

Community detection and graph partitioning

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2013 EPL 103 28003

(<http://iopscience.iop.org/0295-5075/103/2/28003>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

This content was downloaded by: count0

IP Address: 134.102.186.160

This content was downloaded on 19/08/2014 at 14:43

Please note that [terms and conditions apply](#).

Community detection and graph partitioning

M. E. J. NEWMAN

*Department of Physics and Center for the Study of Complex Systems, University of Michigan
Ann Arbor, MI 48109, USA*

received 7 June 2013; accepted in final form 16 July 2013

published online 9 August 2013

PACS 89.75.Hc – Networks and genealogical trees

PACS 02.50.Cw – Probability theory

PACS 02.70.Hm – Spectral methods

Abstract – Many methods have been proposed for community detection in networks. Some of the most promising are methods based on statistical inference, which rest on solid mathematical foundations and return excellent results in practice. In this paper we show that two of the most widely used inference methods can be mapped directly onto versions of the standard minimum-cut graph partitioning problem, which allows us to apply any of the many well-understood partitioning algorithms to the solution of community detection problems. We illustrate the approach by adapting the Laplacian spectral partitioning method to perform community inference, testing the resulting algorithm on a range of examples, including computer-generated and real-world networks. Both the quality of the results and the running time rival the best previous methods.



Copyright © EPLA, 2013

Introduction. – The problem of community detection in networks has received wide attention [1,2]. It has proved to be a problem of remarkable subtlety, computationally challenging and with deep connections to other areas of research including machine learning, signal processing, and spin-glass theory. A large number of algorithmic approaches to the problem have been considered, but interest in recent years has focused particularly on statistical inference methods [3–5], partly because they give excellent results, but also because they are mathematically principled and, at least in some cases, provably optimal [5,6].

In this paper we study two of the most fundamental community inference methods, based on the so-called stochastic block model or its degree-corrected variant [7]. We show that it is possible to map both methods onto the well-known minimum-cut graph partitioning problem, which allows us to adapt any of the large number of available methods for graph partitioning to solve the community detection problem. As an example, we apply the Laplacian spectral partitioning method of Fiedler [8,9] to derive a community detection method competitive with the best currently available algorithms in terms of both speed and quality of results.

Likelihood maximization for the stochastic block model. – The first method we consider is based on the *stochastic block model*, sometimes also called the *planted*

partition model, a well-studied model of community structure in networks [7,10]. This model supposes a network of n vertices divided into some number of groups or communities, with different probabilities for connections within and between groups. We will initially focus on the simplest case of just two groups (of any size, not necessarily equal). In the commonest version of the model edges are placed independently at random between vertex pairs with probability p_{in} for pairs in the same group and p_{out} for pairs in different groups. In this paper we use the slightly different Poisson version of the model described in [7], in which we place between each pair of vertices a Poisson-distributed number of edges with mean ω_{in} for pairs in the same group and ω_{out} for pairs in different groups. In essentially all real-world networks the fraction of possible edges that are actually present in the network is extremely small (usually modeled as vanishing in the large- n limit), in which case the two versions of the model become indistinguishable, but the Poisson version is preferred because its analysis is more straightforward.

At its heart, the statistical inference of community structure is a matter of answering the following question: if we assume an observed network is generated according to our model, what then must the parameters of that model have been? In other words, what were the values of ω_{in} and ω_{out} used to generate the network and, more importantly, which vertices fell in which groups? Even though the model is probably not a good representation of the

process by which most real-world networks are generated, the answer to this question often gives a surprisingly good estimate of the true community structure.

To answer the question, we make use of a maximum likelihood method. Let us label the two groups or communities in our model group 1 and group 2, and denote by g_i the group to which vertex i belongs. The edges in the network will be represented by an adjacency matrix having elements

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge between vertices } i, j, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Then the likelihood of generating a particular network or graph G , given the complete set of group memberships, which we will denote by the shorthand g , and the Poisson parameters, which we will denote by ω , is

$$P(G|g, \omega) = \prod_{i < j} \frac{\omega_{g_i g_j}^{A_{ij}}}{A_{ij}!} \exp(-\omega_{g_i g_j}), \quad (2)$$

where $\omega_{g_i g_j}$ denotes the expected number of edges between vertices in groups g_i and g_j —either ω_{in} or ω_{out} , depending on whether the groups are the same or different. We are assuming there are no self-edges in the network—edges that connect vertices to themselves—so $A_{ii} = 0$ for all i .

Given the likelihood, one can maximize it to find the most likely values of the group labels and parameters, which can be done in a number of different ways. In ref. [7], for example, the likelihood was maximized first with respect to the parameters ω_{in} and ω_{out} by differentiation. Applying this method to eq. (2) gives most likely values of

$$\omega_{\text{in}} = \frac{2m_{\text{in}}}{n_1^2 + n_2^2}, \quad \omega_{\text{out}} = \frac{m_{\text{out}}}{n_1 n_2}, \quad (3)$$

where m_{in} and m_{out} are the observed numbers of edges within and between groups, respectively, for a given candidate division of the network, and n_1 and n_2 are the numbers of vertices in each group. Substituting these values back into eq. (2) gives the *profile likelihood*, which depends on the group labels only. In fact, one typically quotes not the profile likelihood itself but its logarithm, which is easier to work with. Neglecting an unimportant additive constant, the log of the profile likelihood for the present model is

$$\mathcal{Q} = m_{\text{in}} \ln \frac{2m_{\text{in}}}{n_1^2 + n_2^2} + m_{\text{out}} \ln \frac{m_{\text{out}}}{n_1 n_2}. \quad (4)$$

The communities can now be identified by maximizing this quantity over all possible assignments of the vertices to the groups. This is still a hard task, however. There are an exponentially large number of possible assignments, so an exhaustive search through all of them is unfeasible for all but the smallest of networks. One can apply standard heuristics like simulated annealing to the problem, but in this paper we take a different approach.

In the calculation above, the likelihood is maximized over ω first, for fixed group assignments, then over the group assignments. But we can also take the reverse approach, maximizing first over the group assignments, for given ω , and then over ω at the end. This approach is attractive for two reasons. First, as we will show, the problem of maximizing with respect to the group assignments when ω is given is equivalent to the standard problem of minimum-cut graph partitioning, a problem for which many excellent heuristics are already available. Second, after maximizing with respect to the group assignments the remaining problem of maximizing with respect to ω is a one-parameter optimization that can be solved trivially. The net result is that the problem of maximum-likelihood community detection is reduced to one of performing a well-understood task—graph partitioning—plus one undemanding extra step. The resulting algorithm is fast and, as we will see, gives good results.

So consider the problem of maximizing the likelihood, eq. (2), with respect to the group labels g_i , for given values of the parameters ω_{in} and ω_{out} . We will actually maximize the logarithm \mathcal{L} of the likelihood,

$$\mathcal{L} = \ln P(G|g, \omega) = \sum_{i < j} [A_{ij} \ln \omega_{g_i g_j} - \omega_{g_i g_j} - \ln A_{ij}!], \quad (5)$$

which gives the same result but is usually easier.

To proceed we write $\omega_{g_i g_j}$ and $\ln \omega_{g_i g_j}$ as

$$\omega_{g_i g_j} = \delta_{g_i g_j} \omega_{\text{in}} + (1 - \delta_{g_i g_j}) \omega_{\text{out}}, \quad (6)$$

$$\ln \omega_{g_i g_j} = \delta_{g_i g_j} \ln \omega_{\text{in}} + (1 - \delta_{g_i g_j}) \ln \omega_{\text{out}}, \quad (7)$$

where δ_{ij} is the Kronecker delta. Substituting these into eq. (5) and dropping overall additive and multiplicative constants, which have no effect on the position of the maximum, the log-likelihood can be rearranged to read

$$\mathcal{L} = \sum_{i < j} (1 - \delta_{g_i g_j}) (\gamma - A_{ij}), \quad (8)$$

where

$$\gamma = \frac{\omega_{\text{in}} - \omega_{\text{out}}}{\ln \omega_{\text{in}} - \ln \omega_{\text{out}}}, \quad (9)$$

which is positive whenever $\omega_{\text{in}} > \omega_{\text{out}}$, meaning we have traditional community structure in our network. (It is possible to repeat the calculations for the case $\omega_{\text{in}} < \omega_{\text{out}}$ and derive methods for detecting such structure as well, although we will not do that here.)

The quantity $\sum_{i < j} (1 - \delta_{g_i g_j}) A_{ij}$ is the *cut size* of the network partition represented by our two communities, i.e., the number of edges connecting vertices in different communities, which we previously denoted m_{out} , and

$$\sum_{i < j} (1 - \delta_{g_i g_j}) = n_1 n_2, \quad (10)$$

where as previously n_1 and n_2 are the numbers of vertices in communities 1 and 2. Thus, we can also write the log-likelihood in the form

$$\mathcal{L} = -m_{\text{out}} + \gamma n_1 n_2. \quad (11)$$

The maximization of this log-likelihood corresponds to the minimization of the cut size, with an additional penalty term $\gamma n_1 n_2$ that favors groups of equal size. This is similar, though not identical, to the so-called *ratio cut* problem, in which one minimizes the ratio $m_{\text{out}}/n_1 n_2$, which also favors groups of equal size, although the nature of the penalty for unbalanced groups is different.

The catch with maximizing eq. (11) is that we do not know the value of γ , which depends on the unknown quantities ω_{in} and ω_{out} via eq. (9), but we can get around this problem by the following trick. We first perform a limited maximization of (11) in which the sizes n_1 and n_2 of the groups are held fixed at some values that we choose. This means that the term $\gamma n_1 n_2$ is a constant and hence drops out of the problem and we are left to maximize $-m_{\text{out}}$ only, or equivalently minimize the cut-size m_{out} . This problem is now precisely the standard minimum-cut problem of graph partitioning—the minimization of the cut size for divisions of a graph into groups of given sizes¹.

There are $n + 1$ possible choices of the sizes of the two groups, ranging from putting all vertices in group 1 to all vertices in group 2, and everything in between. If we solve the minimum-cut problem for each of these $n + 1$ choices we get a set of $n + 1$ solutions and we know that one of these must be the solution to our overall maximum likelihood problem. It remains only to work out which one. But choosing between them is easy, since we know that the true maximum also maximizes the profile likelihood, eq. (4). So we can simply calculate the profile likelihood for each solution in turn and find the one that gives the largest result.

In effect, this approach narrows the exponentially large pool of candidate divisions of the network to a one-parameter family of just $n + 1$ solutions (parametrized by group size), from which it is straightforward to pick the overall winner by exhaustive search. Moreover, the individual candidate solutions are all themselves solutions of the standard minimum-cut partitioning problem, a problem that has been well studied for many years and about which a great deal is known [11,12]. Although partitioning problems are, in general, hard to solve exactly, there exist many heuristics that give good answers in practical situations. The approach developed here allows us to apply any of these heuristics directly to the maximum-likelihood community detection problem.

Although we have concentrated here on the case of a network with just two communities, our results generalize in straightforward fashion to more than two. The fundamental equations (5)–(9) are unchanged for more than two

communities and eq. (10) generalizes to

$$\sum_{i < j} (1 - \delta_{g_i g_j}) = \sum_{r < s} n_r n_s. \quad (12)$$

Then the likelihood takes the form

$$\mathcal{L} = -m_{\text{out}} + \gamma \sum_{r < s} n_r n_s. \quad (13)$$

If we fix the sizes of the communities, the second term again becomes constant and the maximization of the log-likelihood is equivalent to a minimum-cut partitioning, leading once more to a polynomial-sized family of candidate solutions to the community detection problem. Among these, the correct overall solution is the one that maximizes the profile likelihood, which for an arbitrary number of communities is given by

$$\mathcal{Q} = \sum_{r,s} m_{rs} \log \frac{m_{rs}}{n_r n_s}, \quad (14)$$

where m_{rs} is the number of edges running between groups r and s , or twice that number when $r = s$ [5,7].

Spectral algorithm. As an example of our approach, we demonstrate a fast and simple spectral algorithm for the two-community case based on the Laplacian spectral bisection method for graph partitioning introduced by Fiedler [8,9]. A description of this method can be found, for example, in [13], where it is shown that a good approximation to the minimum-cut division of a network into two parts of specified sizes can be found by calculating the *Fiedler vector*, which is the eigenvector of the graph Laplacian matrix \mathbf{L} corresponding to the second smallest eigenvalue. (The graph Laplacian is the $n \times n$ symmetric matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where \mathbf{A} is the adjacency matrix and \mathbf{D} is the $n \times n$ diagonal matrix with D_{ii} equal to the degree of vertex i .) Having calculated the Fiedler vector, one divides the network into groups of the required sizes n_1 and n_2 by inspecting the vector elements and assigning the n_1 vertices with the largest (most positive) elements to group 1 and the rest to group 2. Although the method gives only an approximation to the global minimum-cut division, practical experience (and some rigorous results) show that it gives good answers under commonly occurring conditions [9].

A nice feature of this approach is that, in a single calculation, it gives us the entire one-parameter family of minimum-cut divisions of the network. We need calculate the Fiedler vector only once, sort its elements in decreasing order, then cut them into two groups in each of the $n + 1$ possible ways and calculate the profile likelihood for the resulting divisions of the network. The one with the highest score is (an approximation to) the maximum-likelihood community division of the network.

Degree-corrected block model. These developments are for the standard stochastic block model. As shown in

¹Not to be confused with the similarly named, but quite different, maximum-flow/minimum-cut problem, which is the problem of finding the minimum cut that separates two given vertices in a network.

ref. [7], however, the standard block model gives poor results when applied to most real-world networks because the model fails to take into account the broad degree distribution such networks possess. This problem can be fixed by a relatively simple modification of the model in which the expected number $\omega_{g_i g_j}$ of edges between vertices i and j is replaced by $k_i k_j \omega_{g_i g_j}$ where k_i is the degree of vertex i . All the developments for the standard block model above generalize in straightforward fashion to this “degree-corrected” model. The log-likelihood and log-profile likelihood become

$$\mathcal{L} = -m_{\text{out}} + \gamma \kappa_1 \kappa_2, \quad \mathcal{Q} = m_{\text{in}} \ln \frac{2m_{\text{in}}}{\kappa_1^2 + \kappa_2^2} + m_{\text{out}} \ln \frac{m_{\text{out}}}{\kappa_1 \kappa_2}, \quad (15)$$

where κ_1 and κ_2 are the sums of the degrees of the vertices in the two groups. In other words, the expressions are identical to those for the uncorrected model except for the replacement of the group sizes n_1, n_2 by κ_1, κ_2 .

The maximization of \mathcal{L} is thus once again reduced to a generalized minimum-cut partitioning problem, with a penalty term proportional to $\kappa_1 \kappa_2$, which again favors balanced groups. Although we do not know the value of γ , we can reduce the problem to a variant of the minimum-cut problem by the equivalent of our previous approach, holding κ_1 and κ_2 constant. And again we can derive a spectral algorithm for this problem based on the graph Laplacian. By a derivation analogous to that for the standard spectral method we can show that a good approximation to the problem of minimum-cut partitioning with fixed κ_1, κ_2 (as opposed to fixed n_1, n_2) is given not by the second eigenvector of \mathbf{L} but by the second eigenvector of the generalized eigensystem $\mathbf{L}\mathbf{v} = \lambda \mathbf{D}\mathbf{v}$, where, as previously, \mathbf{D} is the diagonal matrix of vertex degrees. Once again we calculate the vector and split the vertices into two groups according to the sizes of their corresponding vector elements and once again this gives us a one-parameter family of $n+1$ candidate solutions from which we can choose an overall winner by finding the one with the highest profile likelihood, eq. (15).

Results. – We have tested this method on a variety of networks, and in practice it appears to work well. Figure 1 shows results from tests of the degree-corrected algorithm of the previous section on a large group of synthetic (*i.e.*, computer-generated) networks. These networks were themselves generated using stochastic block models (which are commonly used as a benchmark for community detection [1,10]). Panels (a) and (b) in the figure show the profile likelihood for the families of $n+1$ candidate solutions generated by the spectral calculation, for networks drawn from the ordinary (not degree-corrected) block model with two equally sized groups (a) and unequal groups (b). In each case there is a clear peak in the profile likelihood at the correct group size, suggesting that the algorithm has correctly identified the group membership of most vertices. Panel (c) shows results for networks generated using the degree-corrected block model with

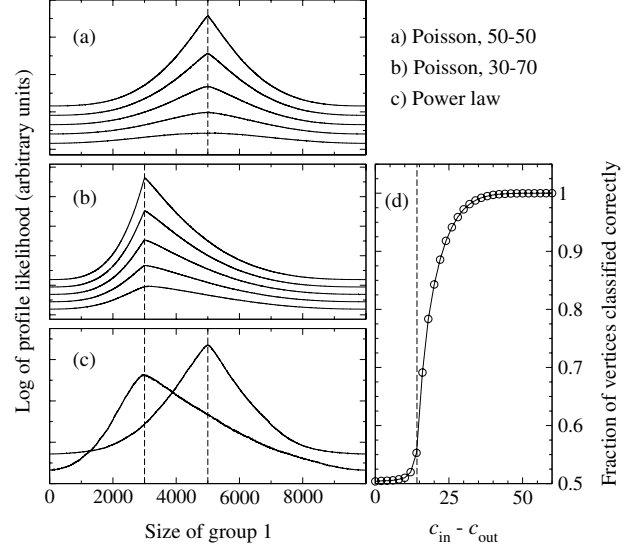


Fig. 1: Results from tests of the method developed here on computer-generated benchmark networks. (a) Profile likelihood as a function of group size for candidate solutions generated by the spectral method described in the text, for single networks of $n = 10000$ vertices drawn from the (uncorrected) stochastic block model with two equally sized groups of 5000 vertices each and a range of strengths of the community structure. Defining $c_{\text{in}} = n\omega_{\text{in}}$ and $c_{\text{out}} = n\omega_{\text{out}}$, the curves are (top to bottom) $c_{\text{in}} = 80, 75, 70, 65$, and 60 , and $c_{\text{out}} = 100 - c_{\text{in}}$. The dashed vertical line indicates the true size of the planted communities. The curves have been displaced from one another vertically for clarity. The vertical axis units are arbitrary because additive and multiplicative constants have been neglected in the definition of the log-likelihood. (b) Profile likelihoods for the same parameter values but groups of unequal sizes 3000 and 7000. (c) Profile likelihoods for single networks of $n = 10000$ vertices generated using the degree-corrected stochastic block model with an expected degree distribution following a power law with exponent -2.5 , minimum degree 10, and $\omega_{\text{in}}, \omega_{\text{out}}$ chosen to give a 10:1 ratio of within-group edges to between-group edges. The two curves are, respectively, for networks with equally sized groups and with groups of size 3000 and 7000. (d) The average fraction of vertices classified correctly for networks generated from the uncorrected block model with 10000 vertices and two equally sized groups. Each point is an average over 100 networks. Statistical errors are smaller than the points in all cases. The vertical dashed line indicates the position of the “detectability threshold” at which the community structure becomes formally undetectable [6,14–20].

a power-law degree distribution. Networks with broad degree distributions similar to those seen in real-world networks present a more realistic challenge to the algorithm, but again the peaks in the curves fall at the correct points, suggesting that the algorithm has correctly determined the group membership of most vertices.

Panel (d) in fig. 1 quantifies the success rate of the algorithm, plotting the fraction of correctly identified vertices as a function of the strength of community structure for the ordinary block model in the case of

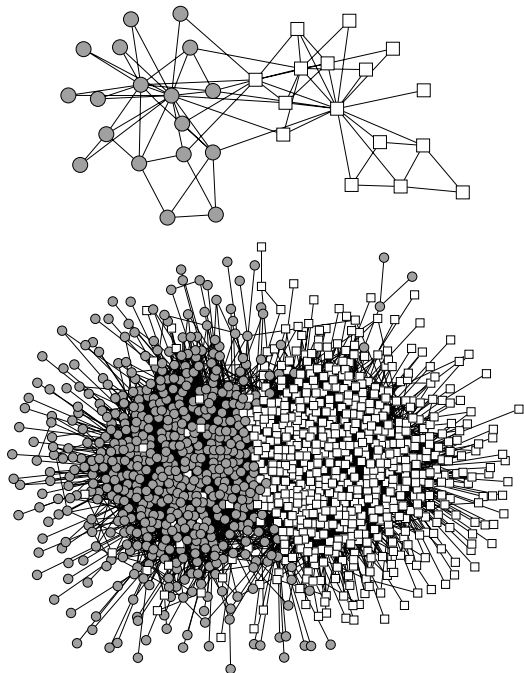


Fig. 2: Division into two groups of two well-known networks from the literature. Top: the karate club network of Zachary [21]. Bottom: the network of political blogs compiled by Adamic and Glance [22]. Shades and shapes of vertices indicate group membership and both divisions are qualitatively similar to the accepted ones.

equally sized groups (which is the most difficult case). As the figure shows, the algorithm correctly identifies most vertices over a large portion of the parameter space. The vertical dashed line represents the “detectability threshold” identified by previous authors [6,14–20], below which every method of community detection must fail. Our algorithm fails below this point also, as it must, but appears to work well essentially all the way down to the transition, and there are reasons to believe this result to be exact, at least for networks that are not too sparse [16].

Figure 2 shows the results of applications of the algorithm to two well-studied real-world networks, Zachary’s “karate club” network [21] and Adamic and Glance’s network of political blogs [22]. Both are known to have pronounced community structure and the divisions found by our spectral algorithm mirror closely the accepted communities in both cases.

In addition to being effective, the algorithm is also fast. The computation of the eigenvector can be done using, for instance, the Lanczos method, an iterative method which takes time $O(m)$ per iteration, where m is the number of edges in the network. The number of iterations required is typically small, although the exact number is not known in general. The search for the division that maximizes the profile likelihood can also be achieved in $O(m)$ time. Of the $n + 1$ different divisions of the network that must be considered, each one differs from the previous one by the

movement of just a single vertex from one group to the other. The movement of vertex i between groups causes the quantities appearing in eq. (15) to change according to

$$\kappa_1 \rightarrow \kappa_1 - k_i, \quad \kappa_2 \rightarrow \kappa_2 + k_i, \quad (16)$$

$$m_{\text{in}} \rightarrow m_{\text{in}} - \Delta m, \quad m_{\text{out}} \rightarrow m_{\text{out}} + \Delta m, \quad (17)$$

where Δm equals the number of edges between i and vertices in group 1 minus the number between i and vertices in group 2. These quantities and the resulting change in the profile likelihood can be calculated in time proportional to the degree of the vertex and hence all n vertices can be moved in time proportional to the sum of all degrees in the network, which is equal to $2m$. Thus, to leading order, the total running time of the algorithm goes as m times the number of Lanczos iterations, the latter typically being small, and in practice the method is about as fast as the best competing algorithms and should be feasible for networks of millions of vertices or more.

Conclusions. – In this paper we have shown that the widely studied maximum likelihood method for community detection in networks can be reduced to a search through a small family of candidate solutions, each of which is itself the solution to a minimum-cut graph partitioning problem, which is a well-studied problem about which much is known. This mapping allows us to use trusted partitioning heuristics to solve the community detection problem. As an example we have adapted the method of Laplacian spectral partitioning to derive a spectral likelihood maximization algorithm and tested its performance on both synthetic and real-world networks. In terms of both accuracy and speed we find the algorithm to be competitive with the best current methods.

A number of extensions of our approach would be possible, including extensions with more general forms for the parameters ω , such as different values of ω_{in} and ω_{out} for different groups, but we leave these for future work.

The author would like to thank CHARLIE DOERING, TAMMY KOLDA, and RAJ RAO NADAKUDITI for useful conversations, and LADA ADAMIC for providing the data for the network of political blogs. This work was funded in part by the National Science Foundation under grant DMS-1107796 and by the Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) under grant FA9550-12-1-0432.

REFERENCES

- [1] GIRVAN M. and NEWMAN M. E. J., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 7821.
- [2] FORTUNATO S., *Phys. Rep.*, **486** (2010) 75.

- [3] AIROLDI E. M., BLEI D. M., FIENBERG S. E. and XING E. P., *J. Mach. Learn. Res.*, **9** (2008) 1981.
- [4] CLAUSET A., MOORE C. and NEWMAN M. E. J., *Nature*, **453** (2008) 98.
- [5] BICKEL P. J. and CHEN A., *Proc. Natl. Acad. Sci. U.S.A.*, **106** (2009) 21068.
- [6] DECELE A., KRZAKALA F., MOORE C. and ZDEBOROVÁ L., *Phys. Rev. Lett.*, **107** (2011) 065701.
- [7] KARRER B. and NEWMAN M. E. J., *Phys. Rev. E*, **83** (2011) 016107.
- [8] FIEDLER M., *Czech. Math. J.*, **23** (1973) 298.
- [9] POTHEN A., SIMON H. and LIOU K.-P., *SIAM J. Matrix Anal. Appl.*, **11** (1990) 430.
- [10] CONDON A. and KARP R. M., *Random Struct. Algorithms*, **18** (2001) 116.
- [11] ELSNER U., Technical Report Technische Universität Chemnitz 97-27 (1997).
- [12] FJÄLLSTRÖM P.-O., *Linköping Electronic Artic. Comput. Inf. Sci.*, **3** (1998).
- [13] NEWMAN M. E. J., *Networks: An Introduction* (Oxford University Press, Oxford) 2010.
- [14] REICHARDT J. and LEONE M., *Phys. Rev. Lett.*, **101** (2008) 078701.
- [15] HU D., RONHOVDE P. and NUSSINOV Z., *Philos. Mag*, **92** (2012) 406.
- [16] NADAKUDITI R. R. and NEWMAN M. E. J., *Phys. Rev. Lett.*, **108** (2012) 188701.
- [17] MOSSEL E., NEEMAN J. and SLY A., preprint arXiv:1202.1499 (2012).
- [18] RADICCHI F., preprint arXiv:1306.1102 (2013).
- [19] FLORETTA L., LIECHTI J., FLAMMINI A. and DE LOS RIOS P., preprint arXiv:1306.2230 (2013).
- [20] PEIXOTO T. P., preprint arXiv:1306.2507 (2013).
- [21] ZACHARY W. W., *J. Anthropol. Res.*, **33** (1977) 452.
- [22] ADAMIC L. A. and GLANCE N., *The political blogosphere and the 2004 US election*, presented at *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem 2005*.