Clustering from sparse pairwise measurements

Alaa Saade, Florent Krzakala, Marc Lelarge, Lenka Zdeborová

Clustering from Sparse Pairwise Measurements

Alaa Saade
Laboratoire de Physique Statistique
École Normale Supérieure, 24 Rue Lhomond
Paris 75005

Florent Krzakala
Sorbonne Universités, UPMC Univ. Paris 06
Laboratoire de Physique Statistique, CNRS UMR 8550
École Normale Supérieure, 24 Rue Lhomond, Paris

Marc Lelarge
INRIA and École Normale Supérieure
Paris, France

Lenka Zdeborová
Institut de Physique Théorique
CEA Saclay and CNRS, France.

***Abstract*—We consider the problem of grouping items into clusters based on few random pairwise comparisons between the items. We introduce three closely related algorithms for this task: a belief propagation algorithm approximating the Bayes optimal solution, and two spectral algorithms based on the non-backtracking and Bethe Hessian operators. For the case of two symmetric clusters, we conjecture that these algorithms are asymptotically optimal in that they detect the clusters as soon as it is information theoretically possible to do so. We substantiate this claim for one of the spectral approaches we introduce.**

I. Introduction

A. Problem and model

Similarity-based clustering is a standard approach to label items in a dataset based on some measure of their resemblance. In general, given a dataset $\{x_i\}_{i\in[n]} \in \mathcal{X}^n$, and a symmetric measurement function $s : \mathcal{X}^2 \to \mathbb{R}$ quantifying the similarity between two items, the aim is to cluster the dataset from the knowledge of the pairwise measurements $s_{ij} := s(x_i, x_j)$, for $1 \le i < j \le n$. This information is conveniently encoded in a similarity graph, which vertices represent items in the dataset, and the weighted edges carry the pairwise similarities. Typical choices for this similarity graph are the complete graph and the nearest neighbor graph (see e.g. [1] for a discussion in the context of spectral clustering).

Here however, we will not assume the measurement function s to quantify the *similarity* between items, but more generally ask that the measurements be *typically different* depending on the cluster memberships of the items, in a way that will be made quantitative in the following. For instance, s could be a distance in an Euclidean space or could take values in a set of colors (i.e. s does not need to be real-valued). Additionally, we will not assume knowledge of the measurements for all pairs of items in the dataset, but only for $O(n)$ of them chosen uniformly at random. Sampling is a well-known technique to speed up computations by reducing the number of non-zero entries [2]. The main challenge is to choose the lowest possible sampling rate while still being able to detect the signal of interest. In this paper, we compute explicitly this fundamental limit for a simple probabilistic model and present three algorithms allowing partial recovery of the signal above this limit. Below the limit, in the case of two clusters, no algorithm can give an output positively correlated with the true clusters. Our three algorithms are respectively a belief propagation algorithm and two spectral algorithms based on the non-backtracking operator and the Bethe Hessian. Although these three algorithms are intimately related, so far, a sketch of rigorous analysis is available only for the spectral properties of the non-backtracking matrix. From a practical perspective however, belief propagation and the Bethe Hessian are much simpler to implement and show even better numerical performance.

To evaluate the performance of our proposed algorithms, we construct a model with n items in k predefined clusters of same average size n/k, by assigning to each item $i \in [n]$ a cluster label $c_i \in [k]$ with uniform probability $1/k$. We assume that the pairwise measurement between an item in cluster a and another item in cluster b is a random variable with density $p_{a,b}$. We choose the observed pairwise measurements uniformly at random, by generating an Erdős-Rényi random graph $G = (V = [n], E) \in \mathcal{G}(n, \alpha/n)$. The average degree α corresponds to the sampling rate: pairwise measurements are observed only on the edges of G, and α therefore controls the difficulty of the problem. From the base graph G, we build a measurement graph by weighting each edge $(ij) \in E$ with the measurement s_{ij}, drawn from the probability density p_{c_i,c_j}. The aim is to recover the cluster assignments c_i for $i \in [n]$ from the measurement graph thus constructed.

We consider the sparse regime $\alpha = O(1)$, and the limit $n \to \infty$ with fixed number of clusters k. With high probability, the graph G is disconnected, so that *exact recovery* of the clusters, as considered e.g. in [3], [4], is impossible. In this paper, we address instead the question of how many measurements are needed to *partially recover* the cluster assignments, i.e. to infer cluster assignments $\hat{c}_i$ such that the following quantity, called *overlap*, is strictly positive:

$$\max_{\sigma\in\mathfrak{S}_k} \frac{\frac{1}{n}\sum_i \mathbf{1}(\sigma(\hat{c}_i) = c_i) - \frac{1}{k}}{1 - \frac{1}{k}}, \tag{1}$$

where $\mathfrak{S}_k$ is the set of permutations of $[k]$. This quantity is monotonously increasing with the number of correctly classified items. In the limit $n \to \infty$, it vanishes for a random guess, and equals unity if the recovery is perfect. Finally, we note an important special case for which analytical results can be derived, which is the case of symmetric clusters: $\forall a, b \in [k]$

$$p_{a,b}(s) := \mathbf{1}(a = b)p_{\text{in}}(s) + \mathbf{1}(a \neq b)p_{\text{out}}(s), \tag{2}$$

where $p_{\rm in}(s)$ (resp. $p_{\rm out}(s)$) is the probability density of observing a measurement s between items of the same cluster (resp. different clusters). For this particular case, we conjecture that all of the three algorithms we propose achieve partial recovery of the clusters whenever $\alpha > \alpha_c$, where

$$\frac{1}{\alpha_c} = \frac{1}{k}\int_{\mathcal{K}} \mathrm{d}s \, \frac{\big(p_{\rm in}(s) - p_{\rm out}(s)\big)^2}{p_{\rm in}(s) + (k-1)p_{\rm out}(s)}\,, \tag{3}$$

where $\mathcal{K}$ is the support of the function $p_{\rm in} + (k-1)p_{\rm out}$. This expression corresponds to the threshold of a related reconstruction problem on trees [5]. In the following, we substantiate this claim for the case of $k=2$ symmetric clusters, and discrete measurement distributions. Note that the model we introduce is a special case of the labeled stochastic block model of [6]. In particular, for the case $k=2$, it was proven in [7] that partial recovery is information theoretically impossible if $\alpha < \alpha_c$. In this contribution, we argue that this bound is tight, namely that partial recovery is possible whenever $\alpha > \alpha_c$, and that the algorithms we propose are optimal, in that they achieve this threshold. Note also that the symmetric model (2) contains the censored block model of [8]. More precisely, if $p_{\rm in}$ and $p_{\rm out}$ are discrete distributions on $\{\pm 1\}$ with $p_{\rm in}(+1) = p_{\rm out}(-1) = 1-\epsilon$, then $\alpha_c = (1-2\epsilon)^{-2}$. In this case, the claimed result is known [9], and to the best of our knowledge, this is the only case where our result is known.

B. Motivation and related work

The ability to cluster data from as few pairwise comparisons as possible is of broad practical interest [4]. First, there are situations where all the pairwise comparisons are simply not available. This is particularly the case if a comparison is the result of a human-based experiment. For instance, in crowd-clustering [10], [11], people are asked to compare a subset of the items in a dataset, and the aim is to cluster the whole dataset based on these comparisons. Clearly, for a large dataset of size n, we can't expect to have all $O(n^2)$ measurements. Second, even if these comparisons can be automated, the typical cost of computing all pairwise measurements is $O(n^2 d)$ where d is the dimension of the data. For large datasets with n in the millions or billions, or large dimensional data, like high resolution images, this cost is often prohibitive. Storing all $O(n^2)$ measurements is also problematic. Our work supports the idea that if the measurements between different classes of items are sufficiently different, a random subsampling of $O(n)$ measurements might be enough to accurately cluster the data.

This work is inspired by recent progress in the problem of detecting communities in the sparse stochastic block model (SBM) where partial recovery is possible only when the average degree α is larger than a threshold value, first conjectured in [12], and proved in [13]–[15]. A belief propagation (BP) algorithm similar to the one presented here is introduced in [12], and argued to be optimal in the SBM. Spectral algorithms that match the performance of BP were later introduced in [16], [17]. The spectral algorithms presented here are based on a generalization of the operators that they introduce.

C. Outline and main results

In Sec. II, we describe three closely related algorithms to solve the partial recovery problem of Sec. I-A. The first one is a belief propagation (BP) algorithm approximating the Bayes optimal solution. The other two are spectral methods derived from BP. We show numerically that all three methods achieve the threshold (3). Next in Sec. III we substantiate this claim for the spectral method based on the non-backtracking operator.

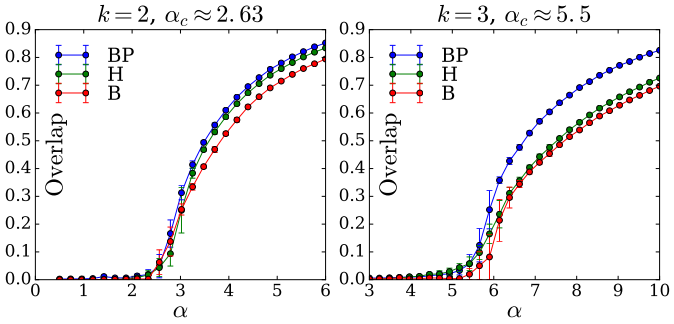

Fig. 1. Performance in clustering model-generated measurement graphs in the symmetric case (2). The overlap is averaged over 20 realizations of graphs of size $n = 10^5$, with $k = 2, 3$ clusters, and Gaussian $p_{\rm in}, p_{\rm out}$ with mean respectively 1.5 and 0, and unit variance. The theoretical transition (3) is at $\alpha_c \approx 2.63$ for $k = 2$, and conjectured to be at $\alpha_c \approx 5.5$ for $k = 3$. All three methods achieve the theoretical transition, although the Bethe Hessian (H) and belief propagation (BP) achieve a higher overlap than the non-backtracking operator (B).

II. Algorithms

A. Belief propagation

We consider a measurement graph generated from the model of Section I-A. From Bayes' rule, we have:

$$\mathrm{P}(\{c_i\}|\{s_{ij}\}) = \frac{1}{Z}\prod_{(ij)\in E} p_{c_i,c_j}(s_{ij})\,, \tag{4}$$

where Z is a normalization. The Bayes optimal assignment, maximizing the overlap (1), is $\hat{c}_i = \operatorname{argmax} \mathrm{P}_i$, the mode of the marginal of node i. We approximate this marginal using belief propagation (BP):

$$\mathrm{P}_i(c_i) \approx \frac{1}{Z_i}\prod_{l\in\partial i}\sum_{c_l=1}^{k} p_{c_i,c_l}(s_{il})\mathrm{P}_{l\to i}(c_l)\,, \tag{5}$$

where ∂i denotes the neighbors of node i in the measurement graph G, Z_i is a normalization, and the $\mathrm{P}_{i\to j}(c_i)$ are the fixed point of the recursion:

$$\mathrm{P}_{i\to j}(c_i) = \frac{1}{Z_{i\to j}}\prod_{l\in\partial i\setminus j}\sum_{c_l=1}^{k} p_{c_i,c_l}(s_{il})\mathrm{P}_{l\to i}(c_l)\,. \tag{6}$$

In practice, starting from a random initial condition, we iterate (6) until convergence, and estimate the marginals from (5). On sparse tree-like random graphs generated by our model, BP is widely believed to give asymptotically accurate results, though a rigorous proof is still lacking. This algorithm is general and applies to any model parameters p_{ab}. For now on, however, we restrict our theoretical discussion to the symmetric model (2). Eq. (6) can be written in the compact form $\mathrm{P} = F(\mathrm{P})$, where $\mathrm{P} \in \mathbb{R}^{2mk}$ and $m = |E|$ is the number of edges in G.

The first step in understanding the behavior of BP is to note that in the case of symmetric clusters (2), there exists a trivial fixed point of the recursion (6), namely $\mathrm{P}_{i\to j}(c_i) = 1/k$. This

fixed point is uninformative, yielding a vanishing overlap. If this fixed point is stable, then starting from an initial condition close to it will cause BP to fail to recover the clusters. We therefore investigate the linearization of (6) around this fixed point, given by the Jacobian J_F.

B. The non-backtracking operator

A simple computation yields

$$J_F = \mathrm{B} \otimes \left(\mathrm{I}_k - \frac{1}{k} U_k \right) , \tag{7}$$

where I_k is the $k \times k$ identity matrix, U_k is the $k \times k$ matrix with all its entries equal to 1, $\otimes$ denotes the tensor product, and B is a $2m \times 2m$ matrix called the non-backtracking operator, acting on the directed edges of the graph G, with elements for (ab) and $(cd) \in E$:

$$\begin{aligned} \mathrm{B}_{(a\to b),(c\to d)} &= w(s_{cd}) \mathbf{1}(a = d) \mathbf{1}(b \neq c) , \\ \forall s, w(s) &:= \frac{p_{\mathrm{in}}(s) - p_{\mathrm{out}}(s)}{p_{\mathrm{in}}(s) + (k-1) p_{\mathrm{out}}(s)} . \end{aligned} \tag{8}$$

Note that to be consistent with the analysis of BP, our definition of the non-backtracking operator is the transpose of the definition of [16]. This matrix generalizes the non-backtracking operators of [9], [16] to arbitrary edge weights. More precisely, for the censored block model [8], we have $s = \pm 1$ and $w(s) = (1 - 2\epsilon)s$ so that B is simply a scaled version of the matrix introduced in [9]. We also introduce an operator $\mathrm{C} \in \mathbb{R}^{n \times 2m}$ defined as

$$\mathrm{C}_{i,j\to l} = w(s_{jl}) \mathbf{1}(i = l) . \tag{9}$$

This operator follows from the linearization of eq. (5) for small $\mathrm{P}_{l\to i}$. Based on these operators, we propose the following spectral algorithm. First, compute the real eigenvalues of B with modulus greater than 1. Let r be their number, and denote by $v_1, ..., v_r \in \mathbb{R}^{2m}$ the corresponding eigenvectors. If $r = 0$, raise an error. Otherwise, form the matrix $Y = [v_1 \cdots v_r] \in \mathbb{R}^{2m \times r}$ by stacking the eigenvectors in columns, and let $X = \mathrm{C}Y \in \mathbb{R}^{n \times r}$. Finally, regarding each item i as a vector in $\mathbb{R}^r$ specified by the i-th line of X, cluster the items, using e.g. the k-means algorithm.

Theoretical guarantees for the case of $k = 2$ clusters are sketched in the next section, stating that this simple algorithm succeeds in partially recovering the true clusters all the way down to the transition (3). Intuitively, this algorithm can be thought of as a spectral relaxation of belief propagation. Indeed, for the particular case of $k = 2$ symmetric clusters, we will argue that the spectral radius of B is larger than 1 if and only if $\alpha > \alpha_c$. As a simple consequence, whenever $\alpha < \alpha_c$, the trivial fixed point of BP is stable, and BP fails to recover the clusters. On the other hand, when $\alpha > \alpha_c$, a small perturbation of the trivial fixed point grows when iterating BP. Our spectral algorithm approximates the evolution of this perturbation by replacing the non-linear operator F by its Jacobian J_F. In practice, as shown on figure 1, the non-linearity of the function F allows BP to achieve a better overlap than the spectral method based on B, but a rigorous proof that BP is asymptotically optimal is still lacking.

C. The Bethe Hessian

The non-backtracking operator B of the last section is a large, non-symmetric matrix, making the implementation of the previous algorithm numerically challenging. A much smaller, closely related symmetric matrix can be defined that empirically performs as well in recovering the clusters, and in fact slightly better than B. For a real parameter $x \geq 1$, define a matrix $\mathrm{H}(x) \in \mathbb{R}^{n \times n}$ with non-zero elements:

$$\mathrm{H}_{ij}(x) = \begin{cases} 1 + \sum_{l \in \partial i} \frac{w(s_{il})^2}{x^2 - w(s_{il})^2} \text{ if } i = j \\ -\frac{x w(s_{ij})}{x^2 - w(s_{ij})^2} \text{ if } (ij) \in E \end{cases} , \tag{10}$$

where ∂i denotes the set of neighbors of node i in the graph G, and w is defined in (8). A simple computation, analogous to [9], allows to show that $(\lambda \geq 1, v)$ is an eigenpair of B, if and only $\mathrm{H}(\lambda)v = 0$. This property justifies the following picture [17]. For x large enough, $\mathrm{H}(x)$ is positive definite and has no negative eigenvalue. As we decrease x, $\mathrm{H}(x)$ gains a new negative eigenvalue whenever x becomes smaller than an eigenvalue of B. Finally, at $x = 1$, there is a one to one correspondence between the negative eigenvalues of $\mathrm{H}(x)$ and the real eigenvalues of B that are larger than 1. We call Bethe Hessian the matrix $\mathrm{H}(1)$, and propose the following spectral algorithm, by analogy with Sec. II-B. First, compute all the negative eigenvalues of $\mathrm{H}(1)$. Let r be their number. If $r = 0$, raise an error. Otherwise, denoting $v_i, ..., v_r \in \mathbb{R}^n$ the corresponding eigenvectors, form the matrix $X = [v_1 \cdots v_r] \in \mathbb{R}^{n \times r}$ by stacking them in columns. Finally, regarding each item i as a vector in $\mathbb{R}^r$ specified by the i-th line of X, cluster the items, using e.g. the k-means algorithm.

In the case of two symmetric clusters, the results of the next section imply that if $\alpha > \alpha_c$, denoting by $\lambda_1 > 1$ the largest eigenvalue of B, the smallest eigenvalue of $\mathrm{H}(\lambda_1)$ is 0, and the corresponding eigenvector allows partial recovery of the clusters. While the present algorithm replaces the matrix $\mathrm{H}(\lambda_1)$ by the matrix $\mathrm{H}(1)$ and is therefore beyond the scope of this theoretical guarantee, we find empirically that the eigenvectors with negative eigenvalues of $\mathrm{H}(1)$ are also positively correlated with the hidden clusters, and in fact allow better recovery (see figure 1), without the need to build the non-backtracking operator B and to compute its leading eigenvalue.

This last algorithm also has an intuitive justification. It is well known [18] that BP tries to optimize the so-called Bethe free energy. In the same way B can be seen as a spectral relaxation of BP, $\mathrm{H}(1)$ can be seen as a spectral relaxation of the direct optimization of the Bethe free energy. In fact, it corresponds to the Hessian of the Bethe free energy around a trivial stationary point (see e.g. [17], [19]).

D. Numerical results

Figure 1 shows the performance of all three algorithms on model-generated problems. We consider the symmetric problem defined by (2) with $k = 2, 3$, fixed p_{in} and p_{out}, chosen to be Gaussian with a strong overlap, and we vary α. All three algorithms achieve the theoretical threshold.

While all the algorithms presented in this work assume the knowledge of the parameters of the model, namely the functions $p_{a,b}$ for $a, b \in [k]$, we argue that the belief propagation

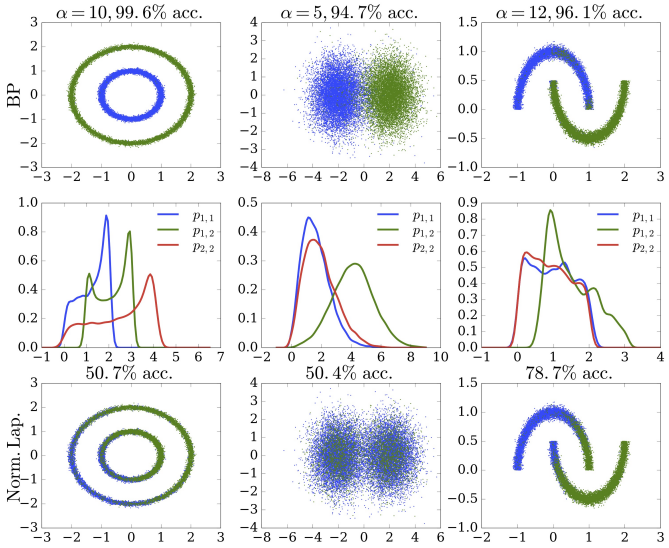

Fig. 2. Clustering of toy datasets using belief propagation. Each dataset is composed of 20000 points, 200 of which come labeled and constitute the training set. We used the Euclidean distance as the measurement function s, and estimated the probability densities p_{ab} on the training set using kernel density estimation (middle row). Although these estimates are very noisy and overlapping, belief propagation is able to achieve a very high accuracy using a random measurement graph G of small average degree α (top row). For comparison, we show in the third row the result of spectral clustering with the normalized Laplacian, using a 3-nearest neighbors similarity graph (see e.g. [1]) built from G, i.e. using only the available measurements.

algorithm is robust to large imprecisions on the estimation of these parameters. To support this claim, we show on figure 2 the result of the belief propagation algorithm on standard toy datasets where the parameters were estimated on a small fraction of labeled data.

III. Properties of the non-backtracking operator

We now state our claims concerning the spectrum of B. We restrict ourselves to the case where $k = 2$ and p_{in} and p_{out} are distributions on a finite alphabet.

Claim 1: Consider an Erdős-Rényi random graph on n vertices with average degree α, with variables assigned to vertices $c_i \in \{1, 2\}$ uniformly at random independently from the graph and measurements s_{ij} between any two neighboring vertices drawn according to the probability density: $p_{c_i,c_j}(s) = \mathbf{1}(c_i = c_j)p_{\text{in}}(s) + \mathbf{1}(c_i \neq c_j)p_{\text{out}}(s)$ for two fixed (i.e. independent of n) discrete distributions $p_{\text{in}} \neq p_{\text{out}}$ on $\mathcal{S}$. Let B be the matrix defined by (8) and denote by $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{2m}|$ the eigenvalues of B in order of decreasing magnitude, where m is the number of edges in the graph. Recall that α_c is defined by (3). Then, with probability tending to 1 as $n \to \infty$:

(i) If $\alpha < \alpha_c$, then $|\lambda_1| \leq \sqrt{\frac{\alpha}{\alpha_c}} + o(1)$.

(ii) If $\alpha > \alpha_c$, then $\lambda_1 = \frac{\alpha}{\alpha_c} + o(1)$ and $|\lambda_2| \leq \sqrt{\frac{\alpha}{\alpha_c}} + o(1)$. Additionally, denoting by v the eigenvector associated with λ_1, $\mathrm{C}v$ is positively correlated with the planted variables $(c_i)_{i\in[n]}$, where C is defined in (9).

Note that for the censored block model, our claim implies Theorem 1 in [9]. The main idea which substantiates our claim is to introduce a new non-backtracking operator with spectral properties close to those of B and then apply the techniques developed in [20] to it. We try to use notations consistent with [20]: for an oriented edge $e = u \to v = (u, v)$ from node u to node v, we set $e_1 = u$, $e_2 = v$ and $e^{-1} = (v, u)$. For a matrice M, its transpose is denoted by M^*. We also define $\sigma_i = 2c_i - 3$ for each $i \in [n]$.

We start by a simple transformation: if t is the vector in $\mathbb{R}^{\vec{E}}$ defined by $t_e = \sigma_{e_1}$ and $\odot$ is the Hadamard product, i.e. $(t \odot x)_e = \sigma_{e_1} x_e$, then we have

$$\mathrm{B}^* x = \lambda x \Leftrightarrow \mathrm{B}^X (t \odot x) = \lambda (t \odot x)\,, \tag{11}$$

with B^X defined by $\mathrm{B}^X_{ef} = \mathrm{B}_{fe}\sigma_{e_1}\sigma_{e_2}$. In particular, B^X and B have the same spectrum and there is a trivial relation between their eigenvectors. With $X_e = \sigma_{e_1} w(s_e)\sigma_{e_2}$, we have:

$$\mathrm{B}^X_{ef} = X_e \mathbf{1}(e_2 = f_1)\mathbf{1}(e_1 \neq f_2)\,.$$

Moreover, note that the random variables $(A_f = \sigma_{f_1}\sigma_{f_2})_{f\in E}$ are random signs with $\mathbf{P}(A_f = 1) = 1/2$ and the random variables $(X_f)_{f\in E}$ are such that

$$\mathbf{E}X_f = \mathbf{E}X_f^2 = \frac{1}{2}\sum_s \frac{(p_{\text{in}}(s) - p_{\text{out}}(s))^2}{p_{\text{in}}(s) + p_{\text{out}}(s)} = \frac{1}{\alpha_c}\,.$$

We now define another non-backtracking operator B^Y. First, letting $\epsilon(s) = \frac{p_{\text{out}}(s)}{p_{\text{in}}(s)+p_{\text{out}}(s)} \in [0, 1]$, we define the sequence of independent random variables $\{\tilde{Y}_e\}_{e\in E}$ with $\mathbf{P}(\tilde{Y}_e = +1|s_e) = 1 - \mathbf{P}(\tilde{Y}_e = -1|s_e) = 1 - \epsilon(s_e)$, so that $\mathbf{E}[\tilde{Y}_e|s_e] = w(s_e)$. We define $Y_e = \tilde{Y}_e\sigma_{e_1}\sigma_{e_2}$ and finally

$$\mathrm{B}^Y_{ef} = Y_f \mathbf{1}(e_2 = f_1)\mathbf{1}(e_1 \neq f_2)\,,$$

so that $\mathbf{E}[\mathrm{B}^Y|G, \{s_e\}_{e\in E}] = B^X$. It turns out that the analysis of the matrix B^Y can be done with the techniques developped in [20]. More precisely, we define P the linear mapping on $\mathbb{R}^{\vec{E}}$ defined by $(Px)_e = Y_e x_{e^{-1}}$ (i.e. in matrix form $P_{ef} = Y_e \mathbf{1}(f = e^{-1})$). Note that $P^* = P$ and since $Y_e^2 = 1$, P is an involution so that P is an orthogonal matrix. A simple computation shows that $(B^Y)^k P = P(B^Y)^{*k}$, hence $(B^Y)^k P$ is a symmetric matrix. This symmetry corresponds to the oriented path symmetry in [20] and is crucial to our analysis. If $(\tau_{j,k}), 1 \leq j \leq 2m$ are the eigenvalues of $(B^Y)^k P$ and $(x_{j,k})$ is an orthonormal basis of eigenvectors, we deduce that

$$(B^Y)^k = \sum_{j=1}^{2m} \tau_{j,k} x_{j,k} (P x_{j,k})^*\,. \tag{12}$$

Since P is an orthogonal matrix $(Px_{j,k}), 1 \leq j \leq 2m$ is also an orthonormal basis of $\mathbb{R}^{\vec{E}}$. In particular, (12) gives the singular value decomposition of $(B^Y)^k$. Indeed, if $t_{j,k} = |\tau_{j,k}|$ and $y_{j,k} = \mathrm{sign}(\tau_{j,k})Px_{j,k}$, then we get $(B^Y)^k = \sum_{j=1}^{2m} t_{j,k} x_{j,k} y_{j,k}^*$, which is precisely the singular value decomposition of $(B^Y)^k$. As shown in [20], for large k, the decomposition (12) carries structural information on the graph.

A crucial element in the proof of [20] is the result of Kesten and Stigum [21], [22] and we give now its extension required here which can be seen as a version of Kesten and Stigum's

work in a random environment. We write $\mathbb{N}^* = \{1, 2, \dots\}$ and $U = \cup_{n\geq 0}(\mathbb{N}^*)^n$ the set of finite sequences composed by $\mathbb{N}^*$, where $(\mathbb{N}^*)^0$ contains the null sequence $\emptyset$. For $u, v \in U$, we note $|u| = n$ for the lenght of u and uv for the sequence obtained by the juxtaposition of u and v. Suppose that $\{(N_u, A_{u1}, A_{u2}, \dots)\}_{u\in U}$ is a sequence of i.i.d. random variables with value in $\mathbb{N} \times \mathbb{R}^{\mathbb{N}^*}$ such that N_u is a Poisson random variable with mean α and the A_{ui} are independent i.i.d. random signs with $\mathbf{P}(A_{u1} = 1) = \mathbf{P}(A_{u1} = -1) = \frac{1}{2}$. We then define the following random variables: first s_u such that $\mathbf{P}(s_u|A_u = 1) = p_{\text{in}}(s_u)$ and $\mathbf{P}(s_u|A_u = -1) = p_{\text{out}}(s_u)$, then $X_u = A_u w(s_u)$ and $Y_u = A_u \tilde{Y}_u$ where $\mathbf{P}(\tilde{Y}_u = 1|s_u) = 1 - \mathbf{P}(Y_u = -1|s_u) = 1 - \epsilon(s_u)$. We assume that for all $u \in U$ and $i > N_u$, $A_{ui} = s_{ui} = X_{ui} = Y_{ui} = 0$. N_u will be the number of children of node u and the sequence $(s_{u1}, \dots, s_{uN_u})$ the measurements on edges between u and its children. We set for $u = u_1 u_2 \dots u_n \in U$,

$$P^X_\emptyset = 1, \qquad P^X_u = X_{u_1} X_{u_1 u_2} \dots X_{u_1 \dots u_n},$$
$$P^Y_\emptyset = 1, \qquad P^Y_u = Y_{u_1} Y_{u_1 u_2} \dots Y_{u_1 \dots u_n}.$$

We define (with the convention $\frac{0}{0} = 0$) ,

$$M_0 = 1, \quad M_t = \sum_{|u|=t} \frac{P^Y_u}{\alpha^t P^X_u}. \tag{13}$$

Then conditionnaly on the variables $(s_u)_{u\in U}$, M_t is a martingale converging almost surely and in L^2 as soon as $\alpha > \alpha_c$. The fact that this martingale is bounded in L^2 follows from an argument given in the proof of Theorem 3 in [6].

In order to apply the technique of [20], we need to deal with the ℓ-th power of the non-backtracking operators. For ℓ not too large, the local struture of the graph (up to depth ℓ) can be coupled to a Poisson Galton-Watson branching process, so that the computations done for M_ℓ above provide a good approximation of the ℓ-th power of the non-backtracking operator and we can use the algebraic tools about perturbation of eigenvalues and eigenvectors, see the Bauer-Fike theorem in Section 4 in [20].

IV. Conclusion

We have considered the problem of clustering a dataset from as few measurements as possible. On a reasonable model, we have made a precise prediction on the number of measurements needed to cluster better than chance, and have substantiated this prediction on an interesting particular case. We have also introduced three efficient and optimal algorithms, based on belief propagation, to cluster model-generated data starting from this transition. Our results suggest that clustering can be significantly sped up by using a number of measurements *linear* in the size of the dataset, instead of quadratic. These algorithms, however, require an estimate of the distribution of the measurements between objects depending on their cluster membership. On toy datasets, we have demonstrated the robustness of the belief propagation algorithm to large imprecisions on these estimates, paving the way for broad applications in real world, large-scale data analysis. A natural avenue for future work is the unsupervised estimation of these distributions, through e.g. an expectation-maximization approach.

Acknowledgment

This work has been supported by the ERC under the European Union's FP7 Grant Agreement 307087-SPARCS and by the French Agence Nationale de la Recherche under reference ANR-11-JS02-005-01 (GAP project).

References

[1] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, p. 395, 2007.

[2] D. Achlioptas and F. McSherry, "Fast computation of low rank matrix approximations," in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*. ACM, 2001, pp. 611–618.

[3] V. Jog and P.-L. Loh, "Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence," *arXiv preprint arXiv:1509.06418*, 2015.

[4] Y. Chen, C. Suh, and A. Goldsmith, "Information recovery from pairwise measurements: A shannon-theoretic approach," in *Information Theory, 2015 IEEE International Symposium on*, 2015, p. 2336.

[5] M. Mézard and A. Montanari, "Reconstruction on trees and spin glass transition," *Journal of statistical physics*, vol. 124, no. 6, pp. 1317–1350, 2006.

[6] S. Heimlicher, M. Lelarge, and L. Massoulié, "Community detection in the labelled stochastic block model," 09 2012. [Online]. Available: http://arxiv.org/abs/1209.2910

[7] M. Lelarge, L. Massoulie, and J. Xu, "Reconstruction in the labeled stochastic block model," in *Information Theory Workshop (ITW), 2013 IEEE*, Sept 2013, pp. 1–5.

[8] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, "Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery," *arXiv:1404.4749*, 2014.

[9] A. Saade, M. Lelarge, F. Krzakala, and L. Zdeborova, "Spectral detection in the censored block model," in *Information Theory (ISIT), 2015 IEEE International Symposium on*, June 2015, pp. 1184–1188.

[10] R. G. Gomes, P. Welinder, A. Krause, and P. Perona, "Crowdclustering," in *Advances in neural information processing systems*, 2011, p. 558.

[11] J. Yi, R. Jin, A. K. Jain, and S. Jain, "Crowdclustering with sparse pairwise labels: A matrix completion approach," in *AAAI Workshop on Human Computation*, vol. 2, 2012.

[12] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Phys. Rev. E*, vol. 84, no. 6, p. 066106, 2011.

[13] E. Mossel, J. Neeman, and A. Sly, "Stochastic block models and reconstruction," *arXiv preprint arXiv:1202.1499*, 2012.

[14] L. Massoulie, "Community detection thresholds and the weak ramanujan property," *arXiv preprint arXiv:1311.3085*, 2013.

[15] E. Mossel, J. Neeman, and A. Sly, "A proof of the block model threshold conjecture," *arXiv preprint arXiv:1311.4115*, 2013.

[16] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, "Spectral redemption in clustering sparse networks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, no. 52, pp. 20 935–20 940, 2013.

[17] A. Saade, F. Krzakala, and L. Zdeborová, "Spectral clustering of graphs with the bethe hessian," in *Advances in Neural Information Processing Systems*, 2014, pp. 406–414.

[18] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Bethe free energy, kikuchi approximations, and belief propagation algorithms," *Advances in neural information processing systems*, vol. 13, 2001.

[19] A. Saade, F. Krzakala, and L. Zdeborová, "Matrix completion from fewer entries: Spectral detectability and rank estimation," in *Advances in Neural Information Processing Systems*, 2015, pp. 1261–1269.

[20] C. Bordenave, M. Lelarge, and L. Massoulié, "Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs," *arXiv*, 2015.

[21] H. Kesten and B. P. Stigum, "A limit theorem for multidimensional galton-watson processes," *The Annals of Mathematical Statistics*, vol. 37, no. 5, pp. 1211–1223, 1966.

[22] ——, "Additional limit theorems for indecomposable multidimensional galton-watson processes," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1463–1481, 12 1966.