

LIMITED RESOLUTION AND MULTIREOLUTION METHODS IN COMPLEX NETWORK COMMUNITY DETECTION

JUSSI M. KUMPULA^{1,*}, JARI SARAMÄKI¹, KIMMO KASKI¹, and JÁNOS KERTÉSZ^{1,2}

¹*Laboratory of Computational Engineering, Helsinki University of Technology,
P.O. Box 9203, FIN-02015 HUT, Finland*

²*Department of Theoretical Physics, Budapest University of Technology and Economics,
Budapest, Hungary
jkumpula@lce.hut.fi

Received 24 July 2007

Accepted 27 July 2007

Communicated by Janos Kertesz

Detecting community structure in real-world networks is a challenging problem. Recently, it has been shown that the resolution of methods based on optimizing a modularity measure or a corresponding energy is limited; communities with sizes below some threshold remain unresolved. One possibility to go around this problem is to vary the threshold by using a tuning parameter, and investigate the community structure at variable resolutions. Here, we analyze the resolution limit and multiresolution behavior for two different methods: a q -state Potts method proposed by Reichard and Bornholdt, and a recent multiresolution method by Arenas, Fernández, and Gómez. These methods are studied analytically, and applied to three test networks using simulated annealing.

Keywords: Complex networks; community detection; limited resolution.

In recent years, it has become apparent that networks are an efficient way to represent and study a large variety of technological, biological and social systems [1, 2]. Structural properties of complex networks representing such systems are closely related to their functionality. In particular, substructures called *communities* are abundant in networks. Communities are, loosely speaking, groups of nodes that are densely interconnected but only sparsely connected with the rest of the network [3, 4, 5, 6] – consider, *e.g.*, groups of individuals interacting with each other in social networks, or functional modules in metabolic networks. It is evident that reliable and computationally tractable methods for detecting these communities in empirical networks are required for studying their structural and functional roles.

Several methods and algorithms have been developed for community detection [7, 8]. One popular class of methods is based on optimizing a global quality function called modularity [9], or on a closely related Hamiltonian [10], which con-

tains the modularity as a special case. Both methods are computationally demanding, especially for large networks, but various approximative methods exist [11, 12, 9, 13, 14]. For many test networks, these methods have been shown to perform well [7, 15]. However, it has recently been shown that the resolution of the modularity based methods is intrinsically limited, that is, modularity optimization fails to find small communities in large networks – instead, small groups of connected nodes turn out merged as larger communities [16]. For the Hamiltonian-based method, there is also a resolution limit due to similar underlying reasons [17] though this method contains a tuning parameter which can be used to study communities of different sizes. Recently, Arenas *et al.* proposed a modification of the modularity optimization method which also provides a parameter that can be used to probe the community structure at different resolutions. Here, we compare these two methods and their resolutions analytically, pointing out similarities and differences. Subsequently we apply them to several test networks using optimization by simulated annealing.

We start by briefly reviewing the concept of modularity, introduced by Newman and Girvan [9]. The modularity Q is defined as follows

$$Q = \frac{2}{K} \sum_{s=1}^m (l_{ss} - [l_{ss}]), \quad (1)$$

where K is the degree sum of the network, l_{ss} is the number of links in community s , $[l_{ss}] \equiv K_s^2/2K$ is the *expected* number of links inside community s , given that the network is random, and K_s is the sum of the degrees of nodes in community s . In modularity optimization, the goal is to assign all nodes into communities such that Q is maximized.

The Hamiltonian-based method introduced by Reichardt and Bornholdt (RB) is based on considering the community indices of nodes as spins in a q -state Potts model, such that if the energy of such a system is minimized, groups of nodes with dense internal connections should end up having parallel spins [10]. The Hamiltonian for the system is defined as follows:

$$\mathcal{H} = - \sum_{s=1}^m (l_{ss} - \gamma [l_{ss}]_{p_{ij}}), \quad (2)$$

where $[l_{ss}]_{p_{ij}}$ is the expected number of links in community s , given the null model p_{ij} , and γ is a tunable parameter. Minimizing \mathcal{H} defines the community structure. When $\gamma = 1$, Eq. (2) becomes Eq. (1) apart from a constant factor. Hence the RB method contains the modularity optimization as a special case, and can be viewed in a more general framework. Changing γ allows to explore the community structure at different resolutions, but communities with large differences in size cannot be simultaneously detected using a single value of γ [17].

Recently Arenas, Fernández and Gómez (AFG) proposed a method [18] for augmenting modularity optimization with a parameter, which similarly to γ above allows tuning the resolution of the method. This approach considers the network to be weighted. The trick introduced by Arenas *et al.* [18] is to add a self-link of

weight r to each node, in which case the modularity becomes

$$Q_w(r) = \frac{1}{W(r)} \sum_{s=1}^m (w_{ss}(r) - [w_{ss}(r)]), \quad (3)$$

where $W(r)$ is total link weight in the network (including self-links), w_{ss} is total link weight inside community s and $[w_{ss}]$ is its expected value. Parameter r adjusts the total weight in the network, which in turn changes the community detection resolution [18]. Sweeping r and observing which communities are most stable with respect to changes in r should reveal the community structure.

Eqs. (2) and (3) suggest that RB and AFG methods are somewhat related, although not equal. The tuning parameters, γ and r , behave qualitatively in the same way: large parameter values allow finding small communities, and small values yield large communities. In fact, in the RB method, the effect of γ in Eq.(2) can be interpreted such that the “effective” number of links in the network equals L/γ , whereas the parameter r in Eq. (3) changes the total weight in the network. However, there is a difference: r also increases the sum of weights within a community, whereas γ has no effect on the number of links within a community. In order to illustrate the connection between these methods, we next derive the “resolution limit” intrinsic for Eq. (3) in the AFG method.

Now suppose that a network consists of “physical” communities, which are somehow known to us. We consider two of these communities, s and t , such that the sum of weights of edges connecting them is w_{st} . If these “physical” communities are merged by the detection method, the modularity $Q_w(r)$ changes by $\Delta Q_w(r) = \frac{1}{W(r)} (w_{st} - [w_{st}(r)])$. The optimization of modularity should merge these communities if $\Delta Q_w(r) > 0$, which yields

$$S_s(r)S_t(r) < 2W(r)w_{st}, \quad (4)$$

where $S_s(r)$ is the total node strength in community s . An analogous result for RB method is $\gamma K_s K_t < 2Ll_{st}$, where K_s is total node degree in community s . Hence the tuning parameters γ and r are not identical, and they affect the optimization outcome differently. However, if we assume that $S_s = S_t \approx n_s \langle s \rangle$, $n_s = n_t$ and $K_s \approx n_s \langle k \rangle$ Eq. (4) reduces to

$$n_s < \sqrt{\frac{Nw_{st}}{\langle s \rangle + r}}, \quad (5)$$

which bears resemblance to the corresponding RB result: $n_s < \sqrt{Nl_{st}/(\gamma \langle k \rangle)}$.

Next, we present some numerical results obtained by sweeping the tuning parameters γ and r of the RB and AFG methods across a range of values, and optimizing the respective energy functions using simulated annealing. Three different test networks are used. We show the behavior of the number of communities detected by the methods as a function of the tuning parameter, and look for “stable” regions where this number remains constant [18]. Earlier, community structures detected using several values of γ in the RB method have been reported in [10], but to our knowledge complete sweeps and stability analysis have not been reported earlier.

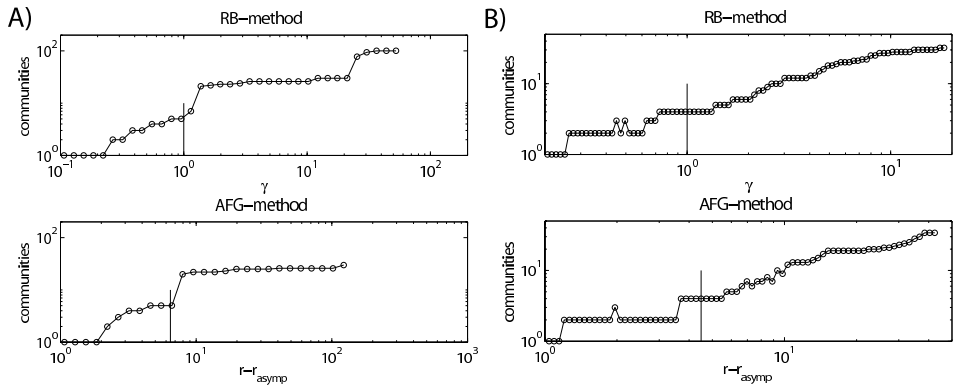


Fig. 1. Number of communities as detected with simulated annealing using the RB (upper) and AFG (lower) methods. A: hierarchical scale-free network [19] of 125 nodes, B: Zachary's karate club. The vertical line denotes the traditional modularity optimization case.

Our first test network is a synthetic, hierarchical scale-free network of $N = 125$ nodes [19]. This unweighted network can be viewed to consist of 5 communities of 25 nodes each, which can be further divided into five-node cliques (for a visualization of this network, see [19] or [18]). Figure 1(A) shows the number of communities detected using the RB and AFG methods. Both methods are able to reveal the large communities at small values of sweeping parameter, although the AFG method seems to perform slightly better. One should note that this might be a feature of the numerical optimization, and not the method itself. We remind the reader that the “traditional” modularity optimization corresponds to $\gamma = 1$ and $r = 0$. These points are shown in the figures as vertical lines. Our results for the AFG method are consistent with those reported in [18].

Our second test network is a small, unweighted network representing Zachary's karate club [20], which has often been used as a “testbed” for community detection. Modularity optimization is known to yield four communities, whereas this club was observed to split into two communities. In [18], the authors demonstrated that AFG method is able to find exactly those communities (by using the weighted version of this network). Results for the unweighted network in Fig. 1(B) show that both methods give similar results and are able to detect the two communities. A closer inspection shows that the communities correspond to the split which eventually happened (except for one individual classified differently by the RB method).

Our third test network is weighted, being larger than the previous examples (986 nodes), and has a more complex community structure, Fig. 2(a). The average degree of this network is $\langle k \rangle = 6$ and it has been generated with a model designed to resemble real, weighted social networks. Visually, the communities are less apparent than in the previous test networks, although it can be seen that there are dense groups of nodes with strong internal links, connected by weaker links. Applying the clique percolation method [6, 21, 22] to this network using clique size = 4 yields communities whose sizes vary from 4 nodes (20 communities) to 43 nodes (1 community). Because the network is weighted, we have used a weighted Hamiltonian instead of (2) for the RB method [17]. Results in Fig. 2(b) show that no clear “stable” regions of the tuning parameters with a constant number of communities

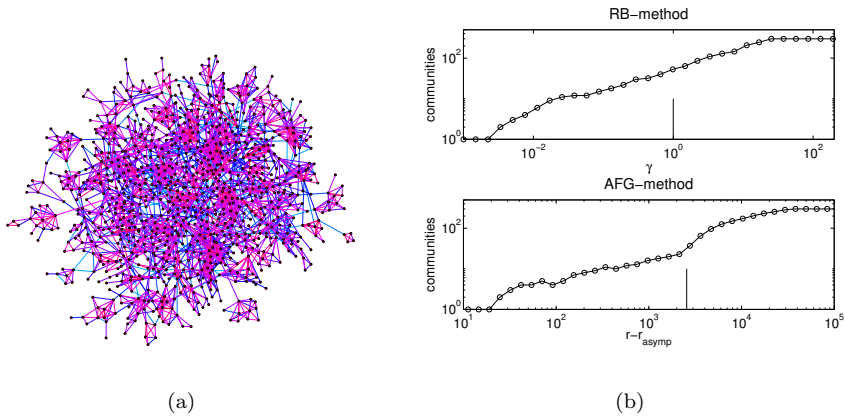


Fig. 2. (Color online) A weighted test network having 986 nodes. Link colors vary from blue (weak) to red (strong), Number of communities for the network as a function of the tuning parameters. Note that we have limited the number of communities to 300.

are apparent. One possible explanation is that this is due to quite non-uniform distribution of community sizes, which may result in large communities continuously being split into smaller ones as the tuning parameters are increased. A similar situation could occur for many large real-world networks. However, by using small values of γ and r it might be possible to study the large-scale community structure, such that the network is split into a small number of large communities.

We have discussed the limited resolution of community detection methods where a global energy-like quantity is optimized, by focusing especially on two methods (RB and AFG) where the resolution can be adjusted using a tuning parameter. Although the tuning parameters of these two methods give rise to qualitatively similar changes in resolution, analytic derivations show that their effect on the resolution limit is somewhat different. These two methods have also been numerically tested by using simulated annealing, with the result that in small test networks, stable regions of tuning parameter values, where the number of communities is constant, can easily be found. These can be viewed to reflect “optimal” communities. However, on a large, weighted test network, where the clique percolation method indicates a broader distribution of community sizes, no such regions are apparent.

Acknowledgments

This work was partially supported the Academy of Finland (Center of Excellence program 2006-2011). JS acknowledges support by the European Commission NEST Pathfinder initiative on Complexity through project EDEN (Contract 043251).

References

- [1] R. Albert and A. L Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**(1) (2002) 47.
- [2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. U. Hwang, Complex networks: Structure and dynamics, *Phys. Rep.* **424**(4-5) (2006) 175–308.

- [3] R. Guimera, S. Mossa, A. Turttschi and L. A. N. Amaral, From the cover: The world-wide air transportation network: Anomalous centrality, community structure, and cities' global roles, *PNAS* **102**(22) (2005) 7794–7799.
- [4] A. Arenas, L. Danon, A. Diazaz-Guilera, P. M. Gleiser and R. Guimera, Community analysis in social networks, *Eur. Phys. J. B* **38** (2004).
- [5] R. Guimera and L. A. N. Amaral, Functional cartography of complex metabolic networks, *Nature* **433** (2005) 895–900.
- [6] Gergely Palla, Imre Derenyi, Illes Farkas and Tamas Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, *Nature* **435**(7043) (2005) 814–818.
- [7] L. Danon, A. Diazaz-Guilera, J. Duch and A. Arenas, Comparing community structure identification, *Journal of Statistical Mechanics: Theory and Experiment* **9** (2005).
- [8] M. E. J. Newman, Detecting community structure in networks, *The European Physical Journal B – Condensed Matter and Complex Systems* **38**(2) (2004) 321–330.
- [9] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E* **69**(2) (2004) 026113.
- [10] J. Reichardt and S. Bornholdt, Statistical mechanics of community detection, *Phys. Rev. E* **74**(1) (2006) 016110.
- [11] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, *Phys. Rev. E* **74** (2006) 036104.
- [12] A. Clauset, M. E. J. Newman and C. Moore, Finding community structure in very large networks, *Phys. Rev. E* **70**(6) (2004) 66111.
- [13] J. Duch and A. Arenas, Community detection in complex networks using extremal optimization, *Phys. Rev. E* **72**(2) (2005) 027104.
- [14] M. E. J. Newman, Fast algorithm for detecting community structure in networks, *Phys. Rev. E* **69**(6) (2004) 066133.
- [15] M. Gustafsson, M. Hornquist and A. Lombardi, Comparison and validation of community structures in complex networks, *Physica A* **367**(7/15) (2006) 559–576.
- [16] S. Fortunato and M. Barthélemy, Resolution limit in community detection, *PNAS* **104**(1) (2007) 36–41.
- [17] J. M. Kumpula, J. Saramäki, K. Kaski and J. Kertesz, Limited resolution in complex network community detection with potts model approach, *Eur. Phys. J. B* **56** (2007) 41–45.
- [18] A. Arenas, A. Fernandez and S. Gomez, Multiple resolution of the modular structure of complex networks, arXiv:physics/0703218v1 (2007).
- [19] E. Ravasz and A. L. Barabasi, Hierarchical organization in complex networks, *Phys. Rev. E* **67**(2) (2003) 26112.
- [20] W. W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* **33**(4) (1977) 452–473.
- [21] I. Derényi, G. Palla and T. Vicsek, Clique percolation in random networks, *Phys. Rev. Lett.* **94**(16) (2005) 160202.
- [22] G. Palla, A.-L. Barabasi and T. Vicsek, Quantifying social group evolution, *Nature* **446**(713) (2007) 664–667.