Spectral redemption in clustering sparse networks

Florent Krzakala[a,b], Cristopher Moore[c], Elchanan Mossel[d,1], Joe Neeman[d], Allan Sly[d], Lenka Zdeborová[e], and Pan Zhang[a,c]

[a]Ecole Supérieure de Physique et de Chimie Industrielles, 75005 Paris, France; [b]Ecole Normale Supérieure, 75005 Paris, France; [c]Santa Fe Institute, Santa Fe, NM 87501; [d]University of California, Berkeley, CA 94720; and [e]Institut de Physique Théorique, Commissariat à l'Energie Atomique Saclay and Unité de Recherche Associée 2306, Centre National de la Recherche Scientifique, 91190 Gif-sur-Yvette, France

Edited* by Peter J. Bickel, University of California, Berkeley, CA, and approved October 23, 2013 (received for review July 2, 2013)

Spectral algorithms are classic approaches to clustering and community detection in networks. However, for sparse networks the standard versions of these algorithms are suboptimal, in some cases completely failing to detect communities even when other algorithms such as belief propagation can do so. Here, we present a class of spectral algorithms based on a nonbacktracking walk on the directed edges of the graph. The spectrum of this operator is much better-behaved than that of the adjacency matrix or other commonly used matrices, maintaining a strong separation between the bulk eigenvalues and the eigenvalues relevant to community structure even in the sparse case. We show that our algorithm is optimal for graphs generated by the stochastic block model, detecting communities all of the way down to the theoretical limit. We also show the spectrum of the nonbacktracking operator for some real-world networks, illustrating its advantages over traditional spectral clustering.

Detecting communities or modules is a central task in the study of social, biological, and technological networks. Two of the most popular approaches are statistical inference, where we fix a generative model such as the stochastic block model to the network (1, 2); and spectral methods, where we classify vertices according to the eigenvectors of a matrix associated with the network such as its adjacency matrix or Laplacian (3).

Both statistical inference and spectral methods have been shown to work well in networks that are sufficiently dense, or when the graph is regular (4–8). However, for sparse networks with widely varying degrees, the community detection problem is harder. Indeed, it was recently shown (9–11) that there is a phase transition below which communities present in the underlying block model are impossible for any algorithm to detect. Whereas standard spectral algorithms succeed down to this transition when the network is sufficiently dense, with an average degree growing as a function of network size (8), in the case where the average degree is constant these methods fail significantly above the transition (12). Thus, there is a large regime in which statistical inference succeeds in detecting communities, but where current spectral algorithms fail.

It was conjectured in ref. 11 that this gap is artificial and that there exists a spectral algorithm that succeeds all of the way to the detectability transition even in the sparse case. Here, we propose an algorithm based on a linear operator considerably different from the adjacency matrix or its variants: namely, a matrix that represents a walk on the directed edges of the network, with backtracking prohibited. We give strong evidence that this algorithm indeed closes the gap.

The fact that this operator has better spectral properties than, for instance, the standard random walk operator, has been used in the past in the context of random matrices and random graphs (13–15). In the theory of zeta functions of graphs, it is known as the edge adjacency operator, or the Hashimoto matrix (16). It has been used to show fast mixing for the nonbacktracking random walk (17), and arises in connection to belief propagation (18, 19), in particular to rigorously analyze the behavior of belief propagation for clustering problems on regular graphs (5). It has also been used as a feature vector to classify graphs (20). However, we are not aware of work using this operator for clustering or community detection.

We show that the resulting spectral algorithms are optimal for networks generated by the stochastic block model, finding communities all of the way down to the detectability transition. That is, at any point above this transition, there is a gap between the eigenvalues related to the community structure and the bulk distribution of eigenvalues coming from the random graph structure, allowing us to find a labeling correlated with the true communities. In addition to our analytic results on stochastic block models, we also illustrate the advantages of the nonbacktracking operator over existing approaches for some real networks.

Spectral Clustering and Sparse Networks

To study the effectiveness of spectral algorithms in a specific ensemble of graphs, suppose that a graph G is generated by the stochastic block model (1). There are q groups of vertices, and each vertex v has a group label $g_v \in \{1,\dots,q\}$. Edges are generated independently according to a $q\times q$ matrix p of probabilities, with $\Pr[A_{u,v}=1]=p_{g_u g_v}$. In the sparse case, we have $p_{ab}=c_{ab}/n$, where the affinity matrix c_{ab} stays constant in the limit $n\to\infty$.

For simplicity we first discuss the commonly studied case where c has two distinct entries, $c_{ab}=c_{\text{in}}$ if $a=b$ and c_{out} if $a\neq b$. We take $q=2$ with two groups of equal size, and assume that the network is assortative, i.e., $c_{\text{in}}>c_{\text{out}}$. The section *More than Two Groups and General Degree Distributions* below discusses our results in more general cases.

The group labels are hidden from us, and our goal is to infer them from the graph. Let $c=(c_{\text{in}}+c_{\text{out}})/2$ denote the average degree. The detectability threshold (9–11) states that in the limit $n\to\infty$, unless

Significance

Spectral algorithms are widely applied to data clustering problems, including finding communities or partitions in graphs and networks. We propose a way of encoding sparse data using a "nonbacktracking" matrix, and show that the corresponding spectral algorithm performs optimally for some popular generative models, including the stochastic block model. This is in contrast with classical spectral algorithms, based on the adjacency matrix, random walk matrix, and graph Laplacian, which perform poorly in the sparse case, failing significantly above a recently discovered phase transition for the detectability of communities. Further support for the method is provided by experiments on real networks as well as by theoretical arguments and analogies from probability theory, statistical physics, and the theory of random matrices.

Author contributions: F.K., C.M., E.M., J.N., A.S., L.Z., and P.Z. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper. The authors are listed in alphabetical order.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: mossel@stat.berkeley.edu.

$$c_{\text{in}} - c_{\text{out}} > 2\sqrt{c}, \quad [1]$$

the randomness in the graph washes out the block structure to the extent that no algorithm can label the vertices better than chance. Moreover, ref. 11 proved that below this threshold it is impossible to identify the parameters c_{in} and c_{out}, whereas above the threshold the parameters c_{in} and c_{out} are easily identifiable.

The adjacency matrix is defined as the $n \times n$ matrix $A_{u,v} = 1$ if $(u,v) \in E$ and 0 otherwise. A typical spectral algorithm assigns each vertex a k-dimensional vector according to its entries in the first k eigenvectors of A for some k, and clusters these vectors according to a heuristic such as the k-means algorithm (often after normalizing or weighting them in some way). In the case $q = 2$, we can simply label the vertices according to the sign of the second eigenvector.

As shown in ref. 8, spectral algorithms succeed all of the way down to the threshold **1** if the graph is sufficiently dense. In that case, A's spectrum has a discrete part and a continuous part in the limit $n \to \infty$. Its first eigenvector essentially sorts vertices according to their degree, whereas the second eigenvector is correlated with the communities. The second eigenvalue is given by

$$\lambda_c = \frac{c_{\text{in}} - c_{\text{out}}}{2} + \frac{c_{\text{in}} + c_{\text{out}}}{c_{\text{in}} - c_{\text{out}}}. \quad [2]$$

The question is when this eigenvalue gets lost in the continuous bulk of eigenvalues coming from the randomness in the graph. This part of the spectrum, like that of a sufficiently dense Erdős–Rényi random graph, is asymptotically distributed according to Wigner's semicircle law (21), $P(\lambda) = \sqrt{4c - \lambda^2}/2\pi c$. Thus, the bulk of the spectrum lies in the interval $[-2\sqrt{c}, 2\sqrt{c}]$. If $\lambda_c > 2\sqrt{c}$, which is equivalent to **1**, the spectral algorithm can find the corresponding eigenvector, and it is correlated with the true community structure.

However, in the sparse case where c is constant while n is large, this picture breaks down due to a number of reasons. Most importantly, the leading eigenvalues of A are dictated by the vertices of highest degree, and the corresponding eigenvectors are localized around these vertices (22). As n grows, these eigenvalues exceed λ_c, swamping the community-correlated eigenvector, if any, with the bulk of uninformative eigenvectors. As a result, spectral algorithms based on A fail a significant distance from the threshold given by **1**. Moreover, this gap grows as n increases: for instance, the largest eigenvalue grows as the square root of the largest degree, which is roughly proportional to $\log n / \log\log n$ for Erdős–Rényi graphs. To illustrate this problem, the spectrum of A for a large graph generated by the block model is depicted in Fig. 1.

Other popular operators for spectral clustering include the Laplacian $L = D - A$, where $D_{uv} = d_u \delta_{uv}$ is the diagonal matrix of vertex degrees, the symmetrically normalized Laplacian $D^{-1/2} L D^{-1/2}$, the stochastic random walk matrix $Q = AD^{-1}$, and the modularity matrix $M_{uv} = A_{uv} - d_u d_v/(2m)$. However, like A, these are prey to localized eigenvectors in the sparse case.

Another simple heuristic is to simply remove the high-degree vertices (e.g., ref. 6), but this throws away a significant amount of information; in the sparse case it can even destroy the giant component, causing the graph to fall apart into disconnected pieces (23). Finally, one can also regularize the adjacency matrix by adding a small constant term (24); however, this introduces a tunable parameter, and we have not explored this here.

Nonbacktracking Operator

The main contribution of this paper is to show how to redeem the performance of spectral algorithms in sparse networks by using a different linear operator. The nonbacktracking matrix B

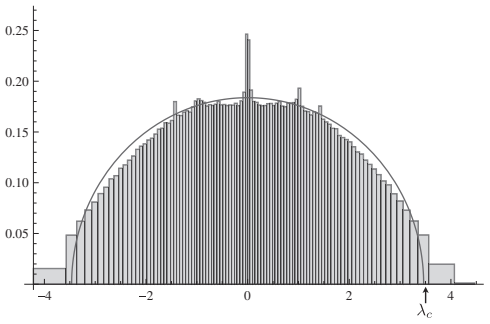

Fig. 1. Spectrum of the adjacency matrix of a sparse network generated by the block model (excluding the zero eigenvalues). Here, $n = 4{,}000$, $c_{\text{in}} = 5$, and $c_{\text{out}} = 1$, and we average over 20 realizations. Even though the eigenvalue $\lambda_c = 3.5$ given by Eq. **2** satisfies the threshold condition **1** and lies outside the semicircle of radius $2\sqrt{c} = 3.46$, deviations from the semicircle law cause it to get lost in the bulk, and the eigenvector of the second largest eigenvalue is uncorrelated with the community structure. As a result, spectral algorithms based on A are unable to identify the communities in this case.

is a $2m \times 2m$ matrix, defined on the directed edges of the graph. Specifically,

$$B_{(u\to v),(w\to x)} = \begin{cases} 1 & \text{if } v = w \text{ and } u \neq x \\ 0 & \text{otherwise.} \end{cases}$$

Using B rather than A addresses the problem described above. The spectrum of B is not sensitive to high-degree vertices, because a walk starting at v cannot turn around and return to it immediately. Other convenient properties of B are that any tree dangling off the graph, or disconnected from it, simply contributes zero eigenvalues to the spectrum, because a nonbacktracking walk is forced to a leaf of the tree where it has nowhere to go. Similarly, one can show that unicyclic components yield eigenvalues that are either 0, 1, or −1.

As a result, B has the following spectral properties in the limit $n \to \infty$ in the ensemble of graphs generated by the block model. The leading eigenvalue is the average degree $c = (c_{\text{in}} + c_{\text{out}})/2$. At any point above the detectability threshold **1**, the second eigenvalue is associated with the block structure and reads

$$\mu_c = \frac{c_{\text{in}} - c_{\text{out}}}{2}. \quad [3]$$

Moreover, the bulk of B's spectrum is confined to the disk in the complex plane of radius $\sqrt{c}$, as shown in Fig. 2. Thus, the second eigenvalue is well-separated from the top of the bulk, i.e., from the third largest eigenvalue in absolute value, as shown in Fig. 3.

The eigenvector corresponding to μ_c is strongly correlated with the communities. Because B is defined on directed edges, at each vertex we sum this eigenvector over all its incoming edges. If we label vertices according to the sign of this sum, then the majority of vertices are labeled correctly (up to a change of sign, which switches the two communities). Thus, a spectral algorithm based on B succeeds when $\mu_c > \sqrt{c}$, i.e., when **1** holds—but, unlike standard spectral algorithms, this criterion now holds even in the sparse case.

We present arguments for these claims in the next section. We will also see that the important part of B's spectrum can be obtained from a $2n \times 2n$ matrix (16, 25, 26)

$$B' = \begin{pmatrix} 0 & D - \mathbb{1} \\ -\mathbb{1} & A \end{pmatrix}. \quad [4]$$

This lets us work with a $2n$-dimensional matrix rather than a $2m$-dimensional one, which significantly reduces the computational complexity of our algorithm.

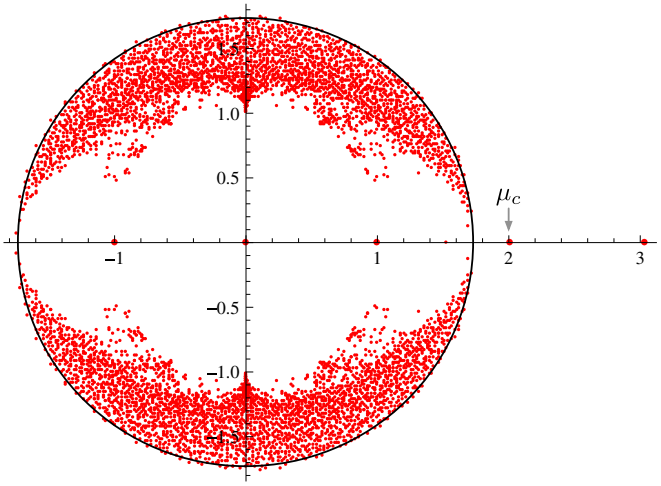

Fig. 2. Spectrum of the nonbacktracking matrix B for a network generated by the block model with same parameters as in Fig. 1. The leading eigenvalue is at $c=3$, the second eigenvalue is close to $\mu_c=(c_{\text{in}}-c_{\text{out}})/2=2$, and the bulk of the spectrum is confined to the disk of radius $\sqrt{c}=\sqrt{3}$. Because μ_c is outside the bulk, a spectral algorithm that labels vertices according to the sign of B's second eigenvector (summed over the incoming edges at each vertex) labels the majority of vertices correctly.

Reconstruction and a Community-Correlated Eigenvector

In this section we sketch justifications of the claims in the previous section regarding B's spectral properties, showing that its second eigenvector is correlated with the communities whenever **1** holds. We assume that $c=O(1)$, so that the graph is locally tree-like.

We start by explicitly constructing a vector g which is correlated with the communities and is an approximate eigenvector with eigenvalue μ_c, as defined in Eq. **3**. We follow ref. 11, which derived a similar result in the case of random regular graphs. For a given integer r, consider the vector $g^{(r)}$ defined by

$$g^{(r)}_{u\to v}=\mu_c^{-r}\sum_{(w,x):d(u\to v,w\to x)=r}\sigma_x, \qquad [5]$$

where $\sigma_u=\pm 1$ denotes u's community, and $d(u\to v,w\to x)$ denotes the number of steps required to go from $u\to v$ to $w\to x$ in the graph of directed edges. By the theory of the reconstruction problem on trees (27, 28), if **1** holds, then for every $u\to v$, the correlation $\left\langle g^{(r)}_{u\to v},\sigma_u\right\rangle$ is bounded away from zero in the limit $n\to\infty$.

Next, we argue that if r is large then g is an approximate eigenvector of B with eigenvalue μ_c. As long as the radius-r neighborhood of v is a tree, we have

$$\left(Bg^{(r)}\right)_{u\to v}=\mu_c^{-r}\sum_{(w,x):d(u\to v,w\to x)=r+1}\sigma_x=\mu_c g^{(r+1)}_{u\to v}. \qquad [6]$$

This is not precisely an eigenvalue equation because $g^{(r)}\neq g^{(r+1)}$; however, it turns out that they are close with high probability. Indeed, we may write $g^{(r)}_{u\to v}-g^{(r+1)}_{u\to v}$ as

$$\mu_c^{-r}\sum_{(w,x):d(u\to v,w\to x)=r}\left[\sigma_x-\mu_c^{-1}\sum_{y\in N(x)\setminus\{w\}}\sigma_y\right].$$

Now, there are (in expectation) c^r terms in this sum, each of which, conditioned on the σ_x s, has mean zero and constant variance. Hence, $\mathbb{E}\left[\left(g^{(r)}_{u\to v}-g^{(r+1)}_{u\to v}\right)^2\right]=O(c^r\mu_c^{-2r})$. Summing over u and v, we have $\mathbb{E}\left|g^{(r)}-g^{(r+1)}\right|^2=O(c^r\mu_c^{-2r}|E|)$. If **1** holds then $\mu_c>\sqrt{c}$ and so with high probability the error term tends to zero for large r. Because $\left|g^{(r)}\right|$ is bounded above zero, Eq. **6** then becomes

$$\left|Bg^{(r)}-\mu_c g^{(r)}\right|=o(1)\left|g^{(r)}\right|,$$

so $g^{(r)}$ is indeed an approximate eigenvector for B with eigenvalue μ_c. Because, as we will discuss shortly, the bulk of B's spectrum is bounded away from μ_c, it follows that the true eigenvector with eigenvalue μ_c is close to $g^{(r)}$, and so it may be used for community detection. Specifically, if we label vertices according to the sign of this eigenvector (summed over all incoming edges at each vertex) we obtain the true communities with significant accuracy.

Summing over incoming and outgoing edges also lets us relate B's spectrum to that of B' (Eq. **4**). Given a $2m$-dimensional vector g, define g^{out} and g^{in} as the n-dimensional vectors

$$g^{\text{out}}_u=\sum_{v\in N(u)}g_{u\to v}\quad\text{and}\quad g^{\text{in}}_u=\sum_{v\in N(u)}g_{v\to u}.$$

If we apply B to g, each incoming edge $v\to u$ contributes d_u-1 times to u's outgoing edges. Similarly, each edge $w\to v$ with $w\neq u$ contributes to the incoming edge $v\to u$. As a result, we have

$$(Bg)^{\text{out}}_u=(d_u-1)g^{\text{in}}_u\quad\text{and}\quad(Bg)^{\text{in}}_u=\sum_{v\in N(u)}g^{\text{in}}_v-g^{\text{out}}_u,$$

or more succinctly,

$$\begin{pmatrix}(Bg)^{\text{out}}\\(Bg)^{\text{in}}\end{pmatrix}=B'\begin{pmatrix}g^{\text{out}}\\g^{\text{in}}\end{pmatrix}.$$

Now suppose that $Bg=\mu g$. If g^{out} and g^{in} are nonzero, then $(g^{\text{out}},g^{\text{in}})$ is an eigenvector of B' with the same eigenvalue μ. In that case, we have

$$\mu g^{\text{in}}=Ag^{\text{in}}-g^{\text{out}}=\left(A-\mu^{-1}(D-\mathbb{1})\right)g^{\text{in}},$$

so μ is a root of the quadratic eigenvalue equation

$$\det\left[\mu^2\mathbb{1}-\mu A+(D-\mathbb{1})\right]=0. \qquad [7]$$

This equation is well known in the theory of graph zeta functions (16, 25, 26). It accounts for $2n$ of B's eigenvalues, the other $2(m-n)$ of which are ± 1.

Next, we argue that the bulk of B's spectrum is confined to the disk of radius $\sqrt{c}$. First note that for any matrix B,

$$\sum_{i=1}^{2m}|\mu_i|^{2r}\le\operatorname{tr}B^r(B^r)^T.$$

On the other hand, for any fixed r, because G is locally tree-like in the limit $n\to\infty$, each diagonal entry $(u\to v,u\to v)$ of $B^r(B^r)^T$

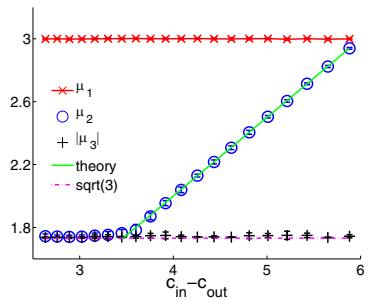

Fig. 3. First, second, and third largest eigenvalues μ_1, μ_2, and $|\mu_3|$, respectively, of B as functions of $c_{\text{in}}-c_{\text{out}}$. The third eigenvalue is complex, so we plot its modulus. Values are averaged over 20 networks of size $n=10^5$ and average degree $c=3$. The green line in the figure represents $\mu_c=(c_{\text{in}}-c_{\text{out}})/2$, and the horizontal lines are c and $\sqrt{c}$, respectively. The second eigenvalue μ_2 is well-separated from the bulk throughout the detectable regime.

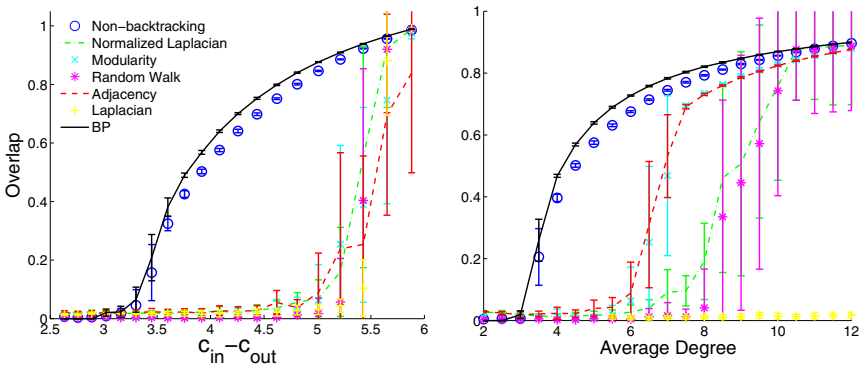

Fig. 4. Accuracy of spectral algorithms based on different linear operators, and of BP, for two groups of equal size. (*Left*) We vary $c_{in}-c_{out}$ while fixing the average degree $c=3$; the detectability transition given by **1** occurs at $c_{in}-c_{out}=2\sqrt{3}\sim 3.46$. (*Right*) We set $c_{out}/c_{in}=0.3$ and vary c; the detectability transition is at $c\sim 3.45$. Each point is averaged over 20 instances with $n=10^5$. Our spectral algorithm based on the nonbacktracking matrix B achieves an accuracy close to that of BP, and both remain large all of the way down to the transition. Standard spectral algorithms (applied to the giant component of each graph, which contains all but a small fraction of the vertices) based on the adjacency matrix, modularity matrix, Laplacian, normalized Laplacian, and the random walk matrix all fail well above the transition, giving a regime where they do no better than chance.

is equal to the number of vertices exactly r steps from v, other than those connected via u. In expectation this is c^r, so by linearity of expectation $\mathbb{E}\,\mathrm{tr}\,B^r(B^r)^T=2mc^r$. In that case, the $2r$th moment in the spectral measure obeys $\mathbb{E}\left(|\mu|^{2r}\right)\le c^r$.

Because this holds for any fixed r, we conclude that almost all of B's eigenvalues obey $|\mu|\le\sqrt{c}$. Proving that all the eigenvalues in the bulk are asymptotically confined to this disk requires a more precise argument and is left for future work.

Finally, the singular values of B are easy to derive for any simple graph, i.e., one without self-loops or multiple edges. Namely, BB^T is block-diagonal: for each vertex v, it has a rank-one block of size d_v that connects v's outgoing edges to each other. As a consequence, B has n singular values d_v-1, and its other $2m-n$ singular values are 1. However, because B is not symmetric, its eigenvalues and its singular values are different—whereas its singular values are controlled by the vertex degrees, its eigenvalues are not. This is precisely why its spectral properties are better than those of A and related operators.

More than Two Groups and General Degree Distributions

The arguments given above regarding B's spectral properties generalize straightforwardly to other graph ensembles. First, consider block models with q groups, where for $1\le a\le q$ group a has fractional size n_a. The average degree of group a is $c_a=\sum_b c_{ab}n_b$. The hardest case is where $c_a=c$ is the same for all a, so that we cannot simply label vertices according to their degree.

The leading eigenvector again has eigenvalue c, and the bulk of B's spectrum is again confined to the disk of radius $\sqrt{c}$. Now B has $q-1$ linearly independent eigenvectors with real eigenvalues, and the corresponding eigenvectors are correlated with the true group assignment. If these real eigenvalues lie outside the bulk, we can identify the groups by assigning a vector in $\mathbb{R}^{q-1}$ to each vertex, and applying a clustering technique such as k means. These eigenvalues are of the form $\mu=c\nu$, where ν is a nonzero eigenvalue of the $q\times q$ matrix

$$T_{ab}=n_a\left(\frac{c_{ab}}{c}-1\right). \qquad [8]$$

In particular, if $n_a=1/q$ for all a, and $c_{ab}=c_{in}$ for $a=b$, and c_{out} for $a\ne b$, we have $\mu_c=(c_{in}-c_{out})/q$. The detectability threshold is again $\mu_c>\sqrt{c}$, or

$$|c_{in}-c_{out}|>q\sqrt{c}. \qquad [9]$$

More generally, if the community-correlated eigenvectors have distinct eigenvalues, we can have multiple transitions where some of them can be detected by a spectral algorithm whereas others cannot.

There is an important difference between the general case and $q=2$. Whereas for $q=2$ it is literally impossible for any algorithm to distinguish the communities below this transition, for larger q the situation is more complicated. In general (for $q\ge 5$ in the assortative case, and $q\ge 3$ in the disassortative one) the threshold **9** marks a transition from an "easily detectable" regime to a "hard detectable" one. In the hard detectable regime, it is theoretically possible to find the communities, but it is conjectured that any algorithm that does so takes exponential time (9, 10). In particular, we have found experimentally that none of B's eigenvectors are correlated with the groups in the hard regime. Nonetheless, our arguments suggest that spectral algorithms based on B are optimal in the sense that they succeed all of the way down to this easy–hard transition.

Because a major drawback of the stochastic block model is that its degree distribution is Poisson, we can also consider random graphs with specified degree distributions. Again, the hardest case is where the groups have the same degree distribution. Let a_k denote the fraction of vertices of degree k. The average branching ratio of a branching process that explores the neighborhood of a vertex, i.e., the average number of new edges leaving a vertex v that we arrive at when following a random edge, is

$$\tilde{c}=\frac{\sum_k k(k-1)a_k}{\sum_k ka_k}=\langle k^2\rangle/\langle k\rangle-1.$$

We assume here that the degree distribution has bounded second moment so that this process is not dominated by a few high-degree vertices. The leading eigenvalue of B is $\tilde{c}$, and the bulk of its spectrum is confined to the disk of radius $\sqrt{\tilde{c}}$, even in the sparse case where $\tilde{c}$ does not grow with the size of the graph. If $q=2$ and the average numbers of new edges linking v to its own group and the other group are $\tilde{c}_{in}/2$ and $\tilde{c}_{out}/2$, respectively, then the approximate eigenvector described in the previous section has eigenvalue $\mu=(\tilde{c}_{in}-\tilde{c}_{out})/2$. The detectability threshold **1** then becomes $\mu>\sqrt{\tilde{c}}$, or $\tilde{c}_{in}-\tilde{c}_{out}>2\sqrt{\tilde{c}}$. The threshold **9** for q groups generalizes similarly.

Deriving *B* by Linearizing Belief Propagation

The matrix B also appears naturally as a linearization of the update equations for belief propagation (BP). This linearization was used previously to investigate phase transitions in the performance of the BP algorithm (5, 9, 10, 29).

We recall that BP is an algorithm that iteratively updates messages $\eta_{v\to w}$ along the directed edges. These messages represent the marginal probability that a vertex v belongs to a given community, assuming that the vertex w is absent from the network. Each such message is updated according to the messages $\eta_{u\to v}$ that v receives from its other neighbors $u\ne w$. The update rule depends on the parameters c_{in} and c_{out} of the block model, as well as the expected size of each community. For the simplest case of two equally sized groups, the BP update (9, 10) can be written as

$$\frac{\eta^+_{v\to w}}{\eta^-_{v\to w}}:=\mathrm{e}^{-h}\frac{\prod_{u\in N(v)-w}\left(\eta^+_{u\to w}c_{in}+\eta^-_{u\to w}c_{out}\right)}{\prod_{u\in N(v)-w}\left(\eta^+_{u\to w}c_{out}+\eta^-_{u\to w}c_{in}\right)}. \qquad [10]$$

Here, + and − denote the two communities. The term e^h, where $h=(c_{in}-c_{out})(n^{BP}_+-n^{BP}_-)$ and $n^{BP}_\pm$ is the current estimate of the

fraction of vertices in the two groups, represents messages from the nonneighbors of v. In the assortative case, it prevents BP from converging to a fixed point where every vertex is in the same community.

The update in Eq. **10** has a trivial fixed point $\eta_{v\to w} = 1/2$, where every vertex is equally likely to be in either community. Writing $\eta^{\pm}_{u\to v} = 1/2 \pm \delta_{u\to v}$ and linearizing around this fixed point gives the following update rule for δ:

$$\delta := \frac{c_{\text{in}} - c_{\text{out}}}{c_{\text{in}} + c_{\text{out}}} B\delta. \qquad [11]$$

More generally, in a block model with q communities, an affinity matrix c_{ab}, and an expected fraction n_a of vertices in each community a, linearizing around the trivial fixed point and defining $\eta^a_{u\to v} = n_a + \delta^a_{u\to v}$ gives a tensor product operator

$$\delta := (T \otimes B)\delta, \qquad [12]$$

where T is the $q \times q$ matrix defined in Eq. **8**.

This shows that the spectral properties of the nonbacktracking matrix are closely related to BP. Specifically, the trivial fixed point is unstable, leading to a fixed point that is correlated with the community structure, exactly when $T \otimes B$ has an eigenvalue greater than 1. However, by avoiding the fixed point where all of the vertices belong to the same group, we suppress B's leading eigenvalue; thus the criterion for instability is $\nu\mu_2 > 1$, where ν is T's leading eigenvalue and μ_2 is B's second eigenvalue. This is equivalent to Eq. **9** in the case where the groups are of equal size.

In general, the BP algorithm provides a slightly better agreement with the actual group assignment, because it approximates the Bayes-optimal inference of the block model. On the other hand, the BP update rule depends on the parameters of the block model, and if these parameters are unknown they need to be learned, which presents additional difficulties (12). In contrast, our spectral algorithm does not depend on the parameters of the block model, giving an advantage over BP in addition to its computational efficiency.

Experimental Results and Discussion

In Fig. 4, we compare the spectral algorithm based on the nonbacktracking matrix B with those based on various classical operators: the adjacency matrix, modularity matrix, Laplacian, normalized Laplacian, and the random walk matrix. We see that there is a regime where standard spectral algorithms do no better than chance, whereas the one based on B achieves a strong correlation with the true group assignment all of the way down to the detectability threshold. We also show the performance of BP, which is believed to be asymptotically optimal (9, 10).

We measure the performance as the overlap, defined as

$$\left(\frac{1}{n}\sum_u \delta_{g_u,\tilde{g}_u} - \frac{1}{q}\right) \Big/ \left(1 - \frac{1}{q}\right). \qquad [13]$$

Here, g_u is the true group label of vertex u, and $\tilde{g}_u$ is the label found by the algorithm. We break symmetry by maximizing over all $q!$ permutations of the groups. The overlap is normalized so that it is 1 for the true labeling, and 0 for a uniformly random labeling.

In Fig. 5 we illustrate clustering in the case $q = 3$. As described above, in the detectable regime we expect to see $q - 1$ eigenvectors with real eigenvalues that are correlated with the true group assignment. Indeed, B's second and third eigenvectors are strongly correlated with the true clustering, and applying k means in $\mathbb{R}^2$ gives a large overlap. In contrast, the second and third eigenvectors of the adjacency matrix are essentially uncorrelated with the true clustering, and similarly for the other traditional operators.

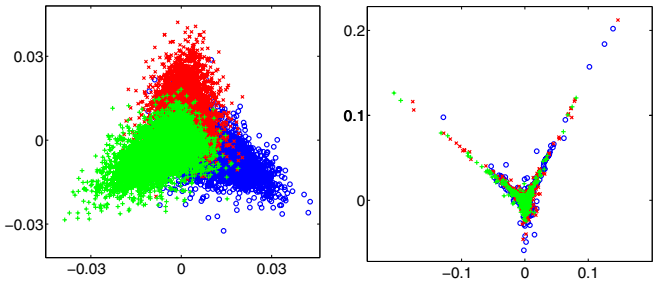

Fig. 5. Clustering for three groups of equal size. (*Left*) Scatter plot of the second and third eigenvectors (X and Y axis, respectively) of the nonbacktracking matrix B, with colors indicating the true group assignment. (*Right*) Analogous plot for the adjacency matrix A. Here, $n = 3 \times 10^4$, $c = 3$, and $c_{\text{out}}/c_{\text{in}} = 0.1$. Applying k means gives an overlap 0.712 using B, but 0.0063 using A.

Finally, we turn to real networks to illustrate the advantages of the nonbacktracking matrix in practical applications. In Fig. 6 we show B's spectrum for several networks commonly used as benchmarks for community detection. In each case we plot a circle whose radius is the square root of the largest eigenvalue. Even though these networks were not generated by the stochastic block model, these spectra look qualitatively similar to the picture discussed above (Fig. 2). This leads to several very convenient properties. For each of these networks we observed that only the eigenvectors with real eigenvalues are correlated to the group assignment given by the ground truth. Moreover, the real eigenvalues that lie outside of the circle are clearly identifiable. This is very unlike the situation for the operators used in standard spectral clustering algorithms, where one must decide which eigenvalues are in the bulk and which are outside.

In particular, the number of real eigenvalues outside the circle seems to be a natural indicator for the true number of clusters in the network, just as for networks generated by the stochastic block model. This suggests that in the network of political books there might be 4 groups rather than 3, in the blog network there might be more than 2 groups, and in the NCAA football network there might be 10 groups rather than 12. However, we note that some real eigenvalues may correspond to small cliques.

A Matlab implementation with demos that can be used to reproduce our numerical results can be found at http://panzhang.net/dea/dea.tar.gz.

Conclusion

Although recent advances have made statistical inference of network models for community detection far more scalable than in the past (e.g., refs. 9, 24, 36, 37), spectral algorithms are highly competitive because of the computational efficiency of sparse linear algebra. However, for sparse networks there is a large regime in which statistical inference methods such as BP can detect communities, whereas standard spectral algorithms cannot.

We closed this gap by using the nonbacktracking matrix B as a starting point for spectral algorithms. We showed that for sparse networks generated by the stochastic block model, B's spectral properties are much better than those of the adjacency matrix and its relatives. In fact, it is asymptotically optimal in the sense that it allows us to detect communities all of the way down to the detectability transition. We also computed B's spectrum for some real-world networks, showing that the real eigenvalues are a good guide to the number of communities and the correct labeling of the vertices.

Our approach can be straightforwardly generalized to spectral clustering for other types of sparse data, such as weighted graphs with real-valued similarities $s(u,v)$ between vertices: then $B_{(u\to v),(w\to x)} = s(u,v)$ if $v = w$ and $u \neq x$, and 0 otherwise. We

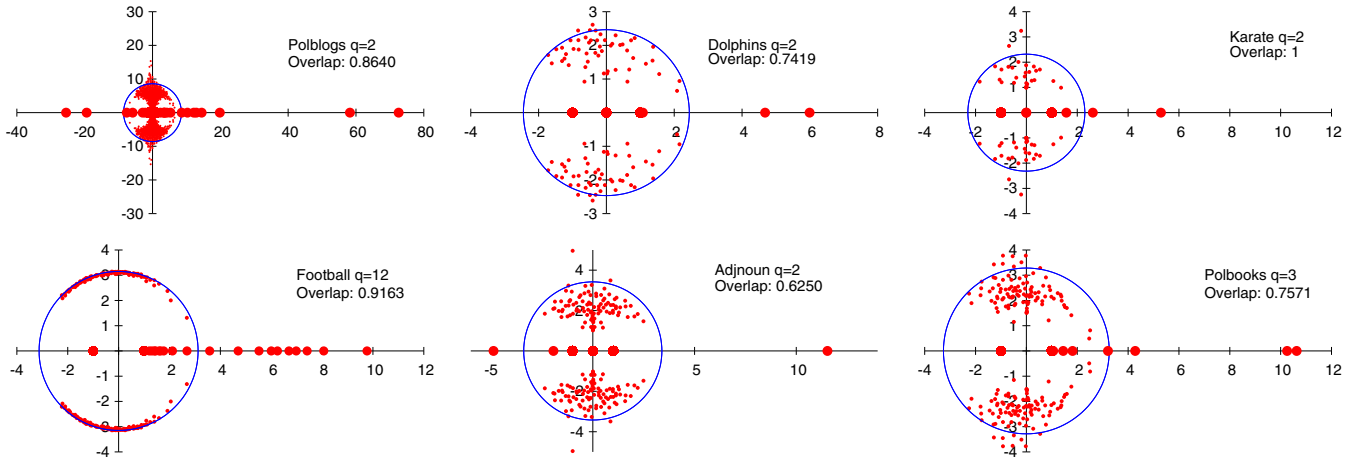

Fig. 6. Spectrum of the nonbacktracking matrix in the complex plane for some real networks commonly used as benchmarks for community detection, taken from refs. 30–35. The radius of the circle is the square root of the largest eigenvalue, which is a heuristic estimate of the bulk of the spectrum. The overlap is computed using the signs of the second eigenvector for the networks with two communities, and using k means for those with three and more communities. The nonbacktracking operator detects communities in all these networks, with an overlap comparable to the performance of other spectral methods. As in the case of synthetic networks generated by the stochastic block model, the number of real eigenvalues outside the bulk appears to be a good indicator of the number q of communities.

believe that, as for sparse graphs, there will be important regimes in which using B will succeed where standard clustering algorithms fail. Given the wide use of spectral clustering throughout the sciences, we expect that the nonbacktracking matrix and its generalizations will have a significant impact on data analysis.

ACKNOWLEDGMENTS. We are grateful to Noga Alon, Brian Karrer, Michael Krivelevich, Mark Newman, Nati Linial, Yuval Peres, and Xiaoran Yan for helpful discussions. C.M. and P.Z. are supported by Air Force Office of Scientific Research and Defense Advanced Research Planning Agency under Grant FA9550-12-1-0432. F.K. and P.Z. have been supported in part by the European Research Council under the European Union's 7th Framework Programme Grant Agreement 307087-SPARCS. E.M. and J.N. were supported by National Science Foundation (NSF) Department of Mathematical Sciences Grant 1106999 and Department of Defense Office of Naval Research Grant N000141110140. E.M was supported by NSF Grant Computing and Communication Foundation 1320105.

1. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. *Soc Networks* 5(2):109–137.
2. Wang YJ, Wong GY (1987) Stochastic blockmodels for directed graphs. *J Am Stat Assoc* 82(397):8–19.
3. Von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17(4):395–416.
4. Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman-Girvan and other modularities. *Proc Natl Acad Sci USA* 106(50):21068–21073.
5. Coja-Oghlan A, Mossel E, Vilenchik D (2009) A spectral approach to analyzing belief propagation for 3-coloring. *Combin Probab Comput* 18(6):881–912.
6. Coja-Oghlan A (2010) Graph partitioning via adaptive spectral techniques. *Combin Probab Comput* 19(02):227–284.
7. McSherry F (2001) Spectral partitioning of random graphs. *Proceedings of 42nd Foundations of Computer Science* (IEEE Computer Society, Las Vegas), pp 529–537.
8. Nadakuditi RR, Newman MEJ (2012) Graph spectra and the detectability of community structure in networks. *Phys Rev Lett* 108(18):188701.
9. Decelle A, Krzakala F, Moore C, Zdeborová L (2011) Inference and phase transitions in the detection of modules in sparse networks. *Phys Rev Lett* 107(6):065701.
10. Decelle A, Krzakala F, Moore C, Zdeborová L (2011) Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys Rev E Stat Nonlin Soft Matter Phys* 84(6 Pt 2):066106.
11. Mossel E, Neeman J, Sly A (2012) Stochastic block models and reconstruction. arXiv: 1202.1499v4.
12. Zhang P, Krzakala F, Reichardt J, Zdeborová L (2012) Comparative study for inference of hidden classes in stochastic block models. *J Stat Mech* 12:P12021. Available at http://iopscience.iop.org/1742-5468/2012/12/P12021. Accessed November 14, 2013.
13. McKay BD (1981) The expected eigenvalue distribution of a large regular graph. *Linear Algebra Appl* 40:203–216.
14. Sodin S (2007) Random matrices, nonbacktracking walks, and orthogonal polynomials. *J Math Phys* 48:123503.
15. Friedman J (2008) A proof of Alon's second eigenvalue conjecture and related problems. *Memoirs of the American Mathematical Society* 195(910).
16. Hashimoto K (1989) Zeta functions of finite graphs and representations of p-adic groups. *Advanced Studies in Pure Mathematics* 15:211–280.
17. Alon N, Benjamini I, Lubetzky E, Sasha S (2007) Non-backtracking random walks mix faster. *Commun Contemp Math* 9(4):585–603.
18. Watanabe Y, Fukumizu K (2010) Graph zeta function in the Bethe free energy and loopy belief propagation. arXiv:1002.3307.
19. Vontobel PO (2010) Connecting the Bethe entropy and the edge zeta function of a cycle code. *Proceedings of the International Symposium on Information Theory* (IEEE, New York), pp 704–708.
20. Ren P, Wilson RC, Hancock ER (2011) Graph characterization via Ihara coefficients. *IEEE Trans Neural Netw* 22(2):233–245.
21. Wigner EP (1958) On the distribution of the roots of certain symmetric matrices. *Ann Math* 67(2):325–327.
22. Krivelevich M, Sudakov B (2003) The largest eigenvalue of sparse random graphs. *Combin Probab Comput* 12(01):61–72.
23. Bollobas B, Svante J, Oliver R (2007) The phase transition in inhomogeneous random graphs. *Random Structures Algorithms* 31(1):3–122.
24. Amini AA, Chen A, Bickel PJ, Levina E (2012) Pseudolikelihood methods for community detection in large sparse networks. arXiv:1207.2340.
25. Bass H (1992) The Ihara-Selberg zeta function of a tree lattice. *Int J Math* 3(06): 717–797.
26. Angel O, Friedman J, Hoory S (2007) The non-backtracking spectrum of the universal cover of a graph. arXiv:0712.0192.
27. Kesten H, Stigum BP (1966) Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann Math Stat* 37(6):1463–1481.
28. Mossel E, Peres Y (2003) Information flow on trees. *Ann Appl Probab* 13:817–844.
29. Richardson T, Urbanke R (2008) *Modern Coding Theory* (Cambridge Univ Press, Cambridge, UK).
30. Adamic L, Glance N (2005) The political blogosphere and the 2004 US Election: Divided they blog. *Proceedings of the 3rd International Workshop on Link Discovery* (Association for Computing Machinery, New York), pp 36–43.
31. Zachary WW (1977) An information flow model for conflict and fission in small groups. *J Anthropol Res* 33:452–473.
32. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E Stat Nonlin Soft Matter Phys* 74(3 Pt 2):036104.
33. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99(12):7821–7826.
34. Lusseau D, et al. (2003) The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav Ecol Sociobiol* 54(4):396–405.
35. Krebs V (2012) Social Network Analysis software & services for organizations, communities, and their consultants. Available at www.orgnet.com/. Accessed November 14, 2013.
36. Ball B, Karrer B, Newman MEJ (2011) Efficient and principled method for detecting communities in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 84(3 Pt 2): 036103.
37. Gopalan P, Mimno D, Gerrish S, Freedman M, Blei D (2012) Scalable inference of overlapping communities. *Advances in Neural Information Processing Systems 25*, eds Bartlett P, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, pp 2258–2266.