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

A Constrained Optimization Method for Community Detection

Ji-Guang Wang[1] Lin Wang[1,2] Yu-Qing Qiu[1]
Yong Wang[1] Xiang-Sun Zhang[1,*]

[1]Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
[2]Computer Science and Information Engineering College,
Tianjin University of Science and Technology, Tianjin 300222, China

Abstract

Community detection is one of the most important problems in complex network research. In recent years, great efforts have been devoted to this problem in term of evaluating the resulting community structure. Our previous work has shown that in addition to the resolution limit of Q, both Q and D suffer from a more serious limitation, termed as extra weak community phenomenon, i.e. some derived communities do not satisfy even the weak community definition. In this paper, we provide a constrained optimization model to overcome extra weak community phenomenon. With an improved simulated annealing algorithm, we solve the constrained optimization model for both Q and D, and then use our new method in several practical community detection problems. The experimental results show that the new method can not only partition large networks into communities properly but also ensure that all resulting communities at least satisfy the weak community definition. In addition, we find that constrained optimization of Q finds fewer but large communities, while constrained optimization of D takes the network apart more detailed.

Keywords Extra Weak Community; Constrained Optimization Model; Community Detection

1 Introduction

Community detection is an important problem in complex network research. As being widely assumed that most networks, such as Internet, social networks, biological networks and so on, show "community structure" i.e., groups of vertices that have a high density of edges within them, and a low density of edges between them [14]. Uncovering the underlying community structure helps to cartographically represent and mine important knowledge from the complex graphical frameworks. For example, communities of world wide web represent the groups of websites with similar topics. They are used to improve search engines, filter contents, and analyze relationships within and among different topics [4]. Communities in biological networks usually represent functional modules. They can be used to predict protein function, explain diseases mechanism, and obtain valuable biological insights [9].

*corresponding to zxs@amt.ac.cn

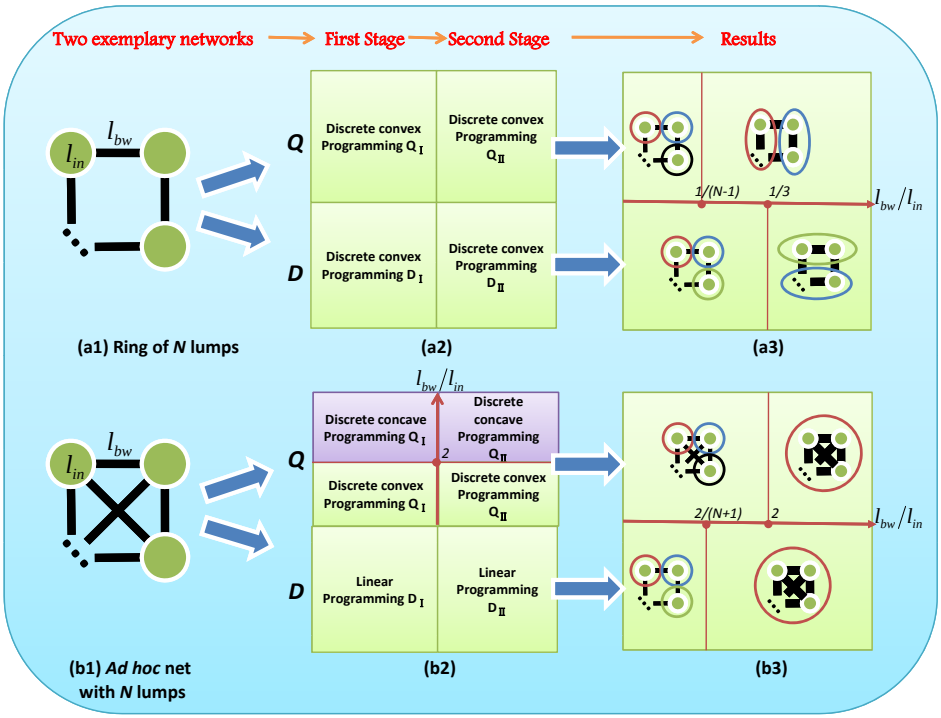

Figure 1: Illustration of the theoretical analysis for modularity measures Q and D by discrete convex programming on two exemplary networks. (a1)-(a3): the diagram for the ring of N lumps, the properties of the two stage optimization problem (Q_I and D_I, Q_{II} and D_{II}), and the analytical solutions. The horizontal axis (in red) in (a3) denotes the value of parameter l_{bw}/l_{in}. When $l_{bw}/l_{in} < 1/(N-1)$, both Q and D can identify the N lumps as communities. When $l_{bw}/l_{in} > 1/3$, both Q and D fail to identify the N communities. When $1/(N-1) < l_{bw}/l_{in} < 1/3$, D can still identify N communities while Q cannot. (b1)-(b3): the diagram for *Ad hoc* net with N lumps, the properties of the two stage optimization problem, and the analytical solutions. The vertical axis (in red) in (b2) denotes the parameter l_{bw}/l_{in}. The property of Q_I and Q_{II} differs in the critical point $l_{bw}/l_{in} = 2$. The horizontal axis (in red) in (b3) also denotes the value of parameter l_{bw}/l_{in}. When $l_{bw}/l_{in} < 2/(N-1)$, both Q and D can identify the N lumps as communities. When $l_{bw}/l_{in} > 2$, both Q and D identify the whole network as a single communities. When $2/(N-1) < l_{bw}/l_{in} < 2$, D identify a single community while Q identify more than one communities. The properties of optimization models (Q_I, Q_{II},D_I, and D_{II}) to maximize the modularity measures Q and D on two exemplary networks can be found in Table S1 of Supplementary Materials 2. The detailed analytical solutions are presented in Table S2 of Supplementary Materials 2 (http://www.aporc.org/doc/wiki/ModularityOptimization).

In recent years, great efforts have been devoted to this problem in term of evaluating the resulting community structure. Newman and Girvan defined a modularity function Q to evaluate the quality of a particular division of a network [14]. Given a network $G = \{V, E\}$, where V and E represent the set of vertexes and the set of links respectively, and its division $P = (\{V_1, E_1\}, \{V_2, E_2\}, \cdots, \{V_n, E_n\})$, where $\cup_{i=1}^{n} V_i = V$, the modularity measure Q is defined by

$$Q = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} \left[\frac{|E_i|}{|E|} - \left(\frac{d(V_i)}{d(V)} \right)^2 \right] \tag{1}$$

where $|E|$ represents the total number of edges in V, $|Ei|$ represents the number of edges in V_i, $d(V)$ is the total degree of vertexes in V, and $d(Vi)$ is the total degree of vertexes in V_i. For each Q_i, the first term represents the observed percentage of edges inside the community, while the second one is its expected value. Generally, a larger Q value corresponds to a more reasonable division, so maximizing Q has been a widely accepted method for detecting community structure of complex networks [9].

However, Q was recently found to suffer from resolution limit, i.e. optimizing Q may fail to identify communities smaller than a scale depending on the total size of the network and on the degree of connections among the communities [5]. To overcome this problem, Li et al proposed another quantitative measure D. Based on the concept of graph density, it is defined by

$$D = \sum_{i=1}^{n} D_i = \sum_{i=1}^{n} \left(\frac{2E_i - \bar{E}_i}{|V_i|} \right) \tag{2}$$

where $\bar{E}_i$ means the number of edges from the ith community to others. Optimization of D does not show the resolution limit that Q suffers from on some examples and improves the quality of community detection [11].

Recently, Radicchi et al. proposed an explicit community definition [16]. In their work, the weak community is defined as that if a subgraph $G_i = \{E_i, V_i\}$ is a community it should satisfy

$$2E_i > \bar{E}_i.$$

This definition gave a basic rule to assess whether a group of nodes are community or not. Based on this definition, Zhang et al. found that in addition to the resolution limit of Q, both Q and D suffer from a more serious limitation, i.e. some derived communities do not satisfy even the weak community definition, which means that these communities, termed as extra weak communities, have sparser connection within them than between them [18]. This phenomena is also called "misidentification" according to the weak community definition. To illustrate this phenomena, a discrete convex optimization framework is used in two artificial networks. One is a ring of dense lumps which consists of N ($N \geq 4$) dense lumps each with m nodes. The other one is a kind of generation of the ad hoc network discussed in [15, 3], which also consist of N dense subgraphs. Under the assumption that all the links in G_s, $s = 1, 2, \cdots, N$, for these two exemplary networks are evenly distributed, the two stages of optimization process for both Q and D can be simplified as discrete convex optimization problems. With the analytical solutions of the discrete convex optimization problems, we compare these two modularity measures in terms of their ability to detect known communities and the extent of misidentification phenomenon in terms

of the network topology structure, i.e., a set of network parameters. The property of the two-stage optimization models with respect to network parameters is illustrated in Figure 1, and the detailed induction can be found in [17, 18].

In this short paper, we focus on the problem pointed out by [18], and then propose a constrained optimization model to improve the performance of both Q and D. Furthermore, we discuss some applications of our new improved methods in several community detection problems. This new methods can ensure all resulting communities at least satisfy the weak definition. In addition, as a byproduct of theoretical analysis and real networks experiments, we found that there is some difference between constrained optimization of Q and D. Constrained optimization of Q finds fewer but large communities, while constrained optimization of D takes the network apart more detailed.

2 A constrained optimization model to eliminate extra weak communities

Modularity function Q and modularity density D are widely used to evaluate the quality of a particular division of a network. However, both Q and D can not ensure all of the derived communities satisfy even the weak community definition. In order to overcome this problem and partition the network reasonable, i.e., each community in the partition satisfies the weak definition, it is natural to build the following constrained optimization problems based on the modularity measures.

For Q, we have

$$\begin{aligned} \max \quad & \textstyle\sum_{i=1}^{n} Q_i \\ \text{s.t.} \quad & 2|E_i| > \bar{E}_i, i = 1,2,\cdots,k \end{aligned} \tag{3}$$

For D, we have

$$\begin{aligned} \max \quad & \textstyle\sum_{i=1}^{n} D_i \\ \text{s.t.} \quad & 2|E_i| > \bar{E}_i, i = 1,2,\cdots,k \end{aligned} \tag{4}$$

Either problem is difficult to solve due to the fact that the space of possible partitions grows exponentially with respect to the size of network. In order to solve them we use the simulated annealing algorithm. Simulated annealing (SA) [10] is a generic probabilistically heuristic method for the global optimization problem, namely finding a good approximation to the global minimum of a given function in a large search space. It is to simulate the cooling process of the heated metal. From an arbitrary initial state the SA reaches the next state with possible minimal energy. At each step, the SA considers some neighbor s' of the current state s, and probabilistically decides either moving the system to state s' or staying in state s. The probability of making the transition from the current state s to a candidate new state s' is specified by an acceptance probability function $P(e,e',T)$, that depends on the energies $e = E(s)$ and $e' = E(s')$ of the two states, and also the temperature T.

We solve the constrained optimization model by improving the simulated annealing technique, that has been used to solve the Q optimization in [9]. Specifically, we always

set the whole network as the initial solution. At each temperature, we provide fn^2 node movements from one community to another community, where n is the number of nodes in the network and f is a coefficient and often taken as 1. It is noted that the node movement must enable the two newly created communities to satisfy the weak definition, otherwise the movement is not accepted. Meanwhile we also provide fn collective movements, which include merging two communities and splitting a community. It is noted that the split must enable the two split communities all satisfy the weak definition, otherwise the split is not accepted. After the movements are evaluated at each temperature, the temperature is decreased with a constant coefficient.

The improved algorithm and software can be found on the webpage: http://www.aporc.org/doc/wiki/ModularityOptimization. Using the improved algorithm, we correctly solved several examples both in artificial and in real biological and social networks in which extra weak communities are misidentified by using the original algorithm.

3 Experimental results on artificial and real networks

Then we used our new method for modularity measures Q and D on the exemplary networks. Although these networks have very special topology structures, the conclusion obtained on them can provide insights into general complex networks.

3.1 Artificial networks

The first numerical example is a set of computer-generated networks [6] which have been widely used to benchmark community detection algorithms. Each network has 128 nodes, which are divided into 4 communities each with 32 nodes. Edges are placed randomly with given probabilities so as to keep the average degree of a node to be 16. The average edge connection of each node to nodes of other communities is denoted by k_{out}. For each k_{out}, 10 random *ad hoc* networks are generated. Then, the partition of each network is obtained by optimizing modularity measures Q and D respectively by a simulated annealing procedure. Figure 2 compares the misidentification problem in both Q and D, where the community numbers (average value over 10 networks) given by the bar plot include both communities satisfying the weak definition and the communities failing to satisfy the weak definition. From this result we can see that for the *ad hoc* networks, modularity density D has no misidentification phenomenon, whereas when $k_{out} > 8$, some communities detected by Q do not satisfy the weak community definition. In other words, Q identifies some subgraphs with inner links even less than half of outward links as communities, which violates our basic community definition. When we apply our new constrained model to partition the networks, the misidentification problems are avoided.

3.2 Real networks

We further extend our theoretical analysis and show the performance of our new constrained model by several examples of real networks. These networks include some well studied complex networks such as metabolic network of *C. elegans* [2], dolphin network [12], email network [8], football network [6], jazz musician network [7], political book network [1], and scientific collaboration networks [13]. In addition, we constructed several bio-molecular networks such as transcriptional regulatory network and protein interaction network to study their modularity properties. The simulated annealing procedure

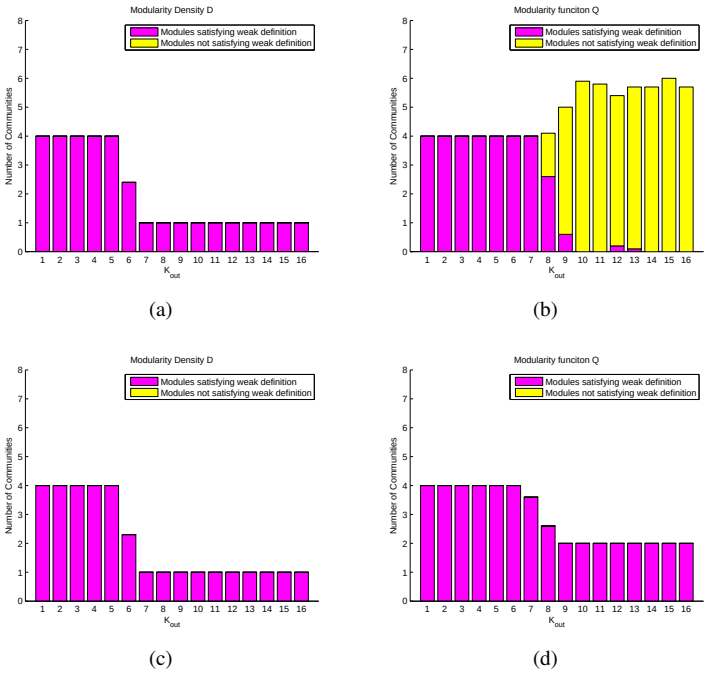

Figure 2: Comparison of Q and D in terms of misidentification on 4-community *ad hoc* networks. (a) The community numbers detected by optimizing modularity density D (average value over 10 networks). (b) The community numbers detected by optimizing modularity function Q (average value over 10 networks). (c) The community numbers detected by optimizing modularity density D with constrains. (d) The community numbers detected by optimizing modularity function Q with constraints.

is used here. The statistics of the network and the partition results are presented in Table 1 and Table 2. We find that for most of small and sparse networks, both modularity measures Q and D work well and all of identified communities satisfy weak definition. But when the studied networks get larger and denser, modularity measures Q and D obviously suffer from the misidentification problem. However, the misidentification phenomena are not appeared through the constrained optimization model which are shown in Table 1 and Table 2. Besides, comparing table 1 and table 2, we find that there is some difference between constrained optimization of Q and D. Constrained optimization of Q finds fewer but large communities, while constrained optimization of D takes the network apart more detailed. The main reason for the difference is that Q focuses on the global characters of networks, while D focuses on the local characters. In particular, for each community in formula (1), the value of Q_i is related to the total number of edges in the whole network, but every D_i in formula (2) is independent from the whole network.

A typical example is the jazz musician network [7] which is a social network to describe the collaboration among jazz bands. The data are from The Red Hot Jazz Archive database which stores 198 bands that performed from 1912 to 1940 with 1275 jazz mu-

Table 1: Experimental results on the real networks by optimizing Q. The misidentification phenomena are highlighted in bold.

	Direct optimization			Considering the weak definition constraints through constrained model		
Network name	Q Value	Number of communities	Satisfying weak definition	Q Value	Number of communities	Satisfying weak definition
C. celegans metabolic [2]	0.45	9	9	0.42	7	7
dolphins [12]	0.53	4	4	0.52	4	4
email [8]	0.57	10	10	0.57	9	9
football [6]	0.60	9	9	0.60	10	10
jazz [7]	0.44	**4**	**3**	0.44	3	3
karate [11]	0.42	4	4	0.40	4	4
politics books [1]	0.53	4	4	0.53	4	4
scienceA [13]	0.75	7	7	0.75	8	8
Yeast TRN	0.48	**14**	**12**	0.47	13	13
Yeast TFR	0.35	**6**	**3**	0.22	3	3

Table 2: Experimental results on the real networks by optimizing D. The misidentification phenomena are highlighted in bold.

	Direct optimization			Considering the weak definition constraints through constrained model		
Network name	D Value	Number of communities	Satisfying weak definition	D Value	Number of communities	Satisfying weak definition
C. celegans metabolic [2]	30.25	15	15	29.76	16	16
dolphins [12]	11.73	5	5	11.96	5	5
email [8]	63.16	**31**	**30**	60.03	28	28
football [6]	43.73	11	11	44.39	11	11
jazz [7]	52.84	**4**	**3**	52.03	4	4
karate [11]	7.85	3	3	7.85	3	3
politics books [1]	21.86	7	7	20.05	5	5
scienceA [13]	28.30	16	16	28.31	16	16
Yeast TRN	15.78	**15**	**14**	19.25	23	23
Yeast TFR	11.50	4	4	11.66	4	4

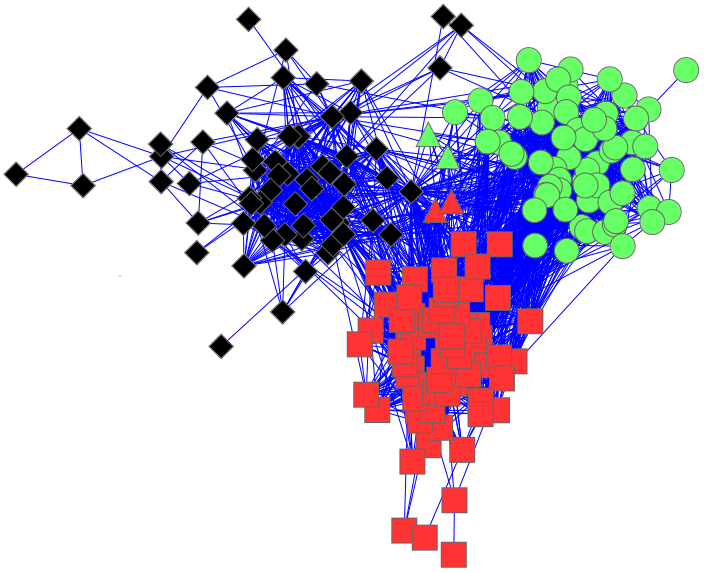

Figure 3: Misidentification in optimization of modularity measures Q in the jazz musician network. We use modularity measures Q to partition this network. The communities with different node shapes are identified by Q based on direct optimization, and communities with different colors are detected by Q based on constrained optimization model.

sicians [7]. In the jazz musician network, the bands are represented by nodes and two bands with at least one shared musician are linked by an edge. Due to the black/white racial segregation and the cities that bands recorded in, the network can be divided into three communities in reality. The community detection results by Q and D are shown in Table 1 and 2. We found that both Q and D partition this network into four communities with one misidentification. We draw the partition results of Q in Figure 3. The community misidentified by Q (**triangles** in Figure 3) has 4 nodes and has fewer inner links than outer links. However using the constrained model we can obtain the correct partition.

4 Conclusions

Community detection plays a fundamental role in complex network studies. We have given a way to overcome one serious problem in community detection, i.e. the occurrence of extra weak community or called misidentification. In this paper, we mainly show the advantage of constrained optimization model over un-constrained one by avoiding unreasonable extra weak community. In addition, as a byproduct, we found that there is some difference between constrained optimization of Q and D. Constrained optimization of Q finds fewer but large communities, while constrained optimization of D takes the network apart more detailed.

Acknowledges

The authors thank the anonymous reviewers for their valuable suggestions to improve the article. The authors are separately supported by the NSFC grants 10631070, 60873205, 10801131, 10701080, the grant kjcx-yw-s7 from CAS, and 2006CB503905 from MST of China.

References

[1] L. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. *ACM New York, NY, USA*, pages 36–43, 2005.

[2] J. Duch and A. Arenas. Community identification using extremal optimization. *Physical Review E*, 72:027104, 2005.

[3] Weinan E, T. Li, and E. Vanden-Eijnden. Optimal partition and effective dynamics of complex networks. *Proceedings of the National Academy of Sciences*, 105:7907–7912, 2008.

[4] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35(3):66–70, 2002.

[5] S. Fortunato and M. Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36, 2007.

[6] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:7821, 2002.

[7] P. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems*, 6:565–573, 2003.

[8] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68:65103, 2003.

[9] R. Guimera and L. A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[10] S. Kirkpatrick, C. Gelatt, and M. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[11] Z. Li, S. Zhang, R. S. Wang, X. S. Zhang, and L. Chen. Quantitative function for community detection. *Physical Review E*, 77(3):036109, 2008.

[12] D et al. Lusseau. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54:396–405, 2003.

[13] M. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98:404, 2001.

[14] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[15] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:26113, 2004.

[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658–2663, 2004.

[17] X. S. Zhang and Wang R. S. Optimization analysis of modularity measures for network community detection. *Proceeding of the second international symposium, OSB'08 Lijiang, China.*, 9:13–20, 2008.

[18] X. S. Zhang, Wang R. S., Wang J., Qiu Y., Wang L., and L. Chen. Modularity optimization in community detection of complex networks. *Europhysics Letters*, 87:38002, 2009.