} = 1$ if $i = j$ and is zero otherwise, and $\tilde{A}_{ij} \in \left\{0, w_{ij}, \sum_{i' \in n^{\text{out}}(i)} w_{ij} / \left|n^{\text{out}}(i)\right|\right\}$. Similarly, a_i^- includes a set of vector components $\tilde{A}_{ji}$ from the in-neighbors of node i, where $\tilde{A}_{ji} = \sum_{i' \in n^{\text{in}}(i)} w_{i'i}\delta_{ji} + w_{ji} / \left|n^{\text{in}}(i)\right|$ and $n^{\text{in}}(i)$ is the set of in-neighbors of node i. Therefore, we can calculate the similarity S between edges e_{ik} and e_{kj} based on the Tanimoto coefficient [17]. The similarity is calculated using

$$S\left(e_{ik}, e_{kj}\right) = \frac{a_i^+ \cdot a_j^-}{\left|a_i^+\right|^2 + \left|a_j^-\right|^2 - a_i^+ \cdot a_j^-}. \tag{2}$$

If the weights of all edges are equal to 1, this equation is equivalent to the Jaccard coefficient. Because our method is based on single-linkage hierarchical clustering, we calculate the distance for all pairs of edges using S. The distance d between edges e_{ik} and e_{kj} is

$$d\left(e_{ik}, e_{jk}\right) = 1 - S\left(e_{ik}, e_{kj}\right). \tag{3}$$

When $|P| = 0$ in step 6, the number of communities is 1 if a path from one node to all other nodes exists. If $|P| = 0$ and $C > 0$, then the network contains several components. Fig. 4 illustrates the entire process of interaction-based edge clustering through a dendrogram.

2.2. Quality function

The best-community structure in our method maximizes the inter-interaction of C and minimizes the intra-interaction of C. Each community should concurrently minimize the influence of the edge density. To satisfy this requirement, we propose the use of a quality function to evaluate the quality of the community structure using the interaction cohesion λ_k and the density D_k. When all weights lie in the range $0 \le w_{ij} \le 1$, the quality function is defined as

$$Q = \sum_k \frac{F_k^{\text{in}}}{T} \lambda_k D_k \tag{4}$$

where F_k^{in} is the internal interaction of C_k, T is $\sum_k F_k^{\text{in}}$, and $F_k = T$ if the number of communities is one. When measuring the quantity and quality of the interactions, Q is the average of $D_k\lambda_k$ weighted by F_k^{in}. Therefore, a community C_k exerts a strong influence on the community structure if it has a high F_k^{in} value. Q lies in the range $0 \le Q \le 1$; higher Q values indicate stronger interaction-based community structures. The internal interaction (F_k^{in}) is the sum of the weights for inter-interaction among members within the same community, as shown in Fig. 3(a). This interaction is defined as

$$F_k^{\text{in}} = \sum_{i,j \in N_k} w_{ij} = \sum_{i \in N_k} \sum_{j \in \mu_k^{\text{in}}(i)} w_{ij} \tag{5}$$

where $\mu_k^{\text{in}}(i) = \{j \mid j \in n(i), j \in N_k\}$ is the set of neighbors of node i that belong to the same community C_k and $n(i)$ is the set of all neighbors of node i. If m_k is the number of edges in C_k and all edge weights are 1, then F_k^{in} is equal to m_k.

In contrast, the external interaction (F_k^{out}) is the sum of the intra-interactions for community C_k, as illustrated in Fig. 3(b). This interaction is defined as

$$F_k^{\text{out}} = \sum_{i \in N_k} \sum_{j \in \mu_k^{\text{out}}(i)} w_{ij} \tag{6}$$

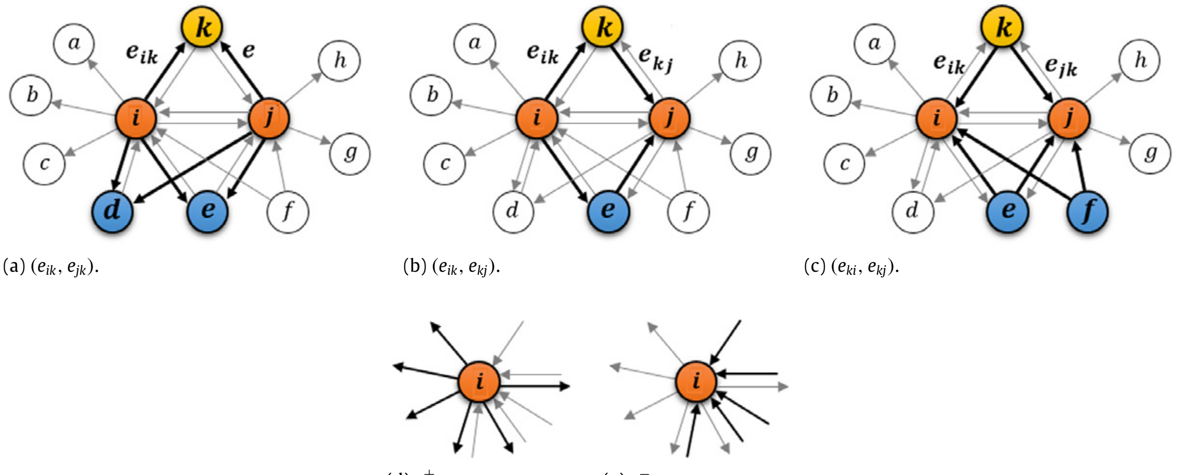
where $\mu_k^{\text{out}}(i) = \{j \mid j \in n(i), j \notin N_k\}$ is the set of neighbors of node i that are not members of C_k. We calculate the interaction cohesion (λ_k) for each community C_k using F_k^{in} and F_k^{out}. This value is the ratio between the amount of internal interaction and the amount of internal and external interactions:

$$\lambda_k = \frac{F_k^{\text{in}}}{F_k^{\text{in}} + F_k^{\text{out}}}. \tag{7}$$

The interaction cohesion lies in the range $0 \le \lambda_k \le 1$. If a community C_k has a high λ_k value, then the members of C_k are connected through highly weighted edges. However, a community only with a high λ_k value is not necessarily a good community. We also measure the interaction density D_k to evaluate the structure of the weighted edges for each C_k; D_k reflects the interaction weights in the edge density. We define D_k as

$$D_k = \frac{F_k^{\text{in}} - W_k^{\min}}{W_k^{\max} - W_k^{\min}} = \frac{F_k^{\text{in}}\left(n_k - 1\right)}{n_k^2\left(n_k - 1\right) - F_k^{\text{in}}} \tag{8}$$

where D_k represents the normalization of F_k^{in} with respect to the minimum and maximum weights of C_k. When the weights range from 0 to 1, we define the maximum weight $W_k^{\max}$ as $n_k\left(n_k - 1\right)$ and the minimum weight $W_k^{\min}$ as $F_k^{\text{in}}\left(n_k - 1\right)/W_k^{\max}$, where n_k is the number of nodes in C_k, $n_k - 1$ is the minimum number of edges in a directed graph, and $n_k(n_k - 1)$ is the maximum number of edges. A good community structure for the purposes of our method maximizes the interaction density, but the structure is influenced by the edge density because $W_k^{\max}$ and $W_k^{\min}$ are based on the number of edges in the community.



(a) (e_{ik}, e_{jk}).

(b) (e_{ik}, e_{kj}).

(c) (e_{ki}, e_{kj}).

(d) a_i^+.

(e) a_i^-.

Fig. 2. Edge pairs (a)–(c) and weighting vectors (d) and (e). To be an edge pair, two edges must share a node k. Depending on the in-neighbors and out-neighbors of node k, the method for calculating the similarity of the edge pair is different. If node i is an in-neighbor of node k, weighting vector (d) is selected; otherwise, (e) is selected.

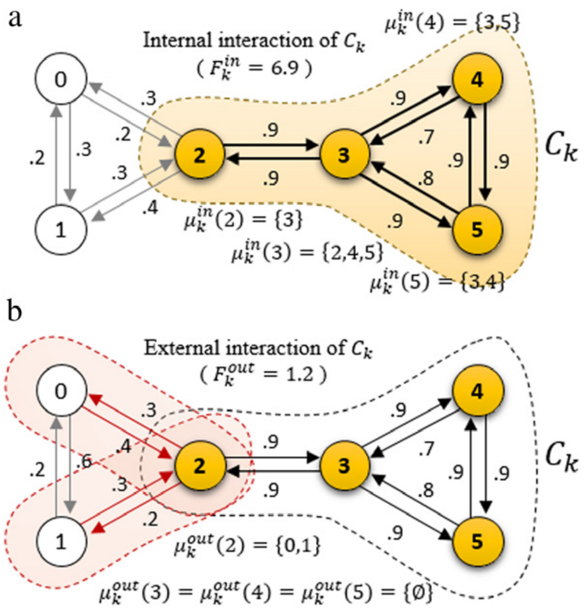

a
Internal interaction of C_k
(F_k^{in} = 6.9)
$\mu_k^{in}(4) = \{3,5\}$
0
1
2
3
4
5
.3
.2
.3
.2
.3
.4
.9
.9
.9
.7
.8
.9
.9
.9
C_k
$\mu_k^{in}(2) = \{3\}$
$\mu_k^{in}(3) = \{2,4,5\}$
$\mu_k^{in}(5) = \{3,4\}$
b
External interaction of C_k
(F_k^{out} = 1.2)
.3
.4
.6
.2
.3
.2
.9
.9
.9
.7
.8
.9
.9
.9
C_k
$\mu_k^{out}(2) = \{0,1\}$
$\mu_k^{out}(3) = \mu_k^{out}(4) = \mu_k^{out}(5) = \{\emptyset\}$

Fig. 3. An example of an internal interaction F_k^{in} and an external interaction F_k^{out}. Given a community such as C_k, (a) depicts the F_k^{in} of community C_k. In this case, F_k^{in} is the sum of the weights of the edges within C_k. In (b), F_k^{out} is the amount of interaction between the members and non-members of C_k.

If all edges in the network have the same weight, then the community structure is similar to the ideal community structure because the interaction density is approximately equal to the edge density. Finally, when using λ_k and D_k, the function Q for evaluating the quality of the community structure is

$$Q = \frac{1}{T}\sum_k \frac{F_k^{\text{in}^3}\,(n_k - 1)}{\left(F_k^{\text{in}} + F_k^{\text{out}}\right)\left(n_k^2\,(n_k - 1) - F_k^{\text{in}}\right)}. \tag{9}$$

Essentially, Q measures the quality of the potential optimal communities whenever two edges are merged in step 4 of the algorithm, as demonstrated in Fig. 4.

3. Experimental results

We test our method on several benchmark networks to compare the results of our method with known community structures. We then apply our method to several complex networks in the real world and identify the optimal interaction

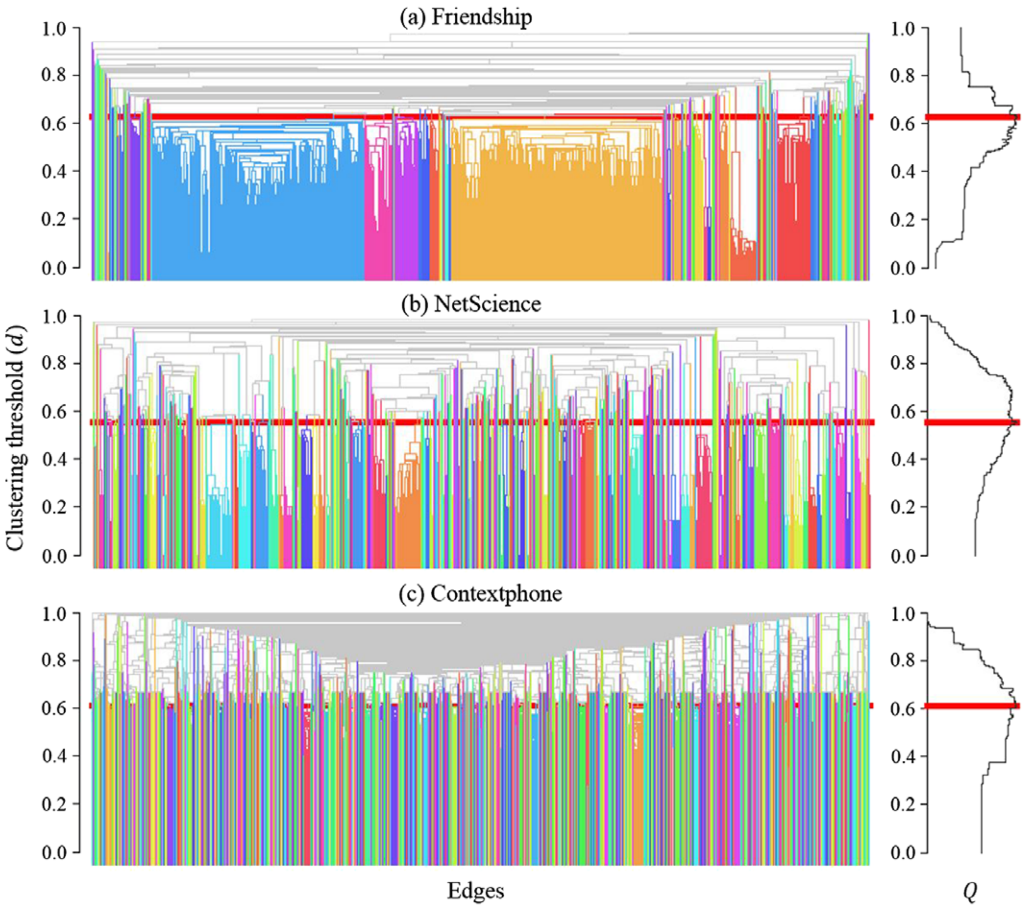


Fig. 4. Edge clustering in the Friendship, NetScience, and Contextphone networks. The dendrogram demonstrates that the proposed method merges the edges with the shortest length (d) among the edge pairs. Whenever a pair of edges is merged, we evaluate the community structure using a quality function (Q).

communities. Additionally, we compare the quality and coverage of the community structure identified by our method with those of the structures identified using other methods.

3.1. Benchmarks

We test our method on several LFR benchmark networks [18]. The LFR benchmark introduces heterogeneity in the degree and community-size distributions of a network. These distributions are governed by power laws with exponents of τ_1 and τ_2, respectively. For the generation of overlapping communities, the fraction of overlapping nodes O_n is specified, and each node has a number of community memberships O_m. LFR also provides a rich set of parameters through which to control the network topology, including the network size N, the mixing parameters μ_t for the network topology and μ_w for the edge weights, the average degree $\langle k \rangle$, the maximum degree $k_{\max}$, and the range of the community size. We focus on directed and weighted LFR benchmark networks [19] because the input for our method is a directed and weighted graph. We measure the relative performance of our method on the LFR benchmark with directed and weighted edges and overlapping communities. As a measure of similarity between the planted partitions, representing a known community structure, and the interaction communities identified by our method, we calculate the normalized mutual information (NMI) [20] to compare the two community structures. For all tests on artificial networks, each data point represents an average over 100 sample networks.

Fig. 5 presents the NMI between the planted partition of the benchmark and the interaction community identified by our method as a function of the mixing parameter μ_w. In all plots, the proposed method demonstrates good performance at low values of μ_w. However, as μ_w approaches 1, the difference between the interaction communities and the known network structure increases. The reason for this difference is that the nodes in the network possess more inter-community edges as μ_w is increased. In other words, the proposed method fails to identify the known community structure because the external interaction density is greater than the internal interaction density. Overall, the performance for $N = 5000$ is better than for $N = 1000$. Additionally, when $N = 1000$, the results are influenced by the network size, whereas the results for $N = 5000$

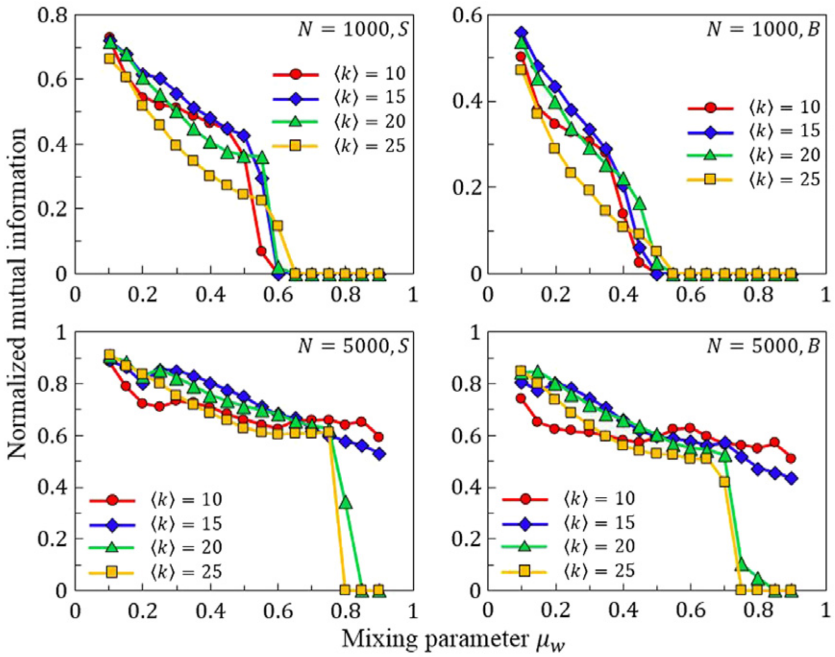


Fig. 5. Tests on directed and weighted LFR benchmark networks. The parameters of the networks are as follows: average degree $\langle k\rangle$, mixing parameters $\mu_t = \mu_w$, maximum degree $k_{\max} = 50$, fraction of overlapping nodes $O_n = 10\%$, node membership $O_m = 2$, and power-law exponents $\tau_1 = 2$ for degree and $\tau_2 = 1$ for community size. The notation S or B indicates that the community sizes are in the range [10, 50] or [20, 100], respectively. We consider two network sizes: $N = 1000$ (top) and $N = 5000$ (bottom).

are unaffected by the network size. The variation in the results as $\langle k\rangle$ is varied is slight, but the results for $\langle k\rangle = 15$ and $\langle k\rangle = 20$ are somewhat better than the other results.

We also test our method on directed and weighted LFR benchmark networks with overlapping communities. The left-hand side plot in Fig. 6 illustrates how the performance of our method decays with an increasing fraction of overlapping nodes, for several different choices of mixing parameters and different community sizes. Overall, the proposed method retains its performance in detecting the known community structure until the fraction of overlapping nodes O_n reaches 50%. We also observe differences between the planted partitions and the interaction communities as O_n is increased. The structure of the interaction communities is similar to the planted partitions for community size range S and $\mu_w = 0.1$ because the internal interaction density of the interaction communities is always greater than the external interaction density, as in the ideal community structure. In the right-hand side plot, we present a test on networks whose nodes are all shared among communities. As we increase the number of community memberships of the nodes, we detect interaction communities that are increasingly different from the known community structure. In particular, the proposed method retains good performance for $\langle k\rangle = 20$, but the results are considerably different in the other cases.

As demonstrated in several tests, differences exist between the known community structures of the benchmark networks and the structures of the interaction communities. In particular, the difference is greater when the fraction of intra-community edges in the known community structure is higher. The primary reason for this difference is that the community structure of an LFR benchmark network is generated based on the internal edge density. In other words, the proposed method minimizes the influence of the edge density when choosing a community structure if the internal interaction density of the communities is high.

3.2. Interaction communities

For the application of our method to complex networks, we select several directed and weighted networks. These networks consist of social, biological and infrastructure networks. Importantly, the members of these networks interact with one another. The weights of the networks are measured in terms of both the quality and quantity of interaction. Table 1 summarizes several properties and statistics of the selected networks. Contextphone [21] and Nodobo [22] are mobile phone networks that are weighted by call frequency. Friendship [23] is a collaboration network, and NetScience [24] is a co-authorship network. OClinks [25] and Twitter [26] are online social networks. The weight of each edge in OClinks represents the number of personal messages, and the weight of each edge in Twitter is the number of mentions and retweets. *C. elegans* [27] is a directed and weighted network that represents the neural network of *Caenorhabditis elegans*. USAirport [28] is an infrastructure network that is weighted by the number of flights.

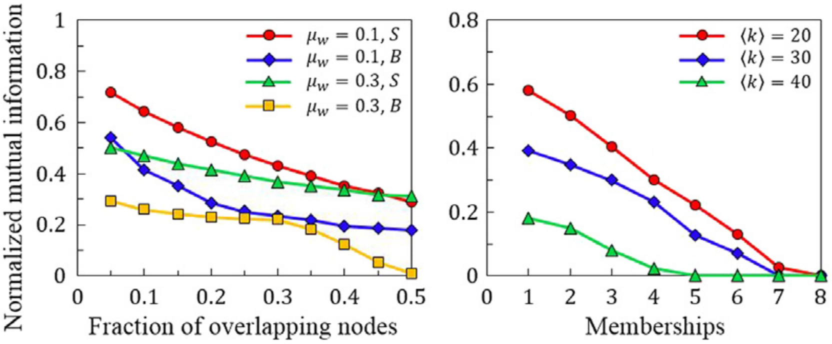


Fig. 6. Tests on directed and weighted LFR benchmark networks with overlapping communities. The common parameters of the networks are as follows: network size $N = 1000$, average degree $\langle k \rangle = 20$, maximum degree $k_{\max} = 50$, and power-law exponents of $\tau_1 = 2$ for degree and $\tau_2 = 1$ for community size. The notation S or B indicates that the community sizes are in the range [10, 50] or [20, 100], respectively. The left-hand side plot presents the normalized mutual information between the planted partitions and the community structure identified by our method as a function of the fraction of overlapping nodes. The four curves correspond to different values of the mixing parameter μ_w and different community size ranges. The right-hand side plot presents test results from networks whose nodes are all shared among communities. Each curve corresponds to a given value of the average degree. The specific parameters are as follows: $N = 2000$, $\mu_w = \mu_t = 0.2$, and fraction of overlapping nodes $O_n = 10\%$.

Table 1
Application of the proposed method to several complex networks. $|N|$ and $|E|$ are the number of nodes and edges in each network, respectively; the clustering threshold is the cutting point of the dendrogram with the highest Q; $|C|$ is the number of communities based on the clustering threshold; and M_{avr} is the average number of members in community C.

Network	$\lvert N\rvert$	$\lvert E\rvert$	Average degree	Clustering threshold	$\lvert C\rvert$ ($\lvert N_k\rvert \geq 3$)	M_{avr}
Contextphone	9573	14 416	3.0118	0.6080	320	11.6125
Nodobo	710	1 150	3.2394	0.5394	35	3.8571
Friendship	81	817	20.1728	0.6266	38	6.2368
NetScience	1589	2 742	3.4512	0.5524	163	4.0552
OClinks	1899	20 296	21.3754	0.5462	389	3.4936
Twitter	3656	185 809	101.6460	0.6190	2928	4.4283
C. elegans	297	2 345	15.7912	0.3578	194	3.7371
USAirport	500	5 960	23.8400	0.3886	192	5.6458

We detect interaction communities in the selected complex networks using the proposed method. In each network, we evaluate the quality of the community structure using the quality function Q while performing interaction-based edge clustering. Afterward, we select one of the community structures as an interaction community. Fig. 7 presents the value of the quality function Q, the number of communities $|C|$, and the average number of community memberships M_{avr} of a node as functions of the clustering threshold in each network. In this figure, the red line represents the highest quality for a community structure at the given clustering threshold $d\left(e_{ik},\ e_{jk}\right)$. If the threshold is 0, then the number of communities is equal to the number of edges, and if the threshold is 1, then the number of communities is the number of components that are not connected to one another. If all nodes in the network are connected, then the number of components is 1. In most networks, the average numbers of memberships and communities are relatively small when Q is the maximum value. Therefore, the community quality is always low when communities have many members. However, the NetScience and Twitter networks have a high Q value when the number of communities is large because the average number of members in most communities is small but the intra-interaction density of the communities is high. The results for the Contextphone and Nodobo networks are relatively similar because they exhibit similar structures, such as star topologies. In this case, the central nodes of the star topologies represent the participants who provided call data.

3.3. Evaluation

The goal of community detection is to identify an appropriate community structure for network analysis. Similarly, the goal of the proposed method is to identify a specific community structure that maximizes the difference between the amount of interaction among the members within the same community and the amount of interaction among members of different communities. This structure should concurrently cover as many nodes as possible. Additionally, the overlapping nodes should be more active than the other nodes. Based on these requirements, we evaluate our method and compare its performance with various other methods. The evaluation criteria include the community quality and coverage as well as the overlap quality and coverage [14]. First, we evaluate the community quality by calculating the average interaction weight of C:

$$\text{Community Quality} = \frac{F_k^{\text{in}} / |N_k|}{|C|} \tag{10}$$

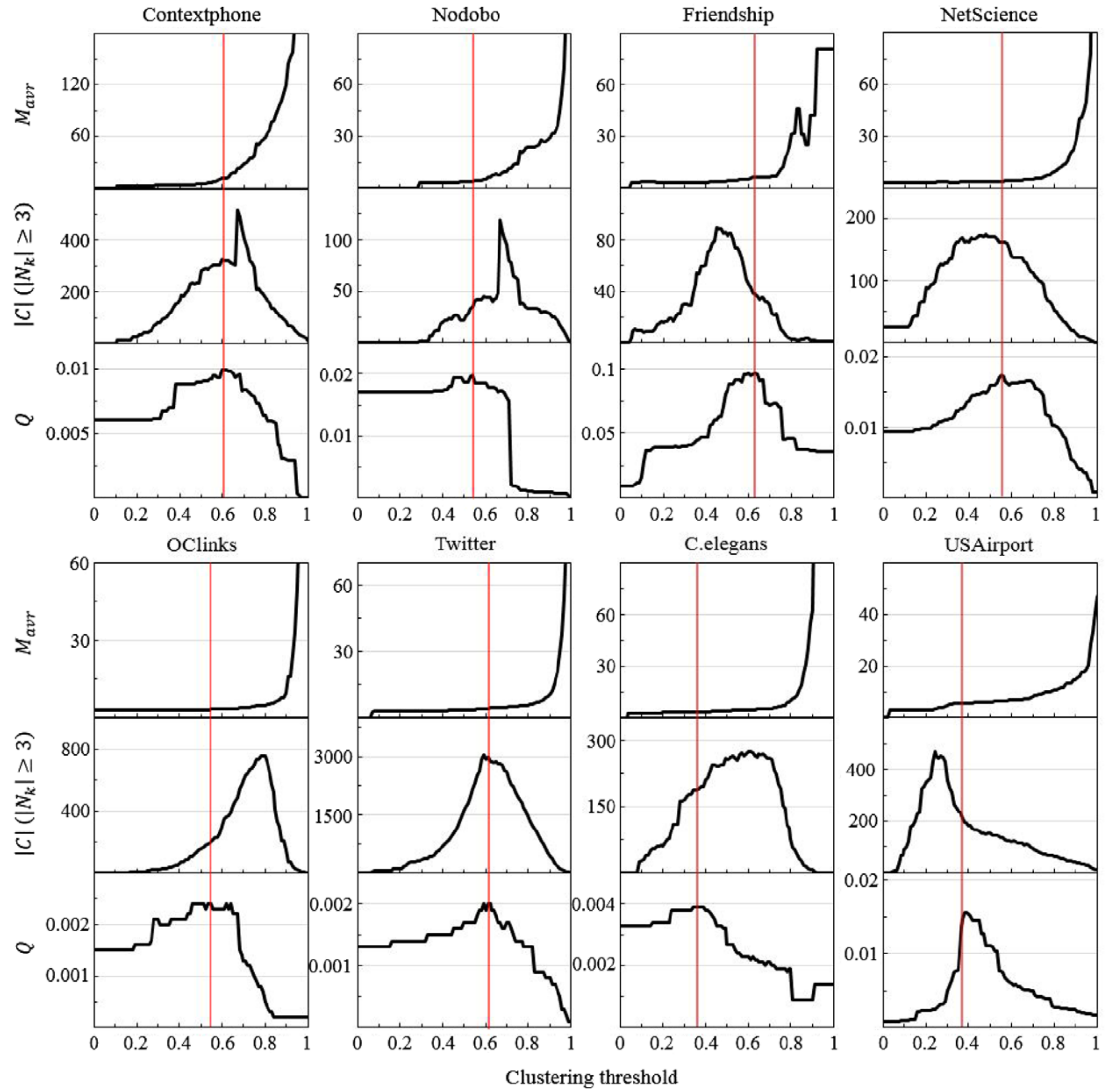


Fig. 7. The variation in the quality function (Q), $|C|$, and M_{avr} as functions of the clustering threshold during the performance of edge clustering to detect interaction communities in several complex networks. $|C|$ is the number of communities, and M_{avr} is the average number of members of a community. The red line indicates the highest value of Q and the cutting point of the dendrogram. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $C = \{C_k \mid |N_k| \geqslant 3\}$ is the set of communities. This equation indicates that members of the same community are more active if the community quality is high [29]. The community coverage is the ratio between the number of nodes that belong to at least one community and the total number of nodes: $\left|\bigcup_k N_k\right| / |N|$. This value indicates how much of the network is analyzed.

Most networks in the real world are composed of overlapping communities. The overlapping nodes that connect several communities are a type of hub. These hubs fulfill various roles and exchange a large amount of information. Therefore, we evaluate the overlap quality by calculating the ratio between the interaction weights of the overlapping nodes and the interaction weights of all nodes. Finally, we calculate the overlap coverage using the average number of community memberships. This quantity represents how much information is extracted from the portion of the network that a method can analyze.

We identify optimal communities using diverse community-detection methods, including the proposed method, and measure the quality and coverage of the communities and overlap. We then renormalize all community and overlap quality values such that the maximum value is 1 for the best-performing method. The community and overlap coverage are also

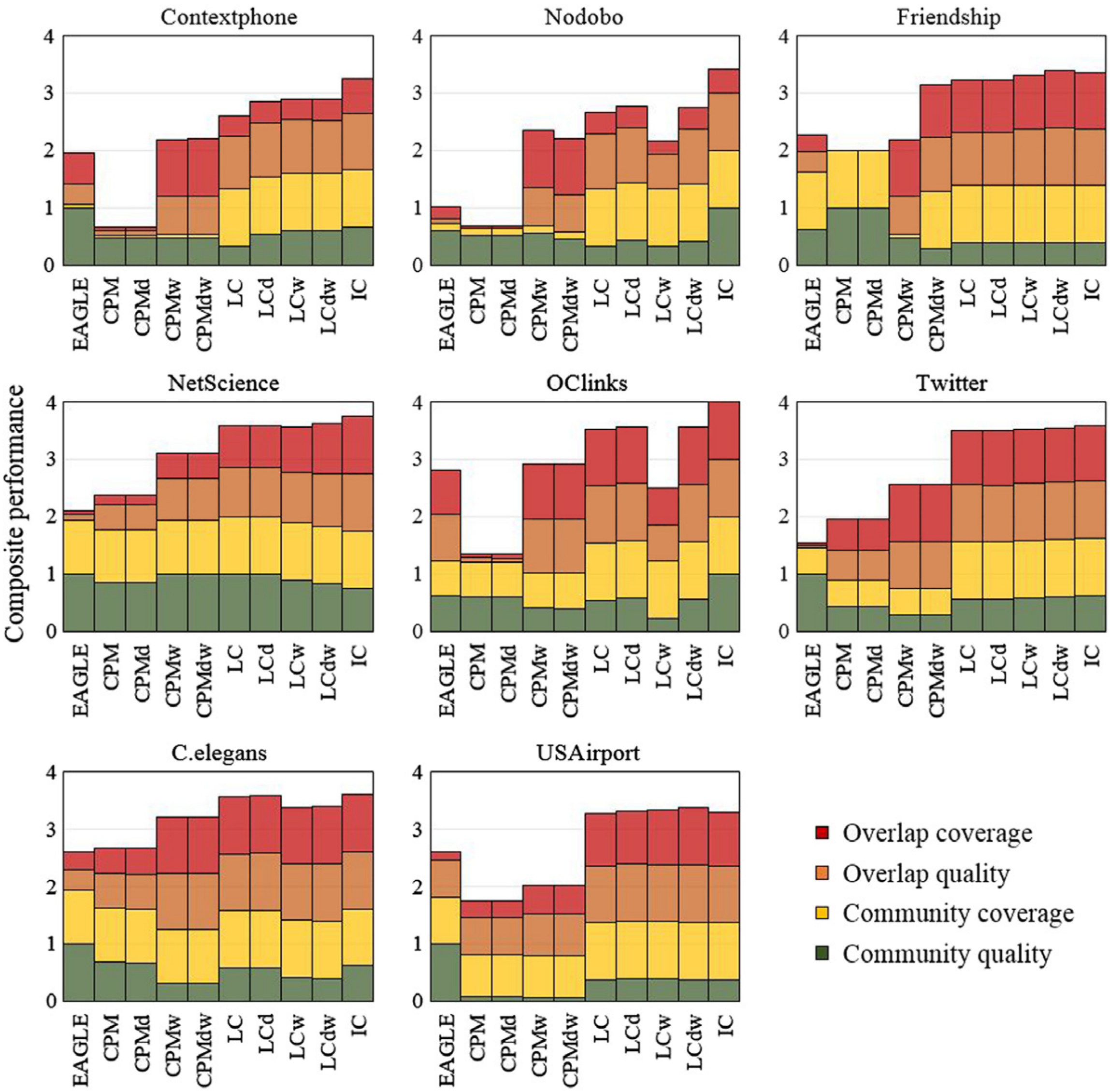


Fig. 8. Comparison of the composite performance of the proposed method (IC) and other methods. Each column shows all the evaluation criteria: the community quality and coverage and the overlap quality and coverage. These criteria measure the accuracy and sensitivity of the community and overlap structure determined by each method. It is evident that the proposed method (IC) yields the best performance in most networks.

renormalized; however, there is typically one algorithm that yields complete coverage, so that these values are already in the range [0, 1].

Fig. 8 presents a comparison of the investigated methods according to their composite performance. The methods evaluated for comparison are EAGLE [13], CPM [11], CPMd [15], CPMw [16], CPMdw, LC [14], LCd, LCw, LCdw, and our method (IC). For community detection, the inputs to EAGLE, CPM, and LC are undirected and unweighted networks. The inputs to CPMd and LCd are directed networks, and the inputs to CPMw and LCw are weighted networks. CPMdw, LCdw, and IC are tested using a directed and weighted network. Although the input to each method is different, the evaluations of the quality and coverage for all methods are performed using the original network. The proposed method outperforms the other methods on most networks. Therefore, the proposed method can efficiently analyze the interactions among members of the networks. In principle, the performance of a method that considers edge directions or weights should be better than the performance of a method that does not. If a method considers both directions and weights, its performance should be higher than that of a method that considers only the directions or only the weights of the edges. The methods based on EAGLE and CPM yield high values of community quality, but their community coverage is low because these algorithms detect only a few core groups. In contrast, our proposed method can achieve good coverage for many of the nodes in a network while retaining community quality. In particular, the overlap quality of the proposed method is better than any of the other methods; the overlapping community structure is stronger for the proposed method than for the other methods because the overlapping nodes are more active than in the other methods.

4. Conclusions

In this paper, we propose a method for the detection of interaction communities in complex networks. Our approach uses the edge directions and weights determined based on the interactions of the members to identify overlapping communities and hierarchical structures. We evaluate the differences between the quantity and quality of the intra- and inter-community interactions to evaluate the quality of the community structure. We apply our method to several complex networks to identify the optimal interaction communities. To compare the identified interaction-community structures with known community structures, we test our method on several LFR benchmark networks. We then evaluate the quality and coverage of our method and demonstrate that the composite performance of our method is better than other methods. Moreover, we observe that our method can efficiently analyze the community structures of complex networks consisting of weighted and directed edges.

Acknowledgment

This research was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2012M3C4A7032185).

References

[1] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[2] M.A. Porter, J.-P. Onnela, P.J. Mucha, Communities in networks, Notices Amer. Math. Soc. 56 (2009) 1082–1097. 1164–1166.
[3] J.J. Ramasco, S.A. Morris, Social inertia in collaboration networks, Phys. Rev. E 73 (2006) 016122.
[4] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99 (2002) 7821–7826.
[5] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in network, Phys. Rev. E 69 (2003) 026113.
[6] M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103 (2006) 8577–8582.
[7] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Community structure in large networks: natural cluster seizes and the absence of large well-defined clusters, Internet Math. 6 (2008) 29–123.
[8] S. Gregory, An algorithm to find overlapping community structure in networks, in: Lecture Notes in Computer Science, vol. 4702, 2007, pp. 91–102.
[9] S. Gregory, A fast algorithm to find overlapping communities in network, in: Lecture Notes in Computer Science, vol. 5211, 2008, pp. 408–423.
[10] 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) P03024.
[11] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814–818.
[12] G. Palla, A. Barabasi, T. Vicsek, Quantifying social group evolution, Nature 446 (2007) 664–667.
[13] H. Shen, X. Cheng, K. Cai, M. Hu, Detect overlapping and hierarchical community structure in networks, Physica A 388 (2008) 1706–1712.
[14] Y. Ahn, J.P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in network, Nature 466 (2010) 761–764.
[15] G. Palla, I.J. Farkas, P. Pollner, I. Derenyi, T. Vicsek, Directed network modules, New J. Phys. 9 (2007) 186.
[16] I.J. Farkas, D. Abel, G. Palla, T. Vicsek, Weighted network modules, New J. Phys. 9 (2007) 180.
[17] T.T. Tanimoto, An elementary mathematical theory of classification and prediction, IBM Internal Report 17, 1957.
[18] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E (2008) 046110.
[19] A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Phys. Rev. E 80 (2009) 016118.
[20] A. Lancichinetti, S. Fortunato, J. Kertesz, Detecting the overlapping and hierarchical community structure in complex networks, New J. Phys. 11 (2009) 033015.
[21] N. Eagle, A. Pentland, D. Lazer, Inferring friendship network structure by using mobile phone data, Proc. Natl. Acad. Sci. USA 106 (2009) 15274–15278.
[22] S. Bell, M. McDiarmid, J. Irvine, Nodobo: mobile phone as a software sensor for social network research, in: Proceedings of Context Awareness for Proactive Systems, 2011.
[23] T. Nepusz, A. Petroczi, L. Negyessy, F. Bazso, Fuzzy communities and the concept of bridgeness in complex networks, Phys. Rev. E 77 (2008) 016107.
[24] M.E. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[25] T. Opsahl, P. Panzarasa, Clustering in weighted networks, Social Networks 31 (2) (2009) 155–163.
[26] http://wiki.gephi.org/index.php/Datasets.
[27] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 292 (1998) 440–442.
[28] V. Colizza, R. Pastor-Satorras, A. Vespignani, Reaction–diffusion processes and metapopulation models in heterogeneous networks, Nat. Phys. 3 (2007) 276–282.
[29] H. Yu, et al., High-quality binary protein interaction map of the yeast interactome network, Science 322 (2008) 104–110.