Modularity optimization in community identification of complex networks

Supplementary Material 1

Xiang-Sun Zhang [a*], Rui-Sheng Wang [b], Yong Wang [a], Ji-Guang Wang [a], Yu-Qing Qiu [a], Lin Wang [a], and Luonan Chen [c]

[a]Academy of Mathematics and Systems Science, CAS, Beijing 100080, China, [b]Department of Physics, The Pennsylvania State University, University Park, PA 16801, USA, [c]Department of Electronics, Information, and Communication Engineering, Osaka Sangyo University, Osaka 574-8530, Japan.

Experimental results on artificial and real networks

In the main text, we give a systematic analysis based on discrete convex programming 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. In this material, we conduct computational experiments on more general networks, which confirm the misidentification problem in Q and D optimization and illustrate the network configurations that the modularity measures suit for. Furthermore, the optimization model (9) in the main text is executed to show its effectiveness and the algorithm for solving the optimization model (9) is illustrated.

Artificial networks

The first numerical example is a set of computer-generated networks [1] 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 [2, 5–7]. The simulated annealing strategy is used because the optimization of modularity measure is a NP-hard problem. The network partition results by optimizing Q and D are evaluated by the fraction of nodes correctly classified into the original 4 communities.

The average accuracy over 10 *ad hoc* networks with respect to k_{out} is summarized in Figure S1-1, from which we can see that when k_{out} is small, i.e. the networks have distinct communities, both modularity measures have good performance and the detected partitions are nearly the same as the known communities. When k_{out} becomes large, especially when $k_{out} > 8$, the known communities in original networks become very ambiguous, and the partitions detected by both measures are quite different with the early cases. When $6 < k_{out} < 10$, Q seems to have better performance than D. This is natural, since we can see from their definitions that D puts more penalty on outward edges of communities than Q, so when the number of outward edges increases, they have different performance. This does not necessarily indicate that the communities detected by Q are better. Figure S1-2 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

*To whom correspondence should be addressed at: E-mails: zxs@amt.ac.cn

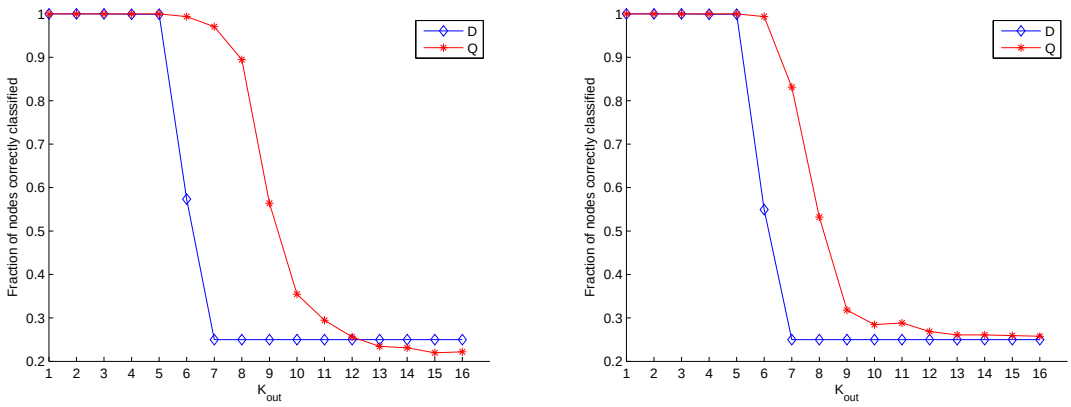

Figure S1-1: Comparison of Q and D in terms of accuracy on 4-community ad hoc networks. It plots the average accuracy over 10 ad hoc networks with respect to parameter k_{out}. (a) The results of direct optimization Q and D. (b) The results of optimization Q and D with constrains which refer to optimization model (9).

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 the optimization model (9) is applied to partition the networks, the misidentification problems are avoided as shown in Figure S1-2.

The *ad hoc* network described above only has four communities, so the resolution limit does not appear in Q optimization. Now we generate another set of *ad hoc* networks with 50 communities of size 8. The average node degree in these communities is 8. The results are evaluated in the same way as the previous example and summarized in Figure S1-3. From Figure S1-3 we can see that, due to the resolution limit in Q, i.e., the tendency to group several known dense subgraphs into one community, Q has a bad performance even for the networks with small k_{out}. This result is consistent with the observation in [3], where Q is found to fail to detect communities smaller than a scale. In contrast, D has a good performance when k_{out} is small. With the increasing of k_{out}, D groups all communities into one since at this time the communities become very ambiguous. From Figure S1-4, we also can see that the communities detected by optimizing D always satisfy the community definition in the weak sense, whereas optimizing Q can lead to "false" communities failing to satisfy the weak community definition. It is same as the above example that the misidentification problems can be wiped out by using optimization model (9) as shown in Figure S1-4. The comparison results in Figures S1-1 and S1-3 also provide us some insights on the network configurations that these two modularity measures suit for. For a network with obvious community structure, Q is a good choice, whereas for a large sparse network with small communities, it is better to use D.

Real networks

We further illustrate our theoretical analysis by several examples of real networks. These networks include some well studied complex networks such as metabolic network of *celegans* [8], dolphin network [9], email network [10], football network [1], jazz musician network [11], political book network [12], 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 is used here. The statistics of the network and the partition results are presented in Table S1-1 and Table S1-2. 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

Table S1-1: Experimental results on the real networks by optimizing Q. The misidentification phenomena are highlighted in bold.

			direct optimization			considering the weak definition constraints through optimization model (9)		
Network name	♯ of nodes	♯ of edges	Q Value	♯ of communities by optimizing Q	♯ of communities satisfying weak definition	Q Value	♯ of communities by optimizing Q	♯ of communities satisfying weak definition
C. celegans metabolic [8]	453	2025	0.45	9	9	0.42	7	7
dolphins [9]	62	159	0.53	4	4	0.52	4	4
email [10]	1133	5451	0.57	10	10	0.57	9	9
football [1]	115	613	0.60	9	9	0.60	10	10
jazz [11]	198	2742	0.44	**4**	**3**	0.44	3	3
karate [4]	34	78	0.42	4	4	0.40	4	4
politics books [12]	105	441	0.53	4	4	0.53	4	4
scienceA [13]	118	200	0.75	7	7	0.75	8	8
Yeast TRN	4441	12873	0.48	**14**	**12**	0.47	13	13
Yeast TFR	162	663	0.35	**6**	**3**	0.22	3	3

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 optimization model (9) which are shown in Table S1-1 and Table S1-2.

A typical example is the jazz musician network [11] 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 musicians [11]. 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 S1-1 and S1-2. 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 S1-5. The community misidentified by Q (**triangles** in Figure S1-5) has 4 nodes and has fewer inner links than outer links. However using the optimization model (9) we can obtain the correct partition.

Furthermore, the experimental results show that the misidentification problem of Q and D is very common in biological networks. We constructed a transcriptional regulatory network (TRN) in yeast from six global ChIP-chip experiments [14–19], which describes the direct interactions between transcription factors (TFs) and target genes (TGs). The yeast TRN network with 163 TFs, 4,405 TGs, and 12,873 transcriptional regulations. The result in Table S1-1 and S1-2 shows that both D and Q suffer from misidentification problem in this large and sparse network. Q has 2 misidentified communities and D has 1 misidentified community. Furthermore, to take a closer look, we extracted the TF regulatory network in yeast (Yeast TFR) from the Yeast TRN by merely considering the regulatory relationships among TFs which results in a smaller and denser network with 162 TFs and 663 regulatory interactions. We find that D partitions the network into four communities where all of them satisfy weak definition of community. While Q partitions the network into six communities where three of them do not satisfy the weak definition. Then we used the Gene Ontology Term Finder `http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.pl` to find the significant shared GO terms to describe the TFs in a community. We simply counted the number of GO terms with p-value less than 1×10^{-9} in every community, then calculated the average number of significant GO terms for a partition. The results show that the partition given by D has 35 significant GO terms while the partition obtained by Q only has 25.6 significant GO terms. It demonstrates that misidentification could leads to misunderstand the network function

Table S1-2: Experimental results on the real networks by optimizing D. The misidentification phenomena are highlighted in bold.

			direct optimization			considering the weak definition constraints through optimization model (9)		
Network name	♯ of nodes	♯ of edges	D Value	♯ of communities by optimizing D	♯ of communities satisfying weak definition	D Value	♯ of communities by optimizing D	♯ of communities satisfying weak definition
C. celegans metabolic [8]	453	2025	30.25	15	15	29.76	16	16
dolphins [9]	62	159	11.73	5	5	11.96	5	5
email [10]	1133	5451	63.16	**31**	**30**	60.03	28	28
football [1]	115	613	43.73	11	11	44.39	11	11
jazz [11]	198	2742	52.84	**4**	**3**	52.03	4	4
karate [4]	34	78	7.85	3	3	7.85	3	3
politics books [12]	105	441	21.86	7	7	20.05	5	5
scienceA [13]	118	200	28.30	16	16	28.31	16	16
Yeast TRN	4441	12873	15.78	**15**	**14**	19.25	23	23
Yeast TFR	162	663	11.50	4	4	11.66	4	4

and D is more effective in detecting biological functional communities than Q. We also checked the functional enrichment of the individual communities identified by Q. For example, we have a misidentified community with 24 nodes in the results given by Q. There are 44 inner edges in this community and 102 edges with other communities. In total there are 29 significant GO terms with p-value less than 1×10^{-9}. We compare it with an identified community of almost similar size satisfying weak definition. This community has 23 nodes, 32 inner links, and 51 outer links and is enriched with 37 significant GO terms. It further provides evidence that misidentification problem may lead to unreasonable network partition and should be carefully considered in designing community detection model.

The simulated annealing algorithm for solving optimization model (9)

Simulated annealing (SA) [7] 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 optimization model (9) by improving the simulated annealing technique, that has been used to solve the Q optimization in [2]. 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 taken as 1 often. 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 moments, 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.

References

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

[2] Guimerà, R, Amaral, L (2005) Functional cartography of complex metabolic networks. *Nature* 433:895.

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

[4] Li, Z, Zhang, S, Wang, R, Zhang, X, Chen, L (2008) Quantitative function for community detection. *Physical Review E* 77:36109.

[5] Colizza, V, Pastor-Satorras, R, Vespignani, A (2007) ReactionCdiffusion processes and metapopulation models in heterogeneous networks. *Nature Physics* 3:276.

[6] Schauer, N et al. (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature Biotechnology* 24:447–454.

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

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

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

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

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

[12] Adamic, L, Glance, N (2005) *The political blogosphere and the 2004 US election: divided they blog* (ACM New York, NY, USA), pp 36–43.

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

[14] Borneman, A et al. (2006) Target hub proteins serve as master regulators of development in yeast. *Genes & Development* 20:435–448.

[15] Borneman, A et al. (2007) Transcription factor binding site identification in yeast: a comparison of high-density oligonucleotide and PCR-based microarray platforms. *Functional & Integrative Genomics* 7:335–345.

[16] Workman, C et al. (2006) A systems approach to mapping DNA damage response pathways. *Science* 312:1054–1059.

[17] Horak, C et al. (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes & Development* 16:3017–3033.

[18] Harbison, C et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104.

[19] Lee, T et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298:799–804.

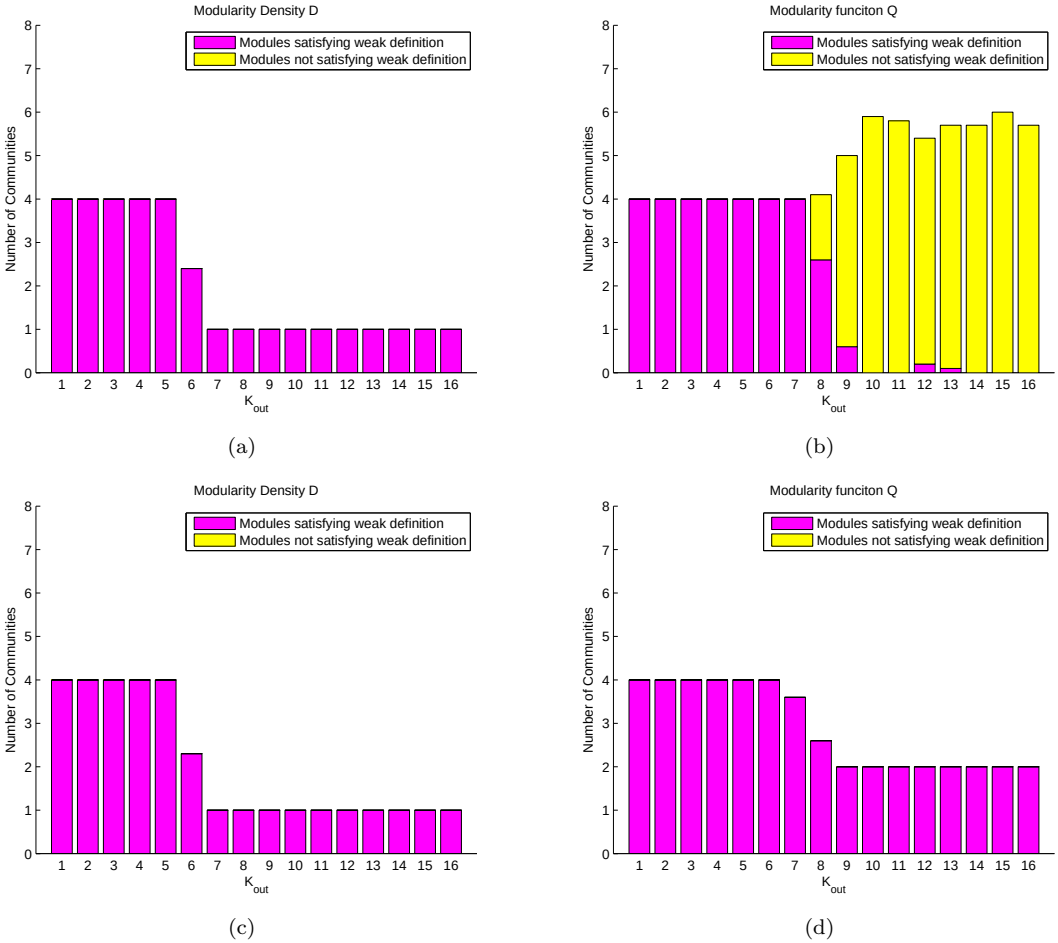

Figure S1-2: Comparison of Q and D in terms of misidentification on 4-community *ad hoc* networks. (a) The community numbers detected by optimizing modularity density D (average value over 10 networks). (b) The community numbers detected by optimizing modularity function Q (average value over 10 networks). (c) The community numbers detected by optimizing modularity density D with the weak condition constrains which refer to optimization problem (9). (d) The community numbers detected by optimizing modularity function Q with the weak condition constrains which refer to optimization problem (9). It shows that in this example D has no misidentification problem, whereas Q suffers from the misidentification problem when $k_{out} > 8$. Under the weak definition constrains, Q is free of misidentification limitation.

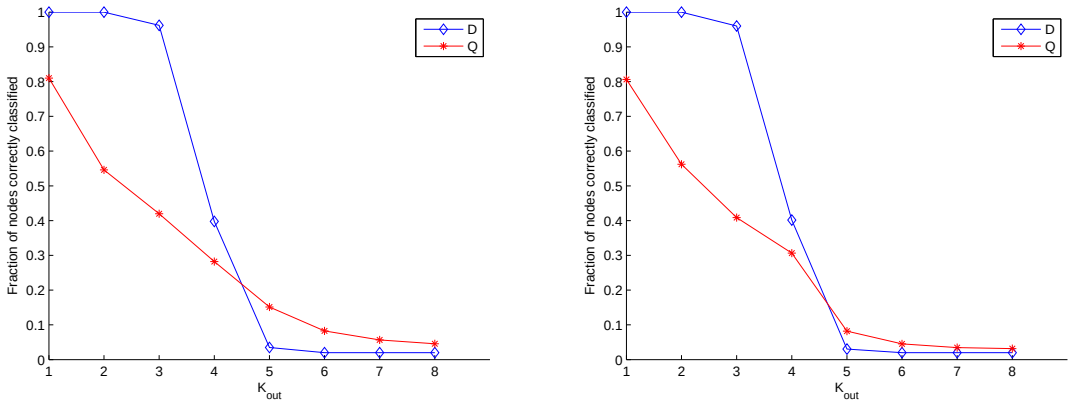

Figure S1-3: Comparison of Q and D in terms of accuracy on 50-community *ad hoc* networks. It shows that modularity function Q suffers from the resolution limit problem when k_{out} is small. (a) The results of direct optimization Q and D. (b) The results of optimization Q and D with constrains which refer to optimization model (9).

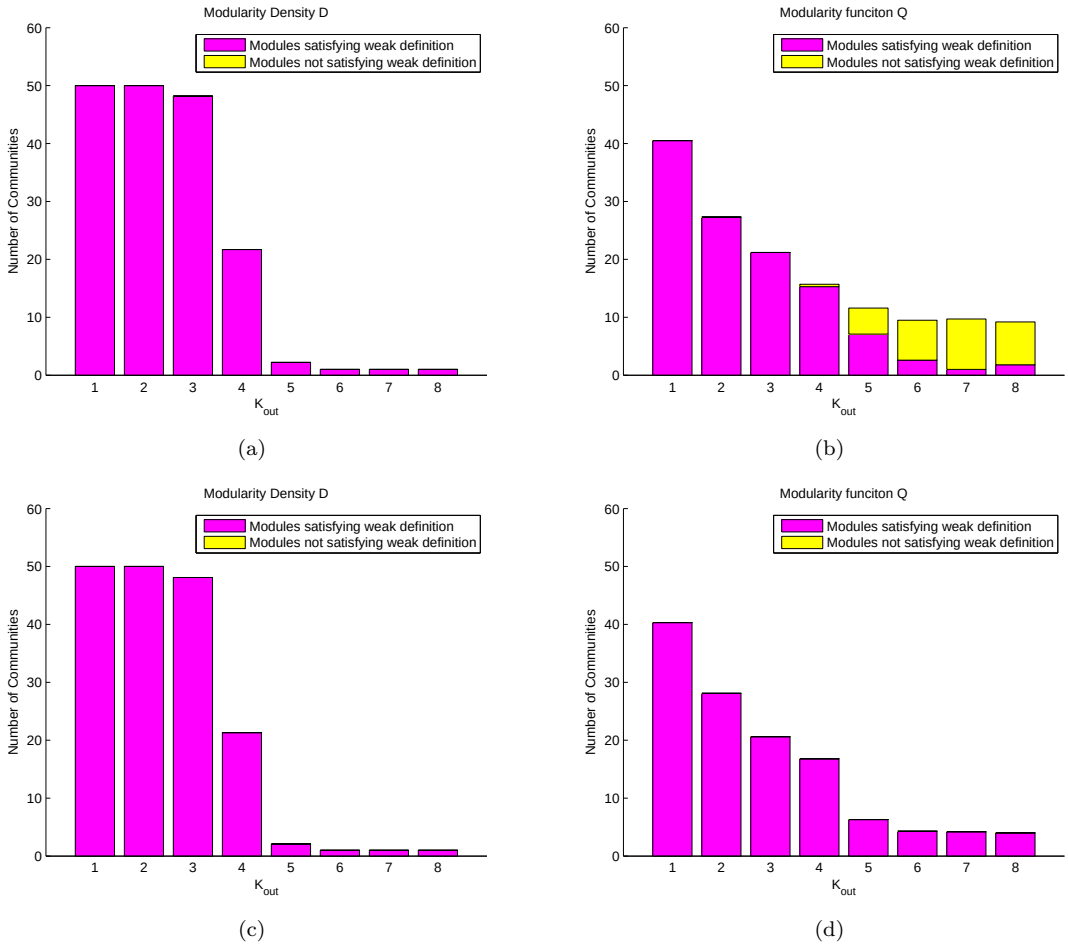

Figure S1-4: Comparison of Q and D in terms of misidentification on 50-community *ad hoc* networks. (a) The community numbers detected by optimizing modularity density D (average value over 10 networks). (b) The community numbers detected by optimizing modularity function Q (average value over 10 networks). (c) The community numbers detected by optimizing modularity density D with the weak condition constrains which refer to optimization problem (9). (d) The community numbers detected by optimizing modularity function Q with the weak condition constrains which refer to optimization problem (9). It shows that D has no misidentification problem in this example, whereas Q suffers from the misidentification problem when $k_{out} > 4$. Under the weak definition constrains, Q is free of misidentification problems. Q also suffers from the resolution limit when $k_{out} = 1$.

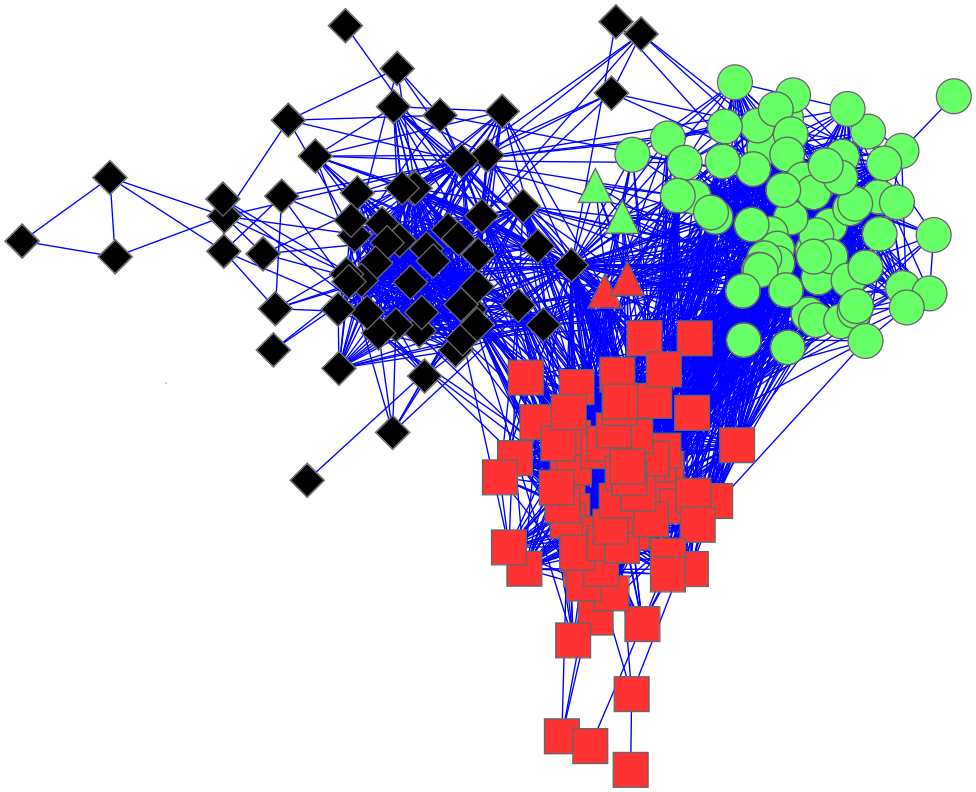

Figure S1-5: Misidentification in optimization of modularity measures Q in the jazz musician network. In this network, the nodes denote the jazz bands and edges represent jazz musician sharing relationships. In reality, there are three communities (white musician community, black musician community located in New York, and black musician community located in Chicago) which represent black/white racial segregation and cities that bands recorded in. 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 optimization model (9). The results show that Q find a community unsatisfying the weak definition. The misidentified community by Q (**triangles**) has 4 nodes with fewer inner links than outer links. However the optimization model (9) can correctly partition the network.