Almost Exact Recovery in Label Spreading

Konstantin Avrachenkov and Maximilien Dreveton

Inria Sophia Antipolis
2004 Route des Lucioles, 06902 Valbonne, France
k.avrachenkov@inria.fr
maximilien.dreveton@inria.fr

Abstract. In semi-supervised graph clustering setting, an expert provides cluster membership of few nodes. This little amount of information allows one to achieve high accuracy clustering using efficient computational procedures. Our main goal is to provide a theoretical justification why the graph-based semi-supervised learning works very well. Specifically, for the Stochastic Block Model in the moderately sparse regime, we prove that popular semi-supervised clustering methods like Label Spreading achieve asymptotically almost exact recovery as long as the fraction of labeled nodes does not go to zero and the average degree goes to infinity.

Keywords: Semi-supervised clustering · community detection · label spreading · random graphs · stochastic block model.

1 Introduction and previous work

Graph clustering consists of partitioning a graph into communities (or clusters) so that nodes in the same cluster are, in some sense, more densely connected than nodes belonging to different clusters. Graph clustering (or community detection) is a fundamental problem in machine learning. Many scientific disciplines rely on graphs to model a large number of interacting agents: atoms or interacting particles in statistical physics, proteins interactions in molecular biology, social networks in sociology, the Internet's webgraph in computer science, etc. Such complex networks typically have clustering structure, whose detection and description is very important for network analysis.

To model complex networks, we can interpret them as random graphs. The simplest random graph model with clustering structure is the Stochastic Block Model (SBM), introduced independently in [6] and [9]. SBM is a generalization of the Erdős-Rényi (ER) random graph [7,8]. In its easiest form, an SBM graph has two communities of equal size, and edges between nodes of the same community are drawn with probability p, and edges between nodes of different communities have a probability q, where $p \neq q$. Of course, this is a very basic model of a graph with clustering structure. Despite its simplicity, the basic SBM poses a number of theoretical challenges for community detection problem and highlights various intuitions and trade-offs.

Community detection in SBM is still a very active topic, and one can find a recent and complete review in [1], mentioning the up to date unsupervised clustering results. In this paper, we will consider a semi-supervised situation, where an oracle reveals the community belonging of a fraction of nodes. In practice, labeling nodes according to their community requires human intervention, thus is expensive (could be months of experiments in a case of protein study), and the fraction of pre-labeled nodes is expected to be the smallest possible. As was noted in the previous publications on graph-based semi-supervised learning (see e.g., [2, 5, 14–16]), it is a very powerful technique allowing to achieve high accuracy with only a small number of labeled data points. Moreover, as those methods are naturally distributed, they can efficiently cluster large graphs.

A popular graph based semi-supervised method is Label Spreading [14]. The main goal of the present work is to provide a theoretical justification why Label Spreading works well, by showing that it achieves almost exact recovery on SBM graphs, in the moderately sparse regime (when the average degree d is of the order of $\log n$), as long as the fraction of labeled points r does not go to zero.

Note that the recovery is said to be exact if all nodes are correctly labeled (almost surely, in the limit as n goes to infinity), and almost exact if the fraction of misclassified nodes goes to 0 (almost surely, when n goes to infinity) [1].

The paper is structured as follows: in Section 2, we describe the minimization procedure we used for semi-supervised graph clustering (Label Spreading) and provide more background references on the semi-supervised learning. In Section 3 we study the case of SBM graphs, using a mean field analysis. We derive the exact expression for the semi-supervised solution of the mean field SBM and explain why exact recovery is possible for the mean field. Then, we show concentration of the limit towards its mean field value and conclude with the recovery result. Section 4 provides discussion and directions for future research.

2 Semi-supervised graph clustering with the normalized Laplacian matrix (Label Spreading)

Let $G = (V, E)$ be a graph, where V is the set of n nodes, and E is the set of m edges. In the following, we will consider weighted undirected networks: each edge $(ij) \in E$ holds a positive weight w_{ij}. Thus, the graph can be fully represented by a symmetric matrix W, where the entry (ij) of W is the weight w_{ij} of an edge between nodes i and j (a weight of zero corresponds to the absence of edge). When the weights are binary, the weight matrix is called the adjacency matrix and is traditionally denoted by A. The degree d_i of a node $i \in V$ is defined as the sum of the weights of all edges going from i, that is $d_i = \sum_j w_{ij}$. The diagonal matrix D with entries d_i is called the degree matrix.

We will consider a graph exhibiting a community structure: hence, the set of nodes can be partitioned into K non overlapping communities (or clusters). By observing only V and E, and supposing K known, we aim to recover the underlying partition in a semi-supervised manner. This means some nodes are already labeled: we know to which community they belong. Let ℓ and u be

respectively the set of labeled node and the set of unlabeled nodes. Without loss of generality, we can suppose that the first $|\ell|$ nodes are labeled, and we define r the ratio of labeled nodes with respect to the total number of nodes ($|\ell| = r|V|$).

Our strategy is to find a matrix X of size $n \times K$ from which we could predict the node's labels. We will refer to the columns $X_{\cdot k}$ as classification functions, and node i will be classified in cluster $k(i)$ if:

$$k(i) = \underset{k' \in \{1,\dots,K\}}{\arg\max}\ X_{ik'}. \tag{1}$$

To make use of the semi-supervised setting, we shall fix the values of X on the labeled data. More precisely, we introduce the $n \times K$ ground-truth matrix Y as:

$$Y_{ik} = \begin{cases} 1 & \text{if node } i \text{ is in community } k \\ 0 & \text{otherwise.} \end{cases}$$

Since $Y_{\ell\cdot}$ is known, where $Y_{\ell\cdot}$ denotes the first $|\ell|$ rows of the matrix Y (corresponding to the labeled nodes), we will enforce $X_{\ell\cdot} = Y_{\ell\cdot}$. The other rows of X, denoted $X_{u\cdot}$, will be chosen to minimize the energy function:

$$E(X) := \operatorname{tr}\big(X^T \mathcal{L} X\big) \tag{2}$$

$$\text{such that } X_{\ell\cdot} = Y_{\ell\cdot} \tag{3}$$

where $\mathcal{L} := I_n - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ is the normalized Laplacian of the graph.

The choice to minimize an energy function to solve a semi-supervised learning problem can be traced back to [16]. In that paper, the authors chose a standard Laplacian-based energy function. In later works (see e.g., [2, 11, 14]) it has been shown that one can achieve a better accuracy with the use of the normalized Laplacian. There is another important argument why we have chosen to focus on the normalized Laplacian method: as it will be clear from the ensuing development, the normalized Laplacian's spectral norm concentrates sufficiently well around its expectation [12].

The minimization problem (2)-(3) can be solved using Lagrange multiplier:

$$L(X) := E(X) + \lambda \operatorname{tr}\Big((X_{\ell\cdot} - Y_{\ell\cdot})^T (X_{\ell\cdot} - Y_{\ell\cdot})\Big). \tag{4}$$

To compute the solution explicitly in a matrix form, we split the weight matrix W (and other matrices like D) into four blocks $\begin{pmatrix} W_{\ell\ell} & W_{\ell u} \\ W_{u\ell} & W_{uu} \end{pmatrix}$, where $W_{\ell\ell}$ is a submatrix corresponding to the first $|\ell|$ rows and columns of matrix W. The solution $X = \begin{pmatrix} X_\ell \\ X_u \end{pmatrix}$ of the optimization problem (2)-(3) can be derived by letting the partial derivatives of the convex function L with respect to X_{ik} (for $i \notin \ell$ and $k \in \{1, \dots, K\}$) being zero, and writing the solution in a matrix form. More

precisely, let us rewrite the Lagrangian given in equation (4) as follows:

$$\begin{aligned} L(X) &= \sum_{k=1}^{K} \Big(X_{.k}^T \, \mathcal{L} \, X_{.k} + \lambda (X_{\ell k} - Y_{.k})^T (X_{\ell k} - Y_{\ell k}) \Big) \\ &= \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j=1}^{n} w_{ij} \left(\frac{X_{ik}}{\sqrt{d_i}} - \frac{X_{jk}}{\sqrt{d_j}} \right)^2 + \lambda \sum_{k=1}^{K} \sum_{i=1}^{\ell} \big(X_{ik} - Y_{ik} \big)^2 . \end{aligned}$$

Thus, for all $k \in \{1, \dots, K\}$, the first order condition $\dfrac{\partial L}{\partial X_{.k}}(X) = 0$ gives

$$\mathcal{L} X + \lambda S (X - Y) = 0,$$

where $S = \begin{pmatrix} I_{|\ell|} & 0 \\ 0 & 0 \end{pmatrix}$ is an $n \times n$ matrix. Using the block notation introduced earlier leads to the following equations:

$$\forall k \in \{1, \dots, K\} : \; \mathcal{L}_{uu} X_{uk} + \mathcal{L}_{u\ell} X_{\ell k} = 0.$$

By recalling the condition $X_{\ell .} = Y_{\ell .}$, the last equation can be rewritten as

$$X_{u .} = -\mathcal{L}_{uu}^{-1} \, \mathcal{L}_{u\ell} \, Y_{\ell .} \tag{5}$$

Note that $\mathcal{L}_{uu}$ is an extracted block from the normalized Laplacian, hence is invertible, and the expression (5) is well defined as soon as each connected component of the graph has at least one labeled node. The expression (5) depends only on the value of the labeled nodes and on the topology of the graph.

3 Analysis on random SBM graphs

Let us set up the notations for SBM. Each node $i \in \{1, \dots, n\}$ will belong to a cluster $\mathcal{C}_i$. Then, an edge is created between a pair of nodes (ij) with a probability that depends only on nodes' clusters:

$$\Pr\big((ij) \in E\big) = \mathrm{P}_{\mathcal{C}_i \mathcal{C}_j} .$$

The adjacency matrix A is thus a random matrix, whose expected value is

$$\mathbf{E} A_{ij} = \mathrm{P}_{\mathcal{C}_i \mathcal{C}_j} . \tag{6}$$

The weighted graph formed by the expected adjacency matrix of an SBM graph, given by (6), will be called mean field model.

It is common to call $p_i = \mathrm{P}_{\mathcal{C}_i \mathcal{C}_i}$ the intra-cluster edge probabilities and $q = \mathrm{P}_{\mathcal{C}_i \mathcal{C}_j}, i \neq j$ the inter-cluster edge probability (we assume that the inter-cluster edge probabilities are all equal to each other). We will denote by n_i the number of nodes in community i, with $n = \sum_{i=1}^{K} n_i$. Finally, d_i will be the average degree of nodes in cluster i.

We will mostly focus on the symmetric SBM with two communities, and will make use of the following assumptions; nonetheless, for each result, we will clearly state which assumption is needed. We think our results stand for more than two communities as well as in the non symmetric case (incorporating so-called Class Prior Knowledge, see for example [5] Section 10.8), to the price of harder and longer computations.

Assumption 1 (Symmetric SBM) *We consider an SBM graph with two communities of equal size* $n_1 = n_2 = \frac{n}{2}$ *and* $p_1 = p_2 =: p$.

Assumption 2 (Growing degrees) *The average degree* d *goes to infinity.*

Assumption 3 (Fixed fraction of labelled nodes) *The fraction of labeled nodes* r *remains constant as* n *grows to* $+\infty$.

Assumption 4 (Labeled nodes uniformly distributed) *Each community has the same fraction of labeled nodes (with respect to the community size), and they are chosen uniformly at random. Moreover, we assume that there is at least one labeled node in each connected component of the graph.*

The second part of Assumption 4 is needed to ensure that the extracted Laplacian $\mathcal{L}_{uu}$ is invertible. We can now state the main result of this paper.

Theorem 1 (Asymptotically almost exact recovery). *Label Spreading algorithm, defined by the minimization scheme* (2)-(3)*, enables asymptotically almost exact recovery for an SBM graph under Assumptions 1-4.*

We will prove Theorem 1 in two steps. First, by doing exact calculation of the mean field solution X^{MF}, we will show that exact (even nonasymptotic) recovery is possible for the mean field model. Then, we will show that the solution of the minimization problem (2)-(3) is asymptotically concentrated sufficiently well around its mean field value. Those two results put together will give the proof of Theorem 1.

3.1 Exact expression for mean field SBM

Recall that by mean field, we are referring to the situation where the random quantities are replaced by their means. In particular, we call mean field model the weighted graph formed by the expected adjacency matrix of an SBM graph.

In all the following, the subscript MF will be added to all quantities referring to the the mean field model. For simplicity of notations and computations, we will assume there is only two communities, but the analysis can be extended to K communities.

Let 1_{n_1} denote the column vector of size $n_1 \times 1$ with all entries equal to one, and by $J_{n_1;n_2} := 1_{n_1} 1_{n_2}^T$ the matrix of size $n_1 \times n_2$ with all entries equal to one. Furthermore, we will use a shorten notation J_{n_1} for $J_{n_1;n_1}$.

Without loss of generality and for the purpose of performance analysis, we implicitly assume that the first n_1 nodes are in cluster 1, whereas the last n_2 nodes are in cluster 2. Thus,

$$A^{MF} := \mathbf{E}A = \begin{pmatrix} p_1\, J_{n_1} & q\, J_{n_1 n_2} \\ q\, J_{n_2 n_1} & p_2\, J_{n_2} \end{pmatrix}.$$

In order for derivations to be more transparent, we also consider the case where diagonal elements of A^{MF} are not zero. This corresponds to a non-standard definition of SBM, where we could have edges $(i;i)$, with probability p_1 or p_2 depending on the community to whom i belongs to. Nonetheless, we could set the diagonals elements of A^{MF} to zero and our results would still hold.

Also without loss of generality and for the convenience of analysis, we will assume that the first rn_1 and the last rn_2 nodes are labeled. Note that if the quantities rn_i are not integers, we take their integer part, but we shall omit it to simplify the notations. Lastly, we introduce $\tilde{n}_i = (1-r)n_i$ the number of unlabeled nodes in cluster i.

Theorem 2 (Exact expression for X^{MF}). *Let* $a = \dfrac{p_1}{d_1}$, $b = c = \dfrac{q}{\sqrt{d_1 d_2}}$, $d = \dfrac{p_2}{d_2}$ *and* $F := \big(1 - \dfrac{p_1 \tilde{n}_1}{d_1}\big)\big(1 - \dfrac{p_2 \tilde{n}_2}{d_2}\big) - \tilde{n}_1 \tilde{n}_2 \dfrac{q^2}{d_1 d_2}$. *Then*

$$X^{MF}_{u.} = \begin{pmatrix} x^{MF}_{11} J_{(1-r)n_1} & x^{MF}_{12} J_{(1-r)n_1} \\ x^{MF}_{21} J_{(1-r)n_2} & x^{MF}_{22} J_{(1-r)n_2} \end{pmatrix},$$

where:

- $x^{MF}_{11} = rn_1\Big(a - n_1 \dfrac{(1-r)a}{F}\big(-a + \tilde{n}_2(ad-bc)\big) + \dfrac{(1-r)bc}{F} n_2\Big)$;
- $x^{MF}_{12} = rn_2\Big(b - \dfrac{(1-r)b}{F} n_1(-a + \tilde{n}_2(ad-bc)\big) + d\dfrac{(1-r)b}{F} n_2\Big)$;
- $x^{MF}_{21} = rn_1\Big(c + rn_1 \dfrac{(1-r)ac}{F} n_1 - n_2 \dfrac{(1-r)c}{F}\big(-d + \tilde{n}_1(ad-bc)\big)\Big)$;
- $x^{MF}_{22} = rn_2\Big(d + \dfrac{(1-r)bc}{F} n_1 - n_2 \dfrac{(1-r)d}{F}(-d + \tilde{n}_1(ad-bc))\Big)$.

Proof. Recall from equation (5) that $X^{MF}_{u.} = -\big(\mathcal{L}^{MF}_{uu}\big)^{-1}\, \mathcal{L}^{MF}_{u\ell}\, Y_{\ell.}$.

First, let us notice that $\Big(D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\Big)^{MF}_{uu} = \begin{pmatrix} aJ_{\tilde{n}_1} & bJ_{\tilde{n}_1 \tilde{n}_2} \\ cJ_{\tilde{n}_2 \tilde{n}_1} & dJ_{\tilde{n}_2} \end{pmatrix}$, where the quantities a, b, c and d are defined in the statement of the theorem. It follows from Proposition 2 in the Appendix that

$$\Big(\mathcal{L}^{MF}_{uu}\Big)^{-1} = I_{\tilde{n}} - \frac{1}{F}\begin{pmatrix} \big(-a + \tilde{n}_2(ad-bc)\big) J_{\tilde{n}_1} & -bJ_{\tilde{n}_1 \tilde{n}_2} \\ -cJ_{\tilde{n}_2 \tilde{n}_1} & \big(-d + \tilde{n}_1(ad-bc)\big) J_{\tilde{n}_2} \end{pmatrix}.$$

Moreover, $-\mathcal{L}^{MF}_{u\ell} = \begin{pmatrix} aJ_{\tilde{n}_1; rn_1} & bJ_{\tilde{n}_1; rn_2} \\ cJ_{\tilde{n}_2; rn_1} & dJ_{\tilde{n}_2; rn_2} \end{pmatrix}$ and $X_{\ell.} = \begin{pmatrix} 1_{rn_1} & 0_{rn_1} \\ 0_{rn_2} & 1_{rn_2} \end{pmatrix}$, thus

$$-\mathcal{L}^{MF}_{u\ell}\, X_{\ell.} = \begin{pmatrix} rn_1 a\, 1_{\tilde{n}_1} & rn_2 b\, 1_{\tilde{n}_1} \\ rn_1 c\, 1_{\tilde{n}_2} & rn_2 d\, 1_{\tilde{n}_2} \end{pmatrix},$$

and the product of $\big(\mathcal{L}^{MF}_{uu}\big)^{-1}$ by $-\mathcal{L}^{MF}_{u\ell}\, X_{\ell.}$ gives the stated result. □

Proposition 1 (Exact recovery in mean field model). *The minimization procedure* (2)-(3) *achieves exact recovery in the mean field model of an SBM graph with two clusters of equal size, with* $p_1 = p_2$ *(Assumption 1),* $p > q$ *(associative communities) and with the same fraction* $r > 0$ *of labeled nodes in each cluster.*

Proof. Recall that the detection rule is given in equation (1). In the two communities case, recovery will be possible (and 100% correct) if and only if $x_{11} > x_{12}$ and $x_{22} > x_{21}$. By symmetry of the problem, it is enough to consider the condition $x_{11} > x_{12}$.

In the symmetric case, with two clusters of equal size ($n_1 = n_2$) and $p_1 = p_2$ (Assumption 1), it is then straightforward to see that

$$x_{11}^{MF} = r\frac{p}{p+q} + \frac{r(1-r)}{F}\frac{p}{(p+q)^2}(rp+(1-r)q) + \frac{r(1-r)}{F}\frac{q^2}{(p+q)^2},$$
$$x_{12}^{MF} = r\frac{q}{p+q} + \frac{r(1-r)}{F}\frac{q}{(p+q)^2}(rp+(1-r)q) + \frac{r(1-r)}{F}\frac{pq}{(p+q)^2}.$$

By subtracting those two lines, a little of algebra shows that

$$x_{11}^{MF} - x_{12}^{MF} = r\frac{p-q}{2q+r(p-q)}.$$

This last quantity is positive as soon as $p > q$, and this ends the proof. □

We can make two remarks:

- First, note that we have not made any assumptions on the scaling of p_i and q with n, except that p_1 and p_2 are equal. In particular, the result holds in the case of logarithmic degree, which will be our main focus later on.
- Second, in the case of the mean field model, the result holds for finite n, thus it is exact (even non-asymptotic) recovery in the mean field model. It is not surprising, since recovery in the mean field model is obvious.

3.2 Concentration towards mean field

Similarly to the concentration result in [3], we establish the concentration of X around its mean field value X^{MF} in terms of the Euclidean norm. For the sake of better readability, we omit the subscripts from $X_{\cdot k}$ and $X_{\cdot k}^{MF}$ ($k \in \{1, \dots, K\}$) in the next theorem and the two following proofs. Similarly, we will shorten X_{uk} (respectively $Y_{\ell k}$) to X_u (respectively Y_ℓ).

Theorem 3. *Under the same assumptions as Theorem 1, for each class, the relative Euclidean distance between the solution* X *given by Label Spreading and its mean field value* X^{MF} *converges in probability to zero. More precisely, with high probability, we can find a constant* $C > 0$ *such that:*

$$\frac{||X - X^{MF}||}{||X^{MF}||} \leq \frac{C}{\sqrt{d}}. \tag{7}$$

Proof. Let us rewrite equation (5) as a perturbation of a system of linear equations corresponding to the mean field solution:

$$\big(\mathbf{E}\mathcal{L} + \Delta\mathcal{L}\big)_{uu}\big(X_u^{MF} + \Delta X_u\big) = -\Big(\mathbf{E}\mathcal{L} + \Delta\mathcal{L}\Big)_{u\ell} Y_\ell,$$

where $\Delta X := X - X^{MF}$ and $\Delta\mathcal{L} := \mathcal{L} - \mathbf{E}\mathcal{L}$.

Recall that a perturbation of a system of linear equations $(A+\Delta A)(x+\Delta x) = b + \Delta b$ leads to the following sensitivity inequality (see e.g., section 5.8 in [10]):

$$\frac{||\Delta x||}{||x||} \leq \frac{\kappa(A)}{1 - \kappa(A)\dfrac{||\Delta A||}{||A||}}\left(\frac{||\Delta b||}{||b||} + \frac{||\Delta A||}{||A||}\right)$$

where $||.||$ is a matrix norm associated to a vector norm $||.||$ (we used the same notations for simplicity) and $\kappa(A) := ||A^{-1}||.||A||$ the conditioning number. In our case, using spectral norm, this gives:

$$\frac{||X - X^{MF}||}{||X^{MF}||} \leq \frac{||\mathbf{E}\,\mathcal{L}_{uu}||.||(\mathbf{E}\,\mathcal{L}_{uu})^{-1}||}{1 - ||(\mathbf{E}\,\mathcal{L}_{uu})^{-1}||.||\Delta\,\mathcal{L}_{uu}||}\left(\frac{||-\Delta\,\mathcal{L}_{u\ell}.Y_\ell||}{||-\mathbf{E}\,\mathcal{L}_{u\ell}.Y_\ell||} + \frac{||\Delta\,\mathcal{L}_{uu}||}{||\mathbf{E}\,\mathcal{L}_{uu}||}\right).$$

Let us first deal with all the non random quantities. The spectral study of $\mathbf{E}\,\mathcal{L}_{uu}$ is done in the Appendix (Proposition 3). In particular, we have :

$$\Big|\Big|\mathbf{E}\,\mathcal{L}_{uu}\Big|\Big| = \max\big\{|\lambda| : \lambda \in \mathrm{Sp}\big(\mathbf{E}\,\mathcal{L}_{uu}\big)\big\} = 1,$$
$$\Big|\Big|\big(\mathbf{E}\,\mathcal{L}_{uu}\big)^{-1}\Big|\Big| = \frac{1}{\min\big\{|\lambda| : \lambda \in \mathrm{Sp}\big(\mathbf{E}\,\mathcal{L}_{uu}\big)\big\}} = \frac{1}{r}\frac{p+q}{p-q}.$$

Note that since p and q have the same dependency in n (from the assumptions, $p = a\dfrac{\log n}{n}$ and $q = b\dfrac{\log n}{n}$), the ratio $\dfrac{p+q}{p-q}$ does not depend on n, and $\Big|\Big|\big(\mathbf{E}\,\mathcal{L}_{uu}\big)^{-1}\Big|\Big|$ is equal to a constant C'. We are left with the following inequality:

$$\frac{||X - X^{MF}||}{||X^{MF}||} \leq C'\frac{1}{1 - C'\,||\Delta\mathcal{L}_{uu}||}\left(\frac{||\Delta\mathcal{L}_{u\ell}.Y_\ell||}{||\mathbf{E}\,\mathcal{L}_{u\ell}.Y_\ell||} + ||\Delta\mathcal{L}_{uu}||\right).$$

Moreover $\mathbf{E}\,\mathcal{L}_{u\ell}.Y_\ell = (1-r)Y_u$, thus $||\mathbf{E}\,\mathcal{L}_{u\ell}.Y_\ell|| = (1-r)\sqrt{(1-r)n}$. So

$$\frac{||X - X^{MF}||}{||X^{MF}||} \leq \frac{C'}{1 - C'\,||\Delta\,\mathcal{L}_{uu}||}\left(\frac{||\Delta\,\mathcal{L}_{u\ell}||}{1-r} + ||\Delta\,\mathcal{L}_{uu}||\right), \tag{8}$$

where we used $||Y_\ell|| = \sqrt{rn}$ (since Y_ℓ is a vector of size rn with entries equal to 1 or -1) and $\sqrt{\dfrac{r}{1-r}} \leq 1$.

The concentration of the normalized Laplacian towards its mean field value has been established in [12]. In particular, the authors showed that w.h.p.

$$\Big|\Big|\mathcal{L} - \mathbf{E}\,\mathcal{L}\Big|\Big| = O\Big(\frac{1}{\sqrt{d}}\Big), \tag{9}$$

where d is the average degree, when $d = \Omega(\log n)$. However, the result of equation (9) is a concentration of the full normalized Laplacian (an $n \times n$ matrix), while here we are interested in concentration of an extracted matrix. Fortunately, concentration still holds, see Proposition 4 in the Appendix. Therefore, the terms $||\Delta\, \mathcal{L}_{uu}||$ and $||\Delta\, \mathcal{L}_{u\ell}||$ in equation (8) can be bounded by $\dfrac{K}{\sqrt{d}}$.

Last, C' being constant and $||\Delta\, \mathcal{L}_{uu}||$ going to zero, we can lower bound the term $\dfrac{C'}{1 - C'\,||\Delta\, \mathcal{L}_{uu}||}$ by $2C'$ for n large enough, leaving us only with

$$\frac{||X - X^{MF}||}{||X^{MF}||} \leq \frac{C}{\sqrt{d}}$$

for a constant C. This ends the proof.

□

Inequality (7) indicates a slow convergence. For example, in the moderately sparse regime where $p(n)$ and $q(n)$ grows as a constant times $\dfrac{\log(n)}{n}$ (an interesting regime to study for SBM), we have established a bound on the convergence rate in the order of $\dfrac{1}{\sqrt{\log n}}$.

3.3 Asymptotically almost exact recovery for SBM

Proof (of Theorem 1). We just established a concentration inequality for X towards X^{MF}. In order to correctly classify a node i, one should hope that the node's value X_i is close enough to its mean field value X_i^{MF}. To be more precise, $|X_i - X_i^{MF}|$ should be smaller than half the community gap. Recall that in the symmetric case, we showed in Proposition 1 that the community gap is equal to $r\dfrac{p-q}{2q + r(p-q)}$, independent of n when p and q have the same dependency on n.

This leads us to define the notion of 'ϵ-bad nodes'. A node $i \in \{1, \ldots, n\}$ is said to be ϵ-bad if $|X_i - X_i^{MF}| > \epsilon$. Let us denote by B_ϵ the set of ϵ-bad nodes. The nodes that are not ϵ-bad, for an ϵ constant strictly smaller than half the community gap, are almost surely correctly classified.

From $||X - X^{MF}||^2 \geq \sum\limits_{i \in B_\epsilon} |X_i - X_i^{MF}|^2$, it comes that $||X - X^{MF}||^2 \geq |B_\epsilon| \times \epsilon^2$. Thus, using Theorem 3, we have w.h.p.:

$$|B_\epsilon| \leq \frac{C}{\epsilon^2}\frac{n}{d}. \tag{10}$$

If we take for ϵ a constant strictly smaller than half the community gap (recall that the community gap does not depend on n), then all nodes that are not in B_ϵ will be correctly classified. Since by (10) we have $|B_\epsilon| = o(n)$, the fraction of misclassified nodes is at most of order $o(1)$. This establishes almost exact recovery, and the proof of Theorem 1 is completed. □

4 Discussion and future works

In this paper, we explicitly showed that Label Spreading can achieve good result, in the sense of almost exact recovery, for community detection on SBM graphs. Our result stands in the case of two symmetric communities, but extension could be done for more than two non-symmetric communities, as well as labeled nodes non uniformly distributed across communities.

The case of sub-linear number of labeled nodes is worthy of further investigation. As was noted in [13], semi-supervised methods like Label Spreading tend to fail in the limit of small labeled data. Indeed, the minimization scheme (2)-(3) rely too heavily on the condition $X_\ell = Y_\ell$ and not enough on the graph structure. For example, in the extreme case where r is equal to zero, then the solution X have all entries equal, and recovery is not possible. But in that case, we should aim to recover the solution given by unsupervised Spectral Clustering method. Such modified versions of Label Spreading could be part of future research, and should greatly improve the results (at least in the limit of r going to zero).

In particular, we could see if such improved methods could achieve exact recovery under weaker conditions than unsupervised methods. It was shown that unsupervised methods can recover the exact community structure of SBM when $p = a\dfrac{\log n}{n}$ and $q = b\dfrac{\log n}{n}$ if and only if $\dfrac{a+b}{2} > 1 + \sqrt{ab}$. Since $\dfrac{a+b}{2} > 1$ is the connectivity requirement for a symmetric SBM, we can see that connectivity is required (as expected), but not sufficient. Lowering this bound in the semi-supervised scenario, and be able to remove this $\sqrt{ab}$ oversampling factor, would be an interesting result, as we would have exact recovery with semi-supervised setting if and only if the SBM graph is connected.

Acknowledgements. This work has been done within the project of Inria – Nokia Bell Labs "Distributed Learning and Control for Network Analysis".

References

1. Abbe, E.: Community detection and stochastic block models. Foundations and Trends® in Communications and Information Theory **14**(1-2), 1–162 (2018)
2. Avrachenkov, K., Gonçalves, P., Mishenin, A., Sokol, M.: Generalized optimization framework for graph-based semi-supervised learning. SIAM International Conference on Data Mining (SDM'12) (2012)
3. Avrachenkov, K., Kadavankandy, A., Litvak, N.: Mean field analysis of personalized pagerank with implications for local graph clustering. Journal of Statistical Physics **173**(3-4), 895–916 (2018)
4. Avrachenkov, K.E., Filar, J.A., Howlett, P.G.: Analytic perturbation theory and its applications, vol. 135. SIAM (2013)
5. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised Learning. Adaptive computation and machine learning, MIT Press (2006)
6. Condon, A., Karp, R.M.: Algorithms for graph partitioning on the planted partition model. In: Hochbaum, D.S., Jansen, K., Rolim, J.D.P., Sinclair, A. (eds.) Randomization, Approximation, and Combinatorial Optimization. Algorithms and Techniques. pp. 221–232. Springer Berlin Heidelberg, Berlin, Heidelberg (1999)

7. Erdős, P., Rényi, A.: On random graphs. Publicationes Mathematicae (Debrecen) **6**, 290–297 (1959)
8. Gilbert, E.N.: Random graphs. Ann. Math. Statist. **30**(4), 1141–1144 (Dec 1959)
9. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: First steps. Social Networks **5**(2), 109 – 137 (1983)
10. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, 2 edn. (2012)
11. Johnson, R., Zhang, T.: On the effectiveness of laplacian normalization for graph semi-supervised learning. Journal of Machine Learning Research **8**(Jul), 1489–1517 (2007)
12. Le, C.M., Levina, E., Vershynin, R.: Concentration and regularization of random graphs. Random Structures & Algorithms **51**(3), 538–561 (2017)
13. Mai, X., Couillet, R.: A random matrix analysis and improvement of semi-supervised learning for large dimensional data. Journal of Machine Learning Research **19**(1), 3074–3100 (Jan 2018)
14. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in neural information processing systems. pp. 321–328 (2004)
15. Zhu, X.: Semi-supervised learning literature survey. Computer Science Department Technical Report, University of Wisconsin-Madison (2006)
16. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML (2003)

A Background results on matrix analysis

A.1 Inversion of the identity matrix minus a rank 2 matrix

Lemma 1 (Sherman-Morrison-Woodbury formula). *Let A be an invertible $n \times n$ matrix, and B, C, D matrices of correct sizes. Then :* $\Big(A+BCD\Big)^{-1} = A^{-1} - A^{-1}B\Big(I+CDA^{-1}B\Big)^{-1}CDA^{-1}$. *In particular, if u, v are two column vectors of size $n \times 1$, we have :* $\Big(A+uv^T\Big)^{-1} = A^{-1} - \dfrac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$.

Proof. See for example [10], section 0.7.4. □

Lemma 2. *Let* $M = \begin{pmatrix} aJ_{n_1} & bJ_{n_1n_2} \\ cJ_{n_2n_1} & dJ_{n_2} \end{pmatrix}$ *for some values a, b, c, d. Let $n = n_1+n_2$. If $I_n - M$ is invertible, we have :*

$$(I-M)^{-1} = I_n - \frac{1}{K}\begin{pmatrix} \big(-a+n_2(ad-bc)\big)J_{n_1} & -bJ_{n_1n_2} \\ -cJ_{n_2n_1} & \big(-d+n_1(ad-bc)\big)J_{n_2} \end{pmatrix}$$

where $K = (1-n_1a)(1-n_2d) - n_1n_2bc$.

Proof. We will use the Sherman-Morrison-Woodbury matrix identity (Lemma 1) with $A = I_n$, $D = \begin{pmatrix} 1\dots1;0\dots0 \\ 0\dots0;1\dots1 \end{pmatrix}$ (on the first line, there are n_1 ones and n_2 zeros), $B = D^T$ and $C = \begin{pmatrix} -a & -b \\ -c & -d \end{pmatrix}$. We can easily verify that $BCD = -M$.

$$
\begin{aligned}
(I-M)^{-1} &= I_n - B(I+CDB)^{-1}CD \\
&= I_n - B\begin{pmatrix} 1-n_1a & -n_2b \\ -n_1c & 1-n_2d \end{pmatrix}^{-1} CD \\
&= I_n - B\frac{1}{(1-n_1a)(1-n_2d)-n_1n_2bc}\begin{pmatrix} 1-n_2d & n_2b \\ n_1c & 1-n_1a \end{pmatrix} CD \\
&= I_n - \frac{1}{K}B\begin{pmatrix} -a+n_2(ad-bc) & -b \\ -c & -d+n_1(ad-bc) \end{pmatrix} D \\
&= I_n - \frac{1}{K}\begin{pmatrix} \big(-a+n_2(ad-bc)\big)J_{n_1} & -bJ_{n_1n_2} \\ -cJ_{n_2n_1} & \big(-d+n_1(ad-bc)\big)J_{n_2} \end{pmatrix}.
\end{aligned}
$$

□

A.2 Spectral study of a rank 2 matrix

Lemma 3 (Schur's determinant identity, [10]). *Let A, D and $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ be squared matrices. If A is invertible, we have :*

$$\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A)\det(D-CA^{-1}B).$$

Proof. Follows from the formula $\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & I_q \end{pmatrix}\begin{pmatrix} I_p & A^{-1}B \\ 0 & D-CA^{-1}B \end{pmatrix}$. □

Lemma 4 (Matrix determinant lemma, [10]). *For an invertible matrix A and two column vectors u and v, we have* $\det(A+uv^T) = (1+v^TA^{-1}u)\det(A)$.

Lemma 5. *Let α and β be two constants. When $M = \alpha I_n + \beta J$ where J is the $n \times n$ matrix with all entries equal to one, we have* $\det M = \alpha^{n-1}(\alpha+\beta n)$.

Proof. Suppose that $\alpha \neq 0$. Then with $v^T = (1,\dots,1)$ and $u = \beta(1,\dots,1)$ vectors of size $1 \times n$, Lemma 4 gives us

$$
\begin{aligned}
\det M &= \det(\alpha I_n)\Big(1+v^T(\alpha I_n)^{-1}u\Big) \\
&= \alpha^n\Big(1+\frac{\beta n}{\alpha}\Big) \\
&= \alpha^{n-1}(\alpha+\beta n),
\end{aligned}
$$

which proves the lemma for $\alpha \neq 0$. To treat the case $\alpha = 0$, see that the function $\alpha \in \mathbf{R} \mapsto \det(\alpha I_n + \beta J)$ is continuous (even analytic) [4], thus by continuous prolongation in $\alpha = 0$, the expression $\alpha^{n-1}(\alpha+\beta n)$ holds for any $\alpha \in \mathbf{R}$. □

Proposition 2. *Let $M = \begin{pmatrix} aJ_{n_1} & bJ_{n_1n_2} \\ cJ_{n_2n_1} & dJ_{n_2} \end{pmatrix}$ for some values a, b, c, d. The eigenvalues of M are :*

– 0 *with multiplicity* $n_1 + n_2 - 2$*;*
– $\lambda_\pm = \frac{1}{2}\big(n_1 a + n_2 d \pm \sqrt{\Delta}\big)$ *where* $\Delta = (n_1 a - n_2 d)^2 + 4 n_1 n_2 bc$.

Proof. The matrix being of rank 2 (except for some degenerate cases), the fact that 0 is an eigenvalue of multiplicity $n_1 + n_2 - 2$ is obvious. By an explicit computation of the characteristic polynomial of M, the two remaining eigenvalues will be given as roots of a polynomial of degree 2.

Let $\lambda \in \mathbf{R}$ and $A := \lambda I_{n_1} - a J_{n_1}$. If $\lambda \notin \{0; a n_1\}$, then A is invertible and by the Schur's determinant identity (Lemma 3) we have

$$\begin{aligned}\det(\lambda I_n - M) &= \det A \ \det\Big(\lambda I_{n_2} - d J_{n_2} - c J_{n_2 n_1} A^{-1} b J_{n_1 n_2}\Big) \\ &= \det A \ \det B.\end{aligned}$$

From Lemma 5, it follows that $\det A = \lambda^{n_1 - 1}\big(\lambda - n_1 a\big)$.

Let us now compute $\det B$. First, we show that $A^{-1} = \frac{1}{\lambda}\Big(I_{n_1} + \frac{a}{\lambda - a n_1} J_{n_1}\Big)$. Indeed, from the Sherman-Morrison-Woodbury formula (Lemma 1) with $u = -a 1_{n_1}$ and $v = 1_{n_1}$, it follows that

$$\begin{aligned}\Big(\lambda I_{n_1} - a J_{n_1}\Big)^{-1} &= \frac{1}{\lambda} I_{n_1} - \frac{1}{\lambda^2} \frac{-a J_{n_1}}{1 + \frac{-a n_1}{\lambda}} \\ &= \frac{1}{\lambda} I_{n_1} + \frac{1}{\lambda} \frac{a}{\lambda - a n_1} J_{n_1},\end{aligned}$$

which gives the desired expression. Thus,

$$\begin{aligned}B &= \lambda I_{n_2} - d J_{n_2} - \frac{bc}{\lambda} J_{n_2 n_1}\big(I_{n_1} + \frac{a}{\lambda - a n_1} J_{n_1}\big) J_{n_1 n_2} \\ &= \lambda I_{n_2} - d J_{n_2} - \frac{bc}{\lambda}\big(n_1 + \frac{a\, n_1^2}{\lambda - a n_1}\big) J_{n_2} \\ &= \lambda I_{n_2} + \Big(- d - \frac{bc n_1}{\lambda - a n_1}\Big) J_{n_2}.\end{aligned}$$

Again, this matrix is of the form $\lambda I_n + \beta J_n$, and we can use Lemma 5 to show that

$$\det B = \lambda^{n_2 - 1}\Big(\lambda + n_2 \beta\Big).$$

Now we can finish the computation of $\det(\lambda I_n - M)$

$$\begin{aligned}\det(\lambda I_n - M) &= \lambda^{n_1 + n_2 - 2}\big(\lambda - n_1 a\big)\Big(\lambda - n_2 d - \frac{bc n_1 n_2}{\lambda - a n_1}\Big) \\ &= \lambda^{n_1 + n_2 - 2}\Big(\lambda^2 + \lambda(-n_1 a - n_2 d) + n_1 n_2 (ad - bc)\Big).\end{aligned}$$

The discriminant of this second degree polynomial expression is given by

$$\begin{aligned}\Delta &= (n_1 a + n_2 d)^2 - 4n_1 n_2 (ad - bc) \\ &= (n_1 a - n_2 d)^2 + 4n_1 n_2 bc.\end{aligned}$$

Thus $\Delta \geq 0$ and the two remaining eigenvalues are given by

$$\lambda_{\pm} = \frac{1}{2}\big(n_1 a + n_2 d \pm \sqrt{\Delta}\big).$$

□

A.3 Spectral study of $\mathbf{E}\mathcal{L}$

Proposition 3 (Eigenvalues of $\mathbf{E}\mathcal{L}_{uu}$, symmetric case). *Assume two communities of equal size, with $p_1 = p_2 (= p)$. The two smallest eigenvalues of $\mathbf{E}\mathcal{L}_{uu}$ are :*

$$\lambda_1 = r\,\frac{p-q}{p+q} \quad \textit{and} \quad \lambda_2 = r.$$

Note that the other eigenvalue of $\mathbf{E}\mathcal{L}_{uu}$ is one (with multiplicity $\lfloor (1-r)n \rfloor - 2$).

Proof. The matrix $\mathbf{E}\mathcal{L}_{uu}$ can be written as $I - M$, where $M = D^{-1/2} A D^{-1/2}$ has a block form like in Proposition 2, with coefficients $a = \dfrac{p_1}{d_1}$, $b = c = \dfrac{q}{\sqrt{d_1 d_2}}$ and $d = \dfrac{p_2}{d_2}$. Note that the blocks sizes are now $\lfloor (1-r)n_i \rfloor$ and not n_i. Under the symmetric assumption, we have $d_1 = d_2 = \dfrac{n}{2}(p+q)$.

Moreover, λ_M is an eigenvalue of M if and only if $1 - \lambda_M$ is eigenvalue of $\mathbf{E}\mathcal{L}_{uu}$. Using the notations of Proposition 2, we have $\Delta = 4(1-r)^2 \dfrac{q^2}{(p+q)^2}$, and the two non-zero eigenvalues of M are given by:

$$\begin{aligned}\lambda_{\pm} &= \frac{1}{2}\Big(2(1-r)\frac{p}{p+q} \pm 2(1-r)\frac{q}{p+q}\Big) \\ &= 1 - r\frac{p \pm q}{p+q}.\end{aligned}$$

□

B Spectral norm of an extracted matrix

Proposition 4. *Let A be a matrix and B an extracted matrix (non necessarily squared: we can remove rows and columns with different indices, and potentially more rows than columns, or vice versa) from A, then : $||B||_2 \leq ||A||_2$.*

Proof. For two subsets I and J of $\{1, \dots, n\}$, let $B = A_{IJ}$ the matrix obtained from A by keeping only the rows (resp. columns) in I (resp. in J). Then $B = M_1 A M_2$ where M_1 and M_2 are two appropriately chosen permutation matrices. Thus their spectral norm is equal to one, and the result $||B||_2 \leq ||A||_2$ follows from the inequality $||B||_2 \leq ||M_1||_2 ||A||_2 ||M_2||_2$. □