J. Appl. Probab. **61,** 385–400 (2024)
*doi:*10.1017/jpr.2023.44

PHASE TRANSITION FOR THE GENERALIZED TWO-COMMUNITY STOCHASTIC BLOCK MODEL

SUNMIN LEE ,* ** AND
JI OON LEE,* *** *Korea Advanced Institute of Science and Technology (KAIST)*

Abstract

We study the problem of detecting the community structure from the generalized stochastic block model with two communities (G2-SBM). Based on analysis of the Stieltjes transform of the empirical spectral distribution, we prove a Baik–Ben Arous–Péché (BBP)-type transition for the largest eigenvalue of the G2-SBM. For specific models, such as a hidden community model and an unbalanced stochastic block model, we provide precise formulas for the two largest eigenvalues, establishing the gap in the BBP-type transition.

Keywords: BBP-type transition; empirical spectral distribution; Wigner-type matrix; hidden community model; unbalanced stochastic block model

2020 Mathematics Subject Classification: Primary 68Q87; 15B51
Secondary 60B20

1. Introduction

One of the most fundamental and natural problems in data science is understanding an underlying structure from data sets that can be viewed as networks. The problem is known as clustering or community detection, and it appears in diverse study fields involving real-world networks.

The stochastic block model (SBM) is one of the most fundamental mathematical models for understanding the community structure in networks. An SBM is a random graph with N nodes, partitioned into K disjoint subsets, called the communities, $C_1, C_2, \ldots, C_K$. An SBM can be characterized via its adjacency matrix consisting of 0 and 1, which is a symmetric (random) matrix $\widetilde{M}$, whose (i, j)-entry $\widetilde{M}_{ij}$ is an independent Bernoulli random variable with parameters depending only on the communities to which the nodes i and j belong. For the clustering of an SBM, it is often useful to analyze the eigenvalues of the adjacency matrix and their associated eigenvectors, known as a spectral method.

One of the most prominent examples of spectral methods is principal component analysis (PCA), in which the behavior of the eigenvectors associated with the extremal eigenvalues is considered to obtain the community structure of the SBM. For an SBM with two communities,

Received 4 July 2022; revision received 21 May 2023.
* Postal address: Department of Mathematical Sciences, KAIST, Daejeon 34141, South Korea.
** Email address: adelialee@kaist.ac.kr
*** Email address: jioon.lee@kaist.edu

the expectation of its adjacency matrix $\widetilde{M}$ has a block structure:

$$\mathbb{E}[\widetilde{M}] = \left(\begin{array}{c|c} P_{11} & P_{12} \\ \hline P_{21} & P_{22} \end{array}\right). \tag{1.1}$$

In the simplest case of a balanced SBM with $P_{11} = P_{22} = p$, $P_{12} = P_{21} = q$ ($p \neq q$), with the two communities of equal size, it can be easily checked that $\mathbb{E}[\widetilde{M}]$ has at most two non-zero eigenvalues, $N(p+q)/2$ and $N(p-q)/2$. Thus, if $N(p-q)/2$ is sufficiently large, the perturbation $\widetilde{M} - \mathbb{E}[\widetilde{M}]$ is negligible for the two largest eigenvalues $\widetilde{M}$, and it is possible to determine the community structure from the eigenvector associated with the second-largest eigenvalue of $\widetilde{M}$. After subtracting $(p+q)/2$ from each entry, the shifted adjacency matrix becomes the sum of a rank-1 deterministic matrix and a random matrix with centered entries, and we can use the eigenvector associated with the largest eigenvalue of the shifted adjacency matrix for clustering. The subtraction can work effectively as $(p+q)/2$ can be estimated by the overall density of the matrix for sufficiently large N.

The sum of a deterministic matrix and a random matrix has been extensively studied in random matrix theory. When the deterministic matrix is rank-1, and the random matrix is a Wigner matrix, it is called a (rank-1) spiked Wigner matrix. The behavior of the largest eigenvalue of a spiked Wigner matrix is known to exhibit a sharp phase transition depending on the ratio between the spectral norms of the deterministic part and the random part. This type of phase transition is called the BBP transition after the seminal work of Baik, Ben Arous, and Péché [6] for spiked (complex) Wishart matrices. From the BBP transition, we can immediately see that detection of the signal is possible via PCA when the signal-to-noise ratio (SNR) is above a certain threshold.

While the BBP transition has been proved for spiked Wigner matrices under various assumptions [9, 10, 17, 24], it is not directly applicable to the SBM, since the entries in a Wigner matrix are independent and identically distributed (up to a symmetry constraint) whereas those in the adjacency matrix of an SBM are not. The proof of the BBP transition with an SBM is substantially harder. For example, unless the SBM is balanced, the empirical spectral distribution (ESD) of $\widetilde{M}$ does not even converge to the semi-circle distribution, which is the limiting ESD of a Wigner matrix; the limiting ESD, in this case, is not given by a simple formula as the semi-circle distribution but by an implicit formula via its Stieltjes transform.

1.1. Main contribution

In this paper we consider a model that generalizes the SBM, called the generalized two-community stochastic block model (G2-SBM), with two communities. In this model, the mean of the matrix has the same block structure as that of the SBM in (1.1), but the entries are not necessarily Bernoulli random variables. With the structure in (1.1), we consider a model with three parameters p_1, p_2, and q. See Definition 2.1 for the precise definition of the G2-SBM.

For the G2-SBM, We prove the BBP-type transition for its largest eigenvalue (Theorem 2.1). The proof is based on analysis of the Stiejtes transform of the ESD, which involves the resolvent of the random part of the G2-SBM. Due to the community structure, the random part is not a Wigner matrix but a generalization of a Wigner matrix, known as a Wigner-type matrix. The local properties of eigenvalues of Wigner-type matrices are now well established by recent developments in random matrix theory; see, e.g., [4, 5, 13].

In our main result, Theorem 2.1, we only state the existence of the critical values and the limiting gap between the two largest eigenvalues but refrain from writing precise formulas for them. We instead apply our results to specific examples naturally arising in applications, the

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

hidden community model and unbalanced stochastic block model, and present the results from numerical experiments. (In terms of the edge probability, the former corresponds to the case $P_{11} = p$ and $P_{12} = P_{21} = P_{22} = q$, while the latter $P_{11} = P_{22} = p$ and $P_{12} = P_{21} = q$ $(p \neq q)$, but the sizes of the submatrices are different.)

1.2. Related works

The local law for Wigner-type matrices and the behavior of quadratic vector equations, which are crucial in the analysis for Wigner-type matrices, were thoroughly investigated in [4, 5]. A related result on the local law at the cusp for a Wigner-type matrix has also been proved [15]. For more results on general Wigner-type matrices, we refer to [13, 16, 27] and references therein.

The phase transition of the largest eigenvalue was first proved by Baik, Ben Arous, and Péché [6] for spiked Wishart matrices and later extended to other models, including the spiked Wigner matrix under various assumptions [9, 10, 17, 24]. If the SNR is below the threshold given by the BBP transition, the largest eigenvalue has no information on the signal and we cannot use the PCA to detect the signal. For this case, the PCA can be improved by an entry-wise transformation that effectively increases the SNR [8, 25]. Reliable detection is impossible below a certain threshold [25], and it is only possible to consider a weak detection, which is a hypothesis testing between the null model (without spike) and the alternative (with spike). We refer to [12, 14, 21] for more detail about weak detection.

The problem of recovering a hidden community from a symmetric matrix for two important cases, Bernoulli and Gaussian entries, was discussed in [19]. A threshold for exact recovery in the SBM was discussed in [1, 2, 11, 18]. The Kesten–Stigum threshold for the SBM was considered in [1, 20, 22, 23]. Similarly, the information-theoretic threshold for community detection in the SBM was considered in [1, 3, 7]; it can be achieved for a large number of communities. For more results and applications relating to the SBM, we refer to [26] and references therein.

1.3. Organization of the paper

The rest of the paper is organized as follows: In Section 2 we define the model and state the main result. In Sections 2.1 and 2.2 we introduce the hidden community model and unbalanced stochastic block model to provide results from numerical experiments around the transition threshold. In Section 3 we prove the main theorem. A summary of our results and future research directions is discussed in Section 5. Appendix A contains the definition of the Wigner-type matrices and preliminary results on this model. The detailed analysis for the specific models can be found in Section 4.

2. Main results

In this section, we precisely define the matrix model we consider in this paper and state our main theorem. We introduce a shifted, rescaled matrix for a generalized stochastic block model with two communities.

Definition 2.1. *(Generalized two-community stochastic block model.)* An $N \times N$ matrix M is a generalized two-community stochastic block model if $M = H + \lambda uu^\top$, where $\lambda \geq 0$ is a constant, $u = (u_1, u_2, \ldots, u_N) \in \mathbb{R}^N$ with $\|u\| = 1$, and $H = [H_{ij}]$ is an $N \times N$ real symmetric random matrix satisfying the following:

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

- There exist $S \subset [N] := \{1, 2, \dots, N\}$ and constants θ_1, θ_2 such that

$$u_i = \begin{cases} \theta_1 & \text{if } i \in S, \\ \theta_2 & \text{if } i \notin S. \end{cases}$$

  We further assume that $|S|/N$, $(1 - (|S|/N)) > c > 0$ for some (N-independent) constant c.

- The upper diagonal entries $H_{ij} (i \le j)$ are centered independent random variables such that:

  - there exist (N-independent) constants α_1 and α_2 such that

$$\mathbb{E}\left[H_{ij}^2\right] = \begin{cases} \alpha_1 N^{-1} & \text{if } i, j \in S, \\ \alpha_2 N^{-1} & \text{if } i, j \notin S, \\ N^{-1} & \text{otherwise;} \end{cases}$$

  - for any (N-independent) $D > 0$, there exists a constant C_D such that, for all $i \le j$, $\mathbb{E}\left[|H_{ij}|^D\right] \le C_D N^{-D/2}$.

For an adjacency matrix $\widetilde{M}$ as in (1.1), if $P_{11} = p_1$, $P_{22} = p_2$, and $P_{12} = P_{21} = q$, then after shifting and rescaling, we find that

$$\alpha_1 = \frac{p_1(1-p_1)}{q(1-q)}, \qquad \alpha_2 = \frac{p_2(1-p_2)}{q(1-q)}. \tag{2.1}$$

(See Section 4 for more detail.)

We remark that the H_{ij} are not necessarily Bernoulli random variables. The assumption of a finite moment means that the model is in the dense regime. The most typical balanced stochastic block model with two communities corresponds to the choice of parameters $|S| = N/2$ and $\alpha_1 = \alpha_2 > 1$.

Our main theorem is the following result on the phase transition for the spectral gap of the G2-SBM.

Theorem 2.1. *Let M be a generalized two-community stochastic block model as defined in Definition 2.1. Denote by λ_1 and λ_2 the largest and the second-largest eigenvalues of M. Assume that $\gamma := N_1/N$ is fixed. Then, there exists a constant λ_c, depending only on θ_1, θ_2, α_1, α_2, and γ, such that:*

- *(subcritical case) if $\lambda < \lambda_c$, then $\lambda_1 - \lambda_2 \to 0$ as $N \to \infty$, almost surely;*
- *(supercritical case) if $\lambda > \lambda_c$, then $\lambda_1 - \lambda_2 \to g$ as $N \to \infty$, almost surely, for some (N-independent) positive constant $g \equiv g(\lambda)$.*

We do not include precise formulas for the critical value λ_c and the gap g in the statement of Theorem 2.1 for the general cases since they are lengthy but not particularly informative. In the rest of the section, we focus on two specific models and check how the main result, Theorem 2.1, applies to them. Since two models defined in Sections 2.1 and 2.2 are symmetric stochastic block models with two communities, the transition occurs above the Kesten–Stigum threshold discussed in [1, 23].

Conjecture 2.1. (Kesten–Stigum threshold.) *Let (X, G) be drawn from an $N \times N$ symmetric SBM with k communities with probability p inside the communities and q across. Define*

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

$\mathrm{SNR} = N(p-q)^2/k(p+(1-k)q)$. *Then, for any* $k \geq 2$, *it is possible to solve weak recovery effectively if and only if* $\mathrm{SNR} > 1$.

Thus, we obtain the regime $p = q + w/\sqrt{N}$, where $w = \Theta(1)$ and $p = p_1$ in the following two models.

2.1. Hidden community model

In the hidden community model, only one of the intra-community connection probabilities is larger than the inter-community connection probability, and the other intra-connection probability coincides with the inter-community connection. The precise definition for such a model is as follows.

Definition 2.2. *(Hidden community model.)* Let $C \subset [n]$ such that $|C| = K$. Let O be an $N \times N$ symmetric matrix with $O_{ii} = 0$, where the O_{ij} are independent for $1 \leq i \leq j \leq N$ and

$$O_{ij} \sim \begin{cases} P & \text{if } i, j \in C, \\ Q & \text{otherwise} \end{cases}$$

for given probability measures P and Q.

We consider the BBP-type transition of the hidden community model with Bernoulli entries, i.e. $P = \mathrm{Bernoulli}(p)$ and $Q = \mathrm{Bernoulli}(q)$ with $p \neq q$, which also corresponds to the case $\alpha_2 = 1$ or $p_2 = q$ in (2.1). It is not hard to find that the transition occurs in the regime $p_1 := p = w/\sqrt{N} + q$ for some (possibly N-dependent) $w = \Theta(1)$. After shifting and rescaling, we find that $\lambda_2 \to 2$ and

$$\lambda_1 \to \begin{cases} \dfrac{\gamma w}{\sqrt{q(1-q)}} + \dfrac{\sqrt{q(1-q)}}{\gamma w} & \text{if } w > \dfrac{\sqrt{q(1-q)}}{\gamma}, \\ 2 & \text{if } w < \dfrac{\sqrt{q(1-q)}}{\gamma}. \end{cases}$$

See Section 4.1 for the detail.

We perform a numerical simulation for the hidden community model. We set $N = 2500$, $\gamma = 1/4$, and $q = 0.2$ for the sparse model and $q = 0.7$ for the dense model. Following the analysis in Section 4.1, we find that an outlier eigenvalue occurs if

$$p > q + \frac{\sqrt{q(1-q)}}{\gamma\sqrt{N}}.$$

Thus, we can show an outlier if $p > 0.232$ for the sparse model and $p \geq 0.737$ for the dense model. In Figure 1, we compare histograms of the eigenvalues of the shifted, rescaled adjacency matrices with $p = 0.21$ and $p = 0.25$ for the sparse model, and $p = 0.71$ and $p = 0.75$ for the dense model. For each case, we show a histogram of the two largest eigenvalues, λ_1 and λ_2, after 100 iterations. As predicted by the analysis, the outlier appears only for the cases $p = 0.25$ and $p = 0.75$. We can check the gap between λ_1 and λ_2 in Figure 1.

2.2. Unbalanced stochastic block model

We next consider the case $p_1 = p_2$ or $\alpha_1 = \alpha_2$ with $\gamma \neq 1/2$ to refer to an unbalanced stochastic block model. As in the hidden community model, the transition occurs in the regime

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

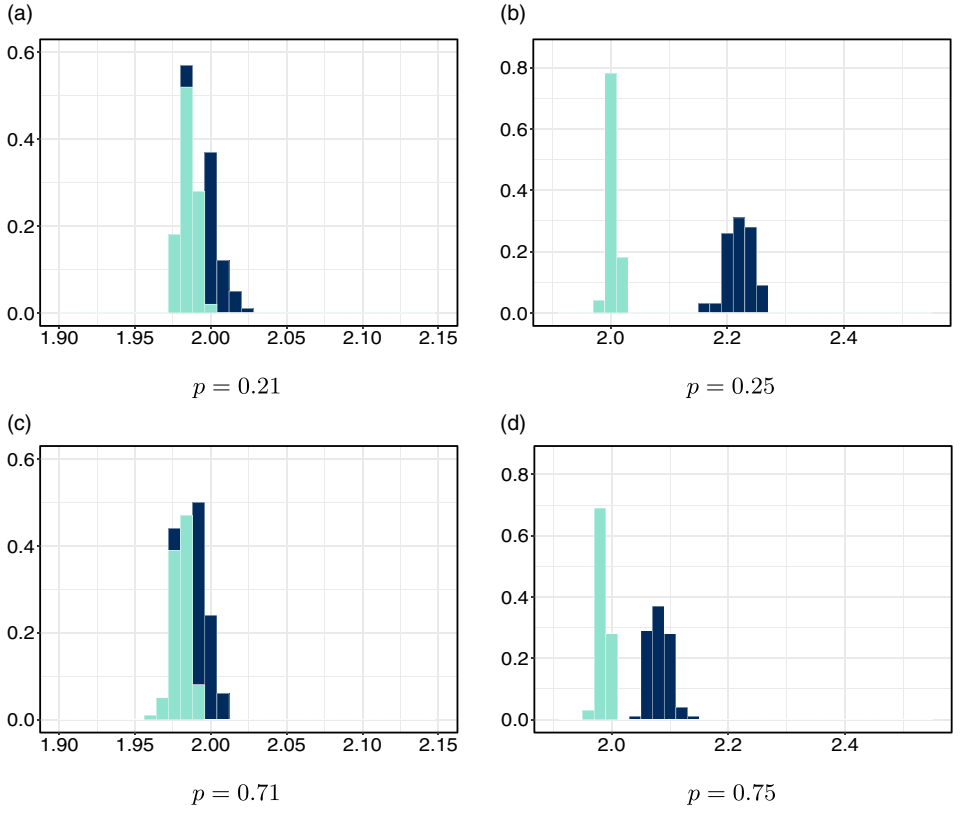

FIGURE 1. Histograms of the two largest eigenvalues of hidden community models generated after 100 iterations for sparse cases (top) and density cases (bottom). In each histogram, dark blue represents the largest eigenvalues, and light green represents the second-largest eigenvalues. The gap between two eigenvalues exists in (b) and (d).

$p_1 = p_2 := p = w/\sqrt{N} + q$. After shifting and rescaling, we find that $\lambda_2 \to 2$ and

$$\lambda_1 \to \begin{cases} \dfrac{w}{2\sqrt{q(1-q)}} + \dfrac{2\sqrt{q(1-q)}}{w} & \text{if } w > 2\sqrt{q(1-q)}, \\ 2 & \text{if } w < 2\sqrt{q(1-q)}. \end{cases}$$

See Section 4.2 for the detail. Note that the transition does not depend on γ.

We perform a numerical simulation for the unbalanced stochastic block model. As in the hidden community model, we set $N = 2500$, $\gamma = 1/4$, and $q = 0.2$ for the sparse model and $q = 0.7$ for the dense model. An outlier eigenvalue occurs if

$$p > q + \frac{2\sqrt{q(1-q)}}{\sqrt{N}}.$$

Thus, we can show an outlier if $p > 0.216$ for the sparse model and $p \geq 0.719$ for the dense model. In Figure 2, we compare histograms of the eigenvalues of the shifted, rescaled adjacency matrices with $p = 0.21$ and $p = 0.23$ for the sparse model, and $p = 0.71$ and $p = 0.73$ for

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

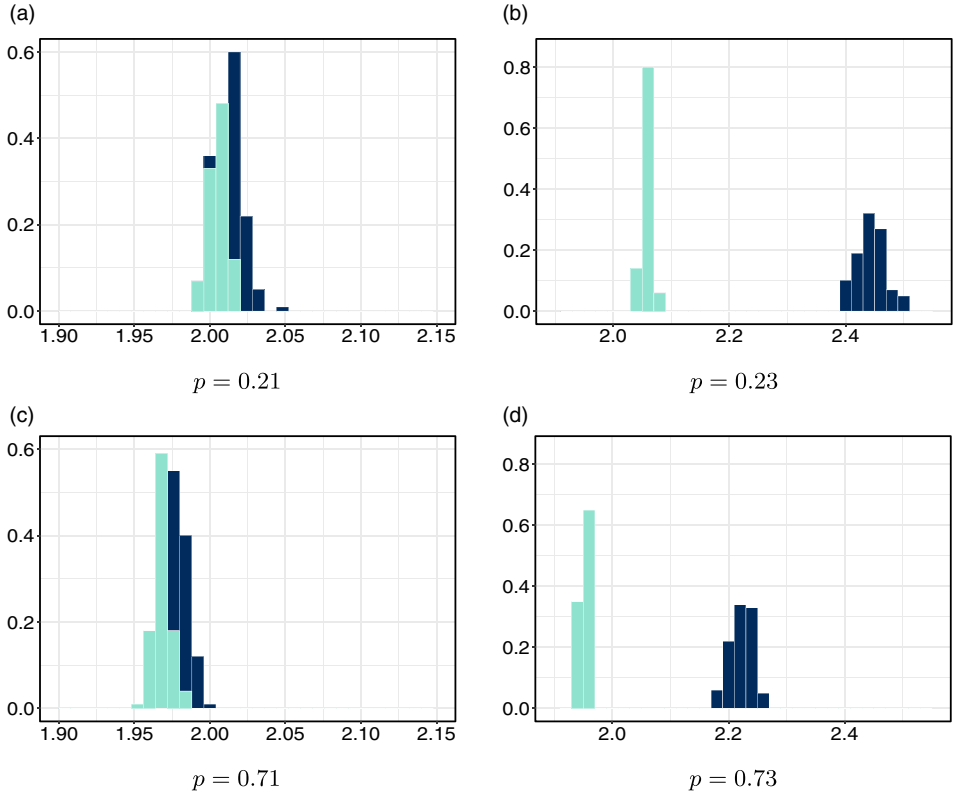

FIGURE 2. Histograms of the two largest eigenvalues of unbalanced stochastic block models generated after 100 iterations for sparse cases (top) and density cases (bottom). In each histogram, dark blue represents the largest eigenvalues, and light green represents the second-largest eigenvalues. The gap between two eigenvalues exists in (b) and (d).

the dense model. Again, for each case, we show histograms of the two largest eigenvalues, λ_1 and λ_2, after 100 iterations. As predicted by the analysis, the outlier appears only for the cases $p = 0.23$ and $p = 0.73$. We can check the gap between λ_1 and λ_2 in Figure 2.

3. Proof of Theorem 2.1

Proof of Theorem 2.1. Recall that we denote by λ_1 and λ_2 the two largest eigenvalues of M. Let μ_1 and μ_2 be the two largest eigenvalues of H. From Corollary 1.11 and its proof in [4] on Wigner-type matrices, we find that μ_1 and μ_2 converge to L_+, the rightmost edge of the limiting ESD of H. More precisely, there exists an N-independent constant $\delta > 0$ such that $|\mu_1 - L_+|,\ |\mu_2 - L_+| \le N^{-\delta}$ with overwhelming probability. (See Definition A.2 for the definition of the overwhelming probability.) By the Cauchy interlacing formula, we have the inequality $\mu_2 \le \lambda_2 \le \mu_1 \le \lambda_1$, which shows that λ_2 also converges to the rightmost edge of the limiting ESD of H.

To prove the limit of $\lambda_1 - \lambda_2$, we first notice that λ_1 is an increasing function of λ due to the following equation:

$$\lambda_1 = \max_{\|x\|=1} \langle x, Mx \rangle = \max_{\|x\|=1} \left(\langle x, Hx \rangle + \lambda |\langle x, u \rangle|^2 \right).$$

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

Choose $\varepsilon \in (0, 1)$, which is not necessarily independent of N. (At the end of the proof, we will let $\varepsilon \to 0$.) Since $\lambda_1 \geq \mu_1$ and $\lambda - \mu_1 \leq \lambda_1 \leq \lambda + \mu_1$, we find that $\lambda_1 - \mu_1 \leq \varepsilon$ whenever $\lambda < \varepsilon$. On the other hand, $\lambda_1 > \mu_1 + 1$ if $\lambda > 2\mu_1 + 1 > 2\mu_1 + \varepsilon$. Thus, since λ_1 is a continuous function of λ and $\lambda_2 \leq \mu_1$, we find that there exists a unique random variable $\widetilde{\lambda}_c \equiv \widetilde{\lambda}_c(N, H, \varepsilon)$ such that

- if $\lambda < \widetilde{\lambda}_c$, then $\lambda_1 - \lambda_2 \leq \varepsilon$;
- if $\lambda > \widetilde{\lambda}_c$, then $\lambda_1 - \lambda_2 \geq \varepsilon$.

We now consider λ_1 by applying the Stieltjes transform method in random matrix theory, for which we use the following definition.

Definition 3.1. *(Stieltjes transform.)* Let μ be a probability measure on the real line. The Stieltjes transform of μ is defined by

$$S_\mu(z) = \int_{\mathbb{R}} \frac{1}{x - z} \, \mathrm{d}\mu(x)$$

for $z \in \mathbb{C} \setminus \mathrm{supp}(\mu)$.

For the noise H, we consider its resolvent $G(z)$ defined by $G(z) := (H - zI)^{-1}$ for $z \in \mathbb{C} \setminus \mathrm{spec}(H)$. Note that the normalized trace $m := N^{-1}\mathrm{Tr}\, G$ is equal to the Stieltjes transform of the ESD of H.

Suppose that z is an eigenvalue of M such that $z > L_+ + \varepsilon$. (In particular, z is not an eigenvalue of H with overwhelming probability.) By definition, $\det\left(H + \lambda uu^\top - zI\right) = 0$, which can be further decomposed into

$$\begin{aligned} 0 = \det(H + \lambda uu^\top - zI) &= \det(H - zI)\left(I + (H - zI)^{-1}\lambda uu^\top\right) \\ &= \det(H - zI) \cdot \det\left(I + (H - zI)^{-1}\lambda uu^\top\right). \end{aligned} \tag{3.1}$$

Thus, if z is not an eigenvalue of H, we find that $\det(H - zI) \neq 0$, and hence

$$\det\left(I + (H - zI)^{-1}\lambda uu^\top\right) = 0.$$

Since $uu^\top$ is a rank-1 matrix, $(H - zI)^{-1}uu^\top$ also has rank one. Therefore, it has only one non-zero eigenvalue, which we call λ_0. Then, λ_0 is -1, for otherwise every eigenvalue of $I + (H - zI)^{-1}\lambda uu^\top$ is non-zero, contradicting (3.1). Furthermore, it is also obvious that $(H - zI)^{-1}\lambda u$ is an eigenvector associated with the eigenvalue -1. Thus,

$$(H - zI)^{-1}\lambda uu^\top (H - zI)^{-1}u = -(H - zI)^{-1}u,$$

which leads us to the equation

$$u^\top (H - zI)^{-1}u = \langle u, G(z)u \rangle = -\frac{1}{\lambda}. \tag{3.2}$$

For the noise matrix H, which is a Wigner-type matrix as considered in [4], we have

$$\langle u, G(z)u \rangle \simeq \sum_{i=1}^{N} m_i(z)u_i^2 \tag{3.3}$$

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

for any z outside the support of the limiting ESD of H, where $\mathbf{m} := (m_1, m_2, \dots, m_N)$ is the solution to the quadratic vector equation (QVE)

$$-\frac{1}{m_i(z)} = z + \sum_{j=1}^{N} \mathbb{E}\big[H_{ij}^2\big] m_j(z) \tag{3.4}$$

for $i, j = 1, 2, \dots, N$. (See Appendix A for a precise statement of (3.3).) We remark that the uniqueness of the solution m for (3.4) is also known [5].

To solve (3.2) using (3.4), we need to estimate $m(z)$ from the assumption on the community structure in Definition 2.1. From the symmetry, we have an ansatz:

$$m_1(z) = m_2(z) = \cdots = m_{N_1}(z), \qquad m_{N_1+1}(z) = \cdots = m_N(z).$$

Then, we can rewrite (3.4) as

$$-\frac{1}{m_i(z)} = \begin{cases} z + \displaystyle\sum_{j=1}^{N_1} \frac{\alpha_1}{N} m_j(z) + \sum_{j=N_1+1}^{N} \frac{1}{N} m_j(z) & \text{if } 1 \le i \le N_1, \\ z + \displaystyle\sum_{j=1}^{N_1} \frac{1}{N} m_j(z) + \sum_{j=N_1+1}^{N} \frac{\alpha_2}{N} m_j(z) & \text{if } N_1 + 1 \le i \le N, \end{cases}$$

which can be further simplified to

$$\begin{aligned} -1 &= z m_1(z) + \alpha_1 \gamma (m_1(z))^2 + (1-\gamma) m_1(z) m_N(z), \\ -1 &= z m_N(z) + \gamma m_1(z) m_N(z) + \alpha_2 (1-\gamma)(m_N(z))^2. \end{aligned} \tag{3.5}$$

We can thus conclude that, if there exists a real $\widehat{z}$ that solves (3.5) under the assumption

$$N_1 m_1(\widehat{z})\theta_1^2 + (N - N_1) m_N(\widehat{z})\theta_2^2 = N\big(\gamma m_1(\widehat{z})\theta_1^2 + (1-\gamma) m_N(\widehat{z})\theta_2^2\big) = -\frac{1}{\lambda} \tag{3.6}$$

obtained from (3.2) and (3.3), then $|\lambda_1 - \widehat{z}| \le N^{-\delta}$ for some (N-independent) $\delta > 0$ with overwhelming probability. On the other hand, if (3.5) has no real solution under the assumption in (3.6), then $|\lambda_1 - L_+| \le N^{-\delta}$ for some (N-independent) $\delta > 0$ with overwhelming probability.

We now consider the behavior of the random critical value $\widetilde{\lambda}_c(\varepsilon)$ via the deterministic equations (3.5) and (3.6). While it is possible to find the deterministic critical value λ_c for the existence of a real solution $\widehat{z}$ for (3.5) and (3.6), we take an indirect approach as follows. First, we notice from the existence of the critical value $\widetilde{\lambda}_c$ and the local law (3.3) that there also exists a deterministic constant $\lambda_c(\varepsilon)$ such that, for any $\lambda > \lambda_c(\varepsilon)$, there exists real $\widehat{z}$ that solves (3.5) and (3.6). Note that such a solution $\widehat{z}$ also satisfies a bound such as $\widehat{z} - L_+ > \varepsilon - N^{-\delta}$ for some δ from the definition of $\widetilde{\lambda}_c$. Considering the limit $\varepsilon \to 0$, we find that λ_c is determined as the largest number such that when $\lambda = \lambda_c$, $z = L_+$ solves (3.5) and (3.6).

So far, we have seen that λ_1 and λ_2 can be approximated by deterministic numbers $\widehat{z}$ and L_+ with overwhelming probability, depending on whether λ is larger than another deterministic number λ_c, again with overwhelming probability. In order to show that these deterministic numbers are N-independent, we notice that the vector $\|u\| = 1$ and hence there exist $\widehat{\theta}_1$ and $\widehat{\theta}_2$, independent of N, such that $N\theta_i^2 = \widehat{\theta}_i^2$ for $i = 1, 2$. This, in particular, implies that (3.6) is independent of N as well. Since (3.5) is also N-independent, we can conclude that $\widehat{z}$, L_+, and λ_c are N-independent. This completes the proof of Theorem 2.1. □

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

As a remark, we explain how to find λ_c. First, we change (3.5) to a single equation involving z and $\overline{m} \equiv \overline{m}(z) := \sum_{i=1}^{N} m_i(z)u_i^2$ only, i.e. $f(z, \overline{m}) = 0$. We have that the upper edge L_+ of the ESD of H is the largest real number such that $f(L_+, \overline{m}) = 0$ has a double root when considered as an equation for $\overline{m}$, which is a consequence of the square-root decay of the limiting ESD at the edge. (For more detail about the behavior of the limiting ESD of Wigner-type matrices, see [4, Theorem 4.1].) The condition can be technically checked easily by solving $f(L_+, \bar{m}) = 0$ and $(\partial/\partial m)f(L_+, \bar{m}) = 0$ simultaneously.

4. Examples from stochastic block models

This section considers stochastic block models corresponding to the G2-SBMs with Bernoulli distribution in our setting. Suppose that $\widehat{H} = [\widehat{H}_{ij}]_{i,j=1}^{N}$ is an SBM such that

$$\widehat{H} = \begin{cases} \widehat{H}_{ij} \sim \text{Bernoulli}(p_1) & \text{if } 1 \le i, j \le N_1, \\ \widehat{H}_{ij} \sim \text{Bernoulli}(p_2) & \text{if } N_1 + 1 \le i, j \le N, \\ \widehat{H}_{ij} \sim \text{Bernoulli}(q) & \text{otherwise.} \end{cases}$$

In what follows, we will say that the (i, j)-entry is in the diagonal block if $1 \le i, j \le N_1$ or $N_1 + 1 \le i, j \le N$, and otherwise it is in the off-diagonal block. In the block matrix form, it can also be expressed as follows:

$$\widehat{H} = \begin{pmatrix} \overbrace{\text{Bernoulli}(p_1)}^{N_1} & \overbrace{\text{Bernoulli}(q)}^{N-N_1} \\ \text{Bernoulli}(q) & \text{Bernoulli}(p_2) \end{pmatrix} \begin{matrix} \} N_1 \\ \} N-N_1 \end{matrix}$$

Our goal is to shift and rescale $\widehat{H}$ to convert it into a G2-SBM $M = H + \lambda uu^\top$ as in Definition 2.1. We first notice that the variances of the entries of $\widehat{H}$ are $p_1(1-p_1)$ and $p_2(1-p_2)$ for the diagonal block and $q(1-q)$ for the off-diagonal block. Since we assume that the variance of entry H_{ij} in the off-diagonal block is N^{-1}, we find that the matrix must be divided by $\sqrt{Nq(1-q)}$. It is then immediate that

$$\alpha_1 = \frac{p_1(1-p_1)}{q(1-q)}, \qquad \alpha_2 = \frac{p_2(1-p_2)}{q(1-q)},$$

as in (2.1).

The mean matrix

$$\mathbb{E}[\widehat{H}] = \begin{pmatrix} p_1 & q \\ q & p_2 \end{pmatrix}$$

is a rank-2 matrix, and thus we need to subtract from each entry a deterministic number that depends on the parameters p_1, p_2, and q. We continue the calculation in Sections 4.1 and 4.2 with two specific cases to obtain the G2-SBM form.

4.1. Hidden community model

Suppose that $p_1 = p$ and $p_2 = q$. It is then easy to find that $\mathbb{E}[\widehat{H}]$ becomes a rank-1 matrix after subtracting q from each entry, i.e. if we let E_0 be the $N \times N$ matrix all of whose entries are q, then

$$\mathbb{E}[\widehat{H}] - E_0 = \begin{pmatrix} p-q & 0 \\ 0 & 0 \end{pmatrix}.$$

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

Thus, we find that

$$M = \frac{1}{\sqrt{Nq(1-q)}}(\widehat{H} - E_0),$$
$$\mathbb{E}[M] = \frac{1}{\sqrt{Nq(1-q)}} \begin{pmatrix} p-q & 0 \\ 0 & 0 \end{pmatrix}.$$

Recall that $N_1 = \gamma N$ and $p = w/\sqrt{N} + q$. Since $\lambda u u^\top = \mathbb{E}[M]$, we get

$$u = \begin{pmatrix} 1/\sqrt{\gamma N} \\ 0 \end{pmatrix} \begin{matrix} \} N_1 \\ \} N - N_1 \end{matrix}$$

i.e. $\theta_1 = 1/\sqrt{\gamma N}$ and $\theta_2 = 0$. We also find that

$$\lambda = \frac{N_1(p-q)}{\sqrt{Nq(1-q)}} = \frac{\gamma w}{\sqrt{q(1-q)}}.$$

Following the proof of Theorem 2.1 in Section 3, we solve the system of equations in (3.5),

$$\begin{aligned} -1 &= z m_1 + \frac{p(1-p)}{q(1-q)} \gamma (m_1)^2 + (1-\gamma) m_1 m_N, \\ -1 &= z m_N + \gamma m_1 m_N + (1-\gamma)(m_N)^2. \end{aligned} \tag{4.1}$$

Since $p - q = O(N^{-1/2})$, we consider the ansatz $m_N = m_1 + O(N^{-1/2})$, which shows, for $m = \gamma m_1 + (1-\gamma) m_N$, that $1 + zm + m^2 = O(N^{-1/2})$. Following the analysis in the last paragraph of Section 3, we find that the upper edge $L_+ = 2 + O(N^{-1/2})$. By Theorem 2.1, it also implies that $\lambda_2 \to 2$ as $N \to \infty$.

In order to determine the location of the largest eigenvalue λ_1, we consider (4.1) under the assumption in (3.6),

$$N\left(\gamma m_1 \theta_1^2 + (1-\gamma) m_N \theta_2^2\right) = m_1 = -\frac{1}{\lambda} = -\frac{\sqrt{q(1-q)}}{\gamma w}. \tag{4.2}$$

We remark that the ansatz $m_N = m_1 + O(N^{-1/2})$ can be directly checked in this case by plugging (4.2) into (4.1) and eliminating z,

$$\left(\frac{p(1-p)}{q(1-q)} - 1\right) \gamma \left(\frac{\sqrt{q(1-q)}}{\gamma w}\right)^2 m_N + m_N + \frac{\sqrt{q(1-q)}}{\gamma w} = 0,$$

whose solution is

$$\begin{aligned} m_N &= -\frac{\sqrt{q(1-q)}/(\gamma w)}{1 + \left(\frac{p(1-p)}{q(1-q)} - 1\right) \gamma \left(\frac{\sqrt{q(1-q)}}{\gamma w}\right)^2} \\ &= -\frac{\sqrt{q(1-q)}}{\gamma w} \left(1 - \frac{\sqrt{N}(1-2q) - w}{N\gamma w + \sqrt{N}(1-2q) - w}\right). \end{aligned}$$

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

To find the location of the largest eigenvalue, we need to check whether the assumption in (4.2) is valid. However, we can instead find the value of z by first assuming that the solution exists. Then,

$$z = \frac{\gamma w}{\sqrt{q(1-q)}} + \frac{\sqrt{q(1-q)}}{\gamma w} + O\big(N^{-1/2}\big).$$

At the critical λ_c for the phase transition in Theorem 2.1, the location of the largest eigenvalue coincides with the location of the upper edge L_+ in the limit $N \to \infty$, or, equivalently, $\gamma w/\sqrt{q(1-q)} = 1$. Thus, we conclude that

$$\lambda_1 \to \begin{cases} \dfrac{\gamma w}{\sqrt{q(1-q)}} + \dfrac{\sqrt{q(1-q)}}{\gamma w} & \text{if } w > \dfrac{\sqrt{q(1-q)}}{\gamma}, \\ 2 & \text{if } w < \dfrac{\sqrt{q(1-q)}}{\gamma}. \end{cases}$$

4.2. Unbalanced stochastic block model

Suppose that $p_1 = p_2 = p$. Following the strategy in Section 4.1, we let E_1 be the $N \times N$ matrix all of whose entries are $(p+q)/2$. Then,

$$\mathbb{E}\big[\widehat{H}\big] - E_1 = \begin{pmatrix} (p-q)/2 & (q-p)/2 \\ (q-p)/2 & (p-q)/2 \end{pmatrix}.$$

Thus, we find that

$$M = \frac{1}{\sqrt{Nq(1-q)}}(\widehat{H} - E_1),$$
$$\mathbb{E}[M] = \frac{1}{\sqrt{Nq(1-q)}} \begin{pmatrix} (p-q)/2 & (q-p)/2 \\ (q-p)/2 & (p-q)/2 \end{pmatrix}.$$

From $\lambda u u^\top = \mathbb{E}[M]$, we get

$$u = \begin{pmatrix} 1/\sqrt{N} \\ -1/\sqrt{N} \end{pmatrix} \begin{matrix} \} N_1 \\ \} N - N_1 \end{matrix}$$

i.e., $\theta_1 = 1/\sqrt{N}$ and $\theta_2 = -1/\sqrt{N}$. Also,

$$\lambda = \frac{N(p-q)}{2\sqrt{Nq(1-q)}} = \frac{w}{2\sqrt{q(1-q)}}.$$

With $\alpha_1 = \alpha_2 = p(1-p)/q(1-q)$, we solve the system of equations in (3.5),

$$\begin{aligned} -1 &= z m_1 + \frac{p(1-p)}{q(1-q)} \gamma (m_1)^2 + (1-\gamma) m_1 m_N, \\ -1 &= z m_N + \gamma m_1 m_N + \frac{p(1-p)}{q(1-q)} (1-\gamma)(m_N)^2. \end{aligned} \tag{4.3}$$

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

Again, we consider the ansatz $m_N = m_1 + O(N^{-1/2})$, which leads us to the result that the upper edge $L_+ = 2 + O(N^{-1/2})$ and $\lambda_2 \to 2$ as $N \to \infty$. The assumption in (3.6) becomes

$$\gamma m_1 + (1-\gamma)m_N = -\frac{1}{\lambda} = -\frac{2\sqrt{q(1-q)}}{w}.$$

If the solution to (4.3) exists, it would be

$$z = \frac{w}{2\sqrt{q(1-q)}} + \frac{2\sqrt{q(1-q)}}{w} + O(N^{-1/2}).$$

At the critical λ_c, $w/2\sqrt{q(1-q)} = 1$, and thus we conclude that

$$\lambda_1 \to \begin{cases} \dfrac{w}{2\sqrt{q(1-q)}} + \dfrac{2\sqrt{q(1-q)}}{w} & \text{if } w > 2\sqrt{q(1-q)}, \\ 2 & \text{if } w < 2\sqrt{q(1-q)}. \end{cases}$$

5. Conclusion and future work

We have considered the generalized stochastic block model with two communities. We showed the phase transition for the G2-SBM where the random part is a Wigner-type matrix, which extends the BBP transition. For the precise formulas, we discussed a hidden community model and an unbalanced stochastic block model with Bernoulli distribution and Gaussian distribution at the Kesten–Stigum threshold. Moreover, referring to [25], along with suitable assumptions on the signal, our theorem can be adapted to both models with a non-Gaussian case.

We believe it is possible to prove the phase transition for the sparse matrix in which the data matrix is not necessarily symmetric, and most elements are composed of zeros. We also hope to extend our result to the G2-SBM with more than two communities.

Appendix A. Local law for Wigner-type matrices

In this section we provide a precise statement of the local law for Wigner-type matrices that was used in (3.3) in Section 3. Wigner-type matrices are defined as follows.

Definition A.1. *(Wigner-type matrix.)* We say an $N \times N$ random matrix $H = (H_{ij})$ is a Wigner-type matrix if the entries of H are independent real symmetric variables satisfying the following conditions:

- $\mathbb{E}(H_{ij}) = 0$ for all i, j.
- The variance matrix $\boldsymbol{s} = (s_{ij})$, where $s_{ij} = \mathbb{E}|H_{ij}|^2$, satisfies $(s^L)_{ij} \geq \rho/N$ and $s_{ij} \leq s_*/N$, $1 \leq i, j \leq N$, for finite parameters ρ, s_*, and L.

For the precise statement of the local law, we use the following definitions, which are frequently used in analysis involving rare events in random matrix theory.

Definition A.2. *(Overwhelming probability.)* An event Ω holds with overwhelming probability if, for any big enough $D > 0$, $\mathbb{P}(\Omega) \leq N^{-D}$ for any sufficient large N.

Definition A.3. *(Stochastic domination.)* Let $\psi^{(N)}(v)$ and $\phi^{(N)}(v)$ be non-negative random variables parametrized by elements v. Consider two families of non-negative random variables,

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

$$\psi = \{\psi^{(N)}(v) \mid N \in \mathbb{N},\ v \in U^{(N)}\}, \qquad \phi = \{\phi^{(N)}(v) \mid N \in \mathbb{N},\ v \in U^{(N)}\},$$

where $U^{(N)}$ is an N-dependent parameter set. Suppose $N_0 \colon (0, \infty)^2 \to \mathbb{N}$ is a given function depending on the model parameters C_D, γ, α_1, and α_2. If, for $\varepsilon > 0$ small enough and $D > 0$ big enough, we have $\mathbb{P}\big(\phi^{(N)} > N^{\varepsilon}\psi^{(N)}\big) \le N^{-D}$ for $N \ge N_0(\varepsilon, D)$, then ϕ is stochastically dominated by ψ, which is denoted by $\phi \prec \psi$.

The following definition of a QVE and its solution uniqueness is discussed on the complex upper half plane $\mathcal{H}$, where $\mathcal{H} = \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$, in [5].

Definition A.4. *(Quadratic vector equation.)* Consider a Banach space $\mathcal{R} := \{w \colon \mathscr{X} \to \mathbb{C} \mid \sup_{x \in \Lambda} |w_x| < \infty\}$ and its subset $\mathcal{R}_+ := \{w \in \mathcal{R} \mid \operatorname{Im} w_x > 0 \text{ for all } x \in \mathscr{X}\}$, where $\mathscr{X}$ is an abstract set of labels. Let $S \colon \mathcal{R} \to \mathcal{R}$ be a non-zero bounded linear operator, and $a \in \mathcal{R}$ a real bounded function. For all $z \in \mathcal{H}$, $-1/m(z) = z + a + S[m(z)]$, and its solution is $m \colon \mathcal{H} \to \mathcal{R}_+$.

Theorem A.1. (Solution of QVE (5).) *Let* $\mathbf{m} := (m_1, m_2, \ldots, m_N) \colon \mathcal{H} \to \mathcal{H}^N$ *be a function on* $\mathcal{H}$*. If a matrix* s *satisfies the conditions in Definition A.1, the QVE* $-1/m_i(z) = z + \sum_{j=1}^{N} s_{ij} m_j(z)$ *for* $i = 1, 2, \ldots, N$ *and* $z \in \mathcal{H}$ *has a unique solution.*

We are now ready to state the local law. Let ρ be the density defined as

$$\rho(\tau) := \lim_{\rho \searrow 0} \frac{1}{\pi N} \sum_{i=1}^{N} \operatorname{Im} m_i(\tau + \imath\eta).$$

(See also [4, Corollary 1.3] for more detail.)

Theorem A.2. (Local law [4].) *Let* H *be a Wigner-type matrix and fix an arbitrary* $\gamma \in (0, 1)$*. Then, uniformly for all* $z = a + bi$ *with* $b \ge N^{\gamma - 1}$*, the resolvent* $G(z) = (H - zI)^{-1}$ *satisfies*

$$max_{i,j} |G_{ij}(z) - m_i(z)\delta_{ij}| \prec \frac{1 + \sqrt{\rho(z)}}{\sqrt{bN}} + \frac{1}{bN}.$$

Furthermore, for any deterministic vector $w \in \mathbb{C}^N$ *with* $\max_i |w_i| \ge 1$*,*

$$\left| \sum_{i,j=1}^{N} \overline{w_i}(G_{ij}(z) - m_i(z)) \right| \prec \frac{1}{\sqrt{bN}}.$$

The local law can be generalized to the anisotropic local law as follows.

Theorem A.3. (Anisotropic law [4].) *Suppose the assumptions in Theorem A.2 hold. Then, uniformly for all* $z = a + bi$ *with* $b \ge N^{\gamma - 1}$*, and for any two deterministic* ℓ^2*-normalized vectors* $w, v \in \mathbb{C}^N$*,*

$$\left| \sum_{i,j=1}^{N} \overline{w_i} G_{ij}(z) v_i - \sum_{i=1}^{N} m_i(z) \overline{w_i} v_i \right| \prec \frac{1 + \sqrt{\rho(z)}}{\sqrt{bN}} + \frac{1}{bN}.$$

Funding information

The authors were supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2019R1A5A1028324).

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18**, 1–86.

[2] Abbe, E., Bandeira, A. S. and Hall, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62**, 471–487.

[3] Abbe, E. and Sandon, C. (2016). Crossing the KS threshold in the stochastic block model with information theory. In *2016 IEEE International Symposium on Information Theory (ISIT)*. Barcelona, Spain. pp. 840–844.

[4] Ajanki, O., Erdős, L. and Krüger, T. (2017). Universality for general Wigner-type matrices. *Prob. Theory Relat. Fields* **169**, 667–727.

[5] Ajanki, O., Erdős, L. and Krüger, T. (2019). *Quadratic Vector Equations on Complex Upper Half-Plane* (Memoirs of the American Mathematical Society **261**). American Mathematical Society, Providence, RI.

[6] Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Prob.* **33**, 1643–1697.

[7] Banks, J., Moore, C., Neeman, J. and Netrapalli, P. (2016). Information-theoretic thresholds for community detection in sparse networks. *Proc. Mach. Learn. Res.* **49**, 383–416.

[8] Barbier, J., Dia, M., Macris, N., Krzakala, F., Lesieur, T. and Zdeborová, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Information Processing Systems* **29** *(NIPS 2016)*, eds D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett. Curran Associates, Inc., Barcelona.

[9] Benaych-Georges, F. and Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227**, 494–521.

[10] Capitaine, M., Donati-Martin, C. and Féral, D. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *Ann. Prob.* **37**, 1–47.

[11] Chen, P.-Y. and Hero, A. (2015). Universal phase transition in community detectability under a stochastic block model. *Phys. Rev. E* **91**, 032804.

[12] Chung, H. W. and Lee, J. O. (2019). Weak detection of signal in the spiked Wigner model. *Proc. Mach. Learn. Res.* **97**, 1233–1241.

[13] Dumitriu, I. and Zhu, Y. (2019). Sparse general Wigner-type matrices: Local law and eigenvector delocalization. *J. Math. Phys.* **60**, 023301.

[14] El Alaoui, A., Krzakala, F. and Jordan, M. I. (2020). Fundamental limits of detection in the spiked Wigner model. *Ann. Statist.* **48**, 863–885.

[15] Erdös, L., Krüger, T. and Schröder, D. (2020). Cusp universality for random matrices I: Local law and the complex Hermitian case. *Commun. Math. Phys.* **378**, 1203–1278.

[16] Erdös, L. and Mühlbacher, P. (2019). Bounds on the norm of Wigner-type random matrices. *Random Matrices Theory Appl.* **8**, 1950009.

[17] Féral, D. and Péché, S. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Commun. Math. Phys.* **272**, 185–228.

[18] Hajek, B., Wu, Y. and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inf. Theory* **62**, 2788–2797.

[19] Hajek, B., Wu, Y. and Xu, J. (2017). Information limits for recovering a hidden community. *IEEE Trans. Inf. Theory* **63**, 4729–4745.

[20] Hajek, B., Wu, Y. and Xu, J. (2018). Recovering a hidden community beyond the Kesten–Stigum threshold in $o(|e| \log^* |v|)$ time. *J. Appl. Prob.* **55**, 325–352.

[21] Jung, J. H., Chung, H. W. and Lee, J. O. (2020). Weak detection in the spiked Wigner model with general rank. Preprint, arXiv:2001.05676.

[22] Mossel, E., Neeman, J. and Sly, A. (2018). A proof of the block model threshold conjecture. *Combinatorica* **38**, 665–708.

[23] Mossel, E. and Xu, J. (2016). Density evolution in the degree-correlated stochastic block model. *Proc. Mach. Learn. Res.* **49**, 1319–1356.

[24] Péché, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Prob. Theory Relat. Fields* **134**, 127–173.

[25] Perry, A., Wein, A. S., Bandeira, A. S. and Moitra, A. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. *Ann. Statist.* **46**, 2416–2451.

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press

[26] STANLEY, N., BONACCI, T., KWITT, R., NIETHAMMER, M. AND MUCHA, P. J. (2019). Stochastic block models with multiple continuous attributes. *Appl. Network Sci.* **4**, 1–22.
[27] ZHU, Y. (2020). A graphon approach to limiting spectral distributions of Wigner-type matrices. *Random Structures Algorithms* **56**, 251–279.

https://doi.org/10.1017/jpr.2023.44 Published online by Cambridge University Press