Random matrix improved community detection in heterogeneous networks

Hafiz Tiomoko Ali and Romain Couillet
LANEAS Group, CentraleSupélec, University of Paris-Saclay, France
hafiz.tiomokoali@centralesupelec.fr, romain.couillet@centralesupelec.fr

***Abstract*—A new spectral method is proposed for community detection in large dense heterogeneous networks. We theoretically support and analyze an approach based on a novel "α-regularization" of the modularity matrix. We provide a consistent estimator for the choice of α inducing the most favorable community detection in worst case scenarios. We further prove that spectral clustering ought to be performed on a $1-\alpha$ regularization of the dominant eigenvectors (rather than on the eigenvectors themselves) to compensate for biases due to degree heterogeneity. Our clustering method is shown to be very promising on real world networks with competitive performances versus state-of-the-art spectral techniques developed for sparse homogeneous networks.**

I. Introduction

One of the prominent features of real world networks is their community structure which subdivides the network graphs in disjoint clusters: nodes are densely connected inside each cluster and sparsely connected across clusters. Extracting these clusters from the observation of the graph, and mapping the nodes to their respective clusters allow for a better understanding and analysis of large complex networks, and is one of the challenging tasks in modern network mining. There are two major methods in the literature to perform this task: statistical inference methods, which rely on some knowledge of the parameters of an underlying graph generative model and spectral algorithms, which classify vertices using the eigenvectors corresponding to outlying eigenvalues of some matrix representation of the network [1].

Large real world networks are usually sparse, in the sense that the average degree of the nodes remains constant with respect to the network size n as $n \to \infty$, and have heterogeneous degree distributions, often modelled as power laws [1]. Standard spectral algorithms based on variants of the adjacency matrix have been proved suboptimal in sparse network community detection, where they are supplanted by more powerful operators arising from statistical physics, notably the recently proposed non-backtracking [2] or Bethe Hessian (BH) matrix approaches [3]. The latter have however been developed under a stochastic block model (SBM) assumption which does not allow for degree heterogeneity inside clusters and tend to fail in the presence of highly varying node degrees in the network. To allow for degree heterogeneity modeling, in this article we instead consider the so-called degree-corrected stochastic block model (DCSBM), initially proposed in [4]. Denoting $\mathcal{G}$ a K-class graph of n vertices with communities $\mathcal{C}_1,\dots,\mathcal{C}_K$ and q_i, $1 \leq i \leq n$, an intrinsic probability for node i to connect to any other network node, the DCSBM assumes an adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ with A_{ij} independent 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 community-related correction factor (in general $C_{aa} > C_{ab}$ for all $a \neq b$).

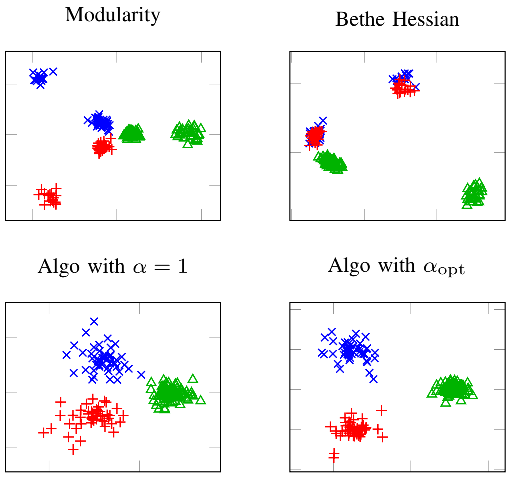

Fig. 1. Two dominant eigenvectors (x-y axes) for $n = 2000$, $K = 3$ classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ of sizes $|\mathcal{C}_1| = |\mathcal{C}_2| = \frac{n}{4}$, $|\mathcal{C}_3| = \frac{n}{2}$, intrinsic probabilities taking two values $q_1 = 0.1$, $q_2 = 0.5$, matrix of weights $\mathbf{C} = \mathbf{1}_3\mathbf{1}_3^{\mathsf{T}} + \frac{100}{\sqrt{n}}\mathbf{I}_3$. Colors correspond to ground truth classes.

To illustrate the aforementioned limitations of spectral methods under the DCSBM model, the top two graphs of Figure 1 provide 2D representations of dominant eigenvector 1 versus eigenvector 2 for the standard modularity matrix and the sparsity-improved BH matrix, when half the nodes connect with low probability q_1 and half the nodes with high probability q_2. For both methods, this erroneously induces the detection of extra communities and even a confusion of genuine communities in the BH approach. The bottom graphs of Figure 1 are the results of the "α-regularized" modularity approach proposed in this article, for an arbitrary and an optimized values of the parameter α, which both restore (with different levels of accuracy) the genuine community structure.

In addition to studying realistic degree corrected structured-graph models, our investigation is motivated by the lack of unification of the spectral methods for community detection in the literature, based on various normalizations of the adjacency matrix: the adjacency matrix itself $\mathbf{A}$ ($A_{ij} = 1$ if nodes i and j are connected, else 0), the modularity matrix $\mathbf{B} = \mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{\mathbf{d}^\mathsf{T}\mathbf{1}_n}$, with $\mathbf{d} = \mathbf{A}\mathbf{1}_n$ the vector of degrees, the unnormalized Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$, with $\mathbf{D} = \operatorname{diag}(\mathbf{d})$, the normalized Laplacian $\mathbf{L}_{\text{norm}} = \mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, the random walk matrix $\mathbf{D}^{-1}\mathbf{A}$, etc. [1]. In this article, we theoretically support the importance to study a generalized version $\mathbf{L}_\alpha \propto \mathbf{D}^{-\alpha}\mathbf{B}\mathbf{D}^{-\alpha}$, with $\alpha \in [0, 1]$. In particular, $\mathbf{L}_0 \propto \mathbf{B}$ and $\mathbf{L}_{1/2} \propto \mathbf{I}_n - \mathbf{L}_{\text{norm}}$. Our key objective is thus to understand the eigenstructure of $\mathbf{L}_\alpha$ and its consequences to community detection.

Our first main result is to show that $\mathbf{L}_\alpha$ is asymptotically well approximated by a *spiked* random matrix $\tilde{\mathbf{L}}_\alpha$, more amenable to analysis than $\mathbf{L}_\alpha$ itself. Spiked random matrices [5] are low rank perturbations of standard random matrices and their spectrum is generally composed of one or many "bulks" of eigenvalues and, whenever a phase transition is met, of additional eigenvalues which isolate from these bulks. In this case, the eigenvectors associated with the isolated eigenvalues are correlated to the low rank matrix eigenvectors. In our context, the approximation $\tilde{\mathbf{L}}_\alpha$ of $\mathbf{L}_\alpha$ tells us that: (i) the phase transition, which corresponds to the *community detectability threshold*, depends heavily on α and (ii) the dominant eigenvectors of $\mathbf{L}_\alpha$ are biased by the degree distribution (unless $\alpha = 0$) and thus need to be properly normalized. A deeper analysis reveals the existence of an optimal choice α_{opt} of α for which the detectability threshold is smallest, thus allowing for the weakest detectable community structure. We finally provide a consistent estimator $\hat{\alpha}_{\text{opt}}$ of α_{opt}.

Due to space limitations, the proofs of the main non-standard results are only sketched in the present manuscript but are available in full in an extended version.

Notations: Vectors (matrices) are denoted by lowercase (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 column 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 $\{z \in \mathbb{C},\ \Im[z] > 0\}$.

II. Main Results

A. Model and assumptions

We consider an n-node graph with K classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of sizes $|\mathcal{C}_k| = n_k$. To enforce a variation of the node degrees inside classes, we define a probability q_i for node i to connect to any node in the graph. We define $\mathbf{C} \in \mathbb{R}^{K \times K}$ a matrix of class weights C_{ab}, independent of the q_is, affecting the connection probability between nodes in $\mathcal{C}_a$ and nodes in $\mathcal{C}_b$. The adjacency matrix $\mathbf{A}$ of the graph has then independent entries (up to symmetry) which are Bernoulli random variables 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 set $A_{ii} = 0$ for all i. In the dense regime under consideration, $P_{ij} = \mathcal{O}(1)$. For convenience of exposition and without loss of generality, we assume that nodes indices are sorted by clusters and discard the nodes having no neighbor.

We shall perform spectral clustering on the family of matrices

$$\mathbf{L}_\alpha \equiv n^{2\alpha - \frac{1}{2}} \mathbf{D}^{-\alpha} \left[\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{\mathbf{d}^\mathsf{T}\mathbf{1}_n} \right] \mathbf{D}^{-\alpha}, \tag{1}$$

where $\mathbf{D} = \mathcal{D}(\mathbf{d})$ with $\mathbf{d} = \mathbf{A}\mathbf{1}_n$.

Clustering is asymptotically trivial when the weights C_{ab} differ by $\mathcal{O}(1)$, as a vanishing error classification rate is easily guaranteed [6]. We shall instead consider the regime where the clustering performance is not asymptotically perfect. This regime is ensured by the following growth rate conditions.

Assumption 1. *As $n \to \infty$, for all $a,b \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, n\}$:*

1) *$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$, where $M_{ab} = \mathcal{O}(1)$; we shall denote $\mathbf{M} = \{M_{ab}\}_{a,b=1}^K$.*
2) *q_i 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 $\mathbf{c} = \{c_k\}_{k=1}^K$.*

B. Preliminary results

First note that, under Assumption 1, since $C_{ab} \to 1$ as $n \to \infty$, $\frac{d_i}{\sqrt{\mathbf{d}^\mathsf{T}\mathbf{1}_n}}$ is a uniformly consistent estimator of q_i for $1 \leq i \leq n$, i.e.,

$$\max_{1 \leq i \leq n} \left| q_i - \frac{d_i}{\sqrt{\mathbf{d}^\mathsf{T}\mathbf{1}_n}} \right| \xrightarrow{\text{a.s.}} 0. \tag{2}$$

Also observe that we can write

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

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 $\sigma_{ij}^2 = q_i q_j (1 - q_i q_j) + \mathcal{O}(n^{-\frac{1}{2}})$.

To study $\mathbf{L}_\alpha$, which is not a standard random matrix model due to the dependency between $\mathbf{D}$, $\mathbf{d}\mathbf{d}^\mathsf{T}$, and $\mathbf{A}$, we shall perform a Taylor expansion of these matrices to retrieve an approximation $\tilde{\mathbf{L}}_\alpha$ of $\mathbf{L}_\alpha$ which is asymptotically consistent in *operator norm*. To this end, note 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 $\mathcal{O}(1)$,[1] so that $\frac{1}{\sqrt{n}}\left(\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\mathsf{T}}{\mathbf{d}^\mathsf{T}\mathbf{1}_n}\right)$ has all terms in $\mathcal{O}(1)$. Denoting $\mathbf{D}_q = \mathcal{D}(\mathbf{q})$, we next find

$$\mathbf{D} = \left(\mathbf{q}^\mathsf{T}\mathbf{1}_n\right)\mathbf{D}_q \left[\mathbf{I}_n + \sqrt{n}\frac{\mathbf{D}_q^{-1}}{\mathbf{q}^\mathsf{T}\mathbf{1}_n}\left(\mathcal{D}\left\{\mathbf{A}_{d,1}\mathbf{1}_n\right\} + \mathcal{D}\left\{\mathbf{A}_{a,1}\mathbf{1}_n\right\}\right)\right],$$

and the first order Taylor expansion of $\mathbf{D}^{-\alpha}$ is easily obtained as the last two terms in bracket are $\mathcal{O}(n^{-\frac{1}{2}})$. Putting all things together, we get the following approximation of $\mathbf{L}_\alpha$.

[1] The notation $\mathcal{O}(\cdot)$ is with respect to the operator norm.

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

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

$$\mathbf{U} = \begin{bmatrix} \frac{\mathbf{D}_q^{1-\alpha}\mathbf{J}}{\sqrt{n}} & \frac{\mathbf{D}_q^{-\alpha}\mathbf{X}1_n}{\mathbf{q}^\mathsf{T}1_n} \end{bmatrix},$$

$$\mathbf{\Lambda} = \begin{bmatrix} \left(\mathbf{I}_K - \mathbf{1}_K\mathbf{c}^T\right)\mathbf{M}\left(\mathbf{I}_K - \mathbf{c}\mathbf{1}_K^T\right) & -\mathbf{1}_K \\ -\mathbf{1}_K^T & 0 \end{bmatrix}.$$

Since $m_\mu^{2\alpha}\tilde{\mathbf{L}}_\alpha$ is a perturbation of $\frac{1}{\sqrt{n}}\mathbf{D}_q^{-\alpha}\mathbf{X}\mathbf{D}_q^{-\alpha}$ (full rank having zero mean and variance $\mathcal{O}(1)$ entries) by $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ (of maximum rank K), it follows a *spiked* random matrix model [7]. As illustrated in Figure 2, the spectrum of $m_\mu^{2\alpha}\tilde{\mathbf{L}}_\alpha$ (or equivalently of $m_\mu^{2\alpha}\mathbf{L}_\alpha$) is asymptotically given by a compact spectrum with a density (red curve in Figure 2) and (sometimes) by additional isolated eigenvalues, hereafter called *spikes*.

Because of the (asymptotic) spiked model nature of $\mathbf{L}_\alpha$, these spikes are only present beyond a phase transition threshold which depends on the norm of $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ (and thus indirectly of $\mathbf{M}$) and on the spread of the compact component of the spectrum of $\mathbf{L}_\alpha$ (which essentially depends on α). In the following, we shall first study the conditions for this phase transition phenomenon and shall find a resulting "best choice" for α, which ensures the most favorable phase transition for worst case matrices $\mathbf{M}$.

Beyond the phase transition, the dominant eigenvectors $\mathbf{u}_i$ of $\mathbf{L}_\alpha$ shall correlate to the space spanned by the columns of $\mathbf{U}$, so in particular to the $\mathbf{D}_q^{1-\alpha}\mathbf{j}_a$, $1 \leq a \leq K$, and not, as one would expect, to the canonical vectors $\mathbf{j}_a$ (determining the classes) themselves. The dominant eigenvectors of $\mathbf{L}_\alpha$ are then *biased* by the degrees carried in $\mathbf{D}_q$ and we thus propose to perform spectral clustering on the normalized vectors $\mathbf{D}^{\alpha-1}\mathbf{u}_i$ rather than on the vectors $\mathbf{u}_i$ themselves (recall that, up to a constant, $\mathbf{D}$ is a consistent estimator for $\mathbf{D}_q$).

These intuitive discussions are made rigorous in the subsequent section.

C. Eigenstructure of $\mathbf{L}_\alpha$

We first aim at defining the aforementioned transition point beyond which eigenvalues isolate from the main spectrum of $\mathbf{L}_\alpha$ and thus non-trivial clustering ought to be achievable. To this end, we first identify the support $\mathcal{S}^\alpha$ of the limiting spectral distribution of $m_\mu^{2\alpha}\mathbf{L}_\alpha$. The latter is defined through its Stieltjes transform $z \to e_2^\alpha(z)$ as follows.[2]

Lemma 1 (Limiting spectral distribution)**.** *For* $z \in \mathbb{C}^+$, *the system*

$$e_1^\alpha(z) = \int \frac{q^{1-2\alpha}}{-z - e_1^\alpha(z)q^{1-2\alpha} + e_2^\alpha(z)q^{2(1-\alpha)}}\mu(dq) \quad (3)$$

$$e_2^\alpha(z) = \int \frac{q^{2(1-\alpha)}}{-z - e_1^\alpha(z)q^{1-2\alpha} + e_2^\alpha(z)q^{2(1-\alpha)}}\mu(dq), \quad (4)$$

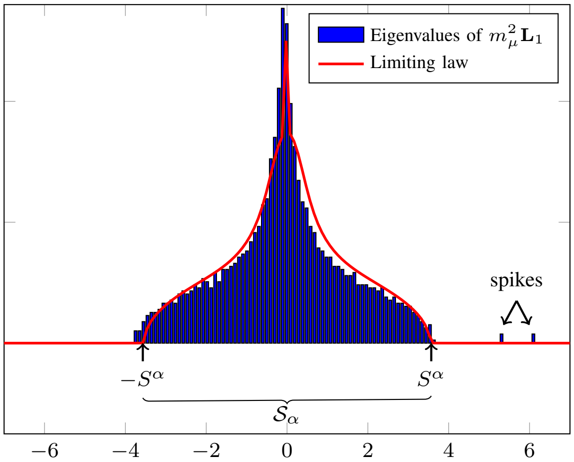

Fig. 2. Eigenvalues of $m_\mu^2\mathbf{L}_1$, $K = 3$, $n = 2000$, $c_1 = 0.3, c_2 = 0.3, c_3 = 0.4$, $\mu = \frac{1}{2}\delta_{q_1} + \frac{1}{2}\delta_{q_2}$, $q_1 = 0.4$, $q_2 = 0.9$, $\mathbf{M}$ defined by $M_{ii} = 12$, $M_{ij} = -4, i \neq j$.

admits a unique solution $(e_1^\alpha(z), e_2^\alpha(z)) \in (\mathbb{C}^+)^2$. *The map* $z \mapsto e_2^\alpha(z)$ *is the Stieljies transform of the limiting spectral distribution of* $m_\mu^{2\alpha}\mathbf{L}_\alpha$ *with compact support* $\mathcal{S}^\alpha = [-S^\alpha, S^\alpha]$.

Our next objective is to determine the existence and, if so, position of the isolated eigenvalues of $m_\mu^{2\alpha}\mathbf{L}_\alpha$. Following now popular spiked model tools, this entails the following result.

Theorem 2. *Let Assumption 1 hold and let* $\lambda_i(\bar{\mathbf{M}})$ *be a non zero eigenvalue of* $\bar{\mathbf{M}} \equiv \left(\mathcal{D}(\mathbf{c}) - \mathbf{c}\mathbf{c}^\mathsf{T}\right)\mathbf{M}$. *Then, for* $\alpha \in [0, 1]$, *there exists a corresponding isolated eigenvalue* $\lambda_i(\mathbf{L}_\alpha)$ *if and only if* [3]

$$\left|\lambda_i(\bar{\mathbf{M}})\right| > \tau^\alpha \equiv -\lim_{x \downarrow S^\alpha} \frac{1}{e_2^\alpha(x)},$$

with e_2^α *defined in Lemma 1. In this case,*

$$\lambda_i(\mathbf{L}_\alpha) = -\frac{1}{m_\mu^{2\alpha}e_2^\alpha(\lambda_i(\bar{\mathbf{M}}))}.$$

The value τ^α defined in Theorem 2 corresponds to a *community detectability threshold* beyond which sufficiently large eigenvalues $\lambda_i(\bar{\mathbf{M}})$ induce isolated eigenvalues in the spectrum of $\mathbf{L}_\alpha$.

When separability is guaranteed, we subsequently show that the *properly normalized* dominant eigenvectors of $\mathbf{L}_\alpha$ tend to be close to linear combinations of the $\mathbf{j}_a$'s (and thus are essentially noisy step functions), the closeness to $\mathbf{j}_a$ being ensured by the amplitude of the $\lambda_i(\bar{\mathbf{M}})$. This result, which again follows from classical spiked random matrix considerations (see e.g., [5], [7]), is given precisely as follows.

Theorem 3. *Under Assumption 1, let* $\lambda_i(\bar{\mathbf{M}})$ *and* $\lambda_i(\mathbf{L}_\alpha)$ *be an eigenvalue pair as defined in Theorem 2. We further assume* $\lambda_i(\mathbf{M})$ *of unit multiplicity and denote* $\mathbf{u}_i$ *the eigenvector*

[2] The Stieltjes transform of a measure ν is defined, for $z \in \mathcal{C} \setminus \operatorname{supp}(\nu)$, as $\int (t-z)^{-1}\nu(dt)$.

[3] The limit $\lim_{x \downarrow S^\alpha} e_2^\alpha(x)$ is well defined in $[-\infty, 0]$ as $x \to e_2^\alpha(x)$ is a growing negative function on the right side of S^α.

associated to $\lambda_i(\mathbf{L}_\alpha)$*. Then, letting* $\bar{\mathbf{u}}_i = \frac{\mathbf{D}^{\alpha-1}\mathbf{u}_i}{\|\mathbf{D}^{\alpha-1}\mathbf{u}_i\|}$ *and* $\mathbf{\Pi} = \sum_{a=1}^{K} \frac{\mathbf{j}_a \mathbf{j}_a^\mathsf{T}}{n_a}$*, as* $n \to \infty$*, almost surely,*

- *If* $\left|\lambda_i(\bar{\mathbf{M}})\right| \leq \tau^\alpha$*,* $\bar{\mathbf{u}}_i^\mathsf{T} \mathbf{\Pi} \bar{\mathbf{u}}_i \to 0$.
- *If* $\left|\lambda_i(\bar{\mathbf{M}})\right| > \tau^\alpha$*,* $\liminf_n \bar{\mathbf{u}}_i^\mathsf{T} \mathbf{\Pi} \bar{\mathbf{u}}_i > 0$.
- *For all* $\epsilon > 0$*, there exists* $T^\alpha > 0$ *such that, if* $\left|\lambda_i(\bar{\mathbf{M}})\right| > T^\alpha$*, then* $\liminf_n \bar{\mathbf{u}}_i^\mathsf{T} \mathbf{\Pi} \bar{\mathbf{u}}_i > 1 - \epsilon$.

From Theorems 2–3, it is clear that the smaller τ^α the more likely the condition $\left|\lambda_i(\bar{\mathbf{M}})\right| > \tau^\alpha$ is met. Spectral clustering with $\mathbf{L}_{\alpha_{\rm opt}}$ where

$$\alpha_{\rm opt} = \operatorname{argmin}_{\alpha \in [0,1]} \left\{ \tau^\alpha \right\},$$

is thus "optimal" in the sense that it allows for non-trivial clustering when $\lambda_i(\bar{\mathbf{M}})$ is only slightly larger than $\tau^{\alpha_{\rm opt}}$.

However, knowing $\alpha_{\rm opt}$ means being capable of estimating $e_2^\alpha(x)$ for all $\alpha \in [0,1]$, as per Lemma 2. This is in fact doable thanks to Equation (2) which ensures that all q_i, and thus μ, can be consistently estimated from the degrees d_i. We thus have the following result.

Lemma 2. *Define* $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\hat{q}_i}$ *with* $\hat{q}_i = \frac{d_i}{\sqrt{\mathbf{d}^\mathsf{T} \mathbf{1}_n}}$ *and* $\hat{e}_i^\alpha(z)$*,* $i \in \{1,2\}$*,* $\hat{\mathcal{S}}^\alpha$ *as in Lemma 1 but for* μ *replaced by* $\hat{\mu}$*. Then, as* $n \to \infty$*,*

$$\hat{\alpha}_{\rm opt} \to \alpha_{\rm opt}$$

almost surely, where $\hat{\alpha}_{\rm opt} = \operatorname{argmin}_{\alpha \in [0,1]} \{\hat{\tau}^\alpha\}$ *with*

$$\hat{\tau}_\alpha \equiv -\frac{1}{\lim_{x \downarrow \hat{S}^\alpha} \hat{e}_2^\alpha(x)}.$$

While all these results provide strong hints on the expected performance of spectral clustering from $\mathbf{L}_\alpha$, it still remains that the actual content of the dominant eigenvectors $\mathbf{u}_i$ is unknown for generic settings. In the longer version of this article, we explicitly retrieve the "noisy plateaus" structure of the $\mathbf{D}^{\alpha-1}\mathbf{u}_i$ and establish the consequences to clustering.

III. Numerical results

This section illustrates the performances of our proposed method as compared to state-of-the-art spectral clustering methods, such as the BH approach (appropriate for sparse networks generated from the SBM). We provide simulations both for synthetic data generated from the DCSBM and for real world benchmarks commonly considered in the literature. The performance evaluation is the overlap to ground truth communities, defined in [2] as

$$\text{Overlap} \equiv \frac{\frac{1}{n}\sum_{i=1}^{n} \delta(g_i \tilde{g}_i) - \frac{1}{K}}{1 - \frac{1}{K}},$$

where g_i and $\tilde{g}_i$ are the true and estimated label of node i, respectively. As the last step of all the spectral algorithms, we have performed 100 trials of the k-means algorithm on the $K-1$ leading eigenvectors (or $\mathbf{D}^{\alpha-1}$-normalized eigenvectors) $\mathbf{u}_1, \ldots, \mathbf{u}_{K-1}$ starting from K points in $\mathbb{R}^{K-1}$ randomly extracted from rows of the matrix $[\mathbf{u}_1, \ldots, \mathbf{u}_{K-1}]$. The selected k-means is chosen as the one with maximal modularity $\frac{1}{\mathbf{d}^\mathsf{T}\mathbf{1}_n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(A_{ij} - \frac{d_i d_j}{\mathbf{d}^\mathsf{T}\mathbf{1}_n} \right) \delta(\tilde{g}_i, \tilde{g}_j)$ [8].

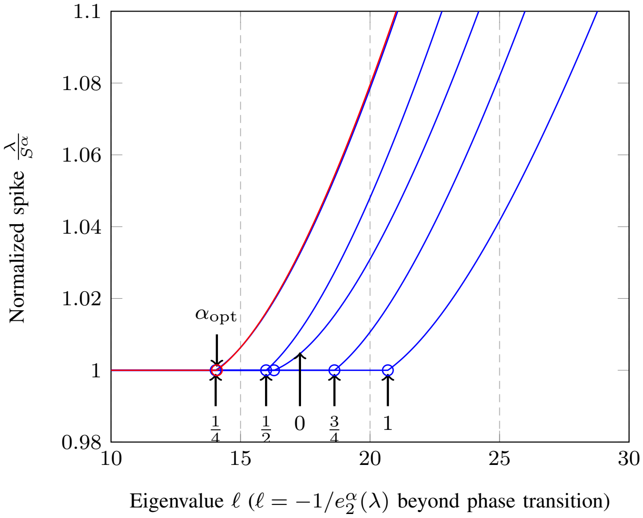

Fig. 3. Largest eigenvalue λ of $m_\mu^{2\alpha}\mathbf{L}_\alpha$ as a function of the largest eigenvalue ℓ of $\bar{\mathbf{M}}$, $\mathbf{M} = \Delta\mathbf{I}_3$, $c_i = \frac{1}{3}$, for $\Delta \in [10, 150]$, μ a power law with exponent 3 and support $[0.05, 0.3]$, for $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\rm opt}\}$ (indicated below the graph). Here, $\alpha_{\rm opt} = 0.28$. Circles indicate phase transition.

A. Synthetic graphs

In this section, we consider graphs generated from the DCSBM. We first consider a rather realistic graph scenario (quite sparse with power law degree distribution). Figure 3 presents the theoretical (asymptotic) ratio between the largest eigenvalue λ of $m_\mu^{2\alpha}\mathbf{L}_\alpha$ and the right edge S^α of the main eigenvalue spectrum support, with respect to the amplitude of the elements of $\mathbf{M}$, which one expects large to achieve good clustering performance. As predicted by our theoretical analysis, in the worst case scenario for $\mathbf{M}$, this ratio is greater than 1 (and thus an eigenvalue escapes $\mathcal{S}^\alpha$) only when $\alpha = \alpha_{\rm opt}$. Besides, for fixed amplitude of the eigenvalues of $\mathbf{M}$, the ratio obtained with $\alpha_{\rm opt}$ is seen to be the largest as compared to other values of α in the range $[0,1]$; this suggests (without any theoretical support) that better clustering is achieved using $\mathbf{L}_{\alpha_{\rm opt}}$ even in non-worst case scenarios.

Figure 4 subsequently shows the overlap performance under the setting of Figure 3. Although we consider in that example a quite sparse regime which, for $n = 2000$, is not very consistent with our assumptions, our proposed method outperforms the BH approach, especially around the worst case values for $\mathbf{M}$. It is worth mentioning that the empirically observed phase transitions closely match the theoretical ones (drawn in circles and the same as in Figure 3) but for the case $\alpha = 1$. This mismatch is due to $\mathbf{C}$ taking values too far from 1, especially for large Δ, thereby no longer conforming with Assumption 1.

We finally present in Figure 5 an example where the BH algorithm fails due to strongly heterogeneous node degrees. Assuming nodes connect with either low $q_1 = 0.1$ or high $q_2 > q_1$ intrinsic probability, we observe a sudden drop of the BH overlap once $q_2 - q_1$ is too large. This phenomenon is consistent with the fact, observed earlier in Figure 1, that BH

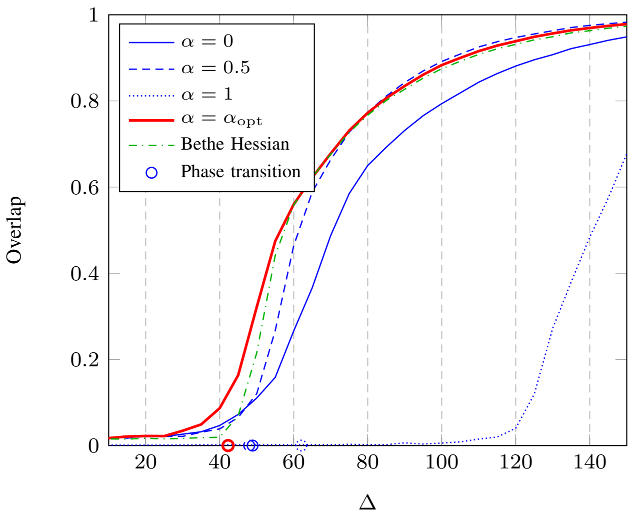

Fig. 4. Overlap for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, μ a power law with exponent 3 and support $[0.05, 0.3]$, $\mathbf{M} = \Delta\mathbf{I}_3$, for $\Delta \in [10, 150]$. Here $\alpha_{\mathrm{opt}} = 0.28$.

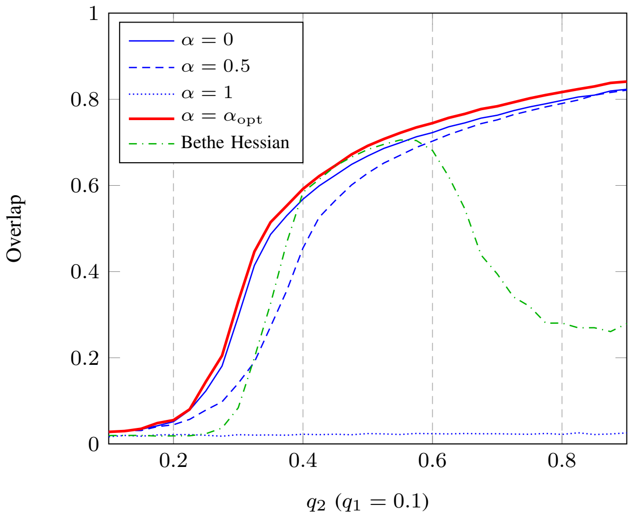

Fig. 5. Overlap for $n = 3000$, $K = 3$, $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ with $q_1 = 0.1$ and $q_2 \in [0.1, 0.9]$, $\mathbf{M}$ defined by $M_{ii} = 10$, $M_{ij} = -10, i \neq j$, $c_i = \frac{1}{3}$.

creates artificial communities out of nodes with the same q_i parameter. This is a practical demonstration of the need for a proper eigenvector normalization to avoid degree biases.

B. Real world benchmarks

We finally confront the overlap performance on real world benchmarks in Table I. The best overlap score for each benchmark is set in boldface and quasi-equal scores in italic. Our approach largely outperforms the BH method on some benchmarks and has competitive performances on others. However note that, for so small network sizes, the performance achieved by $\mathbf{L}_{\hat{\alpha}_{\mathrm{opt}}}$ may be quite unsatisfactory.

IV. Conclusion

We have introduced a new approach to community detection in large dense heterogeneous graphs using tools from random matrix theory. A salient feature of the approach lies in its automatically choosing the regularization of the modularity matrix which offers optimal performances. This work complements the recent breakthroughs in community detection, currently focusing on sparse but homogeneous graphs. A natural next step is to confront both approaches on a common ground, which is a challenging task due to the inadequacy (down to the mathematical tools) between sparse and dense network considerations. Studying the BH matrix in the dense heterogeneous regime is a natural first step into this investigation.

Graph	$\alpha = 0$	$\alpha = \frac{1}{2}$	$\alpha = 1$	$\hat{\alpha}_{\mathrm{opt}}$	BH
Polbooks[9]	*0.743*	**0.757**	0.214	*0.743*	**0.757**
Adjnoun[9]	0.571	**0.714**	0.000	0.571	0.661
Karate[10]	0.176	*0.941*	0.353	0.176	**1.000**
Dolphins[11]	**0.968**	**0.968**	0.387	**0.968**	*0.935*
Polblogs[12]	**0.897**	0.035	0.040	**0.897**	0.304
Football[9]	0.858	*0.905*	*0.905*	*0.905*	**0.924**

TABLE I

Overlap performance on benchmark graphs.

References

[1] Santo Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[2] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang, "Spectral redemption in clustering sparse networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20935–20940, 2013.

[3] Alaa Saade, Florent Krzakala, and Lenka Zdeborová, "Spectral clustering of graphs with the bethe hessian," in *Advances in Neural Information Processing Systems*, 2014, pp. 406–414.

[4] Brian Karrer and Mark EJ Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, pp. 016107, 2011.

[5] Florent Benaych-Georges and Raj Rao Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[6] Lennart Gulikers, Marc Lelarge, and Laurent Massoulié, "A spectral method for community detection in moderately-sparse degree-corrected stochastic block models," *arXiv preprint arXiv:1506.08621*, 2015.

[7] Francois Chapon, Romain Couillet, Walid Hachem, and Xavier Mestre, "The outliers among the singular values of large rectangular random matrices with additive fixed rank deformation," *arXiv preprint arXiv:1207.0471*, 2012.

[8] Mark EJ Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[9] Mark EJ Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, pp. 036104, 2006.

[10] Wayne W Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, pp. 452–473, 1977.

[11] David Lusseau, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten, and Steve M Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.

[12] Lada A Adamic and Natalie Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 36–43.