Journal of Mathematical Research with Applications
May, 2013, Vol. 33, No. 3, pp. 288–296
DOI:10.3770/j.issn:2095-2651.2013.03.003
Http://jmre.dlut.edu.cn

Finding Community Structure in Networks Using a Shortest-Path-Based k-Means Algorithm

Jinglu GAO
School of Mathematics, Jilin University, Jilin 130012*, P. R. China*

Abstract We consider the problem of detecting the community structure in a complex network, groups of nodes with a higher-than-average density of edges connecting them. In this paper we use the simulated annealing strategy to maximize the modularity, which has been indicated as a robust benefit function, associating with a shortest-path-based k-means iterative procedure for network partition. The proposed algorithm can not only find the communities, but also identify the nodes which occupy central positions under the metric of the shortest path within the communities to which they belong. The optimal number of communities can be automatically determined without any prior knowledge about the network structure. The applications to both artificial and real-world networks demonstrate the effectiveness of our algorithm.

Keywords community structure; modularity; shortest path; K-means; simulated annealing.

MR(2010) Subject Classification 05C85; 68M10; 90B18

1. Introduction

Recently, the structure and dynamics of complex networks have been frequently concerned in physics and other fields as a foundation for the mathematical representation of various complex systems [1–3]. Network models have also become popular tools in social science, economics, the design of transportation and communication systems, banking systems, etc, due to our increased capability of analyzing these models [4, 5]. Modular organization of networks, closely related to the ideas of graph partitioning, has attracted considerable attention, and many real-world networks appear to be organized into community structure that are densely connected within themselves but sparsely connected with the rest of the networks. A huge variety of community detection techniques have been developed to partition the network into a small number of communities [6–12], which are based variously on centrality measures, flow models, random walks, optimization and many other approaches.

In traditional clustering literature, the standard k-means algorithm is based on the optimization of a specified objective function with the known number of clusters [13]. Such objective function usually decreases as the number of clusters increases. However, people are sometimes required to determine the number of communities of the optimal network partition. To overcome

Received December 8, 2011; Accepted December 20, 2011
Supported by the National Natural Science Foundation of China (Grant No. 10771085).
E-mail address: jlgao@jlu.edu.cn

this weakness, we choose a robust benefit function of modularity [7–10] as a valid measure for network partition, which has larger values indicating stronger community structure. Then we apply the simulated annealing strategy to search the maximal value of modularity [14, 15]. The cooling process is operated with a k-means iterative procedure based on the measure of shortest path on networks. The proposed algorithm performs efficiently since the process of iteration can accelerate the tendency in maximizing the modularity function. The second merit is that this category of method can not only identify the community structure, but also give the nodes which occupy central positions within the communities which they belong to. The center of a community can convey the information that how an important status it holds among the members in the same group, since people sometimes are interested in the characterization of the communication in small groups and assumed a relation between structural centrality and influence in group processes [16, 17]. Another advantage is that an appropriate number of communities can be automatically determined without any a prior knowledge about the community structure.

We constructed our algorithm — simulated annealing with a shortest-path-based k-means algorithm (SASP) for network partition. The algorithm is tested on two artificial networks, including the ad hoc network and the sample network generated from Gaussian mixture model, with reasonable computational effort and leads to an accurate partitioning result. Applications to two real-world networks, including the karate club network and the dolphins network, confirm the capability of the method.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the measure of shortest path for proximity of nodes in networks and the concept of modularity. After reviewing the basic idea of simulated annealing, we propose our algorithm and the corresponding strategies in Section 3. In Section 4, we apply the proposed method to the representative examples mentioned before. Finally we make the conclusion in Section 5.

2. The measures and properties of network structure

2.1. The shortest path

In [18], a standard implementation of computing the shortest path matrix of a network is proposed. The concept of the shortest path in graph theory has been frequently used, which measures the extent of connection and communication between nodes in networks.

Let $G(S, E)$ be a network with n nodes and m edges, where S is the nodes set, $E = \{e(x, y)\}_{x,y \in S}$ is the weight matrix and $e(x, y)$ is the weight for the edge connecting the nodes x and y. We denote $d(x)$ as the degree of the node x whose entries are given by [19, 20]

$$d(x) = \sum_{z \in S} e(x, z). \tag{1}$$

Let $s(x, y)$ be the shortest path linking node x and node y, which indicates the minimal number of edges traversed to get form x to y. There may not be a unique shortest path between two nodes. The shortest paths between any two nodes x and y in the network can be calculated using the following procedure [18]:

(i) Assign node x distance zero, to indicate that it is zero steps away from itself, and set $d = 0$.

(ii) For each node z whose assigned distance is d, follow each attached edge to the node w at its other end and, if w has not already been assigned a distance, assign it distance $d + 1$. Declare z to be a predecessor of w.

(iii) If w has already been assigned distance $d + 1$, then there is no need to do this again, but z is still declared a predecessor of w.

(iv) Set $d = d + 1$.

(v) Repeat the process from step (ii) until there are no unassigned vertices left.

Now the shortest path from x to y is the path got by stepping from x to its predecessor, and then to the predecessor of each successive node until y is reached. If a node has two or more predecessors, then there are two or more shortest paths, each of which must be followed separately if we wish to know all shortest paths from x to y.

We take a partition of S as $S = \bigcup_{k=1}^{N} S_k$ with $S_k \bigcap S_l = \emptyset$ if $k \neq l$. Obviously, $s(x, y)$ will be small if x and y belong to the same community and large if they belong to different communities. The center $m(S_k)$ of community S_k can be defined as

$$m(S_k) = \arg\min_{x \in S_k} \frac{1}{|S_k|} \sum_{y \in S_k, y \neq x} s(x, y), \quad k = 1, \ldots, N, \tag{2}$$

where $|S_k|$ is the number of nodes in community S_k. This is an intuitive idea to choose the node who reaches others in the same community with the minimum average shortest path as the center.

2.2. The modularity function

In recent years, a concept of modularity proposed by Newman has been widely used as a measure of goodness for community structure [7–10]. A good division of a network into communities is not merely one in which the number of edges running between groups is small. Rather, it is one in which the number of edges between groups is smaller than expected. These considerations lead to the modularity Q defined by

$$Q = (\text{number of edges within communities}) - (\text{expected number of such edges}).$$

It is a function of the particular partition of the network into groups, with larger values indicating stronger community structure [7, 8]. Some existing methods are presented to find good partitions of a network into communities by optimizing the modularity over possible divisions, which has proven highly effective in practice [12].

The definition of the modularity can involve a comparison of the number of within-group edges in a real network and the number in some equivalent randomized model network in which edges are placed without regard to community structure [9]. The null model also has n nodes as the original network. The probability $e^0(x, y)$ for an edge to fall between every pair of nodes x and y is specified. More precisely, $e^0(x, y)$ is the expected number of edges between x and y. A definition allows for the possibility that there may be more than one edge between a pair of vertices, which happens in certain types of networks. For a given partition $\{S_k\}_{k=1}^{N}$, the

modularity can be written

$$Q = \frac{1}{2m} \sum_{k=1}^{N} \sum_{x,y \in S_k} \big(e(x,y) - e^0(x,y)\big), \tag{3}$$

where

$$e^0(x,y) = \frac{\mathrm{d}(x)\mathrm{d}(y)}{2m} \tag{4}$$

and m is again the number of edges in the network and $d(x)$ is defined in (1) (see [9]). This model has been studied in the past in its own right as a model of a network, and also closely related to the configuration model, which has been widely studied in physics literature [20].

3. The algorithm

The first simulated annealing algorithm was motivated by simulating the physical process of annealing solids [14]. The process can be described as follows. Firstly, a solid is heated from a high temperature and then cooled slowly so that the system at any time is approximately in thermodynamic equilibrium. At equilibrium, there may be many configurations with each one corresponding to a specific energy level. The chance of accepting a change from the current configuration to a new configuration is related to the difference in energy between the two states. The simulated annealing strategy is widely applied to optimization problems [15].

Let $E = -Q$. $E^{(n)}$ and $E^{(n+1)}$ represent the current energy, and new energy, respectively. $E^{(n+1)}$ is always accepted if it satisfies $E^{(n+1)} < E^{(n)}$, but if $E^{(n+1)} > E^{(n)}$ the new energy level is only accepted with a probability as specified by $\exp(-\frac{1}{T}\triangle E^{(n)})$, where $\triangle E^{(n)} = E^{(n+1)} - E^{(n)}$ is the difference of energy and T is the current temperature. Worse solutions are accepted based on the change in solution quality which allows the search to avoid becoming trapped at local minima. The temperature is then decreased gradually and the annealing process is repeated until no more improvement is reached or any termination criteria have been met. The initial state is generated at random by N communities, where N is an integer within the range $[N_{\min}, N_{\max}]$. The initial temperature T is set to a high temperature $T_{\max}$. A neighbor of the current state is produced by randomly choosing the strategies of our proposal, then the energy of the new state is calculated. The new state is kept if the acceptance requirement is satisfied. This process will be repeated for R times at the given temperature. A cooling rate $0 < \alpha < 1$ decreased the current temperature until the bound $T_{\min}$ is reached. The whole procedure of the Simulated Annealing with a Shortest-Path-based k-means algorithm (SASP) is summarized below

(i) Set parameters $T_{\max}$, $T_{\min}$, $N_{\min}$, $N_{\max}$, α and R. Choose N randomly within range $[N_{\min}, N_{\max}]$ and initialize the partition $\{S_k^{(0)}\}_{k=1}^{N}$ randomly; Set the current temperature $T = T_{\max}$.

(ii) Compute the centers $\{m(S_k^{(0)})\}_{k=1}^{N}$ according to (2), then calculate the initial energy $E^{(0)}$ using (3); Set $n^* = 0$.

(iii) For $n = 0, 1, \ldots, R$, do the following

(iiia) Generate a set of centers $\{m(S_k^{(n)})\}_{k=1}^{N'}$ according to our proposal below and set $N = N'$;

(iiib) Update the partition $\{S_k^{(n+1)}\}_{k=1}^N$ and the center set $\{m(S_k^{(n+1)})\}_{k=1}^N$ according to

$$S_k^{(n+1)} = \big\{x : k = \arg\min_l s(x, m(S_l^{(n)})\big\}, \quad k = 1, \ldots, N \tag{5}$$

and (2), respectively, then calculate the new energy $E^{(n+1)}$ using (iii);

(iiic) Accept or reject the new state. If $E^{(n+1)} < E^{(n)}$ or $E^{(n+1)} > E^{(n)}$ with $u \sim \mathcal{U}[0,1]$, $u < \exp\{-\frac{1}{T}\triangle E^{(n)}\}$, then accept the new solution by setting $n = n + 1$; Else, reject it;

(iiid) Update the optimal state, i.e., if $E^{(n)} < E^{(n^*)}$, set $n^* = n$.

(iv) Cooling temperature $T = \alpha \cdot T$. If $T < T_{\min}$, go to Step (v); Else, set $n = n^*$, repeat Step (iii).

(v) Output the optimal solution $\{S_k^{(n^*)}\}_{k=1}^N$ and the minimum energy $E^{(n^*)}$ of the whole procedure.

Our proposal to the process of generating a set of new centers in Step (iiia) comprises three functions, including deleting a current center, splitting a current center and perturb a current center. At each iteration, one of the three functions can be randomly chosen and the community centrality

$$c(S_k) = \frac{1}{|S_k|} \sum_{x \in S_k} s(m(S_k), x), \quad k = 1, \ldots, N \tag{6}$$

is used to select a community. Obviously, that the centrality is smaller indicates that the corresponding community is stronger. The three functions are described below

• Delete Center. The community with the maximal centrality S_d is identified using (6) and its center $m(S_d)$ should be deleted from the current center set.

• Split Center. The community with the minimal centrality S_s is chosen since $|S_s|$ may be large. The new center is obtained by

$$m(S_{N+1}) = \arg \min_{x \in S_s, x \neq m(S_s)} s(x, m(S_s)). \tag{7}$$

• Perturb Center. The community with the maximal centrality S_p is identified and we change $m(S_p)$ to another node in S_p with the second minimal average shortest path links with others in S_p.

The number of the iteration steps depends on the initial and terminal temperature, the cooling rate and the repeating times at the given temperature. The global maximum of modularity can also be obtained by searching over the all possible N using the two k-means algorithms. This will cost extremely much since for each fixed N, the k-means procedure should be operated 1000 to 5000 trials due to its local minima. Such a procedure either cannot produce a reasonable partitioning result according to our numerical results below. However, the simulated annealing strategy can avoid repeating ineffectively and lead to a high degree of efficiency and accuracy.

4. Experimental results

4.1. Ad-hoc networks with 128 nodes

The first example is the ad hoc network with 128 nodes. The ad hoc network is a benchmark problem used in many papers [6, 7, 11, 12]. It has a known partition and is constructed as follows.

Suppose we choose $n = 128$ nodes, split them into 4 communities with 32 nodes each. Assume that pairs of nodes belonging to the same communities are linked with probability p_{in}, and pairs belonging to different communities with probability p_{out}. These values are chosen so that the average node degree d is fixed at $d = 16$. In other words, p_{in} and p_{out} are related as

$$31p_{\text{in}} + 96p_{\text{out}} = 16. \tag{8}$$

We denote $S_1 = \{1 : 32\}, S_2 = \{33 : 64\}, S_3 = \{65 : 96\}, S_4 = \{97 : 128\}$. We change z_{out} from 0.5 to 8 and look into the fraction of nodes which are correctly classified. The fraction of correctly identified nodes is shown in Figure 1, compared with the two methods described in [7]. It seems that our method performs better than the two previous methods, especially for the more diffusive cases when z_{out} is large.

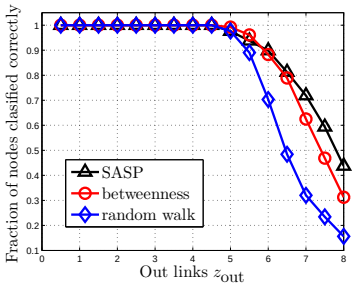


Figure 1 The fraction of correctly classified nodes

4.2. Sample network generated from gaussian mixture model

To further test the validity of the algorithms, we apply them to a sample network generated from a Gaussian mixture model [21]. We generate n sample points $\{\boldsymbol{x}_i\}$ in two dimensional Euclidean space subject to a K-Gaussian mixture distribution $\sum_{k=1}^{K} q_k G(\boldsymbol{\mu}_k, \Sigma_k)$, where $\{q_k\}$ are mixture proportions satisfying $0 < q_k < 1$, $\sum_{k=1}^{K} q_k = 1$, $\boldsymbol{\mu}_k$ and Σ_k are the mean positions and covariance matrices for each component, respectively. Then we generate the network with a thresholding strategy. That is, if $|\boldsymbol{x}_i - \boldsymbol{x}_j| \leq \text{dist}$, we set an edge between the i-th and j-th node; otherwise they are not connected. With this strategy, the topology of the network is induced by the metric. We take $n = 300$ and $K = 3$, then generate the sample points with the means and the covariance matrices

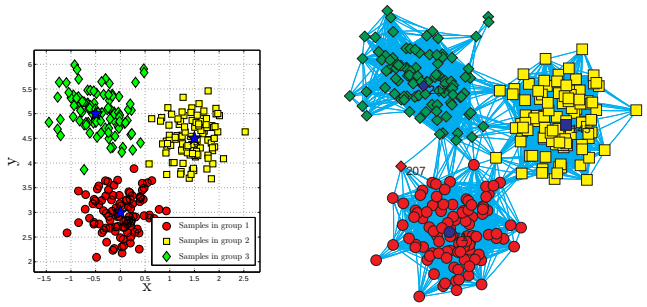
$$\boldsymbol{\mu}_1 = (0.0, 3.0)^T, \boldsymbol{\mu}_2 = (1.5, 4.5)^T, \boldsymbol{\mu}_3 = (-0.5, 5.0)^T, \tag{9a}$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.15 \end{pmatrix}. \tag{9b}$$

Here we pick nodes 1:100 in group 1, nodes 101:200 in group 2 and nodes 201:300 in group 3 for simplicity. With this choice, approximately $q_1 = q_2 = q_3 = 100/300$. We take dist $= 0.7$ in this example. The sample points and the community structure are shown in Figure 2. The results are extremely reasonable which indicate the effectiveness of our algorithm.

Figure 2(a) shows 300 sample points generated from the given 3-Gaussian mixture distribution. The star symbols represent the centers of each Gaussian component. The circle, square and

diamond shaped symbols represent the position of sample points in each component respectively. Figure 2(b) shows the partition for the network generated from the sample points in Figure 2(a) with dist = 0.7. The greatest $Q = 0.6554$ corresponding to 3 communities represented by different colors and the centers $m = \{71,145,217\}$ are shown blue.

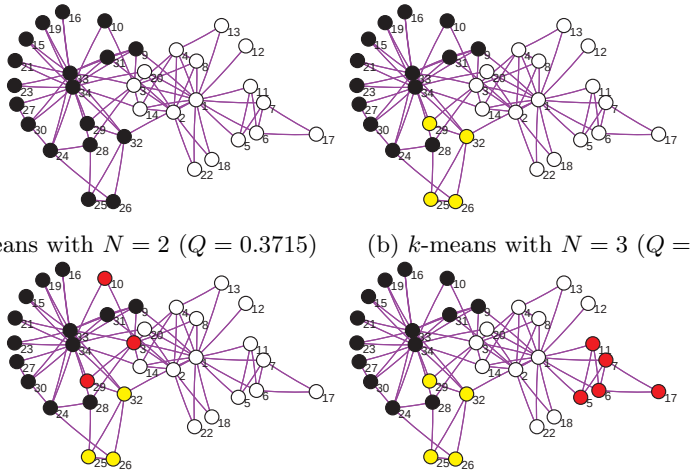


(a) (b)

Figure 2 Sample network generated from gaussian mixture model

4.3. The karate club network

This network was constructed by Wayne Zachary after he observed social interactions between members of a karate club at an American university [22]. Soon after, a dispute arose between the clubs administrator and main teacher and the club split into two smaller clubs. It has been used widely to test the algorithms for finding communities in networks [6–11]. The results obtained by our methods are shown in Figure 3. Our method reaches a higher value of Q than searching over all the possible N using the shortest-path-based k-means algorithm.



(a) k-means with $N = 2$ ($Q = 0.3715$) (b) k-means with $N = 3$ ($Q = 0.3875$)

(c) k-means with $N = 4$ ($Q = 0.3863$) (d) SASP ($N = 4, Q = 0.4151$)

Figure 3 The community structure of the karate club network detected by k-means and SASP methods

4.4. The dolphins network

The dolphins network is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand [23]. The network was compiled

from the studies of the dolphins, with ties between dolphin pairs being established by observation of statistically significant frequent association [7]. The results obtained by our method is shown in Figure 4. The network seems splitting into two large communities by the yellow part and the larger one and the larger one keeps splitting into 3 smaller communities, represented by different colors. The greatest modularity achieved is 0.5265 and corresponds to the 4 communities. The centers Jet, TR77, Grin and Topless are colored blue. The split into two groups appears to correspond to a known division of the dolphin community [24]. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals that the green part consists almost of entirely females and the others almost entirely males.

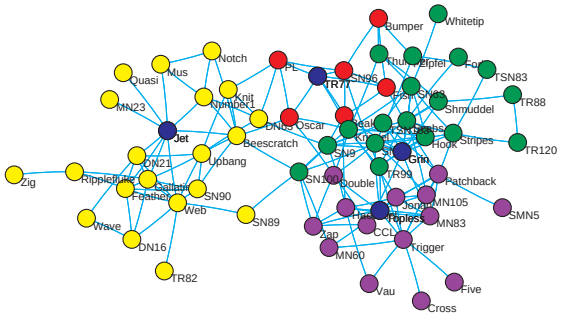


Figure 4 The community structure of the dolphins network detected by our method

5. Conclusions

In this paper, we present a new method to detect the community structure in complex networks. The proposed algorithm, simulated annealing associated with a shortest-path-based k-means algorithm(SASP), is constructed and efficiently tested on several representative examples. The results on artificial networks show that our method has better efficiency and accuracy in most cases. We again point out that our algorithm can not only find the community structure, but also identify the central node of each community during the cooling process. The optimal number of communities can be efficiently determined without any prior knowledge about the community structure. Moreover, the algorithm is successfully applied to two real-world networks, including the karate club network and the dolphins network.

References

[1] R. ALBERT, A. L. BARABÁSI. *Statistical mechanics of complerx networks*. Rev. Mod. Phys., 2002, **74**(1): 47–97.

[2] M. NEWMAN. *The structure and function of net works*. Comput. Phys. Commun., 2002, **147**(1-2): 40–45.

[3] M. NEWMAN, A. L. BARABÁSI, D. J. WATTS. *The Structure and Dynamics of Networks*. Princenton University Press, Princeton, 2005.

[4] A. BARABÁSI, H. JEONG, Z. NEDA, et al. *Evolution of the social network of scientific collaborations*. Physica A, 2002, **311**: 590–614.

[5] E. RAVASZ, A. SOMERA, D. MONGRU, et al. *Hierarchical organization of modularity in metabolci networks*. Science, 2002, **297**(5586): 1551–1555.

[6] M. GIRVAN, M. NEWMAN. *Community structure in social and biological networks*. Proc. Natl. Acad. Sci. USA, 2002, **99**(12): 7821–7826.

[7] M. NEWMAN, M. GIRVAN. *Finding and evaluating community structure in networks*. Phys. Rev. E, 2004, **69**(2): 026113.
[8] M. NEWMAN, *Detecting community structure in networks*. Eur. Phys. J. B, 2004, **38**(2): 321–330.
[9] M. NEWMAN. *Fing community structure in networks using the eigenvectors of matrices*. Phys. Rev. E, 2006 **74**: 036104.
[10] M. NEWMAN. *Modularity and community structure in networks*. Proc. Natl. Acad. Sci. USA, 2006, **103**(23): 8577–8582.
[11] J. DUCH, A. ARENAS. *Community detection in complex networks using extremal optimization*. Phys. Rev. E, 2005, **9** 027104.
[12] L. DANNON, A. DIAZ-GUILERA, J. DUCH, et al. *Comparing community structure identification*. J. Stat. Mech., 2005, **9** P09008.
[13] T. HASTIE, R. TIBSHIRANI, J. FRIEDMAN. *The Element of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Veral, New York, 2001.
[14] N. METROPOLIS, A. W. ROSENBLUTH, M. N. ROSENBLUTH, et al. *Equation of state calculations by fast computing machines*. J. Chem. Phys., 1953, **21**(6): 1087.
[15] S. KIRKPATRICK, C. D. GELATT, M. P. VECCHI. *Optimization by simulated annealing*. 1983, **220**(4598): 671–680.
[16] S. WASSERMAN, K. FAUST. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
[17] V. LATORA, M. MARCHIORI. *A measure of centrality based on the network efficiency*. New J. Phys., 2007, **9**: 188.
[18] M. NEWMAN. *Scientific collaboration networks. II. Shortest patys, weighted networks and centrality*. Phys. Rev. E, 2001, **64**(1): 16132.
[19] L. LOVASZ. *Random walks on graphs: A survey, Combinatiorics*. Paul Erdos is Eighty, 1993, **2**: 1–46.
[20] F. CHUNG. *Spectral Graph Theory, American Mathematical Society*. Rhode Island, 1997.
[21] M. PENROSE. *Random Geometric Graphs*. Oxford University Press, Oxford, 2003.
[22] W. ZACHARY. *An information flow model for conflict and fission in small groups*. J. Anthrop. Res., 1977, **33**(4): 452-473.
[23] D. LUSSEAU. *The emergent properties of a dolphin social netword*. Proceeding of the Royal Society B: Biological Sciences, 2003, **207**: 186–188.
[24] D. LUSSEAU, K. SCHNEIDER, O. BOISSEAU, et al. *The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations*. Behavioral Ecology and Sociobiology, 2003, **54**(4): 396–405.