HIERARCHICAL MULTIRESOLUTION METHOD TO OVERCOME THE RESOLUTION LIMIT IN COMPLEX NETWORKS

CLARA GRANELL[*], SERGIO GÓMEZ[†] and ALEX ARENAS[‡]

Department d'Enginyeria Informàtica i Matemàtiques,
Universitat Rovira i Virgili, Av. Països Catalans 26,
43007 Tarragona, Catalonia, Spain
[*]clara.granell@urv.cat
[†]sergio.gomez@urv.cat
[‡]alexandre.arenas@urv.cat

The analysis of the modular structure of networks is a major challenge in complex networks theory. The validity of the modular structure obtained is essential to confront the problem of the topology-functionality relationship. Recently, several authors have worked on the limit of resolution that different community detection algorithms have, making impossible the detection of natural modules when very different topological scales coexist in the network. Existing multiresolution methods are not the panacea for solving the problem in extreme situations, and also fail. Here, we present a new hierarchical multiresolution scheme that works even when the network decomposition is very close to the resolution limit. The idea is to split the multiresolution method for optimal subgraphs of the network, focusing the analysis on each part independently. We also propose a new algorithm to speed up the computational cost of screening the mesoscale looking for the resolution parameter that best splits every subgraph. The hierarchical algorithm is able to solve a difficult benchmark proposed in [Lancichinetti & Fortunato, 2011], encouraging the further analysis of hierarchical methods based on the modularity quality function.

Keywords: Complex networks; community structure; multiple resolution; modularity.

1. Introduction

The quality function called *modularity* has been largely used in the assessment of the modular structure of networks [Girvan & Newman, 2002; Newman & Girvan, 2004; Newman, 2004a; Clauset *et al.*, 2004; Duch & Arenas, 2005; Danon *et al.*, 2005] and for data clustering and exploration [Newman, 2006; Granell *et al.*, 2011]. Modularity is a global descriptor of a complex network that measures the difference between a given partition of the network and the same partition in an ensemble of the randomized versions of the original network preserving the local strength of every node. The optimization of modularity is coherently related to the definition of modules in the network; a module is defined as the result of the optimal modularity partition. Fortunato and Barthélemy [2007] pointed out a drawback in this function consisting of a certain resolution limit (generalized later in [Kumpula *et al.*, 2007; Good *et al.*, 2010]), beyond which optimization of modularity is unable to identify certain modules, even those easily detectable at first sight, such as cliques almost disconnected from the rest of the network. This effect is known as the resolution limit of modularity. This problem arises because modularity fixes a global scale that could be appropriate for some networks but not for others, specially not suitable for those networks conformed

to coexisting densely large and small communities. After this work, a multiresolution method was introduced in [Arenas *et al.*, 2008], which preserved the use of modularity with the addition of a parameter to control the resistance of nodes to form communities. The idea is that the analysis of communities may be performed at different scales of description, and the resolution limit is overcome just by moving to the right scale. Other methods to overcome the resolution limit are found in [Reichardt & Bornholdt, 2004; Pons & Latapy, 2011; Traag *et al.*, 2011; Berry *et al.*, 2011; Ronhovde & Nussinov, 2009, 2010].

A recent work by [Lancichinetti & Fortunato, 2011] shows that even those methods devoted to avoid the resolution limit, indeed have a resolution limit, and propose the use of an algorithm composed of several approaches called OSLOM [Lancichinetti *et al.*, 2011] to really avoid such resolution problem. The proof that multiresolution schemes still have a resolution limit is performed analytically on the RB (after Reichardt–Bornholdt) method, and extended qualitatively using examples to the AFG (after Arenas–Fernández–Gómez) method and the recent CPM (Constant Potts Model) method.

We have performed extensive simulations using the AFG method and conclude that the authors of [Lancichinetti & Fortunato, 2011] are right, the AFG method also has a resolution limit, and that the benchmark they propose (see Fig. 1), consisting of a giant Erdös–Rényi (ER) network and two small cliques, connected between them by just one link, is impossible to separate in the configuration of one cluster for the giant ER network and one cluster for each of the cliques, in the current proposal of the AFG method. Even though the synthetic benchmarks where multiresolution methods could fail are far from the structure of real networks, it is still challenging to investigate what are the problems and how to solve them.

In this paper, we focus on the AFG method, analyzing its performance in resolution limiting situations, and proposing alternatives to eliminate or, if not possible, diminish the effect of this limit to minimum. An alternative is presented, a hierarchical application of the resolution screening, that avoids the resolution limit. The hierarchical application of a multiresolution method focuses the screening on different clusters of the network as soon as these clusters are detected.

2. Multiresolution AFG Method

In a previous work, the authors proposed a method that allows the full screening of the topological structure at different resolution levels using the original formulation and semantics of modularity as defined in [Girvan & Newman, 2002]. The original modularity allows the comparison of different partitions of the network. Given a network partitioned into communities, C_i being the community to which node i is assigned, the mathematical definition of modularity is

$$Q[w_{ij}, C] = \frac{1}{2w} \sum_i \sum_j \left(w_{ij} - \frac{w_i w_j}{2w} \right) \delta(C_i, C_j),$$

(1)

where w_{ij} is the weight of the link between nodes i and j (zero if no link exists), $w_i = \sum_j w_{ij}$ is the strength of node i and $2w = \sum_i w_i$ is the total strength of the network [Newman, 2004a]. The Kronecker delta function $\delta(C_i, C_j)$ takes the value 1 if nodes i and j are from the same community and 0 otherwise. Several authors have attacked the problem of modularity optimization, with considerable success, by proposing different heuristics [Newman, 2004b; Clauset *et al.*, 2004; Guimerà & Amaral,
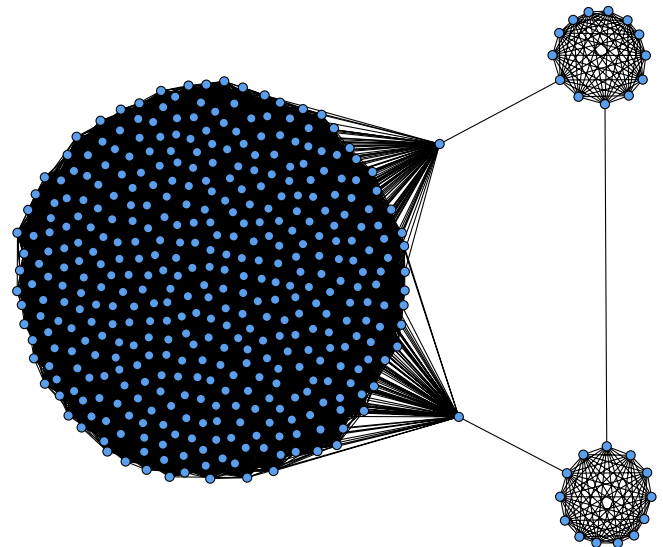

Fig. 1. Benchmark proposed by [Lancichinetti & Fortunato, 2011] to test the resolution limit of multiresolution methods. The large component is a ER network of 400 nodes with $k = 100$ linked to two cliques of 13 nodes each, sharing only one link between them. The goal is to separate these three subgraphs using a community detection algorithm aimed to detect multiple resolutions.

2005; Duch & Arenas, 2005; Pujol *et al.*, 2006; Newman, 2006], see [Fortunato, 2010] for a review.

The AFG method was designed to evaluate the community structure of networks using a kind of magnifying glass of the topology [Arenas *et al.*, 2008]. The mathematical form of this prescription is given by

$$Q_{\mathrm{AFG}}[w_{ij}, C, r] = Q[w_{ij} + r\delta_{ij}, C], \qquad (2)$$

where the resistance r is the parameter controlling the resolution of the partitions we want to find, and $w_{ij} + r\delta_{ij}$ is the new weights matrix after the addition of a self-loop with value r to each node. When r is zero, we recover the standard modularity Q. The definition of Q_{AFG} preserves the original semantics of modularity.

A refinement of the AFG method may be found in [Granell *et al.*, 2011, 2012], where the original formulation of modularity Eq. (1) is replaced by its extension to networks with positive and negative weights [Gómez *et al.*, 2009; Traag & Bruggeman, 2009]. Although the differences are usually small, this is necessary since the access to the macroscale needs the use of negative values of the resistance, even if the original network has only positive weights. Thus, the adequate formulation of modularity Eq. (1) for undirected weighted signed networks which should be used is

$$Q[w_{ij}, C] = \frac{1}{2w^+ + 2w^-}$$
$$\times \sum_i \sum_j \left[w_{ij} - \left(\frac{w_i^+ w_j^+}{2w^+} - \frac{w_i^- w_j^-}{2w^-} \right) \right]$$
$$\times \delta(C_i, C_j), \qquad (3)$$

where

$$w_i^+ = \sum_{j, w_{ij} > 0} w_{ij}, \qquad (4)$$

$$w_i^- = \sum_{j, w_{ij} < 0} |w_{ij}|, \qquad (5)$$

are the positive and negative strengths of node i, and

$$2w^+ = \sum_i w_i^+, \qquad (6)$$

$$2w^- = \sum_i w_i^-, \qquad (7)$$

are the positive and negative total strengths, respectively. Please note that these four strengths are defined to be non-negative. The extension to directed networks [Arenas *et al.*, 2007] is simply obtained by the substitutions in Eq. (3)

$$w_i^\pm \to w_i^{\pm,\mathrm{out}} = \sum_k w_{ik}^\pm, \qquad (8)$$

$$w_j^\pm \to w_j^{\pm,\mathrm{in}} = \sum_k w_{kj}^\pm. \qquad (9)$$

For the sake of simplicity, we will refer to the undirected case for the rest of the paper. In the particular case that the original network does not have negative weights, $r < 0$ and $w_{ii} = 0$, $\forall i$, Eq. (2) reads

$$Q_{\mathrm{AFG}}[w_{ij}, C, r] = \frac{1}{2w + N|r|} \sum_i \sum_j \left[w_{ij} + r\delta_{ij} \right.$$
$$\left. - \left(\frac{w_i w_j}{2w} - \frac{r^2}{N|r|} \right) \right] \delta(C_i, C_j) \qquad (10)$$

$$= \frac{2w}{2w - Nr} Q[w_{ij}, C]$$
$$+ \frac{r}{2w - Nr} \left(N - \frac{1}{N} \sum_{s \in C} n_s^2 \right), \qquad (11)$$

where N is the total number of nodes, n_s is the number of nodes in community s of the partition C, and the nodes and total strengths refer to the original network before the addition of the self-loops. It is interesting to realize that, since all the negative strengths are equal to the absolute value of r, the contribution of the resistance to modularity is equivalent to a constant Potts model [Traag *et al.*, 2011].

Resolving the substructure of networks using a unique parameter as proposed in the AFG has still a resolution problem. As pointed out by [Lancichinetti & Fortunato, 2011], when very different sized modules coexist, multiresolution methods will tend to break the larger groups before finding the smaller ones. The phenomenon is easy to understand with an example: let us imagine an image with a real size elephant and an ant, to see the details of the ant we have to get so close to the image that the elephant image disintegrates in smaller parts, and only part of the elephant is seen when focusing
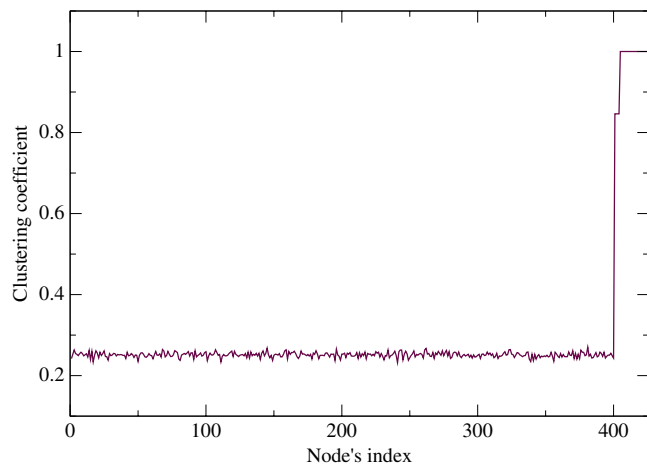
Fig. 2. Clustering coefficient for the benchmark of Fig. 1. Note the sharp transition in the relative local density of links represented by the clustering coefficient. This is indeed a hint to the coexistence of very different topological scales in the network.

on the ant. In terms of modularity, we are trying to unravel those areas which are denser in terms of links with respect to other areas in the network. A way to determine if we could have resolution problems is to plot a link density map and detect if there are sharp contrasts. If very different topological scales coexist, there will also be jumps in the clustering coefficient. In the example provided by [Lancichinetti & Fortunato, 2011], which consists of an ER network of 400 nodes with average degree 100 linked to two cliques of 13 nodes only by a unique link between them (see Fig. 1), the clustering coefficient presents a drastic separation of scales, see Fig. 2. This indicates small zones of the network very densely connected and a wide area not so dense, corresponding to the cliques and the ER, respectively.

3. Hierarchical Multiresolution Method

Our approach to solve the resolution problem takes advantage of the capability of the AFG method to find meaningful communities from the initial steps of the mesoscale analysis. More precisely, we propose the use of an iterative scheme which combines the optimization of modularity close to the macroscale of the network with its splitting in subgraphs, one for each of the previously found communities.

Supposing that our network is undirected, weighted, with positive weights and no self-loops,

the prescription of our algorithm is the following:

(1) Start out from the macroscale partition $\mathcal{M}$, which has only one community containing all nodes. Then, find the upper bound of this macroscale, which is the minimum value of the resistance parameter ($r_{\min}$) needed to find a partition $\mathcal{C}$ of the network with optimal modularity $Q_{\text{AFG}}[w_{ij}, \mathcal{C}, r_{\min}]$ formed by more than one community.
(2) Split the network in the subgraphs defined by the partition $\mathcal{C}$ just found.
(3) Repeat the previous steps with each subgraph until no further subdivisions are needed.

This algorithm defines a hierarchical organization of the nodes, where the values of $r_{\min}$ at each splitting define the ultrametric distances between nodes, i.e. the heights in the dendrogram at which every pair of nodes first meet.

The calculations of $r_{\min}$ and $\mathcal{C}$ may be performed simultaneously, therefore avoiding the costly scanning of the whole mesoscale between the lower and upper bounds of the resistance [Granell *et al.*, 2011, 2012]. This is a consequence of the following properties:

- The value of $r_{\min}$ is negative, with the only exception in which the network is just a clique.
- $Q_{\text{AFG}}[w_{ij}, \mathcal{M}, r] = 0, \forall r < 0$, because:

$$Q_{\text{AFG}}[w_{ij}, \mathcal{M}, r] = \frac{1}{2w + N|r|} \sum_i \sum_j \left[w_{ij} + r\delta_{ij} \right.$$

$$\left. - \left(\frac{w_i w_j}{2w} - \frac{r^2}{N|r|} \right) \right]$$

$$= \frac{1}{2w + N|r|} \left[2w + Nr \right.$$

$$\left. - \left(\frac{(2w)^2}{2w} + \frac{N^2 r^2}{Nr} \right) \right]$$

$$= 0. \tag{12}$$

In fact, modularity Eq. (3) is always zero for $\mathcal{M}$, no matter the network or the value of the self-loops.

- Since $Q_{\text{AFG}}[w_{ij}, \mathcal{M}, r] = 0$ and modularity is a continuous and monotonically increasing function of the resistance for any given $C \neq \mathcal{M}$, the optimal partition $\mathcal{C}$ at $r_{\min}$ must satisfy $Q_{\text{AFG}}[w_{ij}, \mathcal{C}, r_{\min}] = 0$.
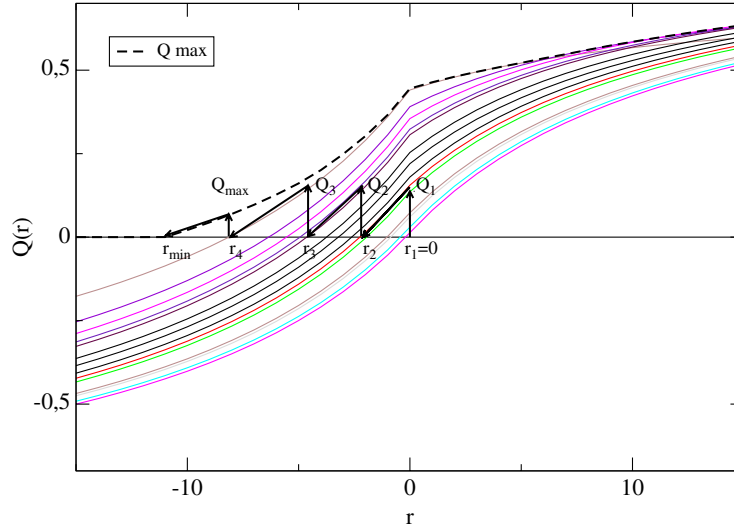
Fig. 3. Example of evolution of the FTR algorithm finding $r_{\min}$ in four iterations of the scheme. We start at $r_1 = 0$, optimizing modularity we find Q_1, we look for the $r_{\min}$ corresponding to the partition found at Q_1 using Eq. (13) and label it r_2, the process follows with $Q_2 \to r_3 \to Q_3 \to r_4 \to Q_{\max}$ up to finding $r_{\min}$, beyond this value the only partition we will find corresponds to the whole network as a unique module. Different curves in color are values of $Q[w_{ij}, \mathcal{C}, r]$ for different partitions.

- For any given partition C, the minimum meaningful value of the resistance $r_{\min}(C)$ is the one for which $Q_{\mathrm{AFG}}[w_{ij}, C, r_{\min}(C)] = 0$. Thus, Eq. (11) leads to

$$r_{\min}(C) = \frac{-2w}{N - \frac{1}{N}\sum_{s \in C} n_s^2} Q[w_{ij}, C]. \quad (13)$$

- The upper bound of the macroscale is given by

$$r_{\min} = \min_C \{r_{\min}(C)\} \quad (14)$$

and $\mathcal{C}$ is the partition which minimizes $r_{\min}(C)$.

All these properties may be combined in the following *fast-tracking resistance* (FTR) algorithm to find the upper bound of the macroscale:

(1) Optimize modularity at $r = 0$, to obtain partition C_{prev}.
(2) Calculate $r_{\min}(C_{\mathrm{prev}})$ using Eq. (13).
(3) Optimize modularity at $r := r_{\min}(C_{\mathrm{prev}})$, to obtain the current partition C_{curr}.
(4) If $C_{\mathrm{curr}} = C_{\mathrm{prev}}$ or $C_{\mathrm{curr}} = \mathcal{M}$, then $r_{\min} := r_{\min}(C_{\mathrm{prev}})$ and $\mathcal{C} := C_{\mathrm{prev}}$.
(5) Otherwise, let $C_{\mathrm{prev}} := C_{\mathrm{curr}}$ and go back to the second step.

In practice, this algorithm converges in a few number of steps. It stops when a value of r is found such that the optimization of modularity does not produce any new partition. In this case, the modularity

of both C_{prev} and $\mathcal{M}$ is zero, and no known partition can be used to obtain a better upper bound of the macroscale. Of course, we cannot claim that we have found the "real" $r_{\min}$, since no optimization heuristic can ensure the finding of the global maximum of modularity (this problem is known to be a NP-hard problem, see [Brandes *et al.*, 2008]), but this is the best approximation one may obtain. To exemplify the functioning of the FTR algorithm we show in Fig. 3 its application to the first hierarchical splitting of Zachary karate club network [Zachary, 1977].

4. Results and Discussion

We have applied the hierarchical multiresolution method explained before to the benchmark proposed by [Lancichinetti & Fortunato, 2011] shown in Fig. 1. We use the FTR algorithm to speed up the process of finding the minimal r at which every subgraphs splits. The aim is to find the partition divided in three communities in which the giant ER and each clique are separated. These three communities should contain the nodes labeled 1 to 400, 401 to 413 and 414 to 426, respectively. As stated in the method, we have started out from the macroscale $\mathcal{M}$ of the network, which contains the 426 nodes. The optimal partition splits into two communities at a value of the resistance parameter -12.5, obtaining a community formed by the nodes from

1 to 400 and another community containing the 26 nodes corresponding to the two cliques. Performing the hierarchical method on the two communities obtained, we find that the community containing the 26 nodes rapidly splits into two communities of 13 nodes, at a value of the resistance equal to -11.69. The partition containing 400 nodes splits in two at a much greater value of the resistance parameter, which is -8.97. After that, a hierarchical multiresolution is applied to any community found, until no further divisions are needed. The results of this example are shown in a dendrogram representation in Fig. 4.

Observing this figure, we find that there is a region of the resistance parameter in which the three communities we were hoping to find coexist. This happens because the two cliques form their own communities much before the community of 400 nodes is split in two. Note that this result cannot be obtained using the original multiresolution AFG method exploring the whole mesoscale,

because of the resolution limit emerging from the coexistence of very different topological scales. The rationale behind the success of the hierarchical method in this situation is the following: the separation of the network in optimal subgraphs, each one split and independently analyzed through the multiresolution scheme, reduces the global resolution limit. This resolution limit depends on the number of nodes and the number of links in the whole structure. The multiresolution method is able to focus the attention on lower scales while other parts of the network are being screened independently at larger resolution values of r.

5. Conclusions

We have presented a hierarchical multiresolution method able to cope with networks where the resolution limit would make other schemes to fail, finding the natural communities as defined by [Fortunato & Barthélemy, 2007]. The method is boosted by a mechanism that allows the determination of the resolution parameter r at which to optimize modularity in a few steps. The results solving the difficult separation of the benchmark proposed in [Lancichinetti & Fortunato, 2011] are encouraging and open the door for further investigation of modularity based community detection methods to escape from the implicit resolution limit.

Acknowledgments

References

Arenas, A., Duch, J., Fernández, A. & Gómez, S. [2007] "Size reduction of complex networks preserving modularity," *New J. Phys.* **9**, 176.

Arenas, A., Fernández, A. & Gómez, S. [2008] "Analysis of the structure of complex networks at different resolution levels," *New J. Phys.* **10**, 053039.

Berry, J. W., Hendrickson, B., LaViolette, R. A. & Phillips, C. A. [2011] "Tolerating the community detection resolution limit with edge weighting," *Phys. Rev. E* **83**, 056119.

Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z. & Wagner, D. [2008] "On modularity clustering," *IEEE Trans. Knowl. Data Engn.* **20**, 172.
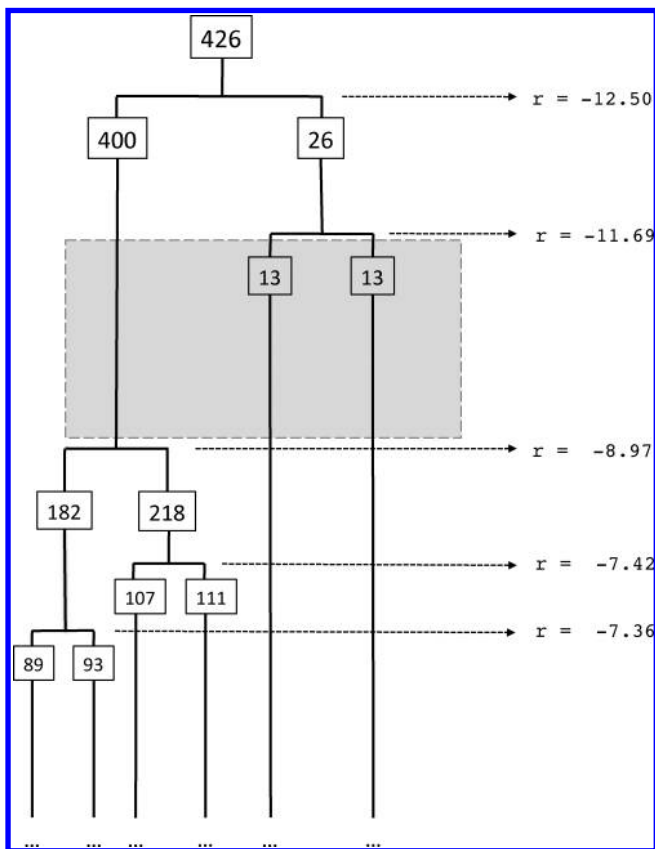
Fig. 4. Dendrogram resulting from the application of the hierarchical multiresolution method on the benchmark of Fig. 1. The gray region shows the range of the resistance parameter in which the three communities searched coexist. Note that the vertical lines are not scaled.

Clauset, A., Newman, M. E. J. & Moore, C. [2004] "Finding community structure in very large networks," *Phys. Rev. E* **70**, 066111.

Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. [2005] "Comparing community structure identification," *J. Stat. Mech.* P09008.

Duch, J. & Arenas, A. [2005] "Community identification using extremal optimization," *Phys. Rev. E* **72**, 027104.

Fortunato, S. & Barthélemy, M. [2007] "Resolution limit in community detection," *Proc. Natl. Acad. Sci. USA* **104**, 36.

Fortunato, S. [2010] "Community detection in graphs," *Phys. Rep.* **486**, 75.

Girvan, M. & Newman, M. E. J. [2002] "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. USA* **99**, 7821.

Gómez, S., Jensen, P. & Arenas, A. [2009] "Analysis of community structure in networks of correlated data," *Phys. Rev. E* **80**, 016114.

Good, B. H., de Montjoye, Y.-A. & Clauset, A. [2010] "Performance of modularity maximization in practical contexts," *Phys. Rev. E* **81**, 046106.

Granell, C., Gómez, S. & Arenas, A. [2011] "Mesoscopic analysis of networks: Applications to exploratory analysis and data clustering," *Chaos* **21**, 016102.

Granell, C., Gómez, S. & Arenas, A. [2012] "Unsupervised clustering analysis: A multiscale complex networks approach," *Int. J. Bifurcation and Chaos* **22**, 1230023-1–10.

Guimerà, R. & Amaral, L. A. N. [2005] "Cartography of complex networks: Modules and universal roles," *J. Stat. Mech.* P02001.

Kumpula, J. M., Saramaki, J., Kaski, K. & Kertesz, J. [2007] "Limited resolution and multiresolution methods in complex network community detection," *Fluct. Noise Lett.* **7**, 209.

Lancichinetti, A. & Fortunato, S. [2011] "Limits of modularity maximization in community detection," ArXiv e-prints, arXiv:1107.1155.

Lancichinetti, A., Radicchi, F., Ramasco, J. J. & Fortunato, S. [2011] "Finding statistically significant communities in networks," *PLoS ONE* **6**, e18961.

Newman, M. E. J. & Girvan, M. [2004] "Finding and evaluating community structure in networks," *Phys. Rev. E* **69**, 026113.

Newman, M. E. J. [2004a] "Analysis of weighted networks," *Phys. Rev. E* **70**, 056131.

Newman, M. E. J. [2004b] "Fast algorithm for detecting community structure in networks," *Phys. Rev. E* **69**, 066133.

Newman, M. E. J. [2006] "Modularity and community structure in networks," *Proc. Natl. Acad. Sci. USA* **103**, 8577.

Pons, P. & Latapy, M. [2011] "Post-processing hierarchical community structures: Quality improvements and multi-scale view," *Theoret. Comput. Sci.* **412**, 892.

Pujol, J. M., Béjar, J. & Delgado, J. [2006] "Clustering algorithm for determining community structure in large networks," *Phys. Rev. E* **74**, 016107.

Reichardt, J. & Bornholdt, S. [2004] "Detecting fuzzy community structures in complex networks with a potts model," *Phys. Rev. Lett.* **93**, 218701.

Ronhovde, P. & Nussinov, Z. [2009] "Multiresolution community detection for megascale networks by information-based replica correlations," *Phys. Rev. E* **80**, 016109.

Ronhovde, P. & Nussinov, Z. [2010] "Local resolution-limit-free potts model for community detection," *Phys. Rev. E* **81**, 046114.

Traag, V. A. & Bruggeman, J. [2009] "Community detection in networks with positive and negative links," *Phys. Rev. E* **80**, 036115.

Traag, V. A., Dooren, P. V. & Nesterov, Y. [2011] "Narrow scope for resolution-limit-free community detection," *Phys. Rev. E* **84**, 016114.

Zachary, W. W. [1977] "An information flow model for conflict and fission in small groups," *J. Anthropol. Res.* **33**, 452–473.