Revisiting the Bethe-Hessian: Improved Community Detection in Sparse Heterogeneous Graphs

Lorenzo Dall'Amico, Romain Couillet, Nicolas Tremblay

Revisiting the Bethe-Hessian: Improved Community Detection in Sparse Heterogeneous Graphs

Lorenzo Dall'Amico
GIPSA-lab, UGA, CNRS, Grenoble INP
lorenzo.dall-amico@gipsa-lab.fr

Romain Couillet
GIPSA-lab, UGA, CNRS, Grenoble INP
L2S, CentraleSupélec, University of Paris Saclay

Nicolas Tremblay
GIPSA-lab, UGA, CNRS, Grenoble INP

Abstract

Spectral clustering is one of the most popular, yet still incompletely understood, methods for community detection on graphs. This article studies spectral clustering based on the Bethe-Hessian matrix $H_r = (r^2 - 1)I_n + D - rA$ for sparse heterogeneous graphs (following the degree-corrected stochastic block model) in a two-class setting. For a specific value $r = \zeta$, clustering is shown to be insensitive to the degree heterogeneity. We then study the behavior of the informative eigenvector of H_ζ and, as a result, predict the clustering accuracy. The article concludes with an overview of the generalization to more than two classes along with extensive simulations on synthetic and real networks corroborating our findings.

1 Introduction

Network theory studies the interaction of connected systems of agents. Real networks tend to be structured in affinity classes and the problem of clustering consists in retrieving these unknown classes from the observed network pairwise interactions [1]. Belief propagation (BP) is an efficient way to reconstruct communities and – under certain conditions (see [2]) – was proved to give *optimal* reconstruction. On the negative side, BP suffers from a possibly long convergence time and a non-trivial implementation. Among the alternative clustering algorithms, spectral techniques proved particularly efficient in terms of speed and analytical tractability [3–6]. In the dense regime, in particular, where the average node degree scales like the size of the network, random matrix theory [4, 7, 8] manages to predict the asymptotic spectral clustering performances and to identify transition points beyond which asymptotic non trivial classification is achievable. This is however not the typical condition for real networks that tend instead to be *sparse*. For a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = n$ nodes, the condition of sparsity means that the average degree d does not depend on the size of the network and in particular $d \ll n$.

Both standard spectral clustering methods and their associated random matrix asymptotics collapse in this regime. As an answer, many intuitions emerged from statistical physics and led to important seminal steps. Notably, two deeply connected matrices recently proved to overcome the problem of sparsity: the $n \times n$ Bethe-Hessian [9] H_r with $r \in \mathbb{R}$ a parameter to be fixed – the study of which is the object of the present article–, and the non symmetric non backtracking operator $B \in \{0, 1\}^{2|\mathcal{E}| \times 2|\mathcal{E}|}$ [10]. Both matrices were introduced and studied under the homogeneous degree stochastic block model (SBM). Narrowing to the case of two communities it was proved both experimentally and theoretically [11, 2, 12, 13] that, if there exists an algorithm able to detect communities better then random guess, then these two matrices can be used to give non-trivial node partition. It is said that both algorithms work *down to the detectability threshold*.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

However, real networks are rarely homogeneous and typically follow a power law degree distribution [14]. The results of [15, 16] generalize the above studies to heterogeneous networks, generated by degree-corrected stochastic block models (DC-SBM) [17] and suggest that both B and H_r provide also in this case non trivial clustering down to the detectability threshold. Yet, a precise characterization of their behavior and performances is still lacking; the present article shows that some aspects of the behavior of B and H_r have indeed been overlooked.

Spectral clustering in sparse heterogeneous networks has also been tackled using various regularized Laplacian matrices [18–20] but, to our knowledge, these are not proved to operate down to the detectability threshold. These structurally different methods are discussed in concluding remarks.

The main message of the present communication is that, under a DC-SBM setting, the choice of r in H_r proposed in [9] for the SBM setting is suboptimal. We propose and theoretically support an improved parametrization $r = \zeta$ that allows the Bethe-Hessian H_ζ to efficiently detect communities in sparse and heterogeneous graphs. In detail, under the DC-SBM setting, a) we propose a spectral algorithm on H_ζ which performs efficiently down to the detectability threshold, with an informative eigenvector not tainted by the degree distribution (unlike in [9]); b) the algorithm is generalized to k-class clustering with a consistent estimation procedure for k; c) substantial performance improvements on the originally proposed Bethe-Hessian are testified by simulations on synthetic and real networks.

The remainder of the article is organized as follows: Section 2 argues on the optimal value $r = \zeta$ for H_r and, based on heuristic arguments, studies the behavior of the informative eigenvector of H_ζ, concluding with an explicit expression of the clustering performance; Section 3 provides an unsupervised method to estimate ζ, drawing on connections with the non-backtracking matrix B; Section 4 extends the algorithm to a k-class scenario; numerical supports are then provided in Section 5 on both synthetic and real networks; concluding remarks close the article.

Reproducibility. A Python implementation of the proposed algorithm along with codes to reproduce the results of the article are available at lorenzodallamico.github.io/codes.

2 Model and Main Results

2.1 Model setting

Consider an undirected binary graph $\mathcal{G}(\mathcal{E}, \mathcal{V})$, with nodes $\mathcal{V} = \{1, \dots, n\}$ ($|\mathcal{V}| = n$) and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ ($|\mathcal{E}| = m$). Let $\boldsymbol{\sigma} \in \{-1, 1\}^n$ be the vector of class labels, both classes being of equal size (i.e., $\sum_i \sigma_i = 0$), and $C = \left(\begin{smallmatrix} c_{\rm in} & c_{\rm out} \\ c_{\rm out} & c_{\rm in} \end{smallmatrix}\right)$. These assumptions are meant to set the problem in a more readable symmetric scenario. Section 4 extends the results to multiple classes of possibly different sizes. In order to account both for sparsity and heterogeneity, we consider the DC-SBM as a generative model for $\mathcal{G}$. Denoting $A \in \{0, 1\}^{n \times n}$ the adjacency matrix defined by $A_{ij} = 1_{(i,j) \in \mathcal{E}}$, the DC-SBM generates edges independently according to:

$$\mathbb{P}(A_{ij} = 1 | \sigma_i, \sigma_j, \theta_i, \theta_j) = \theta_i \theta_j \frac{C_{\sigma_i, \sigma_j}}{n}, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ is the vector of random intrinsic connection "probabilities" of each node. The θ_i's are assumed i.i.d. and independent of the class labels, and we impose $\mathbb{E}[\theta_i] = 1$, $\mathbb{E}[\theta_i^2] = \Phi$. The $1/n$ term bounds the degree of each node to an n-independent value, making the network sparse. Denoting $c = (c_{\rm in} + c_{\rm out})/2$, the detectability condition [16] reads:

$$\alpha \equiv \frac{c_{\rm in} - c_{\rm out}}{\sqrt{c}} \geq \frac{2}{\sqrt{\Phi}} \equiv \alpha_{\rm c}. \tag{2}$$

For $\alpha < \alpha_c$, no algorithm can partition the nodes better than by random guess. Letting $D = \mathrm{diag}(A\mathbb{1})$ be the degree matrix, the Bethe-Hessian is defined as

$$H_r = (r^2 - 1) I_n + D - rA, \quad r \in \mathbb{R}. \tag{3}$$

This matrix was originally proposed in [9] for $r = \sqrt{c\Phi}$, which asymptotically provides non trivial clustering down to the *detectability threshold* (for $\alpha > \alpha_c$). The informative eigenvector of H_r is associated with the second smallest eigenvalue and we denote it $\boldsymbol{x}_r^{(2)}$. The components of $\boldsymbol{x}_{\sqrt{c\Phi}}^{(2)}$ are however strongly tainted by the θ_i's, sensibly altering the algorithm performance.

We show here that for $\alpha \geq \alpha_c$ there exists a value $\zeta \leq \sqrt{c\Phi}$ for which the components of the second eigenvector $\boldsymbol{x}_\zeta^{(2)}$ of H_ζ align to the labels irrespective of the θ_i's, thus largely improving the algorithm performance while maintaining detectability down to the threshold.

2.2 Informative eigenvector of H_r

In the sequel we assume that: (i) being sparse, we can locally approximate the graph by a tree [21] and therefore $\mathbb{P}(\boldsymbol{\sigma}_{\partial_i}|\sigma_i) \simeq \prod_{j\in\partial_i} \mathbb{P}(\sigma_j|\sigma_i)$, with ∂_i the neighbourhood of i; (ii) $n \to \infty$ and c is bounded by an n-independent value while being arbitrarily larger than one, i.e., $n \gg c \gg 1$.

For ease of notation we work here with $D - rA$ rather than H_r, both having the same eigenvectors. The core of our proposed method lies in the following observation, related to the action of H_r on $\boldsymbol{\sigma}$:

$$[(D - rA)\boldsymbol{\sigma}]_i = d_i\sigma_i \left[1 - r\left(\frac{|\partial_i^{(s)}|}{d_i} - \frac{|\partial_i^{(o)}|}{d_i}\right)\right] \tag{4}$$

where $|\partial_i^{(s)}|$ (resp., $|\partial_i^{(o)}|$) stands for the number of neighbors of i belonging to the same (resp., opposite) class as i. We show next that a proper choice of r can annihilate the right-hand side of (4) "on average" or whenever the typical degrees d_i are not too small, turning (4) into an eigenvector equation. To this end, we need to quantify the random variables $|\partial_i^{(s)}|$ and $|\partial_i^{(o)}|$.

From a Bayesian perspective, $\boldsymbol{\sigma}$ and $\boldsymbol{\theta}$ are unknown parameters and A (and thus d_i) known realizations. We may thus write

$$\begin{aligned}\mathbb{P}(\sigma_i|\sigma_j, A_{ij} = 1) &= \frac{\mathbb{P}(\sigma_i, \sigma_j|A_{ij} = 1)}{\mathbb{P}(\sigma_j|A_{ij} = 1)} = 2\iint \mathbb{P}(\sigma_i, \sigma_j, \theta_i, \theta_j|A_{ij} = 1)d\theta_i d\theta_j \\ &\propto \iint \mathbb{P}(A_{ij} = 1|\sigma_i, \sigma_j, \theta_i, \theta_j)\mathbb{P}(\sigma_i, \sigma_j, \theta_i, \theta_j)d\theta_i d\theta_j \propto C(\sigma_i, \sigma_j),\end{aligned}$$

where we used the facts that the classes are of equal size ($\mathbb{P}(\sigma_i)$ is constant), and the θ_i are i.i.d., independent of the classes with $\mathbb{E}[\theta_i] = 1$. Normalizing, one finally obtains $\mathbb{P}(\sigma_i|\sigma_j, A_{ij} = 1) = C(\sigma_i, \sigma_j)/(c_{\rm in} + c_{\rm out})$, which is independent of the degree distribution. We further know that $|\partial_i^{(s)}| + |\partial_i^{(o)}| = d_i$, which is a deterministic observation. Given the locally tree-like structure of the graph, neighbors of the same node are conditionally independent – see (i) – so that $|\partial_i^{(s)}|$ is the sum of d_i i.i.d. Bernoulli random variables with parameter $p = c_{\rm in}/(c_{\rm in} + c_{\rm out})$. We thus obtain

$$\mathbb{E}[[(D - rA)\boldsymbol{\sigma}]_i \mid A] = d_i\sigma_i\left(1 - r\frac{c_{\rm in} - c_{\rm out}}{c_{\rm in} + c_{\rm out}}\right). \tag{5}$$

This equation suggests that, for the expectation of (4) to be an eigenvector equation in the large (but finite) d_i regime, r should be taken equal to

$$r = \frac{c_{\rm in} + c_{\rm out}}{c_{\rm in} - c_{\rm out}} = \frac{2\sqrt{c}}{\alpha} \equiv \zeta_\alpha. \tag{6}$$

with α as in (2) the proper control parameter for the clustering problem (as shown e.g., in [7, 15, 16, 22]). For simplicity of notation the dependence on α of $\zeta = \zeta_\alpha$ will be made explicit only when relevant. Intuitively, this calculus suggests that ζ is the only value of r that ensures that H_r has an informative eigenvector not significantly tainted by the degree distribution. This claim is supported by the following two remarks.

Remark 1 (Consistency of ζ for trivial classification)**.** *In the limit of trivial clustering where $c_{\rm out} \to 0$, $|\partial_i^{(s)}|$ and $|\partial_i^{(o)}|$ tend to their mean. In particular, for $c_{\rm out} = 0$, $\zeta = 1$ and $(D-\zeta A)\boldsymbol{\sigma} = (D-A)\boldsymbol{\sigma} = 0$, so that $\boldsymbol{\sigma}$ is an exact eigenvector of $H_{\zeta=1}$ associated with its zero eigenvalue.*

Remark 2 (Mapping to Ising)**.** *The original intuition behind the Bethe-Hessian matrix arises from a mapping of the community labels into the spins of a Ising Hamiltonian. The "temperature-related" parameter r guarantees a correct mapping* only *for $r = \zeta$. This is elaborated in details in Section A of the supplementary material.*

Although one commonly assumes an assortative model for the communities, by which $c_{\rm in} > c_{\rm out}$, the Bethe-Hessian matrix is oblivious of the sign of $c_{\rm in} - c_{\rm out}$.

Remark 3 (Disassortative networks)**.** *The case where $c_{\rm out} > c_{\rm in}$ does not invalidate the above analysis which results in $\zeta < 0$. Clustering with H_ζ is thus also valid in disassortative networks.*

In practice, for a given (non averaged) realization of the σ_i's, $\boldsymbol{\sigma}$ is not an exact eigenvector of H_ζ. By a perturbation analysis around $\boldsymbol{\sigma}$, we next analyze the behavior of the corresponding informative eigenvector of H_ζ and theoretically predict the overlap performance.

2.3 Performance Analysis

To generalize the averaged analysis of (5), we perturb $\boldsymbol{\sigma}$ by a "noise" term $\boldsymbol{\delta}$ and write $\boldsymbol{x}_\zeta^{(2)} \equiv \boldsymbol{\sigma} + \boldsymbol{\delta}$. Since ζ is however maintained, the associated eigenvalue of $D - \zeta A$, which is zero in (5), now possibly deviates from zero; this eigenvalue is denoted λ_α, i.e.,

$$(D - \zeta_\alpha A)(\boldsymbol{\sigma} + \boldsymbol{\delta}) = \lambda_\alpha(\boldsymbol{\sigma} + \boldsymbol{\delta}). \tag{7}$$

From Remark 1, we already know that $\lim_{\alpha \to \sqrt{2c_{\rm in}}} \lambda_\alpha = 0$.

In the following, expectations are taken for a fixed realization of the network, *i.e.* $\mathbb{E}[\cdot] \equiv \mathbb{E}[\cdot|A]$. Writing $|\partial_i^s| = \mathbb{E}[|\partial_i^s|] + \Delta_i$ and $|\partial_i^o| = \mathbb{E}[|\partial_i^o|] - \Delta_i$, where we exploited the relation $|\partial_i^s| + |\partial_i^o| = d_i$, we obtain:

$$[(D - \zeta_\alpha A)(\boldsymbol{\sigma} + \boldsymbol{\delta})]_i = -2\zeta_\alpha \sigma_i \Delta_i + d_i \delta_i - \zeta_\alpha \sum_{j \in \partial_i} \delta_j. \tag{8}$$

The random variable Δ_i is a sum of d_i independent (centered) Bernoulli random variables, tending in the large c limit to a zero mean Gaussian, i.e.,

$$\Delta_i \sim \mathcal{N}(0, d_i c_{\rm in} c_{\rm out}/(c_{\rm in} + c_{\rm out})^2) \equiv \mathcal{N}(0, d_i f_\alpha^2/\zeta_\alpha^2), \quad f_\alpha \equiv \frac{\sqrt{c_{\rm in} c_{\rm out}}}{c_{\rm in} - c_{\rm out}} = \frac{1}{\alpha}\sqrt{c - \frac{\alpha^2}{4}}. \tag{9}$$

Our analysis of (8) relies on the following claim that we shall justify next.

Assumption 1. *The random variables δ_i, $1 \le i \le n$, are distributed as $\delta_i \sim \mathcal{N}(-\mu_\alpha \sigma_i, f_\alpha^2 \beta_i^2)$ for some $\mu_\alpha \in \mathbb{R}$ depending on α only, and $\beta_i \in \mathbb{R}$ depending on i only. Besides, the δ_i's are "weakly dependent" in the sense that $\mathbb{E}[\delta_i \delta_j] = \mathbb{E}[\delta_i]\mathbb{E}[\delta_j] + O(1/c)$.*

The elements of Assumption 1 rely on the following observations:

- *Weak dependence*: This claim follows from the weak dependence of the Δ_i's, which results from the sparse (and thus locally tree-like) nature of the graph.
- *Gaussianity*: The right-hand side of (8) features 3 random variables, the leftmost being Gaussian and rightmost the sum of d_i variables tending to an (asymptotically independent) Gaussian. It is thus reasonable that δ_i be Gaussian (so to ensure (7)) yet not independent of Δ_i or $\sum_{j \in \partial_i} \delta_j$.
- *Mean of δ_i*: The symmetry of the problem at hand (equal class sizes, same affinity $c_{\rm in}$ for each class), along with the fact that the right-hand side of (4) vanishes in its first order approximation in d_i, suggest that the mean of δ_i does not depend in the first order on d_i but only on σ_i. The amplitude of the mean then depends on α characterized here through μ_α.
- *Variance of δ_i*: The variance appears as the product of two terms: one that depends on i (β_i) and one that depends on α. This follows from assuming that the fluctuations of δ_i follow the fluctuations of Δ_i for which the variance is similarly factorized.

Imposing the norm of the eigenvector $\boldsymbol{x}_\zeta^{(2)} = \boldsymbol{\sigma} + \boldsymbol{\delta}$ to be constant with respect to α and the boundary condition $\mu_{\alpha_c} = 1$ (i.e., there is no information about the classes at the detectability threshold), we find the following explicit expressions for μ_α and β_i.

$$1 - \mu_\alpha = \sqrt{\frac{c\Phi - \zeta_\alpha^2}{c\Phi - 1}}, \quad \beta_i = \frac{2}{\sqrt{d_i}}.$$

Details are provided in Section B of the supplementary material. Figure 1-(a) supports the analysis by comparing this prediction to simulations for a synthetic network with power law degree distribution.

The previous line of argument provides a large dimensional approximation for the performance of spectral clustering based on the eigenvector $\boldsymbol{x}_\zeta^{(2)}$. The performance measure of interest is the

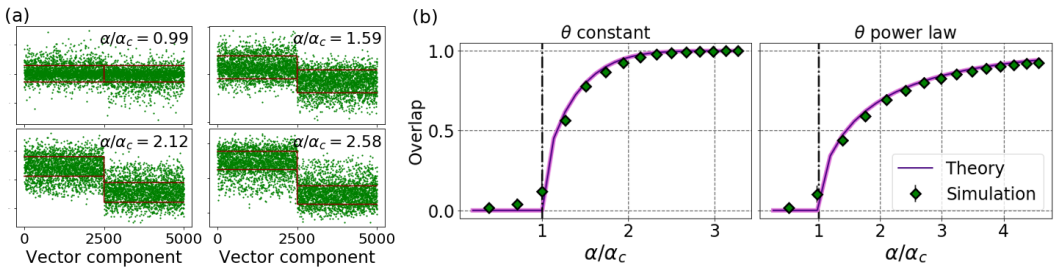

Figure 1: (a) Theoretical values of mean and variance (red line indicates $1 - \mu_\alpha \pm 2 f_\alpha / \sqrt{c}$) vs simulation (green dots) for power-law distributed θ_i's ($\theta_i \sim Z^{-1}[\mathcal{U}(3, 10)]^4$). (b) Theoretical (10) vs simulated overlap, averaged over 10 realizations, for θ_i constant (left), and power-law distributed (right). For both figures, $n = 5000$, $c_{\text{out}} = 6$, $c_{\text{in}} = 7 \to 36$.

overlap, defined as $\text{Ov} \equiv 2 \max_{\mathcal{P}_{\hat{\boldsymbol{\sigma}}}} \left[\frac{1}{n} \sum_{i=1}^n \delta_{\sigma_i, \hat{\sigma}_i} - \frac{1}{2} \right]$ where $\hat{\boldsymbol{\sigma}}$ denotes the vector of estimated labels, $\mathcal{P}_{\hat{\boldsymbol{\sigma}}}$ the set of permutations of the labels, and δ the Kronecker symbol ($\delta_{ij} = 1$ if $i = j$, and 0 otherwise). In this particularly symmetric setting *only* $\hat{\sigma}_i = \text{sign}[(\boldsymbol{x}_\zeta^{(2)})_i]$ where sign is the sign function. (Remark 5 underlines the necessity not to cluster based on sign in asymmetric scenarios). From the expression of μ_α and β_i, we find that, conditionally to A,

$$\mathbb{E}[\text{Ov}] \simeq \frac{1}{n} \sum_{i=1}^n \text{erf} \left[\sqrt{\frac{\alpha^2 d_i}{8c - 2\alpha^2} \left(\frac{c\Phi - \zeta_\alpha^2}{c\Phi - 1} \right)} \right] \tag{10}$$

(proof details are provided in Section B of the supplementary material). Figure 1-(b) compares the prediction of Equation (10) to simulations on networks with $\theta_i = 1$ constant (left) or power-law distributed (right). The observed match on this 5 000-node synthetic network is close to perfect.

As a side remark, our analysis reveals an interesting connection between H_ζ and $D^{-1}A$.

Remark 4 (Relation to the random walk Laplacian)**.** *Similar to A, $D - A$, and $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, the matrix $D^{-1}A$ is claimed* inappropriate *as a spectral community detection matrix for sparse graphs. This is in fact a slight overstatement: as already observed in [20], as the graph under study gets sparser, $D^{-1}A$ still possesses one or possibly more informative eigenvectors, however not necessarily corresponding to dominant isolated eigenvalues (it was in particular noted that for the real network* polblogs *[23] the informative eigenvector is associated to the third and not the second largest eigenvalue). This observation is easily explained in our analysis framework. Similar to our derivation for $D - \zeta A$, the average action of $D^{-1}A$ on the class vector $\boldsymbol{\sigma}$ reads $\mathbb{E}[[D^{-1}A\boldsymbol{\sigma}]_i | A] = \sigma_i / \zeta$ and thus, for large d_i, $\boldsymbol{\sigma}$ is a close eigenvector to $D^{-1}A$, correctly predicting the existence of an informative eigenvalue also for this matrix. However, the associated eigenvalue $1/\zeta$ decays with increasing ζ and thus with harder detection tasks, hence explaining why the informative eigenvectors are associated with eigenvalues found deeper into the spectrum of $D^{-1}A$.*

3 Estimating ζ

While $r = \zeta$ is more appropriate a choice than $r = \sqrt{c\Phi}$, ζ is not readily accessible (as it depends on $c_{\text{in}} - c_{\text{out}}$), unlike $\sqrt{c\Phi}$ that is easily estimated from the d_i's. To estimate ζ, we elaborate on the deep relations between the Bethe Hessian H_r and the non-backtracking operator $B \in \mathbb{R}^{2|\mathcal{E}| \times 2|\mathcal{E}|}$ defined, for all $(ij), (lm) \in \mathcal{E}_D$ the set of directed edges of $\mathcal{G}$, as $B_{(ij)(lm)} = \delta_{jl}(1 - \delta_{im})$.

When r is an eigenvalue of B, then $\det H_r = 0$ [11, 24]. This is convenient as B only has a few isolated real eigenvalues (B is non symmetric) that can send the associated isolated eigenvalues of H_r to zero. This provides us with two alternative methods to estimate ζ.

3.1 Exploiting the eigenvalues outside the bulk of B

It is proved in [15] that, for the DC-SBM and beyond the phase transition ($\alpha > \alpha_c$), the eigenvalues $\gamma_1, \dots, \gamma_{2m}$ of B, decreasingly sorted in modulus, satisfy in the large n setting: $\gamma_1 \to \Phi(c_{\text{in}} + c_{\text{out}})/2$, $\gamma_2 \to \Phi(c_{\text{in}} - c_{\text{out}})/2 > \sqrt{\gamma_1}$ and, for $i > 2$, $\limsup_n |\gamma_i| \leq \sqrt{\gamma_1}$, almost surely.

Since $\zeta = \lim_n \gamma_1/\gamma_2$, denoting $\nu_i(r)$ the eigenvalues of H_r sorted in *increasing* order, this result conveys the following first method to estimate ζ.

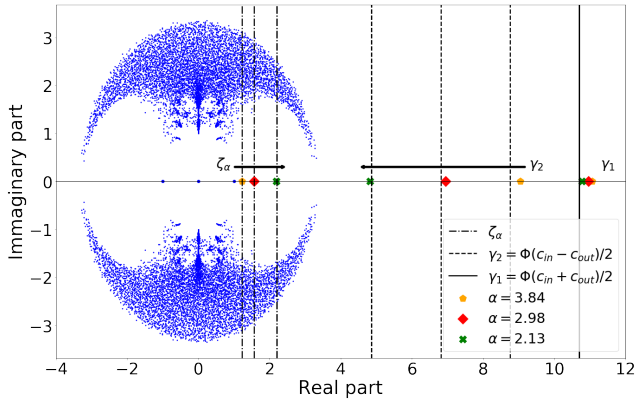

Figure 2: Superposed spectra of B for 3 values of α: $n = 4000$, $c_{\text{in}} = 12, 11, 10$ and $c_{\text{out}} = 1, 2, 3$ ($c_{\text{in}} + c_{\text{out}}$ is fixed); $\boldsymbol{\theta}$ with power law distribution; all eigenvalues displayed in blue except top three dominant real displayed in colors for each $(c_{\text{in}}, c_{\text{out}})$ pair.

Method 1 (First estimation of ζ)**.** *Under the previous notations $\zeta \simeq \gamma_1/\gamma_2$. The eigenvalues γ_1 and γ_2 of B can be estimated by a line search over $r \in (\sqrt{\rho(B)}, \infty)$ on changing signs of $\nu_1(r)$ and $\nu_2(r)$ that correspond to $r = \gamma_1$ and $r = \gamma_2$, respectively.*[1]

3.2 Exploiting the eigenvalues inside the bulk of B

The matrix B can be obtained from the linearization of the belief propagation (BP) equations (see [10] for details). In particular, the linear expansion to first order of the beliefs around their fixed points yields $B\boldsymbol{w} \simeq \zeta \boldsymbol{w}$. According to this argument, one expects the matrix B to have a real eigenvalue equal to ζ with[2] $\zeta \leq \sqrt{c\Phi}$. Figure 2 visually emphasizes this eigenvalue for three different values of α, maintaining c constant. The matrix B thus has four eigenvalues inside its main bulk: $-1, 0, 1$ and ζ. As the community detection problem gets harder, both ζ and γ_2 shift towards the edge of the bulk (from the left for the former and from the right for the latter) and then meet exactly at $\sqrt{c\Phi}$ when $\alpha = \alpha_c$. Then, for $\alpha < \alpha_c$, they reach the complex part of the bulk.

More fundamentally, simulations further suggest that the eigenvector associated with the null eigenvalue of H_ζ is precisely $\boldsymbol{x}_\zeta^{(2)} = \boldsymbol{\sigma} + \boldsymbol{\delta}$ studied in Section 2.3. This indicates that the informative eigenvalue λ_α of $D - \zeta_\alpha A = H_{\zeta_\alpha} - (\zeta_\alpha^2 - 1)I_n$ in Equation (7) coincides with $-(\zeta_\alpha^2 - 1)$. It further explains why $H_{\sqrt{c\Phi}}$, initially proposed in [9], works well close to the detectability threshold as $\zeta \to \sqrt{c\Phi}$ when $\alpha \to \alpha_c$. We thus expect most of the improvement of the choice $r = \zeta$ to emerge in the easier scenarios.

Note that, as was already observed in [9], if $|r| > 1$, then the eigenvalues of the bulk of H_r are strictly positive for $|r| \neq \sqrt{c\Phi}$. As a consequence, $\boldsymbol{x}_\zeta^{(2)}$ is necessarily isolated when $\alpha > \alpha_c$ and so spectral clustering on H_ζ works down to the detectability threshold. To the best of our knowledge, this property is not formally proved, but we point out that it agrees with the shape of the spectrum of B: if the bulk of H_r was negative for some $|r| > 1$, then there would be a 'continuum' of real eigenvalues in $[1, \sqrt{c\Phi}]$ if $r > 1$ (in the assortative case). As this is not the case, the smallest eigenvalue in the bulk of H_r cannot be negative.

Claim 1 (Informative eigenvalue of H_{ζ_α})**.** *The eigenvalue associated to the informative eigenvector of H_{ζ_α} is equal to zero. Equivalently, the eigenvalue λ_α associated to the informative eigenvector of $D - \zeta_\alpha A$ is given by $\lambda_\alpha = -(\zeta_\alpha^2 - 1) = -4f_\alpha^2$ which vanishes for $c_{\text{out}} \to 0$.*

This claim gives rise to a second method to estimate ζ.

[1] The spectral radius of the matrix B, $\rho(B)$, can be estimated as $\rho(B) \simeq \sum_i d_i^2 / \sum_i d_i$.

[2] This eigenvalue is visible in [10, 11] but not commented.

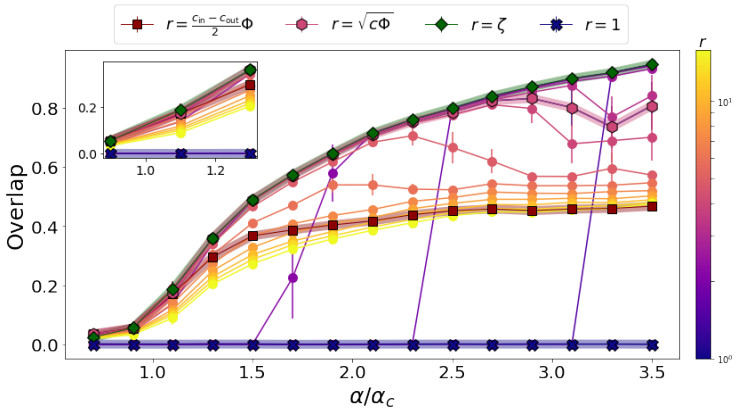

Figure 3: Overlap comparison as a function of α, using the second smallest eigenvector of H_r, for different values of r. In color code the values of r ranging from $r = 1$ (blue) to $r = c\Phi$ (yellow). The red squares indicate $r = (c_{\text{in}} - c_{\text{out}})\Phi/2$, that is equivalent to clustering with the matrix B [10], the purple hexagons represent the Bethe-Hessian of [9], the green diamonds are the proposed Algorithm 1 and the blue crosses are the graph Laplacian. In the top left corner, a zoom of the overlap close to the transition. For these simulations, $n = 5000$, $c_{\text{in}} : 15 \to 9.4$, $c_{\text{out}} : 1 \to 6.6$ (while keeping c fixed), $\theta_i \sim [\mathcal{U}(3, 10)]^4$.

Method 2 (Second estimation of ζ)**.** *Under the previous notations $\nu_2(\zeta) = 0$. The parameter ζ then corresponds to the position of change of sign of $\nu_2(r)$ in the set $r \in (1, \sqrt{\rho(B)})$.*

4 Extension to multiple uneven-sized classes

The analysis performed in the previous sections is resilient to heterogeneous degree distributions and can be generalized to k uneven-sized classes, with last clustering step by *k-means*. To this end, let $\Pi \in \mathbb{R}^{k\times k}$ be diagonal with Π_{ii} the fraction of nodes in class i and assume $C\Pi\mathbb{1} = c\mathbb{1}$. This assumption is a standard hypothesis [10, 22, 11, 25] which ensures that the averaged node connectivity is independent of the class. For $1 \le p \le k$, let $(\tau_p, \boldsymbol{v}^{(p)})$ be the p-th largest eigenpair of $C\Pi$, and $\boldsymbol{u}^{(p)} \in \mathbb{R}^n$ defined as $u_i^{(p)} = v_{\ell_i}^{(p)}$ $\forall\, 1 \le i \le n$ for ℓ_i the class of node i. The vector $\boldsymbol{u}^{(p)}$ contains plateaus with heights corresponding to the values of $\boldsymbol{v}^{(p)}$. Repeating the arguments of Section 2 (see details in Section C of the supplementary material), we obtain k choices for r:

$$\mathbb{E}[[(D - rA)\boldsymbol{u}^{(p)}]_i] = d_i u_i^{(p)} \left[1 - r\frac{\tau_p}{c}\right] \qquad \text{and thus} \quad r = \frac{c}{\tau_p} \equiv \zeta_p, \quad 1 \le p \le k. \tag{11}$$

Since the largest eigenpair $(c, \mathbb{1})$ of $C\Pi$ is not informative of the class structure, only the $k-1$ next largest eigenvectors $\boldsymbol{v}^{(p)}$ of $C\Pi$ are informative. The vector $\boldsymbol{u}^{(p)}$ (corresponding to the p-th largest eigenvalue τ_p) is in one-to-one mapping with $\boldsymbol{v}^{(p)}$ and corresponds to the p-th smallest value of $\zeta_p = c/\tau_p$. Considering $r = \sqrt{c\Phi}$, all the informative eigenvalues of H_r are negative [9]. By decreasing r they progressively become positive: for $r = \zeta_k$ (the largest among ζ_p) the k-th smallest eigenvalue is the first to hit zero. By further decreasing r, all the informative eigenvalues follow, until $r = \zeta_1 = 1$ for which the smallest eigenvalue is null. We conclude that $\boldsymbol{u}^{(p)}$ is associated with the p-th smallest eigenvector $\boldsymbol{x}_{\zeta_p}^{(p)}$ of H_{ζ_p}.

Method 1 and Method 2 both generalize to this scenario. In particular the outer eigenvalues of B converge as $\gamma_p \to \tau_p\Phi$ and the linearization of BP retrieves the fixed points as $\zeta_p = c/\tau_p$.

For $k > 2$, the value $r = \sqrt{c\Phi}$ still plays an important role. It was chosen in [9] because, asymptotically, for this value of r *only* the informative eigenvalues of $H_{\sqrt{c\Phi}}$ are negative. The number of classes is then directly obtained from counting the number of negative eigenvalues of $H_{\sqrt{c\Phi}}$. The relation between H_r and B further guarantees that the number of isolated eigenvalues of B (hence of H_r) is asymptotically equal to the number of detectable classes.

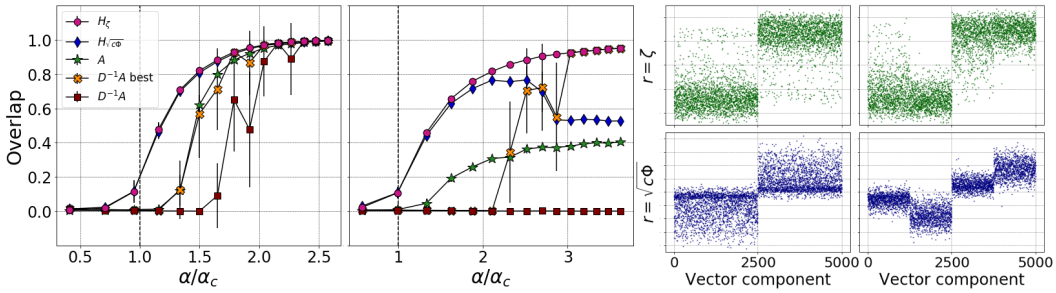

Figure 4: (a) Comparison of spectral clustering for $\theta_i = 1$ (left) and with power law distribution $\theta_i \sim Z^{-1}[\mathcal{U}(3,10)]^4$. "$D^{-1}A$ best" indicates spectral clustering on the best (among the first 25) eigenvector of $D^{-1}A$. Here, $n = 5000$, $c_{\text{out}} = 1$, $c_{\text{in}} = 2 \to 16$. Averaged over 10 samples. The error bars indicate one standard deviation. (b) $\boldsymbol{x}_\zeta^{(2)}$ (top) and $\boldsymbol{x}_{\sqrt{c\Phi}}^{(2)}$ (bottom) for power law distributed θ_i (left) and for $\theta_i = \theta_0$, $i \leq n/4$ and $n/2 \leq i \leq 3n/4$, and $\theta_i = 4\theta_0$ otherwise (right).

Remark 5 (On k-means versus signed-based clustering)**.** *Under a symmetric 2-class of even size setting, the classification of the entries of the informative eigenvector of H_r can be performed based on their signs. This sign-based method first does not generalize to more than two or uneven sized classes, where k-means or expectation-maximization based clustering is required. But it also hinders the fact that the eigenvector entries may be quite concentrated around zero (close to 0^+ or 0^- according to the class) and thus* not clustered*, a situation where k-means has no discriminative power.*

Simulations (and reported results in [9] based on signs rather than k-means) suggest that the informative eigenvector of $H_{\sqrt{c\Phi}}$ precisely suffers this condition. We have demonstrated here instead that the informative eigenvector of H_ζ has the convenient feature of being genuinely clustered.

5 Experimental validation

Our results can be summarized by Algorithm 1, where we recall that $\nu_p(r)$ is the p-th smallest eigenvalue of H_r and where $\boldsymbol{x}_r^{(p)}$ indicates the corresponding eigenvector.

Figure 3 depicts the overlap, as a function of α, of the output of a two-class k-means on the informative eigenvector of H_r, for different values of r, ranging from 1 to $c\Phi$. When α is large enough, small values of r lead to better partitions than large values of r that are more affected by degree heterogeneity. However, for r small, the informative eigenvector is not necessarily corresponding to the second smallest eigenvalue, leading to a meaningless partition. On the contrary, larger values of r show isolated eigenvectors also in the "hard regime". We recall that $r = \zeta$ is an α-dependent parameter: for $\alpha \to \alpha_c$, ζ is "large enough" so that the informative eigenvalue is isolated, while for $\alpha \to \sqrt{2c_{\text{in}}}$ it is "small enough" to give good partitions. Also the value of $r = (c_{\text{in}} - c_{\text{out}})\Phi/2$ is α-dependent and it corresponds to clustering with B as indicated in [10]. While it gives good partitions very close to the transition, this choice of r seems largely sub-optimal for easier tasks.

Figure 4-(a) compares the overlaps obtained with Algorithm 1 versus related spectral clustering methods based on $H_{\sqrt{c\Phi}}$, $D^{-1}A$ and A. Accordingly with Remark 5, *k-means* clustering (rather than sign-based) on the informative eigenvectors is systematically performed. For $\theta_i = 1$, the left display recovers the results of [9], evidencing a strong advantage for H_r versus Laplacian methods. Since the degrees are similar, both $r = \sqrt{c\Phi}$ and $r = \zeta$ induce similar H_r performances. The improvement provided by H_ζ arises in the right display for power-law distributed θ_i, with most of the gain appearing away from the detection threshold. On both displays is also depicted the performance of $D^{-1}A$ based on its second largest eigenvector and on an oracle choice of the informative eigenvector with maximal overlap. These curves confirm Remark 4 on the non-dominant position of the informative eigenvector of $D^{-1}A$ in hard tasks.[3] Figure 4-(b) depicts the informative eigenvectors of $H_{\sqrt{c\Phi}}$ and H_ζ, demonstrating the negative impact of θ_i on $H_{\sqrt{c\Phi}}$, in stark contrast with the resilience of H_ζ.

[3]The low performance of $D^{-1}A$, even in an oracle setting, can be attributed to the high density of eigenvalues in the bulk of the spectrum which induces a "dispersion" of the informative eigenvectors to the eigenvectors associated to neighboring eigenvalues. The class information is thus "spread" across several eigenvectors.

Algorithm 1 Improved Bethe-Hessian Community Detection

Input : adjacency matrix of undirected graph $\mathcal{G}$
Detect the number of classes: $\hat{k} \leftarrow |\{i,\ \nu_i(\sqrt{c\Phi}) < 0\}|$.
for $2 \leq p \leq \hat{k}$ **do**
    $\zeta_p \leftarrow r$ such that $\nu_p(r) = 0$
    $X_p \leftarrow \boldsymbol{x}^{(p)}_{\zeta_p}$
Estimate community labels $\hat{\boldsymbol{\ell}}$ as output of $\hat{k}$-class *k-means* on the rows of $X = [X_2, \ldots, X_{\hat{k}}]$.
return Estimated number $\hat{k}$ of communities and label vector $\hat{\boldsymbol{\ell}}$.

Table 1 next provides a comparison of the algorithm performances on real networks, both labelled and unlabelled, confirming the overall superiority of Algorithm 1, quite unlike $H_{\sqrt{c\Phi}}$ which fails on several examples.[4]

L	n	k	Alg.1	$H_{\sqrt{c\Phi}}$	A	U	n	k	Alg.1	$H_{\sqrt{c\Phi}}$	A
Karate [28]	34	2	**1.00**	**1.00**	**1.00**	Mail	1133	21	**0.50**	0.40	0.32
Dolphins [29]	62	2	**0.97**	0.87	0.65	Facebook	4039	65	**0.77**	0.48	0.38
Polbooks [30]	105	3	**0.77**	0.74	0.57	Power grid	4941	53	**0.92**	0.61	0.31
Football [31]	115	12	**0.92**	**0.92**	**0.92**	Nutella	6301	5	**0.34**	0.15	0.14
Polblogs [23]	1221	2	**0.91**	0.32	0.26	Wikipedia	7115	21	**0.21**	0.18	0.15

Table 1: Performance comparison on real networks. Labelled datasets with k known and overlap comparison: (left). Unlabelled networks [32] with k estimated and modularity comparison. Only assortative features are kept into account.

6 Concluding Remarks

Beyond the demonstration of superiority of H_ζ to $H_{\sqrt{c\Phi}}$, originally proposed in [9], the article provides a consistent understanding of the natural limitations and strengths of the wide class of spectral clustering methods involving combinations of A and D.

Yet, other methods, the performances of which cannot always be compared on even grounds, have been proposed in the literature that marginally relate to the present study. This is notably the case of [18] which performs spectral clustering on $L_\tau = (D + \tau I_n)^{-\frac{1}{2}} A (D + \tau I_n)^{-\frac{1}{2}}$ (with a proposed choice $\tau = c$) which aims at neutralizing the deleterious effects of small d_i. Although evidently affecting the spectrum (and thus the informative structure) of A by the non-linear normalization, simulations on L_τ suggest competitive performances to H_ζ in almost all studied examples. A systematic analysis of this and similarly proposed methods in the literature is clearly called for.

Despite its demonstrated significant performance improvement, Algorithm 1 suffers from a slightly larger computational cost than most competing methods ($O(nk^3)$ instead of the usual $O(nk^2)$ complexity in the case of sparse graph) due to the successive estimations of ζ. We are currently working on improving this computation time.

From a theoretical standpoint, the request for $c \gg 1$ is still inappropriate to many practical networks. A first consequence of smaller values for c is the loss of Gaussianity of the eigenvector entries as already evidenced in Figures 1 and 4 where Gaussianity is clearly lost in the easiest tasks in profit of a "one-sided" distribution. This suggests further improvement of our analysis framework along with the development of algorithms more appropriate than k-means to handle the last clustering step.

Acknowledgments

This work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006), the IDEX GSTATS Chair at University Grenoble Alpes and by CNRS PEPS I3A (Project RW4SPEC). The authors thank Jean-Louis Barrat for fruitful discussions.

[4] In Table 1, the modularity is defined as $\mathcal{M} = \frac{1}{2|\mathcal{E}|} \sum_{i,j=1}^n \left(A_{ij} - \frac{d_i d_j}{2|\mathcal{E}|} \right) \delta(\hat{\ell}_i, \hat{\ell}_j)$, see e.g., [26, 27].

References

[1] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[2] Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Conference on Learning Theory*, pages 356–370, 2014.

[3] Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[4] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[5] Lennart Gulikers, Marc Lelarge, and Laurent Massoulié. A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Advances in Applied Probability*, 49(3):686–721, 2017.

[6] J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[7] Raj Rao Nadakuditi and Mark EJ Newman. Graph spectra and the detectability of community structure in networks. *Physical review letters*, 108(18):188701, 2012.

[8] Hafiz Tiomoko Ali and Romain Couillet. Random matrix improved community detection in heterogeneous networks. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 1385–1389. IEEE, 2016.

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

[10] 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*, 110(52):20935–20940, 2013.

[11] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.

[12] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

[13] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[15] Lennart Gulikers, Marc Lelarge, and Laurent Massoulié. Non-Backtracking Spectrum of Degree-Corrected Stochastic Block Models. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 44:1–44:27, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[16] Lennart Gulikers, Marc Lelarge, Laurent Massoulié, et al. An impossibility result for reconstruction in the degree-corrected stochastic block model. *The Annals of Applied Probability*, 28(5):3002–3027, 2018.

[17] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[18] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

[19] Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

[20] Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *arXiv preprint arXiv:1312.1733*, 2013.

[21] Amir Dembo, Andrea Montanari, et al. Gibbs measures and phase transitions on sparse random graphs. *Brazilian Journal of Probability and Statistics*, 24(2):137–211, 2010.

[22] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[23] 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*, pages 36–43. ACM, 2005.

[24] Audrey Terras. *Zeta functions of graphs: a stroll through the garden*, volume 128. Cambridge University Press, 2010.

[25] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[26] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[27] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103:8577–8582, 2006.

[28] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[29] 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*, 54(4):396–405, 2003.

[30] http://www.orgnet.com/.

[31] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[32] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

Supplementary material

A Mapping to Ising

As introduced in [22] for the SBM ($\boldsymbol{\theta} = \mathbb{1}_n$), the probability to realize a graph under the sparse DC-SBM hypothesis reads:

$$\mathbb{P}(A|\boldsymbol{\sigma},\boldsymbol{\theta}) = \prod_{i,j<i}\left(\theta_i\theta_j\frac{C_{\sigma_i,\sigma_j}}{n}\right)^{A_{ij}}\left(1-\theta_i\theta_j\frac{C_{\sigma_i,\sigma_j}}{n}\right)^{1-A_{ij}} = \prod_{i,j<i}\left(\theta_i\theta_j\frac{C_{\sigma_i,\sigma_j}}{n}\right)^{A_{ij}} + o\left(\frac{1}{n}\right)$$
$$= \prod_{(ij)\in\mathcal{E}}\theta_i\theta_j\frac{C_{\sigma_i,\sigma_j}}{n} + o\left(\frac{1}{n}\right).$$

By making use of the Bayes theorem we can map the probability distribution of the labels to a physical analogue of spins interacting on the graph.

$$\mathbb{P}(\boldsymbol{\sigma}|A) = \int d\boldsymbol{\theta}\,\mathbb{P}(\boldsymbol{\sigma},\boldsymbol{\theta}|A) = \int d\boldsymbol{\theta}\,\mathbb{P}(A|\boldsymbol{\sigma},\boldsymbol{\theta})\frac{\mathbb{P}(\boldsymbol{\sigma})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(A)}$$
$$\underset{n\to\infty}{\sim}\frac{1}{\mathbb{P}(A)2^n}\prod_{(ij)\in\mathcal{E}}\frac{C_{\sigma_i,\sigma_j}}{n}\int d\boldsymbol{\theta}\,\mathbb{P}(\boldsymbol{\theta})\theta_i\theta_j = \frac{1}{Z}\prod_{(ij)\in\mathcal{E}}C_{\sigma_i,\sigma_j} = \frac{1}{Z}e^{-\tilde{\mathcal{H}}(\boldsymbol{\sigma})}$$

where we recovered the Boltzmann distribution with dimensionless Hamiltonian given by

$$\tilde{\mathcal{H}}(\boldsymbol{\sigma}) = -\sum_{(ij)\in\mathcal{E}}\log\left[C_{\sigma_i,\sigma_j}\right] \equiv -\sum_{(ij)\in\mathcal{E}}\mathrm{ath}\left(\frac{1}{r}\right)\sigma_i\sigma_j + const \tag{A.1}$$

where $const$ is a constant that will be absorbed in the normalization factor. This last step gives rise to an Ising Hamiltonian. The following system of equations must then hold for some r:

$$\log[c_{\mathrm{in}}] = \mathrm{ath}\left(\frac{1}{r}\right) + const. \tag{A.2a}$$

$$\log[c_{\mathrm{out}}] = -\mathrm{ath}\left(\frac{1}{r}\right) + const. \tag{A.2b}$$

It is easy to check that $r = \zeta$ is the solution to this system of equations. From this result, one can then follow the derivation of the Bethe-Hessian matrix proposed in [9].

It has to be remarked that to obtain Equation (A.1) we neglected terms coming from non-nearest neighbours, in the limit for $n \to \infty$. The mapping is therefore not exact, but it still constitutes a useful tool to analyze and understand the problem.

Further note that, for disassortative networks, $c_{\mathrm{in}} < c_{\mathrm{out}}$ and thus $\zeta < 0$ as commented in Remark 3 in the main article. This would correspond to an anti-ferromagnetic interaction between the spins, in complete agreement with the mapping provided.

B Mean and variance of the eigenvector

We need to identify the terms β_i and μ_α introduced in Assumption 1 to track the behavior of $\boldsymbol{\delta}$ and thus of the eigenvector $\boldsymbol{\sigma} + \boldsymbol{\delta}$. A first constraint on $\boldsymbol{\delta}$ follows from imposing the normalization of the eigenvector which, in the trivial limit equals $\boldsymbol{\sigma}$, the norm of which is $\sqrt{n}$. As such,

$$\|(1-\mu_\alpha)\boldsymbol{\sigma} + f_\alpha\boldsymbol{\beta}\odot\boldsymbol{N}\|^2 = n \tag{B.1}$$

where $\boldsymbol{\beta} = (\beta_i)_{i=1}^n$, and $\boldsymbol{N}$ is a vector of zero mean and unit variance Gaussian random variables. Denoting $n\tilde{\beta}^2 \equiv \|\boldsymbol{\beta}\odot\boldsymbol{N}\|^2$ and observing that $\tilde{\beta} = O(\beta_i)$ – *i.e.* they have the same scaling with respect to c –, we can rewrite this equation under the form:

$$(1-\mu_\alpha)^2 + f_\alpha^2\tilde{\beta}^2 = 1. \tag{B.2}$$

This provides a first relation between μ_α and $\tilde{\beta}$. To obtain our next equations, we now explore boundary conditions on the model parameters in the limit of trivial clustering and at the phase transition where clustering becomes impossible.

It is established in [16] that there exists a critical value $\alpha_c \equiv 2/\sqrt{\Phi}$ for α below which community detection is (asymptotically) impossible. In particular, for $\alpha = \alpha_c$, the eigenvector $\boldsymbol{\sigma} + \boldsymbol{\delta}$ does not contain any information about the classes and thus $\mu_{\alpha_c} = 1$. From Equation (9), we then find that $f_{\alpha_c} = \sqrt{c\Phi - 1}/2$. Also, from (B.2), we get $\tilde{\beta} = 1/f_{\alpha_c}$. Updating (B.2), we now have an explicit expression for μ_α for all α. Recalling that $4f_\alpha^2 = \zeta_\alpha^2 - 1$ (from (6) and (9)) then gives

$$1 - \mu_\alpha = \sqrt{\frac{c\Phi - \zeta_\alpha^2}{c\Phi - 1}}. \tag{B.3}$$

Getting back to (7) and (8), it now remains to estimate β_i, which we shall perform in the limit $\alpha \to \sqrt{2c_{\rm in}}$ of trivial clustering. To this end, combining both equations, we have

$$2f_\alpha(1-\mu_\alpha)\sqrt{d_i}\tilde{N}_i - \zeta_\alpha \sum_{j\in\partial_i} f_\alpha\beta_j N_j + d_i f_\alpha \beta_i N_i = \lambda_\alpha[(1-\mu_\alpha)\sigma_i + f_\alpha\beta_i N_i]$$

for $\tilde{N}_1, N_1, \ldots, \tilde{N}_n, N_n$ all (non necessarily independent) standard normal random variables. The second left-hand side term is proportional to $\sqrt{d_i}$ (and thus of order $O(\sqrt{c})$) as per the weak independence assumption of the N_k's (Assumption 1). Dividing both sides by $f_\alpha\sqrt{d_i}$ to equate terms of order $O(1)$, the right-hand side now scales as $\lambda_\alpha/(f_\alpha\sqrt{d_i})$. As noted in Remark 1, in the trivial clustering limit where $\alpha \to \sqrt{2c_{\rm in}}$, $\lambda_\alpha \to 0$, but it is not clear whether the right-hand side (after division by $f_\alpha\sqrt{d_i}$) vanishes; we now investigate this term in detail. One may at first observe that, if $c_{\rm out} = \epsilon c_{\rm in}$ for $\epsilon \ll 1$, since c typically scales like d_i, we obtain that $f_\alpha\sqrt{d_i} = \sqrt{\epsilon c_{\rm in}/2} + O(\epsilon)$. Hence, if $c_{\rm in} \gtrapprox \epsilon^{-1}$, the right-hand side vanishes. But imposing this growth condition is in fact not even necessary. If $\lambda_\alpha \propto f_\alpha^\eta$ for some $\eta > 1$, we directly obtain a vanishing right-hand side term; in Section 3 we argued that $\eta = 2$ (see Claim 1).
Denoting $\sum_{j\in\partial_i}\beta_j N_j \equiv \langle\beta\rangle N\sqrt{d_i}$ for some $\langle\beta\rangle > 0$, we may then rewrite

$$2(1-\mu_\alpha)\tilde{N}_i - \zeta_\alpha\langle\beta\rangle N + \sqrt{d_i}\beta_i N_i \to 0 \tag{B.4}$$

in the limit $\alpha \to \sqrt{2c_{\rm in}}$. Besides, $\mu_\alpha \to 0$ while $\zeta_\alpha \to 1$. We already argued that β_i (and thus $\langle\beta\rangle$), which is of the order of $\tilde{\beta}$, scales as $1/f_{\alpha_c} = O(c^{-1/2})$. Thus, in the limit of large degrees, the second term in (B.4) is negligible and the third of order $O(1)$. Equating the large degree limiting variances of the resulting equation finally gives

$$\beta_i = \frac{2}{\sqrt{d_i}}.$$

We now have the mean and the variance of each vector component and we can estimate the expression of the overlap. Considering a node with $\sigma_i = 1$ without loss of generality, in the large c limit, we have the approximate classification error for node i:

$$\mathbb{P}_{\rm err}^i \simeq \frac{1}{\sqrt{2\pi[f_\alpha\beta_i]^2}}\int_{(1-\mu_\alpha)}^{\infty} e^{-x^2/(2[f_\alpha\beta_i]^2)}dx = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{1}{\sqrt{2}[f_\alpha\beta_i]}(1-\mu_\alpha)\right)\right].$$

From this, the expression of the overlap follows.

C Extension to more than two classes

In order to generalize the argument carried on for two classes, first we look into the following quantity

$$\begin{aligned}
\mathbb{P}(\ell_i|\ell_j, A_{ij}=1) &= \frac{\mathbb{P}(\ell_i,\ell_j|A_{ij}=1)}{\mathbb{P}(\ell_j|A_{ij}=1)} = \frac{\iint d\theta_i d\theta_j \mathbb{P}(\ell_i,\ell_j,\theta_i\theta_j|A_{ij}=1)}{\mathbb{P}(\ell_j)} \\
&= \frac{\iint d\theta_i d\theta_j \mathbb{P}(A_{ij}=1|\theta_i,\theta_j,\ell_i,\ell_j)\mathbb{P}(\ell_i)\mathbb{P}(\ell_j)\mathbb{P}(\theta_i)\mathbb{P}(\theta_j)}{Z\pi_{\ell_j}} \\
&= \frac{\pi_{\ell_i}C_{\ell_i,\ell_j}}{c} = \frac{(\Pi C)_{\ell_i,\ell_j}}{c} = \frac{(C\Pi)_{\ell_j,\ell_i}}{c}
\end{aligned}$$

By repeating the same argument on the average behavior of the adjacency matrix we obtain:

$$\begin{aligned}
\langle(A\boldsymbol{u}^{(p)})_i\rangle &= \sum_{j\in\partial(i)}\langle u_j^{(p)}\rangle = \sum_{j\in\partial(i)}\langle v_{\ell_j}^{(p)}\rangle = d_i\sum_{\ell_j}\mathbb{P}(\ell_j|\ell_i,A_{ij}=1)v_{\ell_j}^{(p)} \\
&= \frac{d_i}{c}\sum_{\ell_j}(C\Pi)_{\ell_i,\ell_j}v_{\ell_j}^{(p)} = \frac{d_i}{c}(C\Pi v^{(p)})_{\ell_i} = \frac{d_i}{c}\tau_p v_{\ell_i}^{(p)} = d_i\frac{\tau_p}{c}u_i^{(p)}
\end{aligned}$$

from which the result unfolds. In the simulations on synthetic networks, the off-diagonal terms of the matrix C are drawn from a uniform distribution $\mathcal{U}(c_{\text{out}} - f, c_{\text{out}} + f)$, the element C_{11} is fixed to c_{in} and all the other diagonal terms are determined to ensure $C\Pi\mathbb{1}_k = c\mathbb{1}_k$. The randomness will make the eigenvalues of $C\Pi$ non degenerate and there will not be a unique transition. The line $c_{\text{in}} - c_{\text{out}} = k\sqrt{c}$ indicates the approximated position of the transition.

In Figure 5 we report the spectrum of B in the case of four classes, that shows that the largest isolated real eigenvalues of the matrix B are τ_p for $1 \leq p \leq k$, followed by c/τ_p for $2 \leq p \leq k$. This result can be obtained analytically from the linearization of the belief propagation equations (see [10]).

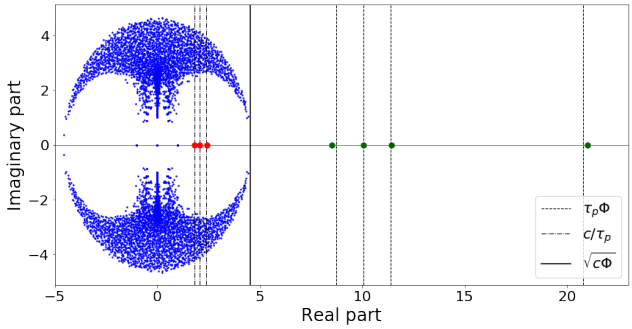

Figure 5: Spectrum of B. In green the isolated real eigenvalues outside the bulk corresponding to $\{\tau_p\Phi\}$, in red those inside the bulk, corresponding to $\{\zeta_p = c/\tau_p\}$; in blue all the others. We used 4 clusters of equal size, $n = 5000$, $c_{\text{in}} = 20$, $c_{\text{out}} = 5$, $f = 1.5$ and $\theta_i \sim \theta = \mathcal{U}(3, 13)^4$.

Figure 6(a) displays the overlap as a function of the hardness of the problem and of the number of classes comparing our algorithm with [9], evidencing a strong advantage in terms of performance for our algorithm. The red square underlines the fact that the two methods coincide *only* at the transition when $k = 2$ and the latter algorithm pays a lot in terms of performance for $k > 2$, even close to the transition. Figure 6(b) shows how $\hat{k} = |\{p,\ v_p(\sqrt{c\Phi}) < 0\}|$ is a good estimator of the number of classes. With $k_d = |\{p,\ \tau_p > \sqrt{c/\Phi}\}|$ we denote the number of theoretically detectable clusters and plot the quantity $2(\hat{k} - k_d)/(\hat{k} + k_d)$, showing small disagreement only close to the transition. The recovery being asymptotically exact, this can be interpreted as a finite size effect.

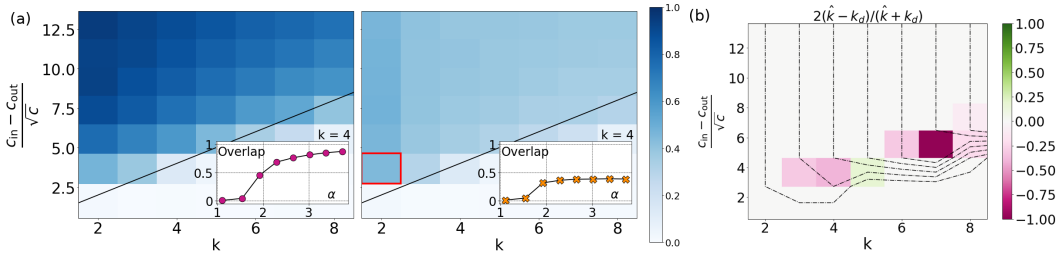

Figure 6: (a) Overlap (color scale) as a function of the number of classes (k) and hardness of the problem for the proposed algorithm (left) and $H_{\sqrt{c\Phi}}$ (right). Here, $n = 10\,000$, $c_{\text{in}} = 4 \to 40$, $c_{\text{out}} = 3$, $f = 2/k$, $\theta_i \sim [\mathcal{U}(3, 13)]^4$. Averaged over 10 samples.
(b) Recovery $(2(\hat{k} - k_d)/(\hat{k} + k_d))$ as a function of k and the hardness of the problem for the same parameters as (a).