The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

A Random Iterative Algorithm for Community Detection*

Junhua Zhang[1,2,†] Shihua Chen[3]

[1]Academy of Mathematics and Systems Science, CAS, Beijing 100190, China
[2]Key Laboratory of Random Complex Structures and Data Science,
Academy of Mathematics and Systems Science, CAS, Beijing 100190, China
[3]College of Mathematics and Statistics, Wuhan University, Wuhan 430072, China

Abstract Research on community structure detection in complex networks has attracted a great deal of attention in recent years. In this paper we propose a random iterative algorithm to uncover meaningful communities. The algorithm starts with initial population creation. Each individual of the population is encoded with the community identifiers of the nodes in the network, so it is a potential solution of the community structure of the network considered. Nodes are randomly assigned into communities at the beginning of the algorithm. At each iteration some nodes are randomly selected, their community identifiers are reassigned according to the modularity function and the measure of information discrepancy based on the shortest path profiles of nodes in the network. In the end, a proper community structure can be detected by the identifiers encoded in the individual with the largest modularity. The algorithm does not need any prior knowledge about the number of communities and can give an appropriate number by maximizing the modularity function. The computational results of the method on real-world networks confirm its capability.

Keywords Community detection; random iteration; complex networks

1 Introduction

Many real-world systems can be represented by networks composed of vertices and edges, such as the Internet [1], social networks [2, 3], biological networks [4, 5], the food webs [6] and et al. For complex networks, apart from small-world property, power-law degree distribution and network transitivity, one common typical property is the community structure, i.e. the division of networks into groups (also called clusters) having dense intra-connections, and sparse inter-connections. Detecting this community structure is fundamental for uncovering the links between structure and function in complex networks, and has a lot of applications in many different disciplines such as biology and social sciences.

Identifying community structure in complex networks has been receiving a great deal of attention and many techniques have been proposed for this purpose in recent years. For example, the hierarchical (agglomerative and divisive) clustering method [2, 7, 8], clique

*This work is partially supported by the Innovation Project of Chinese Academy of Sciences, kjcx-yw-s7.
†Corresponding author: zjh@amt.ac.cn

percolation [9, 10], spectral algorithm [11, 12, 13], Potts model [14, 15, 16], and so on. A huge number of methods have been recently reviewed and evaluated in Refs. [17, 18].

Among all the techniques for detecting communities, there are a great many algorithms proposed based on maximizing a modularity Q, which is introduced by Newman and Girvan [19] and has been broadly used as a valid measure for community structure. Specifically, the modularity function Q is defined as

$$Q = \sum_{c=1}^{k} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right] \tag{1}$$

where the sum is over the k communities of the partition, l_c is the number of links inside community c, L is the total number of links in the network, and d_c is the total degree of the nodes in community c. The modularity function provides a way to determine if a partition is valid to decipher the community structure in a network. Maximization of the modularity function Q over all the possible partitions of a network is usually an effective method [19, 20, 21]. So based on the modularity function, many methods have been developed [13, 22, 23, 24, 25].

As a matter of fact, finding an exact optimal solution for partitioning a network to detect the community structure is believed to be an NP-complete problem and therefore difficult to solve. Many existing methods are computationally exhaustive especially for large networks. In recent years a kind of approach has been proposed based on genetic algorithm or evolutionary computation which provides promising algorithms for solving NP-hard problems. They provide good (acceptable) solutions for community detection in complex networks in a reduced amount of time [26, 27]. The main drawback of these evolutionary techniques is the parameter problem, that is, there are too many parameters needing to be determined in the practical use, besides the population size, the iteration count, some others are the crossover proportion, the mutation rate, the threshold for clean-up and when to start clean-up [26]. This heavily limits its further applications.

In this study we propose a random iterative algorithm to uncover meaningful community structure in complex networks. First of all, we introduce the shortest path profile of each node which can characterize its overall connection information in a network, as well as the measure of information discrepancy (MID) [28] to measure the distance of any two nodes in the network. The algorithm starts with initial population creation. Each individual of the population is encoded with the community identifiers of the nodes, so it is a potential solution of the community structure of the network considered. And nodes are randomly assigned into communities at the beginning of the algorithm. At each iteration some nodes are randomly selected, their community identifiers are reassigned according to the MID measure and the Q value. In the end, a proper community structure can be detected by the identifiers encoded in the individual with the largest modularity Q. The algorithm does not need any prior knowledge about the number of communities and can give an appropriate number by maximizing Q. Comparative to the genetic algorithm [26], the method proposed here need not the operations such as crossover, mutation and clean-up, so there are no many parameters to be determined here, this increases its feasibility in practical use. The computational results of the method on real-world networks confirm its capability.

2 Shortest path profile and the measure of information discrepancy

For each node i of a given connected network with n nodes, we can get an n-dimensional vector $D_i = (d_{i1}, d_{i2}, \ldots, d_{in})$, where d_{ij} denotes the shortest path from node i to node j $(i, j = 1, \ldots, n)$. By normalizing D_i, we can obtain a new vector $P_i = (p_{i1}, p_{i2}, \ldots, p_{in})$, which satisfies $\sum_{j=1}^{n} p_{ij} = 1, i = 1, \ldots, n$, and is referred to as the shortest path (SP) ***profile*** of node i here. The SP profile well deciphers relationship of one node with the whole network. A direct idea is that if two nodes i and j have similar SP profiles, they must have very close link relationship.

To measure the similarity of SP profiles of any two nodes, the MID measure proposed by Fang [28] is used, that is

$$B(P_1, P_2) = \sum_{i=1}^{2} \sum_{j=1}^{n} p_{ij} \cdot \ln \frac{p_{ij}}{\sum_{i=1}^{2} p_{ij}/s} \tag{2}$$

where $P_i = (p_{i1}, p_{i2}, \cdots, p_{in})$ is an n-dimensional distribution and now represents the SP profile of node i, and $0 \cdot \ln \frac{0}{0}$ is defined as 0 as in the Kullback-Leiber entropy [29]. The measure B has a close relationship with Shannon entropy and has many good properties, such as non-negativity, identity, symmetry, boundedness, uniform continuity, monotonicity, maximum, convexity and so on [30]. As a matter of fact, the MID measure has some satisfactory applications in bioinformatics and other fields [31, 32]. Moreover, the advantage of the MID measure over the traditional Euclidean distance for measuring the similarity of nodes based on SP profiles has been discussed in [25].

3 The new random iterative algorithm

The algorithm is carried out by iteration. At every iteration the search process is operated on a population, each individual of which encodes a potential solution of the community structure of the network considered. So our algorithm starts with initial population creation. For a network with n nodes, each individual of the population has n elements, each of which represents the community identifier (*CommId*) of the corresponding node. And nodes are randomly assigned into communities at the beginning of the algorithm. Then at each iteration the modularity function Q is calculated for each individual, and the community identifiers of some randomly selected nodes are adjusted according to the MID measure and the Q value. In the end, the proper community structure can be obtained by the identifiers encoded in the individual with the largest modularity Q.

In detail, given an n-node network $G(V, E)$ consisting of the node set V and the edge set E, the iterative procedure of our algorithm works as follows:

1. For the first iteration, set $s = 1$.
2. An initial population with predetermined size N_p is created: $\mathscr{H} = \{H_1, \cdots, H_{N_p}\}$. Each individual H_i is an n-dimensional vector $H_i = (h_{i1}, \cdots, h_{in})$, where h_{ij} denotes the community identifier of node j in individual i $(i = 1, \cdots, N_p, j = 1, \cdots, n)$. Initially, each h_{ij} is randomly assigned a value between 1 and n.

3. For $i = 2, \cdots, N_p$, some predetermined proportional nodes are randomly selected from the individual H_i. For example, n_r nodes, where n_r is an integer, are denoted by $V_i = \{v_{i1}, \cdots, v_{in_r}\}$. For each v_{ij} (i.e., node j of H_i), we calculate its degree $deg_j = \sum_{k=1}^{n} a_{jk}$, where $[a_{jk}]_{n\times n}$ is the adjacency matrix, that is, $a_{jk} = 1$ if $(j,k) \in E$ and otherwise $a_{jk} = 0$.
   (a) If $deg_j = 1$, set $CommId(v_{ij}) = CommId(v_{j_0})$, where v_{j_0} is the unique neighbor of node v_{ij} in the network $G(V,E)$.
   (b) If $deg_j = 2$, let v_{j_1} and v_{j_2} be the two neighbors of v_{ij} in $G(V,E)$.
      i. If $deg_{j_1} \geq deg_{j_2}$, set $CommId(v_{ij}) = CommId(v_{j_1})$ and $CommId(v_{j_2}) = CommId(v_{j_1})$.
      ii. If $deg_{j_1} < deg_{j_2}$, set $CommId(v_{ij}) = CommId(v_{j_2})$ and $CommId(v_{j_1}) = CommId(v_{j_2})$.
   (c) If $deg_j \geq 3$, let S_j denote the set of all neighbors of node v_{ij} in the network $G(V,E)$, i.e., $S_j = \{v_{j_1}, \cdots, v_{j_m}\}$ $(m \geq 3)$. According to (2) we calculate the MID measure between v_{ij} and its every neighbor : $B_{j,k} = B(P_j, P_{j_k})$ where P_j and P_{j_k} represent the SP profiles of nodes v_{ij} and v_{j_k}, respectively $(k = 1, \cdots, m)$. Let $\bar{B}_j = (1/m)\sum_{k=1}^{m} B_{j,k}$.
   For $k = 1, \cdots, m$, if $B_{j,k} \leq \gamma \bar{B}_j$ (γ is a positive parameter):
      i. and if $deg_j \geq deg_{j_k}$, set $CommId(v_{j_k}) = CommId(v_{ij})$;
      ii. else if $deg_j < deg_{j_k}$, set $CommId(v_{ij}) = CommId(v_{j_k})$.
4. For $i = 1, \cdots, N_p$, the modularity Q in (1) is calculated for each H_i, which can be used to evaluated the community structure encoded in it. We denote these values by $Q_{s,1}, \cdots, Q_{s,N_p}$, respectively. Then reorder the individuals $\{H_i\}_{1\leq i\leq N_p}$ in the population $\mathscr{H}$ according to $Q_{s,i}$ from large to small.
5. Set $s := s + 1$, return to step 3, until $Q_{s,1} - Q_{s-80,1} < 0.005$, the iteration stops. And the community structure of the network $G(V,E)$ can be detected through the community identifiers encoded in the individual H_1 at the current iteration.

The iterative procedure can be easily implemented with Matlab [33] programming. Both the source code and the data for testing can be obtained from the authors.

4 Experiments

We test the performance of the proposed algorithm here by applying it to several well-studied real-world networks. In our experiments we use $N_p = 50$ as the population size. For the value of the parameter γ in (c) of step 3, generally we choose one number from $\{1.0, 1.1, \cdots, 1.5\}$, depending on that which can make the algorithm get a larger Q value. And in the iteration $n_r = [n/8]$ nodes are randomly selected every time to be reassigned community identifiers, where n is the size of the network considered and $[x]$ means the maximal integer no more than x. It is the randomness of the selection of the nodes that different runs may bring forth different results. In the following the best one is reported through about 10 runs for each experiment.

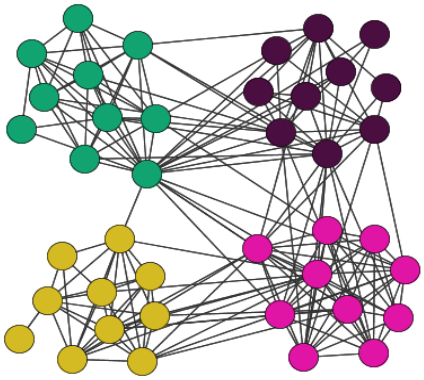

Figure 1: The community structure of journal index network obtained by the proposed method.

4.1 The journal index network

The journal index network constructed by Rosvall and Bergstrom [34] consists of 40 journals as nodes from four different fields: physics, chemistry, biology and ecology, and 189 links connecting nodes if at least one article from one journal cites an article in the other journal during 2004. 10 journals with the highest impact factor in the four different fields were selected.

Exactly the same communities as the actual partition are obtained using our algorithm with $\gamma = 1.1$, where $Q = 0.4783$ (Figure 1). Furthermore, when $\gamma = 1.2$ is used 3 groups are obtained with $Q = 0.4197$ where physical and chemical journals are incorporated into a single module. So do ecological and biological journals when γ takes a value among 1.3, ..., 1.6 and 1.7, therefore we get 2 groups at this time with $Q = 0.3981$. But when γ is a little small or too large, unsatisfactory results occur. At this situation either a single journal (node) is split from its original field or parts of different fields are combined to form a group.

4.2 The scientific collaboration network

The scientific collaboration network collected by Girvan and Newman [7] is a widely used test example for methods of detecting communities in complex networks [7, 12]. This network consists of 118 nodes (scientists) and 200 edges.

Different from the journal index network ((Figure 1)), this network is a little sparse in which more than 20 nodes with only degree 1, so it looks a little like stelliform subgraphs for some parts of the network because some nodes have very high degrees therein. Now the algorithm is used with $\gamma = 1.5$ and we get 7 groups where $Q = 0.7456$. Figure 2 shows the detected community structure which is visually very reasonable.

4.3 The football team network

The third real network we have tested is the college football network of the United States, which represents the game schedule of the 2000 season of Division I of the US

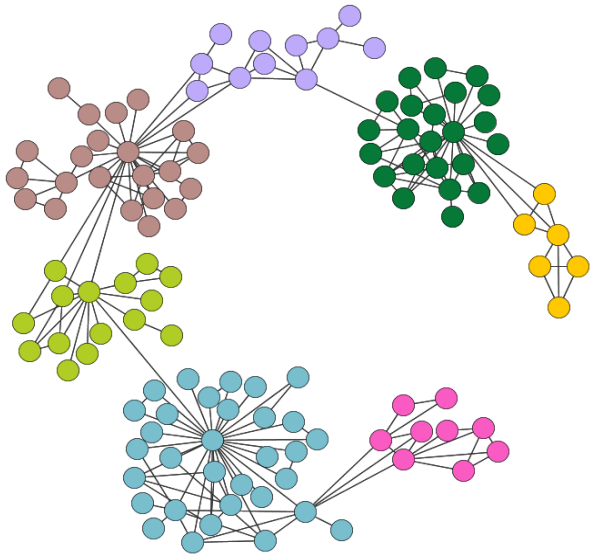

Figure 2: The community structure of scientific collaboration network obtained by our method.

college football league. The nodes in the network represent the 115 teams, while the links represent 613 games played in the course of the year. The teams are divided into 12 conferences of 8-12 teams each and, generally, games are more frequent between members of the same conference than between teams of different conferences. The natural community structure in the network makes it a commonly used benchmark for community-detecting algorithm testing [7, 12].

This network is quite dense because the average degree of every node is more than 10. Using our algorithm with $\gamma = 1.1$ eleven communities are detected with $Q = 0.5932$ (Figure 3). All of them correspond almost exactly to the original conferences except 9 nodes. Three nodes including 60, 64, 98 are classified in the other two groups for their weak link with their original group. Node 111 is classified into an inconsistent group for its more links with current group than its original one just as the papers [7, 31, 25] have obtained. Because there are few edges among all five members of the 12th conference, these five nodes are distributed to other groups due to their more links with those groups. It is very interesting that our algorithm can correctly classify node 59 into its original group although many papers [7, 31, 25] have mislaid it. The community structure found by our model seems to suggest a more precise organization than original conferences.

5 Discussion and conclusion

In this paper we propose a random iterative algorithm for community detection in complex networks. At the beginning of the algorithm we randomly create a population

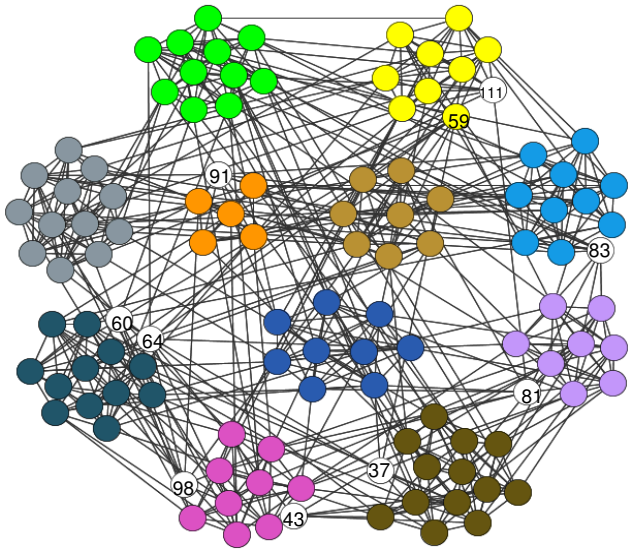

Figure 3: The community structure of football network obtained by our method.

with N_p individuals whose elements are the community identifiers of the nodes in the network considered. During each iteration n_r nodes are randomly selected to be reassigned identifiers according to the interesting SP profile, the MID measure and the modularity Q for all individuals except the first one. In the end, a proper community structure can be detected by the identifiers encoded in the first individual because it has the largest modularity.

The number of the individuals in the population is directly affecting the performance of the algorithm. However increasing the size may take a little more time to execute the algorithm and it does not yield better results after some point. Here an appropriate value $N_p = 50$ is used which takes into account these two factors.

The value of n_r is related to the speed of the algorithm. Larger n_r can make the algorithm faster, but, on the other hand, it is easily subject to a local minimum. After a lot of trial $n_r = [n/8]$ is adopted for our experiments and satisfactory results are obtained.

The parameter γ is an tuning factor of the mean MID measure around a selected node to determine which neighbors need to be reassigned the identifiers. Just like the discussion in the experiment 4.1, adjusting the value of γ in a proper range can sometimes provide multiresolution community structure for us.

The algorithm finishes the search process of finding the best community structure if there is almost no change in the Q value between 80 iterations. In fact, the algorithm is always able to end in 200 iterations in our experiments.

Although the population creation in our algorithm is similar to that in the genetic algorithm [26], the procedure of our method differs from the latter greatly. Namely, our algorithm doesn't need any operation like crossover, mutation or clean-up, thus there are

not so many parameters to be determined here, this increases its feasibility in practical use. We hope that this new method will be a helpful complementarity in the detection of communities in complex networks with practical significance, and we expect that it will be employed with promising results in this field.

References

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. 29(4), 251-262 (1999).

[2] J. Scott, Social Network Analysis: A Handbook, 2nd ed. (Sage Publications, London, 2000).

[3] M. E. J. Newman and J. Park, Phys. Rev. E 68, 036122 (2003).

[4] E. Almaas, B. Kov*á*cs, T. Vicsek, Z. N. Oltvai, and A.-L. Barab*á*si, Nature 427, 839-843 (2004).

[5] F. Rao and A. Caflisch, J. Mol. Biol. 342, 299-306 (2004).

[6] J. A. Dunne, R. J. Williams, and N. D. Martinez, Proc. Natl. Acad. Sci. USA 99, 12917-12922 (2002).

[7] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).

[8] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl Acad. Sci. USA 101, 2658-2663 (2004).

[9] G. Palla, I. Der*é*nyi, I. Farkas, and T. Vicsek, Nature 435, 814 (2005).

[10] J. M. Kumpula, M. Kivel*ä*, K. Kaski, and J. Saram*ä*ki, arXiv:0805.1449 [physics.soc-ph] (2008).

[11] L. Donetti and M.A. Muñoz, J. Stat. Mech., P10012 (2004).

[12] F. Wu and B. A. Huberman, Eur. Phys. J. B 38, 331-338 (2004).

[13] M. E. J. Newman, Phys. Rev. E 74, 036104 (2006).

[14] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. 93, 218701 (2004).

[15] J. Reichardt and S. Bornholdt, Phys. Rev. E 74, 016110 (2006).

[16] P. Ronhovde and Z. Nussinov, arXiv:0803.2548 [physics.soc-ph] (2008).

[17] S. Fortunatoa and C. Castellano, arXiv:0712.2716 [physics.soc-ph] (2007).

[18] M. A. Porter, J.-P. Onnela, and P. J. Mucha, arXiv:0902.3788 [physics.soc-ph] (2009).

[19] M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004).

[20] M.E.J. Newman, Eur. Phys. J. B 38, 321-330 (2004).

[21] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas, J. Stat. Mech., P09008 (2005).

[22] R. Guimer*à* and L.A.N. Amaral, Nature 433, 895-900 (2005).

[23] J. Duch and A. Arenas, Phys. Rev. E 72, 027104 (2005).

[24] S. Zhang, R.S. Wang, and X.-S. Zhang, Phys. Rev. E 76, 046103 (2007).

[25] J. Zhang, S. Zhang, and X.-S. Zhang, Physica A 387, 1675-1682 (2008).

[26] M. Tasgin and H. Bingol, arXiv:cond-mat/0604419 (2006).

[27] A. Gog, D. Dumitrescu, and B. Hirsbrunner, F. Almeida e Costa et al. (Eds.): ECAL 2007, LNAI 4648, 886-894. Springer-Verlag Berlin Heidelberg (2007).

[28] W. Fang, Math. soc. sci. 28, 85-111 (1994).

[29] S. Kullback, Information Theory and Statistics. New York: Wiley (1959).

[30] W. Fang, Infor. Sci. 125, 207-232 (2000).

[31] W. Li, W. Fang, L. Ling, J. Wang, Z. Xuan, and R. Chen, J. Biol. Phys. 28, 439-447 (2002).

[32] M. Zhang, W. Fang, J. Zhang, and Z. Chi, Comput. Biol. Chem. 29, 175-181 (2005).

[33] http://www.mathworks.com/.

[34] M. Rosvall and C.T. Bergstrom, Proc. Natl. Acad. Sci. USA 104, 7327-7331 (2007).