Eigenvector dynamics under perturbation of modular networks

Somwrita Sarkar*
Design Lab, Faculty of Architecture, Design and Planning, University of Sydney, Australia NSW 2006 and ARC Centre of Excellence for Integrative Brain Function

Sanjay Chawla
Qatar Computing Research Institute, Hamad Bin Khalifa University (HBKU), Qatar Foundation, Doha, Qatar and Faculty of Engineering and IT, University of Sydney, Australia NSW 2006

P. A. Robinson
School of Physics, University of Sydney, Australia NSW 2006 and ARC Centre of Excellence for Integrative Brain Function

Santo Fortunato
Department of Computer Science, Aalto University, Finland

(Received 23 October 2015; revised manuscript received 27 April 2016; published 20 June 2016; corrected 30 June 2016)

Rotation dynamics of eigenvectors of modular network adjacency matrices under random perturbations are presented. In the presence of q communities, the number of eigenvectors corresponding to the q largest eigenvalues form a "community" eigenspace and rotate together, but separately from that of the "bulk" eigenspace spanned by all the other eigenvectors. Using this property, the number of modules or clusters in a network can be estimated in an algorithm-independent way. A general argument and derivation for the theoretical detectability limit for sparse modular networks with q communities is presented, beyond which modularity persists in the system but cannot be detected. It is shown that for detecting the clusters or modules using the adjacency matrix, there is a "band" in which it is hard to detect the clusters even before the theoretical detectability limit is reached, and for which the theoretically predicted detectability limit forms the sufficient upper bound. Analytic estimations of these bounds are presented and empirically demonstrated.

DOI: 10.1103/PhysRevE.93.062312

I. INTRODUCTION

Networks have community structure, i.e., groups of nodes with significantly higher internal link density than the density of links joining the groups. Community detection, the problem of correctly estimating the number of communities and their constitution, has attracted significant attention in physics, applied mathematics, and computer science [1,2]. Accurate solutions enhance the understanding of the relationships between network structure and dynamics.

Spectral methods, employing the eigenvectors and eigenvalues of the adjacency, Laplacian, and modularity matrices [1–8], are widely used to identify communities. While the behavior of eigenvalues is widely studied [1,2,5–7,9,10], there is less work on understanding how eigenvectors behave under variations in network structure, even though it is the eigenvector properties that are used to perform community detection. In the present paper, the focus is on the behavior of the eigenvectors of the adjacency matrix and the relationship of this behavior to gaps between eigenvalues of the adjacency matrix. It is to be expected that similar results will hold for Laplacian and modularity matrices also.

A related problem is the algorithm-independent determination of the number of communities, a parameter that many detection methods need as input. Other methods estimate this number, but several runs of the same algorithm on even the same data set can return different numbers and constitutions of communities. The performance of several algorithms in determining this number has been measured [11], and its *a priori* knowledge improves their performance significantly. Algorithm-independent techniques and analytic understanding of systems to determine this number are thus beneficial.

One algorithm-independent way of determining the number of modules is to count the number of eigenvalues q separated from the bulk eigenvalues of a suitable matrix representation [7,9,10,12]. However, for networks with broad distributions of node degree, and numbers and sizes of communities, the eigenvalues can show highly variable behavior. For example, large eigenvalues can reflect both high degree nodes as well as the number of modules. Further, as mentioned above, even though the formal identification of modules is performed based on the properties of the corresponding eigenvectors [1,2,5], the overall behavior of eigenvectors under variations in network structure is much less understood than that of the eigenvalues [1,2,9,13]. Therefore, this warrants further attention to the structure of eigenvectors.

A. Contributions

We investigate rotations of eigenvectors of the adjacency matrix when the network is randomly perturbed: this rotation behavior is dependent on the gaps between eigenvalues of the adjacency matrix and contains accurate community structure information. The first main result of the paper is that in the presence of q communities, the number of eigenvectors corresponding to the q largest eigenvalues form a "community" eigenspace and rotate together, but separately from that of

*somwrita.sarkar@sydney.edu.au

the "bulk" eigenspace spanned by all the other eigenvectors. Using this property, the number of modules or clusters in a network can be estimated in an algorithm-independent way. We investigate this behavior right to the theoretical detectability limit, beyond which modularity persists in the system but cannot be detected [7,9]. The second contribution of the paper is that we present a general derivation of the theoretical detectability limit for q communities, using arguments about upper and lower bounds on the eigenvalues, that was previously shown for the $q = 2$ case [9]. Third, again using the same bounds, we show that for detecting the clusters or modules using the adjacency matrix, there is a "band" in which it is hard to detect the clusters even before the theoretical detectability limit is reached, and for which the theoretically predicted detectability limit forms the upper bound. Analytic estimations of these bounds are presented and empirically demonstrated.

II. BACKGROUND

A symmetric adjacency matrix A represents an undirected graph G with N nodes, with $A_{ij} = 1$ if an edge exists between nodes i and j, and 0 otherwise. A random Bernoulli perturbation E, by definition, is a matrix with half of its entries set to $+1$ and half to -1. E can be either symmetric or asymmetric, but for this paper, we assume a symmetric form: since A is symmetric (undirected graph), we would like $A + E$ to be symmetric. Since A is a simple undirected graph (i.e., no self-loops), we also have the diagonal of E set to 0. E is scaled by a small number ϵ to control the size of the perturbation, and we construct the perturbed matrix $A + \epsilon E$. In practical implementations, we have used a range of values for ϵ, varying it from 0.01 to 0.2 (the figures show results at $\epsilon = 0.05$, for example). The only care to take while choosing ϵ would be that the noise should not be so large as to override the signal. With this condition satisfied, these results hold for any chosen ϵ. We study the rotation of eigenvectors under perturbation, i.e., the angles between an eigenvector of A and the corresponding one in $A + \epsilon E$.

The eigenvalues of A, are arranged as $z_1 \geqslant z_2 \geqslant \cdots \geqslant z_N$ to define gaps $\Delta_i = z_i - z_{i+1}$. The Davis-Kahan-Wedin theorem [13–15] imposes an upper bound on the sine of the angle between v_1 and v_1':

$$\sin\angle(v_1, v_1') \leqslant \frac{2\epsilon||E||}{\Delta_1}, \tag{1}$$

where v_1 and v_1' are the first eigenvectors of the original and perturbed matrices, respectively, and $||E||$ is the spectral norm. When $\Delta_1 \leqslant 2\epsilon||E||$, the theorem is trivially true, as the sine function is bounded above by 1. Thus, $\Delta_1 > 2\epsilon||E||$ for all nontrivial results. If E is symmetric, with mean 0 and unit variance, $||E|| = 2\sqrt{N}$ [16].

A critical point to note before moving ahead is that the behavior of the angle between the eigenvectors of A and those of $A + \epsilon E$ could be discontinuous and are dependent principally on the gaps between eigenvalues of A. Consider this small example [15]. Let

$$A = \begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1-\epsilon \end{bmatrix}, \tag{2}$$

and let

$$\epsilon E = \begin{bmatrix} -\epsilon & \epsilon \\ \epsilon & \epsilon \end{bmatrix}, \tag{3}$$

in which case we get

$$A + \epsilon E = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix}. \tag{4}$$

Now, it is easy to see that while the eigenvalues of A and $A + \epsilon E$ are the same, $1 + \epsilon$ and $1 - \epsilon$, the eigenvectors of A are [0,1] and [1,0], but the eigenvectors of $A + \epsilon E$ are $[1/\sqrt{2}, 1/\sqrt{2}]$ and $[1/\sqrt{2}, -1/\sqrt{2}]$, regardless of how small ϵ is. Thus, it turns out that the behavior of the eigenvectors under perturbation, and the identification of the number of communities, are dependent on the gaps between the eigenvalues. If these gaps are very small, the rotations could be discontinuous and large.

Recently these bounds were improved for matrices of low rank [13,15,17]. The intuition is that if A has low rank structure, the action of E on A will also occur in a lower rank subspace. Thus, $||E|| = O(\sqrt{N})$ [Eq. (1)] can be replaced by a dependence on the rank q of A because $q < O(\sqrt{N})$, leading to tighter bounds on the rotation of eigenvectors as measured by the sine. If a network has q communities, a lower rank matrix of rank q is a suitable representation of the original network matrix. The improvements in Refs. [13,15,17] show that with high probability,

$$\sin\angle(v_1, v_1') \leqslant C_0\left(\frac{\sqrt{q}}{\Delta_1} + \frac{\epsilon||E||}{z_1} + \frac{\epsilon^2||E||^2}{z_1\Delta_1}\right). \tag{5}$$

The improvements also provide a bound on the largest principal angles between subspaces $V = \{v_1, \ldots, v_j\}$ and $V' = \{v_1', \ldots, v_j'\}$, for $1 \leqslant j \leqslant q$, defined as

$$\sin\angle(V, V') = \max_{v\in V; v\neq 0} \min_{v'\in V'; v'\neq 0} \sin\angle(v, v'). \tag{6}$$

The bound on subspaces is given by

$$\sin\angle(V, V') \leqslant C_1\left(\frac{\sqrt{q}}{\Delta_j} + \frac{\epsilon||E||}{z_j} + \frac{\epsilon^2||E||^2}{z_j\Delta_j}\right). \tag{7}$$

These improvements [13,15,17] provide a tighter bound on the angles than the Davis-Kahan bounds.

III. RESULTS

We construct A using a stochastic block model (SBM), following Refs. [6,9,18], with q communities of s nodes, yielding a total number of nodes $N = sq$. Each node i has a community label $g_i \in [1, \ldots, q]$. Edges are then generated independently based on a $q \times q$ probability matrix p, with $Pr[A_{ij} = 1] = p_{g_i g_j}$. In the simplest case, $p_{ab} = p_{\text{in}}$ if $a = b$ and $p_{ab} = p_{\text{out}}$ if $a \neq b$, with $p_{\text{in}} > p_{\text{out}}$. For the sparse case, we define $c_{\text{in}} = Np_{\text{in}}$ and $c_{\text{out}} = Np_{\text{out}}$, or equivalently $\mu_{\text{in}} = sp_{\text{in}}$ and $\mu_{\text{out}} = sp_{\text{out}}$, with c_{in} and c_{out} constant in the limit $N \to \infty$. Thus, A is partitioned into q^2 blocks of size $s \times s$, with q blocks along the diagonal and $q(q-1)$ off-diagonal.

We have empirically shown distributions of eigenvalues and resulting detection of modularity for a distribution of unequal module sizes in previous work [6,18], where the eigenvalues

formed clusters based on the distributions of module sizes. Here results are presented for all communities of the same size, since we also explore the rotation behavior up to the theoretical detectability limit.

A. The theoretical detectability limit

In this section we discuss the behavior of eigenvalues when, keeping probability of connections inside a module or community constant (μ_{in}, p_{in}, or c_{in}), we increase the probability of connections between modules (i.e., steadily increase μ_{out}, p_{out}, or c_{out}).

Now, we can write $A = \bar{A} + X$, where $\bar{A}$, the ensemble average matrix, is also partitioned into q^2 blocks of size s, with q diagonal blocks with all entries equal to p_{in}, and $q(q-1)$ off-diagonal blocks with all entries equal to p_{out}. The fluctuations around the average X, by definition, has mean 0 and finite variance. As mentioned before, the eigenvalues of A are denoted by $\{z_1 \geqslant z_2 \geqslant \cdots \geqslant z_N\}$. Let us denote the eigenvalues of $\bar{A}$ by $\{\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N\}$ and the eigenvalues of X by $\{x_1 \geqslant x_2 \geqslant \cdots \geqslant x_N\}$.

Since A, $\bar{A}$, and X are all symmetric, we use the Weyl inequalities, which imply that spectrum of a symmetric (Hermitian in the general case) matrix will be stable to small perturbations [19]:

$$z_{i+j-1} \leqslant \lambda_i + x_j \tag{8}$$

with $i, j \geqslant 1$ and $i + j - 1 \leqslant N$. We are particularly interested in the following cases, which give us lower and upper bound estimates for z_1 and $z_{2:q}$:

$$z_1 \leqslant \lambda_1 + x_1, i, j = 1, \tag{9}$$

$$z_{2:q} \leqslant \lambda_{2:q} + x_1, i = 2:q, j = 1. \tag{10}$$

This shows that when there are no fluctuations around the mean, then $z_1 = \lambda_1$, and $z_{2:q} = \lambda_{2:q}$, but as fluctuations around the mean increase, z_1 and $z_{2:q}$ increase but are still bounded by the inequalities (9) and (10). Thus, the lower bound on z_i of A is $b_l = \lambda_i$, the upper bound is $b_u = \lambda_i + x_1$.

Thus, the actual z_i lie anywhere between these upper and lower bounds. This implies an argument for detecting the modules, that is, when the mean of these bounds $b = (b_l + b_u)/2 = (\lambda_i + \lambda_i + x_1)/2 = \lambda_i + (x_1/2)$ is subsumed into the bulk distribution and is equal to x_1, it will no longer be possible to detect all the modules (as it is to be reasonably expected that at least one or some of the eigenvalues would have passed into the bulk by the time the mean passes into the bulk). As we will see, this provides us with the generalized theoretical detectability limit, which was proved for two communities in Ref. [9].

Now we compute the eigenvalues of $\bar{A}$ and X. The eigenvalues of $\bar{A}$ can be easily calculated [6]. The first eigenvalue is the largest, with

$$\lambda_1 = s[p_{\text{in}} + (q-1)p_{\text{out}}], \tag{11}$$

$$\lambda_1 = \tfrac{1}{q}[c_{\text{in}} + (q-1)c_{\text{out}}], \tag{12}$$

where λ_1 can also be expressed in terms of c_{in} and c_{out}, with $N \to \infty$. Similarly, the next $q-1$ are

$$\lambda_{2:q} = s(p_{\text{in}} - p_{\text{out}}), \tag{13}$$

$$\lambda_{2:q} = \tfrac{1}{q}[c_{\text{in}} - c_{\text{out}}], \tag{14}$$

where $\lambda_{2:q}$ implies eigenvalues from 2 to q, while the remaining eigenvalues are zero:

$$\lambda_{q+1:N} = 0. \tag{15}$$

We now derive the eigenvalue distribution for the matrix X, which is symmetric, has a mean of 0, and finite variance. By Wigner's semicircle law [16], all the eigenvalues of X will be contained in the interval $[-2\sigma_A\sqrt{N}, 2\sigma_A\sqrt{N}]$, where σ_A represents the standard deviation of entries in the matrix A. For our case, this implies that the eigenvalues of X are spread around 0, but its largest one is $2\sigma_A\sqrt{N}$.

We now derive the variance σ_A^2. A has only 0 and 1 entries. The mean expected value of A is $M = \frac{1}{q}[p_{\text{in}} + (q-1)p_{\text{out}}]$, thus the number of 1 entries, columnwise, is NM, and the number of 0 entries, columnwise, is $N(1-M)$. Calculating variance by its definition,

$$\sigma_A^2 = [NM(1-M)^2 + N(1-M)(0-M)^2]/N \tag{16}$$

$$= M(1-M)^2 + M^2(1-M) \tag{17}$$

$$= M(1-M) \tag{18}$$

$$= \frac{1}{q}[p_{\text{in}} + (q-1)p_{\text{out}}] - \left(\frac{1}{q}[p_{\text{in}} + (q-1)p_{\text{out}}]\right)^2. \tag{19}$$

The variance for each column is the same, so the variance for the whole of A is as shown in Eq. (19). Further, as N grows, $M(1-M) \approx M$, thus we can say $\sigma_A \approx \sqrt{M}$.

Now, following Ref. [16], the largest eigenvalue of X can be computed using $2\sigma_A\sqrt{N}$ as

$$x_1 = 2\sqrt{NM} = 2\sqrt{\lambda_1}. \tag{20}$$

Using Eqs. (9) and (10), the $z_{2:q}$ fall between the lower and upper bounds of $b_l = \lambda_{2:q}$ and $b_u = \lambda_{2:q} + x_1$, with $b = \lambda_{2:q} + (x_1/2)$. Since $\lambda_{2:q}$ provide the lower limit, one threshold is attained when

$$\lambda_{2:q} = 2\sqrt{\lambda_1}. \tag{21}$$

This provides the lower threshold [demonstrated in Figs. 1(b)–1(e)]:

$$\lambda_{2:q} > 2\sqrt{\lambda_1}, \tag{22}$$

$$s(p_{\text{in}} - p_{\text{out}}) > 2\sqrt{s(p_{\text{in}} + (q-1)p_{\text{out}}}, \tag{23}$$

$$qs(p_{\text{in}} - p_{\text{out}}) > 2\sqrt{q^2 s(p_{\text{in}} + (q-1)p_{\text{out}}}, \tag{24}$$

$$c_{\text{in}} - c_{\text{out}} > 2\sqrt{q[c_{\text{in}} + (q-1)c_{\text{out}}]}. \tag{25}$$

This also implies the condition $\lambda_1 > \lambda_{2:q} > 2\sqrt{\lambda_1}$. Since λ_2 is the lowest possible value of z_2, this is the lowest limit, that is, it is ensured that if this condition is satisfied, then the modules will be detected absolutely. This threshold marks the beginning of the "hard" phase, where even though the actual $z_{2:q}$ still sit outside the bulk, it gets progressively harder to

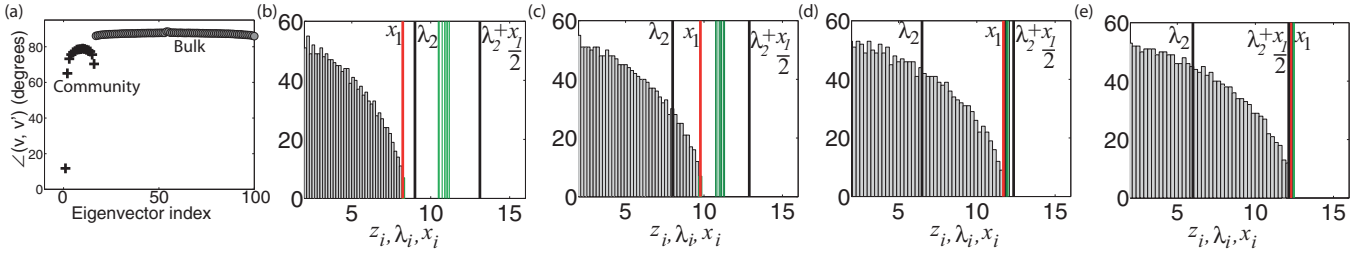

FIG. 1. Eigenvector rotation angles under perturbation and their relationship to eigenvalue gaps. (a) Mean angles of rotation under perturbation for eigenvectors corresponding to the 100 largest (absolute) eigenvalues, for stochastic block models with $N = 10\,000$, $q = 16$, $\mu_{\text{in}} = 10$, $\mu_{\text{out}} = 1$. Points are averages over 10 networks, each perturbed 300 times. (b–e) Spectrum for A showing bulk eigenvalues; red line shows x_1, green lines are $z_{2:q}$, and black lines show their lower bound $[b_l]$ and upper bound $[b]$, $N = 4096$, $q = 8$, $\mu_{\text{in}} = 10$, and $\mu_{\text{out}} = 1$, 2, 3.5 (just before the detectability limit) and 4 (just after the detectability limit), respectively.

detect them as c_{out} increases further; see Figs. 1(b)–1(e) for a demonstration.

Further, let us now consider the eigenvalues $z_{2:q}$: when these eigenvalues move into the bulk, it will provide the upper bound for the detectability, since once all or some of them move into the bulk, it will no longer be possible to detect the communities. As already mentioned, since the eigenvalues z_i are bounded by b_l and b_u, a condition for not detecting all the modules is when b becomes equal to the largest eigenvalue of the bulk distribution, x_1. When we set b equal to the largest eigenvalue of the bulk x_1, as before, we get

$$\tfrac{1}{2}[2\lambda_2 + 2\sqrt{\lambda_1}] = 2\sqrt{\lambda_1}, \tag{26}$$

$$\lambda_2 = \sqrt{\lambda_1}, \tag{27}$$

which will give us the general detectability limit for q communities, as stated in previous works and derived for $q = 2$ [7,9], and demonstrated in Figs. 1(b)–1(e):

$$c_{\text{in}} - c_{\text{out}} = \sqrt{q[c_{\text{in}} + (q-1)c_{\text{out}}]}. \tag{28}$$

This "hard" phase also shows that it will be hard to detect the modules even before the absolute detectability limit. For the stochastic block model demonstrated in Fig. 1, the actual z_i are smaller than b, and as the value of c_{out} is increased, it becomes impossible to detect the modules or their number even before the threshold is reached. For example, at a c_{out} value of 3.5, which is just below the detectability threshold (3.9), it is already not possible to clearly detect the modules [Fig. 1(d)]. This behavior has also been empirically reported in Ref. [7]. With the analytic insight provided here, we establish that the detectability properties begin to deteriorate in a "hard" phase of detection that is characterized by the bounds.

For theoretical interest, if we push the same idea to its other extreme limit, that is, the point where the upper limit $\lambda_2 + x_1$ becomes equal to x_1, we will see that λ_2 goes to zero: this implies no modularity in the system, only the bulk. If c_{out} increases even further, we can hypothesize that the $\langle z_{2:q}\rangle$ will move towards and out of the other end of the bulk, and the groups would be distinguishable again. They would not be communities, though, but anticommunities, as they will be much more connected with other groups than they are internally.

The general detectability limit provides the condition on the gaps between eigenvalues that governs the perturbation behavior of the eigenvectors, used to detect the numbers and compositions of modules, as shown in the next section.

B. Detecting number of communities through rotation of eigenvectors under perturbation

The first eigenvalue of $\bar{A}$ is the largest, $\lambda_1 = s[p_{\text{in}} + (q-1)p_{\text{out}}]$, the next $q-1$ are $\lambda_{2:q} = s(p_{\text{in}} - p_{\text{out}})$, where $\lambda_{2:q}$ implies all eigenvalues from 2 to q, while the remaining eigenvalues are $\lambda_{q+1:N} = 0$. Of particular interest here are the following gaps, where we consider the limit b as an estimate for the value $\langle z_i\rangle$:

$$\delta_1 = \lambda_1 - \lambda_{2:q} = Np_{\text{out}}, \tag{29}$$

$$\Delta_1 = \langle z_1\rangle - \langle z_{2:q}\rangle = \delta_1, \tag{30}$$

$$\delta_2 = \lambda_{2:q} - 2\sqrt{\lambda_1}, \tag{31}$$

$$\Delta_2 = \langle z_{2:q}\rangle - 2\sqrt{\lambda_1} = \lambda_{2:q} - \sqrt{\lambda_1}, \tag{32}$$

$$\delta_3 = 2\sqrt{\lambda_1} - \lambda_{q+1:N} = 2\sqrt{\lambda_1}. \tag{33}$$

We now perturb A with ϵE, getting $A + \epsilon E = (\bar{A} + X) + \epsilon E = \bar{A} + X'$, where once again X', by definition, has mean 0 and finite variance. Thus, $\bar{A}$ is a rank q matrix, with $q < N$. We then substitute $\langle z_i\rangle$ and Δ_i into the new improved bounds [Eqs. (5) and (7)].

First, for v_1, substituting the values of $\langle z_1\rangle$ and Δ_1 into Eq. (5), if we fix q, p_{in}, and p_{out}, all three terms decrease as N grows. Thus, the rotation of the first eigenvector is bounded to a small angle, as seen in Fig. 1(a).

Second, eigenvectors $2\ldots q$ span a subspace and rotate together. Defining the subspace $V = \{v_2, \ldots, v_q\}$ and $V' = \{v_2', \ldots, v_q'\}$, the largest principal angle between all pairs of vectors is governed by Δ_2. Substituting the values of $\langle z_{2:q}\rangle$ and Δ_2 into Eq. (7), with q, p_{in}, and p_{out} fixed, again implies that all three terms decrease as N grows. Thus, the rotation of eigenvectors $2, \ldots, q$ is also bounded to a small angle [Fig. 1(a)]. Since Δ_1 and $\langle z_1\rangle$ are larger than Δ_2 and $\langle z_2\rangle$, the sine of the angle between V and V' is larger than that between v_1 and v_1', but still bounded to a small angle with high probability governed by Δ_2. Figure 1(a) shows that eigenvectors $v_{2:q}$ indeed have this behavior: they rotate as

a group, showing that the subspace V behaves as one and is different from the subspaces v_1 and $V'' = \{v_{q+1}, \ldots, v_n\}$. We call V the *community eigenspace* and V'' the *bulk eigenspace*, for which applying the same theorems will lead to the largest angles of rotation with the sine approaching 90°.

The results in Fig. 1(a) show a clear sharp separation between the community eigenspace and the bulk eigenspace, revealing the correct number of modules in the network. This behavior changes as we approach the detectability limit for sparse modular networks, a threshold beyond which modularity exists in the network, but cannot be detected.

Figures 1(b)–1(e) shows how the zone between $\lambda_{2:q}$ and $\lambda_{2:q} + x_1/2$ gradually moves into the bulk as c_{out} or μ_{out} are increased, keeping c_{in} or μ_{in} constant. Δ_2 gradually decreases as $\lambda_{2:q} + x_1/2$ moves towards the bulk and becomes equal to x_1, providing the detectability threshold.

Figure 2 shows the differences between the mean angles of rotation of the first eigenvector of the bulk [the $(q+1)$-th, though choosing any vector for this plot would still reveal q due to the structure in Fig. 1(a)] and of all the other eigenvectors of A, $D_{q+1} - D_i, i = 2{:}N$, for $\mu_{\text{out}} = 0$ to 10 keeping $\mu_{\text{in}} = 10$. This difference brings out the behavior of the two subspaces clearly: for eigenvectors 2 to q, $D_{q+1} - D_i$ decreases as μ_{out} is increased, whereas for the eigenvectors of the bulk the behavior is different.

At the detectability threshold (vertical line), this difference is the smallest for all the vectors, and after the threshold it starts to expand again, for both the community and the bulk eigenvectors. The larger this difference, the clearer the separation between the community and the bulk eigenspaces, and the easier it is to detect the number of groups.

Note that once the "hard" phase of detection sets in, there is a steep drop and a near zero gap to the detectability limit, even before the exact limit is reached. At the detectability threshold, the rotation angles of the community and the bulk eigenspace merge into one. If μ_{out} increases further, we can hypothesize that the $\langle z_{2:q} \rangle$ will move towards and out of the other end of the bulk, and the groups would be distinguishable again. They would not be communities, though, but anticommunities, as they will be much more connected with other groups than they are internally.

In addition to the above results, we also empirically observe an oscillatory behavior that is not explained by the theorems above: the pairing up of eigenvectors corresponding to "mirrored" eigenvalues in both the community and bulk eigenspaces. For example, in Fig. 1(a), with 16 modules, the angles of rotation under perturbation for the first 16 eigenvectors are separated from the bulk. The angles of rotation of the community eigenvectors, e.g., the 2nd and the 16th, the 3rd, the 15th, and so on, and those of the bulk, e.g., the 17th, the 10 000th, and so on, are similar. We observed this behavior across a large range of parameters and networks. The angles of rotation first increase for the first half of the eigenvectors, and then decrease again for the last half of the eigenvectors in the subspace, resulting in a symmetric pairing up of eigenvectors from the two halves.

Figure 3(a) shows the distribution of eigenvalues in the community and bulk eigenspaces along with the mean eigenvalue of each eigenspace. We characterize the distribution of eigenvalues in the two eigenspace distributions by defining deviations of each eigenvalue from the mean of the eigenvalues in each space. These are defined as $w_i = |z_{2:q} - \langle z_{2:q} \rangle|, i = 2{:}q$, and $w_i = |z_{q+1:N} - \langle z_{q+1:N} \rangle|, i = q+1{:}N$. Over m networks, we get vectors $\mathbf{w}_i \in \Re^m$, $i = 1{:}N$, where each $\mathbf{w}_i$ is a measure of the deviation from the mean eigenvalues for the groups defined above. We compute a distance matrix W with W_{ij} equal to the Euclidean distance between the vectors $\mathbf{w}_i$ and $\mathbf{w}_j$, with $i, j = 1$ to N. Figures 3(b) and 3(c) show W for the bulk and the community spaces, respectively. The main diagonals in both show that the eigenvalues successively close to each other are at very low distance, but the main cross-diagonal shows the same low-distance relationship for pairs of eigenvalues symmetrically disposed about the mean eigenvalue of the eigenspace, showing self-similar behavior in both the bulk and in the community eigenspace; i.e., Fig. 3(c) is simply a blow up of the bottom-right corner of Fig. 3(b). Since the gaps between eigenvalues govern the rotation behavior, we empirically relate observations on the distributions of eigenvalues, gaps between them, and their symmetric distribution around the means: not

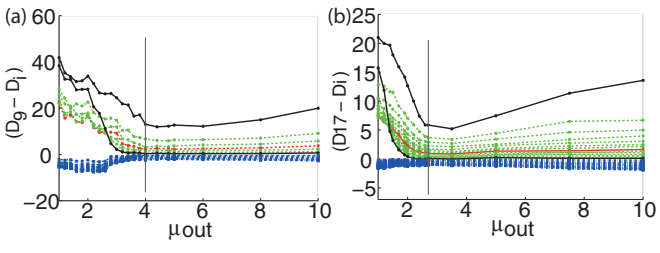

FIG. 2. Difference between mean angles of rotation of the $(q+1)$-th and of the top 50 eigenvectors of stochastic block models under perturbations, $D_{q+1} - D_i, i = 2{:}50$ as μ_{out} is varied from 0 to 10, keeping $\mu_{\text{in}} = 10$. (a) $N = 4096$, $q = 8$, and detectability threshold $\mu^D_{\text{out}} = 4$. (b) $N = 10000$, $q = 16$, and detectability threshold $\mu^D_{\text{out}} = 2.7$. Each point is an average over 10 networks, each perturbed 300 times. Rotation lines for eigenvectors 2nd and qth in black, all other community eigenvectors in green, the middle one between the 2nd and the qth in red, bulk eigenvectors in blue.

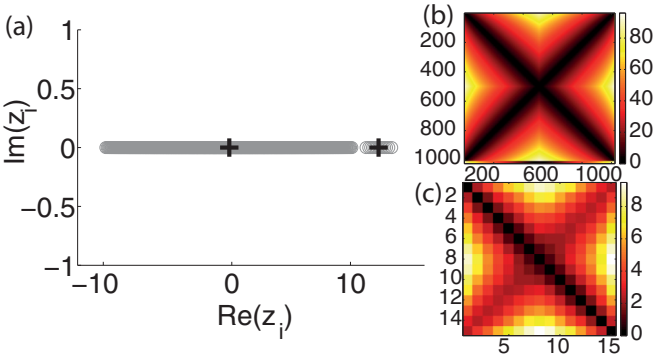

FIG. 3. Symmetric distribution of eigenvalues as basis for "pairing up" of eigenvectors under rotation. (a) Eigenvalue distribution for a network generated by a stochastic block model with $N = 10000$, $q = 16$, $\mu_{\text{in}} = 10$, $\mu_{\text{out}} = 1$. Black markers show mean eigenvalues for the community and bulk eigenspaces. (b, c). Distance matrix $\mathbf{W}$ for bulk and community eigenspaces, respectively.

only do the rotation angles depend on the gaps, but the pairing of eigenvectors in each eigenspace seems to be related to the pairing of eigenvalues via their symmetric distributions around the mean eigenvalue in that eigenspace. This behavior is lost at the detectability limit, where the two spaces merge to become one, as shown both by the eigenvalue gaps [Figs. 1(b)–1(e)] and the rotation of eigenvectors under perturbation (Fig. 2).

C. Tests on real and benchmark networks

Finally, the SBM, while very useful for deriving analytical results, does not accurately represent the structure of real world networks. Therefore, we tested the approach on Lancichinetti-Fortunato-Radicchi (LFR) benchmark graphs [20] that can be used to generate networks that have some properties of real world networks such as broad degree and community size distributions and benchmark real world networks with known community structure. We generated a number of these networks, parametrically defining a complex mix of parameters such as a high number of very small communities with differing sizes and varying degree.

In Figs. 4(a) and 4(b), we use LFR networks with 1000 nodes and 24 and 43 communities, respectively, and vary the mixing parameter μ, expressing the ratio between the external degree of a node and its total degree. The number of communities is correctly estimated by the rotation of eigenvectors under perturbation.

The plot in Fig. 4(a) shows the community and bulk eigenspaces, respectively, with the community eigenspace defined by the first 24 eigenvectors corresponding to the 24 largest eigenvalues, respectively, and the bulk eigenspace defined by the rest. We note that not only is there a sharp separation gap between the 24th and 25th eigenvectors, there is also the corresponding oscillatory behavior in the community and bulk eigenspaces, revealing exactly the 24 communities. Exactly the same reading holds for Fig. 4(b), showing an LFR network with 43 communities. These communities can then be detected by using any algorithm, but most obviously, by using a lower dimensional space defined by 28 (or 43) eigenvectors in each case and defining each vertex as a point in 28- or 43-dimensional space and using cosine or K-means clustering to detect the modules accurately [5]. The gap is less prominent in Fig. 4(b) because μ is higher and communities are more mixed and harder to detect.

We also apply the approach to some real world benchmark networks. Figures 4(c) and 4(d) show the American college football network [21] and the network of political books [22]. The American College Football network [21] is a network of American football games between college teams during a regular season (Fall 2000). Teams are organized into conferences, with each conference containing around 8–12 teams. Games are more frequent between teams of the same conferences than between teams of different conferences. This results in communities. The network of political books [22] is data incorporating books about recent U.S. politics sold by the online bookseller Amazon.com. Edges between books represent frequent copurchasing of books by the same buyers. The network was compiled by V. Krebs and is unpublished but can found on Krebs's web site. Two main communities exist, representing two main political party divisions, since members supporting one party are more likely to purchase books representing that party's ideology. In both cases, the correct number of communities is predicted, and lower dimensional spectral detection can similarly be employed to detect the communities accurately.

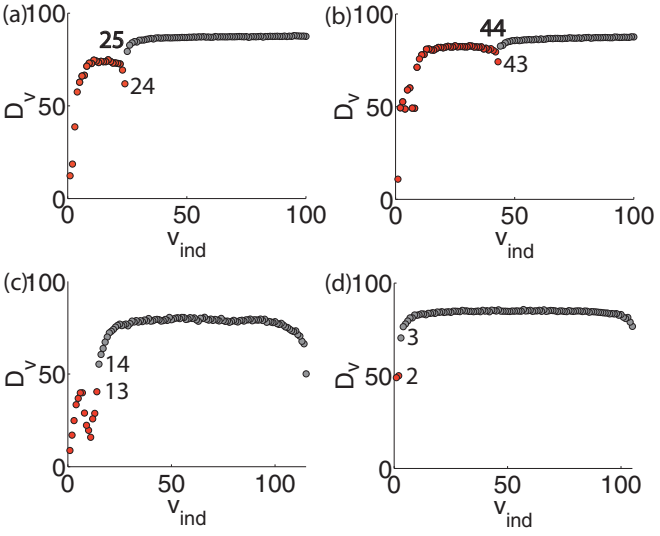

FIG. 4. Number of communities (red circles) detected in LFR and real networks by eigenvector rotations. Each network is perturbed 300 times; the eigenvector index v_{ind} is plotted against its angle of rotation D_v. (a) LFR network with $N = 1000$, $q = 24$, $\langle k \rangle = 25$, $k_{\max} = 60$, $\mu = 0.3$. (b) LFR network with $N = 1000$, $q = 43$, $\langle k \rangle = 25$, $k_{\max} = 60$, $\mu = 0.4$. (c) American college football network: $q = 13$, $N = 115$ [21]. (d) Political books network: $q = 2$, $N = 105$ [22].

IV. CONCLUSIONS AND DISCUSSION

We presented the dynamics of rotation of eigenvectors of adjacency matrices of modular networks under random perturbations. In the presence of q communities, the number of eigenvectors corresponding to the q largest eigenvalues form a "community" eigenspace and rotate together, but separately from that of the "bulk" eigenspace spanned by all the other eigenvectors. Using this property, the number of modules or clusters in a network can be accurately estimated in an algorithm-independent way. Results are shown to hold to a point where a "hard" phase of detectability sets in before the theoretical detectability limit for sparse modular networks. Analytic insight is presented into the bounds of this hard phase, using which a general derivation of the detectability threshold is presented for q communities, using arguments based on these bounds. This is consistent with previous results, and a proof of which was previously provided for two communities. A plausibility argument is presented for the observed symmetric pairing up of eigenvalues and eigenvectors in the two eigenspaces. The approach presented demonstrates that the rotation behavior of the eigenvectors of the adjacency matrix under perturbations reveals information about the community structure of a network.

ACKNOWLEDGMENTS

This work is supported by a Henry Halloran Trust Research Fellowship, University of Sydney, Australian Research Council (ARC) Center of Excellence for Integrative Brain Function Grant CE140100007, ARC Laureate Fellowship Grant FL1401000225, and ARC Discovery Project Grant DP130100437.

[1] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
[2] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[3] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).
[4] M. E. J. Newman, Phys. Rev. E **74**, 036104 (2006).
[5] S. Sarkar and A. Dong, Phys. Rev. E **83**, 046114 (2011).
[6] S. Sarkar, J. A. Henderson, and P. A. Robinson, PLoS ONE **8**(1), e54383 (2013).
[7] F. Krzakala, C. Moore, C. Mosseld, J. Neemand, A. Slyd, L. Zdeborova, and P. Zhanga, Proc. Natl. Acad. Sci. USA **110**, 20935 (2013).
[8] M. E. J. Newman, Phys. Rev. E **88**, 042822 (2013).
[9] R. R. Nadakuditi and M. E. J. Newman, Phys. Rev. Lett. **108**, 188701 (2012).
[10] S. Chauhan, M. Girvan, and E. Ott, Phys. Rev. E **80**, 056114 (2009).
[11] R. K. Darst, Z. Nussinov, and S. Fortunato, Phys. Rev. E **89**, 032809 (2014).
[12] M. Newman, arXiv:1308.6494.
[13] S. Rourke, V. Vu, and K. Wang, arXiv:1311.2657.
[14] C. Davis and M. Kahan, Bull. Am. Math. Soc. **75**, 863 (1969).
[15] V. Vu, Random Struct. Alg. **39**, 526 (2011).
[16] Z. Furedi and J. Kolmos, Combinatorica **1**, 233 (1981).
[17] V. Vu, arXiv:1004.2000.
[18] S. Sarkar, A. Dong, J. A. Henderson, and P. A. Robinson, J. Mech. Des. **136**, 011006 (2013).
[19] T. Tao, 254A, Notes 3a: Eigenvalues and sums of Hermitian matrices, https://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices/, Technical report (2010).
[20] A. Lancichinetti, S. Fortunato, and F. Radicchi, Phys. Rev. E **78**, 046110 (2008).
[21] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[22] V. Krebs, http://www.orgnet.com.