Electronic Journal of Statistics
Vol. 16 (2022) 3315–3342
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1971

Estimating the number of communities by spectral methods

Can M. Le

Department of Statistics,
University of California, Davis
e-mail: canle@ucdavis.edu

Elizaveta Levina

Department of Statistics,
University of Michigan
e-mail: elevina@umich.edu

Abstract: Community detection is a fundamental problem in network analysis with many methods available to estimate communities. Most of these methods assume that the number of communities is known, which is often not the case in practice. We study a simple and very fast method for estimating the number of communities based on the spectral properties of certain graph operators, such as the non-backtracking matrix and the Bethe Hessian matrix. We show that the method performs well under several models and a wide range of parameters, and is guaranteed to be consistent under several asymptotic regimes. We compare this method to several existing methods for estimating the number of communities and show that it is both more accurate and more computationally efficient.

MSC2020 subject classifications: Primary 62H12; secondary 62H30.
Keywords and phrases: Community detection, stochastic block model, network analysis.

Received January 2021.

Contents

1. Introduction

The problem of clustering similar objects into groups is a fundamental problem in data analysis. In network analysis, it is known as community detection ([34, 3, 10, 4]). Given a network, which consists of a set of nodes and a set of edges between them, the goal of community detection is to cluster the nodes into groups (communities) so that nodes in the same community share a similar connectivity.

One of the simplest ways of modeling a community structure is the stochastic block model (SBM), proposed by [17]. Given the number of communities K, n node labels c_i are drawn independently from a multinomial distribution with parameter $\pi = (\pi_1, ..., \pi_K)$. The edges between pairs of nodes (i, j) are then drawn independently from a Bernoulli distribution with parameter $P_{c_i c_j}$ and collected in the $n \times n$ adjacency matrix A, with $A_{ij} = 1$ if nodes i and j are connected by an edge, and 0 otherwise. A limitation of the stochastic block model is that all nodes in the same communities are equivalent and follow the same degree distribution, whereas many real networks contain a small number of high-degree nodes, the so called hubs. To address this limitation, [19] proposed the degree-corrected stochastic block model (DCSBM). It assigns a degree parameter θ_i to each node i, and edges between nodes are drawn independently with probabilities $\theta_i \theta_j P_{c_i c_j}$. The community detection task is to recover the labels c_i given the adjacency matrix A.

A large number of methods have been proposed for finding the underlying community structure ([28, 33, 3, 10, 37, 12, 4, 20, 42, 30, 38]). Most of these methods require the number of communities K as input, but in practice K is often unknown. To address this problem, a few likelihood-based methods have been proposed to estimate K under either the SBM or the DCSBM ([14, 21, 35, 39, 44]). These methods use BIC-type criteria for choosing the number of communities from a set of possible values, which requires computing the likelihood, done using either MCMC or the variational method, which are both computationally very challenging for large networks. A different approach based on the distribution of leading eigenvalues of an appropriately scaled version of the adjacency matrix was proposed by [9, 23]. Under the SBM, distributions of the leading eigenvalues converge to the Tracy-Widom distribution; this fact is used to determine K through a sequence of hypothesis tests. Since the rate of convergence is slow for relatively sparse networks, a bootstrap correction procedure was employed, which also leads to a high computational cost. Cross-validation approaches were proposed by [13] and [24]. While they have good properties under the SBM and the DCSBM, they require estimating communities on many random network splits, and are computationally costly.

To the best of our knowledge, all existing methods are either restricted to a specific model or computationally intensive. In this paper we study a fast and reliable method that uses spectral properties of either the Bethe Hessian or the non-backtracking matrices. Under a simple SBM in the sparse regime, these matrices have been used to recover the community structure ([20, 38, 11]); It was observed in the physics literature that the informative eigenvalues (i.e., those corresponding to eigenvectors which encode the community structure) of these matrices are well separated from the bulk and can be used to estimate the number of communities, but the properties of this estimator have never been investigated, either theoretically or empirically. We show that the number of "informative" (to be defined explicitly below) eigenvalues of these matrices directly estimates the number of communities, and the estimate performs well under different network models and over a wide range of parameter values, outperforming existing methods designed specifically for estimating K under either SBM or DCSBM. This method is extremely computationally efficient, since all it requires is computing a few leading eigenvalues of just one typically sparse matrix, and to the best of ourknowledge, is by far the fastest available accurate method for estimating the number of communities.

Several new methods for estimating the number of communities K have been developed concurrently with the present paper. For example, [36] use a variant of the Chinese restaurant process to generate community assignments, which automatically yields a choice of K; this method is implemented via a Monte Carlo sampling scheme, which is computationally intensive. A method based on semidefinite programming, another very computationally intensive technique, was derived and proved to be consistent for assortative networks by [45]. Improving on [44], the authors of [18] proposed a corrected BIC criterion in [44] to correct for under-estimation. More recently, [26] combined spectral clustering with binary segmentation to derive a new estimate of K. Compared to all these new methods, the estimators based on Bethe Hessian or non-backtracking matrices we study is still the most computationally efficient, arguably the simplest, and competitive on estimation accuracy (see [26] for some numerical comparisons). The theoretical analysis of the Bethe-Hessian and the nonbactracking matrices we provide in this paper explain this performance and cover a wider range of settings, including sparse, dense, assortative and disassortative networks; no other method is known to be applicable under a wider range of settings, and most are narrower.

2. Preliminaries

Recall A is the $n \times n$ symmetric network adjacency matrix. Let $d_i = \sum_{j=1}^{n} A_{ij}$ be the degree of node i. Treating A as a random matrix, let $\mathbb{E}\, A$ be the expectation of A (conditioned on c_i and θ_i), and let $d = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\, d_i$ be the average expected node degree.

2.1. The non-backtracking matrix

Let m be the number of edges in an undirected network, $2m = \sum_{i,j=1}^{n} A_{ij}$. To construct the non-backtracking matrix, we represent the edge between node i and node j by two directed edges, one from i to j and the other from j to i. The $2m \times 2m$ non-backtracking matrix $\tilde{B}$, indexed by these directed edges, is defined by

$$\tilde{B}_{i\to j,k\to l} = \begin{cases} 1 & \text{if } j = k \text{ and } i \neq l \\ 0 & \text{otherwise.} \end{cases}$$

It is well-known [5, 20] that the spectrum of $\tilde{B}$ consists of ± 1 and eigenvalues of an $2n \times 2n$ matrix

$$B = \begin{pmatrix} A & I_n - D \\ I_n & 0_n \end{pmatrix}. \tag{2.1}$$

Here 0_n is the $n \times n$ matrix of all zeros, I_n is the $n \times n$ identity matrix, and $D = \operatorname{diag}(d_i)$ is $n \times n$ diagonal matrix with degrees d_i on the diagonal. It was observed by [20] that if a network has K communities then the first K largest (in absolute value) eigenvalues of B are real-valued and well separated from the bulk, which is contained in a circle of radius $\sqrt{\rho(B)}$, where $\rho(B)$ is the spectral radius of B. We refer to these K eigenvalues as informative eigenvalues of B. It was also shown by [20] that the spectral norm of the non-backtracking matrix is approximated by

$$\tilde{d} = \Big(\sum_{i=1}^{n} d_i\Big)^{-1} \Big(\sum_{i=1}^{n} d_i^2\Big) - 1. \tag{2.2}$$

For a special case of a sparse SBM with a bounded expected node degree, [11] proved that the leading eigenvalues of B concentrate around non-zero eigenvalues of $\mathbb{E}\, A$ and the bulk is contained in a circle of radius $\sqrt{\rho(B)}$, and used the corresponding leading eigenvectors to recover the community labels. The spectrum of B for denser Erdős-Rényi graphs was later analyzed in [43]. In particular, if $d \gg n^{5/6}$, then every eigenvalue of $(d-1)^{-1/2}B$ is within a vanishing distance from a limiting spectrum supported on the unit circle of the complex plane (hereafter, we use $a_n \gg b_n$ or $b_n \ll a_n$ to denote that there exists a sufficiently large constant $C > 0$ such that $a_n > Cb_n$ for all but possibly a finite set of values of n). In Theorem A.1 below we extend this result to much sparser and more general random graphs and require only that $d \gg \log n$.

2.2. The Bethe Hessian matrix

The Bethe Hessian matrix is defined by

$$H(r) = (r^2 - 1)I - rA + D, \tag{2.3}$$

where $r \in \mathbb{R}$ is a parameter. In graph theory, the determinant of $H(r)$ is the Ihara-Bass formula for the graph zeta function. It vanishes if r is an eigenvalue

TABLE 1
Spectral methods for estimating the number of communities.

Method	Parameter	Estimated number of communities $\hat{K}$
NB	None	$\left\lvert\left\{\lambda(B) \in \mathbb{R} : \lambda(B) \geq \sqrt{\rho(B)}\right\}\right\rvert$
BHm	$r_m = \left(\frac{\sum_{i=1}^n d_i^2}{\sum_{i=1}^n d_i} - 1\right)^{1/2}$	$\max\{k : \lambda_{n-k}(H(r_m)) \leq 0\}$
BHmc	$r_m = \left(\frac{\sum_{i=1}^n d_i^2}{\sum_{i=1}^n d_i} - 1\right)^{1/2}$	$\max\{k : t\lambda_{n-k+1}(H(r_m)) \leq \lambda_{n-k}(H(r_m))\}$
BHa	$r_a = \left(\frac{1}{n}\sum_{i=1}^n d_i\right)^{1/2}$	$\max\{k : \lambda_{n-k}(H(r_a)) \leq 0\}$
BHac	$r_a = \left(\frac{1}{n}\sum_{i=1}^n d_i\right)^{1/2}$	$\max\{k : t\lambda_{n-k+1}(H(r_a)) \leq \lambda_{n-k}(H(r_a))\}$

of the non-backtracking matrix [16, 6, 5]. The Bethe Hessian was used for community detection by [38] Under the SBM, they argued that the best choice of r is $r_c = \pm\sqrt{d}$, depending on whether the network is assortative or disassortative; for a more general network, they take $r_c = \pm\sqrt{\rho(B)}$. For assortative sparse networks with K communities and a bounded d, they empirically showed that the K eigenvalues of $H(r_c)$ whose corresponding eigenvectors encode the community structure are negative, while the bulk of $H(r_c)$ are positive. Thus, the number of negative eigenvalues of $H(r_c)$ corresponds to the number of communities. In Theorem 4.3 below, we prove that this method is indeed consistent for graphs with $d \gg \log n$. See also the discussion following Theorem 4.3 for more intuition of why the number of negative eigenvalues of H coincides with the number of communities.

3. Spectral estimates of the number of communities

The spectral properties of the non-backtracking and the Bethe Hessian matrices lead to natural estimates of the number of communities, but they have not been previously considered in this context. We next outline several spectral methods to determine the number of communities K. They are based on simple counts of eigenvalues of either the non-backtracking matrix or the Bethe Hessian matrix, and therefore do not require any adjustment for different models such as SBM or DCSBM. We list them in Table 1, and proceed to explain the motivation for each one.

3.1. Estimating K from the non-backtracking matrix

As we will show in Theorems 4.1 and 4.2 under the SBM, the informative eigenvalues of the non-backtracking matrix are real-valued and separated from the bulk of radius $\sqrt{\rho(B)}$. Therefore we can estimate K by counting the number of real eigenvalues of B that are at least $\sqrt{\rho(B)}$. We denote this method by NB (for non-backtracking). As shown by Theorem 4.2 and numerical results in Section 5, this estimate of K also works under much more general models with

low-rank structure such as DCSBM. When the network is balanced (communities have similar sizes and edge densities), NB performs well; however, the accuracy of NB drops if the communities are unbalanced in either size or edge density. Since B is not symmetric, computing the eigenvalues of B is slightly more demanding than that of the Bethe Hessian matrix for large networks.

3.2. Estimating K from the Bethe Hessian matrix

The number of communities corresponds to the number of negative eigenvalues of $H(r)$; the challenge is in choosing an appropriate value of r. It was argued by [38] that when $r = \sqrt{\rho(B)}$, the informative eigenvalues of $H(r)$ are negative, while the bulk are positive; by [20], $\rho(B)$ can be approximated by $\tilde{d}$ from (2.2). Following these results, we first choose r to be $r_m = \tilde{d}^{1/2}$ and call the corresponding method BHm. Simulations show that using $r = r_m$ and $r = \sqrt{\rho(B)}$ produce similar results; we choose $r = r_m$ because computing r_m is less demanding than computing $\sqrt{\rho(B)}$.

Another choice of r is $r_a = \sqrt{(d_1 + \cdots + d_n)/n}$, which was proposed by [38] for recovering the community structure under the SBM; we call the corresponding method BHa. We have found that when the network is balanced, NB, BHm and BHa perform similarly; when the network is unbalanced, BHa produces better results.

Both BHm and BHa tend to underestimate the number of communities, especially when the network is unbalanced. In that setting, some informative eigenvalues of $H(r)$ become positive, although they may still be far from the bulk. Based on this observation, we correct BHm and BHa by also using positive eigenvalues of $H(r)$ that are much close to zero than to the bulk. Namely, we sort eigenvalues of $H(r)$ in non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and estimate K by

$$\hat{K} = \max\{k : t\lambda_{n-k+1} \leq \lambda_{n-k}\}, \tag{3.1}$$

where $t \geq 1$ is a tuning parameter. Note that if $\lambda_{n-k_0+1} < 0$ then $\hat{K} \geq k_0$ because $\lambda_{n-k_0+1} \leq \lambda_{n-k_0}$, therefore the number of negative eigenvalues of $H(r)$ is always upper bounded by $\hat{K}$. Heuristically, if the bulk follows the semi-circular law and $\lambda_{n-k} \geq 0$ is given, then the probability that $0 \leq \lambda_{n-k+1} \leq \lambda_{n-k}/t$ is less than $1/t$. When $1/t$ is sufficiently small, we may suspect that λ_{n-k+1} is an informative eigenvalue. In practice we find that $t \in [4, 6]$ works well; we will set $t = 5$ for all computations in this paper. Simulations show that $\hat{K}$ performs well, especially for unbalanced networks. The resulting methods are denoted by BHmc and BHac, respectively. We will also use BH to refer to all the methods that use the Bethe Hessian matrix. For a summary of these methods, see Table 1.

4. Consistency

The consistency of the non-backtracking matrix based method (NB) for estimating the number of communities in the sparse regime under the stochastic

block model with certain regularity conditions follows directly from Theorem 4 of [11]. We state this consistency result here for completeness. The proof given by [11] is combinatorial in nature and this approach unfortunately does not extend to any other regimes or the Bethe-Hessian matrix.

Theorem 4.1 (Consistency in the sparse regime)**.** *Consider a stochastic block model with* $\pi = (\pi_1, ..., \pi_K)$ *and* $P = (P_{kl}) = \frac{1}{n}P^{(0)}$ *for some fixed* $K \times K$ *symmetric matrix* $P^{(0)}$ *of rank* K*. Assume that* $(\mathrm{diag}(\pi)P)^r$ *has positive entries for some positive integer* r*. Further, assume that* $E(d_i) = d > 1$ *for all* i*, and the absolute values of all* K *non-zero eigenvalues of* P *are strictly larger than* $\sqrt{d}$*. Then with probability tending to one as* $n \to \infty$*, the number of real eigenvalues of* B *that are at least* $\sqrt{\rho(B)}$ *is equal to* K*.*

To better understand the condition on the eigenvalues of P, consider the simple model $G(n, \frac{a}{n}, \frac{b}{n})$. This model assumes that there are two communities of equal sizes and nodes are connected with probability a/n if they are in the same community, and b/n otherwise. Since the two non-zero eigenvalues of P are $(a+b)/2$ and $(a-b)/2$, the condition on eigenvalues of P is $(a-b)^2 > 2(a+b)$. This matches the phase transition condition for the detectability in the sparse regime [29, 31, 27].

Next, we prove the consistency of the proposed methods in the denser regime $d \gg \log n$, sometimes referred to as semi-dense in contrast to the dense regime of $d = O(n)$. For this regime, we make the following assumptions. Hereafter, we use C to denote a positive constant that is sufficiently large and its value can change from line to line.

Assumption 4.1. All nodes have the same expected degree satisfying

$$\mathbb{E}\sum_{j=1}^{n} A_{ij} = d \geq C \log n, \quad 1 \leq i \leq n.$$

Assumption 4.2. Matrix $\mathbb{E}\,A$ is of rank K and nonzero eigenvalues of $\mathbb{E}\,A$ satisfy

$$|\lambda_1(\mathbb{E}\,A)| \geq |\lambda_2(\mathbb{E}\,A)| \geq \cdots \geq |\lambda_K(\mathbb{E}\,A)| \geq 4d^{1/2} + C(d^{1/4} + (\log n)^{1/2}).$$

Assumption 4.3. The expected degree d in Assumption 4.1 satisfies

$$d^5 \max_{i,j} \mathbb{E}\,A_{ij} \leq n^{-1/13}.$$

Following [11], we assume in Assumption 4.1 that all nodes have the same expected degree. This corresponds perhaps to the most challenging setting where expected degrees alone do not contain information about the latent structure of interest. As in [11] and [43], this assumption allows us to simplify our theoretical analysis of the non-backtracking matrix considerably, although numerical results in Section 5 show that the method still performs well and remains competitive when this assumption no longer holds. If some communities have different expected degrees, we can first use node degrees to identify them and divide the

network into sub-networks of similar expected node degrees and apply our results on the sub-networks. Note that for the degree-corrected stochastic block model, if the underlying stochastic block model satisfies this assumption and the degree parameters are drawn from the same distribution, then the degree-corrected stochastic block model itself will also satisfy the assumption.

The lower bound on $\lambda_K(\mathbb{E}\,A)$ in Assumption 4.2 is of the form $|\lambda_K(\mathbb{E}\,A)| \geq 4(1+o(1))\sqrt{d}$ when $d \gg \log n$. Under $G(n, \frac{a}{n}, \frac{b}{n})$, this bound is $(a-b)^2 \geq 32(1+o(1))(a+b)$. For a comparison, exact community recovery under $G(n, \frac{a}{n}, \frac{b}{n})$ with known number of communities requires $(a-b)^2 > 2(a+b+2\sqrt{ab})\log n$ (see e.g. [1, Theorem 13]).

Assumption 4.3 guarantees a sharp bound on $\|A - \mathbb{E}\,A\|$, which is established by [7]. We use this bound in the proofs of Theorem 4.2 and Theorem 4.3 below. For the Erdős-Rényi model, Assumption 4.3 is equivalent to $d \leq n^{2/13}$. It is unclear if this condition can be removed from the result of [7] and consequently from Theorem 4.2 and Theorem 4.3.

Theorem 4.2 (Consistency of NB based method in the semi-dense regime)**.** *Consider random graphs that satisfy Assumptions 4.1, 4.2 and 4.3. Then with probability at least* $1-1/n$*, the nonbacktracking matrix has exactly* K *real eigenvalues with magnitude at least* $(1+\varepsilon)\sqrt{d}$ *and the remaining eigenvalues are of magnitude smaller than* $(1+\varepsilon)\sqrt{d}$*, where*

$$\varepsilon = C\left[\left(\frac{\log n}{d}\right)^{1/4} + \left(\frac{1}{d}\right)^{1/8}\right].$$

According to Theorem 4.2, the K informative eigenvalues of the nonbacktracking matrix are separated from the bulk by a circle of radius $(1+\varepsilon)\sqrt{d}$, where ε is vanishing if d grows faster than $\log n$. Unlike in Theorem 4.1, K is allowed to depend on n in Theorem 4.2.

To compute this estimator in practice, we simply set $\varepsilon = 0$ and estimate d with the average observed degree $\bar{d} = (d_1 + \cdots + d_n)/n$. It is straightforward to show that $\bar{d}$ is close to d with high probability.

Let us briefly describe the main ideas in the proof of Theorem 4.2. Denote $\Gamma = \mathbb{E}(D-I)$. As pointed out in [43], B and the following conjugation matrix admit the same spectrum:

$$\begin{pmatrix} \Gamma^{-1/2} & 0_n \\ 0_n & I_n \end{pmatrix} \begin{pmatrix} A & I_n - D \\ I_n & 0_n \end{pmatrix} \begin{pmatrix} \Gamma^{1/2} & 0_n \\ 0_n & I_n \end{pmatrix}$$
$$= \begin{pmatrix} \Gamma^{1/2} & 0_n \\ 0_n & \Gamma^{1/2} \end{pmatrix} \begin{pmatrix} \Gamma^{-1}A\Gamma^{1/2} & \Gamma^{-1}(I_n - D) \\ I_n & 0_n \end{pmatrix}$$

Under Assumption 4.1, $\Gamma = (d-1)I_n$ and the right-hand side of the above equality greatly simplifies. Consequently, it is sufficient to study the spectrum of the following matrix

$$\begin{pmatrix} \frac{1}{\sqrt{d-1}}A & \frac{1}{d-1}(I_n - D) \\ I_n & 0_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{d-1}}A & -I_n \\ I_n & 0_n \end{pmatrix} + \begin{pmatrix} 0_n & \frac{1}{d-1}(\mathbb{E}\,D - D) \\ 0_n & 0_n \end{pmatrix}.$$

The last term on the right-hand side of the above equality can be viewed as a noise term. Thus, the spectrum of $(d-1)^{-1/2}B$ is a perturbation of the spectrum of matrix

$$\begin{pmatrix} \frac{1}{\sqrt{d-1}}A & -I_n \\ I_n & 0_n \end{pmatrix},$$

which is directly related to the spectrum of A via an explicit mapping; see Appendix A.1 for detail.

The main difficulty in this analysis comes from the fact that the above matrix is not symmetric, so many standard perturbation analysis techniques from random matrix theory do not apply. To address this problem, [43] uses Bauer-Fike theorem and the replacement principle [40] to show that for Erdős-Rényi random graphs, the above idea works if $d \gg n^{5/6}$. Using a more direct analysis, we are able to replace this condition with the much weaker condition $d \gg \log n$ and extend the validity of the result way beyond the Erdős-Rényi model. For detail, see Theorem A.1 in Appendix A, which may also be of independent interest.

Note that when Assumption 4.1 does not hold, the spectrum of the non-backtracking matrix depends on the node degree distribution through matrix

$$\begin{pmatrix} \Gamma^{1/2} & 0_n \\ 0_n & \Gamma^{1/2} \end{pmatrix}.$$

This explains why the performance of the method based on spectrum of B is influenced by the severe heterogeneity of node degrees (simulations show that other methods are affected as well).

For the Bethe Hessian, no formal results have been previously established. We show in the following theorem that both BHm and BHa methods produce consistent estimators of $K = \operatorname{rank}(\mathbb{E}\,A)$, provided that the following stronger version of Assumption 4.2 holds.

Assumption 4.4. Matrix $\mathbb{E}\,A$ is of rank K and nonzero eigenvalues of $\mathbb{E}\,A$ satisfy

$$\lambda_1(\mathbb{E}\,A) \geq \lambda_2(\mathbb{E}\,A) \geq \cdots \geq \lambda_K(\mathbb{E}\,A) \geq 4d^{1/2} + C(d^{1/4} + (\log n)^{1/2}).$$

Note that Assumption 4.2 allows networks to be disassortative, meaning probabilities of connections between communities are higher than within communities, in which case the eigenvalues of $\mathbb{E}\,A$ may be negative. In contrast, Assumption 4.4 requires all eigenvalues of $\mathbb{E}\,A$ to be non-negative.

Theorem 4.3 (Consistency of the Bethe Hessian matrix method)**.** *Consider random graphs that satisfy Assumptions 4.1, 4.3 and 4.4. Then with probability at least* $1 - 1/n$*, the Bethe Hessian* $H(r)$ *with* $r = (1+\varepsilon)r_m$ *or* $r = (1+\varepsilon)r_a$ *and* $\varepsilon = C\sqrt{\log n/d}$ *has exactly* K *negative eigenvalues.*

To describe the main idea in the proof of this result, let us rewrite $H(r)$ as follows:

$$H(r) = (r^2-1)I - r(A - \mathbb{E}\,A) + D - r\,\mathbb{E}\,A =: \hat{H}(r) - r\,\mathbb{E}\,A.$$

Using a recent sharp concentration bound [7], it can be shown that both $r\|A - \mathbb{E}\,A\|$ and $\|(r^2-1)I + D\|$ are of order $2d$ if $r = \sqrt{d}$. Moreover, under some

conditions, $\hat{H}(r)$ is positive semi-definite and $\hat{H}(r)$ is of smaller order than $r\,\mathbb{E}\,A$ if they are restricted to the subspace formed by the first K eigenvectors of $\mathbb{E}\,A$. If $\operatorname{rank}(\mathbb{E}\,A) = K$ and the network is assortative, this implies $H(r)$ has exactly K negative eigenvalues.

Note that to show the positive semi-definiteness of $\hat{H}(r)$, we need to compare $(1/d)D$ and I_n, and Assumption 4.1 is convenient for that purpose. It also indicates that the accuracy of the proposed methods may drop as the node degree heterogeneity increases. This is confirmed by numerical results in Section 5, although our methods remain competitive and often outperform existing methods even when Assumption 4.1 is violated.

Again in practice, we set $\varepsilon = 0$ to compute the estimator.

Theorem 4.2 and Theorem 4.3 show the consistency of the proposed methods. Besides Assumption 4.1, they mainly require that the K-th eigenvalue of $\mathbb{E}\,A$ is at least $4d^{1/2}$ and $d \gg \log n$. To put them in perspective, let us discuss some existing theoretical results for estimating K. Using a sequence of hypothesis tests, [9] shows that K can be consistently estimated if d grows linearly in n. This requirement is relaxed to $d \gg n^{1/2}$ in [13] and $d \gg n^{1/3}\log^{4/3} n$ in [24], where network cross-validation is implemented. Another computationally demanding method based on semi-definite programming by [45] requires $d \gg \log n$, but it is only consistent for assortative networks. The same condition is needed for the likelihood-based method in [44], the corrected BIC criterion in [18] and the binary segmentation method in [26], although they are computationally more intensive than the proposed methods in this paper. Note also that none of these results covers the sparse setting $d = O(1)$ in which the method based on the spectrum of the non-backtracking matrix remains consistent. Thus, compared to these existing methods, our estimators are computationally more efficient and often require weaker assumptions.

5. Numerical results

In this section, we briefly compare the empirical accuracy of estimating the number of communities by using the non-backtracking matrix (NB), and all the versions based on the Bethe Hessian matrix (BHm, BHmc, BHa, and BHac), described in Section 3.1 and Section 3.2. We compare them with two other methods representative of approaches in the literature to estimating the number of communities in networks: the network cross-validation method (NCV) proposed by [13] and a likelihood-based BIC-type method (VLH, for variational likelihood) proposed by [44]. We use NCVbm and NCVdc to denote the versions of the NCV method specifically designed for the SBM and the DCSBM, respectively; VLH is only designed to work under the SBM, so it is not included in the DCSBM comparisons. To make comparisons with VLH computationally feasible, instead of using the variational method to estimate the posterior of the community labels as done by [44], we first estimate the node labels by the pseudo-likelihood method proposed by [4] and then compute the posterior following [44]. In small-scale simulations where both approaches are computa-

tionally feasible (results omitted) we found that substituting pseudo-likelihood for the variational method has very little effect on the estimate of K. The tuning parameter of VLH is set to one following [44]. We do not include the method of [9] in these comparisons due to its high computational cost. Note that our theoretical analysis assumes for simplicity that all expected node degrees are equal (Theorems 4.1, 4.2 and 4.3); however, we allow different expected node degrees in simulations. In this section, $d = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\, d_i$ denotes the average expected node degree.

5.1. Synthetic networks

To generate synthetic networks, we fix the labels $c \in \{1, ..., K\}^n$ so that $c_i = k$ if $n\pi_{k-1} + 1 \le i < n\pi_k$, where $\pi_0 = 0$. The label matrix $Z \in \mathbb{R}^{n\times K}$, given by $Z_{ik} = \mathbf{1}(c_i = k)$, encodes c by representing each node's label with a row of K elements, exactly one of which is equal to 1, and the rest are equal to 0. Let $\tilde{P}$ be a $K \times K$ matrix with the diagonal $w = (w_1, ..., w_K)$ and off-diagonal entries β, and $M = Z\tilde{P}Z^T$. Under the stochastic block model, we generate entries of A using the edge probability matrix $E(A) = \rho_n M$; the average degree d is controlled by ρ_n. The parameter w controls the relative edge densities within communities, and β controls the out-in probability ratio. Smaller values of β and larger values of d make the problem easier. For the DCSBM, we generate the degree parameters θ_i from a distribution that takes two values, $\mathbb{P}(\theta = 1) = 1-\gamma$ and $\mathbb{P}(\theta = 0.2) = \gamma$. Parameter γ controls the fraction of "hubs", the high-degree nodes allowed under the DCSBM, and setting $\gamma = 0$ gives back the regular SBM. Given $\theta = (\theta_i, ..., \theta_n)$, the edges are generated independently with probabilities $E(A) = \rho_n \mathrm{diag}(\theta) M \mathrm{diag}(\theta)$, where $\mathrm{diag}(\theta)$ is a diagonal matrix with θ_i's on the diagonal.

The number of nodes is set to $n = 1200$, the out-in probability ratio $\beta = 0.2$, and we vary the average degree d, weights w, and community sizes determined by the vector π. We consider three different values for the number of communities, $K = 2$, 4, and 6. For each setting, we generate 200 replications of the network and record the accuracy, defined as the fraction of times a method correctly estimates the true number of communities K. The methods NCV and VLH require a pre-specified set of K values to choose from; we use the set $\{1, 2, ..., 8\}$ for synthetic networks and $\{1, 2, ..., 15\}$ for real-world networks.

We start by varying the average degree d, which controls the overall difficulty of the problem, while keeping community sizes equal. Figure 1 shows the performance of all methods for the balanced community density case, $w_i = 1$ for all $1 \le i \le K$. Figure 2 shows the unbalanced case, with $w = (1, 2)$ for $K = 2$, $w = (1, 1, 2, 3)$ for $K = 4$, and $w = (1, 1, 1, 1, 2, 3)$ for $K = 6$. In every figure, the top row corresponds to the SBM ($\gamma = 0$) and the bottom row to the DCSBM ($\gamma = 0.9$, meaning 10% of nodes are hubs).

In general, we see that when everything is balanced (Figure 1), all spectral methods perform fairly similarly and outperform both cross-validation (NCV) and the BIC-type criterion (VLH). Also, for larger K and especially under

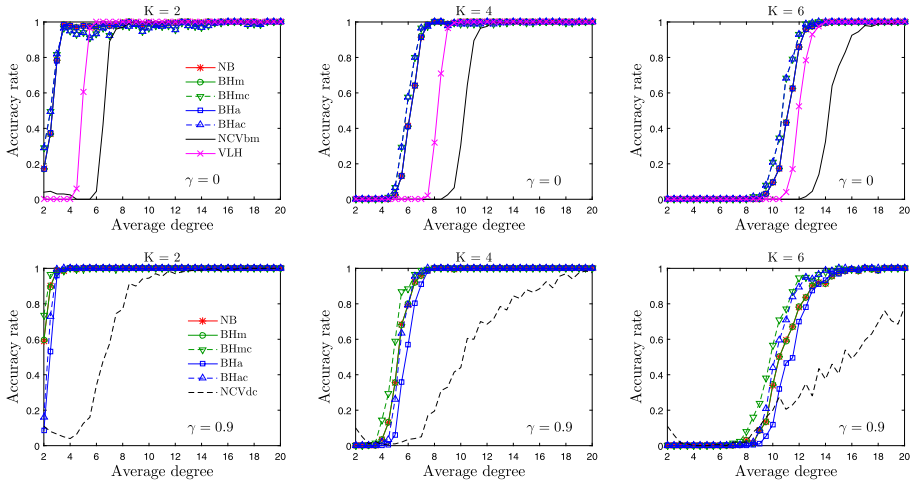

FIG 1. *The accuracy of estimating K as a function of the average degree. All communities have equal sizes, and $w_i = 1$ for all $1 \le i \le K$.*

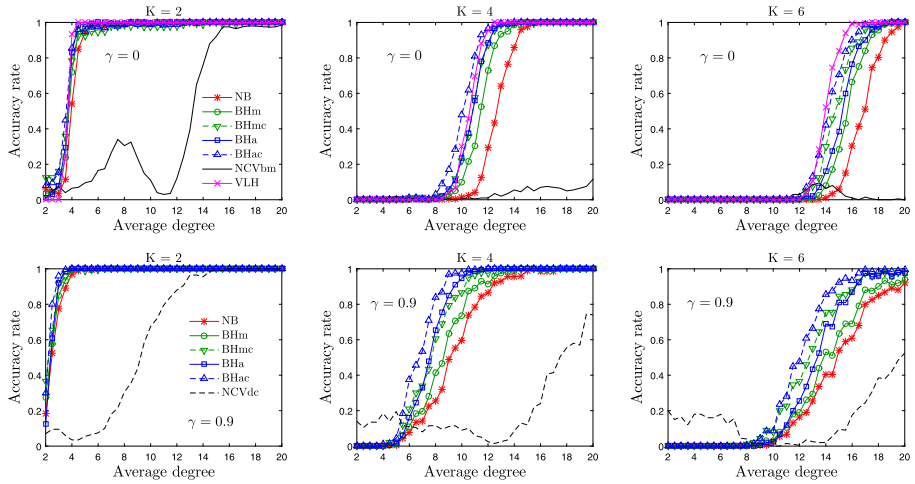

FIG 2. *The accuracy of estimating K as a function of the average degree. All communities have equal sizes; $w = (1, 2)$ for $K = 2$, $w = (1, 1, 2, 3)$ for $K = 4$, and $w = (1, 1, 1, 1, 2, 3)$ for $K = 6$.*

DCSBM, the corrected versions are somewhat better than the uncorrected ones, and the best Bethe Hessian methods are better than the non-backtracking estimator.

For networks with equal size communities but different edge densities within communities (Figure 2), cross-validation performs poorly, but VLH relatively improves. For larger K the spectral methods are also distinguishable, with all

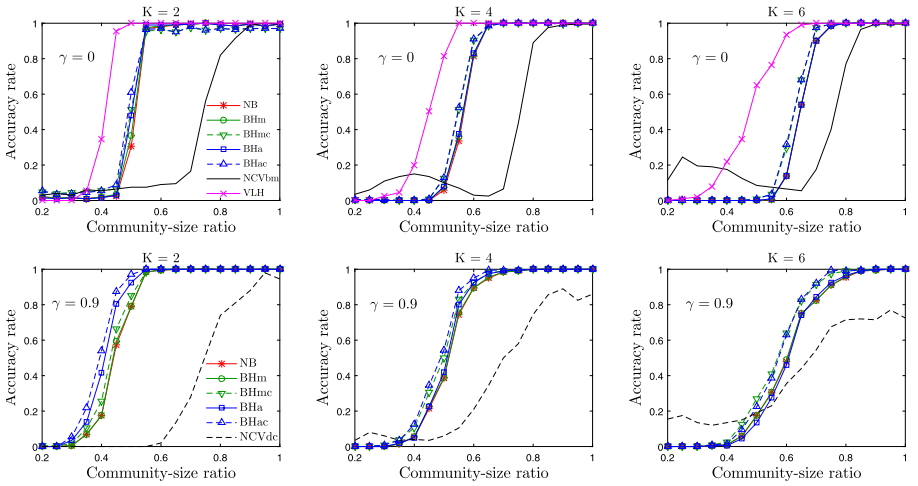

FIG 3. *The accuracy of estimating K as a function of the community-size ratio r: $\pi_1 = r/K$, $\pi_K = (2-r)/K$, and $\pi_i = 1/K$ for $2 \le i \le K-1$. In all plots, $w_i = 1$ for $1 \le i \le K$; the average degrees are $\lambda_n = 10$ (left), 15 (middle), and 20 (right).*

BH methods dominating NB, and corrected versions providing improvement. Overall, BHac is the best spectral method, with VLH comparable for the SBM in this case. The BHac method is the best overall for DCSBM where VLH is not applicable.

Communities of different sizes present a challenge for community detection methods in general, and the presence of relatively small communities makes the problem of estimating K difficult. To test the sensitivity of all the methods to this factor, we change the proportions of nodes falling into each community setting $\pi_1 = r/K$, $\pi_K = (2-r)/K$, and $\pi_i = 1/K$ for $2 \le i \le K-1$, and varying r in the range $[0.2, 1]$. As r increases, the community sizes become more similar, and are all equal when $r = 1$. Figure 3 shows the performance of all methods as a function of r. The top row corresponds to the SBM ($\gamma = 0$), the bottom row to the DCSBM ($\gamma = 0.9$), and the within-community edge density parameters $w_i = 1$ for all $1 \le i \le K$. Here we see that VLH is less sensitive to r than the spectral methods, but unfortunately it is not available under the DCSBM. Cross-validation is still dominated by spectral methods except for very small values of r, where all methods perform poorly. The corrections still provide a slight improvement for Bethe Hessian based methods, although all spectral methods perform fairly similarly in this case.

5.2. Real world networks

Finally, we apply the proposed methods on several popular network datasets which come with the "ground truth" node labels and the corresponding number of communities. We note that the network structure itself can indicate a differ-

TABLE 2
Estimates of the number of communities in real-world networks.

Dataset	NB	BHm	BHmc	BHa	BHac	NCVbm	NCVdc	VLH	Truth
College football	10	10	10	10	10	14	13	9	12
Political books	3	3	4	4	4	8	2	6	3
Dolphins	2	2	2	2	2	4	3	2	2
Karate club	2	2	2	2	2	3	3	4	2
Political blogs	8	7	8	7	8	10	2	1	2

ent number of communities than those given in the ground truth, since those are typically derived from one specific node attribute and there may be other communities or sub-communities corresponding to different attributes. However, these ground truth labels still provide a reasonable baseline against which to compare estimators.

The college football network [15] represents 115 US college football teams and the games they played in 2000. The "ground truth" communities are the 12 conferences that the teams belong to. The political books network [32], compiled around 2004, consists of 105 books about US politics; an edge is "frequently purchased together" on Amazon. The $K = 3$ communities are "conservative", "liberal", or "neutral", labelled manually based on contents. The dolphin network [25] is a social network of 62 dolphins, with edges representing social interactions, and $K = 2$ communities are based on a split which happened after one dolphin left the group. Similarly, the karate club network [46] is a social network of 34 members of a karate club, with edges representing friendships, and $K = 2$ communities based on a split following a dispute. Finally, the political blogs network [2], collected around 2004, consists of blogs about US politics, with edges representing web links, and $K = 2$ communities are "conservative" and "liberal", based on manual labelling. For this dataset, as is commonly done in the literature, we only consider its largest connected component of 1222 nodes.

Table 2 shows the estimated number of communities in these networks. All spectral methods estimate the correct number of communities for dolphins and the karate club, and do a reasonable job for the college football and political books data. For political blogs, all methods but NCV and VLH estimate a much larger number of communities, suggesting the estimates correspond to smaller sub-communities with more uniform degree distributions that have been previously detected by other authors. We also found that the VLH method was highly dependent on the tuning parameter, and the estimates by NCVbm and NCVdc varied noticeably from run to run due to their use of random partitions.

6. Discussion

The numerical experiments suggest that the spectral methods provide extremely fast and reliable estimates of the number of communities K for balanced networks, with the Bethe Hessian based method with the threshold choice r_a and the correction described in (3.1) the best choice in most scenarios. With communities of significantly different sizes, they tend to underestimate K by combining

small communities together, which seems to be an intrinsic limitation of spectral methods. This suggests that their estimates can be used as a lower bound on K and a starting point for a more elaborate and computationally demanding likelihood-based method like VLH, in the same way that spectral clustering can be used to initialize a more sophisticated community detection method. Having a small set of plausible values of K to focus on can significantly reduce the computational cost and improve the accuracy of estimating the number of communities.

For semi-dense networks, we show in Theorems 4.2 and 4.3 that estimating the number of communities is possible below the exact recovery threshold. For example, under $G(n, \frac{a}{n}, \frac{b}{n})$, our results require $(a-b)^2 \geq 32(1+o(1))(a+b)$ while exact community recovery is feasible if $(a-b)^2 > 2(a+b+2\sqrt{ab})\log n$. Determining the exact condition under which estimating the number of communities is possible is an interesting and challenging question and we leave it for future research.

Appendix A: Proof of Theorem 4.2

Following [43], we will work with the following rescaled conjugation of the non-backtracking matrix B defined in (2.1) (which has the same eigenvalues as $B/\sqrt{\alpha}$ where $\alpha = d-1$)

$$\begin{pmatrix} \frac{1}{\sqrt{\alpha}}A & \frac{1}{\alpha}(I-D) \\ I & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\alpha}}A & -I \\ I & 0 \end{pmatrix} + \begin{pmatrix} 0 & \frac{1}{\alpha}(\mathbb{E}\,D - D) \\ 0 & 0 \end{pmatrix} =: H + E. \quad \text{(A.1)}$$

The key result for proving Theorem 4.2 is Theorem A.1 below, which establishes a connection between spectra of $H+E$ and H. The spectrum of H is closely related to the spectrum of the adjacency matrix, and is discussed in Section A.1.

To prove Theorem A.1, we only need a crude bound on $\|A - \mathbb{E}\,A\|$ that is known to hold for very general graph models, including SBM, DCSBM and inhomogeneous Erdos-Renyi models [22]. For clarity, we put this bound in Assumption A.1 below. We will replace it with a sharper bound in Theorem A.2 to prove Theorem 4.2.

Assumption A.1. With probability at least $1 - 1/n$, the following inequality holds

$$\|A - \mathbb{E}\,A\| \leq C\sqrt{d}.$$

It is easy to see that Assumption A.1 implies $\|E\| = O(1/\sqrt{d})$ with high probability while [43] shows that H is diagonalizable as follows.

A.1. Spectrum of H

Denote by $v_1, ..., v_n$ and $\lambda_1, \lambda_2, ..., \lambda_n$ eigenvectors and corresponding eigenvalues of $A/\sqrt{\alpha}$ ordered so that $|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_n|$. For each i, H has two

eigenvalues μ_{2i-1} and μ_{2i} that are solutions of equation $\mu^2 - \lambda_i \mu + 1 = 0$, that is

$$\mu_{2i-1} = \frac{\lambda_i + \sqrt{\lambda_i^2 - 4}}{2}, \quad \mu_{2i} = \frac{\lambda_i - \sqrt{\lambda_i^2 - 4}}{2}. \tag{A.2}$$

The corresponding left (unit) eigenvectors of H are

$$y_{2i-1}^* = \frac{1}{\sqrt{1 + |\mu_{2i-1}|^2}}(-\mu_{2i-1} v_i^T, v_i^T), \quad y_{2i}^* = \frac{1}{\sqrt{1 + |\mu_{2i}|^2}}(-\mu_{2i} v_i^T, v_i^T)$$

and their inner product is

$$\langle y_{2i-1}, y_{2i} \rangle = \begin{cases} \frac{\lambda_i^2 + \lambda_i \sqrt{\lambda_i^2 - 4}}{4}, & \text{if } |\lambda_i| < 2 \\ \frac{2}{|\lambda_i|}, & \text{if } |\lambda_i| \geq 2 \end{cases} = \begin{cases} \frac{\lambda_i \mu_{2i-1}}{2}, & \text{if } |\lambda_i| < 2 \\ \frac{2}{|\lambda_i|}, & \text{if } |\lambda_i| \geq 2. \end{cases} \tag{A.3}$$

The corresponding right eigenvectors of H are proportional to

$$x_{2i-1} = \frac{\sqrt{1 + |\mu_{2i-1}|^2}}{\mu_{2i} - \mu_{2i-1}} \begin{pmatrix} v_i \\ \mu_{2i} v_i \end{pmatrix}, \quad x_{2i} = \frac{\sqrt{1 + |\mu_{2i}|^2}}{\mu_{2i-1} - \mu_{2i}} \begin{pmatrix} v_i \\ \mu_{2i-1} v_i \end{pmatrix}, \tag{A.4}$$

with inner product

$$\langle x_{2i-1}, x_{2i} \rangle = \begin{cases} \frac{\lambda_i^2 + \lambda_i \sqrt{\lambda_i^2 - 4}}{\lambda_i^2 - 4}, & \text{if } |\lambda_i| < 2 \\ \frac{2|\lambda_i|}{4 - \lambda_i^2}, & \text{if } |\lambda_i| \geq 2 \end{cases} = \begin{cases} \frac{2\lambda_i \mu_{2i-1}}{\lambda_i^2 - 4}, & \text{if } |\lambda_i| < 2 \\ \frac{2|\lambda_i|}{4 - \lambda_i^2}, & \text{if } |\lambda_i| \geq 2 \end{cases} \tag{A.5}$$

Note that x_{2i-1} and x_{2i} are not unit vectors. Their squared norms are

$$\|x_{2i-1}\|^2 = \|x_{2i}\|^2 = \begin{cases} \frac{4}{4 - \lambda_i^2}, & \text{if } |\lambda_i| < 2 \\ \frac{\lambda_i^2}{\lambda_i^2 - 4}, & \text{if } |\lambda_i| \geq 2. \end{cases} \tag{A.6}$$

It is convenient to not normalize x_{2i-1} and x_{2i} because H admits the decomposition

$$H = \sum_{i=1}^{n} \left(\mu_{2i-1} x_{2i-1} y_{2i-1}^* + \mu_{2i} x_{2i} y_{2i}^* \right).$$

Note that from the formulas above we have

$$x_{2i-1} \perp y_{2i}, \quad x_{2i} \perp y_{2i-1}, \quad \langle x_{2i-1}, y_{2i-1} \rangle = \langle x_{2i}, y_{2i} \rangle = 1.$$

The space $\mathbb{C}^{2n}$ can be decomposed as a direct sum of orthogonal two-dimensional subspaces $\text{span}\{x_{2i-1}, x_{2i}\} = \text{span}\{y_{2i-1}, y_{2i}\}$, which are invariant under the action of H. Moreover, the orthogonal projection onto $\text{span}\{x_{2i-1}, x_{2i}\}$ is given by $x_{2i-1} y_{2i-1}^* + x_{2i} y_{2i}^*$.

A.2. Spectrum of $H + E$

The main difficulty of analyzing the spectrum of $H + E$ is that H and E are not symmetric so standard Weyl's inequalities do not apply even though $\|E\|$ is small. Wang and Wood [43] use the Bauer-Fike theorem instead and show that for Erdős-Rényi random graphs, the perturbation of E is negligible if the average degree is at least of order $n^{5/6}$. This strong assumption is likely an artifact of their proof because the Bauer-Fike bound is often not tight. In fact, by a direct and more careful analysis we show in the following theorem that the spectrum of $H + E$ is close to the spectrum of H for much sparser graphs.

Theorem A.1 (Connection between spectra of non-backtracking and adjacency matrices)**.** *There exists a constant $C > 0$ such that the following holds. Consider random graphs satisfying Assumptions 4.1 and A.1. Then with probability at least $1 - 1/n$, for each eigenvalue β of $H + E$, there exists an eigenvalue μ of H such that*

$$|\beta - \mu| \le Cd^{-1/8}.$$

For proving Theorem 4.2, we replace Assumption A.1 with the following shaper bound on $\|A - \mathbb{E}\,A\|$, which holds under stronger assumptions. This bound follows directly from [7] and [41]; see also [43].

Theorem A.2 (Concentration of adjacency matrix)**.** *There exists a constant $C_1, C_2 > 0$ such that the following holds. Assume that*

$$d \ge C_1 \log n \quad \textit{and} \quad d^5 \max_{i,j} \mathbb{E}\,A_{ij} \le n^{-1/13}.$$

Then with probability at least $1 - 1/n$, we have

$$\|A - \mathbb{E}\,A\| \le 2\sqrt{d} + C_2\sqrt{\log n}.$$

We are now ready to prove Theorem 4.2.

Proof of Theorem 4.2. Let $\lambda_1(\mathbb{E}\,A), ..., \lambda_K(\mathbb{E}\,A)$ be the nonzero eigenvalues of $\mathbb{E}\,A$ and $\lambda_1(A), \cdots, \lambda_n(A)$ be eigenvalues of A, ordered so that $|\lambda_1(\mathbb{E}\,A)| \ge \cdots \ge |\lambda_K(\mathbb{E}\,A)| > 0$ and $|\lambda_1(A)| \ge \cdots \ge |\lambda_n(A)|$. Then by Weyl's inequality and Theorem A.2, with probability at least $1 - 1/n$ we have

$$\begin{aligned} |\lambda_i(A)| &\le 2\sqrt{d} + C\sqrt{\log n} \quad \text{for} \quad i \ge K + 1, \\ |\lambda_i(A) - \lambda_i(\mathbb{E}\,A)| &\le 2\sqrt{d} + C\sqrt{\log n} \quad \text{for} \quad 1 \le i \le K. \end{aligned}$$

Since $|\lambda_K(\mathbb{E}\,A)| \ge 4\sqrt{d} + 4C(\sqrt[4]{d} + \sqrt{\log n})$ by Assumption 4.2, it follows that $|\lambda_i(A)| \ge 2\sqrt{d} + 2C(\sqrt[4]{d} + \sqrt{\log n})$ for $1 \le i \le K$. Therefore for $1 \le i \le K$, from (A.2) we have

$$\begin{aligned} \max\{|\mu_{2i-1}(H)|, |\mu_{2i}(H)|\} &\ge 1 + \frac{2C(\sqrt[4]{d} + \sqrt{\log n})^{1/2}}{d^{1/4}} > 1, \\ \min\{|\mu_{2i-1}(H)|, |\mu_{2i}(H)|\} &= \frac{1}{\max\{|\mu_{2i-1}(H)|, |\mu_{2i}(H)|\}} < 1. \end{aligned}$$

Similarly, for $i \geq K+1$ we have

$$\max\{|\mu_{2i-1}(H)|, |\mu_{2i}(H)|\} < 1 + 2C\left(\frac{\log n}{d}\right)^{1/4}.$$

Theorem A.1 and the continuity of eigenvalues with respect to small perturbation then imply that for $1 \leq i \leq K$,

$$\begin{aligned}\max\{|\mu_{2i-1}(H+E)|, |\mu_{2i}(H+E)|\} &\geq 1 + \frac{2C(\sqrt[4]{d} + \sqrt{\log n})^{1/2}}{d^{1/4}} - Cd^{-1/8} \\ &\geq 1 + 2C\left(\frac{\log n}{d}\right)^{1/4} + Cd^{-1/8},\end{aligned}$$

while the remaining eigenvalues of $H+E$ have magnitude at most

$$1 + 2C\left(\frac{\log n}{d}\right)^{1/4} + Cd^{-1/8}.$$

Since $B = \sqrt{\alpha}(H+E)$ by (2.1) and (A.1), it follows that the nonbacktracking matrix has exactly K eigenvalues with magnitude at least $(1+\varepsilon)\sqrt{d}$ and the remaining eigenvalues are of magnitude smaller than $(1+\varepsilon)\sqrt{d}$.

To show that the K largest eigenvalues in magnitude of B are real, we use the following deterministic inclusion bound for the spectrum of B; see [5, Theorem 3.7]. Let $d_{\min} \geq 2$ and $d_{\max}$ be the minimal and maximal degrees of a graph. Then the spectrum of B satisfies

$$\sigma(B) \subseteq \left\{\lambda \in \mathbb{C} : \sqrt{d_{\min}-1} \leq |\lambda| \leq \sqrt{d_{\max}-1}\right\} \cap \{\lambda \in \mathbb{R} : 1 \leq |\lambda| \leq d_{\max}-1\}.$$

In our setting, we bound $d_{\max}$ using standard Bernstein's inequality: with probability at least $1 - 1/n$,

$$\sqrt{d_{\max}-1} \leq \sqrt{d + C\sqrt{d\log n}} \leq (1+\varepsilon)\sqrt{d}.$$

Since all complex eigenvalues of B are contained in a circle of radius at most $\sqrt{d_{\max}-1}$, the K largest eigenvalues of B in magnitude, which are outside the circle of radius $(1+\varepsilon)\sqrt{d}$, must be real. The proof is complete. □

The rest of this section is devoted to proving Theorem A.1. Besides the facts listed in Section A.1, we need the following elementary lemmas, the proofs of which are postponed until the end of this section.

Lemma A.3. *Let x, y, v be unit vectors with $|\langle x, y\rangle| \leq 1-\varepsilon$ for some $\varepsilon \in [0,1]$, $v \in \operatorname{span}\{x, y\}$ and $a, b \in \mathbb{C}$ be any complex numbers. Then*

$$\|ax + by\|^2 \geq \varepsilon(|a|^2 + |b|^2), \quad |\langle v, x\rangle|^2 + |\langle v, y\rangle|^2 \geq \varepsilon.$$

Lemma A.4. *Let x_{2i-1}, x_{2i} be right eigenvectors of H given by (A.4). Then for any $a, b \in \mathbb{C}$ and $1 \leq i \leq n$ we have*

$$\|ax_{2i-1} + bx_{2i}\| \geq \max\{|a|, |b|\}.$$

Lemma A.5. *Let x_{2i-1}, x_{2i} be right eigenvectors of H given by (A.4) and denote $W_i = span\{x_{2i-1}, x_{2i}\}$. Then for any $1 \le i \le n$ we have*

$$\sup_{w \in W_i} \|Hw\| \le 4 \max\{|\lambda_i|, 1\} \cdot \|w\|.$$

We are now ready to prove Theorem A.1.

Proof of Theorem A.1. Denote by P_i the orthogonal projection onto $\text{span}\{x_{2i-1}, x_{2i}\}$. Let u be a unit eigenvector of $H+E$ with corresponding eigenvalue β and $u_i = P_i u / \|P_i u\|$. Note first that

$$u = \sum_i P_i u = \sum_i (x_{2i-1} y_{2i-1}^* + x_{2i} y_{2i}^*) P_i u.$$

This allows us to write Eu as follows:

$$Eu = \beta u - Hu = \sum_i \left[(\beta - \mu_{2i-1}) x_{2i-1} y_{2i-1}^* + (\beta - \mu_{2i}) x_{2i} y_{2i}^* \right] P_i u.$$

Note that the terms in above sum belong to orthogonal subspaces of $\mathbb{C}^{2n}$. Therefore

$$\begin{aligned} \|E\|^2 &\ge \sum_i \left\| \left[(\beta - \mu_{2i-1}) x_{2i-1} y_{2i-1}^* + (\beta - \mu_{2i}) x_{2i} y_{2i}^* \right] u_i \right\|^2 \|P_i u\|^2 \\ &= \sum_i T_i \|P_i u\|^2 \end{aligned} \tag{A.7}$$

where T_i denotes the first factor of the corresponding term in the sum.

Let $\varepsilon \in (0, 1/4)$ be a small number to be chosen later. Consider first the eigenvalues λ_i with magnitude not close to 2, namely those satisfying $||\lambda_i| - 2| > \varepsilon$. From (A.5) and (A.6) we have

$$|\langle y_{2i-1}, y_{2i} \rangle| = \frac{|\langle x_{2i-1}, x_{2i} \rangle|}{\|x_{2i-1}\| \cdot \|x_{2i}\|} = \begin{cases} |\lambda_i|/2, & \text{if } |\lambda_i| < 2 - \varepsilon \\ 2/|\lambda_i|, & \text{if } |\lambda_i| > 2 + \varepsilon \end{cases} \le 1 - \varepsilon/3. \tag{A.8}$$

It also follows from (A.6) that $\|x_{2i-1}\| = \|x_{2i}\| > 1$. Since $u_i \in \text{span}\{x_{2i-1}, x_{2i}\} = \text{span}\{y_{2i-1}, y_{2i}\}$, if $||\lambda_i| - 2| > \varepsilon$ then by (A.8) and Lemma A.3 (applied to $\|x_{2i-1}\|^{-1} x_{2i-1}$, $\|x_{2i}\|^{-1} x_{2i}$ first and then to y_{2i-1}, y_{2i}) we have

$$\begin{aligned} T_i &\ge \varepsilon/3 \cdot \left(|\beta - \mu_{2i-1}|^2 |y_{2i-1}^* u_i|^2 + |\beta - \mu_{2i}|^2 |y_{2i}^* u_i|^2 \right) \cdot \|x_{2i}\|^2 \\ &\ge \varepsilon^2/9 \cdot \min\{|\beta - \mu_{2i-1}|^2, |\beta - \mu_{2i}|^2\}. \end{aligned} \tag{A.9}$$

We now consider two cases of u, namely whether the following inequality holds:

$$\sum_{||\lambda_i| - 2| > \varepsilon} \|P_i u\|^2 > \varepsilon. \tag{A.10}$$

We will show that in both cases there exists an eigenvalue of H that is close to β. Assume first that (A.10) holds. Then from (A.7), (A.9) and (A.10) we have

$$\begin{aligned}
\|E\|^2 &\geq \sum_{||\lambda_i|-2|>\varepsilon} T_i \cdot \|P_i u\|^2 \\
&\geq \sum_{||\lambda_i|-2|>\varepsilon} \varepsilon^2/9 \cdot \min\{|\beta-\mu_{2i-1}|^2, |\beta-\mu_{2i}|^2\} \cdot \|P_i u\|^2 \\
&\geq \varepsilon^2/9 \cdot \min_{||\lambda_i|-2|>\varepsilon} \{|\beta-\mu_{2i-1}|^2, |\beta-\mu_{2i}|^2\} \cdot \sum_{||\lambda_i|-2|>\varepsilon} \|P_i u\|^2 \\
&\geq \varepsilon^3/9 \cdot \min_{||\lambda_i|-2|>\varepsilon} \{|\beta-\mu_{2i-1}|^2, |\beta-\mu_{2i}|^2\}.
\end{aligned}$$

It follows that there exists i with $||\lambda_i|-2| > \varepsilon$ such that

$$\min\{|\mu_{2i-1}-\beta|^2, |\mu_{2i}-\beta|^2\} \leq \frac{9\|E\|^2}{\varepsilon^3}. \tag{A.11}$$

We now consider the second case of u when (A.10) does not hold, or equivalently

$$\sum_{||\lambda_i|-2|\leq\varepsilon} \|P_i u\|^2 > 1-\varepsilon. \tag{A.12}$$

We partition the set of indices i satisfying $||\lambda_i|-2| \leq \varepsilon$ as a union of J and I, where J is the set of indices i such that $||\lambda_i|-2| \leq \varepsilon$ and $\max\{|y_{2i-1}^* u_i|, |y_{2i}^* u_i|\} > \varepsilon$, and I is the set of indices i such that $||\lambda_i|-2| \leq \varepsilon$ and $\max\{|y_{2i-1}^* u_i|, |y_{2i}^* u_i|\} \leq \varepsilon$. It follows from (A.12) that at least one of the following inequalities hold:

$$\sum_{i\in J} \|P_i u\|^2 > \varepsilon, \quad \sum_{i\in I} \|P_i u\|^2 > 1-2\varepsilon.$$

If the first inequality holds then by (A.7) and Lemma A.4 we have

$$\begin{aligned}
\|E\|^2 &\geq \sum_{i\in J} T_i \cdot \|P_i u\|^2 \\
&\geq \sum_{i\in J} \max\left\{|(\beta-\mu_{2i-1})y_{2i-1}^* u_i|^2, |(\beta-\mu_{2i})y_{2i}^* u_i|^2\right\} \cdot \|P_i u\|^2 \\
&\geq \min_{i\in J} \max\left\{|(\beta-\mu_{2i-1})y_{2i-1}^* u_i|^2, |(\beta-\mu_{2i})y_{2i}^* u_i|^2\right\} \cdot \sum_{i\in J} \|P_i u\|^2 \\
&\geq \varepsilon \cdot \min_{i\in J} \max\left\{|(\beta-\mu_{2i-1})y_{2i-1}^* u_i|^2, |(\beta-\mu_{2i})y_{2i}^* u_i|^2\right\}.
\end{aligned}$$

Since $\max\{|y_{2i-1}^* u_i|, |y_{2i}^* u_i|\} > \varepsilon$ for $i \in J$, it follows that there exists $i \in J$ such that

$$\min\{|\beta-\mu_{2i-1}|^2, |\beta-\mu_{2i}|^2\} \leq \frac{\|E\|^2}{\varepsilon^3}. \tag{A.13}$$

We now assume that the following inequality holds:

$$\sum_{i\in I} \|P_i u\|^2 > 1-2\varepsilon. \tag{A.14}$$

This inequality implies that $|\beta|$ is bounded. Indeed, from identities $(H+E)u = \beta u$ and $u = \sum_i \|P_i u\| u_i$ we get

$$\sum_i \|P_i u\| H u_i + Eu = \beta \sum_i \|P_i u\| u_i. \tag{A.15}$$

Note that $Hu_i \in \operatorname{span}\{x_{2i-1}, x_{2i}\}$ because $u_i \in \operatorname{span}\{x_{2i-1}, x_{2i}\}$ and $\{x_{2i-1}, x_{2i}\}$ are eigenvectors of H. Denote $P_I = \sum_{i\in I} P_i$ and apply P_I to both sides of (A.15), we have

$$\sum_{i\in I} \|P_i u\| H u_i + P_I E u = \beta \sum_{i\in I} \|P_i u\| u_i.$$

If $i \in I$ then H is bounded on $\operatorname{span}\{x_{2i-1}, x_{2i}\}$ by Lemma A.5. Therefore from (A.14) we obtain

$$(1-2\varepsilon)^{1/2}|\beta| \;\le\; \Big\|\beta \sum_{i\in I} \|P_i u\| u_i\Big\| \;\le\; \Big\| \sum_{i\in I} \|P_i u\| H u_i\Big\| + \|P_I E u\| \;\le\; C + \|E\|.$$

Since $\varepsilon \le 1/4$ and $\|E\| \le 1$, this implies $|\beta| \le 2C$. Applying $P_{I^c} = \sum_{i\notin I} P_i$ to both sides of (A.15), using (A.14) and the boundedness of β, we have

$$\Big\| \sum_{i\in I^c} \|P_i u\| H u_i\Big\| \le \|P_{I^c} E u\| + |\beta| \cdot \Big\| P_{I^c} \sum_{i\in I^c} \|P_i u\| u_i\Big\| \le \|E\| + C\sqrt{2\varepsilon}. \tag{A.16}$$

Therefore using $(H+E)u = \beta u$ and inequalities (A.14), (A.16) we have

$$\begin{aligned} \Big\|\beta u - (H+E)\sum_{i\in I} \|P_i u\| u_i\Big\| &= \Big\| \sum_{i\in I^c} \|P_i u\| H u_i + E \sum_{i\in I^c} \|P_i u\| u_i\Big\| \\ &\le (\|E\| + C\sqrt{2\varepsilon}) + \|E\| \\ &\le 2C(\sqrt{\varepsilon} + \|E\|). \end{aligned} \tag{A.17}$$

Denote $\bar{x}_{2i-1} = \|x_{2i-1}\|^{-1} x_{2i-1}$ and $\bar{x}_{2i} = \|x_{2i}\|^{-1} x_{2i}$. Since $\bar{x}_{2i-1} \perp y_{2i}$, $\bar{x}_{2i} \perp y_{2i-1}$ and $\max\{|y_{2i-1}^* u_i|, |y_{2i}^* u_i|\} \le \varepsilon$ for $i \in I$, it follows that $|\langle u_i, \bar{x}_{2i-1}\rangle| \ge 1-2\varepsilon$ and $|\langle u_i, \bar{x}_{2i}\rangle| \ge 1-2\varepsilon$. By multiplying $\bar{x}_{2i}$ with a complex number of magnitude one if necessary, we may assume that $\langle u_i, \bar{x}_{2i}\rangle \ge 1-2\varepsilon$ for $i \in I$, and consequently

$$\|u_i - \bar{x}_{2i}\|^2 \le 4\varepsilon. \tag{A.18}$$

We are now ready to show that β is close to an eigenvalue of H. By (A.18), (A.14), (A.17), the fact that β and μ_{2i} are bounded for $i \in I$, and triangle inequality we have

$$\begin{aligned} \Big\| \sum_{i\in I} \|P_i u\| (\mu_{2i} - \beta) u_i\Big\| &= \Big\| \sum_{i\in I} \|P_i u\| \mu_{2i} u_i - \sum_{i\in I} \|P_i u\| \beta u_i\Big\| \\ &\le \Big\| \sum_{i\in I} \|P_i u\| \mu_{2i} \bar{x}_{2i} - \sum_{i\in I} \|P_i u\| \beta u_i\Big\| + C\sqrt{4\varepsilon} \\ &\le \Big\| \sum_{i\in I} \|P_i u\| \mu_{2i} \bar{x}_{2i} - \sum_{i=1}^{n} \|P_i u\| \beta u_i\Big\| + C(\sqrt{4\varepsilon} + \sqrt{2\varepsilon}) \end{aligned}$$

$$
\begin{aligned}
&= \left\| H \sum_{i \in I} \|P_i u\| \; \bar{x}_{2i} - \beta u \right\| + C(\sqrt{4\varepsilon} + \sqrt{2\varepsilon}) \\
&\leq \left\| H \sum_{i \in I} \|P_i u\| \; u_i - \beta u \right\| + C(2\sqrt{4\varepsilon} + \sqrt{2\varepsilon}) \\
&\leq \left\| (H + E) \sum_{i \in I} \|P_i u\| \; u_i - \beta u \right\| \\
&\quad + C(2\sqrt{4\varepsilon} + \sqrt{2\varepsilon}) + \|E\| \\
&\leq 2C(\sqrt{\varepsilon} + \|E\|) + C(2\sqrt{4\varepsilon} + \sqrt{2\varepsilon}) + \|E\| \\
&\leq 8C(\sqrt{\varepsilon} + \|E\|).
\end{aligned}
$$

Together with (A.14) this implies

$$
\min_{i \in I} |\beta - \mu_{2i}|^2 \leq \frac{1}{1 - 2\varepsilon} \cdot \sum_{i \in I} \|P_i u\|^2 |\beta - \mu_{2i}|^2 \leq C(\varepsilon + \|E\|^2). \tag{A.19}
$$

Finally, it follows from (A.11), (A.13) and (A.19) that if β is an eigenvalue of $H + E$ then there exists an eigenvalue μ of H such that

$$
|\beta - \mu| \; \leq \; \frac{C(\|E\| + \varepsilon^2)}{\varepsilon^{3/2}} \; = \; 2C\|E\|^{1/4}
$$

for $\varepsilon = \|E\|^{1/2}$. It follows from Assumption A.1 that $\|E\| = O(1/\sqrt{d})$ and therefore the proof is complete. □

Proof of Lemma A.3. We prove the first inequality:

$$
\begin{aligned}
\|ax + by\|^2 &= |a|^2 + |b|^2 + 2 \cdot \mathrm{Re}\{\bar{a} b \langle x, y \rangle\} \\
&\geq |a|^2 + |b|^2 - 2|ab|(1 - \varepsilon) \\
&= (1 - \varepsilon)(|a| - |b|)^2 + \varepsilon(|a|^2 + |b|^2) \\
&\geq \varepsilon(|a|^2 + |b|^2).
\end{aligned}
$$

To prove the second inequality, denote $z = x - y$ and $w = x + y$. Then z, w are perpendicular and $x = (z + w)/2$, $y = (w - z)/2$. Therefore

$$
|\langle v, x \rangle|^2 + |\langle v, y \rangle|^2 = v^*(xx^* + yy^*)v = v^*(zz^* + ww^*)v/2.
$$

Note that the restriction of $zz^* + ww^*$ on $\mathrm{span}\{x, y\}$ is a positive definite matrix with eigenvalues $\|z\|^2$ and $\|w\|^2$ because z and w are perpendicular. By the first inequality

$$
\min\{\|z\|^2, \|w\|^2\} = \min\{\|x - y\|^2, \|x + y\|^2\} \geq 2\varepsilon.
$$

Since $v \in \mathrm{span}\{x, y\}$, it follows that

$$
v^*(zz^* + ww^*)v/2 \geq 2\varepsilon v^* v/2 = \varepsilon.
$$

The proof is complete. □

Proof of Lemma A.4. We decompose x_{2i-1} as $x_{2i-1} = z + w$ where $z \perp x_{2i}$ and $w \in \text{span}\{x_{2i}\}$. Then

$$\|ax_{2i-1} + bx_{2i}\|^2 = |a|^2\|z\|^2 + \|aw + bx_{2i}\|^2 \geq |a|^2\|z\|^2.$$

To calculate z, denote $\bar{x}_{2i-1} = \|x_{2i-1}\|^{-1}x_{2i-1}$, $\bar{x}_{2i} = \|x_{2i}\|^{-1}x_{2i}$ and $\tau = \langle \bar{x}_{2i-1}, \bar{x}_{2i}\rangle$. From (A.5) and (A.6) we get

$$\tau = \begin{cases} -\frac{\lambda_i^2 + \lambda_i\sqrt{\lambda_i^2-4}}{4}, & \text{if } |\lambda_i| < 2 \\ -\frac{2}{|\lambda_i|}, & \text{if } |\lambda_i| \geq 2. \end{cases}$$

Since $\|x_{2i-1}\| = \|x_{2i}\|$, it follows that

$$z = x_{2i-1} - \langle x_{2i-1}, \bar{x}_{2i}\rangle \bar{x}_{2i} = x_{2i-1} - \tau x_{2i}.$$

Therefore by (A.6), we obtain

$$\begin{aligned} \|z\|^2 &= \|x_{2i-1}\|^2 + |\tau|^2\|x_{2i}\|^2 - 2\text{Re}(\tau\langle x_{2i-1}, x_{2i}\rangle) \\ &= \|x_{2i}\|^2\left(|\tau|^2 + 1 - 2\text{Re}(\tau^2)\right) \\ &= \begin{cases} \frac{4}{4-\lambda_i^2}\left(\frac{\lambda_i^2}{4} + 1 - \frac{\lambda_i^4 - 2\lambda_i^2}{4}\right), & \text{if } |\lambda_i| < 2 \\ \frac{\lambda_i^2}{\lambda_i^2-4}\left(1 - \frac{4}{\lambda_i^2}\right), & \text{if } |\lambda_i| \geq 2 \end{cases} \\ &= \begin{cases} \lambda_i^2 + 1, & \text{if } |\lambda_i| < 2 \\ 1, & \text{if } |\lambda_i| \geq 2 \end{cases} \\ &\geq 1. \end{aligned}$$

This implies $\|ax_{2i-1} + bx_{2i}\| \geq |a| \cdot \|z\| \geq |a|$. By decomposing x_{2i} instead of x_{2i-1} and repeating the same argument, we obtain $\|ax_{2i-1} + bx_{2i}\| \geq |b|$. The proof is complete. □

Proof of Lemma A.5. Since $\bar{x}_{2i-1} = \|x_{2i-1}\|^{-1}x_{2i-1}$ and y_{2i} form an orthonormal basis of W_i, it is enough to bound $\|H\bar{x}_{2i-1}\|$ and $\|Hy_{2i}\|$. Note that the restriction H_i of H on W_i has the formula

$$H_i = \mu_{2i-1}x_{2i-1}y_{2i-1}^* + \mu_{2i}x_{2i}y_{2i}^*.$$

Therefore $\|H_i\bar{x}_{2i-1}\| = \|\mu_{2i-1}\bar{x}_{2i-1}\| \leq |\lambda_i|$. For the more involved calculation of Hy_{2i} we will repeatedly use identities

$$\mu_{2i-1}\mu_{2i} = 1, \quad \mu_{2i-1} + \mu_{2i} = \lambda_i \tag{A.20}$$

which follow directly from the formulas of μ_{2i-1} and μ_{2i} in (A.2).

The case $|\lambda_i| < 2$. From (A.3), (A.4) and identities $|\mu_{2i-1}| = |\mu_{2i}| = 1$, $\mu_{2i-1}\mu_{2i} = 1$ we have

$$H_iy_{2i} = \frac{\lambda_i\mu_{2i-1}^2}{\sqrt{2}(\mu_{2i} - \mu_{2i-1})}\begin{pmatrix} v_i \\ \mu_{2i}v_i \end{pmatrix} + \frac{\sqrt{2}\mu_{2i}}{\mu_{2i-1} - \mu_{2i}}\begin{pmatrix} v_i \\ \mu_{2i-1}v_i \end{pmatrix}$$

$$= \frac{1}{\sqrt{2}(\mu_{2i}-\mu_{2i-1})}\begin{pmatrix}(\lambda_i\mu_{2i-1}^2-2\mu_{2i})v_i\\(\lambda_i\mu_{2i-1}-2)v_i\end{pmatrix}.$$

Using (A.20) we get

$$\begin{aligned}\lambda_i\mu_{2i-1}^2-2\mu_{2i} &= (\mu_{2i-1}+\mu_{2i})\mu_{2i-1}^2-2\mu_{2i}\\ &= \mu_{2i-1}^3+\mu_{2i-1}-2\mu_{2i}\\ &= (\mu_{2i-1}-\mu_{2i})(\mu_{2i-1}^2+1).\end{aligned}$$

Similarly,

$$\lambda_i\mu_{2i-1}-2=(\mu_{2i-1}+\mu_{2i})\mu_{2i-1}-2=\mu_{2i-1}^2-1=\mu_{2i-1}(\mu_{2i-1}-\mu_{2i}).$$

Therefore

$$\|Hy_{2i}\|^2=(|\mu_{2i-1}^2+1|^2+|\mu_{2i-1}|^2)/2\le 5/2.$$

The case $\lambda_i \ge 2$. In this case μ_{2i-1} and μ_{2i} are real positive numbers. Then from (A.3), (A.4) and (A.20) we have

$$H_iy_{2i}=\frac{2\mu_{2i-1}\sqrt{1+\mu_{2i-1}^2}}{\lambda_i(\mu_{2i}-\mu_{2i-1})}\begin{pmatrix}v_i\\\mu_{2i}v_i\end{pmatrix}+\frac{\mu_{2i}\sqrt{1+\mu_{2i}^2}}{\mu_{2i-1}-\mu_{2i}}\begin{pmatrix}v_i\\\mu_{2i-1}v_i\end{pmatrix}.$$

It follows from (A.20) that

$$\sqrt{1+\mu_{2i-1}^2}=\mu_{2i-1}\sqrt{1+\mu_{2i}^2}.$$

Therefore

$$H_iy_{2i}=\frac{\sqrt{1+\mu_{2i}^2}}{\lambda_i(\mu_{2i}-\mu_{2i-1})}\begin{pmatrix}(2\mu_{2i-1}^2-\lambda_i\mu_{2i})v_i\\(2\mu_{2i-1}-\lambda_i)v_i\end{pmatrix}=-\frac{\sqrt{1+\mu_{2i}^2}}{\lambda_i}\begin{pmatrix}(\mu_{2i-1}+\lambda_i)v_i\\v_i\end{pmatrix}.$$

Note that $\mu_{2i}\le 1$ and $\mu_{2i-1}\le\lambda_i$ by (A.2). Hence

$$\|H_iy_{2i}\|^2=\frac{(1+\mu_{2i}^2)(1+(\mu_{2i-1}+\lambda_i)^2)}{\lambda_i^2}\le 10.$$

The case $\lambda_i \le -2$. In this case μ_{2i-1} and μ_{2i} are real negative numbers. Then from (A.3), (A.4) and (A.20) we have

$$H_iy_{2i}=\frac{2\mu_{2i-1}\sqrt{1+\mu_{2i-1}^2}}{\lambda_i(\mu_{2i-1}-\mu_{2i})}\begin{pmatrix}v_i\\\mu_{2i}v_i\end{pmatrix}+\frac{\mu_{2i}\sqrt{1+\mu_{2i}^2}}{\mu_{2i-1}-\mu_{2i}}\begin{pmatrix}v_i\\\mu_{2i-1}v_i\end{pmatrix}.$$

It follows from (A.20) that

$$\sqrt{1+\mu_{2i-1}^2}=-\mu_{2i-1}\sqrt{1+\mu_{2i}^2}.$$

Therefore

$$H_i y_{2i} = \frac{\sqrt{1+\mu_{2i}^2}}{\lambda_i(\mu_{2i}-\mu_{2i-1})}\begin{pmatrix}(2\mu_{2i-1}^2-\lambda_i\mu_{2i})v_i\\(2\mu_{2i-1}-\lambda_i)v_i\end{pmatrix} = -\frac{\sqrt{1+\mu_{2i}^2}}{\lambda_i}\begin{pmatrix}(\mu_{2i-1}+\lambda_i)v_i\\v_i\end{pmatrix}.$$

Note that $\mu_{2i}^2 \le \lambda_i^2$ and $|\mu_{2i-1}| \le 1$ by (A.2). Hence

$$\|H_i y_{2i}\|^2 = \frac{(1+\mu_{2i}^2)(1+(\mu_{2i-1}+\lambda_i)^2)}{\lambda_i^2} \le 10\lambda_i^2.$$

The proof is complete. □

Appendix B: Proof of Theorem 4.3

Proof of Theorem 4.3. We first rewrite the Bethe Hessian as follows:

$$H(r) = (r^2-1)I - r(A-\mathbb{E}\,A) + D - r\,\mathbb{E}\,A =: \hat{H}(r) - r\,\mathbb{E}\,A.$$

We show that eigenvalues of $\hat{H}(r)$ are non-negative and are of smaller order than non-zero eigenvalues of $r\,\mathbb{E}\,A$. This in turn implies that K eigenvalues of $H(r)$ are negative while the rest are positive.

By Theorem A.2, with probability at least $1-1/n$ we have

$$\|A-\mathbb{E}\,A\| \le 2\sqrt{d} + C\sqrt{\log n}. \tag{B.1}$$

To bound the node degrees, we use the standard Bernstein's inequality: with probability at least $1-1/n$,

$$\|D-\mathbb{E}\,D\| \le C\sqrt{d\log n}, \quad |r^2-(1+\varepsilon)^2 d| \le C\sqrt{d\log n}. \tag{B.2}$$

For square matrices X, Y we use $X \succeq Y$ to signify that $X-Y$ is positive semidefinite. Then by (B.1), (B.2) and Assumption 4.2, we have

$$\begin{aligned}\hat{H}(r) &\succeq \left[(r^2-1) - r\left(2\sqrt{d}+C\sqrt{\log n}\right) + (1+\varepsilon)^2 d - C\sqrt{d\log n}\right] I\\ &\succeq \left[\left(r-\sqrt{d}\right)^2 + (2\varepsilon+\varepsilon^2)d - C\sqrt{d\log n}\right] I\\ &\succeq 0\end{aligned} \tag{B.3}$$

because $\varepsilon = C\sqrt{\log n/d}$.

For a subspace $U \subseteq \mathbb{R}^n$, we denote by $\dim(U)$ the dimension of U, and by $U^\perp$ the orthogonal complement of U. Also, let $\mathrm{col}(\mathbb{E}\,A)$ be the column space of $\mathbb{E}\,A$. Using the Courant min-max principle (see e.g. [8, Corollary III.1.2]) and (B.3), we have

$$\begin{aligned}\rho_{n-K}(H(r)) &= \max_{\dim(U)=n-K}\ \min_{x\in U,\|x\|=1}\langle H(r)x,x\rangle\\ &\ge \min_{x\in \mathrm{col}(\mathbb{E}\,A)^\perp,\|x\|=1}\langle H(r)x,x\rangle \ge 0.\end{aligned}$$

Therefore the $n-K$ largest eigenvalues of $H(r)$ are non-negative.

It remains to show that the K smallest eigenvalues of $H(r)$ are negative. From (B.1), (B.2), and a triangle inequality, we have

$$\|\hat{H}(r)\| \leq 4d + C\sqrt{d\log n}. \tag{B.4}$$

On the other hand, from (B.2) and Assumption 4.2 we get

$$\lambda_K(r\,\mathbb{E}\,A) \geq (1+\varepsilon)\sqrt{d}\left(4\sqrt{d} + C\sqrt{\log n}\right) \geq 4d + C\sqrt{d\log n}. \tag{B.5}$$

Combining (B.4), (B.5), and using the Courant min-max principle again, we conclude that the K smallest eigenvalues of $H(r)$ are negative, which completes the proof. □

References

[1] E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18:1–86, 2018. MR3827065

[2] L. A. Adamic and N. Glance. The political blogosphere and the 2004 US election. In *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.

[3] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Machine Learning Research*, 9:1981–2014, 2008.

[4] A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013. MR3127859

[5] O. Angel, J. Friedman, and S. Hoory. The non-backtracking spectrum of the universal cover of a graph. *Transactions of the American Mathematical Society*, 367(6):4287–4318, 2015. MR3324928

[6] H. Bass. The Ihara-Selberg zeta function of a tree lattice. *Int J Math*, 3(06):717–797, 1992. MR1194071

[7] F. Benaych-Georges, C. Bordenave, and A. Knowles. Spectral radii of sparse random matrices. *Ann. Inst. H. Poincaré Probab. Statist.*, 56(3):2141–2161, 2020. MR4116720

[8] R. Bhatia. *Matrix Analysis*. Springer-Verlag New York, 1996. MR1477662

[9] P. Bickel and P. Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B, to appear*, 2013. MR3453655

[10] P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA*, 106:21068–21073, 2009.

[11] C. Bordenave, M. Lelarge, and L. Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs. *The Annals of Probability*, 46(1):1–71, 2018. MR3758726

[12] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 23:35.1–35.23, 2012.

[13] K. Chen and J. Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018. MR3803461

[14] J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statist. Comput.*, 18:173–183, 2008. MR2390817

[15] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, 99(12):7821–7826, 2002. MR1908073

[16] K. Hashimoto. Zeta functions of finite graphs and representations of p-adic groups. *Advanced Studies in Pure Mathematics*, 15:211–280, 1989. MR1040609

[17] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983. MR0718088

[18] J. Hu, H. Qin, T. Yan, and Y. Zhao. Corrected bayesian information criterion for stochastic block models. *To be published in Journal of the American Statistical Association*, 2019. MR4189756

[19] B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107, 2011. MR2788206

[20] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci.*, 110(52):20935–20940, 2013. MR3174850

[21] P. Latouche, E. Birmelé, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Stat. Modelling*, 12:93–115, 2012. MR2953099

[22] C. M. Le, E. Levina, and R. Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 2017. MR3689343

[23] J. Lei. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1):401–424, 2016. MR3449773

[24] T. Li, E. Levina, and J. Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020. MR4108931

[25] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of doubtful sound features a large propor- tion of long-lasting associations. can geographic isola- tion explain this unique trait? *Behavioral Ecology and Sociobiology*, 54:396–405, 2003.

[26] S. Ma, L. Su, and Y. Zhang. Determining the number of communities in degree-corrected stochastic block models. *arXiv:1809.01028*, 2018. MR4253762

[27] L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the Forty-sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703. ACM, 2014. MR3238997

[28] McSherry. Spectral partitioning of random graphs. *Proc. 42nd FOCS*, pages 529–537, 2001. MR1948742

[29] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. arXiv:1202.1499, 2012.
[30] E. Mossel, J. Neeman, and A. Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, DOI:10.1007/s00440-014-0576-6, 2014. MR3383334
[31] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018. MR3876880
[32] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006. MR2282139
[33] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582, 2006.
[34] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004. MR2282139
[35] T. P. Peixoto. Parsimonious module inference in large networks. *Phys. Rev. Lett.*, 110:148701, 2013.
[36] M. A. Riolo, G. T. Cantwell, G. Reinert, and M. E. J. Newman. Efficient method for estimating the number of communities in a network. *PHYSICAL REVIEW E*, 96:032310, 2017.
[37] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block model. *Annals of Statistics*, 39(4):1878–1915, 2011. MR2893856
[38] A. Saade, F. Krzakala, and L. Zdeborová. Spectral clustering of graphs with the Bethe Hessian. *Advances in Neural Information Processing Systems 27*, pages 406–414, 2014.
[39] D. F. Saldana, Y. Yu, and Y. Feng. How many communities are there? *Journal of Computational and Graphical Statistics*, 26(1):171–181, 2017. MR3610418
[40] T. Tao and V. Vu. Random matrices: universality of esds and the circular law. *Ann. Probab.*, 38(5):2023–2065, 2010. MR2722794
[41] V. Vu. Random discrete matrices. *Horizons of Combinatorics*, pages 257–280, 2008. MR2432537
[42] V. Vu. A simple SVD algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018. MR3734334
[43] K. Wang and P. M. Wood. Limiting empirical spectral distribution for the non-backtracking matrix of an Erdos-Renyi random graph. *arXiv:1710.11015*, 2017.
[44] R. Wang and P. Bickel. Likelihood-based model selection for stochastic block models. *Ann. Statist.*, 45(2):500–528, 2017. MR3650391
[45] B. Yan, P. Sarkar, and X. Cheng. Provable estimation of the number of blocks in block models. *Proceedings of Machine Learning Research*, 84:1185–1194, 2018.
[46] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.