Eur. Phys. J. B (2012) 85: 352
DOI: 10.1140/epjb/e2012-30301-2

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Multi-resolution modularity methods and their limitations in community detection

J. Xiang[1], X.G. Hu[2], X.Y. Zhang[1], J.F. Fan[1], X.L. Zeng[1], G.Y. Fu[1], K. Deng[3], and K. Hu[4,a]

[1] Department of Basic Sciences, The First Aeronautical Institute of The Air Force, Xinyang 464000, Henan, P.R. China
[2] School of Information Engineering, HuangShan University, HuangShan 245021, Anhui, P.R. China
[3] Department of Physics, Jishou University, Jishou 416000, Hunan, P.R. China
[4] Department of Physics, Xiangtan University, Xiangtan 411105, Hunan, P.R. China

Received 8 April 2012 / Received in final form 26 June 2012
Published online 22 October 2012 –

Abstract. Community detection is of considerable importance for understanding the structure and function of complex networks. Recently, many multi-resolution methods have been proposed to uncover community structures of networks at different scales. Here, different multi-resolution methods are derived from modularity using self-loop assignment schemes, and then a set of multi-resolution modularity methods of this type are presented. These methods are carefully investigated by theoretical analysis of the transition points of the multi-resolution processes and experimental tests in model networks. Compared with the degree-dependent self-loop assignment, the mean-degree-dependent self-loop assignment can quicken the disconnecting of (small) communities with small vertex degrees, and can slow down the breakup of (large) communities with large vertex degrees. Moreover, we show that all these methods will encounter a limitation which is independent of the network size: large communities will break up before small communities are revealed by increasing their resolution parameters when the distribution of community sizes is very broad. Also, the tolerance of different methods against the limitation is different.

1 Introduction

In recent years, community structure in complex networks, an important topological property common to many real-world networks including social, ecological, biological and technological networks, has attracted much attention in physics and many interdisciplinary fields [1,2]. Networks with community structure generally consist of densely connected groups of vertices (known as communities or modules) surrounded by sparsely connected regions [2]. Detecting such communities in networks can provide a useful coarse-grained representation of complex networks, and will help in understanding their structure and function of the networks [3–7]. Since the work of Girvan and Newman [8], a large number of community detection algorithms have been proposed to detect and analyze the community structure in complex networks based on various approaches such as dissimilarity or similarity measures [9–11], spectral analysis [12,13], resistance models [14], random walk dynamics [15–19], label propagation [20–22], statistical models [23,24] and modularity optimization [25–29], as well as the combination of various methods (see Refs. [2,30,31] for reviews).

It is interesting that most of the popular methods for community detection in networks often consist of the optimization of quality functions that can evaluate community divisions in the networks, such as the famous modularity function initially proposed by Newman and Girvan [32]. Given a community division of a network, the modularity function Q can be written as

$$Q = \sum_s \frac{k_s^{in}}{2M} - \left(\frac{k_s}{2M}\right)^2, \tag{1.1}$$

where M is the total number of edges in the network, k_s^{in} the inner degree of group s, k_s the total degree of group s, and the sum runs over all communities in the given network. For the weighted case, M becomes the total weight of edges in the network, k_s^{in} the inner weighted degree of group s and k_s the total weighted degree of group s [33]. The modularity evaluates the fraction of edges within communities in the network minus the expected value in a random graph with the same degree distribution (i.e. in the configuration *null* model). In general, the larger the modularity, the better the division. In recent years, the modularity has become one of the most popular quality functions, is used widely by various community detection algorithms, and has proven able to give useful and relevant

[a] e-mail: huke1998@yahoo.cn

results in many complex networks. However, the modularity optimization also suffers from some difficulties [34,35], such as the resolution limit [36], which can be expressed by

$$k_s k_t < 2M \cdot e_{st}, \tag{1.2}$$

where k_s and k_t are the total degrees of communities s and t, e_{st} is the number of edges between them, and M is the total number of edges in the network. That is to say, if the inequality is satisfied, the (small) communities s and t will be put together into one group by modularity-based methods, even if they are cliques linked only by a single edge, because this results in greater modularity if so. The resolution limit means that the modularity optimization fails to detect communities below a certain characteristic scale, which is not compatible with the multi-scale structures in many real-world networks.

To attack the resolution problem, various methods have been proposed, such as the recursive partitioning of networks [37] or the recursive re-weighting methods of the inter- and intra-community edges [38,39]. In particular, it has been shown that various multi-resolution methods with tunable resolution parameters can help to discover community structures at different scales [23,40–44]. For example, Zhang et al. [42] proposed a method with tunable resolution by seeding the kernel in graphs. Li et al. [43] investigated community detection based on Potts model and the network's spectral characterization, and showed that the local, uniform behavior of spins in Potts model can naturally reveal the hierarchical community structures in networks. From the Potts-spin-model point of view, Reichardt and Bornholdt (RB) [23] proposed a general ansatz for the quality function, and discussed their multi-resolution modularity functions as special cases. This scheme was also applied recently by Ronhovde and Nussinov [45] and by Traag et al. [46]. Arenas, Fernández and Gómez (AFG) [44] also proposed a multi-resolution modularity method by providing each vertex with a self-loop of the same magnitude r, which is equivalent to modifying the modularity function by the parameter r. In this paper, we will focus on these multi-resolution methods that can directly adjust the resolution of modularity, though there have been few discussions about them to date [47]. We will present a critical analysis of the applicability of the methods to the problem of community detection. It maybe provides valid guidelines for the design of new effective methods in practical applications. Also, we believe that these methods are worthy of more thorough investigation before being applied to community detection problems in the real world.

The paper is organized as follow. In Section 2, different multi-resolution modularity methods (AFG and RB) will be derived using self-loop assignment scheme, i.e. by providing each vertex with a suitable self-loop, and then a set of methods of this type will be proposed based on that scheme. In Section 3, we will focus on the transition points where small communities become visible and where large communities begin to break up, and we will compare the general processes of these multi-resolution methods by theoretical analysis and experimental tests on model networks. In Section 4, a kind of intrinsic contradiction in the methods will be highlighted and discussed – increasing the values of the resolution parameters may in some cases cause large communities to break up before small communities become apparent. Finally, we will come to our conclusion.

2 Multi-resolution modularity methods

In this section, we will introduce multi-resolution modularity methods. These methods can be derived from the Potts spin model [46], and are closely related to the time scales of random walks on networks [48]. Here, we introduce these methods using self-loop assignment schemes, i.e. by providing each vertex with a suitable self-loop, and then, based on this scheme, a set of multi-resolution modularity methods of this type will be presented. The self-loop assignment scheme makes the multi-resolution modularity functions mathematically very similar to the modularity, and thus these can be optimized directly by the existing modularity optimization algorithms with minimum code development. Of course, the multi-scale modularity functions can also be optimized by other methods such as Potts-model-based heuristics [23].

2.1 AFG method

The multi-resolution modularity method proposed by Arenas, Fernández and Gómez (AFG) is a representation based on the self-loop assignment scheme [44]. Using a tunable parameter, it rescales the topology of the network by directly assigning a self-loop of the same magnitude to each vertex. Here, we will instead provide each vertex with a self-loop of strength $\bar{k}r$, where $\bar{k}$ is the mean degree of vertices in the network and r is a tunable parameter. We refer to this method as the *Mean-Degree-Dependent Self-loop Assignment* scheme (MDD-SA). With the introduction of the self-loop $\bar{k}r$, the multi-resolution modularity function in the AFG method can be written as

$$\begin{aligned} Q^{AFG}(r) &= \sum_s \left(\frac{k_s^{in} + n_s \bar{k} r}{2M + N\bar{k}r} - \left(\frac{k_s + n_s \bar{k} r}{2M + N\bar{k}r} \right)^2 \right) \\ &= \sum_s \left(\frac{k_s^{in} + \bar{k}_s r}{2M(1+r)} - \left(\frac{k_s + \bar{k}_s r}{2M(1+r)} \right)^2 \right) \\ &= \frac{r}{(1+r)} + \frac{1}{(1+r)} \\ &\quad \times \sum_s \left(\frac{k_s^{in}}{2M} - (1+r) \left(\frac{\overline{\overline{k_s(r)}}}{2M} \right)^2 \right), \end{aligned} \tag{2.1}$$

where k_s^{in} is the inner degree of the group s, k_s is the total degree of the group s, n_s is the number of vertices in the group s, N is the number of vertices in the network, $\bar{k}_s = n_s \bar{k}$, $2M = N\bar{k}$, and the effective total degree of the group $\bar{\bar{k}}_s = \overline{\overline{k_s(r)}} = (k_s + \bar{k}_s r)/(1+r)$ is a function of r, which can be regarded as the r-weighted average value

of k_s and $\bar{k}_s$. Ignoring the constant term and multiplier before the summation, which are independent of community divisions in networks, we can see that it is similar to the modularity (2.3). In contrast, MDD-SA simultaneously changes the form of the null model to $\left(\overline{\overline{k_s(r)}}/2M\right)^2$ and adds a pre-factor $(1+r)$ which tunes the contribution of the null model by the parameter r. According to the above modularity function, the resolution inequality becomes

$$\begin{aligned}&(k_s+\bar{k}_s r)(k_t+\bar{k}_t r)<2M(1+r)e_{ts}\\ \text{or}\quad &\overline{\overline{k_s(r)}}\times\overline{\overline{k_t(r)}}<2Me_{ts}/(1+r).\end{aligned}\tag{2.2}$$

2.2 RB method

Reichardt and Bornholdt (RB) [23] have shown that the problem of community detection can be mapped onto finding the ground state of an infinite range Potts spin glass, where the spin states correspond to the community indices. Similarly to the modularity, the RB Potts model makes use of a null model as a comparison. The larger the deviation from the null model, the better the community divisions. The classical null model is the configuration null model used in the modularity (1.1), preserving the degree distribution of the corresponding networks. The RB modularity with the configuration null model, which has the modularity (1.1) as a special case, can be written as

$$Q^{RB}=\sum_s\frac{k_s^{in}}{2M}-\gamma\left(\frac{k_s}{2M}\right)^2,\tag{2.3}$$

where γ is the pre-factor for tuning the contribution of the null model, and the remaining notation is the same as in modularity (1.1). Accordingly, the resolution inequality becomes

$$k_sk_t<2M\cdot e_{st}/\gamma,\tag{2.4}$$

where the notation is the same as for inequality (1.2). While the RB method was not designed to study the resolution of modularity, it still provides a workaround for the issue.

Interestingly, the RB method can be obtained by the *Degree-Dependent Self-loop Assignment* scheme (DD-SA), i.e. by assigning to each vertex a self-loop proportional to its degree. Specifically, we are to provide each vertex with a self-loop of strength k_ir, where k_i is the degree of the vertex i in the network and r is a tunable parameter. As a result, the modularity (1.1) can be re-written as

$$\begin{aligned}Q^{RB}(r)&=\sum_s\left(\frac{k_s^{in}+k_sr}{2M+2Mr}-\left(\frac{k_s+k_sr}{2M+2Mr}\right)^2\right)\\&=\frac{r}{1+r}+\frac{1}{1+r}\sum_s\left(\frac{k_s^{in}}{2M}-(1+r)\left(\frac{k_s}{2M}\right)^2\right),\end{aligned}\tag{2.5}$$

where the remaining notation is the same as in modularity (1.1). Clearly, the modularity (2.5) is equivalent to the modularity (2.3) and the relation $\gamma=1+r$ applies to all divisions in networks as the factors before the summation do not change the results of the optimization. It suggests that the optimization of the RB modularity can also be done by the existing optimization algorithms if we rescale the network topology using the self-loop scheme. Unlike MDD-SA, however, we see that DD-SA only adds a pre-factor which tunes the contribution of the null model by the parameter r and does not change the form of the null model.

2.3 General methods based on self-loop assignment

Based on the above observation, we here introduce a set of general multi-resolution modularity methods by assigning each vertex a self-loop of strength $[\bar{k}(1-\mu)+k_i\mu]r$, where $\bar{k}$ is the mean degree of vertices in network, k_i the degree of the vertex i, r a parameter tuning the resolution of modularity, and μ $(0\leqslant\mu\leqslant1)$ a parameter to control the contributions of $\bar{k}$ and k_i (i.e. the contributions of MDD-SA and DD-SA) to the null model. As a result, the modularity can be re-written as

$$\begin{aligned}Q(\mu,r)&=\sum_s\left(\frac{k_s^{in}+[\bar{k}_s(1-\mu)+k_s\mu]r}{2M+2Mr}\right.\\&\quad\left.-\left(\frac{k_s+[\bar{k}_s(1-\mu)+k_s\mu]r}{2M+2Mr}\right)^2\right)\\&=\frac{r}{1+r}+\frac{1}{1+r}\sum_s\left(\frac{k_s^{in}}{2M}-(1+r)\right.\\&\quad\left.\times\left(\frac{k_s+[\bar{k}_s(1-\mu)+k_s\mu]r}{2M(1+r)}\right)^2\right)\\&=\frac{r}{1+r}+\frac{1}{1+r}\sum_s\left(\frac{k_s^{in}}{2M}-(1+r)\right.\\&\quad\left.\times\left(\frac{\overline{\overline{k_s(\mu,r)}}}{2M(1+r)}\right)^2\right),\end{aligned}\tag{2.6}$$

and the resolution inequality accordingly becomes

$$\begin{aligned}&\overline{\overline{k_s(\mu,r)}}\times\overline{\overline{k_t(\mu,r)}}<2Me_{ts}/(1+r),\\&\text{where}\quad\bar{\bar{k}}_s=\overline{\overline{k_s(\mu,r)}}=\frac{k_s+[\bar{k}_s(1-\mu)+k_s\mu]r}{1+r}.\end{aligned}\tag{2.7}$$

Clearly, both the inequalities (2.2) and (2.4) are contained within it.

By tuning the value of the parameter μ, we can obtain the multi-resolution methods from AFG to RB. For $\mu=0$, it corresponds to AFG, while RB corresponds to $\mu=1$. With regards to the general multi-resolution methods (including AFG and RB as special cases), modularity (2.6) is the same as modularity (1.1) for $r=0$. When $r<0$, we can find the superstructures above those at $r=0$. When $r>0$, we can obtain the substructures

under those at $r = 0$. All the multi-resolution methods will give the same results on their initial iteration (the whole network is regarded as one group) and on their final iteration (the whole network is separated into a set of single-vertex groups). However, results from intermediate iterations generally differ from each other.

Moreover, note that all these multi-resolution methods are equivalent in homogeneous networks where all the vertices have the same degree and must therefore produce the same results. In the following sections we will analyze the processes of these methods on non-homogeneous networks with hierarchical community structures.

3 Comparison of multi-resolution processes

For convenience of analysis, we construct a type of simple model network with two types of pre-defined communities of different sizes (see Fig. 1a). Specifically, the networks consist of cliques (fully-connected sub-graphs) of two sizes. The numbers of the *large* cliques and the *small* cliques in the networks are respectively m and $2m$ (m is equivalent to the number of clique-sets with one large clique and two small cliques), and the sizes of them are denoted respectively by n_l and $n_s (n_l > n_s)$. The cliques are connected one by one with single edges, generating a simple, ring-like configuration, and the large cliques are evenly spaced with two small cliques separating them. Here, we discuss unweighted and un-directed networks, and the small cliques are invisible for the modularity Q. As an example, we generated a network of this type with $m = 6$, $n_l = 16$ and $n_s = 6$ (see Fig. 1). The mean degree of the network ($5 < \bar{k} < 15$) is clearly between the vertex degrees of the large cliques and the small cliques. This important property will be used in the following theoretical analysis. Moreover, in the following experimental tests, we will detect the optimal community divisions of the networks corresponding to different values of resolution parameters, by using the fast greedy algorithm of Blondel et al. [29], and check the results by using extremal-optimization algorithm [26].

3.1 RB process (DD-SA)

From simplicity to complexity, we first study the processes of the RB and AFG methods (i.e. $\mu = 1$ and 0) by focusing on several of their transition points, and then discuss the general methods ($0 \leqslant \mu \leqslant 1$). The RB method corresponds to the multi-resolution method with $\mu = 1$. For simplicity, we define $\gamma = r + 1$. When $\gamma = 1$, we can obtain the same result as the modularity Q. By tuning the γ-value, we can adjust the resolution of the modularity. Here, we focus on $\gamma > 1$. As shown in Figure 1b, the small communities become detectable with increasing γ. Thus we can find 18 cliques, i.e. all the predefined communities. The exact point at which this occurs is when the value of $\gamma = r + 1$ satisfies

$$\gamma > \gamma_{a1} = r_{a1} + 1 = \frac{2Me_{st}}{k_s k_t}, \tag{3.1}$$

where the notation is the same as in the inequality (2.4). γ_{a1} is the first transition point of the RB process in the network. (See the dotted γ_1-line for γ_{a1} in Fig. 1b)

After the small communities become detectable, we can see a stable community division in the network which is not affected by increasing γ, untilthe large communities begin to be broken up. Here, we consider bi-partitions of the communities [49,50]. The expected γ-value for the equal-sized bi-partition of the *large* community l as a random sub-graph can be denoted as (see Appendix)

$$\gamma_{a2} = r_{a2} + 1 \simeq 2M\frac{4E_{AB}}{k_l^2}, \tag{3.2}$$

where M is the total number of edges in the network, E_{AB} is the expected number of edges between two equally-sized parts A and B of the community, and k_l is the total degree of the community. γ_{a2} is the second transition point of the RB process in the network. (See the dotted γ_2-line for γ_{a2} in Fig. 1b). Finally, when the γ-value ($\gamma > \gamma_{\max} = 2M/k_{\min}^2$, $k_{\min}$ is the minimal degree of vertices in the network) is large enough, the whole network will be separated into a set of single-vertex groups by maximizing the multi-resolution modularity.

3.2 AFG process (MDD-SA)

The multi-resolution method with $\mu = 0$ corresponds to the AFG method. We will apply it to the above example network, and compare with the RB process. Again, we define $\gamma = r + 1$. As γ increases, the small cliques are disconnected first, and thus we can find all the predefined communities (see the dotted γ_1-line in Fig. 1c), when the value of γ satisfies

$$\gamma > \gamma_{b1} = r_{b1} + 1 = \frac{2Me_{ts}}{\overline{\overline{k_s}} \times \overline{\overline{k_t}}}, \tag{3.3}$$

where $\overline{\overline{k_s}} = \overline{\overline{k_s(r_{b1})}} = (k_s + \bar{k}_s r_{b1})/(1 + r_{b1})$ is the effective total degree of the clique s when $r = r_{b1}$, which can be regarded as the r_{b1}-weighted average value of k_s and $\bar{k}_s (\bar{k}_s = n_s\bar{k})$, and the remaining notation is the same as in equation (2.2). It is not difficult to obtain the exact solution of the inequality. For the sake of comparison, we have formatted the expression similarly to equation (3.1).

Comparing with equation (3.1), we can infer that $\gamma_{b1} < \gamma_{a1}$, because there are $k_s < \bar{\bar{k}}_s$ and $k_t < \bar{\bar{k}}_t$ for the small cliques in this network. In other words, $\gamma_{b1} < \gamma_{a1}$ because the vertex degree in small cliques (s and t) is less than that of the whole network. This is proved by the γ_1-lines in Figures 1b–1c. As a result, compared with DD-SA (RB), MDD-SA (AFG) can quicken the disconnecting of small communities in which the degree of vertices is less than the mean degree of the whole network. Therefore, Figure 1 shows that these small communities can be found earlier by AFG than RB when the γ-value is increased (note that MDD-SA will delay the disconnecting of the small communities if the vertex degree inside them is larger than the mean degree of the whole network).

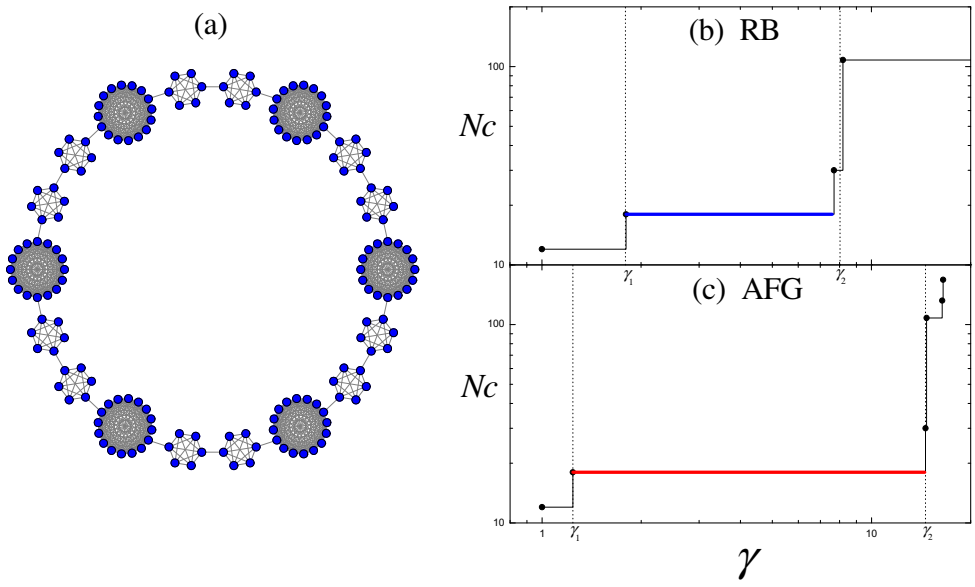

Fig. 1. (Color online) (a) Network with $m = 6$, $n_l = 16$ and $n_s = 6$, described in text. (b) and (c) the curves of the γ-value and the number Nc of groups found by maximizing the RB and AFG modularity in the network. The bold solid lines between γ_1 and γ_2 corresponds to all the cliques being revealed simultaneously, (b) by RB and (c) by AFG.

After the small communities become visible, as shown in Figure 1c, a stable community division of the network that is not affected by an increase in γ can be found, until the large communities begin to be broken up. Similarly, the expected γ-value for the equally-sized bi-partition of the community l as a random sub-graph can be denoted as (see Appendix)

$$\gamma_{b2} = r_{b2} + 1 \simeq 2M \frac{4E_{AB}}{\left(\bar{\bar{k}}_l\right)^2}, \tag{3.4}$$

where M is the total number of edges in the network, E_{AB} is the expected number of edges between the two equally-sized parts A and B, and $\bar{\bar{k}}_l = \overline{\overline{k_l(r_{b2})}} = (k_l + \bar{k}_l r_{b2})/(1 + r_{b2})$ is the effective total degree of the community l when $r = r_{b2}$, which is the r_{b2}-weighted average value of k_l and $\bar{k}_l$ ($\bar{k}_l = n_l \bar{k}$). (See the dotted γ_2-line for γ_{b2} in Fig. 1c.) For the sake of comparison, we give an expression for γ_{b2} that is similar to equation (3.2). When $\gamma > \gamma_{b2}$, the large community will break up. When the value of γ is large enough, the whole network will be separated into a set of single-vertex groups by AFG.

Comparing with equation (3.2), we can find $\gamma_{a2} < \gamma_{b2}$, because $k_l > \bar{\bar{k}}_l$ for large cliques in this network (or because the vertex degree in the community l is larger than that of the whole network). This is proved by the γ_2-lines in Figures 1b−1c. With regards to disconnecting small communities, MDD-SA (AFG) differs from DD-SA (RB) in that it will s*low down* the breakup of the large communities in which the vertex degree is greater than the mean degree of vertices in the whole network. Therefore, we can see in Figure 1 that the large cliques break up first in the RB method, before this occurs in the AFG method (note that MDD-SA will quicken the breakup of the large communities if the vertex degree in them is less than the mean degree of the whole network).

3.3 General processes ($0 \leqslant \mu \leqslant 1$)

In this section, we will test the multi-resolution methods with $0 \leqslant \mu \leqslant 1$ on the model network above. We will still focus on the two transition points of γ_1 and γ_2. It is easy to give the expressions of γ_1 and γ_2, similar to (3.3) and (3.4),

$$\gamma_1(\mu) = r_1(\mu) + 1 = \frac{2Me_{ts}}{\bar{\bar{k}}_s \bar{\bar{k}}_t}, \tag{3.5}$$

$$\gamma_2(\mu) = r_2(\mu) + 1 = 2M \frac{4E_{AB}}{\left(\bar{\bar{k}}_l\right)^2}, \tag{3.6}$$

where the notation is the same as in (2.7) and (3.4). In the clique-ring network, as shown in Figure 2a, the real and expected values of γ_1 are accurately consistent, and the real and expected values of γ_2 are very close as anticipated (the expected values of γ_2 can be estimated by different approaches, see Appendix). Generally, the values of γ_1 and γ_2 will vary with μ. In Figure 2a, we can see that $\gamma_1(\gamma_{b1} \leqslant \gamma_1 \leqslant \gamma_{a1})$ increases with increasing values of μ, while $\gamma_2(\gamma_{a2} \leqslant \gamma_2 \leqslant \gamma_{b2})$ decreases with increasing values of μ, which can be obtained by theoretical analysis of equations (3.5) and (3.6) in this test network. Consequently, the length of the plateau from γ_1 to γ_2 decreases when μ is increased. We highlight that the parameter μ can adjust the contributions of MDD-SA and DD-SA to the null model in the modularity (2.6). Comparing the multi-resolution methods from $\mu = 0$ to $\mu = 1$, especially AFG and RB, we can find a phenomenon of interest: MDD-SA tends to lengthen the plateau from γ_1 to γ_2

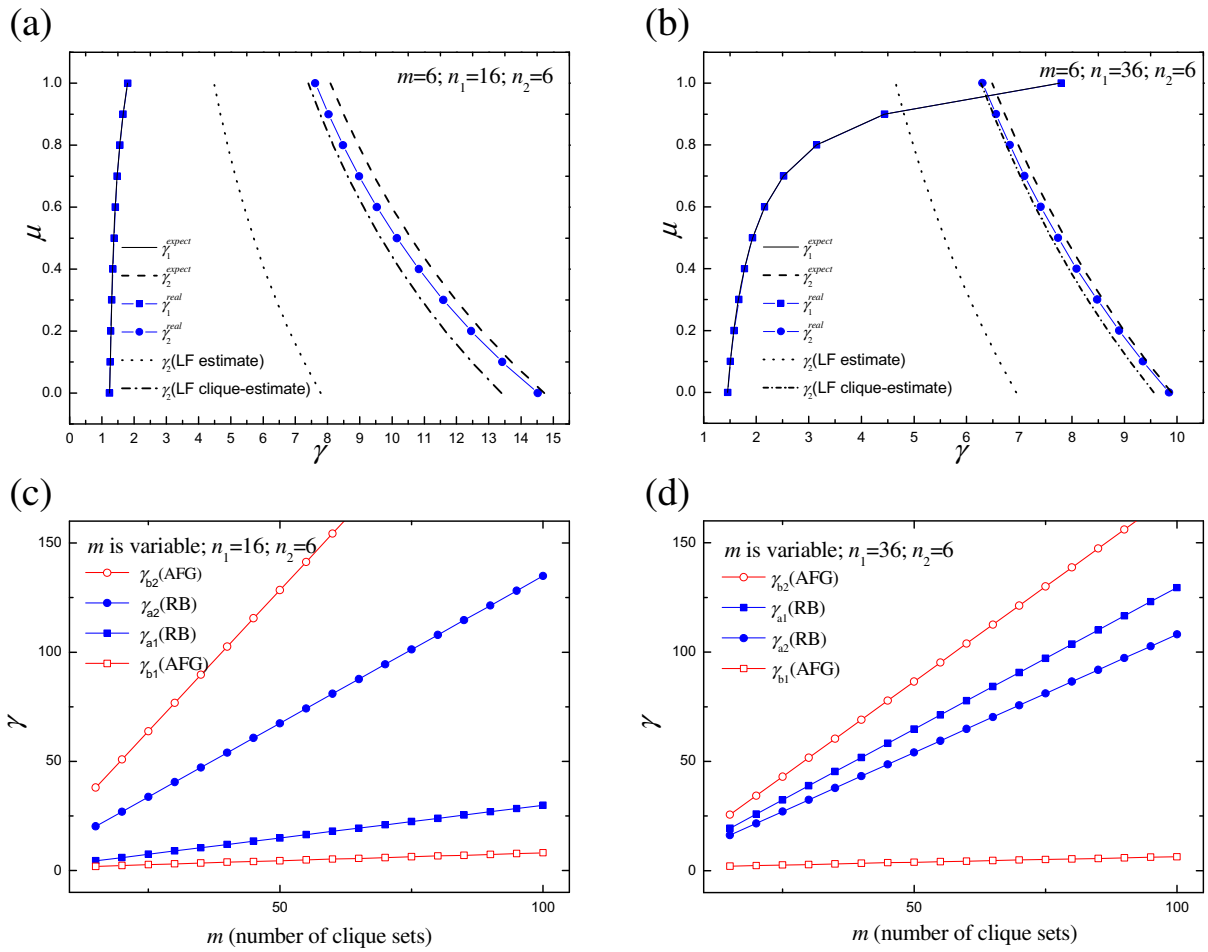

Fig. 2. (Color online) (a)–(b) Critical parameters γ_1 and γ_2 as a function of μ in the clique-ring networks described in the text: (a) with $m = 6$, $n_1 = 16$ and $n_2 = 6$ and (b) with $m = 6$, $n_1 = 36$ and $n_2 = 6$. The real values for γ_1 and γ_2 are denoted by γ_1^{real} and γ_2^{real} respectively. γ_1^{expect} is the theoretical line for γ_1. γ_2^{expect} is the theoretical estimate for γ_2 by the Simple estimate, and γ_2 (LF estimate) and γ_2 (LF clique-estimate) are the theoretical estimates for γ_2 using the LF estimate (see Appendix). (c)–(d) Critical parameters γ_1 and γ_2 as a function of m (which is proportional to the size of the networks) by the RB and AFG methods in the clique-ring networks in text, preserving the difference in community size: (c) with $n_1 = 16$ and $n_2 = 6$ and (d) with $n_1 = 36$ and $n_2 = 6$.

in the networks. In general, it is believed that the length of the plateau is related to the stability of the community division found, while these results clearly indicate that it is also closely related to the multi-resolution methods themselves.

3.4 Further analysis

3.4.1 Effect of network size

Further, we study the two transition points of γ_1 and γ_2 by increasing the size of the networks, preserving the difference between the sizes of the small cliques and of large cliques. In Figure 2c, we can see that the inequalities $\gamma_{b1}(AFG) < \gamma_{a1}(RB)$ and $\gamma_{a2}(RB) < \gamma_{b2}(AFG)$ remain valid when the network size is increased. For other μ-values, the γ_1-lines will be between γ_{b1} and γ_{a1}, while the γ_2-lines will be between γ_{a2} and γ_{b2}. These results can be obtained by analyzing the self-loop assignment rule in the modularity (2.6). Moreover, as shown in Figure 2c, both γ_1 and γ_2 (for AFG and RB) will increase with the size of the network, which is to be expected according to equations (3.1)–(3.4), and the intervals between γ_1 and γ_2 will increase with the size of the network for both RB and AFG. These can be extended to other multi-methods ($0 \leqslant \mu \leqslant 1$).

3.4.2 Effect of community size difference

In Figure 3, we show that the critical parameters γ_1 and γ_2 will vary when the difference in community sizes is increased. Similarly, the inequalities $\gamma_{b1} < \gamma_{a1}$ and $\gamma_{a2} < \gamma_{b2}$ are always satisfied in all the networks. We can see that γ_1 increases with increasing community size difference, while γ_2 decreases with increasing community size difference. This results in reduced intervals between γ_1

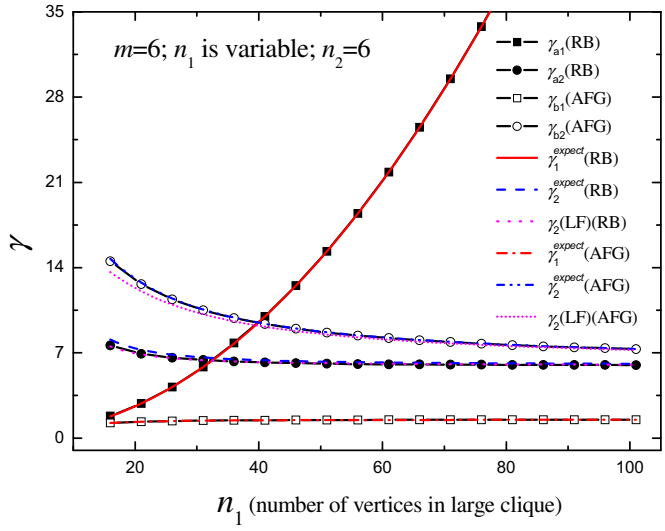

Fig. 3. (Color online) Critical parameters γ_1 and γ_2 as a function of n_1 in the clique-ring networks described in the text, by the RB and AFG methods. γ_{a1} and γ_{a2} correspond to the real values for γ_1 and γ_2 by RB, γ_{b1} and γ_{b2} by AFG. γ_1^{expect} and γ_2^{expect} are the theoretical lines for γ_1 and γ_2 respectively by the Simple estimate, and γ_2(LF) is the theoretical estimate for γ_2 by the LF estimate (see Appendix).

and γ_2. This is true for both AFG and RB, and can easily be extend to all the multi-resolution methods of this type. Furthermore, for other μ-values, we can also infer that the γ_1-lines will be between γ_{b1} and γ_{a1} while the γ_2-lines will be between γ_{a2} and γ_{b2}. Interestingly, when the community size difference is large enough (i.e. $\gamma_2 < \gamma_1$) it will be impossible to simultaneously reveal the small and large communities in the networks, e.g. by the RB method. This will be discussed in the following section.

4 Contradiction in multi-resolution modularity methods

It is worth noticing that when the difference in size between communities is very large, the multi-resolution methods will encounter a kind of limitation due to the existence of an intrinsic contradiction in them [49], which indicates a characteristic scale of these multi-resolution methods. That is to say, if $\gamma_1 > \gamma_2$, as shown in Figure 2b, for some values of μ, especially for $\mu = 1$ (i.e. RB method), the large communities will break up before the small communities can be found by the multi-resolution methods. In the clique-ring networks, the authors have concluded that $k_s < \sqrt{k_l}$ (k_s and k_l are the total degrees of the small and large communities) is a sufficient condition that the RB method suffers from the limitation on the networks. This can be regarded as a characteristic scale of RB in the networks [49]. The limitation is independent of the size of the whole network, but instead depends on the degree of interconnectedness of the small communities and the difference in community sizes [49]. For the other multi-resolution methods, it is not easy to give general formulas similar to that of RB due to their special self-loop assignment rules. Therefore we will study the limitation by experimental tests.

4.1 Effect of network size

We have investigated the two transition points γ_1 and γ_2 by increasing the size of the networks while preserving the difference in community size. Because the limitation is independent of the network sizes, the inequality $\gamma_1 < \gamma_2$ or $\gamma_1 > \gamma_2$ will be satisfied in all the larger networks if it is satisfied in small networks (as shown in Figs. 2c−2d). Further, we can see $\gamma_1 < \gamma_2$ for both AFG and RB in Figure 2c, while $\gamma_{b1} < \gamma_{b2}$ for AFG and $\gamma_{a1} > \gamma_{a2}$ for RB in Figure 2d. Clearly, the limitation is closely related to the multi-resolution methods.

4.2 Effect of community size difference

In general, the larger the difference in the community size, the more likely it is that the multi-resolution methods will encounter the limitation. As we see in Figure 3, the RB method frequently suffers from the limitation, while the AFG method seems robust in the tests.This can be explained by their special self-loop assignment rules, which indicates that the AFG method's tolerance against the limitation is stronger than the RB method. This can explain the perfect performance of AFG in some tests such as the toy model in [44]. But it does not mean that the AFG method never suffers from the above limitation in any network. For example, in network containing large communities as random sub-graphs (not cliques), the limitation appears much more easily than in the clique-ring networks for all the methods.

4.3 Communities as random sub-graphs

We further generate a community-ring network with $m = 6$, $n_1 = 36$ and $n_2 = 6$ by retaining the community structure of the clique-ring network while making vertices in large communities have the same degree of 9, and vertices in small communities have the degree of 6 (see Fig. 4a). In this network, small communities are still cliques (fully-connected sub-graphs), but the small cliques nearby are connected by five edges. The large communities are random sub-graphs (not cliques) with a vertex degree of 9, so random fluctuation can take effect [51]. In Figure 4b, we can see that $\gamma_1^{real} > \gamma_2^{real}$ for all values of μ. This clearly shows that all the multi-resolution methods encounter the above limitation problem – the large communities will break up in the network before the small communities become visible. Moreover, comparing γ_2^{expect} and γ_2^{real} in Figure 4b, we can find that the large communities break up earlier than expected by γ_2^{expect}. So it is certain that the limitation will appear in networks when $\gamma_2^{expect} < \gamma_1$ (or γ_1^{expect}) [49].

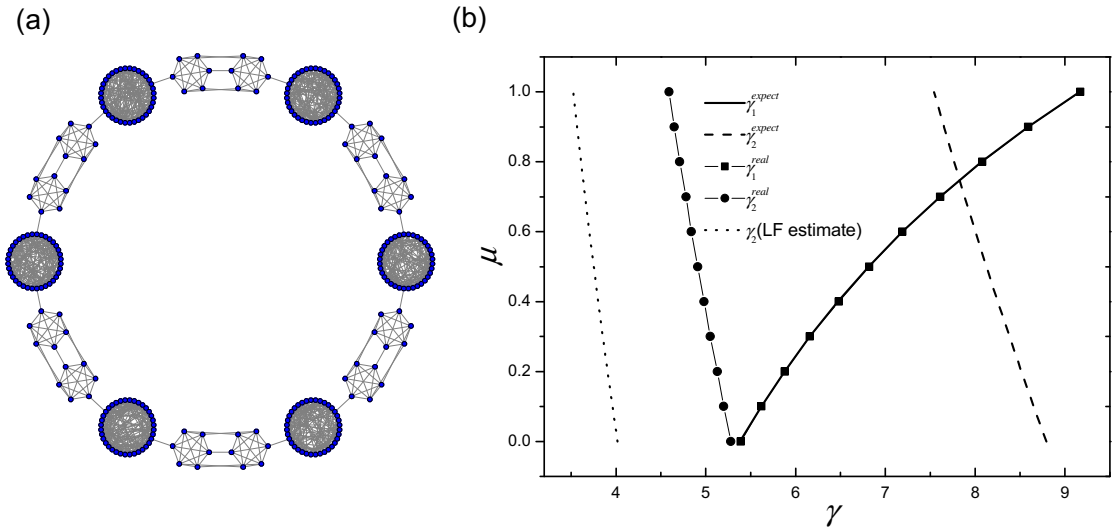

Fig. 4. (Color online) (a) Network described in the text. (b) Critical parameters γ_1 and γ_2 as a function of μ in the network described in the text. The real values for γ_1 and γ_2 are denoted by γ_1^{real} and γ_2^{real} respectively. γ_1^{expect} is the theoretical line for γ_1. γ_2^{expect} and γ_2(LF estimate) are the theoretical estimates for γ_2 by the Simple estimate and the LF estimate respectively (see Appendix).

4.4 Discussion

Concerning the above limitation problem of the multi-resolution modularity, Lancichinetti and Fortunato [50] also had an interesting discussion recently. When the difference in community size is large enough, the merger and breakup of the planted communities in the networks will appear simultaneously, which is independent of the network sizes. In this case, no suitable γ-values can be found. The limitation may exist in various multi-resolution methods based on global optimization, because of the existence of an intrinsic and irreconcilable conflict – the (large) communities as a whole prefer small resolution parameters, while the disconnecting of (small) communities requires an increase in the resolution parameter [49,50], even if they have perfect performance in some tests. After all, it is very difficult to simultaneously see the details of ants and elephants as a whole on a limited viewfinder [52] (the size of which may be related to the length of the interval from γ_1 to γ_2), though the resolution may be adjustable by resolution parameters. (It is interesting that the size of the "viewfinder" of the AFG method is larger than RB in our tests.)

In consideration of this problem, Granell et al. recently presented a new hierarchical multi-resolution method according to the AFG method [52]. The idea of the method is to split the multi-resolution method for optimal subgraphs of the network, focusing the analysis on each part independently. Moreover, the method makes use of the extension definition of modularity that can deal with negative weights better [53,54], though in the original AFG method the "resistance" parameter may take negative values. One may notice that the hierarchical method is only suitable for hierarchical networks. Indeed, the limitation mentioned above is often shown through hierarchical networks, but the limitation as well as the multi-scale community structures are not restricted to networks of this type. However, we believe that it is still interesting and will encourage further investigation of multi-resolution methods to avoid the implicit limitation.

5 Conclusion

In this paper a set of multi-resolution modularity methods using the self-loop assignment scheme (i.e. providing each vertex with a suitable self-loop) was introduced and analyzed. All the schemes in the set are equivalent in homogeneous networks (where all the vertices have the same degree), and RB & AFG are included as special cases. We hope that the study in the paper can help further understand these multi-resolution methods in community detection.

Focusing on the transition points where small communities become visible and where large communities begin to break up, we analyzed the general processes of these multi-resolution methods, both by theoretical analysis and by experimental tests on model networks. Comparing the multi-resolution methods, we show that the mean-degree-dependent self-loop assignment can quicken the disconnecting of the (small) communities with small vertex degrees, and can delay the breakup of the (large) communities with large vertex degrees. As a result, the length of the plateau from γ_1 to γ_2, which is considered to be related to the "stability" of the corresponding community division, is clearly confirmed to be closely related to these multi-resolution methods themselves, e.g. AFG can enlarge the plateau.

We also show that all the multi-resolution methods may encounter a kind of intrinsic limitation (or contradiction) regardless of the size of the whole network: when

increasing the values of the resolution parameters, large communities may break up before small communities become visible in some cases. Interestingly, the AFG method's tolerance against the limitation is clearly stronger than the RB method in our tests due to the difference in their self-loop assignment rules. However, similarly to RB, the AFG method still suffers from the above limitation because of an intrinsic and irreconcilable conflict – the (large) communities as a whole prefer small resolution parameters, while the disconnecting of (small) communities requires larger resolution parameters.

It may be possible to design new multi-resolution methods with stronger tolerances against these limitations, but similar difficulties may still be encountered due to the reasons discussed above. To solve the problem, one of the solutions may be to pin the natural communities without inner sub-communities, for example, by the theoretical or experimental evaluation of the statistical significance of the communities. As we see, the Simple estimate is only suited to fully-connected sub-graphs, while the LF estimate also prefers dense sub-graphs. In the future, better theoretical methods are expected which may be closely related to the minimal cut-size problem of classical graph partitioning [51].

This work has been supported by the National Natural Science Foundation of China (Grant Nos. 11147121), the Scientific Research Fund of Education Department of Hunan Province of China (Grant Nos. 11B128), Scientific research project of Huangshan University (Grant No. 2011xkj007) and partly by the Doctor Startup Project of Xiangtan University (Grant No. 10QDZ20).

Appendix

Here, we discuss the theoretical estimate of the parameter γ_2 at the transition point where the communities break up by using the multi-resolution modularity methods. Given a community l with n_l vertices, the total degree and its inner degree are denoted k_l and k_l^{in} respectively. The ratio between them is denoted $\alpha_l = k_l^{in}/k_l$.

Take the RB method as an example and consider the community as a random sub-graph. Now we randomly divide it into two parts A and B, where the degrees of the two parts are denoted k_A and k_B respectively ($k_l = k_A + k_B$). The community l will break up, when

$$E_{AB} - \gamma\frac{k_A k_B}{2M} < 0 \quad \text{or} \quad \gamma > \gamma_2 = \frac{2ME_{AB}}{k_A k_B} = \frac{2ME_{AB}}{(k_l/2)^2}, \tag{A.1}$$

where M is the total number of edges in the network and E_{AB} is the number of edges between the parts A and B. Clearly, evaluating the value of E_{AB} is the core of the estimate of γ_2. Since the modularity of the bi-partition in random graphs is optimal when two parts are of about equal size [50], we here consider the equally-sized bi-partition of the community, that is, $k_A \approx k_B \approx k_l/2$.

Simple estimate: most simply, we can suppose that there exists the relation $\alpha_i \approx \alpha_l$ for all vertices in the community l, where $\alpha_i = k_i^{in}/k_i$ is the ratio between the inner degree k_i^{in} and the degree k_i of vertex i in the community. Statistically, we can obtain the expected value of E_{AB},

$$E_{AB}^{expect} = \frac{\alpha_l k_A k_B}{(k_l - \bar{k})} = \frac{\alpha_l k_l^2}{4(k_l - \bar{k})}, \tag{A.2}$$

where $\bar{k}$ is the average degree of vertices in the community and k_l minus $\bar{k}$ is used to exclude the effect of self-links. Now we can rewrite (A.1) as

$$\gamma > \gamma_2^{expect} = \frac{2M\alpha_l}{k_l}\frac{n_l}{n_l - 1} \approx \frac{2M\alpha_l}{k_l}, \quad \text{when } n_l \text{ is very large.} \tag{A.3}$$

If the inequality is satisfied, then the community l will break up or have broken up. The tests in text have shown that (A.2) and (A.3) can give proper estimates of E_{AB} and γ_2 for full-connected sub-graphs (i.e. cliques).

Notice that the inequality (A.3) may be regarded as a sufficient condition that the community breaks up, but it is not a necessary one due to the simple hypothesis above, as well as random fluctuations in communities [51]. For convenience, we supposed that $\alpha_i \approx \alpha_l$ for all vertices in the community l, however it is also possible that $\alpha_i < \alpha_l$ for some parts of the community. These parts are less stable than those with $\alpha_i \geqslant \alpha_l$, so they may be separated from the community before the inequality is satisfied. This problem also exists in the LF estimate. Moreover, E_{AB}^{expect} can only give a simple estimate of E_{AB}. The number of possible equally-sized bi-partitions of the community is very large. There must exist bi-partitions whose E_{AB} is much smaller than the expected value, especially in large, random sub-graphs. This will lead to the breakup of the sub-graphs before the value of γ approaches γ_2^{expect}. In other words, the splitting of random communities will be much easier than expected by the inequality (A.3).

LF estimate: we can directly make use of the results in the graph partitioning problem [51]. As Lancichinetti and Fortunato (LF) showed [50], one can obtain the theoretical estimate of E_{AB} by considering the community l under study as a random sub-graph:

$$E_{AB}(LF\ estimate) \approx M_l\left(\frac{1}{2} - 0.765\frac{\left\langle\sqrt{k}\right\rangle_l}{\langle k\rangle_l}\right), \tag{A.4}$$

where $M_l = k_l^{in}/2 = \alpha_l k_l/2$ is the number of edges within the community l, and $\langle\ldots\rangle_l$ denotes expectation value over the ensemble of random graphs with the same degree distribution as the community. Using (A.4), for example, we can rewrite (A.1) as

$$\gamma > \gamma_2(LF\ estimate) = \frac{2M2\alpha_l}{k_l}\left(\frac{1}{2} - 0.765\frac{\left\langle\sqrt{k}\right\rangle_l}{\langle k\rangle_l}\right). \tag{A.5}$$

Clearly, if the community is a clique, then $E_{AB} \approx M_l/2$. So the estimate should be corrected: $\gamma > \gamma_2(LF\ clique - estimate) = 2M\alpha_l/k_l$, which is similar to (A.3).

According to the Simple estimate (A.2) or the LF estimate (A.4), one can obtain the theoretical estimates of γ_2 for other multi-resolution modularity methods. As shown in the text, the Simple estimate always provides an overestimation of E_{AB} (or γ_2), while the LF estimate gives an underestimation of E_{AB} (or γ_2). As shown in Figures 2–3, the Simple estimate (A.2) and the LF clique-estimate can give good results in the clique-ring networks, and these estimates can improve with increasing community size (see Fig. 3). Because the Simple estimate is just a simple and rough calculation and prefers fully-connected sub-graphs, while the LF estimate is also more valid in large and dense (sub-) graphs [50,51]. For non-full-connected communities, such as the one in Figure 4, the results of the Simple estimate (A.2) and the LF estimate (A.4) have a clear deviation from the real values of γ_2, but they still have similar qualitative trends.

References

1. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.U. Hwang, Phys. Rep. **424**, 175 (2006)
2. S. Fortunato, Phys. Rep. **486**, 75 (2010)
3. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Science **297**, 1551 (2002)
4. P. Chen, S. Redner, J. Informetr. **4**, 278 (2010)
5. R. Guimera, L.A. Nunes Amaral, Nature **433**, 895 (2005)
6. C. Castellano, F. Cecconi, V. Loreto, D. Parisi, F. Radicchi, Eur. Phys. J. B **38**, 311 (2004)
7. S.-H. Zhang, X.-M. Ning, C. Ding, X.-S. Zhang, BMC Syst. Biol. **4**, 1 (2010)
8. M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002)
9. H. Zhou, Phys. Rev. E **67**, 061901 (2003)
10. Y. Pan, D.-H. Li, J.-G. Liu, J.-Z. Liang, Physica A **389**, 2849 (2010)
11. E.A. Leicht, P. Holme, M.E.J. Newman, Phys. Rev. E **73**, 026120 (2006)
12. A. Capocci, V.D.P. Servedio, G. Caldarelli, F. Colaiori, Physica A **352**, 669 (2005)
13. L. Donetti, M.A. Munoz, J. Stat. Mech. **2004**, P10012 (2004)
14. F. Wu, B.A. Huberman, Eur. Phys. J. B **38**, 331 (2004)
15. M. Rosvall, C.T. Bergstrom, Proc. Natl. Acad. Sci. USA **105**, 1118 (2008)
16. D. Jin, B. Yang, C. Baquero, D. Liu, D. He, J. Liu, J. Stat. Mech. **2011**, P05031 (2011)
17. C. Piccardi, Plos One **6**, e27028 (2011)
18. V. Zlatić, A. Gabrielli, G. Caldarelli, Phys. Rev. E **82**, 066109 (2010)
19. J.-C. Delvenne, S.N. Yaliraki, M. Barahona, Proc. Natl. Acad. Sci. USA **107**, 12755 (2010)
20. M.J. Barber, J.W. Clark, Phys. Rev. E **80**, 026129 (2009)
21. U.N. Raghavan, R. Albert, S. Kumara, Phys. Rev. E **76**, 036106 (2007)
22. L. Šubelj, M. Bajec, Eur. Phys. J. B **85**, 1 (2012)
23. J. Reichardt, S. Bornholdt, Phys. Rev. E **74**, 016110 (2006)
24. B. Karrer, M.E.J. Newman, Phys. Rev. E **83**, 016107 (2011)
25. A. Medus, G. Acuña, C.O. Dorso, Physica A **358**, 593 (2005)
26. J. Duch, A. Arenas, Phys. Rev. E **72**, 027104 (2005)
27. G. Agarwal, D. Kempe, Eur. Phys. J. B **66**, 409 (2008)
28. S. Lehmann, L.K. Hansen, Eur. Phys. J. B **60**, 83 (2007)
29. V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, J. Stat. Mech. **2008**, P10008 (2008)
30. J.P. Bagrow, J. Stat. Mech. **2008**, P05001 (2008)
31. L. Danon, A. Díaz-Guilera, J. Duch, A. Arenas, J. Stat. Mech. **2005**, P09008 (2005)
32. M.E.J. Newman, M. Girvan, Phys. Rev. E **69**, 026113 (2004)
33. M.E.J. Newman, Phys. Rev. E **70**, 056131 (2004)
34. B.H. Good, Y.-A. de Montjoye, A. Clauset, Phys. Rev. E **81**, 046106 (2010)
35. X.S. Zhang, R.S. Wang, Y. Wang, J. Wang, Y. Qiu, L. Wang, L. Chen, Europhys. Lett. **87**, 38002 (2009)
36. S. Fortunato, M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007)
37. J. Ruan, W. Zhang, Phys. Rev. E **77**, 016104 (2008)
38. D. Lai, H. Lu, C. Nardini, Phys. Rev. E **81**, 066118 (2010)
39. J.W. Berry, B. Hendrickson, R.A. LaViolette, C.A. Phillips, Phys. Rev. E **83**, 056119 (2011)
40. M. Sales-Pardo, R. Guimerà, A.A. Moreira, L.A.N. Amaral, Proc. Natl. Acad. Sci. USA **104**, 15224 (2007)
41. M.Á. Serrano, M. Boguñá, A. Vespignani, Proc. Natl. Acad. Sci. USA **106**, 6483 (2009)
42. J. Zhang, K. Zhang, X.-K. Xu, C.K. Tse, M. Small, New J. Phys. **11**, 113003 (2009)
43. H.-J. Li, Y. Wang, L.-Y. Wu, Z.-P. Liu, L. Chen, X.-S. Zhang, Europhys. Lett. **97**, 48005 (2012)
44. A. Arenas, A. Fernández, S. Gómez, New J. Phys. **10**, 053039 (2008)
45. P. Ronhovde, Z. Nussinov, Phys. Rev. E **80**, 016109 (2009)
46. V.A. Traag, P. Van Dooren, Y. Nesterov, Phys. Rev. E **84**, 016114 (2011)
47. J.M. Kumpula, J. Saramäki, K. Kaski, J. Kertész, Eur. Phys. J. B **56**, 41 (2007)
48. R. Lambiotte, in Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), `arXiv:1004.4268v1 [physics.soc-ph]` (2010)
49. J. Xiang, K. Hu, Physica A **391**, 4995 (2012)
50. A. Lancichinetti, S. Fortunato, Phys. Rev. E **84**, 066122 (2011)
51. J. Reichardt, S. Bornholdt, Phys. Rev. E **76**, 015102 (2007)
52. C. Granell, S. Gómez, A. Arenas, Int. J. Bifur. Chaos **22**, 1250171 (2012)
53. V.A. Traag, J. Bruggeman, Phys. Rev. E **80**, 036115 (2009)
54. S. Gómez, P. Jensen, A. Arenas, Phys. Rev. E **80**, 016114 (2009)