The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Optimization analysis of modularity measures for network community detection

Xiang-Sun Zhang*[1] Rui-Sheng Wang[2]

[1]Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190
[2]School of Information, Renmin University of China, Beijing 100872, China

Abstract Complex networks have always been a hot research topic in recent years. In addition to some statistic properties such as scale-free nature, small-world property, modularity is another characteristic common to many types of complex networks. Detecting modular (or community) structure of complex networks is an important but challenging task, where quantitative measures for evaluating the modularity of networks play critical roles. In this paper, we make an analytic comparison of two existing modularity measures for network community detection based on optimization techniques, which can provide insights on their applications in real networks.

Keywords Complex network; Community detection; Modularity measure; Optimization

1 Introduction

Many systems in real world can be represented as a network, in which a set of nodes denote the objects of interest and links (edges) that connect nodes describe the relations between them. Examples range from social networks (scientific collaboration networks, food networks, etc.), technological networks (telecommunication systems, power grid networks, etc.), to biological networks such as protein interaction networks, gene regulatory networks, metabolic networks [3]. These different types of complex networks have been revealed to have common topological features such as scale-free nature, small-word property [1]. In addition to various statistic properties, many complex networks have community or modular structure, i.e. networks consist of groups of nodes, within which nodes are densely connected and meanwhile between which there are only sparse connections. Uncovering such community structure can not only help us to understand the topological structure of large-scale networks, but also reveal the functionality of each component. A close related subject to community structure in complex network is the modular organization of biological systems [9], which means that biological systems are composed of interacting, separable, functional modules. Identifying these modules is essential to understand the organization of biological systems and cellular processes.

So far, there are a number of algorithms have been proposed to detect communities in complex networks. networks, such as betweenness-based methods [3], spectral methods [6], information theoretical methods [10], machine learning methods [11], etc. Another

*Correspondence: zxs@amt.ac.cn

class of methods come forth after a modularity function Q was developed by Newman [7]. Q is a modularity measure to evaluate whether the community structure in a network is distinct or not. For a partition of a network, Q can also measure if this partition is good enough to capture the modularity structure underlying the network. Now the modularity function Q has been widely used, and a large class of methods directly based on maximizing modularity have been proposed [6]. However, Q has been exposed to resolution limits. Fortunato and Barthélemy recently claimed that modularity Q contains an intrinsic scale that depends on the total size of links in the network [2]. Modules smaller than this scale may not be resolved even in the extreme case that they are complete graphs connected by single bridges. In a recent study [5], Li et al. proposed a novel quantitative measure D for evaluating the community structure of networks. Based on the concept of graph density, this measure can overcome the resolution limits in Q and improve the quality of module detection.

Although the basic qualitative definition of community is that a group of nodes with dense connection inside and sparse connection to the outside, there are some quantitative definitions proposed to describe it. One of them is the weak definition of community in [8]. In this paper, based on this weak community definition, we make an analysis comparison of the modularity measures Q and D, which will provide insights on their applications in real networks. For this, although there are some research work on revealing fuzzy or overlapping community structure [12], the quantitative measures Q and D actually state the problem of network community detection as follows:

Partition a network into non-overlapping individual modules such as the connection within modules is as dense as possible, and the connection between modules is as sparse as possible.

2 Formulation of community detection

Given a network as $N = (V,E)$, where $V = \{v_1, v_2, \cdots, v_n\}$ is the set of nodes and E is the set of edges, $[e_{ij}]$ is its adjacency matrix with $e_{ij} = 1$ if $(v_i, v_j) \in E$ and otherwise $e_{ij} = 0$. Suppose the network is partitioned into K communities $N_1, N_2, \cdots, N_K, K \in \{1, 2, \cdots, n\}$, we use binary integer variables x_{ij} to represent if the node v_i is in community N_j. $x_{ij} = 1$ indicates that node v_i is in community N_j, otherwise $x_{ij} = 0$. The weak definition of community given in [9] is that the following inequality

$$\sum_{s,t \in V} e_{st} x_{sj} x_{tj} \geq \sum_{s,t \in V} e_{st} x_{sj} (1 - x_{tj}) \tag{1}$$

holds for each N_j, $N_j \neq \emptyset$, $j = 1, \cdots, K$, where $K \leq n$.

For this definition, the problem of network community detection is then to find a solution $\{x_{ij}\}$ which leads to a set of nonempty communities that satisfy the condition (1). We call it as the *basic* problem of community detection. Though not completely consistent with the weak definition, two quantitative measures *modularity function Q* [6] and *modularity density D* [5] also give quantitative descriptions of network community detection when adopted as objective functions:

$$Q(N_1, \cdots, N_K) = \sum_{i=1}^{K} \left[\frac{|E_i|}{|E|} - \left(\frac{d_i}{2|E|} \right)^2 \right] \tag{2}$$

and

$$D(N_1,\cdots,N_K)=\sum_{i=1}^{K}\left(\frac{2|E_i|}{|V_i|}-\frac{|\bar{E}_i|}{|V_i|}\right), \tag{3}$$

where d_i represents total degrees of all nodes in N_i and $\bar{E}_i$ is all edges linking V_i and $V\setminus V_i$. With these two measures, the basic problem can be transformed (approximately, since Q and D are not completely consistent with the weak definition) into mathematical programs.

For the modularity function Q, we want find a partition of the network to solve:

$$\max_{K}\max_{N_1\cup\cdots\cup N_K=N}\sum_{i=1}^{K}\left[\frac{|E_i|}{|E|}-\left(\frac{d_i}{2|E|}\right)^2\right]. \tag{4}$$

With the defined binary integer variable x_{ij}, it corresponds to the following mathematical programming:

$$\begin{aligned}\max\ &\sum_{j=1}^{K}\left[\frac{\sum_{s,t\in V}e_{st}x_{sj}x_{tj}}{\sum_{(s,t)\in E}e_{st}}-\left(\frac{\sum_{s,t\in V}e_{st}x_{sj}}{\sum_{(s,t)\in E}e_{st}}\right)^2\right]\\ s.t.\ &\sum_{j=1}^{K}x_{ij}=1\\ &x_{ij}\in\{0,\,1\},\ i=1,\cdots,n,\ j=1,\cdots,K\end{aligned} \tag{5}$$

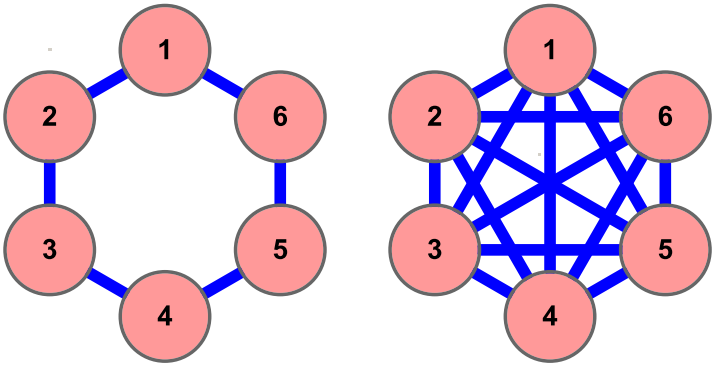
For the modularity density D, the optimization problem is:

$$\max_{K}\max_{N_1\cup\cdots\cup N_K=N}\sum_{i=1}^{K}\left(\frac{2|E_i|}{|V_i|}-\frac{|\bar{E}_i|}{|V_i|}\right) \tag{6}$$

It can be formulated as the following mathematical programming:

$$\begin{aligned}\max\ &\sum_{j=1}^{K}\left[\frac{\sum_{s,t\in V}e_{st}x_{sj}x_{tj}}{\sum_{t\in V}x_{tj}}-\frac{\sum_{s,t\in V}e_{st}x_{sj}(1-x_{tj})}{\sum_{t\in V}x_{tj}}\right]\\ s.t.\ &\sum_{j=1}^{K}x_{ij}=1\\ &x_{ij}\in\{0,\,1\},\ i=1,\cdots,n,\ j=1,\cdots,K\end{aligned} \tag{7}$$

It is obvious that modularity measures Q and D help to formulate the community detection problem into closed optimization models. But note that both the problem (5) and (7) are nonlinear integer programming without knowledge about the convexity or concavity of the objective functions, so they are hard to be analyzed theoretically or solved numerically. That is why most of the papers discussing the Q and D properties use some special networks, such as the *ring of cliques* [2] and the *ad hoc network* [7, 4] which in fact borrow convexity or concavity properties.



(a) The ring of lumps (b) The ad hoc network

Figure 1: Diagrams of two exemplary networks.

Two exemplary networks with known community structure are widely used in network research. One is a ring of dense lumps (Figure 1(a)), whose adjacency matrix is defined by

$$A^L = \begin{pmatrix} A & M & 0 & . & . & 0 & 0 & M \\ M & A & M & . & . & 0 & 0 & 0 \\ 0 & M & A & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & A & M & 0 \\ 0 & 0 & 0 & . & . & M & A & M \\ M & 0 & 0 & . & . & 0 & M & A \end{pmatrix} \tag{8}$$

where $L \geq 4$, A is an $m \times m$ adjacency matrix to represent a connected subnetwork called as lump, then A^L is an $Lm \times Lm$ matrix. M stands for a random matrix with l non-zero elements. Note that these random matrices don't have to be identical, provided that they have the same number of non-zero elements.

The second exemplary network is a special version of the ad hoc network (a computer-generated network, see Figure1(b)). Its adjacency matrix takes the form:

$$A^L = \begin{pmatrix} A & M & M & . & . & M & M & M \\ M & A & M & . & . & M & M & M \\ M & M & A & . & . & M & M & M \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ M & M & M & . & . & A & M & M \\ M & M & M & . & . & M & A & M \\ M & M & M & . & . & M & M & A \end{pmatrix} \tag{9}$$

Again, M stands for a random matrix with l non-zero elements. These random matrices don't have to be identical, but they should have the same number of non-zero elements.

3 Analysis for special examples

For an arbitrary partition of a network $P = \{V_1, V_2, \cdots, V_k\}$, rewrite the problems (4) and (6) as:

$$Q_p : \max_k \bar{Q}(k) = \max_k \max_{\sum_{i=1}^k |V_i| = n} Q(V_1, V_2, \cdots, V_k); \tag{10}$$

and

$$D_p : \max_k \bar{D}(k) = \max_k \max_{\sum_{i=1}^k |V_i| = n} D(V_1, V_2, \cdots, V_k); \tag{11}$$

These are two-step optimization problems. We denote $\bar{Q}(k)$ and $\bar{D}(k)$ as the solutions from the first-step optimization problems: with a fixed k, partition the whole network into k subnetworks $N_1 = (V_1, E_1), \cdots, N_k = (V_k, E_k)$ to maximize the quantitative functions Q and D. And $\max_k \bar{Q}(k)$ and $\max_k \bar{D}(k)$ are the second-step optimization problems.

3.1 The ring network of lumps

(1) Modularity function Q

Suppose that we partition the whole network into k communities with each community containing L_i lumps, $L_1 + \cdots + L_k = L$. Then

$$\begin{aligned} Q_p &= \max_k \max_{\sum_{i=1}^k L_i = L} \sum_{i=1}^k \left[\frac{L_i|A| + 2(L_i - 1)l}{L|A| + 2Ll} - \left(\frac{L_i + 2(L_i - 1)l + 2l}{L|A| + 2Ll} \right)^2 \right] \\ &= \max_k \max_{\sum_{i=1}^k L_i = L} \sum_{i=1}^k \frac{-1}{(L|A| + 2Ll)^2} [(|A| + 2l)^2 L_i^2 - (L|A|^2 + 2L|A|l + 2L|A|l \\ &\qquad + 4Ll^2)L_i + (2L|A|l + 4Ll^2)] \end{aligned}$$

Note that the first-step optimization problem is a discrete convex program. A function (or a programming) whose variables take discrete values (or, say, the sample values) is called as discrete convex (concave) function (or programming) if they can be embedded into a continuous convex (concave) function (or programming). To let the computation here make sense, the value of L is chosen as $L = 2^k$ and $k = 1, 2, 4, \cdots, L/2^{s+1}, L/2^s, L/2^{s-1}, \cdots, L$. Denote $F = \{1, 2, 4, \cdots, L/2^{s+1}, L/2^s, L/2^{s-1}, \cdots, L\}$. Solving the K-K-T equation of the above first-step optimization problem leads to $L_1 = \cdots = L_k = \frac{L}{k}$, then

$$\begin{aligned} Q_p &= \max_k \{-(|A| + 2l)^2 \frac{L^2}{k} + (L|A|^2 + 4L|A|l + 4Ll^2)L - (2L|A|l - 4Ll^2 k)\} \\ &\equiv \max_k \bar{Q}(k). \end{aligned}$$

It is easy to see that $\bar{Q}(k)$ is a discrete concave function, then the solution is given by the derivative of $\bar{Q}(k)$ at zero. That is, from

$$\bar{Q}'(k) = \frac{(|A|+2l)^2 L^2}{k^2} - 2Ll(|A|+2l) = 0 \tag{12}$$

we have solution

$$k^* = \langle \sqrt{\frac{|A|+2l}{2l}} \sqrt{L} \rangle_F \tag{13}$$

where $\langle \sqrt{\frac{|A|+2l}{2l}} \sqrt{L} \rangle_F$ means the integer in F nearest to $\sqrt{\frac{|A|+2l}{2l}} \sqrt{L}$.

(2) Modularity density D

$$\begin{aligned} D_p &= \max_k \max_{\sum_{i=1}^k L_i = L} \left\{ \sum_{i=1}^k \left(\frac{L_i|A| + 2(L_i - 1l)}{L_i m} - \frac{2l}{L_i m} \right) \right\} \\ &= \max_k \max_{\sum_{i=1}^k L_i = L} \sum_{i=1}^k \left(\frac{-4l}{L_i m} + \frac{|A|+2l}{m} \right) \end{aligned}$$

where m is the rank of A. The first-step optimization is a convex programming problem with solution $L_1 = \cdots = L_k = \frac{L}{k}$, then

$$D_p = \max_k \left\{ -\frac{4l}{m} \frac{k^2}{L} + k \frac{|A|+2l}{m} \right\} \tag{14}$$

and the solution is $k^* = \langle \frac{(|A|+2l)L}{8l} \rangle_F$.

3.2 The ad hoc network

(1) Modularity function Q

$$\begin{aligned} Q_p &= \max_k \max_{\sum_{i=1}^k L_i = L} \sum_{i=1}^k \left[\frac{L_i|A| + L_i(L_i-1)l}{L|A| + L(L-1)l} - \left(\frac{L_i|A| + L_i(L_i-1)l + L_i(L-L_i)l}{L|A| + L(L-1)l} \right)^2 \right] \\ &= \max_k \max_{\sum_{i=1}^k L_i = L} \sum_{i=1}^k \{ L_i^2 (l - |A|)(|A| + (L-1)l) - L_i(|A| - l)(L|A| + L(L-1)l) \} \end{aligned}$$

Note that the first-step optimization is a convex programming if $l < |A|$, then it has solution $L_1 = \cdots = L_k = \frac{L}{k}$. We further have

$$Q_p = \max_k \left\{ \frac{L^2}{k} (l - |A|)(|A| + (L-1)l) - L(|A| - l)(L|A| + L(L-1)l) \right\} \tag{15}$$

which again is a convex problem and the solution is $k^* = L$.

When $|A| < l$, $\bar{Q}_A$ is a concave programming, the solution is reached at the boundary. Note that $\bar{Q}(k)$ is a monotonously decreasing function, then $k^* = 1$.

(2) Modularity density D

$$\begin{aligned} D_p &= \max_k \max_{\sum_{i=1}^k L_i = L} \sum_{i=1}^k \left\{ \frac{L_i|A| + L_i(L_i-1)l}{L_i m} - \frac{L_i(L-L_i)l}{L_i m} \right\} \\ &= \max_k \max_{\sum_{i=1}^k L_i = L} \frac{1}{m} \sum_{i=1}^k \{|A| + 2L_i l - (L+1)l\} \end{aligned}$$

Now the first-step optimization is a simple linear programming problem with any feasible solution as the optimal solution. Then

$$D_p = \max_k \{k(|A| - (L+1)l) + 2Ll\} \tag{16}$$

which again is a linear function, then

$$k^* = \begin{cases} L \ \ if \ \ l < |A|/(l+1), \\ 1 \ \ if \ \ l > |A|/(l+1). \end{cases} \tag{17}$$

When $l = \frac{|A|}{L+1}$, any k is a solution.

4 Conclusion and discussion

Partitioning a network into communities is a problem related to optimization modeling and but has never been carefully studied from the view of optimization theory. It is also a very difficult problem not only for the NP-completeness of its computational model but also for the lack of deep understanding of the problem definition and quantitative measure properties.

In this short paper, we clearly described the basic problem of community detection, and introduced two closed optimization models based on modularity measures Q and D to approximately solve the basic problem. Two special network structures, the ring of dense lumps and the ad hoc network, are discussed to make the optimization models to be convex or concave, then to solve the solution theoretically. These results are being used in the authors' research on comparison of different quantitative measures. It should be noted that, although Q and D can approximately solve the basic problem, they both are found to have some problems, especially the so called resolution limit ([2]). In the case of a solution affected by resolution limit, the communities found by the corresponding algorithm can be split into smaller sub-communities which are still satisfying the weak definition. To overcome the limits and improve community detection, we give the following definition of community detection problem

Partition a network into as many non-overlapping modules as possible such that each model satisfies the weak definition.

Based on this definition, a more reasonable quantitative measure for characterizing modularity is expected to be obtained.

Acknowledgment

This work is supported by the Ministry of Science and Technology of China, under Grant no.2006CB503905, National Natural Science Foundation of China under Grant no.10631070 and no.10701080, and the JSPS-NSFC collaboration project 10711140116.

References

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47, 2002.

[2] S. Fortunato and M. Barthelemy, Resolution limit in community detection, *Proc. Natl. Acad. Sci. USA*, 104(1), 36-41, 2007.

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

[4] W.E.T. Li, and E. Vanden-Eijnden, Optimal partition and effective dynamics of complex networks, *Proc. Natl. Acad. Sci. USA*, 105(23): 7907-7912, 2008.

[5] Z. Li, S. Zhang, R.S. Wang, X.S. Zhang and L. Chen, Quantitative function for community detection, *Phys. Rev. E.*, 77, 036109, 2008.

[6] M.E.J. Newman, Modularity and community structure in networks, *Proc. Natl. Acad. Sci. USA*, 103(23), 8577-8582, 2006.

[7] M.E.J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E.*, 69(2), 026113, 2004.

[8] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, Defining and identifying communities in networks, *Proc. Natl. Acad. Sci. USA*, 101(9): 2658-2663, 2004.

[9] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabasi, Hierarchical organization of modularity in metabolic Networks, *Science*, 297(5586): 1551-1555, 2002.

[10] M. Rosvall, C.T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. USA*, 104, 7327-7331, 2007.

[11] R.S. Wang, S. Zhang, Y. Wang, X.S. Zhang, L. Chen. Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures. *Neurocomputing*, doi:10.1016/j.neucom.2007.12.043.

[12] S. Zhang, R.S. Wang, X.S. Zhang. Uncovering fuzzy community structure in complex networks. *Phys. Rev. E.*, 76, 046103, 2007.