FAST COMMUNITY DETECTION ALGORITHM BASED ON RELATIONSHIP STRENGTH COUPLING IN SOCIAL NETWORKS

Ye Li, He-Jin Mao, Xing-Yu Qi, Luo-Kun Ren, Hui-Jia Li

School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100080, China.
Hjli@amss.ac.cn.

Keywords: community structure; social network; k-relationship strength; optimal number of communities; polynomial time.

Abstract

Community structure detection is one of the most interesting issues in the study of social networks. However, there are seldom polynomial time algorithms which is able to uncover community structure accurately. Inspired by ideas of famous Modularity optimization, in this paper, we proposed a novel type of k-strength relationship which naturally represents the coupling distance between two nodes. Community structure detection algorithm is presented using a generalized Modularity measure based on the k-strength matrix. To obtain the optimal number of communities, we then propose a new parameter-free framework using the eigenvalue gap of the specific transition matrix. Finally, we apply our algorithm both on artificial and real networks. Theoretical analysis and experiments show that the algorithm is able to uncover communities fast and accurately, which can be easily applied on large-scale real networks.

1 Introduction

Community structure detection [1–3] is a main focus of complex network studies and has attracted a great deal of attention from various scientific fields. Intuitively, community refers to a group of nodes in the network that are more densely connected internally than with the rest of the network. The studies for community detection are potentially useful in real social networks because nodes in a community are more likely to have same properties and all these communities may be functional groups. The methods for detecting community in social networks are similar to the graph partitioning in graph theory [4, 5]. For example, in parallel computing, the pattern of required communications can be represented as a graph or network in which the nodes represent processes and edges join process pairs that need to communicate. The problem is to allocate the processes to processors in such a way as roughly to balance the load on each processor, while at the same time minimizing the number of edges that run between processors so that the amount of inter processor communication is maximized. In general, finding an exact solution to a partitioning task of this kind is an NP-complete problem, so it is prohibitively difficult to be solved accurately for large graphs. Inspired of this, a variety of heuristic algorithms have been developed that give acceptably good solutions in many cases, the best known being perhaps the Kernighan-Lin algorithm which runs in time $O(n^3)$ on sparse graphs [11].

Many algorithms on community detection had been proposed recently [6–8] and some of them are designed by the parameters of the networks, for example, eigenvectors of graph matrix, maximal modularity Q, clustering coefficient etc. Some algorithms are designed by dynamical characters of networks, such as random walk and spreading mechanism etc, however, those methods only deal with the special cases. For example, a network is divided into two communities according to the second smallest eigenvector of Laplace matrix. Yet, the second smallest eigenvector does not work when there are more than two communities. A network can be cut into several communities by the maximum modularity [9]. Unfortunately, computing the maximum modularity Q is proved to be NP-complete [10]. It means not all the communities are detected by computing the values of Q even though there are many heuristic algorithms. The random walk [12, 13], each node to be a walker and the walker will randomly choose a neighbor and currently stands on to localize in each time, has a probability to reach any other nodes, a dendrogram is got and the communities can be detected with the help of modularity Q. But it is difficult to specify the optimum random-walking time. Signal sending [14] is to transfer the topological relationship of nodes into the geometrical structure in n-dimensional Euclidean spaces, how to choose a proper p and partite all nodes into p-cluster is the weakness even it is empirical by the aid of F-statistics. Other methods depend on the probability of the communities in dynamic social networks such as [15, 16], and the values of modularity to find the proper communities such as [7, 8].

Since there are seldom polynomial time algorithms detecting the communities precisely, some valuable researches are focus on how to obtain much lower computing complexity for the detection algorithm compare with much more accuracy. In this paper, in order to

design fast and accurate algorithm to detect communities in weighted social networks, we proposed a novel definition, i.e. k-strength relationship, which naturally represents the coupling distance between two nodes. Community structure detection algorithm is presented using a generalized Modularity measure based on the k-strength relationship matrix in various types of social networks. Furthermore, to obtain the the optimal number of communities, we propose a new parameter-free method using the eigenvalue gap of specific transition matrix. Finally, we apply our algorithm on both benchmark network and real networks to evaluate its efficiency. Theoretical analysis and experiments show that the algorithm can uncover communities fast and accurately, which be easily extended to large scale real networks.

The outline of the paper is as follows. In Section 2 we introduce the fundamental definitions, such as k-strength relationship and its generalized Modularity measure. In Section 3, we present the details of our framework, including the procedures of algorithm and the analysis of computational complexity. Section 4 describes a novel method which able to determine the optimal number of communities naturally. Then we give some representative experiments on both benchmark and real networks to validate the effectiveness and efficiency of the algorithm in Section 5 . Finally, Section 6 concludes this paper.

2 Methods

2.1 Definitions

In many societies, such as the economic systems, the agents in system influence one another directly: a rush to buy or sell a particular asset can promote the other to do the same. In most common cases, the agents are influenced only by their neighbors who joint by direct relationships. All the buyer and seller formed an inseparable structure and have very little interactions outside the structure. Such structure is known as communities in social networks.

We denote agents by nodes in network and the influence each other by a weight between two nodes. In the following, a network is denoted by G with N-node set V, m-link set E and G also is an undirected without loop or multi-edges. The adjacency matrix A of G is a $N \times N$ zero-one matrix denoted by $A = (a_{ij})_{n\times n}$, where a_{ij}=1 if there is a link between i and j, and a_{ij}=0 otherwise. The adjacency matrix in an undirected graph is symmetric. If the network is weighted, we denote the weight of each link by w_{ij} and the weight matrix is $W = (w_{ij})_{N\times N}$. For a given positive integer k, we denote a path from node i to j by a k-path if it is a walk with $k+1$ nodes and without cycle on it. $A^k = (a^k_{ij})_{n\times n}, a^k_{ij} = \sum_{l=1}^{N} a^{k-1}_{il} \times a_{lj}$ is the number of k-paths from node i to j ($i \neq j$), if $i = j$, set a^k_{ij}=0. We denote $S^k = (s^k_{ij})_{N\times N}$ be a matrix of G for a given positive integer k. S^k is defined as a k^{th}**-strength matrix** of G. It is recursively defined as following:

If $k = 0$,

$$S^0 = A, \tag{1}$$

If $k = 1$,

$$S^1 = (w_{ij})_{N\times N}, \tag{2}$$

For all $k \geq 2$, let display

$$S^k = (s^k_{i,j})_{N\times N}, s^k_{i,j} = \sum_{s=1}^{a^k_{i,j}} \frac{1}{k} \sum_{l=1}^{k} w_{i^s_{l-1} i^s_l}, \tag{3}$$

where $i = i^s_0, i^s_1, \cdots, i^s_{k-1}, i^s_k = j$ are k-path for$s = 1, 2, \cdots, a^k_{i,j}$. To compute the values of $s^k_{i,j}$, $S^0 = A$ is fixed as the network determined. All the k-path between each pair of nodes can be obtained by $A^k, S^{i,j}$ is an additive polynomial. Therefore, we can compute the value of $S^k_{i,j}$ precisely.

Each k-strength matrix induces a ***k-strength relationship***: $R_k = \{(i, j, s_{i,j})|s_{i,j} = \sum_{l=1}^{k} s^l_{i,j}\}$. That is, $s_{i,j}$ in R_k are the elements of $S = S^1 + S^2 + \cdots + S^k = (\sum_{l=1}^{k} s^l_{i,j} = s_{i,j})_{N\times N}$. We denote S a ***k-strength matrix of G*** and the networks induced by S is a k-strength relationship networks. It is involved with a global idea of the mean-field theory on the definition of k-strength relationship networks. Each node knows all the others' information (the weights of nodes). It might be quite reasonable in many real systems. For example, the traders on Shanghai Stock Exchange are influenced by others on the same floor, but they can also be reminded by the trading patterns occurring on London or Paris. Therefore, some mature trading behavior patterns will be formed in economic systems. It is also very common in social networks to express the strength of friendship among people. For instance, in acquaintance network, the relationship is the tightness of acquaintance and higher the value is, more often the communication occurs. Another useful definition in our framework is minimal q-cut of a graph, which denotes the cut edges own the smallest sum of weight.

Here, q is a positive integer, $\{C_1, C_2, \cdots, C_q\}$ with $|C_i| = k_i$ and $\cup_{i=1}^{q} C_i \subseteq V(G)$ be a vertex subset such that the remaining of G after deleting all C_i is a disconnected and the sum of link weights among the remaining is the minimum. It was found that a minimum cut is a partition of G when $\cup_{i=1}^{q} C_i \subseteq V(G)$.

Guttmann had designed an algorithm to detect minimum cut in complete graphs [17] and inspired of his idea, we will detect the communities using the strength relationships. Our framework is also based on maximizing Modularity which firstly proposed by Newman, and we generalized it on strength relationship matrix of G. Suppose there are $q(q \leq N/2)$ communities in G

(how to confirm the value of q is really a tough problem, we will propose a framework in section 4 to solve it), $\mathcal{C} = \{C_1, C_2, \cdots, C_q\}$. The generalized Modularity Q in strength relationship matrix is defined by

$$Q = \max_{q} \sum_{i=1}^{q} (c_{i,i} - c_i^2) \tag{4}$$

where $c_{i,i} = \sum_{i,j} \frac{s_{i,j}}{\Delta} \delta_{i,j}, c_i = \sum_j s_{i,j}$ and $\Delta = \sum_{i,j} s_{i,j}$. $\delta_{i,j} = 1$,if the nodes i and j are in the same community, $\delta_{i,j} = 0$, otherwise. $c_{i,i}$ denotes the fraction of strength with both ends in the same partition C_i, c_i is the proportion of strength with one end in C_i and the other not. If the network is unweighted (binary network), the Q is just Newman's modularity. Based on this form, the new measure can capture the properties of the real social systems. One can find that both direct and undirect information between two nodes can be used within our framework. When two nodes are exchanging their information in social networks, the chains will formed within the same community. Thus, it might be more reasonable to describe relationship between nodes using the strength relationship, such as same ideas in a society are more likely be connected closely and transmitted one by one. It means a tightly connected social community implies a faster rate of information transmission or rumor spreading rate than a sparsely connected one, because more paths there are, faster transmissions rate there is.

2.2 Determining the k-strength relationship matrix

As described above, the k-strength relationship matrix is fundamental to the whole framework. In this section, we focus on determining the k-strength relationship matrix. The following theorem not only provides the process of computing all the elements, but also reveals the important time complexity information.

Theorem 1: *The k-strength relationship matrix W^k can be obtained in polynomial time.*

Proof: Suppose adjacent matrix of network G is A, the number of all k-length paths from i to j is $a_{i,j}^k$, and $A^k = A^{k-1} \times A = (a_{i,j}^k)$ in [16]. A path is called k-path if its length is k. Denote the k-path by $\{i_0, i_1, \cdots, i_{k-1}, i_k\}$ with $k+1$ nodes and $i_s \neq i_j$ for all s and j, that is, there is no cycle in the path.

In order to get the elements in k-strength relationship matrix, we define an operation $\oplus$ on weight matrix of G. $\oplus : W^k = W^{k-1} \oplus W = (w_{i,j}^k)_{N \times N}$, where $w_{i,j}^k$ is defined as: If $\sum_{l=1}^{N} (w_{i,l}^{k-1} \times w_{l,j} \neq 0)$, it means there are k links connect node i and j. Equivalent, there are at least one term in $\sum_{l=1}^{N} (w_{i,l}^{k-1} \times w_{l,j} \neq 0)$. Without of generally, we suppose, there are h terms not zeros,$w_{i,l^1}^{k-1} \times w_{l^1,j} \neq 0, w_{i,l^s}^{k-1} \times w_{l^s,j} \neq 0, w_{i,l^h}^{k-1} \times w_{l^h,j} \neq 0$. Then $w_{i,j}^k = \sum_{s=1}^{h} (w_{i,l^s}^{k-1} + w_{l^s,j})$; Otherwise, $w_{i,j}^k = 0$ (that is, there is no link joint i and j).

The value of $w_{i,j}^k$ is the sum of all weights in each k-path from i to j. We can take not so much effort to obtain $s_{i,j}^k = w_{i,j}^k/k$. That is $s_{i,j}^k = \sum_{s=1}^{a_{i,j}^k} \frac{1}{k} \sum_{l=1}^{k} w_{i_{l-1}^s, i_l^s} = w_{i,j}^k/k$.

All the k-path can be lay out when $\sum_{l=1}^{N} (w_{i,l}^{k-1} \times w_{l,j})$ is determined. If $w_{i,l^1}^{k-1} \times w_{l^1,j} \neq 0, w_{i,l^s}^{k-1} \times w_{l^s,j} \neq 0, w_{i,l^h}^{k-1} \times w_{l^h,j} \neq 0$ for each positive integer $k \geq 2$; Denote a k-path connect nodes i and j by $P_{i,j}^k$, it is easily to find there are $a_{i,j}^k$ k-paths joint node i and j and hence,$P_{i,j}^k = \{P_{i,l^1}^{k-1} \vee (l^1, j), P_{i,l^2}^{k-1} \vee (l^2, j), \cdots, P_{i,l^h}^{k-1} \vee (l^h, j)\}$, where $P_{i,l}^{k-1} \vee (l, j)$ means the all k-path formed by the $(k-1)$-paths in set $P_{i,l}^{k-1}$ join the link (i, j).

Finally, we can lay out all the k-paths inductively. That is, $\oplus$ is a polynomial time algorithm. Totally, the strength matrix is got by computing the weight matrix with computing complexity time $O(n^2 m)$ since the multiplication of each pairs of N-rank matrixes costs at most $N \times N$ and at most m links. Output $S^k = (s_{i,j}^k) N \times N$ for a fixed k. Outline all the paths from i to j with length k, $i = i_0^s, i_1^s, \cdots, i_{k-1}^s, i_k^s = j$ for $s = 1, 2, \cdots, a_{i,j}^k$.

The proof is end.

2.3 Community detection algorithm

A minimal q-cut $\widetilde{E}$ of G is an edge set with minimal sum of weight that the remaining graph of deleting the edges set, $G - \widetilde{E}$, is an isolated graph. A directly method to determine the partitions is investigating all the components of the remaining graph $\mathcal{C} = \{C_1, C_2, \cdots, C_c\}$. We need chose the components such that $\sum_{i<j} w(C_i, C_j)$ is minimum or $\sum_{i=1}^{c} w(C_i, C_i)$ is maximum by maximum flow and minimum cut theorem [18]. However, the minimal q-cut problem is NP-complete and it is difficult to find a polynomial algorithm. Fortunately, Guttmann-Beck and Hassin designed an algorithm in complete graphs and proved the approximate solution is less than three times the optimal [19]. Inspired by this nice idea, we obtain the detailed procedures in Algorithm 1.

2.4 Computational complexity

For a given positive number q, we can solve the transport problem in time $O(N)$ since it is a 0-1 transportation problem. There are $C_N^q O(qN)$ subsets of V. Altogether the time complexity is $O((q+1)N)$. Here, two important claims are proposed which useful to the analysis:

Claim 1: If $\{C_1, C_2, \cdots, C_q\}$ is a partition of $\widetilde{G}$, if and only if it is a partition of G.

Proof: It is easy to verify that the claim holds, since $\widetilde{G}$ and G has the same vertex set.

Claim 2: Suppose $\{C_1, C_2, \cdots, C_q\}$ are q partition

Algorithm 1 Community detection algorithm

Require: a social network $G = (V, E)$;

Ensure: a minimum q-partition with maximal modularity value.;

1: **Step 1:** Shrink each one degree node to its neighbor until there is no one degree node in G. This operation does not affect the community detection, because the one degree node has no other choice but to its unique neighbor, so the one degree node will be in the same community with its neighbor. We still write the network as G.

2: **Step 2:** Determine the optimal number of communities q using the theorem 2(we will show it in the following section).

3: **Step 3:** For a fixed q , the minimum q-cut problem is polynomial time solvable in $O(|V|^{q^2})$ [17].

4: Therefore, we suppose we had a partition $\mathcal{C} = \{C_1, C_2, \cdots, C_q\}$ in $\widetilde{G} = (V, \widetilde{E})$, and $|C_i| = k_i, \sum_{i=1}^{q} k_i = N, C_i \cap C_j = \emptyset$. We will detect he minimum q-cut in $\widetilde{G}$, and then prove it is also is the minimum q-cut in G. Let $v_i \in C_i$ for $i = 1, 2, \cdots, q$. $x_{i,j} = \begin{cases} 1 & , \quad u_j \in C_i \\ 0 & , \quad otherwise \end{cases}$

Begin

For $\{v_1, v_2, \cdots, v_q\} \subset V, v_i \in C_i$.

For $u_j \in V - \{v_1, v_2, \cdots, v_q\}$,the following transport problem is optimal.

$$min : \sum_{i=1}^{q} \sum_{j=1}^{N-q} w(C_i, u_j)(1 - x_{i,j})$$

$$subject\ to \begin{cases} \sum_{j=1}^{N-q} x_{i,j} = k_i - 1 & , \quad i = 1, 2, \cdots, q \\ \sum_{i=1}^{q} x_{i,j} = 1 & , \quad j = 1, 2, \cdots, N - q \\ x_{i,j} \in \{0, 1\} & , \quad i = 1, 2, \cdots, q \text{ and } j = 1, 2, \cdots, N - q \end{cases}$$

End

$C_i = C_i \cup \{u_j | x^*_{i,j} = 1, 1 \le j \le N - q\}$ for $1 \le i \le q$.

End

Back to begin

5: **Step 4:** Output:$\{C_1, C_2, \cdots, C_q\}$ with $v_i \in C_i$ is a minimum q-cut on $\widetilde{G}$.

of $\widetilde{G}$ such that $\sum_{i<j, C_i, C_j \subset \widetilde{G}}^{q} w(C_i, C_j)$ is a minimum. Then there is a minimum partition of $\widetilde{G}$, say $\{\overline{C_1}, \cdots, \overline{C_q}\}$, is also a minimum partition G such that$\sum_{i<j, \overline{C_i}, \overline{C_j} \subset G}^{q} w(\overline{C_i}, \overline{C_j})$ is a minimum.

Proof: By the definition of $\widetilde{G}, \Delta = \sum_{i<j, C_i, C_j \subset \widetilde{G}}^{q} w(C_i, C_j)$. Since $w(C_i, C_j) = \sum_{i \in C_i, j \in C_j} s^1_{ij} + \sum_{i \in C_i, j \in C_j} \sum_{k=2} s^k_{ij}$ and the value of Δ is fixed because $\{C_1, C_2, \cdots, C_q\}$ is the minimum partition of $\widetilde{G}$. We know that $\sum_{i \in C_i \subset \widetilde{G}, j \in C_j \subset \widetilde{G}} s^1_{ij} = \sum_{i \in C_i \subset G, j \in C_j \subset G} s^1_{ij}$ by the definition of k-strength relationship. Therefore, we construct a minimum partition $\{\overline{C_1}, \cdots, \overline{C_q}\}$ in $\widetilde{G}$ by Algorithm 1 such that$\sum_{i \in \overline{C_i} \subset \widetilde{G}, j \in \overline{C_j} \subset \widetilde{G}} s^1_{i,j}$ is a minimum, then,$\{\overline{C_1}, \cdots, \overline{C_q}\}$ is the minimum q partition of G.

The proof is end.

2.5 Determine the number of the communities

It is not difficult to find out that if a piece of information is drop into a reality community, the information will stay within the community more often as the dense connections. According to the stochastic theory, the spectral properties of Markov process are able to naturally reveal the "stability" of a specific partition [12] [13]. Inspired by it, we propose an efficient method to determine the optimal number of communities in a social network.

Theorem 2: *Let P be the generalized transition matrix of G. Then the optimal number of communities is opt* $= \arg\min_z(\frac{\log|\lambda_{z-1}|}{\log|\lambda_z|})$, *where λ_z is the z largest eigenvalue of matrix G.*

Proof: The transition probability present an ability that a node diffuse the information or disease to others, it positive related to the links' weight. The transition probability matrix $P = (p_{i,j})$ is defined as

$$p_{i,j} = \frac{r_{i,j}}{\sum_{j=1}^{N} r_{i,j}}, \tag{5}$$

where $r_{i,j} = \langle s_{i,j} \rangle_k$ is the average k-strength relationship value across all k. Via this representation, our framework can be utilized for the purposes of community detection analysis. Let P be the transition probability matrix, we have:

$$P = D_C^{-1} C, \tag{6}$$

where D_C is the diagonal degree matrix of $R = (r_{i,j})$. Let $p^{(\tau)}_{i,j}$ be the probability of hitting unit j after τ steps starting from unit i, we have:

$$p^{(\tau)}_{i,j} = (P^\tau)_{i,j}. \tag{7}$$

For this ergodic Markov process, P^τ corresponds to the probability of transitions between states over a period of τ time steps. To compute the transition matrix P^τ, the eigenvalue decomposition of P is used. If λ_k with $k = 0, ..., N-1$ denote the eigenvalues of P, and its right and left eigenvectors f_k and h_k are scaled to satisfy the orthonormality relation [21]:

$$f_k h_l = \delta_{kl}, \tag{8}$$

the spectral representation of P is given by

$$P = \sum_k \lambda_k f_k h_k, \tag{9}$$

and consequently

$$P^\tau = \sum_k \lambda_k^\tau f_k h_k. \tag{10}$$

We assume that eigenvalues of P are sorted such that $\lambda_0 = 1 > |\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_{N-1}|$. The convergence of every initial distribution to the stationary distribution $P^{(0)}$ corresponds to the fact that the spin of whole system ultimately reaches exactly the same value as time goes on. This perspective belongs to a timescale $\tau \to \infty$, at which all eigenvalues λ_k^τ go to 0 except for the largest one, $\lambda_0^\tau = 1$. In the other extreme of a timescale $\tau = 0$, P^τ becomes the identity matrix. All of its columns are different, and the system disintegrates into as many communities as the elements there are.

For the purposes of community identification, intermediate timescales are of interest, on which many but not all of the eigenvalues are practically zero. If we want to identify z communities, we expect to find P^τ at a timescale, the eigenvalues λ_k^τ of which are significantly different from zero only for the range $k = 0, ..., z-1$. This is achieved by determining τ such that $|\lambda_z|^\tau \approx 0$. Using a parameter $\zeta \ll 1$ which is considered to be practical zero, we require $|\lambda_z|^\tau = \zeta$ to determine the appropriate hitting time for the whole system entering into a metastable state with z different communities:

$$\tau(z) = \frac{\log \zeta}{\log |\lambda_z|}. \tag{11}$$

The vanishing of the smaller eigenvalues at a given timescale describes the loss of different states, and the removal of the structural features encoded in the corresponding weaker eigenvectors. We define the stability of z community structure, N_z, as the ratio between the hitting time and exiting time of z-state, $\tau(z)$ and $\tau(z-1)$:

$$N_z = \frac{\log \zeta / \log |\lambda_z|}{\log \zeta / \log |\lambda_{z-1}|} = \frac{\log |\lambda_{z-1}|}{\log |\lambda_z|}. \tag{12}$$

Because of $\log \zeta / \log |\lambda_z| \leq \log \zeta / \log |\lambda_{z-1}|$, it is easy to show $0 \leq N_z \leq 1$, and a smaller N_z implies a better community structure. For real networks, the label of the smallest N_z can be used to estimate the natural number of communities, *opt*, in a given network:

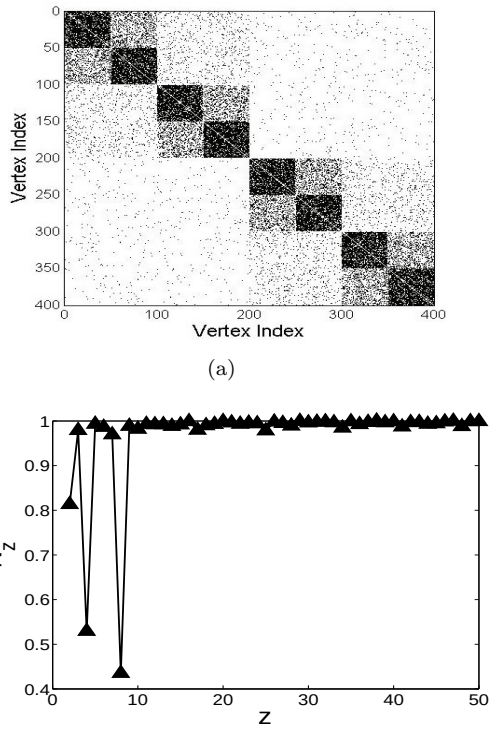


Figure 1: (a) A hierarchical network with 3-level community structure with 400 nodes. Due to heavy link density, it most likely contains eight small communities. Each two small communities are contained in a moderate community and finally the whole network is partitioned into two big sparse ones. (b) The stability N_z versus the number of communities z.

$$opt = \arg\min_z(N_z). \tag{13}$$

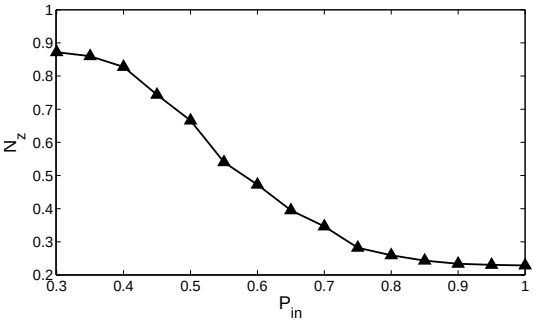
where arg represents the optimal z, at which N_z is minimized.

The proof is end. The complexity of this part mainly depends on computing the eigenvalue and corresponding eigensystem of matrix G. This procedure costs $O(n^2 \log n)$ times.

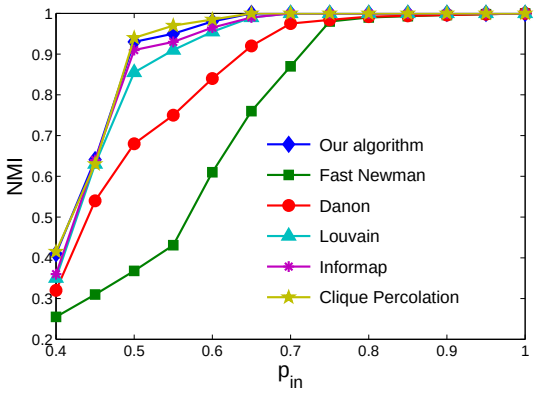
To show that our method can discover the optimal community number of hierarchical structures [20] [22] in different scales, Fig.1 give a representative example of the multi-level community structures. In this case, the number of N_z ($z > 1$) approaching to zero reveals the actual number of hierarchical levels hidden in a network. Furthermore, the significance of such levels can be quantified by their corresponding values of N_z.

3 Results

In this section, we will test the performance of our algorithm. Three experiments are designed and implemented for two main purposes: (1) to evaluate the

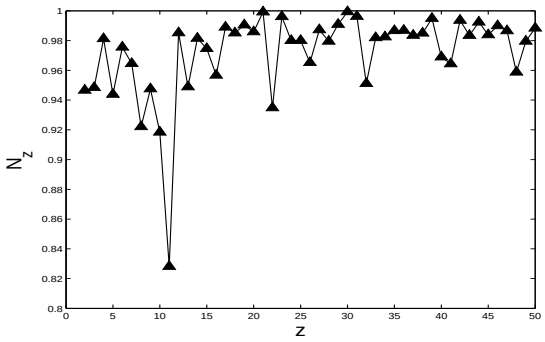


(a)

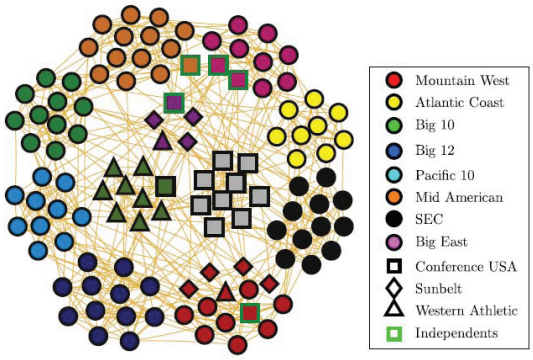


(b)

Figure 2: (a) $N_z(z = 4)$ values of networks versus different P_{in}. (b) Comparison of accuracy of our algorithm with other five existing algorithms.



(a)



(b)

Figure 3: (a) Computational results of N_z with different z on US football network. (b)Computational results of our algorithm on the football team network. The nodes with the same shapes and colors are teams in the same group, and the dense subgraphs in the layout are communities detected by the algorithm.

accuracy of the algorithm; (2) to apply it to real large-scale networks.

3.1 Benchmark network

We empirically demonstrate the effectiveness of our algorithm through comparison with other five well-known algorithms on the artificial benchmark networks. These algorithms include: Newman's fast algorithm [1], Danon et al.'s method [23], the Louvain method [24], Infomap [25], and the clique percolation method [26]. We utilize widely used Ad-Hoc network model, which can produce a randomly synthetic network containing 4 predefined communities and each has 32 nodes. The average degree of nodes is 16, and the ratio of intra-community links is denoted as P_{in}. As P_{in} decreases, the community structures of Ad-Hoc networks become more and more ambiguous, and correspondingly, their N_4 values climb from 0 to 1, as shown in Fig.2(a).

We use the normalized mutual information (NMI) measure [28] to qualify the partition found by each algorithm. We ask the question whether the intrinsic scale can be correctly uncovered. The experimental results are illustrated in Fig.2(b), where y-axis represents NMI value, and each point in curves is obtained by averaging the values obtained on 50 synthetic networks.

As we can see, all algorithms work well when P_{in} is more than 0.7 with NMI larger than 0.85. Compared with other five algorithms, our algorithm performs the best. Its accuracy is only slightly worse than that of the clique percolation when $0.5 \leq P_{in} \leq 0.65$. However, the complexity of the clique percolation is more than $O(n^3)$ and nearly the same as the time consuming Breadth First Search(BFS). By contrast, the time complexity of our method is very low($O(n^2)$) and can be easily implemented.

In addition to the Ad-Hoc network, we have also tested for ring of cliques networks with 1000 cliques and 10 nodes in each. The cliques benchmark network has been introduced to show that the modularity measure (and some other measures) faces resolution limit problem. The network is created by connecting k complete networks on a ring. The model has 2 parameters namely, the number of cliques and the size of cliques. As the result, the algorithm was able to find the communities, i.e. cliques, perfectly.

3.2 US Football network

The United States college football team network has been widely used as a benchmark example [1] [27] due to its natural community structure. We used the data gathered by Girvan and Newman [1]. It is a representation of the schedule of Division I American Football games in the 2000 season in USA. The nodes in the network represent the 115 teams, while the edges represent 613 games played in the course of the year. The whole network can be naturally divided into 12 distinct groups. As a result, games are generally more frequent between members of the same group than between members of different groups.

First, we calculate N_z and the results are illustrated in Fig.3(a). Results show that the optimal number of communities is $opt = 12$, which perfectly agree with the true situation. Then we apply our algorithm to the football team network and partitions the network into 12 communities, which is shown in Fig.3(b). The correct rate of our method is more than 93%, which means that the detected community structure is in a high agreement with the true community structure. Actually, methods based on optimization of modularity Q usually can just find 11 communities and the correct rate is low due to the fuzziness of the network. We concludes that the ability of our method to reveal a natural characteristic is valuable for many real networks. Furthermore, the misidentified nodes can be viewed as the interesting overlapping nodes which described as yellow triangles. The nodes are all fuzzily lie at the boundary communities and can be viewed as some relative independent clubs which can be interpreted readily by the human eye.

4 Conclusion

In this paper, we have designed an efficient algorithm to detect community in social network using a new definition, i.e. k-strength relationship, which naturally represent the coupling degree between two nodes. Theoretical analysis shows this algorithm is polynomial time which much better than most existing ones. Then, to obtain the the optimal number of communities, we propose a new parameter-free method using the eigenvalue gap of specific transition matrix. Finally, we apply our algorithm on both benchmark network and real networks to evaluate its efficiency. Theoretical analysis and experiments show that the algorithm is able to uncover the communities fast and accurately, which can be easily extended to large scale real networks.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable suggestions which are very helpful for improving the manuscript. The authors are separately supported by NSFC grants 71401194, 91324203, 11131009 and "121" Youth Development Fund of CUFE grants QBJ1410.

References

[1] M.E.J.Newman, Phys. Rev. E., 2004, 69,pp.066133.
[2] M.E.J.Newman, and M.Girvan, Phys. Rev. E.,2004, 69,pp.026113.
[3] M.E.J.Newman, Proc. Natl. Acad. Sci.,2006, 103, pp.8577-8582.
[4] M. R. Garey, and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, San Francisco, 1979.
[5] J. Scott, Social Network Analysis: A Handbook(Sage Publications, London, 2000, 2nded).
[6] M.Girvan and M.E.J.Newman, Proc. Natl. Acad. Sci., 2002, 99,pp.7821-7826.
[7] X.S.Zhang, R.S.Wang, Y.Wang, J.Wang, Y.Qiu, L.Wang,and L.Chen., Eur.Phys.Lett., 2009, 87, pp.38002.
[8] X.S.Zhang, Z.Li, R.S.Wang,and Y.Wang. J.Comb.Optim., 2012, 23,(4),pp.425-442.
[9] F. Radicchi, C. Castellano,and F. Cecconi., Proc. Natl. Acad. Sci., 2004, 101, pp.2658-2663.
[10] M. Latapy and P.Pons., Proceedings of the 20th International Symposium on Computer and Information Sciences, Lecture Notes in Computer Science. , 2005, 3733,pp.284-293.
[11] B.W.Kernighan and S. Lin., Bell System Technical Journal. , 1970, 49,pp.291-307.
[12] H.J.Li and X.S.Zhang., Eur.Phys.Lett., 2013, 103,pp.58002.
[13] H.J.Li,Y.Wang,L.Y.Wu,J.Zhang, and X.S.Zhang., Phys.Rev.E., 2012, 86,pp.016109.
[14] U. Brandes, et al. arXiv.org:physics., 2006,pp.0608255.
[15] L.-C. Huang, T.J. Yen, and S. C. T. Chou., 2011 international conference on advances in social networks analysis and mining, IEEE computer society, 2011,pp. 110-117.
[16] Peter J. Mucha, et al., Science,2010, 328,pp.876-878.
[17] N. Guttmann-Beck and Hassin., Algorithmica, 2000, 27, pp.198-207.
[18] M. R. Garey and D. S. Jonson, Computers and intractability:A guide to the theory of NP-completeness, Freeman, San Francisco, CA, 1979.
[19] D.P.Maki., Mathematical Models and Applications, with Emphasis on social, life, and Management sciences., Prentice Hall, 1973.
[20] Z.Y.Xia, Z.Bu., Knowl-Based Syst.,2012, 26, pp.30-39.
[21] W.N.E, T.Li, E.Vanden-Eijnden, Proc. Natl. Acad. Sci., 2008, 105,pp.7907-7912.
[22] E. Ravasz and A. L. Barabási, Phys. Rev. E, 2003, 67,pp. 026112.
[23] L.Danon, J,Duch, D.Guilera,and A.Arenas, J. Stat. Mech., 2005, 29, P09008.
[24] V.D.Blondel,J.L.Guillaume,R.Lambiotte,and E.Lefebvre, J. Stat. Mech., 2005, 10, P10008.

[25] M.Rosvall and C.T.Bergstrom, Proc. Natl. Acad. Sci., 2008, 105,(4), pp.1118-1123.
[26] G.Palla, I.Derényi, I.Farkas,and T.Vicsek, Nature., 2005, 435,pp.814-818.
[27] Z.P.Li,S.H.Zhang,R.S.Wang,X.S.Zhang,and L.Chen, Phys. Rev. E., 2008, 77,pp.036109.
[28] A.Lancichinetti and S.Fortunato, Phys. Rev. E., 2009, 80,pp.056117.