Modularity optimization in community identification of complex networks

Xiang-Sun Zhang,[1] Rui-Sheng Wang,[2] Yong Wang,[1] Ji-Guang Wang,[1] Yu-Qing Qiu,[1] Lin Wang,[1] and Luonan Chen[3]

[1] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China*
[2] *Department of Physics, The Pennsylvania State University, University Park, PA 16801, USA*
[3] *Department of Electronics, Information, and Communication Engineering,*
Osaka Sangyo University, Osaka 574-8530, Japan

(Dated: February 12, 2009)

Detecting community structure of complex networks is a fundamental but challenging topic in Network Science. Modularity measures play critical roles as quality indices in partitioning a network into communities. This paper explored the modularity optimization by a quasi-analytic analysis and revealed its startling and complex behavior. It shows that except the resolution limit phenomenon that restricts the applicability of the widely used modularity function Q, there is a more serious limitation called misidentification for the Q and the modularity density D, a recent suggested modularity measure for alleviating the resolution limit effect. Misidentification means that the resulting communities do not satisfy the basic definition of a community: a community is a subgraph of a network whose nodes are connected tightly inside and sparsely to the outside. The cause leading misidentification is analyzed and a complete modularity optimization model and a realistic algorithm are given. Experiment results on both simulated data and real networks show the effectiveness of the new model and algorithm.

Many systems in real world can be represented as a network, in which nodes denote the objects of interest and edges that connect nodes describe the relationships between them. Examples range from social networks, technological networks to biological networks such as email network and protein interaction network. These different types of complex networks have been revealed to have common topological features such as scale-free and small-world [1]. Importantly, many complex networks have community or modular structure, i.e., networks consist of specific, relatively separate dense subgraphs [2]. In a widely used community definition [3–7], a community is a subgraph of a network whose nodes are connected tightly inside and sparsely to the outside. Uncovering such community structure not only helps us understand the topological structure of large-scale networks, but also reveals the functionality of each component. This is, for example, confirmed by the modular organization of biological networks [5], where the communities are sets of components with similar functions and the modular structure is the result of evolutionary constraints. With its critical role in the network science, there have been a batch of significant papers related to the community structure study in top journals among them we name a few that are related to this paper: [2–7, 10, 11, 15, 21, 23].

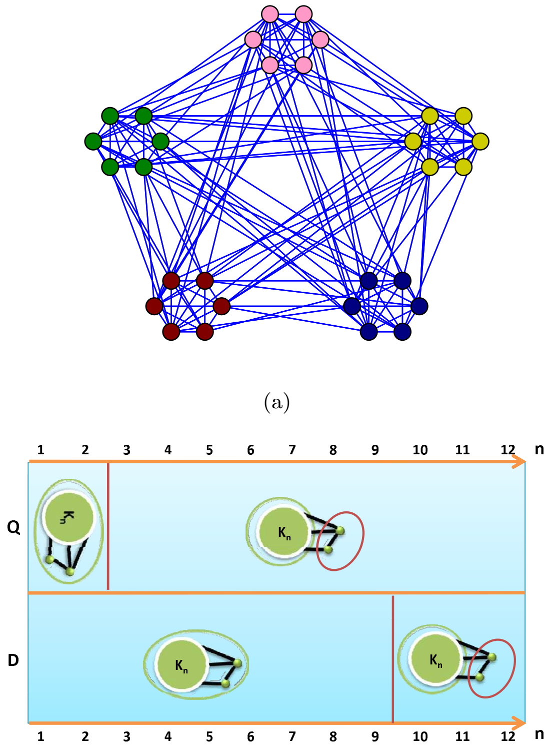
Radicchi et al. introduced two quantitative community definitions, in weak sense and in strong sense [3]. The weak definition is widely used: given a network $G = (V, E)$ with vertex set V and edge set E, we denote $A = [a_{ij}]$ as its adjacency matrix. Letting $V_s \subset V$ be a subgraph and $\overline{V}_s = V \backslash V_s$ be the set of nodes in the rest of the network, then V_s is a community in a weak sense if

$$L(V_s, V_s) > L(V_s, \overline{V}_s), \tag{1}$$

where $L(V_s, V_s) = \sum_{i\in V_s}\sum_{j\in V_s} a_{ij}$, $L(V_s, \overline{V}_s) = \sum_{i\in V_s}\sum_{j\notin V_s} a_{ij}$.

There are two streams of methods that have been proposed to detect communities in complex networks. One class includes betweenness-based methods [2], random walk methods [7], information theoretical methods [6], machine learning methods [8], etc. The concentration of this class of research is on the concrete algorithm design. The efficiency of the designed algorithms is evaluated by known test problems. The second class of methods is to build an optimization model that optimizes certain modularity measures which are related with the community definition. This class is more theoretically oreinted than the first class. One popular measure is the modularity function Q developed by Newman [9]. It is a quantitative measure for evaluating how good a community partition of a network is. A large number of methods have been devised for community detection based on optimizing Q [4, 10–20]. However, Q has been exposed to resolution limits, i.e., communities smaller than certain scale may not be resolved by optimization of Q even in the extreme case that they are complete graphs connected by single bridges [21]. In a recent study [22], Li et al. proposed another quantitative measure D called modularity density to evaluate the community structure of networks. This measure is based on the concept of graph density. Optimization of D does not show the resolution limit for the examples where Q fails. But there is no theoretical analysis to show the advantage of D over Q. Furthermore, we find that Q and D have much more complex behaviors than that the existing literature tells. For example, except the resolution limit there is a misidentification phenomena. To explain the new phenomenon, we first give the Q, D definition in detail.

Suppose that we have a partition $P_K = (G_1, G_2, \cdots, G_K) = ((V_1, E_1), \cdots, (V_K, E_K))$, where K is the number



(a)

(b)

FIG. 1: Illustration of the misidentification phenomena.

of candidate communities of the partition. The modularity function Q [9] is defined as

$$Q = \sum_{s=1}^{K} \left[\frac{L(V_s,V_s)}{2L} - \left(\frac{L(V_s,V_s)+L(V_s,\overline{V}_s)}{2L} \right)^2 \right] \equiv \sum_{s=1}^{K} Q_s, \qquad (2)$$

where $L = L(V,V)/2$ is the total number of links in the network, and $L(V_s,V_s) + L(V_s,\overline{V}_s)$ is the total degree of the nodes in community s. This measure compares the number of edges inside a given community with the expected value in a randomized graph of the same size and same degree sequence. It provides a way to determine if a partition is good enough to decipher the community structure of a network. Generally, a bigger Q corresponds to a better community structure.

Similarly, for a given partition P_K, the modularity density D [22] is defined as

$$D = \sum_{s=1}^{K} \frac{L(V_s,V_s) - L(V_s,\overline{V}_s)}{|V_s|} \equiv \sum_{s=1}^{K} D_s, \qquad (3)$$

where $|V_s|$ is the number of nodes in community s. It should be noted that node information in a community is incorporated into the denominator of the modularity density D. This is different from Q.

We use the term "misidentification" to denote a case when optimization of Q and D leads to the resulting communities failing to satisfy the weak community definition in (1). An illustrative example is given in FIG.1(a), where there are five 6-cliques, any two of which are connected by eight links. Let P_5 denote the partition of five communities where each clique is a community, and P_1 denote the partition that the whole network is a community. Simply computing the modularity function Q shows that $Q(P_5) = 44/155 > 0 = Q(P_1)$, which implies that optimizing Q partitions the network into five communities. However, such communities have 15 inner-links and 32 out-links and thereby do not satisfy the weak community definition. The consequence of such misidentification is that some detected communities may have sparser connection within them than between them. Another simple example in FIG.1(b) shows that both Q and D suffer from misidentification. In this example, there are 3 links between n-clique K_n and a 2-clique. Obviously, 2-clique does not satisfy the weak definition. Optimizing Q partitions the network into 2 communities when $n \geq 3$, while D partitions into 2 when $n \geq 10$. Experiments on real networks further indicate that the misidentification of Q and D is common in many types of artificial and real networks and sometimes leads to unreasonable results especially in biological networks (see the details in Supplementary Material (SM) 1, see SM on http://zhangroup.aporc.org).

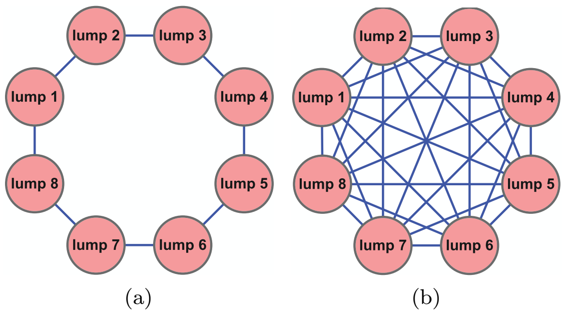
In this paper, to analyze the complex behavior of Q and D including both resolution limit and misidentification we write the Q-optimization and D-optimization as two-stage nonlinear programming problems:

$$Q_{II}: \ \max_{K} Q_I(K) = \max_{K} \max_{P_k} \sum_{s=1}^{K} Q_s \qquad (4)$$

$$D_{II}: \ \max_{K} D_I(K) = \max_{K} \max_{P_k} \sum_{s=1}^{K} D_s \qquad (5)$$

For two widely used exemplar sparse networks called a ring of dense lumps and the *ad hoc* network, the problems (4) and (5) become to two-stage discrete convex/concave programming, then we can solve the detailed solution with a statistical nature (for the detailed analysis, see SM 2). Since the analysis is based on the two exemplary networks, we call our analysis as quasi-analytic.

For the network "ring of lumps" which consists of N ($N \geq 8$ *and* $N = 2^k, k = \{3, 4, 5 \cdots\}$) dense lumps, each with m nodes and l_{in} links (FIG.2(a)). And between two adjacent lumps there are l_{bw} links. When $l_{bw} = 1$ and the lumps are cliques, the ring of lumps becomes a ring of cliques which has been discussed in [21, 22]. Thus, a ring of dense lumps is characterized by three parameters: l_{bw}, l_{in} and N. Parameters l_{bw}, l_{in} are more related



(a)

(b)

FIG. 2: Illustration of the exemplar networks: the ring of lumps and the *ad hoc* network.

to the network structure than N, then we call them as the structure parameters, while N is a scale parameter.

Note that the first-step and second-step optimization problems are all discrete convex programs for Q and D in the case of the ring of lumps (see SM 2). Then they are solvable analytically with specified N value. The optimization solution is

$$K_Q^* = \langle \sqrt{\frac{l_{in} + l_{bw}}{l_{bw}}} \sqrt{N} \rangle_F \tag{6}$$

where $\langle \sqrt{\frac{l_{in}+l_{bw}}{l_{bw}}} \sqrt{N} \rangle_F$ means the point in the integer set $F = \{1, 2, 4, \cdots, N/2^{s+1}, N/2^s, N/2^{s-1}, \cdots, N\}$ nearest to $\sqrt{\frac{l_{in}+l_{bw}}{l_{bw}}} \sqrt{N}$. Similarly, optimizing D gives solution

$$K_D^* = \langle \frac{(l_{in} + l_{bw})N}{4l_{bw}} \rangle_F \tag{7}$$

The solution K_Q^* and K_D^* are unscrambled in FIG.3 in terms of the resolution limit and misidentification phenomena.

From FIG.3, we see that when $l_{bw} < l_{in}/(9N/16 - 1)$, both modularity measures can identify the known communities (with each lump as a community). When l_{bw} is larger than $l_{in}/(9N/16-1)$ and less than $l_{in}/2$, D still can identify the known communities but Q identifies a collection of lumps as a community, i.e., the resolution limit problem appears. D suffers from the resolution limit in the parameter interval $l_{bw} > l_{in}/2, l_{bw} < l_{in}$ which depends on the network structure; Q has resolution limit in the interval $l_{bw} > l_{in}/(9N/16 - 1), l_{bw} < l_{in}$ which depends on the network structure and the scale. For a large scale network, Q is easy to fail to find smaller communities which is less than a scale related to the network structure and scale. When the parameter $l_{bw} > l_{in}$, the lumps are no more qualified communities, the situation becomes more complicated. That is, both Q and D may produce misidentification result, we will not discuss it in details in this paper but only point out that the complex behavior of the modularity functions should be carefully treated.

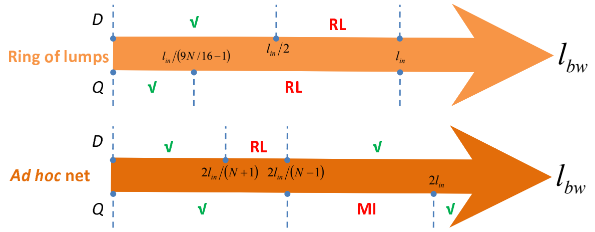


FIG. 3: Illustration of the theoretic analysis for modularity measures Q and D by optimization methodology on two exemplary networks. In this figure, RL and MI respectively mean resolution limit and misidentification, which have been defined above. Green '$\surd$' means the partition is correct.

In the case of the ring of lumps with $l_{in}/(9N/16 - 1) < l_{bw} < l_{in}$, Q identifies a collection of lumps as a community. The number of identified communities by Q is

$$K_Q = \langle \sqrt{\frac{l_{in} + l_{bw}}{l_{bw}}} \sqrt{N} \rangle_F, \tag{8}$$

where K_Q is proportional to $\sqrt{N}$. This result can be verified by the example in [21], where the authors only consider the case $l_{bw} = 1$. In fact, when $l_{bw} = 1$, we have $l_{in}/(N - 1) < 1$ that leads to $l_{in} < (N - 1)$ which implies that the number of cliques is larger than about $\sqrt{Nl_{in} + N - 1}$ which is consistent with the result in [21].

Now we turn to discuss the second exemplar network, the *ad hoc* network [9, 23], which also consists of N dense subgraphs (FIG.2(b)), but has l_{bw} links between each pair of dense subgraphs. Hence, the total number of links in this network is $L = Nl_{in} + N(N-1)l_{bw}/2$. In addition, we assume that all the links in G_s, $s = 1, 2, \cdots, N$, are evenly distributed.

Using the same computation framework in the case of the ring of lumps, the solutions of Q and D optimization on the *ad hoc* network are obtained and explained in FIG.3, where both modularity measures Q and D correctly identify the known communities when $l_{bw} < 2l_{in}/(N + 1)$, while in the interval $(2l_{in}/(N + 1), 2l_{in}/(N - 1))$ Q still works but D falls into the resolution limit although this interval is decreasing with the increase of the network scale. For $l_{bw} > 2l_{in}/(N - 1)$, D works well again but Q misidentifies the communities until $l_{bw} = 2l_{in}$ since it takes subgraphs G_ss as communities which actually do not satisfy the weak community definition.

es the whole Therefore, we conclude that for a class of networks represented by the *ad hoc* network, modularity function Q suffers from more limitations than modularity density D.

The reason causing misidentification is investigated through the relationship between Q,D and the weak community definition. Given a partition $P =$

$\{G_1, G_2, \cdots, G_K\}$=$\{(V_1, E_1)$,$(V_2, E_2)$,$\cdots$,$(V_K, E_K)\}$, $K = 1, \cdots, |V|$, we denote $Q(P), D(P)$ as the values of Q, D on the partition P. For modularity measure D and Q, we have the following proposition (Detailed proof is provided in SM 3).

Proposition Let us denote $D(P) = \sum_{i=1}^{K} D_i$, $Q(P) = \sum_{i=1}^{K} Q_i$, $K \geq 2$. If for any i, G_i satisfies the weak definition [3], then we have $D(P) > 0$ and $Q(P) > 0$.

We note that the reverse of the proposition is not correct, i.e., if $D(P) > 0$ or $Q(P) > 0$, then it is not necessary for all G_i to satisfy the weak definition. An example is shown in FIG.1(a). This indicates that maximizing Q or D makes their value as much positive as possible but not guarantees to produce a feasible partition.

To force a partition as a feasible one, i.e., each community in the partition satisfies the weak definition, it is naturally to have the following constrained Q (D) optimization problem based on the modularity measures:

$$\begin{aligned} &\max\ \textstyle\sum_{s=1}^{K} Q_s \\ &\text{s.t.}\ \ L(V_s, V_s) > L(V_s, \overline{V}_s),\ s = 1, \cdots, k \end{aligned} \tag{9}$$

This is a problem in NP class, to solve them we need to improve the simulation annealing algorithm in [10]. The improved algorithm and software can be find on http://zhangroup.aporc.org. Using the improved algorithm, we correctly solved several examples both in artificial and in real biological and social networks which are misidentified by using the original algorithm (See SM 1).

In conclusion, we systematically analyze the modularity optimization methodology for network community identification in this paper. First, we show that the resolution limit is closely related to the special structure of the network. And the modularity Q is much more sensitive to the resolution limit effect for some special network structures than the modularity density D. The analysis is based on a discrete convex/concave programming framework. Secondly we found that both the modularity measures Q and D suffer from a misidentification problem. We pointed out that the misidentification problem revealed is caused by the fact that optimizing the modularity Q and the density D is not equivalent to search communities that satisfy the weak community definition. In other words, maximizing Q and D can not guarantee all resulting subgraphs as qualified communities. It is noted that Q is much sensitive to the misidentification-potential structure than the D. Furthermore the theoretical analysis on special networks shows that the resolution limit and misidentification alternatively appear in different regions of the parameters. The analysis provides to the users with an insight into limitation and applicability of Q and D.

To overcome the limitations, a constrained optimization model and an revised simulated annealing algorithm are given. Experiments on both simulated data and practical data show that the new algorithm really eliminates the misidentification phenomena. If D is used in the new model and new algorithm, then we can maximally reduce both the resolution limit and misidentification phenomena. Further research is needed to choose better modularity functions or build more complete optimization models.

[1] R. Albert and A. Barabási, Reviews of Modern Physics **74**, 47 (2002).

[2] M. Girvan and M. Newman, Proceedings of the National Academy of Sciences **99**, 7821 (2002).

[3] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proceedings of the National Academy of Sciences **101**, 2658 (2004).

[4] M. Newman, Proceedings of the National Academy of Sciences **103**, 8577 (2006).

[5] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabasi, Science **297**, 1551 (2002).

[6] M. Rosvall and C. Bergstrom, Proceedings of the National Academy of Sciences **104**, 7327 (2007).

[7] M. Rosvall and C. Bergstrom, Proceedings of the National Academy of Sciences **105**, 1118 (2008).

[8] R. Wang, S. Zhang, Y. Wang, X. Zhang, and L. Chen, Neurocomputing **72**, 134 (2008).

[9] M. Newman and M. Girvan, Physical Review E **69**, 26113 (2004).

[10] R. Guimerà and L. Amaral, Nature **433**, 895 (2005).

[11] R. Guimera, S. Mossa, A. Turtschi, and L. Amaral, Proceedings of the National Academy of Sciences **102**, 7794 (2005).

[12] S. Porta, P. Crucitti, and V. Latora, Environment and Planning B: Planning and Design **33**, 705 (2006).

[13] X. Zhu, M. Gerstein, and M. Snyder, Genes & Development **21**, 1010 (2007).

[14] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, Graph-Theoretic Concepts in Computer Science pp. 121–132 (2007).

[15] O. Sporns, C. Honey, and R. Kötter, PLoS ONE **2** (2007).

[16] S. Jalan and J. Bandyopadhyay, Physical Review E **76**, 46107 (2007).

[17] A. Schwarz, A. Gozzi, and A. Bifone, Magnetic Resonance Imaging **26**, 914 (2008).

[18] A. Lancichinetti, S. Fortunato, and F. Radicchi, Physical Review E **78** (2008).

[19] P. Schuetz and A. Caflisch, Physical Review E **78**, 17 (2008).

[20] L. Nayak and R. K. De, Journal of Biomedical Informatics **40**, 726 (2007).

[21] S. Fortunato and M. Barthelemy, Proceedings of the National Academy of Sciences **104**, 36 (2007).

[22] Z. Li, S. Zhang, R. Wang, X. Zhang, and L. Chen, Physical Review E **77**, 36109 (2008).

[23] E. Weinan, T. Li, and E. Vanden-Eijnden, Proceedings of the National Academy of Sciences **105**, 7907 (2008).