COMMUNITY DETECTION IN SPARSE REALISTIC GRAPHS: IMPROVING THE BETHE HESSIAN

Lorenzo Dall'Amico[1], *Romain Couillet*[1,2]

[1]GIPSA-lab, University of Grenoble–Alpes, [2]CentraleSupélec, University of Paris–Saclay

ABSTRACT

This article improves over the recently proposed Bethe Hessian matrix for community detection on sparse graphs, assuming here a more realistic setting where node degrees are inhomogeneous. We notably show that the parametrization proposed in the seminal work on the Bethe Hessian clustering can be ameliorated with positive consequences on correct classification rates. Extensive simulations support our claims.

Index Terms— community detection; Bethe Hessian; spectral clustering; statistical physics.

1. INTRODUCTION

Community detection on graphs [1, 2] is a cornerstone topic in machine learning, much related to unsupervised classification (or clustering) [3], and consists in grouping nodes of strong affinity in distinct classes. Theoretically speaking, given a statistical generative model for a graph $\mathcal{G}$ with classes, the first question to consider is their detectability and the capability to associate each node to its genuine class.

The most popular and versatile approach to perform community detection on graphs is the belief propagation algorithm; however, the latter is computationally expensive, offers no convergence guarantee and is theoretically hard to analyze. Most convincing (since well performing, theoretically analyzable and computationally appealing) among the proposed alternative approaches to community retrieval are *spectral methods* that consist in reading the community classes directly off the dominant eigenvectors of a matrix representation of $\mathcal{G}$, thereby reminiscent of spectral clustering [3]. Assuming a two-class *stochastic block model* (SBM) for the generative graph model with n nodes – where the probability for node i to connect to node j equals $p_{\rm in} \in [0,1]$ if they belong to the same class or $p_{\rm out} \in [0,1]$ otherwise, and every edge is drawn independently – a natural spectral community detection method consists in extracting the class information from the dominant eigenvectors of the adjacency matrix $A \in \{0,1\}^{n\times n}$, where $A_{ij} = 1$ if nodes i and j are connected, and $A_{ij} = 0$ otherwise.

Couillet's work is supported by the GSTATS UGA IDEX DataScience chair and the ANR RMT4GRAPH Project (ANR-14-CE28-0006).

It was indeed shown that, as $n \to \infty$ and $p_{\rm in}, p_{\rm out}$ are independent of n, which is referred to as a *dense graph* community detection problem, spectral clustering on A is "optimal" in the sense that:

(a) there exists a minimal value for $(p_{\rm in}-p_{\rm out})/\sqrt{p_{\rm in}+p_{\rm out}}$ below which community detection is infeasible;

(b) spectral clustering on A returns non-trivial classification (that is on average better than random guess) as soon as this threshold is exceeded.

In statistical physics terms, this asymptotic decidability thresholding effect is referred to as a *phase transition phenomenon*.

Yet, the conditions under which spectral clustering on A is optimal rely on two key ingredients:

(i) the statistical block model for $\mathcal{G}$ is quite elementary;

(ii) the graph is dense (the node degrees scale with n).

Both conditions are deemed *unrealistic* as not representative of real world graphs. To address issue (i), a line of works was developed [4, 5] in a K-class *degree corrected* stochastic block model (DC-SBM), where

$$P(A_{ij} = 1) = q_i q_j C(x_i, x_j)$$

with $q_i > 0$ some *intrinsic* connectivity amplitude for node i, such that $\mathbb{E}[q_i] = 1$, $x_i \in \{1, \ldots, K\}$ the label of the class of node i, and $C(x_i, x_j)$ some class-wise affinity parameter. In [5], it is shown that spectral clustering on A is no longer optimal in that the phase transition phenomenon in general arises well below the detectability power of A; an improvement is then proposed in [5] which shows that there exists $\alpha > 0$ depending on the law of the q_i's such that performing spectral clustering on $D^{-\alpha}AD^{-\alpha}$ rather than A drastically pushes the phase transition to smaller discriminative values of $C(a,b)$.

Addressing limitation (ii) is theoretically much harder. Assuming that the probability for $A_{ij} = 1$ scales like $1/n$, i.e., the average nodal degree is of order $O(1)$ with respect to n, it has long been seen in simulations that spectral clustering on A is largely suboptimal. Via the impulse of *statistical physics* tricks, mostly consisting in either approximating (linearizing) belief propagation or mapping the community detection problem into an Ising model analog, new spectral clustering algorithms were proposed that are shown in

practice (and sometimes in theory [6]) to dramatically improve over spectral clustering on A; this is notably the case of spectral clustering on the *non-backtracking operator* B [7] and on the (closely related) *Bethe Hessian matrix* H_r [8] parametrized by $r \in \mathbb{R}$.

In this article we focus on spectral clustering performed over the matrix H_r, and propose an improved choice $r = r_c$ for the parameter r, which differs from the parameter initially suggested in [8]. This choice results in two important improvements: (i) our algorithm is not sign-based (unlike in [8]) and is able to make a clear distinction between classes; (ii) H_{r_c} is the only H_r matrix resilient to degree heterogeneity, hence more suited to applications to real networks [9]. In the next section we formally introduce the interrelated matrices B and H_r.

2. MODEL AND MAIN RESULTS

2.1. Preliminaries

Consider a 2-class symmetric n-node graph $\mathcal{G} = \mathcal{G}(\mathcal{E}, \mathcal{V})$ (with $\mathcal{E}$ the set of edges, $|\mathcal{E}| = m$, and $\mathcal{V}$ the set of nodes, $|\mathcal{V}| = n$) generated from a sparse DC-SBM model, i.e., with adjacency matrix $A \in \{0,1\}^{n\times n}$ defined by

$$P(A_{ij} = 1) = q_i q_j \frac{C(x_i, x_j)}{n} \tag{1}$$

for $q_1, \ldots, q_n > 0$ random and independently drawn with $\mathrm{E}[q_i] = 1$, $x_i \in \{-1, 1\}$ the class label of node i, and $C(x_i, x_j) = c_{\rm in} > 0$ if $x_i = x_j$ or $C(x_i, x_j) = c_{\rm out} < c_{\rm in}$ if $x_i \neq x_j$. When $q_i = 1$ for each i, we fall back on the homogeneous degree SBM model. The problem of retrieving classes information from the graph is feasible only above a certain threshold. It was proved in [10] that class reconstruction is asymptotically possible if and only if the following condition is met

$$\sqrt{\Phi}(c_{\rm in} - c_{\rm out}) > 2\sqrt{c} \tag{2}$$

where $\Phi = \mathbb{E}[q^2]$ and $c = (c_{\rm in} + c_{\rm in})/2$. From now on we will assume to work in a regime where the condition is satisfied.

Standard spectral clustering methods, well adapted in dense graphs, are however known to perform poorly close to the above threshold condition. By default of efficient mathematical methods, the first well-performing spectral method arose from statistical physics intuitions.

The main idea is as follows. One may map the clustering problem to a "minimal-energy" configuration of interacting particles (the nodes i) under a given temperature-related parameter r. The energy $E(\{x\}; r)$ formulation follows the so-called dimensionless *Ising Hamiltonian* on $\mathcal{G}$ and reads

$$E(\{x\}; r) = -\sum_{i,j\in\mathcal{V}: A_{ij}=1} \mathrm{ath}\left(\frac{1}{r}\right) x_i x_j. \tag{3}$$

Studying the system equilibria (at local minima of the energy) leads one to evaluate the Hessian matrix of the free energy $F(r) = -\log(Z)$, with Z the partition function of the Boltzmann distribution $P(\{x\}; r) = Z^{-1} e^{-E(\{x\};r)}$. Under a sparse (tree-like) network hypothesis that assumes a factorizable form for the joint probability of the x_i's, the Hessian may be approximated by the *Bethe Hessian matrix* H_r defined at "temperature" r (up to a multiplicative constant) by:

$$H_r = (r^2 - 1)I_n + D - rA \tag{4}$$

with $D = \mathrm{diag}(d_1, \ldots, d_n)$ ($d_i = [A1_n]_i$) the degree matrix.

However, selecting the parameter r (or temperature) inducing minimal energy configuration is a delicate matter. At high temperature ($r \to \infty$), the free energy is dominated by the entropy contribution and the x_i's become independent, thereby not raising any clustering of the particles. On the opposite, at low temperature the spins align in a non-trivial way and the solution can be found, provided the detectability threshold (2) is overtaken.

With an intuitive argument, the authors in [8] exploit the relation between the Bethe Hessian and the related *non-backtracking operator* $B \in \mathbb{R}^{2m\times 2m}$ defined by

$$B_{(ij)(kl)} = \delta_{jk}(1 - \delta_{il}) \tag{5}$$

for all $\{i, j, k, l\} \in \mathcal{V}$ such that $A_{ij}A_{kl} = 1$. The Bethe Hessian H_r relates to B in that eigenvalues of B correspond to values of r for which H_r is singular. It was shown in [10] that these eigenvalues $\{\gamma_i\}$ (sorted by decreasing amplitudes) of B satisfy the following:

$$\gamma_1 = \Phi\frac{c_{\rm in} + c_{\rm out}}{2}, \quad \gamma_2 = \Phi\frac{c_{\rm in} - c_{\rm out}}{2}, \quad |\gamma_{i>2}| \leq \sqrt{c\Phi}. \tag{6}$$

Thus B has two isolated real eigenvalues, and all others are contrived to a circle of radius $\sqrt{c\Phi}$ on the complex plane.

Exploiting the mapping between B and H_r, in the precise SBM setting (where $q_i = 1$ and $\Phi = 1$), [8] proposes take $r = \sqrt{c}$, i.e., to take r to be radius of the main eigenvalue bulk of B. This choice is based on observing that the sign of the elements of the second smallest eigenvector are correlated with the class labels $\{x_i\}$. Furthermore, [8] claims that, for heterogeneous degree distributions (so in particular for the DC-SBM setting), this choice should remain optimal, i.e., $r = \sqrt{\rho(B)}$ for $\rho(\cdot)$ the spectral radius.

Unlike in [8], we claim in the following that the choice $r = \sqrt{\rho(B)}$ is not optimal in the DC-SBM case. This is first seen in simulations where $H_{\sqrt{\rho(B)}}$ often performs far from optimally for either the DC-SBM model or for realistic graphs. Even for rather sparse scenarios, spectral clustering on the (supposedly suboptimal) matrices $D^{-\alpha}AD^{-\alpha}$ (even for $\alpha = 0$) often outperforms H_r.

2.2. Main result

Let us start by considering how the Bethe-Hessian matrix acts on the exact labels vector x:

$$(H_r x)_i = (r^2-1)x_i + d_i x_i - r \sum_{k \in \mathcal{N}(i)} x_k$$

where $k \in \mathcal{N}(i) \Leftrightarrow A_{ik} = 1$ (that is $\mathcal{N}(i)$ is the set of neighbors of i). Denoting $\partial_i^S \equiv \{j,\ A_{ij} = 1 \text{ and } x_i = x_j\}$ the set of neighbors of node i belonging to the *same* class and, similarly, $\partial_i^O \equiv \{j,\ A_{ij} = 1 \text{ and } x_i \neq x_j\}$ the set of neighbors of i from the *opposite* class, this simply reads

$$(H_r x)_i = x_i \left[(r^2-1) + d_i - r\left(|\partial_i^S| - |\partial_i^O| \right) \right].$$

We now make the assumption (or rather the heuristic approximation) that, although the average node degree is of order $O(1)$ with respect to n, one can approximately claim that

$$\frac{|\partial_i^S|}{d_i} \simeq \frac{c_{\rm in}}{c_{\rm in}+c_{\rm out}}, \quad \frac{|\partial_i^O|}{d_i} \simeq \frac{c_{\rm out}}{c_{\rm in}+c_{\rm out}}$$

at least for those nodes i having many neighbors (close to the decidability threshold (2), the approximation is mostly adequate to scenarios where $c_{\rm out}$ is rather large). This result can be interpreted as follows. Given a graph, the degree d_i of node i is fixed, and the probability of two neighboring nodes to be in the same class provided that they are connected is equal to $c_{\rm in}/(c_{\rm in}+c_{\rm out})$. Since the graph is sparse – hence tree-like – we can consider the neighbors of a same node as "independent" from one another. We then obtain

$$(H_r x)_i \simeq x_i \left[(r^2-1) + d_i \left(1 - r\frac{c_{\rm in}-c_{\rm out}}{c_{\rm in}+c_{\rm out}} \right) \right] \qquad (7)$$

where, under the DC-SBM model (1), the d_i's may in general be quite different. Thus, in order to retrieve an approximate eigenvector equation for x, one must set

$$r \equiv r_{\rm c} = \frac{c_{\rm in}+c_{\rm out}}{c_{\rm in}-c_{\rm out}} \qquad (8)$$

in which case $H_{r_{\rm c}} x \simeq (r_{\rm c}^2-1)x$. As such, for $r = r_{\rm c}$, one expects to see one dominant eigenvector of $H_{r_{\rm c}}$ *not tainted by the degrees* d_i.

Remark 1 (Homogeneous case)**.** *Note that in the homogeneous case where the q_i's are all equal and thus the d_i's are expected to be approximately the same,* (7) *is an approximate eigenvector equation for all r's. And thus $r_{\rm c}$ is not a particularly preferred candidate.*

This remark explains the origin of the good performances of the Bethe-Hessian for $r = \sqrt{c}$ in the SBM case. In the DC-SBM case instead, from our above arguments, proper spectral clustering is only achieved for $r = r_{\rm c}$. The chosen generalization of [8] to $r = \sqrt{\rho(B)}$ is thus inappropriate. Comparing the two values, it is easily shown that $r_{\rm c} \leq \sqrt{\rho(B)}$ with equality right at the transition (2). Somewhat counterintuitively, we thus propose a value of r *inside* the main bulk of eigenvalues of B. Besides, r_c approximately corresponds to an actual eigenvalue of B, as proved in [7] by exploiting the tree-like structure of sparse graphs.

This said, it is still important to note that the authors in [8] have identified (mostly through simulations) the eigenvector carrying the class information as the one associated to the second smallest eigenvalue of H_r for all positive r's inducing an asymptotic phase transition. This observation seems also to hold in the DC-SBM case.

As such, our final claim may then be formulated as:

Claim 1 (Spectral Clustering on $H_{r_{\rm c}}$)**.** *Assume a sparse DC-SBM model for a graph $\mathcal{G}$. Then, community detection on $\mathcal{G}$ is efficiently performed, irrespective of the heterogeneity of the degrees, by performing spectral clustering on the eigenvector attached to the second smallest eigenvalue of $H_{r_{\rm c}}$ with $r_{\rm c}$ given in Equation* (8)*.*

2.3. Estimation of $r_{\rm c}$ and relation to $D^{-1}A$

A subsequent difficulty for practical application is that $c_{\rm in} - c_{\rm out}$, and thus $r_{\rm c}$, is not directly accessible. Several solutions here exist to retrieve a good approximation for $r_{\rm c}$. One may for instance iteratively perform spectral clustering on H_r starting with, say, $r = \sqrt{\rho(B)}$ (which can be estimated by $\sum_i d_i^2 / \sum_i d_i$), obtain a first estimate of the class components, from which $c_{\rm in}$ and $c_{\rm out}$ are further estimated, and so on. Another initialization option follows from

$$(D^{-1}Ax)_i = \sum_{k \in \mathcal{N}(i)} \frac{x_k}{d_i} = \frac{|\partial_i^S| - |\partial_i^O|}{d_i} x_i \simeq \frac{c_{\rm in}-c_{\rm out}}{c_{\rm in}+c_{\rm out}} x_i. \qquad (9)$$

As such, $r_{\rm c}$ can be retrieved, with the same approximation made above on $H_{r_{\rm c}} x$, as a corresponding isolated (inverse) eigenvalue of $D^{-1}A$.[1]

3. NUMERICAL RESULTS

This section provides numerical support for our claimed results. We start first by considering synthetic DC-SBM graphs with various laws for the q_i's. For comparison fairness and adaptability to uneven class cardinalities, spectral clustering is systematically performed using the k-means algorithm rather than on a sign-based method (as opposed to [8])

[1] This, in passing, raises the question as to why $D^{-1}A$ would not be an equally valid matrix for spectral clustering as $H_{r_{\rm c}}$. The answer possibly lies in the fact that, for small $c_{\rm in} - c_{\rm out}$ (difficult clustering), the informative eigenvector is associated with a small (and thus non-dominant, not isolated) eigenvalue of $D^{-1}A$.

We first focus on the case of two even size classes. While the coming observations have been verified to be equally valid for various heterogeneous settings, we will here depict the most interesting and visible case where $q_i \in \{0.4, 1.6\}$ with $P(q_i = 0.4) = P(q_i = 1.6) = \frac{1}{2}$. In this case, both H_{r_c} and $H_{\sqrt{\rho(B)}}$ essentially have the same performance in terms of overlap (which measures the distance to random guess on a $[0, 1]$ scale), both overtaking that of $D^{-1}A$. However, a careful control of the second smallest eigenvectors of H_{r_c}, $H_{\sqrt{\rho(B)}}$ and second largest of $D^{-1}A$ (Figure 1) reveals that the second suffers from the presence of two distinct values for the q_i's by exhibiting four 'plateaus' rather than two. This is not the case of either H_{r_c} or $D^{-1}A$. Yet, when asked to retrieve exactly two classes, k-means usually performs a correct partitioning, hence the equal overlap performance. Drawing on this observation, Figure 2 compares the performance of the same three methods and for the same choice of q_i but now for two classes of uneven sizes $\frac{n}{3}$ and $\frac{2n}{3}$, respectively. In this more asymmetric situation, the performance of $H_{\sqrt{\rho(B)}}$ is strongly affected by the q_i's that k-means wrongly confuses for the genuine class divisions. Spectral clustering on H_{r_c} does not suffer this limitation.

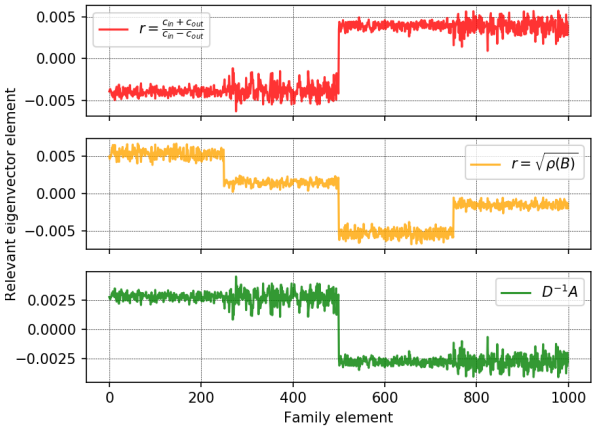

Fig. 1. Second dominant eigenvector of H_{r_c}, $H_{\sqrt{\rho(B)}}$, and $D^{-1}A$ for q_i distributed as $\frac{1}{2}\delta_{0.4} + \frac{1}{2}\delta_{1.6}$. In this case the q_i's, $i = 1, \ldots, n$, are sorted in four $\frac{n}{4}$-sized consecutive blocks as $(1.6, .4, 1.6, .4)$.

Table 1 provides a comparative overlap performance, on the same real graphs as in [8], of $H_{\sqrt{\rho(B)}}$, for the iterated method discussed in Subsection 2.3 with initialization at $r = \sqrt{\rho(B)}$ (indicated as $\sqrt{\rho(B)}^+$) or $r = 1/\lambda_2(D^{-1}A)$ (indicated as $\lambda_2^+(D^{-1}A)$), and for the oracle optimal $r = r_{\text{opt}} \in \mathbb{R}$. Consistently with our intuitive findings, it is observed that r_{opt} is systematically rather far from $\sqrt{\rho(B)}$ while $r = 1/\lambda_2(D^{-1}A)$ and further iterates are close. In terms of overlap, the proposed methods outperform the $H_{\sqrt{\rho(B)}}$ approach,

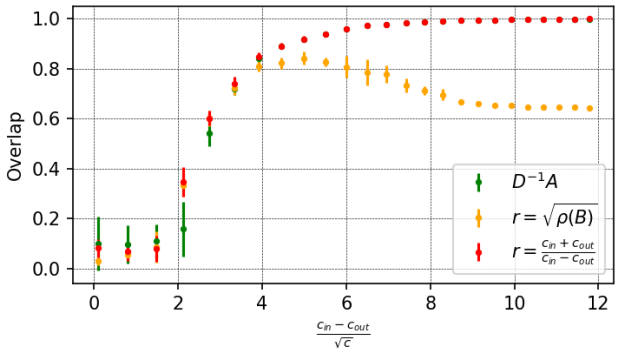

Fig. 2. Overlap performance of the three methods for an uneven population $|\mathcal{C}_1| = 2|\mathcal{C}_2|$ of classes $\mathcal{C}_1$ and $\mathcal{C}_2$, with q_i's distributed as $\frac{1}{2}\delta_{0.4} + \frac{1}{2}\delta_{1.6}$.

although some overlaps remain much lower than optimal, despite the proximity of r to r_{opt}; this is likely due both to a finite-dimensional effect as well as to the specificities of the possibly far-from-DCSBM looking graphs.

Graph / r	$\sqrt{\rho(B)}$	$\sqrt{\rho(B)}^+$	$\lambda_2^+(D^{-1}A)$	r_{opt}
Polblogs	0.32	0.59	0.03	0.90
	(8.01)	*(8.01)*	*(1.09)*	*(1.15)*
		(1.94)	*(1.00)*	
Karate	1	0.94	0.94	1
	(1.78)	*(2.14)*	*(1.15)*	*(1.78)*
		(1.19)	*(1.19)*	
Dolphins	0.93	0.97	0.97	0.97
	(1.61)	*(1.61)*	*(1.04)*	*(1.08)*
		(0.97)	*(1.08)*	

Table 1. Overlap performance on benchmark graphs and, in parentheses, starting and final values of r for the iterated estimates (r^+).

4. CONCLUDING REMARKS

This article proposes an improvement over the recently developed Bethe Hessian approach to community detection on sparse graphs. We showed that the proposed new parametrization, while performing similarly on homogeneous graphs, brings significant gains on more realistic heterogeneous graphs, as confirmed by simulations on real networks.

The cornerstone of our approach however lies in a fortunate cancelling of the heterogeneity effect on the matrix second smallest eigenvector for a precise parameter setting. Estimating the latter satisfactorily, a point of crucial practical importance, requires a more thorough analysis.

A line of further improvement lies in a deeper understanding of the link to statistical physics and its extension to more than two-class clustering.

5. REFERENCES

[1] Santo Fortunato and Darko Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1 – 44, 2016.

[2] Mark EJ Newman, "Detecting community structure in networks," *The European Physical Journal B*, vol. 38, no. 2, pp. 321–330, 2004.

[3] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[4] Amin Coja-Oghlan and André Lanka, "Finding planted partitions in random graphs with general degree distributions," *SIAM Journal on Discrete Mathematics*, vol. 23, no. 4, pp. 1682–1714, 2009.

[5] Hafiz Tiomoko Ali and Romain Couillet, "Improved spectral community detection in large heterogeneous networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8344–8392, 2017.

[6] 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*. IEEE, 2015, pp. 1347–1357.

[7] 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.

[8] 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.

[9] Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[10] Lennart Gulikers, Marc Lelarge, and Laurent Massoulié, "Non-backtracking spectrum of degree-corrected stochastic block models," *arXiv preprint arXiv:1609.02487*, 2016.