PHYSICAL REVIEW E **88**, 010801(R) (2013)

Detectability of communities in heterogeneous networks

Filippo Radicchi*
Departament d'Enginyeria Quimica, Universitat Rovira i Virgili, 43007 Tarragona, Spain
(Received 6 May 2013; published 12 July 2013)

Communities are fundamental entities for the characterization of the structure of real networks. The standard approach to the identification of communities in networks is based on the optimization of a quality function known as *modularity*. Although modularity has been at the center of an intense research activity and many methods for its maximization have been proposed, not much is yet known about the necessary conditions that communities need to satisfy in order to be detectable with modularity maximization methods. Here, we develop a simple theory to establish these conditions, and we successfully apply it to various classes of network models. Our main result is that heterogeneity in the degree distribution helps modularity to correctly recover the community structure of a network and that, in the realistic case of scale-free networks with degree exponent $\gamma < 2.5$, modularity is always able to detect the presence of communities.

DOI: 10.1103/PhysRevE.88.010801 PACS number(s): 89.75.Hc, 02.70.Hm, 64.60.aq

Communities are organizational modules that provide a coarse grained view of a complex network [1–3]. Depending on the nature of the network, communities can have different yet fundamental meanings: In biological networks, communities are likely to group entities having the same biological function [4–6], in the graph of the World Wide Web they may correspond to groups of pages dealing with the same or related topics [7], in food webs they may identify compartments [8], etc. Since communities play an important role for the characterization of the structure of networks, the development of computer algorithms for the detection of communities in networks represents one of the most active areas in network science [3]. In particular, the use of the so-called *modularity* has attracted a great deal of attention in recent years [9,10]. Modularity is a quality function that estimates the relevance of a given network partition (i.e., a division of the network in a given set of communities) by comparing the observed number of internal connections of the communities with the expected number of such edges in a random annealed version of the network. The best community division of the network is then given by the partition that maximizes the modularity function.

Although subjected to some intrinsic limitations [11,12], modularity has become a standard tool for community detection and several methods for modularity maximization have been developed [3]. In this paper, we focus our attention on spectral optimization of the modularity function, i.e., a maximization method that is essentially based on the determination of the principal eigenpair of the so-called modularity matrix [10]. This represents a way to approximate the configuration corresponding to the maximum of the modularity function, whose determination would be otherwise a NP complete problem [13], by relaxing the indices that assign the nodes to the various communities from integer to real valued numbers. This method provides, in general, solutions that are consistent with those obtained by other more sophisticated maximization techniques [10,14], thus the following results can be reasonably considered as valid for any type of community detection algorithm based on modularity maximization.

We consider here the simple case of a symmetric and weighted network formed only by two communities of size N. The adjacency matrices that contain the information about the internal structure of these two groups are denoted, respectively, with $A_{1,1}$ and $A_{2,2}$, while the connections among nodes of different groups are listed in the matrix $A_{1,2} = A_{2,1}^T$. The adjacency matrix of the entire network can be thus written in the following block form:

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}. \quad (1)$$

According to the definition of the modularity function, in the random annealed version of the network, which preserves on average the node degrees, the probability that two nodes are connected is proportional to the product of the degrees of the two nodes [9,15]. The entire information of this null model is contained in the square matrix

$$P = \frac{1}{\langle s_1|1\rangle + \langle s_2|1\rangle} \begin{pmatrix} |s_1\rangle\langle s_1| & |s_1\rangle\langle s_2| \\ |s_2\rangle\langle s_1| & |s_2\rangle\langle s_2| \end{pmatrix},$$

where $|s_i\rangle = |s_{i,i}\rangle + |s_{i,j}\rangle$ is the strength vector of the ith group, with $|s_{i,j}\rangle = A_{i,j}|1\rangle$ equal to a vector whose components are equal to the sum of the weights of all edges connecting nodes of the ith group to nodes of the jth group, and $|1\rangle$ is the vector whose components are all equal to one.

Let us focus our attention on the spectrum of the modularity matrix $Q = A - P$ [10]. Note that by definition $Q|1\rangle = |0\rangle$, where $|0\rangle$ is the vector with all components equal to zero, thus any other eigenvector $|v\rangle$ of the modularity matrix Q, i.e., $Q|v\rangle = \lambda|v\rangle$, must be orthogonal to the vector $|1\rangle$. We can rewrite the eigenvector $|v\rangle = |v_1,v_2\rangle$, where $|v_i\rangle$ is the part of the eigenvector $|v\rangle$ that corresponds to the ith group. The orthogonality with respect to the eigenvector $|1\rangle$ reads as $\langle v_1|1\rangle + \langle v_2|1\rangle = 0$, while the normality of the eigenvector means that $\langle v_1|v_1\rangle + \langle v_2|v_2\rangle = 1$. The eigenvalue problem becomes equivalent to

$$A_{1,1}|v_1\rangle + A_{1,2}|v_2\rangle - \frac{\langle s_1|v_1\rangle + \langle s_2|v_2\rangle}{\langle s_1|1\rangle + \langle s_2|1\rangle}|s_1\rangle = \lambda|v_1\rangle \quad (2)$$

*f.radicchi@gmail.com

and

$$A_{2,2}|v_2\rangle + A_{2,1}|v_1\rangle - \frac{\langle s_1|v_1\rangle + \langle s_2|v_2\rangle}{\langle s_1|1\rangle + \langle s_2|1\rangle}|s_2\rangle = \lambda|v_2\rangle. \tag{3}$$

If we multiply them for $\langle 1|$, and then take their difference, we obtain

$$\begin{aligned}\lambda(\langle v_1|1\rangle - \langle v_2|1\rangle) &= (\langle s_{1,1}|v_1\rangle - \langle s_{1,2}|v_1\rangle - \langle s_{2,2}|v_2\rangle + \langle s_{2,1}|v_2\rangle)\\ &\quad - \alpha(\langle s_1|v_1\rangle - \langle s_2|v_2\rangle),\end{aligned} \tag{4}$$

where we have defined $\alpha = \frac{\langle s_1|1\rangle - \langle s_2|1\rangle}{\langle s_1|1\rangle + \langle s_2|1\rangle}$. For simplicity, in the following we will consider only cases in which $\alpha \simeq 0$, i.e., cases in which both groups have a comparable total number of edges.

We define the detectability regime as the regime in which the two preimposed communities can be detected by means of modularity spectral optimization. This regime is characterized by the fact that the components of the principal eigenvector corresponding to the nodes of one of the modules have coherent signs, while the two portions of the eigenvector corresponding to different groups are opposite in sign [10]. If we suppose that these modules are uncorrelated graphs (i.e., without further internal subcommunity structure) with prescribed in- and out-strength vectors, we expect that this eigenvector is such that

$$|v_1\rangle = n_1\,(|s_{1,1}\rangle - |s_{1,2}\rangle) = n_1|\Delta s_1\rangle \tag{5}$$

and

$$|v_2\rangle = n_2\,(|s_{2,2}\rangle - |s_{2,1}\rangle) = n_2|\Delta s_2\rangle, \tag{6}$$

where n_1 and n_2 are proportionality constants, while $|\Delta s_1\rangle$ and $|\Delta s_2\rangle$ are, respectively, the vectors whose entries are given by the difference of the in and out strengths of the nodes in the groups 1 and 2. Equations (5) and (6) simply state that the coordinates of $|v_1\rangle$ and $|v_2\rangle$ are linearly proportional to the difference of the in- and out-strength vectors, a solution that appears natural if we interpret the modularity function as the stationary solution of a random walk between the two communities [16–18]. This conjecture is indeed perfectly verified in numerical estimations of the largest eigenvector of the modularity matrix (see Fig. 1 and [19]), and thus Eqs. (5) and (6) can be used as a reasonable ansatz for the solution of our problem. If we finally insert Eqs. (5) and (6) into Eq. (4), we can give an estimate of the largest eigenvalue $\lambda_{\max}$ of the modularity matrix in the regime of detectable communities, and write

$$\lambda_{\max} = \frac{1}{2}\left(\frac{\langle\Delta s_1|\Delta s_1\rangle}{\langle\Delta s_1|1\rangle} + \frac{\langle\Delta s_2|\Delta s_2\rangle}{\langle\Delta s_2|1\rangle}\right), \tag{7}$$

where we have used the orthogonality condition $\langle v_1|1\rangle + \langle v_2|1\rangle = 0$, which leads to $n_1 = -n_2\langle\Delta s_2|1\rangle/\langle\Delta s_1|1\rangle$. Note that expression (7) is valid for given strength vectors. If we instead assume that the entries of these vectors are random variates obeying the statistical distributions $P(\Delta s_1)$ and $P(\Delta s_2)$, we can write

$$\lambda_{\max} = \frac{1}{2}\left(\frac{m_2^{(1)}}{m_1^{(1)}} + \frac{m_2^{(2)}}{m_1^{(2)}}\right), \tag{8}$$

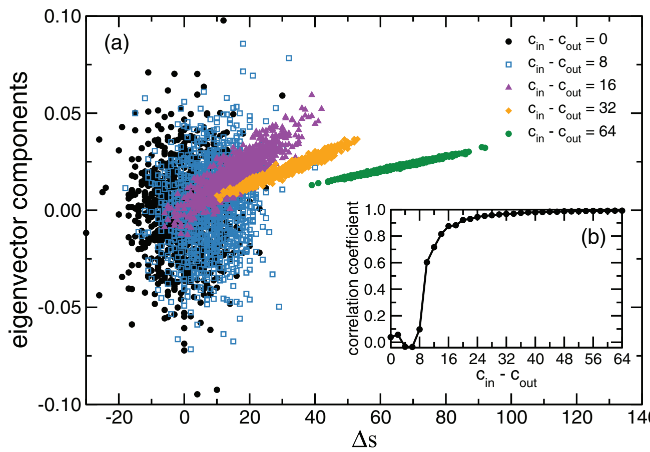

FIG. 1. (Color online) (a) Components of the principal eigenvector of the modularity matrix as functions of the difference between internal and external degrees for a system composed of two Poissonian networks of size $N = 1024$, with average internal degree $c_{\rm in}$ and average external degree $c_{\rm out}$ such that $c_{\rm in} + c_{\rm out} = 64$. We plot here only the components of one of the two modules, but analogous results (with opposite sign) are valid for the other module. Different colors and symbols correspond to different combinations of $c_{\rm in}$ and $c_{\rm out}$. (b) Correlation coefficient between the eigenvector components and the difference between internal and external degrees as a function of $c_{\rm in} - c_{\rm out}$. While the correlation coefficient is not significant in the region in which communities are not detectable [i.e., $c_{\rm in} - c_{\rm out} < \sqrt{c_{\rm in} + c_{\rm out}}$, see Eq. (12)], it becomes significant in the detectability regime (i.e., $c_{\rm in} - c_{\rm out} \geqslant \sqrt{c_{\rm in} + c_{\rm out}}$).

where $m_1^{(i)}$ and $m_2^{(i)}$ are, respectively, the first and the second moments of the distribution $P(\Delta s_i)$.

It is important to stress that Eqs. (7) and (8) give us an estimate of the largest eigenvalue of Q only in the detectability regime. If the structure of the entire graph is instead such that the two modules are not detectable by means of modularity maximization, there will be another principal eigenvector orthogonal to the previous one, and thus not showing the presence of the two modules. Since we have supposed that both modules are randomly generated graphs, the other eigenvalue that is competing with $\lambda_{\max}$ for being the highest eigenvalue of the modularity matrix is given by the second largest eigenvalue of the annealed random network associated to Q [14]. In intuitive terms, this means that, in the regime in which the groups are undetectable, the signal present in A is not sufficiently high, and Q is in spectral terms indistinguishable from P. In the following, we will consider some examples of network ensembles where both these eigenvalues can be analytically estimated, and thus the detectability problem can be explicitly solved.

Regular graphs. In this case, each node has exactly $c_{\rm in}$ random connections with other nodes in its group, and $c_{\rm out}$ random connections outside its own group. Equation (4) reduces to

$$\begin{aligned}(\lambda_{\max} - c_{\rm in} + c_{\rm out})\langle v_1|1\rangle = 0,\\ (\lambda_{\max} - c_{\rm in} + c_{\rm out})\langle v_2|1\rangle = 0,\end{aligned} \tag{9}$$

thus either (i) $\langle v_1|1\rangle = \langle v_2|1\rangle = 0$ and $\lambda_{\max} \neq c_{\rm in} - c_{\rm out}$ or (ii) $\lambda_{\max} = c_{\rm in} - c_{\rm out}$, $\langle v_1|1\rangle \neq 0$, and $\langle v_2|1\rangle \neq 0$. In case

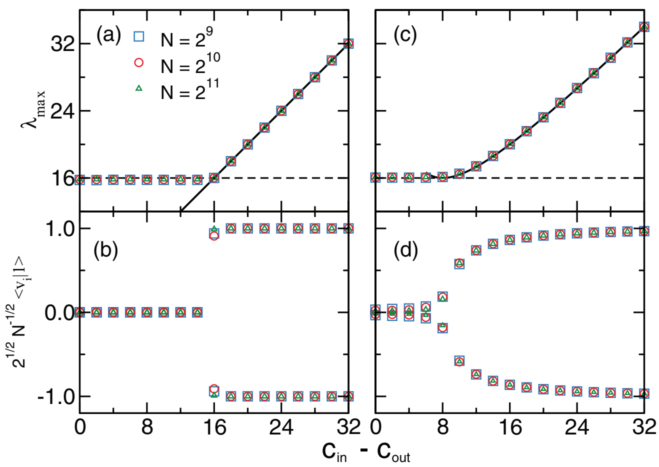

FIG. 2. (Color online) Numerical estimation of the largest eigenpair of the modularity matrix Q in the case of two random regular graphs [(a) and (b)] and in the case of two Poissonian graphs [(c) and (d)]. In both cases, we consider three different network sizes $N = 256$ (blue squares), $N = 512$ (red circles) and $N = 1024$ (green triangles), and we fix $c_{\rm in} + c_{\rm out} = 64$. The results presented here have been obtained by averaging over 100 different realizations of the models. (a), (c) Largest eigenvalue of Q as a function of $c_{\rm in} - c_{\rm out}$. The full black line corresponds to $c_{\rm in} - c_{\rm out}$ in (a), and to Eq. (11) in (c). The dashed black lines are given by $2\sqrt{c_{\rm in} + c_{\rm out}}$. (b), (d) Size-independent inner products $\langle v_1|1\rangle$ and $\langle v_2|1\rangle$ as functions of $c_{\rm in} - c_{\rm out}$.

(ii), one can also prove that the only possible solution is $|v_1\rangle = \pm|1\rangle/\sqrt{2N}$ and $|v_2\rangle = \mp|1\rangle/\sqrt{2N}$ [19]. The same result can be also obtained using Eq. (8) that reduces to Eq. (9) by setting $P(\Delta s_1) = P(\Delta s_2) = \delta(c_{\rm in} - c_{\rm out})$. The term of comparison for $\lambda_{\rm max}$ in the case of regular graphs is given by the second largest eigenvalue of the adjacency matrix of a random regular graph with valency $c = c_{\rm in} + c_{\rm out}$, that is in good approximation equal to $2\sqrt{c}$ [20,21]. Equation (9) tells us that, independently of the system size, the two modules can be either fully detectable or not detectable at all. The sudden transition between these two regimes happens at the point in which

$$c_{\rm in} - c_{\rm out} = 2\sqrt{c_{\rm in} + c_{\rm out}}. \quad (10)$$

This theoretical prediction is in perfect agreement with the results of the numerical simulations reported in Fig. 2.

Poissonian networks. This is the case whose entire spectrum has been analytically determined by Nadakuditi and Newman [14]. Internal degrees in both groups are drawn from a Poisson distribution with average $c_{\rm in}$, and external degrees are also Poissonian variates with average $c_{\rm out}$. In this case, the distributions $P(\Delta s_1)$ and $P(\Delta s_2)$ are two identical Skellam distributions, with first moment equal to $m_1 = c_{\rm in} - c_{\rm out}$, and second moment equal to $m_2 = (c_{\rm in} + c_{\rm out}) + (c_{\rm in} - c_{\rm out})^2$ [22]. We can reduce Eq. (8) to

$$\lambda_{\rm max} = c_{\rm in} - c_{\rm out} + \frac{c_{\rm in} + c_{\rm out}}{c_{\rm in} - c_{\rm out}}. \quad (11)$$

This result is identical to the prediction obtained in [14]. The term of comparison for the largest eigenvalue of the modularity matrix is given by the second largest eigenvalue of a random graph with average degree $c = c_{\rm in} + c_{\rm out}$, that is $2\sqrt{c}$ [14,23], and this finally leads to the detectability threshold

$$c_{\rm in} - c_{\rm out} = \sqrt{c_{\rm in} + c_{\rm out}}, \quad (12)$$

as already obtained in [14,24]. The results of numerical simulations perfectly agree with our theoretical prediction (see Fig. 2). The prediction appears to be not visibly dependent on the system size, and already for small networks Eq. (12) represents a very good estimate of the transition point. We note that, as in the case of regular graphs, for a large portion of the region $c_{\rm in} > c_{\rm out}$, modularity fails to recover the community structure of the graph. It is, however, interesting to stress that the detectability threshold is two times smaller than the one registered for regular graphs, and thus the heterogeneity in node degrees seems to enhance the ability of modularity to detect communities.

LFR benchmark graphs. As a final example, we consider a special case of the benchmark graphs introduced by Lancichinetti *et al.* [25]. We set $|s_{1,1}\rangle = |s_{2,2}\rangle = c_{\rm in}|s\rangle$ and $|s_{1,2}\rangle = |s_{2,1}\rangle = c_{\rm out}|s\rangle$, where the entries of the vector $|s\rangle$ are random variates in the range 1 to N taken from a power-law distribution with exponent γ. Equation (8) becomes

$$\lambda_{\rm max} = (c_{\rm in} - c_{\rm out})\, \frac{\zeta_N(\gamma - 2)}{\zeta_N(\gamma - 1)}, \quad (13)$$

where $\zeta_N(x) = \sum_{n=1}^{N} n^{-x}$ is the Riemann zeta function truncated at the Nth term. The term of comparison for the largest eigenvalue of the modularity matrix is still given by the second largest eigenvalue of the annealed random graph associated with the modularity matrix. We do not have an exact guess on how this quantity depends on parameters of the network model, but we can use the upper bound of the largest eigenvalue of random scale-free graphs to get more insights. According to the predictions by Chung *et al.* adapted to the present case, the largest eigenvalue $\mu_{\rm max}$ of our random scale-free graphs is equal to the maximum between $\mu_{\rm max}^{(1)} = (c_{\rm in} + c_{\rm out})\, \frac{\zeta_N(\gamma-2)}{\zeta_N(\gamma-1)}$ and $\mu_{\rm max}^{(2)} = \sqrt{(c_{\rm in} + c_{\rm out}) s_{\rm max}}$, with $s_{\rm max}$ the largest degree in the network [23,26]. In the limit of sufficiently large N, we have that for $\gamma < 2.5$, the dominating eigenvalue is $\mu_{\rm max}^{(1)}$; for $\gamma > 2.5$, the largest eigenvalue is instead $\mu_{\rm max}^{(2)}$. This has very important implications when compared to our prediction of $\lambda_{\rm max}$ given in Eq. (13): (i) For $\gamma < 2.5$, $\lambda_{\rm max}$ grows as fast as $\mu_{\rm max}$ with the system size, thus the detectability threshold should approach zero as N increases. (ii) For $\gamma > 2.5$ instead, $\mu_{\rm max}$ grows faster than $\lambda_{\rm max}$ as N increases. The detectability threshold should grow with the system size, and eventually converge to a finite fixed value (for instance, for $\gamma \to \infty$ we must recover the result valid for the case of regular graphs).

The results of numerical simulations support our thesis (see Fig. 3). When we plot $\lambda_{\rm max} \frac{q_1}{q_2}$ as a function of $c_{\rm in} - c_{\rm out}$, with q_1 and q_2, respectively, the first and second moments of the strength distribution of the network, we see that when $\gamma < 2.5$ this quantity slowly approaches, as N increases, the linear behavior $c_{\rm in} - c_{\rm out}$ as predicted by Eq. (13). This means that, in the limit of infinite large systems, modularity is able to detect the presence of the network blocks for every $c_{\rm in} > c_{\rm out}$.

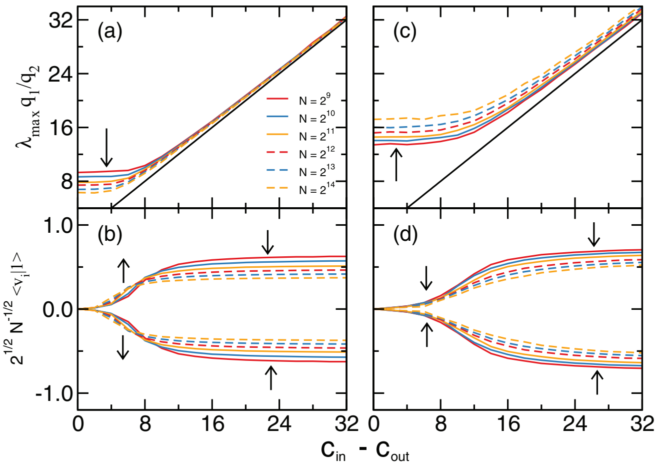

FIG. 3. (Color online) Numerical estimation of the largest eigenpair of the modularity matrix Q in the case of two random scale-free graphs with degree exponents $\gamma = 2.2$ [(a) and (b)] and $\gamma = 2.8$ [(c) and (d)]. In both cases, we consider several network sizes ranging from $N = 256$ to 8192, and we set $c_{in} + c_{out} = 64$. The results presented here have been obtained by averaging over 100 different realizations of the models. To suppress fluctuations, we restricted to the case of networks for which the square of the largest strength is smaller than the sum of all strengths (i.e., $s_{max}^2 < \sum_i s_i$) [23,26], although the qualitative outcome does not depend on this choice. For clarity, we placed arrows in the various panels to indicate the direction of increasing N. (a), (c) Largest eigenvalue λ_{max} multiplied by q_1/q_2 as a function of $c_{in} - c_{out}$. The full black line corresponds to $c_{in} - c_{out}$. (b), (d) Size-independent inner products $\langle v_1|1\rangle$ and $\langle v_2|1\rangle$ as functions of $c_{in} - c_{out}$.

Instead, for $\gamma > 2.5$, the lower part of the curve tends to move away from the linear behavior as the system size grows. This implies that there will be, also in the limit of infinitely large systems, always a part of the $c_{in} > c_{out}$ region in which the two blocks are undetectable via modularity maximization.

To summarize, we identified the necessary conditions that communities need to satisfy in order to be detectable by means of modularity maximization. Our results are valid for the case of two groups with a comparable number of edges, and when the information about the number of such groups is used as ingredient in the maximization of the modularity function. Our main result is that in random network ensembles with preimposed community structure, the eigenvector of the modularity matrix that identifies the presence of the block structure is associated with an eigenvalue approximately equal to the ratio between the second and the first moments of the distribution of the difference between internal and external node strengths. If this eigenvalue is larger than the second largest eigenvalue of the null model associated to the modularity function, then modularity is able to detect such a structure, otherwise not. This represents a limitation in the case of graphs with homogeneous degrees. Increasing the heterogeneity of the network accelerates instead the ability of modularity to recover the correct community structure. For example, adding noise to a regular graph makes the detectability threshold two times smaller. More importantly, if the heterogeneity of the node degrees is sufficiently high, as in the case of real networked systems, then modularity is always able to detect communities.

The author thanks A. Arenas, R. Darst, and S. Fortunato for helpful discussions on the subject of this article. The author acknowledges support from the Spanish Ministerio de Ciencia e Innovación through the Ramón y Cajal program.

[1] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[2] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. USA **101**, 2658 (2004).
[3] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[4] V. Spirin and L. Mirny, Proc. Natl. Acad. Sci. USA **100**, 12123 (2003).
[5] D. M. Wilkinson and B. A. Huberman, Proc. Natl. Acad. Sci. USA **101**, 5241 (2004).
[6] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[7] Y. Dourisboure, F. Geraci, and M. Pellegrini, ACM Trans. Web **3**, 1 (2009).
[8] D. B. Stouffer, M. Sales-Pardo, M. I. Sirer, and J. Bascompte, Science **335**, 1489 (2012).
[9] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[10] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).
[11] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).
[12] B. H. Good, Y. A. de Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).
[13] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE T. Knowl. Data Eng. **20**, 172 (2008).
[14] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**, 188701 (2012).
[15] M. Molloy and B. Reed, Random Struct. Algor. **6**, 161 (1995).
[16] R. Lambiotte, J. C. Delvenne, and M. Barahona, arXiv: 0812.1770.
[17] R. Lambiotte, in *WiOpt* (IEEE, Washington, DC, 2010), pp. 546–553.
[18] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, Science **328**, 876 (2010).
[19] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.88.010801 for the complete solution for the case of regular graphs and numerical estimations of the principal eigenvector components for the case of scale-free graphs.
[20] N. Alon, Combinatorica **6**, 83 (1986).
[21] J. Friedman, in *Stoc '03* (ACM, New York, NY, 2003), pp. 720–724.
[22] J. G. Skellam, J. R. Stat. Soc. **109**, 296 (1946).
[23] F. Chung, L. Lu, and V. Vu, Proc. Natl. Acad. Sci. USA **100**, 6313 (2003).
[24] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. Lett. **107**, 065701 (2011).
[25] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[26] F. Chung, L. Lu, and V. Vu, Ann. Comb. **7**, 21 (2003).