Spectra of random graphs with community structure and arbitrary degrees

Xiao Zhang,[1] Raj Rao Nadakuditi,[2] and M. E. J. Newman[1,3]
[1]*Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*
[2]*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109, USA*
[3]*Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109, USA*

(Received 20 February 2014; published 28 April 2014)

Using methods from random matrix theory researchers have recently calculated the full spectra of random networks with arbitrary degrees and with community structure. Both reveal interesting spectral features, including deviations from the Wigner semicircle distribution and phase transitions in the spectra of community structured networks. In this paper we generalize both calculations, giving a prescription for calculating the spectrum of a network with both community structure and an arbitrary degree distribution. In general the spectrum has two parts, a continuous spectral band, which can depart strongly from the classic semicircle form, and a set of outlying eigenvalues that indicate the presence of communities.

DOI: 10.1103/PhysRevE.89.042816 PACS number(s): 89.75.Hc, 02.70.Hm, 64.60.aq

I. INTRODUCTION

Spectral analysis of networks provides a useful complement to traditional analyses that focus on local network properties like degree distributions, correlation functions, or subgraph densities [1,2]. Spectral analysis can return nonlocal information about network structure such as optimal partitions [3,4], community structure [5], and nonlocal centrality measures [6] and has been widely used in the study of real-world network data since the 1970s. In additional to the development of practical algorithms and methods based on network spectra, such as spectral partitioning schemes and community detection algorithms, a considerable amount of work has been done on the analytic calculation of spectra for synthetic networks generated using random models [7–15]. Study of these model networks can help us to understand how particular features of network structure are reflected in spectra and to anticipate the performance of spectral algorithms.

Recent work on the spectra of networks with community structure, for instance, has demonstrated the presence of a "detectability threshold" as a function of the strength of the embedded structure [14]. When the community structure becomes sufficiently weak it can be shown that the spectrum loses all trace of that structure, implying that any method or algorithm for community detection based on spectral properties must fail at this transition point. A limitation of this work, however, is that the synthetic networks studied have Poisson degree distributions, which makes the calculations easier but is known to be highly unrealistic; real-world degree distributions are very far from Poissonian.

In other work a number of authors have studied the spectra of synthetic networks having broad degree distributions, such as the power-law distributions observed in many real-world networks [7–11,15]. Among other results, it is found that while the spectrum for Poisson degree distributions follows the classic Wigner semicircle law, in the more general case it departs from the semicircle, sometimes dramatically.

In this paper, we combine these two previous lines of investigation and study the spectra of networks that possess general degree distributions and simultaneously contain community structure. To do this, we make use of a recently proposed network model that generalizes the models studied before. We derive an analytic prescription for calculating the adjacency matrix spectra of networks generated by this model, which is exact in the limit of large network size and large average degree. (The opposite limit, of constant average degree, is tackled by completely different means and for a different model in Ref. [16].) In general the spectra have two components. The first is a continuous spectral band containing most of the eigenvalues but having a shape that deviates from the semicircle law seen in networks with Poisson degree distribution. The second component consists of outlying eigenvalues, outside the spectral band and normally equal in number to the number of communities in the network.

II. MODEL

The previous calculations described in the introduction make use of two classes of model networks. For networks with community structure, calculations were performed using the *stochastic block model*, in which vertices are divided into groups and edges placed between them independently at random with probabilities that depend on the group membership of the vertices involved [14,17–21]. This model gives community structure of tunable strength but vertices have a Poisson distribution of degrees within each community.

For networks without community structure but with non-Poisson degree distributions, most calculations have been performed using the so-called configuration model, a random graph conditioned on the actual degrees of the vertices [22,23], or a variant of the configuration model in which one fixes only the expected values of the degrees and not their actual values [24].

The calculations presented in this paper make use of a model proposed by Ball *et al.* [25] that simultaneously generalizes both the stochastic block model and the configuration model, so that both are special cases of the more general model. The model of Ball *et al.* is defined as follows. We assume an undirected network of n vertices labeled $i = 1 \ldots n$, each of which is associated with a q-component real vector $\mathbf{k}_i$ where q is a parameter we choose. Then the number of edges between vertices i and j is an independent, Poisson-distributed random variable with mean $\mathbf{k}_i \cdot \mathbf{k}_j / 2m$, where m is a normalizing constant given by

$$2m = \left| \sum_{i=1}^{n} \mathbf{k}_i \right|. \tag{1}$$

Physically the value of m represents the average total number of edges in the whole network. Its inclusion is merely conventional—one could easily omit it and renormalize $\mathbf{k}_i$ accordingly, and in fact Ball *et al.* did omit it in their original formulation of the model. However, including it will simplify our notation later, as well as making the connection between this model and the configuration model clearer.

The expected number of edges between vertices must be non-negative and Ball *et al.* ensured this by requiring that the elements of the vectors $\mathbf{k}_i$ all be non-negative, but this is not strictly necessary since one can always rotate the vectors globally through any angle (thereby potentially introducing some negative elements) without affecting their products $\mathbf{k}_i \cdot \mathbf{k}_j$. In this paper we will only require that all products be non-negative, which includes all cases studied by Ball *et al.* but also allows us to consider some cases they did not.

Note that it is possible in this model for there to be more than one edge between any pair of vertices (because the number of edges is Poisson distributed) and this may seem unrealistic, but in almost all real-world situations we are concerned with networks that are sparse, in the sense that only a vanishing fraction of all possible edges is present in the network, which means that $\mathbf{k}_i \cdot \mathbf{k}_j / 2m$ will be vanishing as n becomes large. We will assume this to be the case here, in which case the chances of having two or more edges between the same pair of vertices also vanishes and for practical purposes the network contains only single edges.

The average degree c of a vertex in the network is

$$c = \frac{2m}{n} = \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{k}_i \right|, \tag{2}$$

and hence increases in proportion to the average of $\mathbf{k}_i$. In this paper we will consider networks where the vectors $\mathbf{k}_i$ can have a completely general distribution, which gives us a good deal of flexibility about the structure of our network, but consider for example a network in which the vectors have arbitrary lengths, but each one points toward one of the corners of a regular q simplex in a (hyper)plane perpendicular to the direction $(1,1,1,\ldots)$. For such a choice the vectors have the form $\mathbf{k}_i = k_i \mathbf{v}_r$, where k_i is the magnitude of the vector and $\mathbf{v}_r$ is one of q unit vectors that will denote the group r that vertex i belongs to. Then

$$\mathbf{k}_i \cdot \mathbf{k}_j = k_i k_j \mathbf{v}_r \cdot \mathbf{v}_s = k_i k_j [\delta_{rs} + (1 - \delta_{rs}) \cos\phi], \tag{3}$$

where ϕ is the angle between unit vectors $\mathbf{v}_r$ and $\mathbf{v}_s$ (all vectors being separated by the same angle in a regular simplex). Thus for this choice of parametrization we can increase the expected number of edges from i to all other vertices by increasing the magnitude k_i of the vector $\mathbf{k}_i$, hence increasing the vertex's degree. At the same time we can independently control the relative probability of connections within groups (when $r = s$) and between them ($r \neq s$) by varying the angle ϕ.

If we set $\phi = 0$ [so that all $\mathbf{v}_r$ point in the $(1,1,1,\ldots)$ direction] then this model becomes equivalent to the variant of the configuration model in which the expected vertex degrees are fixed and there is probability $k_i k_j / 2m$ of connection between each pair of vertices, regardless of community membership. (Alternatively, if we set the number of groups q to 1, so that the vectors $\mathbf{k}_i$ become scalars k_i then we also recover the configuration model.) If we allow ϕ to be nonzero but make all k_i equal to the same constant value a, then the model becomes equivalent to the standard stochastic block model, having a probability $p_{\text{in}} = a^2/2m$ of connection between vertices in the same community and a smaller probability $p_{\text{out}} = (a^2/2m)\cos\phi$ between vertices in different communities. For all other choices, the model generalizes both the configuration model and the stochastic block model, allowing us to have nontrivial degrees and community structure in the same network, as well as other more complex types of structure (such as overlapping groups—see Ref. [25]).

III. CALCULATION OF THE SPECTRUM

In this section we calculate the average spectrum of the adjacency matrix $\mathbf{A}$ for networks generated from the model above, in the limit of large system size. The adjacency matrix is the symmetric matrix with elements A_{ij} equal to the number of edges between vertices i and j. The elements are Poisson independent random integers for our model, although crucially they are not identically distributed. The spectra of matrices with Poisson elements of this kind can be calculated using methods of random matrix theory. Our strategy will be first to calculate the spectrum of the matrix

$$\mathbf{X} = \mathbf{A} - \langle \mathbf{A} \rangle, \tag{4}$$

where $\langle \mathbf{A} \rangle$ is the average value of the adjacency matrix within the model, which has elements $\langle A_{ij} \rangle = \mathbf{k}_i \cdot \mathbf{k}_j / 2m$. Since $\mathbf{k}_i$ is a q-element vector, this implies that $\langle \mathbf{A} \rangle$ has rank q and hence its eigenvector decomposition has the form

$$\langle \mathbf{A} \rangle = \sum_{r=1}^{q} \alpha_r \mathbf{u}_r \mathbf{u}_r^T, \tag{5}$$

where $\mathbf{u}$ are normalized eigenvectors and α_r are the corresponding eigenvalues.

The matrix $\mathbf{X}$ is a "centered" random matrix, having independent random elements with zero mean, which makes the calculation of its spectrum particularly straightforward. Once we have calculated the spectrum of this centered matrix we will then add the rank-q term $\langle \mathbf{A} \rangle$ back in as a perturbation:

$$\mathbf{A} = \mathbf{X} + \langle \mathbf{A} \rangle. \tag{6}$$

As we will see, the only property of the centered matrix needed to compute its spectrum is the variance of its elements, and since the variance of a Poisson distribution is equal to its mean, we can immediately deduce that the variance of the ij element of $\mathbf{X}$ is $\mathbf{k}_i \cdot \mathbf{k}_j / 2m$.

A. Spectrum of the centered matrix

In this section we calculate the spectral density $\rho(z)$ of the centered matrix $\mathbf{X}$, Eq. (4). The spectral density is defined by

$$\rho(z) = \frac{1}{n} \sum_{i=1}^{n} \delta(z - \lambda_i), \tag{7}$$

where λ_i is the ith eigenvalue of $\mathbf{X}$ and $\delta(z)$ is the Dirac delta. The starting point for our calculation is the well-known Stieltjes-Perron formula, which gives the spectral density directly in terms of the matrix as

$$\rho(z) = -\frac{1}{n\pi}\,\mathrm{Im}\,\mathrm{Tr}\langle(z-\mathbf{X})^{-1}\rangle, \tag{8}$$

where $z-\mathbf{X}$ is shorthand for $z\mathbf{I}-\mathbf{X}$ with $\mathbf{I}$ being the identity.

To calculate the trace, we follow the approach of Bai and Silverstein [26], making use of the result that the ith diagonal component of the inverse of a symmetric matrix $\mathbf{B}$ is [15]

$$[\mathbf{B}^{-1}]_{ii} = \frac{1}{B_{ii} - \mathbf{b}_i^T\mathbf{B}_i^{-1}\mathbf{b}_i}, \tag{9}$$

where B_{ii} is the ith diagonal element of $\mathbf{B}$, $\mathbf{b}_i$ is the ith column of the matrix, and $\mathbf{B}_i$ is the matrix with the ith row and column removed. In the limit of large system size, and provided that the degrees of vertices become large as the network does, the distribution of values of $[\mathbf{B}^{-1}]_{ii}$ becomes narrowly peaked about its mean, and one can write the mean value as

$$\langle[\mathbf{B}^{-1}]_{ii}\rangle = \frac{1}{\langle B_{ii}\rangle - \langle\mathbf{b}_i^T\mathbf{B}_i^{-1}\mathbf{b}_i\rangle}. \tag{10}$$

If, as in our case, the elements of $\mathbf{B}$ are independent random variables with mean zero, then

$$\begin{aligned}\langle\mathbf{b}_i^T\mathbf{B}_i^{-1}\mathbf{b}_i\rangle &= \sum_{jk}\langle[\mathbf{B}_i^{-1}]_{jk}\rangle\langle[\mathbf{b}_i]_j[\mathbf{b}_i]_k\rangle\\ &= \sum_j\langle[\mathbf{B}_i^{-1}]_{jj}\rangle\langle[\mathbf{b}_i]_j^2\rangle,\end{aligned} \tag{11}$$

where we have made use of $\langle[\mathbf{b}_i]_j[\mathbf{b}_i]_k\rangle = \langle[\mathbf{b}_i]_j\rangle\langle[\mathbf{b}_i]_k\rangle = 0$ when $j\neq k$.

In our particular example we have $\mathbf{B} = z-\mathbf{X}$, which means that

$$[\mathbf{b}_i]_j = -X_{ij} \tag{12}$$

(since $i\neq j$ by definition, the ith row having been removed from the matrix), so

$$\begin{aligned}\langle\mathbf{b}_i^T\mathbf{B}_i^{-1}\mathbf{b}_i\rangle &= \sum_j\langle[\mathbf{B}_i^{-1}]_{jj}\rangle\langle X_{ij}^2\rangle\\ &= \sum_j\langle[\mathbf{B}_i^{-1}]_{jj}\rangle\frac{\mathbf{k}_i\cdot\mathbf{k}_j}{2m}\\ &= \frac{1}{2m}\mathbf{k}_i\cdot\sum_j\mathbf{k}_j\langle[(z-\mathbf{X})^{-1}]_{jj}\rangle,\end{aligned} \tag{13}$$

where the last equality applies in the limit of large system size (for which it makes a vanishing difference whether we drop the ith row and column from the matrix or not, so $\mathbf{B}_i$ can be replaced with $z-\mathbf{X}$ for all i). Then, noting that $\langle B_{ii}\rangle = z-\langle X_{ii}\rangle = z$, Eq. (10) becomes

$$\langle[(z-\mathbf{X})^{-1}]_{ii}\rangle = \frac{1}{z-\mathbf{k}_i\cdot\sum_j\mathbf{k}_j\langle[(z-\mathbf{X})^{-1}]_{jj}\rangle/2m}. \tag{14}$$

Summing this expression over i we then get the trace we were looking for, which we will write in terms of a new function

$$\begin{aligned}g(z) &= \frac{1}{n}\mathrm{Tr}\langle(z-\mathbf{X})^{-1}\rangle\\ &= \frac{1}{n}\sum_{i=1}^{n}\langle[(z-\mathbf{X})^{-1}]_{ii}\rangle\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{z-\mathbf{k}_i\cdot\mathbf{h}(z)},\end{aligned} \tag{15}$$

where we have for convenience defined the vector function

$$\mathbf{h}(z) = \frac{1}{2m}\sum_i\mathbf{k}_i\langle[(z-\mathbf{X})^{-1}]_{ii}\rangle. \tag{16}$$

The quantity $g(z)$ (which is just the trace divided by n) is called the *Stieltjes transform* of the matrix $\mathbf{X}$, and it will play a substantial role in the remainder of our calculation.

It remains to calculate the function $\mathbf{h}(z)$, which is now straightforward. Multiplying Eq. (14) by $\mathbf{k}_i$ and substituting into (16), we get

$$\mathbf{h}(z) = \frac{1}{2m}\sum_i\frac{\mathbf{k}_i}{z-\mathbf{k}_i\cdot\mathbf{h}(z)}. \tag{17}$$

The solution for the spectral density involves solving this equation for $\mathbf{h}(z)$, then substituting the answer into Eq. (15) to get the Stieltjes transform $g(z)$. Then the spectral density itself can be calculated from Eq. (8):

$$\rho(z) = -\frac{1}{\pi}\,\mathrm{Im}\,g(z). \tag{18}$$

Alternatively, we can simplify the calculation somewhat by rewriting Eq. (14) as

$$z\langle[(z-\mathbf{X})^{-1}]_{ii}\rangle - \langle[(z-\mathbf{X})^{-1}]_{ii}\rangle\,\mathbf{k}_i\cdot\mathbf{h}(z) = 1, \tag{19}$$

then summing over i and dividing by n to get $zg(z) - c\|\mathbf{h}(z)\|^2 = 1$, or

$$g(z) = \frac{1+c\|\mathbf{h}(z)\|^2}{z}, \tag{20}$$

where $c = 2m/n$ as previously, which is the average degree of the network, and $\|\mathbf{h}(z)\|$ denotes the vector magnitude of $\mathbf{h}(z)$, i.e., $\mathbf{h}\cdot\mathbf{h}$ (not the complex absolute value). Then the spectral density itself, from Eq. (18), is

$$\rho(z) = -\frac{c}{\pi z}\,\mathrm{Im}\,\|\mathbf{h}(z)\|^2. \tag{21}$$

If we further suppose that the parameter vectors $\mathbf{k}_i$ are drawn independently from some probability distribution $p(\mathbf{k})$, which plays roughly the role played by the degree distribution in other network models, then in the limit of large network size Eq. (17) can be written as

$$\mathbf{h}(z) = \frac{1}{c}\int\frac{\mathbf{k}\,p(\mathbf{k})\,\mathrm{d}^q k}{z-\mathbf{k}\cdot\mathbf{h}(z)}. \tag{22}$$

Equations (21) and (22) between them give us our solution for the spectral density. These equations can be regarded as generalizations of the equations for the configuration model given in Ref. [15] and similar equations have also appeared in applications of random matrix methods to other problems [27–31].

B. Examples

As an example of the methods of the previous section, consider a network of n vertices with two communities of $\frac{1}{2}n$ vertices each. Let the first group consist of vertices $1\ldots\frac{1}{2}n$ and the second of vertices $\frac{1}{2}n+1\ldots n$. Vertices in the first group will have parameter vector $\mathbf{k}_i=(\kappa_i,\theta)$ and those in the second group will have $\mathbf{k}_i=(\kappa_{i-n/2},-\theta)$, where the quantities κ_i and θ are positive constants that we choose and $\kappa_i\geqslant\theta$ for all i, to ensure that the expected values $\langle A_{ij}\rangle=\mathbf{k}_i\cdot\mathbf{k}_j/2m$ of the adjacency matrix elements are non-negative.

This particular parametrization is attractive for a number of reasons. First, it already takes the form of the rank-2 eigenvector decomposition of Eq. (5), which simplifies our calculations—the two (unnormalized) eigenvectors are the n-element vectors $(\boldsymbol{\kappa},\boldsymbol{\kappa})$ and $(1,1,\ldots,-1,-1,\ldots)$ where $\boldsymbol{\kappa}$ is the $(\frac{1}{2}n)$-element vector with elements $\kappa_1,\ldots,\kappa_{n/2}$. Also the expected degrees take a particularly simple form. The expected degree of vertex i for $i\leqslant\frac{1}{2}n$ is

$$\frac{1}{2m}\sum_{j=1}^{n}\mathbf{k}_i\cdot\mathbf{k}_j=\frac{1}{2m}\left[\sum_{j=1}^{n/2}(\kappa_i\kappa_j+\theta^2)+\sum_{j=n/2+1}^{n}(\kappa_i\kappa_{j-n/2}-\theta^2)\right]=\frac{\kappa_i}{m}\sum_{j=1}^{n/2}\kappa_j. \tag{23}$$

But, applying Eq. (1), we have $m=\sum_{j=1}^{n/2}\kappa_j$ and hence the expected degree of vertex i is simply κ_i. By a similar calculation it can easily be shown that for $i>\frac{1}{2}n$ the expected degree is $\kappa_{i-n/2}$, and the average degree in the whole network is

$$c=\frac{1}{n/2}\sum_{i=1}^{n/2}\kappa_i. \tag{24}$$

The parameter θ also has a simple interpretation in this model: it controls the strength of the community structure. For instance, when $\theta=0$ vertices in the two communities are equivalent and there is no community structure at all.

To calculate the spectrum for this model, we substitute the values of $\mathbf{k}_i$ into Eq. (22) to get equations for the two components of the vector function $\mathbf{h}(z)$ thus:

$$h_1(z)=\frac{1}{c}\int\kappa p(\kappa)\left[\frac{1}{z-\kappa h_1(z)-\theta h_2(z)}+\frac{1}{z-\kappa h_1(z)+\theta h_2(z)}\right]d\kappa, \tag{25}$$

$$h_2(z)=\frac{\theta}{c}\int p(\kappa)\left[\frac{1}{z-\kappa h_1(z)-\theta h_2(z)}-\frac{1}{z-\kappa h_1(z)+\theta h_2(z)}\right]d\kappa, \tag{26}$$

where $p(\kappa)$ is the probability distribution of the quantities κ_i. Equation (26) has the trivial solution $h_2(z)=0$, so the two equations simplify to a single one:

$$h_1(z)=\frac{1}{c}\int\frac{\kappa p(\kappa)\,d\kappa}{z-\kappa h_1(z)}, \tag{27}$$

and then

$$\rho(z)=-\frac{c}{\pi z}\,\mathrm{Im}\,h_1^2(z), \tag{28}$$

which is independent of the parameter θ. These results are identical to those for the corresponding quantities in the ordinary configuration model with no community structure and expected degree distribution $p(\kappa)$, as derived in Ref. [15], and hence we expect the spectrum of the centered adjacency matrix to be the same for the current model as it is for the configuration model with the same distribution of expected degrees.

To give a simple example application, suppose that there are only two different values of κ. Half the vertices in each community have a value κ_1 and the other half κ_2. Then $p(\kappa)=\frac{1}{2}[\delta(\kappa-\kappa_1)+\delta(\kappa-\kappa_2)]$, where $\delta(x)$ is the Dirac delta, and $c=\frac{1}{2}(\kappa_1+\kappa_2)$. With this choice

$$h_1(z)=\frac{1}{\kappa_1+\kappa_2}\left[\frac{\kappa_1}{z-\kappa_1h_1(z)}+\frac{\kappa_2}{z-\kappa_2h_1(z)}\right], \tag{29}$$

which can be rearranged to give the cubic equation:

$$\kappa_1\kappa_2h_1^3-(\kappa_1+\kappa_2)zh_1^2+\left[\frac{2\kappa_1\kappa_2}{\kappa_1+\kappa_2}+z^2\right]h_1-z=0, \tag{30}$$

which can be solved exactly for $h_1(z)$ and hence we can derive an exact expression for the spectral density. The expression itself is cumbersome (like the solutions of most cubic equations), but Fig. 1 shows an example for the choice $\kappa_1=60$, $\kappa_2=120$, along with numerical results for the spectrum of a single random realization of the model. As the figure shows, the two agree well. (The histogram in the left-hand part of the figure represents the spectrum of the centered matrix. The two outlying eigenvalues that appear to the right belong to the full, noncentered adjacency matrix and are calculated in the following section.)

Note also that in the special case where $\kappa_1=\kappa_2=c$, so that κ is constant over all vertices, Eq. (29) simplifies further to

$$h_1(z)=\frac{1}{z-ch_1(z)}, \tag{31}$$

which is a quadratic equation with solutions

$$h_1(z)=\frac{z\pm\sqrt{z^2-4c}}{2c}, \tag{32}$$

and hence the spectral density is

$$\rho(z)=\frac{\sqrt{4c-z^2}}{2\pi c}, \tag{33}$$

where we take the negative square root in Eq. (32) to get a positive density. Equation (33) has the form of the classic semicircle distribution for random matrices. This model is equivalent to the standard stochastic block model and (33) agrees with the expression for the spectral density derived for that model by other means in Ref. [14].

C. Spectrum of the adjacency matrix

So far we have derived the spectral density of the centered adjacency matrix $\mathbf{X}=\mathbf{A}-\langle\mathbf{A}\rangle$. We can use the results of these

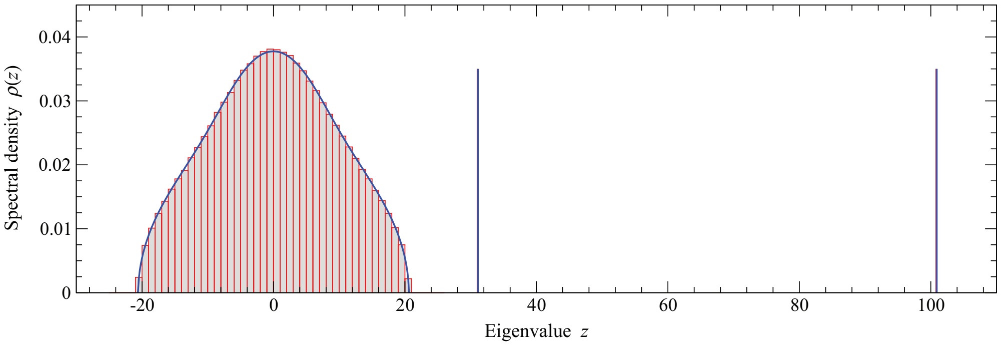

FIG. 1. (Color online) The spectrum of the adjacency matrix for the case of a network with two groups of equal size and $\mathbf{k}_i = (\kappa_i, \pm\theta)$, where $\theta = 50$, $\kappa_{i+n/2} = \kappa_i$, and κ_i, is either 60 or 120 with equal probability. Blue represents the analytic solution, Eqs. (29) and (39). Red is the numerical diagonalization of the adjacency matrix of a single network with $n = 10\,000$ vertices generated from the model with the same parameters. The numerically evaluated positions of the two outlying eigenvalues (the red spikes) agree so well with the analytic values (blue spikes) that the red is mostly obscured behind the blue.

calculations to compute the spectrum of the full adjacency matrix by generalizing the method used in Ref. [15], as follows.

Using Eq. (5) we can write the adjacency matrix as

$$\mathbf{A} = \mathbf{X} + \langle\mathbf{A}\rangle = \mathbf{X} + \sum_{r=1}^{q} \alpha_r \mathbf{u}_r \mathbf{u}_r^T. \tag{34}$$

Let us first consider the effect of adding just one of the terms in the sum to the centered matrix $\mathbf{X}$, calculating the spectrum of the matrix $\mathbf{X} + \alpha_1 \mathbf{u}_1 \mathbf{u}_1^T$. Let $\mathbf{v}$ be an eigenvector of this matrix with eigenvalue z:

$$\big(\mathbf{X} + \alpha_1 \mathbf{u}_1 \mathbf{u}_1^T\big)\mathbf{v} = z\mathbf{v}. \tag{35}$$

Rearranging this equation we have $\alpha_1 \mathbf{u}_1 \mathbf{u}_1^T \mathbf{v} = (z - \mathbf{X})\mathbf{v}$ and, multiplying by $\mathbf{u}_1^T (z - \mathbf{X})^{-1}$, we find

$$\mathbf{u}_1^T (z - \mathbf{X})^{-1} \mathbf{u}_1 = \frac{1}{\alpha_1}. \tag{36}$$

Note that the vector $\mathbf{v}$ has canceled out of the equation, leaving us with an equation in z alone. The solutions for z of this equation give us the eigenvalues of the matrix $\mathbf{X} + \alpha_1 \mathbf{u}_1 \mathbf{u}_1^T$.

Expanding the vector $\mathbf{u}_1$ as a linear combination of the eigenvectors $\mathbf{x}_i$ of the matrix $\mathbf{X}$, the equation can also be written in the form

$$\sum_{i=1}^{n} \frac{\big(\mathbf{x}_i^T \mathbf{u}_1\big)^2}{z - \lambda_i} = \frac{1}{\alpha_1}, \tag{37}$$

where λ_i are the eigenvalues of $\mathbf{X}$. Figure 2 shows a graphical representation of the solution of this equation for the eigenvalues z. The left-hand side of the equation, represented by the solid curves, has simple poles at $z = \lambda_i$ for all i. The right-hand side, represented by the horizontal dashed line, is constant. Where the two intersect, represented by the dots, are the solutions for z. From the geometry of the figure we can see that the values of z must fall between consecutive values of λ_i—we say that the z's and λ's are *interlaced*. If we number the eigenvalues λ_i in order from largest to smallest so that $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, and similarly for the n solutions z_i to Eq. (37), then $z_1 \geqslant \lambda_1 \geqslant z_2 \geqslant \lambda_2 \geqslant \cdots \geqslant z_n \geqslant \lambda_n$. In the limit of large system size, as the λ_i become more and more closely spaced in the spectrum of the matrix, this interlacing places tighter and tighter bounds on the values of z_i, and asymptotically we have $z_i = \lambda_i$ and the spectral density of $\mathbf{X} + \alpha_1 \mathbf{u}_1 \mathbf{u}_1^T$ is the same as that of $\mathbf{X}$ alone.

There is one exception, however, in the highest-lying eigenvalue z_1, which is bounded below by λ_1 but unbounded above, meaning it need not be equal to λ_1 and may lie outside the band of values occupied by the spectrum of the matrix $\mathbf{X}$. To calculate this eigenvalue we observe that the matrix $\mathbf{X}$ being random, its eigenvectors $\mathbf{x}_i$ are also random and hence $\mathbf{x}_i^T \mathbf{u}_1$ is a zero-mean random variable with variance $1/n$. Taking the average of Eq. (37) over the ensemble of networks, the numerator on the left-hand side gives simply a factor of $1/n$ and we have

$$\frac{1}{\alpha_1} = \frac{1}{n}\left\langle \sum_{i=1}^{n} \frac{1}{z - \lambda_i} \right\rangle = \frac{1}{n}\langle \mathrm{Tr}(z - \mathbf{X})^{-1} \rangle = g(z). \tag{38}$$

The solution to this equation gives us the value of z_1.

This then gives us the complete spectrum for the matrix $\mathbf{X} + \alpha_1 \mathbf{u}_1 \mathbf{u}_1^T$. It consists of a continuous spectral band with spectral density equal to that of the matrix $\mathbf{X}$ alone, which is calculated from Eq. (21), plus a single eigenvalue outside the band whose value is the solution for z of $g(z) = 1/\alpha_1$.

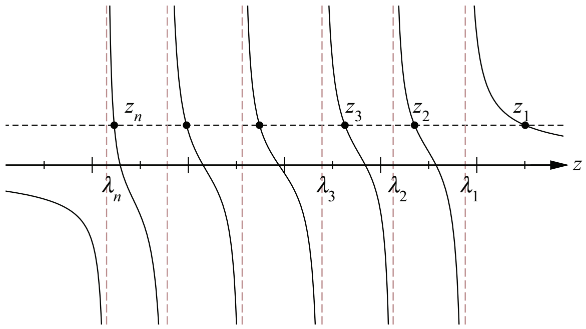

FIG. 2. (Color online) A plot of the left-hand side of Eq. (37) as a function of z has simple poles at $z = \lambda_i$ for all i. The solutions of the equation fall at the points where the curve crosses the horizontal dashed line representing the value of $1/\alpha_1$. From the geometry of the figure we can see that the solutions must lie in between the values of the λ_i, interlacing with them, so that $z_1 \geqslant \lambda_1 \geqslant z_2 \geqslant \cdots \geqslant z_n \geqslant \lambda_n$.

We could have made the same argument about any single term $\alpha_r \mathbf{u}_r \mathbf{u}_r^T$ appearing in Eq. (34) and derived the corresponding result that the continuous spectral band is unchanged from the centered matrix but there can be an outlying eigenvalue z_r given by

$$g(z_r) = \frac{1}{\alpha_r}. \tag{39}$$

The calculation of the spectrum of the full adjacency matrix requires that we consider all terms in Eq. (34) simultaneously, but in practice it turns out that it is enough to consider them one by one using Eq. (39). The argument for this is in two parts as follows.

(1) We have shown that the spectral density of the continuous band in the spectrum of the matrix $\mathbf{X} + \alpha_1 \mathbf{u}_1 \mathbf{u}_1^T$ is the same as that for the matrix $\mathbf{X}$ alone, and there is one additional outlying eigenvalue, which we denote z_1. Now we can add another term $\alpha_2 \mathbf{u}_2 \mathbf{u}_2^T$ and repeat our argument for the matrix $\mathbf{X} + \alpha_1 \mathbf{u}_1 \mathbf{u}_1^T + \alpha_2 \mathbf{u}_2 \mathbf{u}_2^T$, finding the equivalent of Eq. (37) to be

$$\sum_{i=2}^{n} \frac{(\mathbf{x}_i^T \mathbf{u}_2)^2}{z' - z_i} + \frac{(\mathbf{x}_1^T \mathbf{u}_2)^2}{z' - z_1} = \frac{1}{\alpha_2}, \tag{40}$$

where z' is the eigenvalue of the new matrix and z_i are the solutions of (37). As before, this implies there is an interlacing condition and that the spectral density of the perturbed matrix is the same within the spectral band as that for the unperturbed matrix. We can repeat this argument as often as we like and thus demonstrate that the shape of the spectral band never changes, so long as the number of perturbations (which is also the rank of $\langle \mathbf{A} \rangle$) is small compared to the size of the network, i.e., $q \ll n$.

(2) This argument pins down all but the top two eigenvalues of $\mathbf{X} + \alpha_1 \mathbf{u}_1 \mathbf{u}_1^T + \alpha_2 \mathbf{u}_2 \mathbf{u}_2^T$. These two we can calculate by a variant of our previous argument. We average Eq. (40) over the ensemble, noting again that $\langle (\mathbf{x}_i^T \mathbf{u}_2)^2 \rangle = 1/n$ and find that

$$\frac{1}{n} \sum_{i=2}^{n} \frac{1}{z' - z_i} + \frac{1/n}{z' - z_1} = \frac{1}{\alpha_2}. \tag{41}$$

For large n the first sum is once again equal to the Stieltjes transform $g(z')$ and hence the top two eigenvalues are solutions for z' of

$$g(z') + \frac{1/n}{z' - z_1} = \frac{1}{\alpha_2}. \tag{42}$$

But $g(z')$ and α_2 are of order 1, while the term $n^{-1}/(z' - z_1)$ is of order $1/n$ and hence can in most circumstances be neglected, giving $g(z') = 1/\alpha_2$, which recovers Eq. (39). The only time this term cannot be neglected is when z' is within a distance of order $1/n$ from z_1, in which case we have a simple pole in the left-hand side of the equation as z' approaches z_1. Thus the left-hand side has the form sketched in Fig. 3, following $g(z)$ closely for most values of z, but diverging suddenly when very close to z_1. Equation (42) then has two solutions, as indicated by the dots in the figure, one given by $g(z) = 1/\alpha_2$ and one that is asymptotically equal to z_1, which is the solution of $g(z) = 1/\alpha_1$.

We can repeat this argument as many times as we like to demonstrate that the outlying eigenvalues are just the q solutions of Eq. (39) for each value $r = 1 \ldots q$. Thus our final solution for the complete spectrum of the adjacency matrix has two parts: a continuous spectral band, given by Eqs. (21) and (22), and q outlying eigenvalues, given by the solutions of Eq. (39), with $g(z)$ given by Eq. (20).

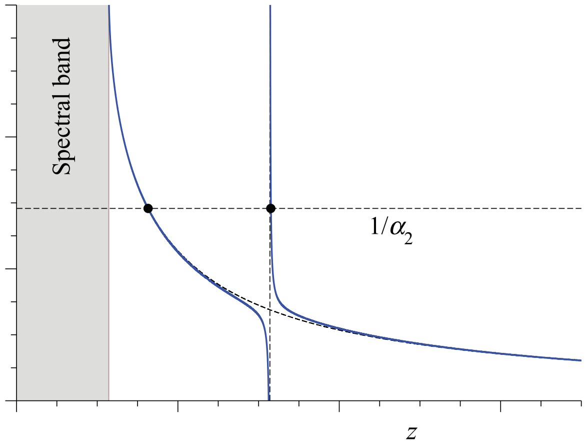

FIG. 3. (Color online) A graphical representation of the solution of Eq. (42). The left-hand side of the equation, represented by the solid blue curve, follows closely the form of the Stieltjes transform $g(z)$, except within a distance of order $1/n$ from z_1, where it diverges. The horizontal dashed line represents the value $1/\alpha_2$ and the solutions to (42), of which there are two, fall at the intersection of this line with the solid curve, as indicated by the dots. One of these solutions coincides closely with z_1, the other is the solution of $g(z) = 1/\alpha_2$.

D. Examples

Let us return to the examples of Sec. III B and apply the methods above to the calculation of their outlying eigenvalues. Recall that we looked at networks with two communities and chose parameter vectors $\mathbf{k}_i = (\kappa_i, \theta)$ for vertices in the first community and $\mathbf{k}_i = (\kappa_{i-n/2}, -\theta)$ for those in the second. For such networks the vector function $\mathbf{h}(z)$ reduces to a single scalar function $h_1(z)$ that satisfies Eq. (27). At the same time, Eq. (20) tells us that for this model $zg(z) = 1 + ch_1^2(z)$ and hence from Eq. (39) the positions of the outlying eigenvalues are solutions of

$$1 + ch_1^2(z) - \frac{z}{\alpha_r} = 0, \tag{43}$$

for $r = 1 \ldots q$. Locating the outliers is thus a matter of solving (27) for h_1, substituting the result into (43), and then solving for z.

Consider, for instance, the choice we made in Sec. III B, where there were just two values of κ, denoted κ_1 and κ_2, with half the vertices in each community taking each value. Then h_1 obeys the cubic equation (30), which can be solved exactly, and hence we can calculate the position of the outliers. Figure 1 shows the results for the choice $\kappa_1 = 60$, $\kappa_2 = 120$, $\theta = 50$, along with numerical results for the same parameter values. As the figure shows, analytic and numerical calculations again agree well—so well, in fact, that the difference between them is quite difficult to make out on the plot.

We also looked in Sec. III B at the simple case where $\kappa = c$ for all vertices, so that they all have the same expected degree, in which case the model becomes equivalent to the standard stochastic block model and the continuous spectral band takes the classic semicircle form of Eq. (33). For this model we have $\alpha_1 = c$ and $\alpha_2 = \theta^2/c$. Using Eq. (32) for $h_1(z)$ and solving (43) for z, we then find the top two eigenvalues of the adjacency matrix to be

$$z_1 = c + 1, \quad z_2 = \frac{\theta^2}{c} + \frac{c^2}{\theta^2}, \tag{44}$$

which agrees with the results given previously for the stochastic block model in Ref. [14].

E. Detectability of communities

One of the primary uses of network spectra is for the detection of community structure [5,14]. As we have seen, the number of eigenvalues above the edge of the spectral band is equal to the number of communities in the network, and hence the observation of these eigenvalues can be taken as evidence of the presence of communities and their number as an empirical measure of the number of communities. The identity of the communities themselves—which vertices belong to which community—can be deduced, at least approximately, by looking at the elements of the eigenvectors [5].

However, as shown previously in Ref. [14] for the simplest two-community block model, the position of the leading eigenvalues varies as one varies the strength of community structure, and for sufficiently low (but still nonzero) strength an eigenvalue may meet the edge of the spectral band and hence become invisible in the spectrum, meaning it can no longer be used as evidence of the presence of community structure. Moreover, as also shown in Ref. [14], the elements of the corresponding eigenvector become uncorrelated with group membership at this point, so that any algorithm which identifies communities by examining the eigenvector elements will fail. The point where this happens, at least in the simple two-community model, coincides with the known “detectability threshold” for community structure, at which it is believed all algorithms for community detection must fail [19–21].

We expect qualitatively similar behavior in the present model as well. Consider the Stieltjes transform $g(z)$ defined in Eq. (15). Inside the spectral band the transform is complex by definition—see Eq. (18). Above the band it is real and monotonically decreasing in z, as we can see by evaluating the trace in the basis in which $\mathbf{X}$ is diagonal:

$$g(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{z-\lambda_i}, \tag{45}$$

where λ_i are the eigenvalues of $\mathbf{X}$ as previously. Above the band, where $z > \lambda_i$ for all i, every term in this sum is monotonically decreasing, and hence so is $g(z)$. This implies via Eq. (39) that larger values of α_r give larger eigenvalues and that the largest real value $g_{\max}$ of the Stieltjes transform occurs exactly at the band edge. Moreover, as shown in Ref. [15], the edge of the band is marked generically by a square-root singularity in the spectral density, which implies that $g_{\max}$ is finite—see Fig. 3 for a sketch of the function. Thus when we make the community structure in the network weaker, meaning we decrease the values of the α_r, we also decrease the outlying eigenvalues of the adjacency matrix and eventually the lowest of those eigenvalues will meet the edge of the band and disappear at the point where $1/\alpha_r = g_{\max}$. If we continue to weaken the structure, more eigenvalues will disappear, in order—smallest first, then second smallest, and so forth.

Thus we expect there to be a succession of detectability transitions in the network, $q - 1$ of them in all, where q again is the number of communities. At the first of these transitions the qth largest eigenvalue will meet the band edge and disappear, meaning there will only be $q - 1$ outlying eigenvalues left and hence there will be observational evidence of only $q - 1$ communities in the network, even if in fact we know there to be q. At the next transition the number will decrease further to $q - 2$, and so forth. One thus loses the ability to detect community structure in stages, one community at a time. Final evidence of any structure at all disappears at the point where the second largest eigenvalue meets the band edge.

Consider, for instance, the example network from Sec. III B again, in which there are two groups with parameter vectors of the form $(\kappa_i, \pm\theta)$, where the parameters κ_i control the expected degrees and θ controls the strength of the community structure. As before, let us study the case where the κ_i take just two different values with equal probability, so that h_1 satisfies the cubic equation (30) (and $h_2 = 0$). Then we can calculate the maximal real value of $g(z)$ as follows.

Like $g(z)$, the function $h_1(z)$ is real outside the continuous spectral band but complex inside it, as one can see from Eq. (28). The band edge is thus the point at which the solution of the cubic equation becomes complex, which is given by the zero of the discriminant of the cubic. Take, for example, the case where $\kappa_1 = \kappa$ and $\kappa_2 = 2\kappa$ for some constant κ. Then, employing the standard formula, the discriminant of (30) is

$$\frac{\kappa^5}{27}\left[27\left(\frac{z^2}{\kappa}\right)^3 - 216\left(\frac{z^2}{\kappa}\right)^2 + 252\left(\frac{z^2}{\kappa}\right) - 512\right]. \tag{46}$$

This is zero when, and hence the band edge falls at, $z = \sqrt{x\kappa}$, where $x \simeq 7.058$ is the sole real solution of the cubic equation $27x^3 - 216x^2 + 252x - 512 = 0$. Substituting into Eq. (30), we then find that the value of h_1 at the band edge is $y/\sqrt{\kappa}$ where $y = 0.723$ is the smallest real solution of the cubic equation $2y^3 - 3\sqrt{x}y^2 + (x + \frac{4}{3})y - \sqrt{x} = 0$. Then, using Eq. (20) and the fact that the average degree is $c = \frac{3}{2}\kappa$, the value of $g(z)$ at the band edge is

$$g_{\max} = \frac{2 + 3y^2}{2\sqrt{x\kappa}}. \tag{47}$$

In this case there is only one parameter α_r with $r \geqslant 2$, which is $\alpha_2 = \theta^2/c$. Hence there is a single threshold at which we lose the ability to detect communities, falling at

$$\frac{c}{\theta^2} = \frac{2 + 3y^2}{2\sqrt{x\kappa}}, \tag{48}$$

or

$$\theta = \sqrt{\frac{3\sqrt{x\kappa^3}}{2 + 3y^2}} \simeq 1.494\kappa^{3/4}. \tag{49}$$

If θ is smaller than this value then spectral methods will fail to detect the communities in the network. We have checked this behavior numerically and find indeed that spectral community detection fails at approximately this point.

IV. CONCLUSIONS

In this paper we have given a prescription for calculating the spectrum of the adjacency matrix of an undirected random network containing both community structure and a nontrivial degree distribution, generated using the model of Ball *et al.* [25]. In the limit of large network size the spectrum consists in general of two parts: (1) a continuous spectral band containing the bulk of the eigenvalues and (2) q outlying eigenvalues above the spectral band, where q is the number of communities in the network. We give expressions for both the shape of the band and the positions of the outlying eigenvalues that are exact in the limit of a large network and large vertex degrees, although their evaluation involves integrals that may not be analytically tractable in practice, in which case we must resort to numerical evaluation. We have compared the spectra calculated using our method with direct numerical diagonalizations and find the agreement to be excellent. Based on our results we also argue that there should be a series of $q-1$ "detectability transitions" as the community structure gets weaker, at which one's ability to detect communities becomes successively impaired. The positions of these transitions correspond to the points at which the outlying eigenvalues meet the edge of the spectral band and disappear. With the disappearance of the second-largest eigenvalue in this manner, all trace of the community structure vanishes from the spectrum and the network is indistinguishable from an unstructured random graph.

ACKNOWLEDGMENTS

The authors thank C. Moore for useful comments. This work was funded in part by the National Science Foundation under Grants No. CCF-1116115 and No. DMS-1107796, by the Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) under Grant No. FA9550-12-1-0432, by AFOSR Young Investigator Award FA9550-12-1-0266, by ONR Young Investigator Award N000141110660, by the Army Research Office under MURI Grant No. W911NF-11-1-0391, and by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002. Opinions, interpretations, recommendations, and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

[1] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[3] M. Fiedler, Czech. Math. J. **23**, 298 (1973).
[4] A. Pothen, H. Simon, and K.-P. Liou, SIAM J. Matrix Anal. Appl. **11**, 430 (1990).
[5] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[6] P. F. Bonacich, Am. J. Sociol. **92**, 1170 (1987).
[7] I. J. Farkas, I. Derényi, A.-L. Barabási, and T. Vicsek, Phys. Rev. E **64**, 026704 (2001).
[8] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. E **64**, 051903 (2001).
[9] F. Chung, L. Lu, and V. Vu, Proc. Natl. Acad. Sci. USA **100**, 6313 (2003).
[10] S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. E **68**, 046109 (2003).
[11] R. Kühn, J. Phys. A **41**, 295002 (2008).
[12] S. Chauhan, M. Girvan, and E. Ott, Phys. Rev. E **80**, 056114 (2009).
[13] T. Rogers, C. Pérez Vicente, K. Takeda, and I. Pérez Castillo, J. Phys. A **43**, 195002 (2010).
[14] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**, 188701 (2012).
[15] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. E **87**, 012803 (2013).
[16] R. Kühn and J. van Mourik, J. Phys. A **44**, 165205 (2011).
[17] P. W. Holland, K. B. Laskey, and S. Leinhardt, Soc. Networks **5**, 109 (1983).
[18] A. Condon and R. M. Karp, Random Struct. Algorithms **18**, 116 (2001).
[19] J. Reichardt and M. Leone, Phys. Rev. Lett. **101**, 078701 (2008).
[20] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, Phys. Rev. Lett. **107**, 065701 (2011).
[21] D. Hu, P. Ronhovde, and Z. Nussinov, Philos. Mag. **92**, 406 (2012).
[22] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).
[23] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[24] F. Chung and L. Lu, Proc. Natl. Acad. Sci. USA **99**, 15879 (2002).
[25] B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **84**, 036103 (2011).
[26] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. (Springer, Berlin, 2010).
[27] S. Molchanov, L. Pastur, and A. Khorunzhii, Theor. Math. Phys. **90**, 108 (1992).
[28] D. Shlyakhtenko, Int. Math. Res. Not. **1996**, 1013 (1996).
[29] G. Anderson and O. Zeitouni, Probab. Theor. Relat. Fields **134**, 283 (2006), proposition 3.4.
[30] G. Casati and V. Girko, Random Oper. Stoch. Eqs. **1**, 279 (1993).
[31] Z. Bai and L. Zhang, J. Multivariate Anal. **101**, 1927 (2010).