Motif-based communities in complex networks

A Arenas[1,2]‡, A Fernández[1], S Fortunato[3] and S Gómez[1]

[1] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Avinguda dels Països Catalans 26, 43007 Tarragona, Spain
[2] Institute for Biocomputation and Physics of Complex Systems (BIFI), Universidad de Zaragoza, Corona de Aragón 42, Edificio Cervantes, 50009 Zaragoza, Spain
[3] Complex Networks Lagrange Laboratory (CNLL), Institute for Scientific Interchange (ISI), Viale S. Severo 65, 10133 Torino, Italy

E-mail: `alexandre.arenas@urv.cat`, `alberto.fernandez@urv.cat`, `fortunato@isi.it` and `sergio.gomez@urv.cat`

Abstract. Community definitions usually focus on edges, inside and between the communities. However, the high density of edges within a community determines correlations between nodes going beyond nearest-neighbours, and which are indicated by the presence of motifs. We show how motifs can be used to define general classes of nodes, including communities, by extending the mathematical expression of Newman-Girvan modularity. We construct then a general framework and apply it to some synthetic and real networks.

PACS number: 89.75

‡ Author to whom any correspondence should be addressed

1. Introduction

Modular structure in complex networks has become a challenging subject of study starting with its very definition [1]. One of the most successful approaches has been the introduction of the quality function called *modularity* [2, 3], that accomplishes two goals: (i) it implicitly defines modules, and (ii) it provides with a quantitative measure to find them. It is based on the intuitive idea that random networks are not expected to exhibit modular structure (communities) beyond fluctuations.

A lot of work has been done to devise reliable techniques to maximize modularity [4, 5, 6, 7, 8, 9]. However, very little has been done to analyze the concept of modularity itself and its reliability as a method for community detection. To a large extent, the success of modularity as a quality function to analyze the modular structure of complex networks, relies on its intrinsic simplicity. The researcher interested in this analysis is endowed with a non-parametric function to be optimized: modularity. The result of the analysis will provide a partition of the network into communities such that the number of edges within each community is larger than the number of edges one would expect to find by random chance. As a consequence, each community is a subset of nodes more connected between them than with the rest of the nodes in the network. Recently, it has been shown that modularity is not the panacea of the community detection problem; in particular it suffers from a resolution limit that avoids grasping the modular structure of networks at low scales [10]. Moreover, modularity is strongly focused on communities, so it cannot be used in general to detect groups of nodes revealed by alternative connectivity patterns. The only exception is represented by "anti-communities", i.e. groups of nodes with a few edges inside and many edges connecting different groups. The presence of anti-communities indicates that the network has a multipartite structure. Anti-communities could be detected by modularity minimization [11], although the results are not so good, as we mention in section 3.

In general, detecting multipartite structure from first principles requires a definition of the classes that is quite different (in fact, opposite) with respect to standard community definitions. Let us consider bipartite networks, where nodes/actors are connected through other entities, for example collaboration in a work, attendance to an event, etc. In these specific cases, nodes of the same class (e.g. actors) are not directly linked, or share but a few edges, and usually some projection of the network in a subnetwork of only a class of nodes is needed for subsequent analysis. For example, in a projection into the actors space, two actors could be connected if they share a team, and the weight of this link could be either one (unweighted projection) or the number of shared teams (weighted projection). However any projection implies knowledge about the different classes of nodes. The definition of community must be generalized to deal with these cases. Doing it within a modularity-based framework requires a different formulation of modularity [12, 13].

We remark that bipartite networks are characterized by the fact that any path with even length starting from a node of either class ends in the same class, due to the

absence of internal edges in each class. So, if the two classes are A and B and we start from a node i_A of class A, the first step leads to one of its neighbours, say i_B, which is in B, the next step to a neighbour j_A of i_B, which is in A, and so on. In this way, paths of even length starting and ending in the same class may reveal bipartite structure, if there are many of them. On the other hand, in a graph with modular structure, there are many edges inside each module, so one expects accordingly a large number of paths between the nodes. In particular, one expects a large number of *cycles*, i.e. closed paths.

We deduce that short paths, or motifs, of a network, could be used to define and identify both communities and more general topological classes of nodes. Here we propose a general framework to classify nodes based on motifs. Classes will be defined based on the principle that they "contain" more motifs than a null model representing a randomized version of the network at study. We adopt the null model of modularity, i.e. a random network with the same degree/strength sequence of the original network, because modularity lends itself to a simple generalization, which makes calculations straightforward. We shall derive different extensions of modularity, where the building blocks will be the motifs and not just the edges, as in the original expression. After that, we shall maximize the new functions to detect the classes.

We stress that we use a modularity-based framework only as an illustrative example of how motifs could be defined to detect general node classes in networks, but in general our framework can be useful to any other method designed to detect substructure in networks. Note that the extended quality functions, that we shall introduce, also obey the principle of the resolution limit, which states that modularity will not be able to resolve substructures beyond a certain size limit, just like the original modularity [10]. However this limit is now motif-dependent and then several resolution of substructures can be achieved by changing the motif.

The rest of the paper is structured as follows: in the next section we present the mathematical formalism of the generalized modularities; then, we test the framework on synthetic and real networks; finally we discuss the results obtained.

2. Mathematical formulation of motif modularity

The original definition of modularity by Newman and Girvan [2] only deals with unweighted and undirected networks. Later on, Newman generalized it to cope with weighted networks [3]. In this work we start from an extension of modularity to weighted directed networks [14], which reduces to the previous one for undirected networks, and which is calculated as follows:

$$Q(C) = \frac{1}{2w} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2w} \right) \delta(C_i, C_j), \qquad (2.1)$$

where w_{ij} is the weight of the connection from the ith to the jth node, $w_i^{\text{out}} = \sum_j w_{ij}$ and $w_j^{\text{in}} = \sum_i w_{ij}$ stand for their output and input strengths respectively, $2w = \sum_{ij} w_{ij}$ is the total strength of the network, C_i is the index of the community which node i belongs to, and the Kronecker δ is 1 if nodes i and j are in the same community, 0

otherwise. For undirected networks, $w_i^{\text{out}} = w_i^{\text{in}} \equiv w_i$, thus recovering the weighted undirected definition of modularity in [3]. The larger the value of modularity, the better the corresponding partition of the network into modules.

In the next subsections we develop the mathematical formulation of a motif modularity which generalizes the standard one in (2.1). First, the most general framework is explained, and then the formalism is applied to several classes of motifs.

2.1. General motif modularity

Let $\mathcal{M} = (V_{\mathcal{M}}, E_{\mathcal{M}})$ be a *motif* (connected undirected graph, or weakly connected directed graph), where $V_{\mathcal{M}}$ is the set of M nodes of the motif, and $E_{\mathcal{M}} \subseteq V_{\mathcal{M}} \times V_{\mathcal{M}}$ is the set of its edges.

Let $\{w_{ij} \geq 0 \mid i, j = 1, \ldots, N\}$ be the weights of a (directed or undirected) network of N nodes, where $w_{ij} = 0$ if there is no edge from the ith to the jth node, and $w_{ij} \in \{0, 1\}$ if the network is unweighted. The nodes of the motif will be labeled by the indices $i_1, i_2, \ldots, i_M$, all of them running between 1 and N.

Given a certain partition C of an unweighted network in communities, the number of motifs fully included within the communities is given by

$$\Psi_{\mathcal{M}}(C) = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_M=1}^{N} \prod_{(a,b) \in E_{\mathcal{M}}} w_{i_a i_b} \, \delta(C_{i_a}, C_{i_b}) \,. \tag{2.2}$$

Degenerated motifs, i.e. those where some nodes are counted more than once, are included in this sum. The formula also holds for weighted networks, which can be inferred from the mapping between weighted networks and unweighted multigraphs [3].

The maximum value of $\Psi_{\mathcal{M}}(C)$ corresponds to the partition in a single community containing all the nodes:

$$\Psi_{\mathcal{M}} = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_M=1}^{N} \prod_{(a,b) \in E_{\mathcal{M}}} w_{i_a i_b} \,. \tag{2.3}$$

For a random network preserving the nodes' strengths, these quantities are respectively

$$\Omega_{\mathcal{M}}(C) = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_M=1}^{N} \prod_{(a,b) \in E_{\mathcal{M}}} w_{i_a}^{\text{out}} w_{i_b}^{\text{in}} \, \delta(C_{i_a}, C_{i_b}) \tag{2.4}$$

and

$$\Omega_{\mathcal{M}} = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_M=1}^{N} \prod_{(a,b) \in E_{\mathcal{M}}} w_{i_a}^{\text{out}} w_{i_b}^{\text{in}} \,. \tag{2.5}$$

Now, by analogy with the standard modularity, we define the *motif modularity* as the fraction of motifs inside the communities minus the fraction in a random network which preserves the nodes' strengths:

$$Q_{\mathcal{M}}(C) = \frac{\Psi_{\mathcal{M}}(C)}{\Psi_{\mathcal{M}}} - \frac{\Omega_{\mathcal{M}}(C)}{\Omega_{\mathcal{M}}} \,. \tag{2.6}$$

The introduction of *nullcase weights* n_{ij}, *masked weights* $w_{ij}(C)$ and *masked nullcase weights* $n_{ij}(C)$,

$$n_{ij} = w_i^{\text{out}} w_j^{\text{in}}, \tag{2.7}$$

$$w_{ij}(C) = w_{ij}\delta(C_i, C_j), \tag{2.8}$$

$$n_{ij}(C) = n_{ij}\delta(C_i, C_j), \tag{2.9}$$

allows the simplification of the previous expressions, in particular motif modularity:

$$Q_{\mathcal{M}}(C) = \frac{\sum\limits_{i_1 i_2 \cdots i_M} \prod\limits_{(a,b)\in E_{\mathcal{M}}} w_{i_a i_b}(C)}{\sum\limits_{i_1 i_2 \cdots i_M} \prod\limits_{(a,b)\in E_{\mathcal{M}}} w_{i_a i_b}} - \frac{\sum\limits_{i_1 i_2 \cdots i_M} \prod\limits_{(a,b)\in E_{\mathcal{M}}} n_{i_a i_b}(C)}{\sum\limits_{i_1 i_2 \cdots i_M} \prod\limits_{(a,b)\in E_{\mathcal{M}}} n_{i_a i_b}}. \tag{2.10}$$

Motif modularity may be further generalized by relaxing the condition that all nodes of the motif should be fully inside the modules. This is done just by removing some of the maskings in (2.10) as required, and possibly with the addition of some Kronecker δ functions between non-adjacent nodes of the motif. In this way, it is possible to define classes of nodes different from communities, as we shall see in subsection 2.3.

2.2. Cycle modularity

Among the simplest possible motifs, triangles are the ones which have deserved more attention in the networks literature. For instance, it has been shown that real networks have higher clustering coefficients than expected in random networks [15]. Thus, it would be desirable to be able to find "communities of triangles". Our approach consists in the definition of a *triangle modularity* $Q_{\triangle}(C)$, based on the triangular motif $E_{\triangle} = \{(1,2),(2,3),(3,1)\}$, which reads:

$$Q_{\triangle}(C) = \frac{\sum\limits_{ijk} w_{ij}(C)w_{jk}(C)w_{ki}(C)}{\sum\limits_{ijk} w_{ij}w_{jk}w_{ki}} - \frac{\sum\limits_{ijk} n_{ij}(C)n_{jk}(C)n_{ki}(C)}{\sum\limits_{ijk} n_{ij}n_{jk}n_{ki}}. \tag{2.11}$$

Triangle modularity is trivially generalizable to cycles of length ℓ, making use of the cyclical motif $E_{\mathcal{C}^{(\ell)}} = \{(1,2),(2,3),\ldots,(\ell-1,\ell),(\ell,1)\}$. The number of these motifs within the communities is given by

$$\Psi_{\mathcal{C}^{(\ell)}}(C) = \sum\limits_{i_1 i_2 \cdots i_\ell} w_{i_1 i_2}(C)w_{i_2 i_3}(C)\cdots w_{i_{\ell-1} i_\ell}(C)w_{i_\ell i_1}(C). \tag{2.12}$$

The full formula for the cycle modularity $Q_{\mathcal{C}^{(\ell)}}(C)$ follows immediately from it.

If the network is directed, other non-cyclical motifs exist. We skip them, since their derivation is straightforward.

2.3. Path modularity

A *path* $\mathcal{P}^{(\ell)}$ of length ℓ is simply the linear motif $E_{\mathcal{P}^{(\ell)}} = \{(1,2),(2,3),\ldots,(\ell,\ell+1)\}$. We remark that cycles are closed paths, but here we shall only consider open paths.

The number of paths of length ℓ fully inside the communities is given by

$$\Psi_{\mathcal{P}^{(\ell)}}(C) = \sum_{i_1 i_2 \cdots i_{\ell+1}} w_{i_1 i_2}(C) w_{i_2 i_3}(C) \cdots w_{i_\ell i_{\ell+1}}(C) \,. \qquad (2.13)$$

Note that this expression equals the sum of the components of the ℓth power of the masked weight matrix.

The path of length $\ell = 1$ corresponds to the simplest motif $E_{\mathcal{P}^{(1)}} = \{(1,2)\}$, which is just a single edge, so its motif modularity (2.10) equals the standard definition of modularity (2.1).

Paths of length 2 are also useful for the analysis of bipartite networks, provided one removes the constraint that all nodes of the path belong to the same module. If one allows that the middle node of a path of length 2 could be any node of the network, whereas the first and third nodes are kept within the same group, the path can be used to discover relationships between nodes of different groups. If a network is bipartite, for instance, there will be many paths of length 2 starting from a class and returning to it from the other class. If only the extremes of the path $\tilde{\mathcal{P}}^{(\ell)}$ are required to be inside the community, their total number is given by

$$\Psi_{\tilde{\mathcal{P}}^{(\ell)}}(C) = \sum_{i_1 i_2 \cdots i_{\ell+1}} w_{i_1 i_2} w_{i_2 i_3} \cdots w_{i_\ell i_{\ell+1}} \delta(C_{i_1}, C_{i_{\ell+1}}) \,. \qquad (2.14)$$

In this case, the calculation makes use of the ℓth power of the weight matrix (instead of the masked weight matrix), and the masking is applied to the sum of their components.

3. Examples and tests

When one is faced with the problem of community detection in a particular network, the first thing to do should be to answer the following question: what sort of connectivity patterns or motifs are pertinent in this study? According to the answer, it is straightforward to select one of the possible motif modularities. We present in this section examples of the application of the previous framework to two synthetic networks. Finally, we perform two tests on real networks for which the real partitions observed are known.

The synthetic networks that we have generated for this purpose are the clique & circle network and the star network. In figure 1 we show these networks as well as the classes found using different motif modularities. Suppose we want to find node classes by means of triangles. When we optimize the triangle modularity for the clique & circle network, the clique forms a community whereas the nodes of the circle are separated into five singleton communities. This is due to the absence of triangles within the circle. On the contrary, the standard modularity identifies the circle as a community.

The second example, the star network, is a case where the path motifs prove to be useful. This network can be seen as a simple bipartite network with eight actors (the leaf nodes) and just one event (the hub node). In this case, recalling what we have said in the previous section, the path modularity of length 2 with a free intermediate node is the proper motif modularity to use. The results confirm that the star is decomposed in
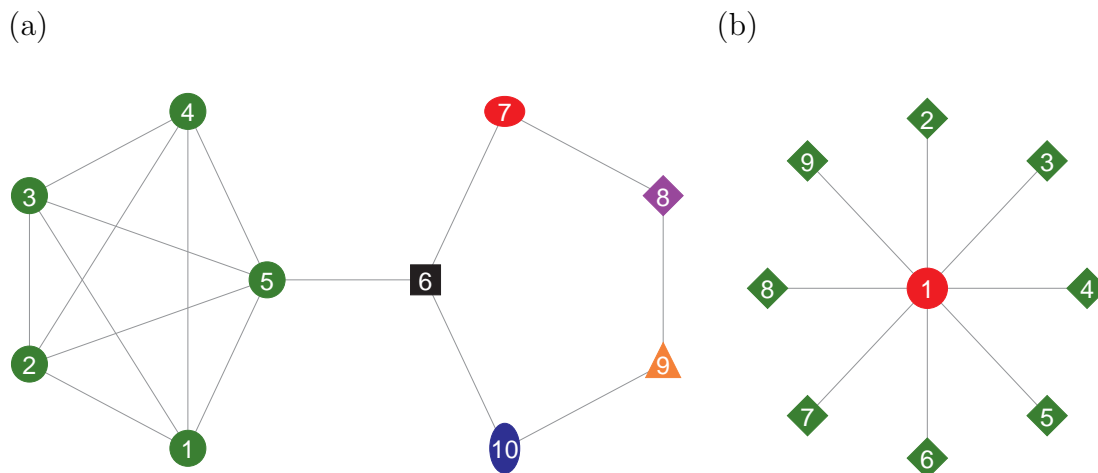
Figure 1. Results for two synthetic networks: (a) Clique & circle network, with triangle modularity; (b) Star network, with paths of size 2 modularity with free intermediate node (see text for details). Members of the same class are depicted using equal symbol and color.

two classes, one for the leaves and another for the hub. The same partition is obtained for any even path length with free intermediate nodes, while for odd path lengths all nodes are joined in a single community. This holds as well if one maximizes the standard modularity; however, the correct partition of the network can be recovered by modularity minimization.

The real networks used for testing are the Zachary Karate Club network [16] and the Southern Women Event Participation network [17, 18]. A description of each network can be found in their respective references. For the mathematical analysis presented here the interesting fact regarding these networks is that we know the real splittings occurred in the Zachary network, as well as the most plausible classification assigned in the literature to the Women Event Participation data, as reported by Freeman [18]. In figure 2 we show both networks as well as their respective partitions.

For the Zachary Karate Club network, the nature of the data suggests to try an optimization of path modularities, since the decision of following any of the two leaders during the splitting of the club surely depended on higher order friendship relationships (friends of friends, and so on). When a path modularity of length 1 is considered (i.e. the classical definition of modularity), the best partition obtained splits each one of the two real communities into two sub-communities, yielding a partition in four communities. But when one looks for a more compact structure of the communities, which can be accomplished by increasing the length of the paths, the optimization of path modularity delivers the real splitting observed, for all path lengths we have used (from 2 to 6). The same result is obtained when the paths are replaced by cycles (lengths from 4 to 9). Triangles give almost the exact partition, but with two exceptions: nodes 10 and 12 become isolated, because they do not belong to any triangle.
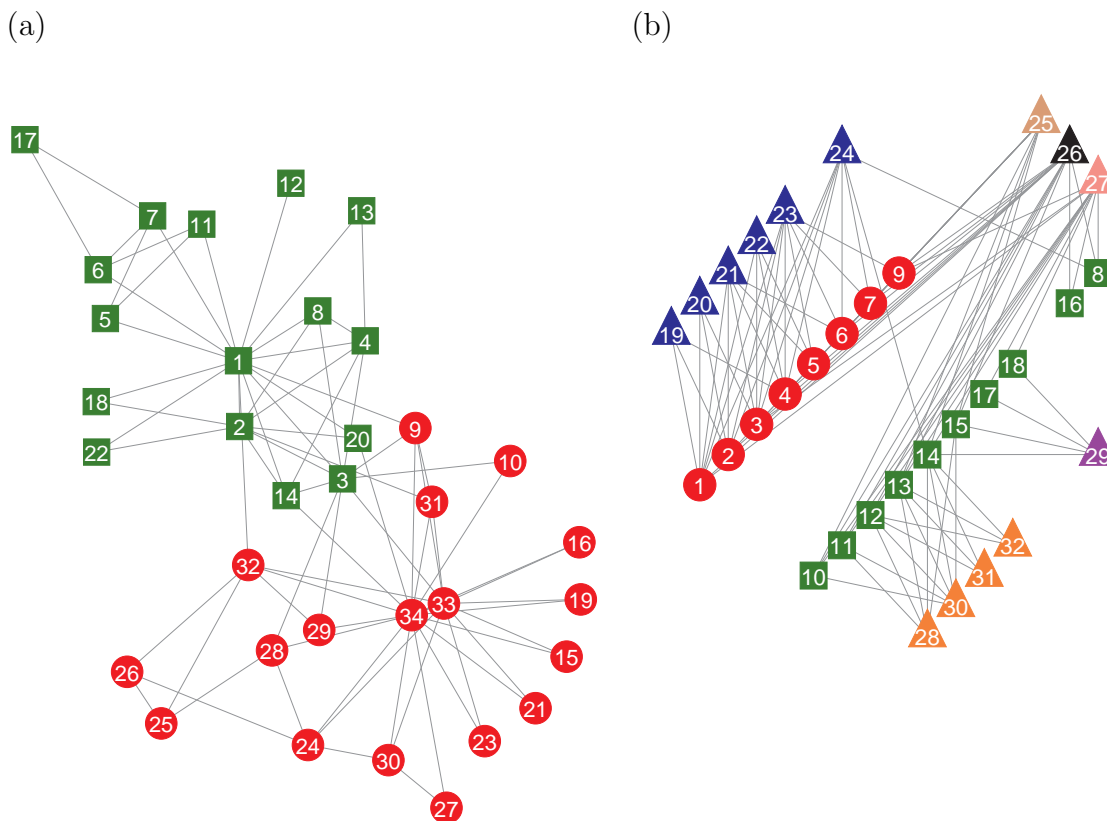
(a)   (b)

Figure 2. Results for two real networks: (a) Zachary Karate Club network. We depict the real splitting obtained when using several path and cycle modularities; (b) Southern Women Event Participation network. We depict the results of the analysis of this multipartite network without any projection, simply applying modularity of path free intermediate of length 2. Remarkably the results show clearly the role differentiation of women and events, as well as the splitting of women according to the events participation that has been reported in the literature.

The second network tested is a multipartite network. In this case, as well as for the star network, the use of path modularity of length 2 with a free intermediate node is crucial, and it accounts for the role differentiation between women and events. The results not only reveal the two roles of events and women, but also recover their internal split according to their participation in events, a classification made by social scientists [18] (with the same exception of one woman, as in the weighted projection and bipartite methods in [12]). In this case, the minimization of standard modularity is only able to separate women and events, with no further subdivision.

4. Conclusions

In this work we have shown that a general classification of node groups in networks is possible if one uses motifs as elementary units, instead of simple edges. To show that, we generalized Newman-Girvan modularity by replacing edges with motifs. The new

versions of modularity obtained have been tested on synthetic and real networks, and are able to recover expected connectivity patterns in networks, both when the networks have modular structure and when they have multipartite structure. However, the principle goes beyond the use of modularity and could inspire promising alternative frameworks.

References

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

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

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

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

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

[6] Duch J and Arenas A 2005 Community identification using Extremal Optimization *Phys. Rev. E* **72** 027104

[7] Guimerà R and Amaral L A N 2005 Functional cartography of metabolic networks *Nature* **433** 895

[8] Pujol J M, Béjar J and Delgado J 2006 Clustering Algorithm for Determining Community Structure in Large Networks *Phys. Rev. E* **74** 016107

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

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

[11] Newman M E J 2006 Finding community structure in networks using the eigenvectors of matrices *Phys. Rev. E* **74** 036104

[12] Guimerà R, Sales-Pardo M, Amaral L A N 2007 Module identification in bipartite and directed networks *Phys. Rev. E* **76** 036102

[13] Barber M J 2007 Modularity and community detection in bipartite networks Preprint arXiv:0707.1616

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

[15] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 Network Motifs: Simple Building Blocks of Complex Networks *Science* **298** 824

[16] Zachary W W 1977 An information flow model for conflict and fission in small groups *J. Anthr. Res.* **33** 452

[17] Davis A, Gardner B B and Gardner M R 1941 *Deep South* (Chicago: The University of Chicago Press)

[18] Freeman L 2003 Finding Social Groups: A Meta-Analysis of the Southern Women Data *Dynamic*

Social Network Modeling Analysis: Workshop Summary and Papers ed R Breiger, K Carley and P Pattison (Washington DC: The National Academies Press) p 39