Journal of Computational Information Systems 8: 23 (2012) 9807–9814
Available at http://www.Jofcis.com

Discovery of Community Structure in Complex Networks Based on Resistance Distance and Center Nodes

Bin LU [1,*], Changyu LIU [2], Yonghong WANG [3]

[1] *School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China*

[2] *Communication & Computer Network Lab of GD, South China University of Technology, Guangzhou 510642, China*

[3] *Guangzhou Information Engineering Vocational School, Guangzhou 510610, China*

Abstract

On the basis of analyzing the classical algorithms of network clustering and community detecting, a new algorithmic method of discovery of community was brought up, which has different thinking from previous algorithms. In this method a general network is looked upon as a corresponding electrical network firstly, the link in between the connected pairs of nodes is substituted by a fixed resistor, and then compute all the resistance distance between nodes. Simultaneously, according to the properties of small world and scale-free, each community has some center nodes with the larger degree. The algorithm decides the center nodes of community by using both the resistance distance and degrees of nodes; whereafter compares the resistance distance in between other nodes and center nodes for clustering the network. The experiment results show that the algorithm is stable and reliable and can accurately obtain the division of community structure.

Keywords: Complex Network; Community Structure; Discovery of Community; Resistance Distance; Center Node

1 Introduction

As the physical significance and mathematic feature of networks are intensively studied, people find many properties in real-world networks, such as the power-law degree distributions, average path length, betweenness and high clustering coefficients [1, 2]. There are other some topology characteristics of complex network evolution model,such as small-world and scale-free characteristics. And the so-called community structure is also another important topology property of complex network [3, 4], which have attracted a large amount of interest in the world recently. A community in networks is a group of nodes that are similar to each other and dissimilar from the rest of the network. Now, many people think a whole network is made up of communities,

*Corresponding author.
Email address: lbhqu@163.com (Bin LU).

December 1, 2012

and nodes within every community connect to each other densely and nodes in different community sparsely [5, 6]. Many complex systems, such as social networks [5], biochemical networks [7], and information networks such as the web [8, 9], have already been shown to possess strong community structure.

There is of great significance to researching the community structures of networks, for they often correspond to some functional units. And finding communities and attempting to analyze them become a very important way to know all kinds of network organization structures, and has lots of practical applications. For example, in a network of biomolecular, nodes in the same community generally take a specific role or have specific function, that can help us better understand the organizational principles of the cell, and identify important biological pathways for further studies [10, 11]; identifying communities from a collaboration network may suggest evolution and development of research areas and so on [12].

So, plenty of algorithms are put forward to discover communities in complex networks. Kernighan-Lin algorithm is a famous heuristic optional algorithm [13]. It introduces a gain function for network division, and its optimization goal is to obtain the minimum of the subtraction of within- and between-community edges. It uses a greedy strategy, so consequently the algorithm ends up in a local optimum rather than a global optimum. Moreover, the algorithm is very sensitive to the choice of the initial solution, and it partition a network into only two divide, so it has big limitations for practical application. Another well-known algorithm is Girvan and Newman's divisive approach (GN algorithm) [6], which is based on the iterative removal of edges with high "betweenness" scores that appears to identify such structure with some sensitivity. The main idea here is that edges that run between communities have higher betweenness values than those that lie within communities. By successively recalculating and removing edges with highest betweenness values, the network breaks down into disjoint connected components. Subsequently, Newman proposed a fast algorithm (FN algorithm) [14], which also used a greedy strategy. In the algorithm Newman quantified these statistical results by a measure called modularity (Q), and he evaluated the quality of community structure by optimizing the value of modularity (Q), then transformed community discovery to an optimization problem. But optimizing Q is a NP-hard problem, so it is almost certainly the case that no polynomial-time algorithm exists that will find the modularity optimum in all cases.

In recent years, spectral clustering methods rise gradually [15, 16], which apply to a complex network for partition now. These methods attempt to globally optimize cost functions such as the Normalized Cut measure, and the majority of these methods make great efforts to balance the size of the clusters while minimizing the interaction between dissimilar nodes. For instance, Newman made the modularity expressed in terms of the eigenvectors of a characteristic matrix for the network by a modularity matrix [17], and proposed a spectral approach used to partition community structure. All these methods depend on the eigenvectors of the Laplacian matrix or its relatives of a graph.

In this paper, we bring up a different method for the discovery of communities, which is based on the theory of resistive electrical network [18] and the properties of small world and scale-free. In the method, a general network expressed by an undirected weighted connected graph is first considered as an electrical network, and then the vertices of the graph correspond to junctions in the electrical network and the edges of the graph to unit resistors. Then the effective resistance between pairs of vertices is a graphical distance. At the same time, we know, if a network has communities, each community has more than one center nodes with the larger degree. So, our method attempt to use the resistance distance and the degree of nodes to determine center nodes of communities together, cluster networks, and then find community structure in networks.

2 The Method

Many researchers found a general network was similar to an electrical network one after another [18, 19]. So a general network expressed by a connected graph can be transforms into an electrical network.

In general, let an undirected weighted graph $G = (V, E, W)$ be a network, where V is a set of n vertices, E is a set of m edges between vertices, and W is a set of weight values on the edges and usually the edges are weighted by 1. According to the knowledge of electrical network, a network can be regard as an electrical network, as shown in Fig.1.

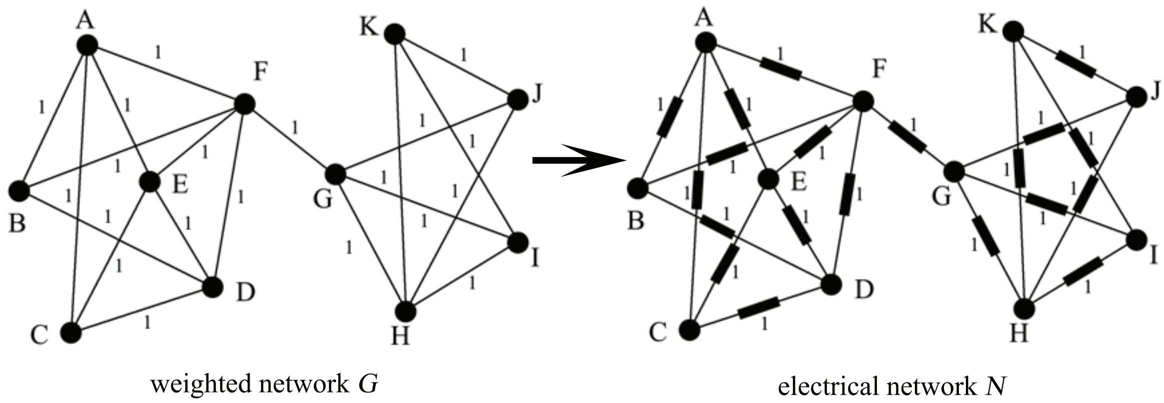

Fig. 1: network G transforming into corresponding electrical network N

Fig.1 illustrates a sample network turning into an electrical network. If every edge of G is replaced by a fixed resistor and weight values are looked upon as resistance values, an electrical network N comes out corresponding to G. And then, the resistance distance between the vertices of G, denoted by R_{ij}, defined to be the effective electrical resistance between nodes i and j of the network , that is, $R = (R_{ij})$ is called the resistance distance matrix of the network N. The quantities R_{ij} are calculated by methods of the theory of resistive electrical networks based on Ohm's and Kirchhoff's laws.

According to Ohm's law, there is an important property about the resistance distance. That is, in general, the more the path in between two nodes, the smaller the resistance distance in between them, and vice visa. This is corresponding to the clustering thought such as "a high density of within-group edges and a lower density of between-group edges", and this is the one of principal thought of our method.

However, by applying the Ohm's law, the computation of the resistance between any two nodes is very complex and fussy. The computation requires us to consider all the paths between them, thus it is not practical to process in batches automatically by computer for a network with a whole lot of nodes. Luckily, there are some other ways to compute resistance distance[19, 20]. Let $A = (A_{ij})$ be the adjacency matrix of G, and let $D = (D_{ij})$ be the diagonal degree matrix of and let $L = (L_{ij})$ be the Laplacian matrix of G. Usually, A is equal to W. Thereupon, the effective resistance in between node i and j can be computed as follows:

$$D_{i,j} = \sum_{j=1}^{n} A_{ij}. \tag{1}$$

$$L = D - A. \tag{2}$$

$$R_{i,j} = L^{+}_{i,i} + L^{+}_{j,j} - 2L^{+}_{i,j}. \tag{3}$$

Where L^{+} represents pseudo-inverse of L, $L^{+} = (L + \frac{1}{n}J)^{-1} - \frac{1}{n}J$ and J is the square matrix of order n all of whose elements are unity.

At first, we attempt to consider the two nodes with the farthest resistance distance as the center nodes of two communities. However, through careful experiment it partitions a network into only two divide like the Kernighan-Lin algorithm and is not accurate enough to partition a network by only applying the resistance distance sometimes. So, we bring to mind another thought about complex network. That is, the small-world and scale-free properties. A complex network with obvious community structures is prone to a small-world model, and every community is corresponding to a small world. And in each community, there are more than one node with the larger degree, and other nodes with the smaller degree. This severe unevenness is corresponding to the scale-free characteristic of the complex network as well. Then, those nodes with the larger degree, we call center nodes of communities, can be used to the discovery of community. In other words, we can study the relationship between a node and center nodes of communities, instead of studying any node belongs to which community. There are less than one center nodes, and finding the center nodes means finding communities.

Under ideal condition, the node with the largest degree can be look upon as the center node of the first community, and the node with the second largest degree can be look upon as the center node of the second community, and the rest similar. But it is possible that the largest nodes are in the same community and cannot partition into divides accurately in practice. Then, we partition a network by using both the resistance distance and the degree of nodes.

Suppose our network contains k communities. For a particular division of the network into k groups, we must determine k nodes as the center nodes of communities. Let C be a set used to save the center nodes of communities. Then, the number of elements in C is k, and C_i expresses the center node of the i-th community. Borrowing ideas from [21], an evaluation formula about whether to determine node i as the center node of a community is as follows:

$$M_i = \sum_{j \in C} D^{\alpha}_{ii} \times R^{\beta}_{ij}. \tag{4}$$

Where α and β respectively represents weight coefficient of degree of nodes and resistance distance, and $\alpha + \beta = 1$.

According to Eq.(4), we can obtain a search strategy, and its key steps are as follows:

(1) Compute the degrees of each node, and make the node with the largest degree as the center node of the first community, store it in C.

(2) From the node with the largest degree (which is not in C), compute the M-value by Eq.(4). Store the node with the maximal M-value in C.

(3) Repeat the step 2, until there are k nodes in C.

After obtain the center nodes, we compare the resistance distance in between the rest node and center nodes in turn. Then, for any node i, it belongs to community j if R_{i,C_j} is minimal in all the resistance distance in between node i and center nodes.

3 Experiments and Discussion

We present a number of tests of our method on computer-generated graphs and on real-world networks for which the community structure is already known. The results indicate that our method is stable and reliable. In each case we find that our algorithm reliably detects the known structure. In consideration of clear expression, two sample examples are presented to explain our method in detail.

First, we apply our method to a set of artificial, computer-generated graph G depicted in Fig.1. We cluster G into 2 clusters, that is to say $k = 2$, and the result is shown in Fig.2. With calculation of the degree of nodes, we know the node with the largest degree is node F, whose degree is 5, and then let it be the center node of the first community. Next, let both α and β be equal to 0.5, and then the maximal M-value appears at the node H, and its value is 2.7534; let node H be the center node of the second community. Finally two communities are partitioned by applying our method, and the simulation effect is shown in Fig.2.

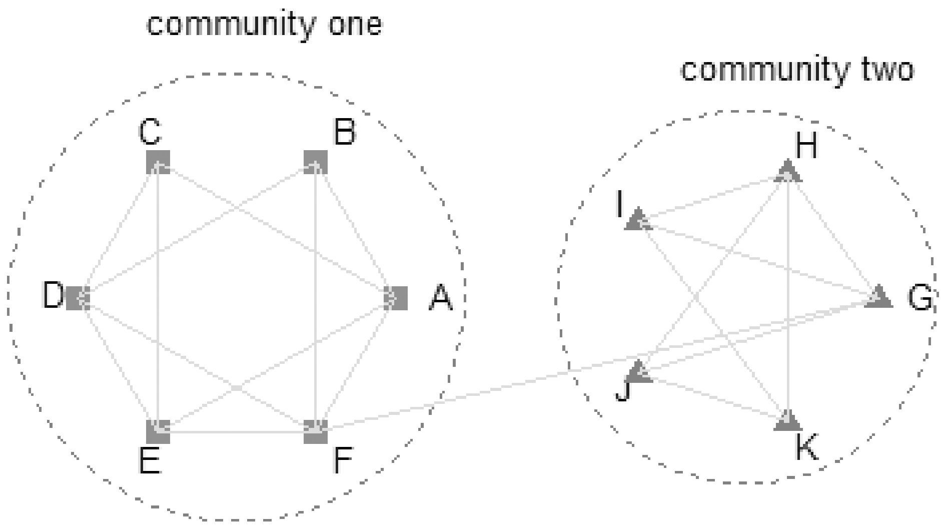

Fig. 2: Simulation effect of community finding on the network G

Nevertheless, if center nodes of communities are measured only by degree of nodes, then node F with the largest degree would be chosen to be the center node of the first community. Next, there are five nodes with the largest degree, which are node A, D, E, H, and G. It is very possible to choose node A, D, or E as the center node of the second community. But in fact node A, D, and E are in the same community with node F, so it is inaccurate to partition the network into communities.

If center nodes of communities are measured only by resistance distance, then, By calculating, we know the resistance distance in between node C and node K is maximal in all the nodes, whose value is 2.2738. Let the two nodes be center nodes of two communities, partition the network accurately but only into two divides. This method cannot partition the network into more two divides.

Second, we apply our method to sets of a few real-world networks also, such as Zachary's karate club friendship network[22]. The karate club network represents friendships between 34 members of a karate club at a US university, as recorded over a two-year period by Zachary. During the course of the study, the club split into two factions centered around the administrator and the teacher as a result of a dispute within the organization, and the members of one fraction left to start their own club. In the karate club network, the node with the largest degree is node 34, whose degree is 17. Let it be the center node of the first community. Let both α and β be equal to 0.5 likewise, and then the maximal M-value appears at the node 1, and its value is 2.0152; let node 1 be the center node of the second community. Finally two communities were partitioned by applying our method, shown in Fig.3. In Fig.3, squares represent individuals who ended up aligning with the clubs administrator after the fission of the club, up triangle those who aligned with the instructor.

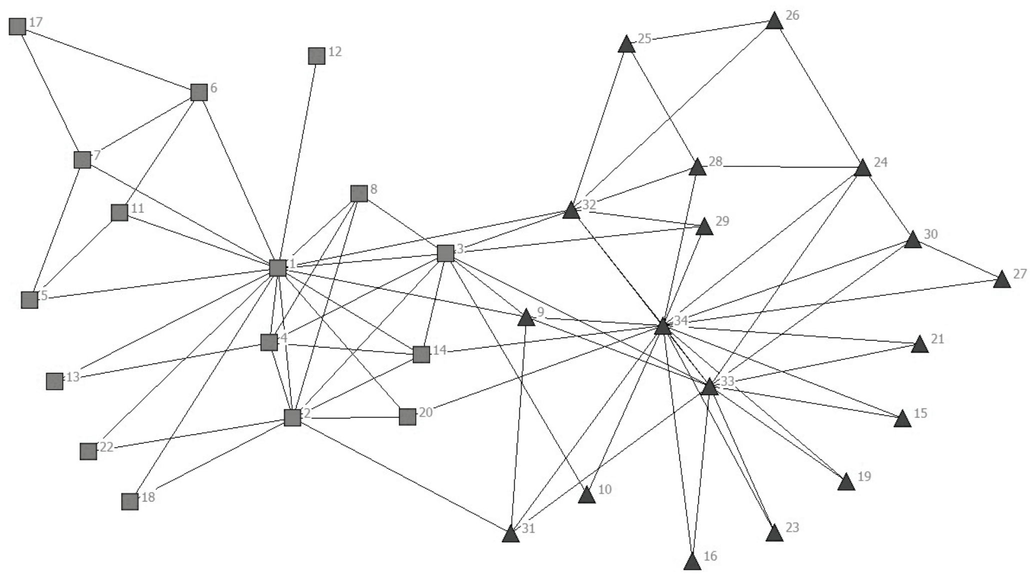

Fig. 3: Clustering result of Zacharys karate club network

We also measured center nodes of communities only by degree of nodes. Because there was only one node with the largest degree and the second largest degree respectively, and they were in the different communities. So, the two communities can be partitioned accurately.

When we measured center nodes of communities only by resistance distance, We knew the resistance distance in between node 12 and node 17 was maximal in all the nodes, whose value was 1.8333. Let the two nodes be center nodes of two communities, the two communities could not be partitioned accurately, for the two nodes were in the same community in fact. Consequently, the two factors must be considered for center node measurement.

In addition, our algorithm is more accurate than both GN algorithm and FN algorithm. GN algorithm divided node 3 of Zachary's karate club network into another community in error, and FN algorithm divided node 10 of Zachary's karate club network into another community by mistake. However, the result of our algorithm coincided with the reality.

In other test, we apply our method to another real-world network, bottlenose dolphin social network, which was analyzed by Lusseau et al.[23]. There were 62 bottlenose dolphins living in Doubtful Sound, New Zealand, with social ties between dolphin pairs established by direct observation over a period of several years. Then, in the network there were 62 nodes and edges were set between animals that were seen together more often than expected by chance. By our algorithm, node 15, the dolphin named as jet, and node 18, the dolphin named as Grin, became the center nodes of two communities. Then the dolphins separated in two groups accurately, represented by the circles and squares, and one group includes 41 members, and another group includes 21 members, shown in Fig.4.

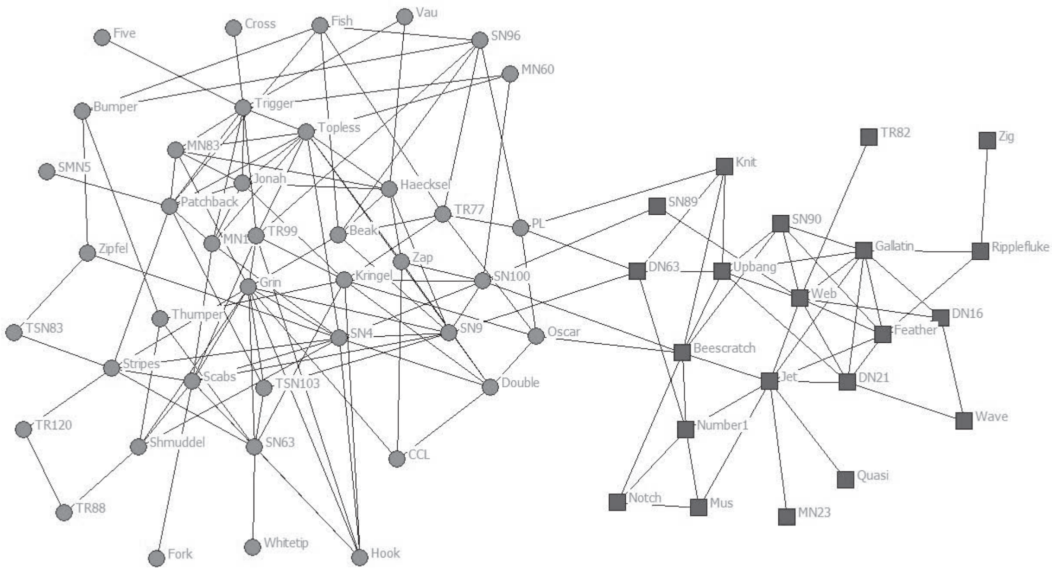

Fig. 4: Clustering result of Dolphin social network

4 Conclusions

In this paper a new algorithmic method has been described for discovering community structure from networks, which has different thinking from previous algorithms. The algorithmic method has been applied to a variety of real-world network data sets and computer-generated data sets. For all cases, our algorithmic method is able to detect right community structures, and the method is proved to be applicable and effectual. But it has a disadvantage that the cluster number must be preestablished. The next work is to decide the value range of the two parameters α and β, and research the properties of our method by more experiments. We find that the method was more effective to the networks with obvious community structures. We believe, our method can bring people much enlightenment and will have more wide range of applications, and it will be an important one in the big family of algorithms of detecting community structure of network.

References

[1] R. Albert, A. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74(2002), 47.

[2] M. E. J. Newman, The structure of scientific collaboration networks, in: Proc. Natl. Acad. Sci. USA, 98(2001), pp. 404-409.

[3] D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks. Nature, 393(1998), 440-442.

[4] A. L. Barabási, R Albert, Emergence of scaling in random networks. Science, 286(1999), 509-512.

[5] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E, 69(2), (2004), 026113.

[6] M. Girvan, M. E. J. Newman, Community structure in social and biological networks, in: Proc. Natl. Acad. Sci. USA, 99(12), (2002), 7821-7826.

[7] P. Holme, M. Huss and H. Jeong, Subnetwork hierarchies of biochemical pathways, Bioinformatics, 19(2003), 532-538.

[8] G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Co-etzee, Self-organization and identification of Web communities, IEEE Computer, 35(3), (2002), 66-71.

[9] L. Lu, F. X. Zhu, B Hu. A novel method to detect latent community in blogspace, Journal of Computational Information Systems, 6(7), (2010), 2151-2157.

[10] V. Spirin and L. A. Mirny, Protein complexes and functional modules in molecular networks, in: Proc. Natl. Acad. Sci. USA, 100(2), (2003), 12123-12128.

[11] D. Wilkinson, B. Huberman, A method for finding communities of related genes, in: Proc. Natl. Acad. Sci. USA, 101(2004), 5241-5248.

[12] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, Tracking evolving communities in large linked networks, Proc. Natl. Acad. Sci. USA, 101(2004), 5249-5253.

[13] B. W. Kernighan, S. Lin, An efficient heuristic procedure for partitioning graphs, Bell System Technical Journal, 49(2), (1970), 291-307.

[14] M. E. J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E, 69(2004), 066133.

[15] A. Ng, M. Jordan, and Y. Weiss, On spectral clustering: analysis and an algorithm, in: NIPS'01, 2001, pp. 849-856.

[16] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), (2000), 888-905.

[17] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E, 74(2006), 036104.

[18] D. J. Klein, M. Randic, Resistance distance, J. Math. Chem., 12(1), (1993), 81-95.

[19] W. J. Xiao, I. Gutman, Resistance distance and Laplacian spectrum, Theoretical Chemistry Accounts, 110(2003), 284-289.

[20] R. B. Bapat, I. Gutman, W. J. Xiao, A simple method for computing resistance distance, Z. Naturforsch A, 58(2003), 494-498.

[21] M. Liu, Research of New Clustering Algorithms to Mine Community Structure in Complex Networks, Changchun: Jilin university, 2010. (in Chinese).

[22] W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research, 33(1977), 452-473.

[23] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behavioral Ecology and Sociobiology, 54(2003), 396-405.