Performance Analysis of spectral community detection in realistic graph models

Hafiz Tiomoko Ali, Romain Couillet

PERFORMANCE ANALYSIS OF SPECTRAL COMMUNITY DETECTION IN REALISTIC GRAPH MODELS

Hafiz TIOMOKO ALI, Romain COUILLET

CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

ABSTRACT

This article proposes a spectral analysis of dense random graphs generated by (a modified version of) the degree-corrected stochastic block model, for a setting where the inter block probabilities differ by $\mathcal{O}(n^{-\frac{1}{2}})$ with n the number of nodes. We study a normalized version of the graph modularity matrix which is shown to be asymptotically well approximated by an analytically tractable (spiked) random matrix. The analysis of the latter allows for the precise evaluation of (i) the transition phase where clustering becomes asymptotically feasible and (ii) the alignment between the dominant eigenvectors and the block-wise canonical basis, thus enabling the estimation of misclassification rates (prior to post-processing) in simple scenarios.

Index Terms— networks, community detection, spectral analysis, graphs, random matrices.

1. INTRODUCTION

In many real world networks representable through graphs, the nodes can be grouped into communities based on their common features or interests. Discovering these groups and mapping the nodes to each group is one of the challenging tasks in network mining. To this end, various methods have been proposed, based on statistical inference (belief propagation, Bayesian inference, block modeling, model selection, information theory), spectral clustering, graph partitioning, modularity-based approaches, dynamic methods (random walks, synchronisation), etc. [1]. Most of these are however difficult to analyze when it comes to realistic networks, so that few theoretical guarantees are known to date. We focus in this article on spectral clustering methods which are both computationally inexpensive and theoretically tractable, while maintaining competitive performance versus the allegedly optimal belief propagation schemes if the network is dense, i.e., when the typical node degree is of order $\mathcal{O}(n)$. When the latter is instead $\mathcal{O}(1)$, the graph is considered sparse and spectral algorithms tend to be suboptimal, failing completely in some cases [2, 3] (other methods have been proposed to handle these cases, e.g., [4]). We shall assume here a dense network scenario.

The network model under present study is based on the stochastic block model (SBM), which extends the classical Erdös-Renyi graph model [5] to community structured graphs. As the SBM does not allow for degree heterogeneity inside blocks, thereby missing an important feature of realistic networks, we consider here the degree-corrected stochastic block model (DC-SBM), first proposed in [6]. Denoting $\mathcal{G}$ a K-class graph of n vertices with communities $\mathcal{C}_1, \ldots, \mathcal{C}_K$ and letting q_i, $1 \leq i \leq n$, be the intrinsic probability for node i to connect to any other network node, the DC-SBM assumes an adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, with A_{ij} independent

This work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006).

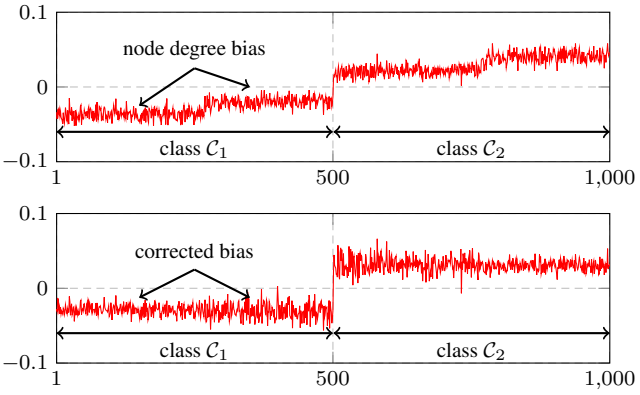

Fig. 1. Second leading eigenvector of $\mathbf{A}$ (top) and first leading eigenvector of $\mathbf{L}$ (bottom) with q_i following a bimodal distribution , two classes with equal proportions, $n = 1000$.

Bernoulli random variables with parameter $P_{ij} = q_i q_j C_{ab}$, for $i \in \mathcal{C}_a$ and $j \in \mathcal{C}_b$, where C_{ab} is a class-wise correction factor. In particular, if, for some $q_0 \in (0,1)$, $q_i = q_0$ for each i, one falls back into the classical homogeneous SBM. In the present dense network regime, $P_{ij} = \mathcal{O}(1)$; if the coefficients C_{ab} differ by $\mathcal{O}(1)$, clustering is asymptotically trivial as a vanishing misclassification rate is easily guaranteed as $n \to \infty$. We thus consider here the non-trivial regime where $C_{ab} = \mathcal{O}(1)$ but differ only by $\mathcal{O}(n^{-\frac{1}{2}})$.

Spectral clustering on the adjacency matrix of a DC-SBM however fails to cluster the nodes as the leading eigenvectors tend to follow a mixture of the degree distribution and class-wise canonical vectors, instead of purely aligning to the latter, therefore leading to ambiguities in classification and a trend to over-clustering (see top of Figure 1 and [7]). We thus work here on a normalized version $\mathbf{L}$ of the adjacency (precisely the modularity) matrix defined, for $\mathcal{D}(\mathbf{x}) = \mathrm{diag}(\mathbf{x})$ and $\mathbf{1}_n = [1, \ldots, 1]^\mathsf{T}$, by

$$\mathbf{L} = \frac{1}{\sqrt{n}} \mathcal{D}\left(\hat{\mathbf{q}}\right)^{-1} \left[\mathbf{A} - \frac{\hat{\mathbf{q}}\hat{\mathbf{q}}^\mathsf{T}}{\frac{1}{n}\hat{\mathbf{q}}^\mathsf{T}\mathbf{1}_n} \right] \mathcal{D}\left(\hat{\mathbf{q}}\right)^{-1}$$

where $\hat{\mathbf{q}} = [\hat{q}_1, \ldots, \hat{q}_n]^\mathsf{T}$, with $\hat{q}_i = \frac{1}{n}\left[\mathbf{A}\mathbf{1}_n\right]_i$.[1] We shall see (as already observed in [7]) that the dominant eigenvectors of $\mathbf{L}$ are strongly aligned to the class-wise canonical eigenvectors, thus recover the lost clustering ability of $\mathbf{A}$ (see bottom of Figure 1).

Being more challenging to analyze than $\mathbf{A}$ itself (due to its entries no longer being independent), our approach will first be to show that $\mathbf{L}$ is asymptotically well approximated by an analytically tractable random matrix, that falls in the family of the spiked random

[1] The notation $\mathbf{L}$ for this matrix is due to its Laplacian-looking form.

matrices [8], i.e., is formed by a deterministic low rank perturbation of a standard random matrix model. As shown in [8], the spectrum of these matrices is essentially composed of (one or several) clusters of eigenvalues and of finitely many isolated ones, and there exists a phase transition phenomenon by which, the larger the amplitude of the low rank matrix eigenvalues, the more eigenvalues tend to isolate from the aforementioned clusters and the more information is contained within their associated eigenvectors. In the present spectral clustering setting, using advanced tools from random matrix theory, we shall provide a precise analysis of these leading eigenvectors, henceforth shedding new light on the relation between the DC-SBM parameters and the classification performance of spectral clustering. All proofs are deferred to an extended version of this article.

Notations: Vectors are denoted with lowercase boldface letters and matrices by uppercase boldface letters. $\{\mathbf{v}_a\}_{a=1}^n$ is the column vector $\mathbf{v}$ with (scalar or vector) entries $\mathbf{v}_a$ and $\{\mathbf{V}_{ab}\}_{a,b=1}^n$ is the matrix $\mathbf{V}$ with (scalar or matrix) entries $\mathbf{V}_{ab}$. The operator $\mathcal{D}(\mathbf{v}) = \mathcal{D}\left(\{\mathbf{v}_a\}_{a=1}^n\right)$ is the diagonal matrix having (scalar or vector) $\mathbf{v}_1, \ldots, \mathbf{v}_n$ down its diagonal. The vector $\mathbf{1}_n \in \mathbb{R}^n$ stands for the vector filled with ones. The Dirac measure at x is δ_x. The vector $\mathbf{j}_a$ is the canonical vector of class $\mathcal{C}_a$ defined by $(\mathbf{j}_a)_i = \delta_{i\in\mathcal{C}_a}$ and $\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_K] \in \{0,1\}^{n\times K}$. The set $\mathbb{C}^+$ is $\{z \in \mathbb{C},\ \Im[z] > 0\}$.

2. MAIN RESULTS

We divide this section into a first analysis of the heterogeneous model, before particularizing the results to the homogeneous case.

2.1. Heterogeneous model

Consider an undirected random graph with n nodes belonging to one of K classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$ with cardinalities $|\mathcal{C}_k| = n_k$. Each node has an intrinsic probability q_i to get connected to any other vertex in the graph. Besides, we define $\mathbf{C} \in \mathbb{R}^{K\times K}$ a matrix of weights C_{ab} affecting the connection probability between all nodes in $\mathcal{C}_a$ and all nodes in $\mathcal{C}_b$. We shall assume for simplicity of exposition (but with no generality restriction) that the nodes are ordered by class, i.e., nodes 1 to n_1 constitute class $\mathcal{C}_1$, nodes n_1+1 to n_2 form class $\mathcal{C}_2$, and so on. The adjacency matrix $\mathbf{A}$ of the graph thus has independent entries (up to symmetry) with A_{ij} Bernoulli with probability $P_{ij} = q_i q_j C_{ab} \in (0,1)$ when $i \in \mathcal{C}_a$ and $j \in \mathcal{C}_b$ and we take $A_{ii} = 0$ (without loss of generality) for all $1 \le i \le n$.

We shall perform spectral clustering on the matrix $\mathbf{L}$ defined by

$$\mathbf{L} = \frac{1}{\sqrt{n}} \hat{\mathbf{D}}^{-1} \left[\mathbf{A} - \frac{\hat{\mathbf{q}}\hat{\mathbf{q}}^\mathsf{T}}{\frac{1}{n}\hat{\mathbf{q}}^\mathsf{T}\mathbf{1}_n} \right] \hat{\mathbf{D}}^{-1} \tag{1}$$

where $\hat{\mathbf{D}} = \mathcal{D}(\hat{\mathbf{q}})$ and

$$\hat{q}_i = \frac{1}{n}\sum_{j=1}^n A_{ij} = \frac{1}{n}\left[\mathbf{A}\mathbf{1}_n\right]_i .$$

In order to achieve non-trivial (asymptotic) misclassification rates, we shall assume the following growth rate conditions.

Assumption 1. *As $n \to \infty$,*

1. *$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$ for $a, b \in \{1, \ldots, K\}$, where $M_{ab} = \mathcal{O}(1)$; we shall denote $\mathbf{M} = \{M_{ab}\}_{a,b=1}^K$.*
2. *$q_i \in (0,1)$, $i \in \{1, \ldots, n\}$, are i.i.d. random variables with probability measure μ having compact support in $(0,1)$. We shall denote $m_\mu = \int t\mu(dt)$.*
3. *$\frac{n_i}{n} \to c_i > 0$ and we will denote $\mathbf{c} = \{c_k\}_{k=1}^K$.*

Under Assumption 1, it is easily shown that

$$\max_{1\le i\le n} |\hat{q}_i - m_\mu q_i| \to 0$$

almost surely, so that the $\hat{q}_i$ are, up to a constant, uniformly consistent estimators for the (a priori unknown) q_i.

Let us provide some intuition on the coming results. Note first that we may write[2]

$$\frac{1}{\sqrt{n}}\mathbf{A} = \underbrace{\frac{1}{\sqrt{n}}\mathbf{q}\mathbf{q}^\mathsf{T}}_{\mathbf{A}_{d,\sqrt{n}}} + \underbrace{\frac{1}{n}\left\{\mathbf{q}_{(a)}\mathbf{q}_{(b)}^\mathsf{T} M_{ab}\right\}_{a,b=1}^K}_{\mathbf{A}_{d,1}} + \underbrace{\frac{1}{\sqrt{n}}\mathbf{X}}_{\mathbf{A}_{r,1}}$$

where $\mathbf{q}_{(i)} = [q_{n_1+\ldots+n_{i-1}+1}, \ldots, q_{n_1+\ldots+n_i}]^\mathsf{T} \in \mathbb{R}^{n_i}$ ($n_0 = 0$) and $\mathbf{X} = \{X_{ij}\}_{i,j=1}^n$ has independent (up to symmetry) entries of zero mean and variances $\sigma_{ij}^2 = q_i q_j (1 - q_i q_j) + \mathcal{O}(n^{-\frac{1}{2}})$. Observing that $\mathbf{A}_{d,\sqrt{n}}$, $\mathbf{A}_{d,1}$, and $\mathbf{A}_{r,1}$ have spectral norms respectively of order $\mathcal{O}(\sqrt{n})$, $\mathcal{O}(1)$, and $\mathcal{O}(1)$, we may expand $\sqrt{n}\hat{\mathbf{D}}^{-1} = \mathcal{D}(n^{-\frac{1}{2}}\mathbf{A}\mathbf{1}_n)^{-1}$ via a Taylor expansion around the dominant term $\mathcal{D}(\mathbf{A}_{d,\sqrt{n}}\mathbf{1}_n)^{-1}$. Pre- and post-multiplying $n^{-\frac{1}{2}}\mathbf{A}$ by the Taylor expansions, we then retrieve a corresponding Taylor expansion for $\mathbf{L}$ and the following estimate.

Theorem 1. *Let Assumption 1 hold and let $\mathbf{L}$ be given by* (1)*. Then, as $n \to \infty$, $\|\mathbf{L} - \tilde{\mathbf{L}}\| \to 0$ in operator norm almost surely, where*

$$\tilde{\mathbf{L}} = \frac{1}{m_\mu^2}\left[\frac{1}{\sqrt{n}}\mathbf{T}\mathbf{D}^{-1}\mathbf{X}\mathbf{D}^{-1}\mathbf{T}^\mathsf{T} + \mathbf{U}\mathbf{M}\mathbf{U}^\mathsf{T}\right]$$

with $\mathbf{D} = \mathcal{D}(\mathbf{q})$, $\mathbf{T} = \mathbf{I}_n - \frac{\mathbf{1}_n\mathbf{q}^\mathsf{T}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}$ and $\mathbf{U} = \mathbf{J} - \mathbf{1}_n\mathbf{c}^\mathsf{T}$.

The matrix $\tilde{\mathbf{L}}$ follows an additive spiked random matrix model similar, but formally different, to that studied in e.g., [9]. This model is characterized by the fact that, under Assumption 1-2., as $n \to \infty$, the eigenvalues of $\tilde{\mathbf{L}}$ converge, to one another in one or several "bulks", but for a maximum of K of them (the rank of $\mathbf{U}\mathbf{M}\mathbf{U}^\mathsf{T}$) that can be found in-between bulks or on either side of the bulks. The alignment between the eigenvectors associated to those isolated eigenvalues and the eigenvectors of $\mathbf{U}\mathbf{M}\mathbf{U}^\mathsf{T}$ can be evaluated and will largely depend on the eigenvalues of $\mathbf{U}\mathbf{M}\mathbf{U}^\mathsf{T}$ as we shall presently observe. Interestingly, $\mathbf{U}$ is constituted by the class-vectors $\mathbf{j}_i$, while $\mathbf{M}$ contains information about the inter- and intra-class affinities. Consequently, the isolated eigenvalue-eigenvector pairs are expected to correlate to the class basis $\mathbf{J}$ as soon as the eigenvalues of $\mathbf{M}$ are sufficiently large. Our next objective is to explore this phenomenon through a careful analysis of the tractable approximate $\tilde{\mathbf{L}}$ of $\mathbf{L}$. Before introducing our main results though, we need the following intermediary result.

Lemma 1 (A deterministic equivalent)**.** *Define the resolvent $\mathbf{Q}_z = (n^{-\frac{1}{2}}\mathbf{T}\mathbf{D}^{-1}\mathbf{X}\mathbf{D}^{-1}\mathbf{T}^\mathsf{T} - z\mathbf{I}_n)^{-1}$. Then, for $z \in \mathbb{C}^+$, the system*

$$e_1(z) = \int \frac{1}{-zt - e_1(z) + e_2(z)t}\mu(dt)$$
$$e_2(z) = \int \frac{t}{-zt - e_1(z) + e_2(z)t}\mu(dt)$$

[2]Here subscript 'd, n^k' stands for deterministic term of order n^k and 'r, n^k' for random term of order n^k.

admits a unique solution $(e_1(z), e_2(z)) \in (\mathbb{C}^+)^2$*, and* $z \mapsto e_2(z)$ *is the Stieljies transform of a continuous probability measure of compact support* $\mathcal{S}$*.*[3] *Furthermore, for all* $z \in \mathbb{C} \setminus \mathcal{S}$*,*

$$\mathbf{Q}_z \leftrightarrow \left(-z\mathbf{I}_n - \mathbf{T}\left[e_1(z)\mathbf{D}^{-1} - e_2(z)\mathbf{I}_n\right]\mathbf{T}\right)^{-1}$$

where the notation $\mathbf{A} \leftrightarrow \mathbf{B}$ *stands for* $\frac{1}{n}\operatorname{tr}\mathbf{DA} - \frac{1}{n}\operatorname{tr}\mathbf{DB} \to 0$ *and* $\mathbf{d}_1^\mathsf{T}(\mathbf{A}-\mathbf{B})\mathbf{d}_2 \to 0$ *almost surely, for all deterministic Hermitian matrix* $\mathbf{D}$ *and deterministic vectors* $\mathbf{d}_i$ *of bounded norms.*

Identifying the isolated eigenvalues of $\mathbf{L}$ then boils down (by Theorem 1) to finding the large ρ solutions to $det(\tilde{\mathbf{L}} - \rho\mathbf{I}) = 0$. Following standard techniques (e.g., [8, 9]) along with Lemma 1, we then have the following limiting result.

Theorem 2 (Isolated Eigenvalues)**.** *Let Assumption 1 hold and, for* $z \in \mathbb{C} \setminus \mathcal{S}$ *(given in Lemma 1), define the* $K \times K$ *matrix*

$$\mathbf{G}_z = \mathbf{I}_K + e_2(z)\left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}$$

with $e_2(z)$ *given in Lemma 1. Let* $\rho \in \mathbb{R} \setminus \mathcal{S}$ *be such that* $\mathbf{G}_\rho$ *has a zero eigenvalue of multiplicity* κ_ρ*. Then there exists* $\lambda_i, \ldots, \lambda_{i+\kappa_\rho-1}$ *eigenvalues of* $m_\mu^2\mathbf{L}$ *converging to* ρ*.*

Theorem 2 is equivalent to saying that $-1/e_2(\rho)$ should be an eigenvalue of $(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$. Consequently, to show the existence and location of isolated eigenvalues, we need to solve in $\rho \notin \mathcal{S}$ the equation $-\ell e_2(\rho) = 1$ for each non zero eigenvalue ℓ of $(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$. Precisely, let us write $\mathcal{S} = \bigcup_{m=1}^M [S_{m,-}, S_{m,+}]$ with $S_{1,-} \leq S_{1,+} < S_{2,-} \leq \ldots < S_{M,+}$ and define $S_{0,+} = -\infty$ and $S_{M+1,-} = \infty$. Then, recalling that the Stieltjes transform of a real supported measure is necessarily increasing on $\mathbb{R}$, there exist isolated eigenvalues of $m_\mu^2\mathbf{L}$ in $(S_{m,+}, S_{m+1,-})$, $m \in \{0, \ldots, M\}$, for all large n almost surely, if and only if there exists eigenvalues ℓ of $(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$ such that

$$\lim_{x \downarrow S_{m,+}} e_2(x) < -\ell^{-1} < \lim_{x \uparrow S_{m+1,-}} e_2(x). \tag{2}$$

In particular, when $\mathcal{S} = [S_-, S_+]$ is composed of a single connected component (as when $\mathcal{S}$ is the support of the semi-circle law), then isolated eigenvalues of $m_\mu^2\mathbf{L}$ may only be found beyond S_+ if $\lim_{x \downarrow S_+} -\ell e_2(x) > 1$ or below S_- if $\lim_{x \uparrow S_-} -\ell e_2(x) < 1$, for some non-zero eigenvalue ℓ of $(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$.

Remark 1 (Maximum number of eigenvalues)**.** *As* $\mathbf{1}_K^\mathsf{T}(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}) = 0$*,* $\mathbf{1}_K$ *is a left-eigenvector of* $(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$ *with eigenvalue* 0*, and thus* $(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$ *is of maximum rank* $K-1$*, meaning that a maximum of* $K-1$ *isolated eigenvalues can be found in between or away from the bulks constituting* $\mathcal{S}$*.*

Let us now turn to the study of the eigenvectors. Our objective here is to correlate the eigenvectors associated with the eigenvalues determined in Theorem 2 to the canonical base vectors $\mathbf{j}_1, \ldots, \mathbf{j}_K$.

Theorem 3 (Eigenspace Projections)**.** *Let* $\lambda_i, \ldots, \lambda_{i+\kappa_\rho-1}$ *be a group of isolated eigenvalues of* $m_\mu^2\mathbf{L}$ *converging to* ρ *as per Theorem 2 and* $\ell = -1/e_2(\rho)$*. Further denote* $\mathbf{\Pi}_\rho$ *the projector on the eigenspace of* $\mathbf{L}$ *associated to these eigenvalues. Then,*

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{\Pi}_\rho\mathbf{J} = \frac{1}{\ell^2 e_2'(\rho)}\mathbf{\Upsilon}_\rho\left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right) + o(1)$$

almost surely, where

$$\mathbf{\Upsilon}_\rho = \sum_{j=1}^{\kappa_\rho} \frac{[\mathbf{V}_{r,\rho}]_j [\mathbf{V}_{l,\rho}]_j^\mathsf{T}}{[\mathbf{V}_{l,\rho}]_j^\mathsf{T}[\mathbf{V}_{r,\rho}]_j}$$

with $\mathbf{V}_{r,\rho}, \mathbf{V}_{l,\rho} \in \mathbb{R}^{K \times \kappa_\rho}$ *respectively sets of right and left eigenvectors of* $\left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}$ *associated with the eigenvalue* ℓ *(and* $[\mathbf{X}]_j$ *the* j*-th column of* $\mathbf{X}$*), and* $e_2'(\rho)$ *the derivative of* $e_2(z)$ *(defined in Lemma 1) along* z *evaluated at* ρ*.*

Remark 2 (Class-wise eigenvector means)**.** *Letting* $\mathbf{u}$ *be a unit multiplicity isolated eigenvector of* $\mathbf{L}$*, write*

$$\mathbf{u} = \sum_{a=1}^K \alpha_a \frac{\mathbf{j}_a}{\sqrt{n_a}} + \sigma_a \mathbf{w}_a$$

with $\alpha_a = n_a^{-\frac{1}{2}}\mathbf{u}^\mathsf{T}\mathbf{j}_a$ *and* $\mathbf{w}_a$ *a (noise) vector orthogonal to* $\mathbf{j}_a$ *and supported on the class-*$\mathcal{C}_a$ *indices. The coefficients* α_a *characterize the alignment between* $\mathbf{u}$ *and the basis vectors* $\mathbf{j}_a$*, and are thus key elements to understand the performance of spectral clustering based on* $\mathbf{u}$*. Now observe that Theorem 3 allows for an estimate of each* α_a*. Indeed, with* $\mathbf{\Pi}_\rho = \mathbf{u}\mathbf{u}^\mathsf{T}$ *in this unit multiplicity case,*

$$|\alpha_a|^2 = \left|\mathbf{u}^\mathsf{T}\frac{\mathbf{j}_a}{\sqrt{n_a}}\right|^2 = \left[\frac{1}{n}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\mathbf{J}^\mathsf{T}\mathbf{\Pi}_\rho\mathbf{J}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\right]_{aa}^2$$

thus allowing to retrieve α_a *up to a sign, while*

$$\alpha_a\alpha_b = \left[\frac{1}{n}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\mathbf{J}^\mathsf{T}\mathbf{\Pi}_\rho\mathbf{J}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\right]_{ab}$$

from which the sign of α_a *can be recovered.*

Remark 3 (Total noise)**.** *For arbitrary multiplicity* κ_ρ*, note that*

$$\begin{aligned} N_\rho &= \kappa_\rho - \operatorname{tr}\left(\frac{1}{n}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\mathbf{J}^\mathsf{T}\mathbf{\Pi}_\rho\mathbf{J}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\right) \\ &= \kappa_\rho\left(1 - \frac{1}{\ell^2 e_2'(\rho)}\right) \in (0, \kappa_\rho) \end{aligned}$$

measures the overall "noise" induced by the graph randomness in the κ_ρ *eigenvectors associated with* ρ *(*0 *for perfect alignment to* $\mathbf{J}$*, and* κ_ρ *for a complete misalignment) and is thus an important metric to assess the spectral clustering performance. In particular, for* $\kappa_\rho = 1$*,* $N_\rho = \sum_{k=1}^K \sigma_a^2$ *defined in Remark 2.*

As a consequence of these two remarks, we have the following corollary of Theorem 3.

Corollary 1 (Clustering Performance for $K = 2$, $n_1 = n_2$ and $\mathbf{M} = \left[\begin{smallmatrix} \alpha & \beta \\ \beta & \alpha \end{smallmatrix}\right]$)**.** *When* $K = 2$*,* $n_1 = n_2$*, by exchangeability arguments, with the definitions of Remark 2, we easily obtain that* $\alpha_1 = -\alpha_2$*, while* $\sigma_1^2 = \sigma_2^2$*, for* $\mathbf{u}$ *the (hypothetically) unique isolated eigenvector of* $\mathbf{L}$*. Clustering then boils down to deciding whether* u_i *is positive or negative for each* i*. Conjecturing asymptotic Gaussianity of the entries of* $\mathbf{u}$*, we then obtain the probability* $\mathbb{P}_c$ *of correct clustering as*

$$\mathbb{P}_c = 1 - \Phi\left(-\sqrt{\frac{1-N_\rho}{N_\rho}}\right) \tag{3}$$

where $N_\rho = 1 - (\ell^2 e_2'(\rho))^{-1}$ *and* $\Phi(x) = \frac{1}{2\pi}\int_{-\infty}^x e^{-\frac{t^2}{2}}dt$*.*

[3] That is, $e_2(z) = \int (t-z)^{-1}\nu(dt)$ for some real supported measure ν.

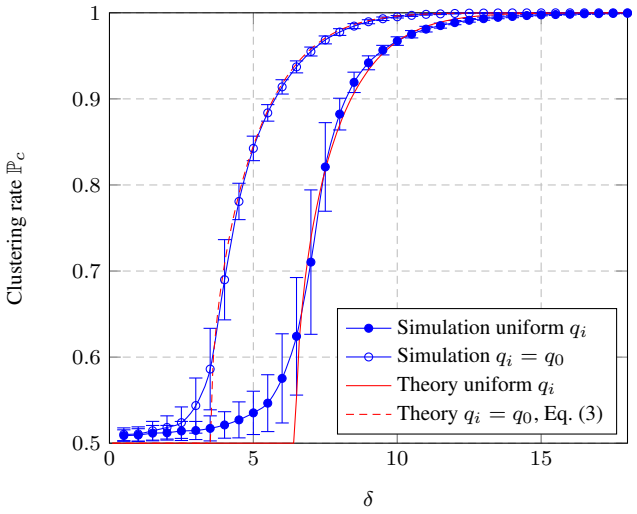

Fig. 2. Performance of community detection, for q_i uniformly distributed in $[.2, .8]$, $\mathbf{M} = \delta \mathbf{I}_2$, $c_1 = c_2 = \frac{1}{2}$, and for $q_i = q_0 = .5$. Simulations for $n = 2000$.

Figure 2 displays the empirical and theoretical correct clustering rates under the conditions of Corollary 1 for $\mathbf{M} = \delta \mathbf{I}_2$ and varying δ, with μ the uniform distribution in $[.2, .8]$ and $\mu = \delta_{.5}$. The phase transition point beyond which clustering is feasible is seen to be shifted to larger values of δ for the uniform distribution, which is a consequence of $\mathcal{S}$ being larger for more spread out measures μ, thus preventing the appearance of spiked eigenvalues. Under the same setting, with $\delta = 20$, Figure 3 displays the leading eigenvector of $\mathbf{L}$ along with the theoretically discovered class-wise means (from Remark 2) and standard deviations (from Corollary 1).

2.2. Homogeneous model

Let $\mu = \delta_{q_0}$, i.e., $q_i = q_0 \in (0, 1)$ for all i, which leads back to the homogeneous SBM. We assume to be unaware of the model homogeneity so that we keep $\hat{q}_i = n^{-1}[\mathbf{A}\mathbf{1}_n]_i$ as an estimator for q_0, instead of e.g., the more appropriate $\hat{q}_0 = n^{-2}\mathbf{1}_n^\mathsf{T}\mathbf{A}\mathbf{1}_n$.

Here, the expression of $e_2(z)$ becomes explicitly

$$e_2(z) = -\frac{z}{2\left(q_0^{-2}-1\right)} - \frac{\sqrt{z^2 - 4\left(q_0^{-2}-1\right)}}{2\left(q_0^{-2}-1\right)}$$

where the branch of the square root is chosen such that $e_2(z)$ is a Stieltjes transform (i.e., $e_2(z) \to 0$ as $|z| \to \infty$ and is analytic

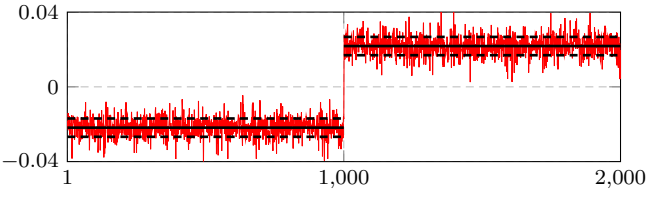

Fig. 3. Leading eigenvector of $\mathbf{L}$ for q_i uniformly distributed in $[.2, .8]$, $\mathbf{M} = \delta \mathbf{I}_2$, $c_1 = c_2 = \frac{1}{2}$, theoretical class-wise means (black) and (one) standard deviations (black-dashed), $n = 2000$.

on $\mathbb{C} \setminus \mathcal{S}$). The associated measure is the popular semi-circle law with $\mathcal{S} = [S_-, S_+] = [-2(q_0^{-2}-1)^{\frac{1}{2}}, 2(q_0^{-2}-1)^{\frac{1}{2}}]$. Besides, $\lim_{x \downarrow S_+} e_2(x) = -(q_0^{-2}-1)^{-\frac{1}{2}}$ and $\lim_{x \uparrow S_-} e_2(x) = (q_0^{-2}-1)^{-\frac{1}{2}}$. Thus, for ℓ a non-zero eigenvalue of $(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$, Condition (2) reduces to $|\ \ell\ | > \sqrt{q_0^{-2}-1}$ so that ℓ must grow large as $q_0 \to 0$ (sparser regime). If the condition is met, the isolated eigenvalues in the spectrum of $m_\mu^2 \mathbf{L}$ then have limit $\rho = \frac{q_0^2(\ell^2-1)+1}{q_0^2 \ell}$.

For the eigenspace projections, Theorem 3 becomes

$$\frac{1}{n}\mathbf{J}^\mathsf{T}\mathbf{\Pi}_\rho \mathbf{J} \to \left(\frac{q_0^2(\ell^2+1)-1}{\ell^2 q_0^2}\right) \sum_{j=1}^{\kappa_\rho} \frac{[\mathbf{V}_{r,\rho}]_j [\mathbf{V}_{l,\rho}]_j^\mathsf{T} \left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)}{[\mathbf{V}_{l,\rho}]_j^\mathsf{T} [\mathbf{V}_{r,\rho}]_j}.$$

In particular, the overall energy of the κ_ρ eigenvectors "noise" is

$$\kappa_\rho - \operatorname{tr}\left(\frac{1}{n}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\mathbf{J}^\mathsf{T}\mathbf{\Pi}_\rho \mathbf{J}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\right) \to \frac{\kappa_\rho}{\ell^2}\left(q_0^{-2}-1\right)$$

implying that, as $|\ell| \to \infty$ or $q_0 \to 1$, the eigenvectors of $\mathbf{L}$ tend to align perfectly to the basis vectors of $\mathbf{J}$.

These results are particularly interesting to adapt to the popular toy model (see e.g., [10]) where $K = 2$ and

$$\{P_{ij}\}_{i,j=1}^n = \begin{pmatrix} p_{\text{in}}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\mathsf{T} & p_{\text{out}}\mathbf{1}_{n_1}\mathbf{1}_{n_2}^\mathsf{T} \\ p_{\text{out}}\mathbf{1}_{n_2}\mathbf{1}_{n_1}^\mathsf{T} & p_{\text{in}}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\mathsf{T} \end{pmatrix}.$$

Letting $\Delta = \sqrt{n}p_{\text{in}}^{-1}(p_{\text{in}} - p_{\text{out}}) > 0$, $P_{ij} = q_0^2(1 + n^{-\frac{1}{2}}M_{ab})$ with $q_0^2 = p_{\text{in}}$ and $\mathbf{M} = \left(\begin{smallmatrix} 0 & -\Delta \\ -\Delta & 0 \end{smallmatrix}\right)$. There, the non zero eigenvalue of $(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T})\mathbf{M}$ is $\ell = 2c_1c_2\Delta$. Condition (2) here reads $(c_1c_2\Delta)^2 \geq 4(p_{\text{in}}^{-1} - 1)$. If met, the leading eigenvector $\mathbf{u}$ satisfies

$$\frac{1}{n}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}}\mathbf{J}^\mathsf{T}\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{J}\mathcal{D}(\mathbf{c})^{-\frac{1}{2}} \to (1-N)\begin{pmatrix} c_2 & -\sqrt{c_1c_2} \\ -\sqrt{c_1c_2} & c_1 \end{pmatrix}$$

where $N = (p_{\text{in}}^{-1} - 1)(2c_1c_2\Delta)^{-2}$.

3. CONCLUDING REMARKS

Although focused here on the normalized modularity matrix, this article has proposed a general framework for the study of the isolated eigenvectors of dense network matrix models (approximation by a tractable random matrix, spike analysis, and eigenvector parameter estimation). Our results so far nonetheless only allow to assess the performance of spectral clustering in elementary scenarios (as per Corollary 1); a more complete analysis would demand a deeper study of the class-wise variances σ_a^2 for each eigenvector (see Remark 2) along with the joint eigenvector fluctuations.

A key observation concerns the detrimental spectrum spreading of the normalized modularity matrix induced by degree heterogeneity, a phenomenon that simulations suggest is less present in the adjacency matrix itself. As the latter however introduces a node degree bias in the eigenvectors, a trade-off between resilience to node degree bias and to spectrum spreading needs be found when deciding on the choice of the matrix to operate.

Finally, our study yet involves dense networks, which are inappropriate models to many practical networks. Community detection over sparse networks however comes along with more stringent technical difficulties and spectral clustering on (derivatives of) the adjacency matrix is known to be suboptimal. In this scenario, the analysis of more involved matrix models, such as the non-backtracking matrix [4], is required. These considerations are left to future work.

4. REFERENCES

[1] S. Fortunato, "Community detection in graphs," Tech. Rep., Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Janv 2010.

[2] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborova, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications.," *Physical Review E, statistical, non linear and soft matter physics*, vol. 84, no. 6 pt 2:066106, 2011.

[3] C. Ohlan, "Graph partitioning via adaptive spectral techniques," *Combinatorics, Probabilities and Computing*, vol. 19, no. 02, pp. 227–284, 2010.

[4] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, , L. Zdeborova, and P. Zhang, "Spectral redemption in clustering sparse networks," *Combinatorics, Probabilities and Computing*, vol. 110, pp. 2093520940, 2013.

[5] P. Erdos and A. Rényi, "On random graphs," *Bull.Inst.Internat.Statist*, vol. 38, no. 4, pp. 343–347, 1961.

[6] B. Karrer and M. E. J. Newmann, "Stochastic block models and community structure in networks," *Phy.Rev*, vol. 83, no. 016107, Jan 2011.

[7] L. Gulikers, M. Lelarge, and L. Massoulie, "A spectral method for community detction in moderately-sparse degree corrected stochastic block models," Submitted to Physics.

[8] F. Benaych-Georges and R.R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," vol. 111, pp. 120–135, 2012.

[9] F. Chapon, R. Couillet, W. Hachem, and X. Mestre, "The ouliers among the singular values of large rectangular random matrices with additive fixed rank deformation.," 2013.

[10] R. R. Nadakuditi and M. E. J. Newmann, "Graph spectra and the detectability of community structure in networks," *Phy.Rev.Lett*, vol. 108, no. 188701, 2012.