

RESEARCH ARTICLE

# Almost exact recovery in noisy semi-supervised learning

Konstantin Avrachenkov<sup>1</sup>  and Maximilien Dreveton<sup>2</sup> 

<sup>1</sup>Inria Sophia Antipolis, 2004 Rte des Lucioles, Valbonne, France

<sup>2</sup>School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

**Corresponding author:** Maximilien Dreveton; Email: [maximilien.dreveton@epfl.ch](mailto:maximilien.dreveton@epfl.ch)

**Keywords:** Degree corrected stochastic block model; Graph-based method; Graph clustering; Noisy data; Semi-supervised learning

**MSC:** Primary 62F12; Secondary 62H30; 68T10

## Abstract:

Graph-based semi-supervised learning methods combine the graph structure and labeled data to classify unlabeled data. In this work, we study the effect of a noisy oracle on classification. In particular, we derive the maximum a posteriori (MAP) estimator for clustering a degree corrected stochastic block model when a noisy oracle reveals a fraction of the labels. We then propose an algorithm derived from a continuous relaxation of the MAP, and we establish its consistency. Numerical experiments show that our approach achieves promising performance on synthetic and real data sets, even in the case of very noisy labeled data.

## 1. Introduction

Semi-supervised learning (SSL) aims at achieving superior learning performance by combining unlabeled and labeled data. Since typically the amount of unlabeled data is large compared to the amount of labeled data, SSL methods are relevant when the performance of unsupervised learning is low, or when the cost of getting a large amount of labeled data for supervised learning is too high. Unfortunately, many standard SSL methods have been shown to not efficiently use the unlabeled data, leading to unsatisfactory or unstable performance [11, Chap. 4], [9, 12]. Moreover, noise in the labeled data can further degrade the performance. In practice, the noise can come from a tired or non-diligent expert carrying out the labeling task or even from adversarial data corruption.

In this paper, we investigate the problem of graph clustering, where one aims to group the nodes of a graph into different classes. Our working model is the two-class degree corrected stochastic block model (DC-SBM), with side information on some node's community assignment given by a noisy oracle. The DC-SBM was introduced in [18] to account for degree heterogeneity and block structure. Let  $n$  be the number of nodes. Each node  $i \in [n]$  is given a community label  $Z_i \in \{-1, 1\}$  chosen uniformly at random and a parameter  $\theta_i > 0$ . Given  $Z = (Z_1, \dots, Z_n)$  and  $\theta = (\theta_1, \dots, \theta_n)$ , an undirected edge is added between nodes  $i$  and  $j$  with probability  $\min(1, \theta_i \theta_j p_{\text{in}})$ , if  $Z_i = Z_j$ , and with probability  $\min(1, \theta_i \theta_j p_{\text{out}})$ , otherwise. This model reduces to the standard stochastic block model (SBM) [1] if  $\theta_i = 1$  for every node  $i$ . The unsupervised clustering problem consists of inferring the latent community structure  $Z$  given one observation of a DC-SBM graph. We make the problem semi-supervised by introducing a noisy oracle. For every node, this oracle reveals the correct community label with probability  $\eta_1$ , a wrong community label with probability  $\eta_0$ , and reveals nothing with probability  $1 - \eta_1 - \eta_0$ .

We first derive the maximum a posteriori (MAP) estimator for SSL-clustering in a DC-SBM graph given the *a priori* information induced by a noisy oracle and graph structure. We note that, despite

its simplicity, this result did not appear previously in the literature, neither for a perfect oracle nor for SBM. In particular, we show that the MAP is the solution to a minimization problem that involves a trade-off between three factors: a cut-based term (as in the unsupervised scenario), a regularization term (penalizing solutions with unbalanced clusters), and a loss term (penalizing predictions that differ from the oracle information).

As solving the MAP estimator is NP-hard, we propose a continuous relaxation and derive an SSL version of a spectral method based on the adjacency matrix. We establish a bound on the ratio of misclassified nodes for this continuous relaxation, and we show that this ratio goes to zero under the hypothesis that the average degree diverges and an almost perfect oracle (see [Corollary 3.2](#) for a rigorous statement). As a result, the proposed SSL method guarantees almost exact recovery (recovering all but  $o(n)$  labels when  $n$  goes to infinity) even when a part of the side information is incorrect. We note that even though we work with the case of two clusters, most of our results are extendable to the setting of more than two clusters at the expense of more cumbersome notations.

One can make several parallels between our continuous relaxation and state-of-the-art techniques. Indeed, SSL-clustering often relies on minimization frameworks (see [\[5, 11\]](#) for an overview). The idea of minimizing a well-chosen energy function was proposed in [\[30\]](#), under the constraint of keeping the labeled nodes' predictions equal to the oracle labels. As we show in the numerical section, this hard constraint is unsuitable if the oracle reveals false information. Consequently, Belkin *et al.* [\[8\]](#) introduced an additional loss term in the energy function to allow the prediction to differ from the oracle information. We recover this loss term with an additional theoretical justification because it comes from a relaxation of the MAP.

Moreover, the regularization term is necessary to prevent the solution from being flat and making classification rely on second-order fluctuations. This phenomenon was previously observed by [\[23\]](#) in the limit of an infinite amount of unlabeled data, as well as by [\[21\]](#) in the large dimension limit. The regularization term here consists of subtracting a constant term from all the entries of the adjacency matrix. It resembles previous regularization techniques, like the centering of the adjacency matrix proposed in [\[22\]](#). However, contrary to [\[22\]](#), we study a noisy framework without assuming a large-dimension asymptotic regime. Moreover, we solve exactly the relaxed minimization problem instead of giving a heuristic with an extra parameter.

It was shown in [\[25\]](#) that even with a perfect oracle revealing a constant fraction of the labels, the phase transition phenomena for exact recovery in SBM (recovering all the correct labels with high probability) remains unchanged. Thus, for the exact recovery problem, one could discard all the side information and simply use unsupervised algorithms when the number of data points goes to infinity. Of course, wasting potentially valuable information is not entirely satisfactory. Thus, in the present work, we consider the case of almost exact recovery and an oracle with noisy information. In [\[7, 17\]](#) criteria different from the exact recovery have also been considered in the framework of SSL.

The paper is structured as follows. We introduce the model and main notations in [Section 2](#), along with the derivation of the MAP estimator ([Section 2.2](#)). A continuous relaxation of the MAP is presented in [Section 3](#) as well as the guarantee of its convergence to the true community structure (Subsection 3.2). We postpone some proofs to the Appendix and leave in the main text only those we consider important to the material exposition. We conclude the paper with numerical results ([Section 4](#)), emphasizing the effect of the noise on the clustering accuracy. In particular, we outperform state-of-the-art graph-based SSL methods in a difficult regime (few label points or large noise).

Lastly, the present paper is a follow-up work on [\[3\]](#). However, there are very important developments. In [\[3\]](#) we have only established almost exact recovery on SBM for Label Spreading [\[29\]](#) heuristic algorithm with a linear number of labeled nodes (see [\[3, Assumption 3\]](#)). In the present work, we extend the analysis to DC-SBM, investigate the effect of noisy labeled data, and allow a potentially sublinear number of labeled nodes. We also add experiments with real and synthetic data that illustrate our theoretical results.

## 2. MAP estimator in a noisy semi-supervised setting

### 2.1. Problem formulation and notations

A homogeneous DC-SBM is parametrized by the number of nodes  $n$ , two class-affinity parameters  $p_{in}, p_{out}$ , and a pair  $(\theta, Z)$  where  $\theta \in \mathbf{R}^n$  is a vector of intrinsic connection intensities and  $Z \in \{-1, 1\}^n$  is the community labeling vector. Given  $(p_{in}, p_{out}, \theta, Z)$ , the graph adjacency matrix  $A = (a_{ij})$  is generated as

$$A_{ij} = A_{ji} \sim \begin{cases} \text{Ber}(\theta_i \theta_j p_{in}), & \text{if } Z_i = Z_j, \\ \text{Ber}(\theta_i \theta_j p_{out}), & \text{otherwise,} \end{cases} \tag{2.1}$$

for  $i \neq j$ , and  $A_{ii} = 0$ . We assume throughout the paper that  $Z_i \sim \text{Uni}(\{-1, 1\})$ , and that the entries of  $\theta$  are independent random variables satisfying  $\theta_i \in [\theta_{min}, \theta_{max}]$  with  $\mathbb{E}\theta_i = 1$ ,  $\theta_{min} > 0$ , and  $\theta_{max}^2 \max(p_{in}, p_{out}) \leq 1$ . In particular, when all the  $\theta_i$ 's are equal to one, the model reduces to the SBM:

$$A_{ij} = A_{ji} \sim \begin{cases} \text{Ber}(p_{in}), & \text{if } Z_i = Z_j, \\ \text{Ber}(p_{out}), & \text{otherwise.} \end{cases} \tag{2.2}$$

In addition to the observation of the graph adjacency matrix  $A$ , an oracle gives us extra information about the cluster assignment of some nodes. This can be represented as a vector  $s \in \{0, -1, 1\}^n$ , whose entries  $s_i$  are independent and distributed as follows:

$$s_i = \begin{cases} Z_i, & \text{with probability } \eta_1, \\ -Z_i, & \text{with probability } \eta_0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

In other words, the oracle (2.3) reveals the correct cluster assignment of node  $i$  with probability  $\eta_1$  and gives a false cluster assignment with probability  $\eta_0$ . It reveals nothing with probability  $1 - \eta_1 - \eta_0$ . The quantity  $\mathbb{P}(s_i \neq Z_i \mid s_i \neq 0)$  is the rate of mistakes of the oracle (i.e., the probability that the oracle reveals false information given that it reveals something), and is equal to  $\eta_0 / (\eta_1 + \eta_0)$ . The oracle is informative if this quantity is less than 1/2, which is equivalent to  $\eta_1 > \eta_0$ . In the following, we will always assume that the oracle is informative.

**Assumption 2.1.** *The oracle is informative, that is,  $\eta_1 > \eta_0$ .*

Given the observation of  $A$  and  $s$ , the goal of clustering is to recover the community labeling vector  $Z$ . For an estimator  $\widehat{Z} \in \{-1, 1\}^n$  of  $Z$ , the relative error is defined as the proportion of misclassified nodes

$$L(\widehat{Z}, Z) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(\widehat{Z}_i \neq Z_i). \tag{2.4}$$

Note that, unlike unsupervised clustering, we do not take a minimum over the permutations of the predicted labels since we should be able to learn the correct community labels from the informative oracle.

*Notations* Given an oracle  $s$ , we let  $\ell$  be the set of labeled nodes, that is  $\ell := \{i \in V : s_i \neq 0\}$ , and denote  $\mathcal{P}$  the diagonal matrix with entries  $(\mathcal{P})_{ii} = 1$ , if  $i \in \ell$ , and  $(\mathcal{P})_{ii} = 0$ , otherwise.

The notation  $I_n$  stands for the identity matrix of size  $n \times n$ , and  $\mathbf{1}_n$  (resp.,  $\mathbf{0}_n$ ) is the vector of size  $n \times 1$  of all ones (resp., of all zeros).

For any matrix  $A = (a_{ij})_{i \in [n], j \in [m]}$  and two sets  $S \subset [n], T \subset [m]$ , we denote  $A_{S,T} = (a_{ij})_{i \in S, j \in T}$  the matrix obtained from  $A$  by keeping elements, whose row indices are in  $S$  and column indices are in  $T$ . We denote by  $\|x\|$  the Euclidean norm of a vector  $x$  and by  $\|A\|$  the spectral norm of a matrix  $A \in \mathbf{R}^{n \times m}$ . Finally,  $A \odot B$  refers to the entry-wise matrix product between two matrices  $A$  and  $B$  of the same size.

**2.2. MAP estimator for semi-supervised recovery in DC-SBM**

Given a realization of a DC-SBM graph adjacency matrix  $A$  and the oracle information  $s$ , the MAP estimator is defined as

$$\widehat{Z}^{\text{MAP}} = \arg \max_{z \in \{-1,1\}^n} \mathbb{P}(z | A, s). \tag{2.5}$$

This estimator is known to be optimal (in the sense that if it fails then any other estimator would also fail, see, e.g., [16]) for the exact recovery of all the community labels. **Theorem 2.2** provides an expression of the MAP.

**Theorem 2.2** *Let  $G$  be a graph drawn from DC-SBM as defined in (2.1) and  $s$  be the oracle information as defined in (2.3). Denote  $M = (F_1 - F_0) \odot A + F_0$ , where  $F_0 = (f_{ij}^{(0)})$  and  $F_1 = (f_{ij}^{(1)})$  such that  $f_{ij}^{(a)} = \log \frac{\mathbb{P}(A_{ij}=a | z_i=z_j)}{\mathbb{P}(A_{ij}=a | z_i \neq z_j)}$  for  $a \in \{0, 1\}$ . The MAP estimator defined in (2.5) is given by*

$$\widehat{Z}^{\text{MAP}} = \arg \min_{z \in \{-1,1\}^n} \left( z^T M z + \log \left( \frac{\eta_1}{\eta_0} \right) \| \mathcal{P}z - s \|^2 \right). \tag{2.6}$$

For a perfect oracle ( $\eta_0 = 0$ ) this reduces to

$$\widehat{Z}^{\text{MAP}} = \arg \min_{\substack{z \in \{-1,1\}^n \\ z_\ell = s_\ell}} z^T M z. \tag{2.7}$$

The proof of **Theorem 2.2** is standard and postponed to **Appendix A**. We note that, despite being *a priori* standard, this result did not appear previously in the literature (neither for the standard SBM nor for the perfect oracle).

The minimization problem (2.6) consists of a trade-off between minimizing a quadratic function  $z^T M z$  and a penalty term. This trade-off reads as follows: for each labeled node such that the prediction contradicts the oracle, a penalty  $\log \left( \frac{\eta_1}{\eta_0} \right) > 0$  is added. In particular, when the oracle is uninformative, that is  $\eta_1 = \eta_0$ , then this term is null, and Expression (2.6) reduces to the MAP for unsupervised clustering.

The following **Corollary 2.3**, whose proof is in **Appendix A**, provides the expression of the MAP estimator for a standard SBM.

**Corollary 2.3.** *The MAP estimator for semi-supervised clustering on SBM graph with  $p_{\text{in}} > p_{\text{out}}$  and with an oracle  $s$  defined in (2.3) is given by*

$$\widehat{Z}^{\text{MAP}} = \arg \min_{z \in \{-1,1\}^n} \left( -z^T \left( A - \tau 1_n 1_n^T \right) z + \lambda^* \| \mathcal{P}z - s \|^2 \right), \tag{2.8}$$

where  $\tau = \frac{\log \left( \frac{1 - p_{\text{out}}}{1 - p_{\text{in}}} \right)}{\log \left( \frac{p_{\text{in}}(1 - p_{\text{out}})}{p_{\text{out}}(1 - p_{\text{in}})} \right)}$  and  $\lambda^* = \frac{\log \left( \frac{\eta_1}{\eta_0} \right)}{\log \left( \frac{p_{\text{in}}(1 - p_{\text{out}})}{p_{\text{out}}(1 - p_{\text{in}})} \right)}$ . For the perfect oracle, this reduces to

$$\widehat{Z}^{\text{MAP}} = \arg \min_{\substack{z \in \{-1,1\}^n \\ z_\ell = s_\ell}} z^T \left( -A + \tau 1_n 1_n^T \right) z. \tag{2.9}$$

### 3. Almost exact recovery using a continuous relaxation

As finding the MAP estimate is NP-hard [26], we perform a continuous relaxation (Section 3.1). We then give an upper bound on the number of misclassified nodes in Section 3.2.

#### 3.1. Continuous relaxation of the MAP

For the sake of presentation simplicity, we focus on the MAP for SBM, that is, minimization problem (2.8). We perform a continuous relaxation mirroring what is commonly done for spectral methods [24], namely

$$\widehat{X} = \arg \min_{\substack{x \in \mathbf{R}^n \\ \sum_i \kappa_i x_i^2 = \sum_i \kappa_i}} \left( -x^T A_\tau x + \lambda (s - \mathcal{P}x)^T (s - \mathcal{P}x) \right), \tag{3.1}$$

where  $A_\tau = A - \tau 1_n 1_n^T$  and  $\kappa = (\kappa_1, \dots, \kappa_n)$  is a vector of positive entries. We choose to constrain  $x$  on the hyper-sphere  $\|x\|^2 = n$  by letting  $\kappa_i = 1$ , but other choices would lead to a similar analysis. In particular, in the numerical Section 4 we will compare this choice with a degree-normalization approach (i.e.,  $\kappa_i = d_i$ ).

We further note that for the perfect oracle, the corresponding relaxation of (2.9) is

$$\widehat{X} = \arg \min_{\substack{x \in \mathbf{R}^n \\ x_\ell = s_\ell \\ \|x\|^2 = n}} \left( -x^T A_\tau x \right). \tag{3.2}$$

Given the classification vector  $\widehat{X} \in \mathbf{R}^n$ , node  $i$  is classified into cluster  $\widehat{Z}_i \in \{-1, 1\}$  such that

$$\widehat{Z}_i = \begin{cases} 1, & \text{if } \widehat{X}_i > 0, \\ -1, & \text{otherwise.} \end{cases} \tag{3.3}$$

Let us solve the minimization problem (3.1). By letting  $\gamma \in \mathbf{R}$  be the Lagrange multiplier associated with the constraint  $\|x\|^2 = n$ , the Lagrangian of the optimization problem (3.1) is

$$-x^T A_\tau x + \lambda (s - \mathcal{P}x)^T (s - \mathcal{P}x) - \gamma (x^T x - n).$$

This leads to the *constrained* linear system

$$\begin{cases} (-A_\tau + \lambda \mathcal{P} - \gamma I_n) x = \lambda s, \\ x^T x = n, \end{cases} \tag{3.4}$$

whose unknowns are  $\gamma$  and  $x$ .

While [22] let  $\gamma$  to be a hyper-parameter (hence the norm constraint  $x^T x = n$  is no longer verified), the exact optimal value of  $\gamma$  can be found explicitly following [14]. Firstly, we note that if  $(\gamma_1, x_1)$  and  $(\gamma_2, x_2)$  are solutions of the system (3.4), then (see Lemma D.1 for the derivations)

$$\mathcal{C}(x_1) - \mathcal{C}(x_2) = \frac{\gamma_1 - \gamma_2}{2} \|x_1 - x_2\|^2,$$

**Algorithm 1.** *Semi-supervised learning with regularized adjacency matrix.*

**Input:** Adjacency matrix  $A$ , oracle information  $s$ , parameters  $\tau$  and  $\lambda$ .

**Procedure:**

Let  $\gamma^*$  be the smallest solution of Equation (3.6).

Compute  $\widehat{X}$  as the solution of Equation (3.5).

**for**  $i = 1 \cdots n$  **do**

$\widehat{Z}_i = \text{sign}(\widehat{X}_i)$ .

**end for**

**return**  $\widehat{Z}$ .

where  $\mathcal{C}(x) = -x^T A_\tau x + \lambda(s - \mathcal{P}x)^T (s - \mathcal{P}x)$  is the cost function minimized in (3.1). Hence, among the solution pairs  $(\gamma, x)$  of the system (3.4), the solution of the minimization problem (3.1) is the vector  $x$  associated with the smallest  $\gamma$ .

Secondly, the eigenvalue decomposition of  $-A_\tau + \lambda\mathcal{P}$  reads as

$$-A_\tau + \lambda\mathcal{P} = Q\Delta Q^T,$$

where  $\Delta = \text{diag}(\delta_1, \dots, \delta_n)$  with  $\delta_1 \leq \dots \leq \delta_n$  and  $Q^T Q = I_n$ . Therefore, after the change of variables  $u = Q^T x$  and  $b = \lambda Q^T s$ , the system (3.4) is transformed to

$$\begin{cases} \Delta u &= \gamma u + b, \\ u^T u &= n. \end{cases}$$

Thus, the solution  $\widehat{X}$  of the optimization problem (3.1) satisfies

$$(-A_\tau + \lambda\mathcal{P} - \gamma_* I_n) \widehat{X} = \lambda s, \tag{3.5}$$

where  $\gamma_*$  is the smallest solution of the *explicit secular equation* [14]

$$\sum_{i=1}^n \left( \frac{b_i}{\delta_i - \gamma} \right)^2 - n = 0. \tag{3.6}$$

We summarize this in Algorithm 1. Note that for the sake of generality, we let  $\lambda$  and  $\tau$  be hyper-parameters of the algorithm. If the model parameters are known, we can use the expressions of  $\lambda$  and  $\tau$  derived in Corollary 2.3. The choice of  $\lambda$  and  $\tau$  is further discussed in Section 4. We must use power iterations or Krylov subspace methods to apply Algorithm 1 to large data sets. The main computational bottleneck in those methods will be the matrix-vector product  $A_\tau v$ . The matrix  $A_\tau$  is not sparse. Since  $A_\tau$  is a sum of a sparse matrix and a rank-one matrix, the computation of  $A_\tau v = Av - \tau(1_n^T v)1_n$  can be done efficiently by subtracting the same scalar  $\tau(1_n^T v)$  from all the entries of the result of the sparse matrix-vector multiplication.

### 3.2. Ratio of misclassified nodes

This section gives bounds on the number of unlabeled nodes misclassified by Algorithm 1. We then specialize the results for some particular cases.

**Theorem 3.1** *Consider a DC-SBM with a noisy oracle as defined in (2.1) and (2.3). Let  $\bar{d} = \frac{n}{2}(p_{\text{in}} + p_{\text{out}})$  and  $\bar{\alpha} = \frac{n}{2}(p_{\text{in}} - p_{\text{out}})$ . Suppose that  $\tau > p_{\text{out}}$  and that  $\eta_0 n \sqrt{\eta_1 + \eta_0} \ll \lambda$ , and let  $\widehat{Z}$  be the output of*

Algorithm 1. Then, for any  $r > 0$ , there exists a constant  $C$  such that the proportion of misclassified unlabeled nodes satisfies

$$L(\widehat{Z}_u, Z_u) \leq C \left( \frac{p_{in} + p_{out}}{p_{in} - p_{out}} \right)^2 \left( \frac{\bar{\alpha} + \lambda}{\lambda} \right)^2 \frac{1}{(\eta_1 + \eta_0)(\eta_1 - \eta_0)^2 \bar{d}},$$

with probability at least  $1 - n^{-r}$ .

The value of  $\lambda$  in Theorem 3.1 serves as a hyper-parameter of the algorithm and may not necessarily be equal to the value  $\lambda^*$  computed in Corollary 2.3. Consequently, one can opt for a  $\lambda$  significantly larger than  $\eta_0 n \sqrt{\eta_0 + \eta_1}$ , even if the  $\lambda^*$  from Corollary 2.3 is not much larger than  $\eta_0 n \sqrt{\eta_0 + \eta_1}$ . Selecting  $\lambda > \lambda^*$  indicates an excessive reliance on the information provided by the oracle, but it has a benign effect on the error bound of the unlabeled nodes given in Theorem 3.1.

The core of the proof relies on the concentration of the adjacency matrix toward its expectation. This result, as presented in [19], holds under loose assumptions: it is valid for any random graph whose edges are independent of each other. To use this result for  $\bar{d} = o(\log n)$ , one needs to replace the matrix  $A_\tau$  by  $A'_\tau = A' - \tau 1_n 1_n^T$ , where  $A'$  is the adjacency matrix of the graph obtained after reducing the weights on the edges incident to the high degree vertices. We refer to [19, Sect. 1.4] for more details. This extra technical step is not necessary when  $\bar{d} = \Omega(\log n)$ . Moreover, concentration also occurs if we replace the adjacency matrix with the normalized Laplacian in Eq. (3.5). In that case, we obtain a generalization of the Label Spreading algorithm [29], [11, Chap. 11].

In the following, the mean-field graph refers to the weighted graph formed by the expected adjacency matrix of a DC-SBM graph. Furthermore, we assume without loss of generality that the first  $n/2$  nodes are in the first cluster and the last  $n/2$  are in the second cluster. Therefore,  $\mathbb{E}A = ZBZ^T$  with  $B = \begin{pmatrix} p_{in} & p_{out} \\ p_{out} & p_{in} \end{pmatrix}$  and  $Z = \begin{pmatrix} 1_{n/2} & 0_{n/2} \\ 0_{n/2} & 1_{n/2} \end{pmatrix}$ . In particular, the coefficients  $\theta_i$  disappear because  $\mathbb{E}\theta_i = 1$ . We consider the setting in which the diagonal elements of  $\mathbb{E}A$  are not zeros. This accounts for modifying the definition of DC-SBM, where we can have self-loops with probability  $p_{in}$ . Nevertheless, we could set the diagonal elements of  $\mathbb{E}A$  to zeros and our results would still hold at the expense of cumbersome expressions. Note that the matrix  $\mathbb{E}A$  has two non-zero eigenvalues:  $\bar{d} = n \frac{p_{in} + p_{out}}{2}$  and  $\bar{\alpha} = n \frac{p_{in} - p_{out}}{2}$ .

Proof of Theorem 3.1. We prove the statement in three steps. We first show that the solution  $\widehat{X}$  of the constrained linear system (3.4) is concentrated around the solution  $\bar{x}$  of the same system for the mean-field model. Then, we compute  $\bar{x}$  and show that we can retrieve the correct cluster assignment from it. We finally conclude with the derivation of the bound.

- (i) Similarly to [4] and [3], let us rewrite Eq. (3.5) as a perturbation of a system of linear equations corresponding to the mean-field solution. We thus have

$$(\mathbb{E}\tilde{\mathcal{L}} + \Delta\tilde{\mathcal{L}})(\bar{x} + \Delta x) = \lambda s,$$

where  $\tilde{\mathcal{L}} = -A_\tau + \lambda \mathcal{P} - \gamma_* I_n$ ,  $\Delta x := \widehat{X} - \bar{x}$  and  $\Delta\tilde{\mathcal{L}} := \tilde{\mathcal{L}} - \mathbb{E}\tilde{\mathcal{L}}$ .

We recall that a perturbation of a system of linear equations  $(A + \Delta A)(x + \Delta x) = b$  leads to the following sensitivity inequality (see, e.g., [15, Sect. 5.8]):

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\Delta A\|}{\|A\|},$$

where  $\|\cdot\|$  is the operator norm associated with a vector norm  $\|\cdot\|$  (we use the same notations for simplicity) and  $\kappa(A) := \|A^{-1}\| \cdot \|A\|$  is the condition number. In our case, the above inequality can be rewritten as follows:

$$\frac{\|\widehat{X} - \bar{x}\|}{\|\bar{x}\|} \leq \left\| \left( \mathbb{E} \tilde{\mathcal{L}} \right)^{-1} \right\| \cdot \|\Delta \tilde{\mathcal{L}}\|, \tag{3.7}$$

employing the Euclidean vector norm and spectral operator norm. The spectral study of  $\mathbb{E} \tilde{\mathcal{L}}$  (see [Corollary B.3](#) in [Appendix B.1](#)) gives:

$$\left\| \left( \mathbb{E} \tilde{\mathcal{L}} \right)^{-1} \right\| = \frac{1}{\min \{|\lambda| : \lambda \in \text{Sp}(\mathbb{E} \tilde{\mathcal{L}})\}} = \frac{1}{-t_2^+ - \bar{\gamma}_*},$$

where  $t_2^+$  is defined in [Corollary B.3](#) in [Appendix B.1](#) and  $\bar{\gamma}_*$  is the solution of [Eq. \(3.6\)](#) for the mean-field model. [Lemma B.4](#) in [Appendix B.1.1](#) leads to

$$\left\| \left( \mathbb{E} \tilde{\mathcal{L}} \right)^{-1} \right\| \leq \frac{1}{\lambda + \bar{\alpha}}. \tag{3.8}$$

The last ingredient needed is the concentration of the adjacency matrix around its expectation. We have

$$\left\| \tilde{\mathcal{L}} - \mathbb{E} \tilde{\mathcal{L}} \right\| \leq \|(\gamma_* - \bar{\gamma}_*) I_n\| + \|A - \mathbb{E} A\| \leq |\gamma_* - \bar{\gamma}_*| + \|A - \mathbb{E} A\|.$$

[Proposition B.5](#) in [Appendix B.1.2](#) shows that

$$|\gamma_* - \bar{\gamma}_*| \leq \left( 1 + \frac{(\bar{\alpha} + \lambda)^3}{2\sqrt{\eta_1 + \eta_0}(\eta_1 - \eta_0)\bar{\alpha}^2\lambda} \right) \sqrt{\bar{d}}.$$

Moreover, when  $d = \Omega(\log n)$ , it is shown in [\[13\]](#) that for every  $r > 0$  there exists a constant  $C'$  such that  $\|A - \mathbb{E} A\| \leq C' \sqrt{\bar{d}}$  holds with probability at least  $1 - n^{-r}$ . If  $\bar{d} = o(\log n)$ , the same result holds with a proper preprocessing on  $A$ , and we refer the reader to [\[19\]](#) for more details. We will omit this extra step in the proof to keep notations short. Using this concentration bound, we have

$$\begin{aligned} \left\| \tilde{\mathcal{L}} - \mathbb{E} \tilde{\mathcal{L}} \right\| &\leq \left( C' + \frac{27(\bar{\alpha} + \lambda)^3}{\sqrt{2}\sqrt{\eta_1 + \eta_0}(\eta_1 - \eta_0)\bar{\alpha}^2\lambda} \right) \sqrt{\bar{d}} \\ &\leq \left( C' + \frac{27}{\sqrt{2}} \right) \frac{(\lambda + \bar{\alpha})^3}{\bar{\alpha}^2\lambda} \frac{\sqrt{\bar{d}}}{\sqrt{\eta_1 + \eta_0}(\eta_1 - \eta_0)} \end{aligned}$$

for some constant  $C'$ . Let  $C = C' + \frac{27}{\sqrt{2}}$ . By combining the above with inequality [\(3.8\)](#), the inequality [\(3.7\)](#) becomes

$$\frac{\|\widehat{X} - \bar{x}\|}{\|\bar{x}\|} \leq C \frac{(\lambda + \bar{\alpha})^2}{\bar{\alpha}^2\lambda} \frac{\sqrt{\bar{d}}}{\sqrt{\eta_1 + \eta_0}(\eta_1 - \eta_0)}. \tag{3.9}$$

- (ii) Node  $i$  in the mean-field model is correctly classified by decision rule [\(3.3\)](#) if the sign of  $\bar{x}_i$  equals the sign of  $Z_i$ . [Corollary C.2](#) in [Appendix C](#) shows that this is indeed the case for the unlabeled nodes.



(iii) Finally, for an unlabeled node  $i$  to be correctly classified, the node's value  $\widehat{X}_i$  should be close enough to its mean-field value  $\bar{x}_i$ .

In particular, part (ii) shows that if  $|\widehat{X}_i - \bar{x}_i|$  is smaller than some non-vanishing constant  $\beta$ , then an unlabeled node  $i$  will be correctly classified. An unlabeled node  $i$  is said to be  $\beta$ -bad if  $|\widehat{X}_i - \bar{x}_i| > \beta$ . We denote by  $S_\beta$  the set of  $\beta$ -bad nodes. The nodes that are not  $\beta$ -bad are almost surely correctly classified, and thus  $L(\widehat{Z}_u, Z_u) \leq \frac{|S_\beta|}{n}$ .

From  $\|\widehat{X} - \bar{x}\|^2 \geq \sum_{i \in S_\beta} |\widehat{X}_i - \bar{x}_i|^2$ , it follows that  $\|\widehat{X} - \bar{x}\|^2 \geq |S_\beta| \times \beta^2$ . Thus, using inequality (3.9) and the norm constraint  $\|\bar{x}\|^2 = n$ , we have with probability at least  $1 - n^{-r}$ ,

$$|S_\beta| \leq \frac{1}{\beta^2} \left( \frac{C}{\eta_1 - \eta_0} \frac{\bar{\alpha} + \lambda}{\bar{\alpha}\lambda} \sqrt{\bar{d}} \right)^2 n,$$

for some constant  $C$ . We end the proof by noticing that  $\frac{\bar{d}}{\bar{\alpha}} = \frac{p_{in} + p_{out}}{p_{in} - p_{out}}$ . □

**Corollary 3.2. (Almost exact recovery in the diverging degree regime)** Consider a DC-SBM such that  $\bar{d} \gg 1$ ,  $\frac{p_{in} + p_{out}}{p_{in} - p_{out}} = O(1)$ ,  $\sqrt{\eta_0 + \eta_1}(\eta_1 - \eta_0) \gg \frac{1}{\sqrt{\bar{d}}}$ , and  $\eta_0 n \sqrt{\eta_0 + \eta_1} \ll \lambda$ . Suppose that  $\tau > p_{out}$  and  $\lambda \gtrsim \bar{\alpha}$ . Then, Algorithm 1 correctly classifies almost all the unlabeled nodes.

*Proof.* With the corollary's assumptions  $(\eta_1 - \eta_0)^2 \bar{d} \rightarrow +\infty$  and  $\frac{\bar{\alpha} + \lambda}{\lambda} = O(1)$ , by Theorem 3.1 the fraction of misclassified nodes is of the order  $o(1)$ . □

The quantity  $(\eta_1 - \eta_0)n$  is the expected difference between the number of nodes correctly labeled and the number of nodes wrongly labeled by the oracle. In particular, Corollary 3.2 allows for a sub-linear number of labeled nodes since  $\eta_0$  and  $\eta_1$  can go to zero.

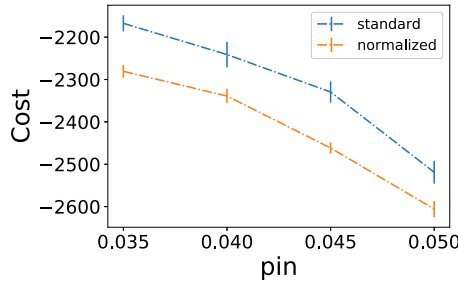
**Corollary 3.3. (Detection in the constant degree regime)** Consider a DC-SBM such that  $p_{in} = \frac{c_{in}}{n}$  and  $p_{out} = \frac{c_{out}}{n}$  where  $c_{in}, c_{out}$  are constants. Suppose that  $\sqrt{\eta_0 + \eta_1}(\eta_1 - \eta_0)$  is a non-zero constant, and let  $\tau > 2p_{out}$  and  $\lambda \gtrsim 1$ . Assume furthermore that  $\eta_0 n \sqrt{\eta_0 + \eta_1} \ll \lambda$ . Then, for  $\frac{(c_{in} - c_{out})^2}{c_{in} + c_{out}}$  bigger than some constant, w.h.p. Algorithm 1 performs better than a random guess.

*Proof.* According to Theorem 3.1, the fraction of misclassified nodes is smaller than  $\frac{1}{2}$  when  $\frac{(c_{in} - c_{out})^2}{c_{in} + c_{out}}$  is larger than  $\frac{4C}{(\eta_1 - \eta_0)^2} \left( \frac{\bar{\alpha} + \lambda}{\lambda} \right)^2$ , which is indeed lower-bounded by a constant. □

The quantity  $\frac{(c_{in} - c_{out})^2}{c_{in} + c_{out}}$  can be interpreted as the signal-to-noise ratio. It is unfortunate that Corollary 3.3 does not allow us to control the constant in the statement of the corollary. This constant comes from the concentration of the adjacency matrix. Similar remarks were made in [19] for the analysis of spectral clustering in the constant degree regime for SBMs graph.

#### 4. Numerical experiments

This section presents numerical experiments both on simulated data sets generated from DC-SBMs and on real networks. In particular, we discuss the impact of the oracle mistakes (defined by the ratio  $\frac{\eta_0}{\eta_0 + \eta_1}$ ) on the performance of the algorithms. The code for the numerical experiments is available on GitHub at <https://github.com/mdreveton/ssl-sbm>



**Figure 1.** Cost in Algorithm 1 with the standard and normalized versions of the constraint, on 50 realizations of SBM with  $n = 500, p_{out} = 0.03$  and 50 labeled nodes with 10% noise.

**4.1. Synthetic data sets**

*4.1.1. Choice of  $\lambda$  and  $\tau$*

Let us denote by  $\sigma_1$  and  $\sigma_2$  the largest and second largest eigenvalues of  $A$ . We choose  $\tau = \frac{4}{n}(\sigma_1 + \sigma_2)$  and  $\lambda = \frac{\log \frac{\eta_1}{\eta_0}}{\log \frac{\sigma_1 + \sigma_2}{\sigma_1 - \sigma_2}}$ , if  $\eta_0 \neq 0$ , and  $\lambda = \frac{\log(n\eta_1)}{\log \frac{\sigma_1 + \sigma_2}{\sigma_1 - \sigma_2}}$ , otherwise. The heuristic for this choice is as follows. For an SBM graph, we have  $\sigma_1 \approx \frac{n}{2}(p_{in} + p_{out})$  and  $\sigma_2 \approx \frac{n}{2}(p_{in} - p_{out})$ , hence  $\frac{4}{n}(\sigma_1 + \sigma_2) = 2p_{in} > p_{out}$ , and  $\tau$  satisfies the condition of Theorem 3.1. For  $\lambda$ , we have  $\frac{\log \frac{\eta_1}{\eta_0}}{\log \frac{\sigma_1 + \sigma_2}{\sigma_1 - \sigma_2}} \approx \frac{\log \frac{\eta_1}{\eta_0}}{\log \frac{p_{in}}{p_{out}}}$ , which is indeed close to the expression of  $\lambda$  derived in Corollary 2.3 if  $p_{in}, p_{out} = o(1)$ .

*4.1.2. Choice of relaxation*

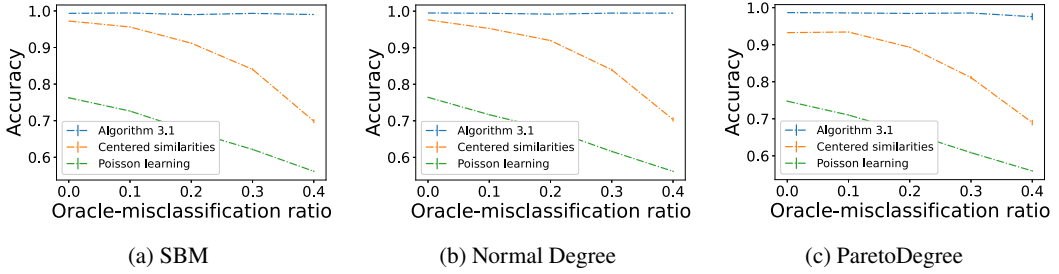
We first compare the choice of the constraint in the continuous relaxation (3.1). Specifically, we compare the choice  $\sum_i x_i^2 = n$  (we refer to it as *standard relaxation*) versus  $\sum_i d_i x_i^2 = 2|E|$  (we refer to it as *degree-normalized relaxation*). This leads to two versions of Algorithm 1, whose cost obtained on SBMs graph with a noisy oracle is presented in Figure 1. In particular, we observe that the normalized choice leads to a smaller cost. Therefore, in the following we will only consider the version of Algorithm 1 solving the relaxed problem (3.1) with constraint  $\sum_i d_i x_i^2 = 2|E|$  instead of  $\sum_i x_i^2 = n$ , as it gives better numerical results.

*4.1.3. Experiments on synthetic graphs*

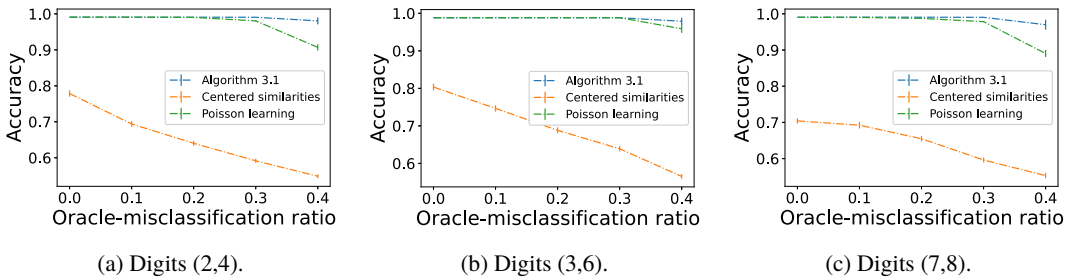
We first consider clustering on DC-SBM. We set  $n = 2000, p_{in} = 0.04$ , and  $p_{out} = 0.02$ . We consider three scenarios.

- In Figure 2(a) we consider a standard SBM ( $\theta_i = 1$  for all  $i$ );
- In Figure 2(b) we generate  $\theta_i$  according to  $|\mathcal{N}(0, \sigma^2)| + 1 - \sigma\sqrt{2/\pi}$  where  $|\mathcal{N}(0, \sigma^2)|$  denotes the absolute value of a normal random variable with mean 0 and variance  $\sigma^2$ . We take  $\sigma = 0.25$ . Note that this definition enforces  $\mathbb{E}\theta_i = 1$ .
- In Figure 2(c) we generate  $\theta_i$  from Pareto distribution with density function  $f(x) = \frac{am^a}{x^{a+1}} 1(x \geq m)$  with  $a = 3$  and  $m = 2/3$  (chosen such that  $\mathbb{E}\theta_i = 1$ ).

We compare the performance of Algorithm 1 with that of the algorithm of [22] (referred to as *Centered similarities*) and the *Poisson learning* algorithm described in [10]. We chose these two algorithms as references since they perform very well on real data sets and are designed to avoid flat solutions. Results are shown in Figure 2. We observe that when the oracle noise is low, the performance of Algorithm 1 is comparable to *Centered similarities*. But, when the noise becomes non-negligible,



**Figure 2.** Average accuracy obtained by different semi-supervised clustering methods on DC-SBM graphs, with  $n = 2000$ ,  $p_{in} = 0.04$ , and  $p_{out} = 0.02$  with different distributions for  $\theta$ . The number of labeled nodes is equal to 40. Accuracies are computed on the unlabeled nodes, and are averaged over 100 realizations; the error bars show the standard error.



**Figure 3.** Average accuracy obtained on a subset of the MNIST data set by different semi-supervised algorithms as a function of the oracle-misclassification ratio, when the number of labeled nodes is equal to 10. Accuracy is averaged over 100 random realizations, and the error bars show the standard error.

the performance of *Centered similarities* deteriorates, while the accuracy of Algorithm 1 remains high. We notice that *Poisson learning* gives poor results on synthetic data sets.

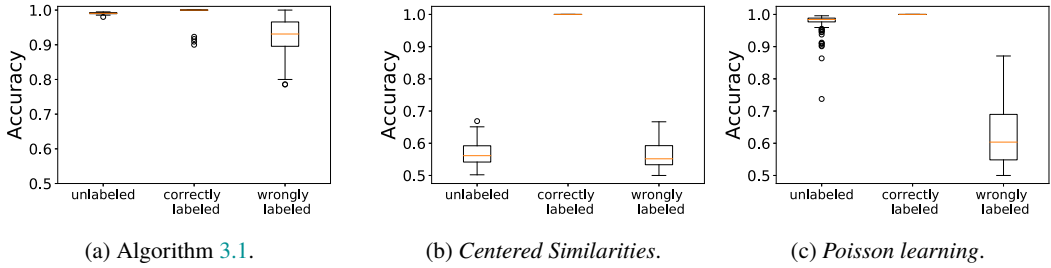
### 4.2. Experiments on real data

We next use real data to show that even if real networks are not generated by the DC-SBM, Algorithm 1 still performs well.

#### 4.2.1. MNIST

As a real-life example, we perform simulations on the standard MNIST data set [20]. As preprocessing, we select 1000 images corresponding to two digits and compute the  $k$ -nearest-neighbors graph (we take  $k = 8$ ) with Gaussian weights  $w_{ij} = \exp(-\|x_i - x_j\|^2/s_i^2)$  where  $x_i$  represents the data for image  $i$  and  $s_i$  is the average distance between  $x_i$  and its  $K$ -nearest neighbors. Figure 3 gives accuracy for different digit pairs. While the performance of *Poisson learning* is excellent, it can suffer from the oracle noise. On the other hand, the accuracy of Algorithm 1 remains unchanged.

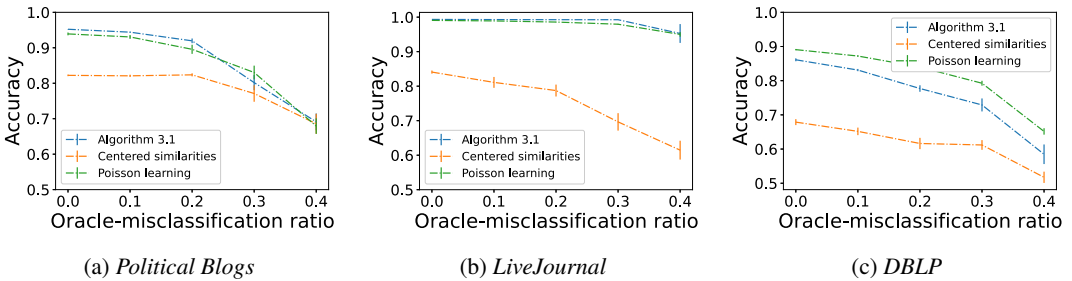
To further highlight the influence of the noise, we plot in Figure 4 the accuracy obtained by the three algorithms on the unlabeled nodes, the correctly labeled nodes, and the wrongly labeled nodes. We observe that the hard constraint  $X_\ell = s_\ell$  imposed by *Centered similarities* forces the correctly labeled nodes to be correctly classified. In contrast, the wrongly labeled nodes are not classified much better than a random guess. This heavily penalizes the unlabeled nodes' accuracy in an extremely noisy setting. On the contrary, Algorithm 1 allows for a smoother recovery: the unlabeled, correctly labeled, and wrongly labeled nodes have roughly the same classification accuracy. While some correctly labeled nodes are misclassified, many wrongly labeled nodes become correctly classified, and the unlabeled



**Figure 4.** Average accuracy obtained on the unlabeled, correctly labeled, and wrongly labeled nodes by the oracle. Simulations are done on the 1,000 digits (2,4). The noisy oracle correctly classifies 24 nodes and misclassifies 16 nodes, and the boxplots show 100 realizations.

**Table 1.** Parameters of the real data sets.  $n_1$  (resp.,  $n_2$ ) corresponds to the size of the first (resp., second) cluster, and  $|E|$  is the number of edges of the network.

Data set	$n_1$	$n_2$	$ E $
Political Blogs [2]	636	586	16,717
LiveJournal [27]	1,426	1,340	24,138
DBLP [27]	7,373	5,953	34,281



**Figure 5.** Average accuracy obtained on real networks by different semi-supervised algorithms as a function of the oracle-misclassification ratio. The number of labeled nodes is 30 for Political Blogs and LiveJournal, and 100 for DBLP. Accuracy is averaged over 50 random realizations, and the error bars show the standard error.

nodes are better recovered. Finally, *Poisson learning* shows a performance somewhere in between these two extreme cases: its accuracy on the unlabeled nodes is excellent, but it fails at correctly classifying the erroneously labeled nodes.

4.2.2. Common benchmark networks

Finally, we perform simulations on three benchmark networks: *Political Blogs*, *LiveJournal*, and *DBLP*. These networks are commonly used for graph clustering since the “ground truth” clusters are known. For *LiveJournal* and *DBLP*, we consider only the two largest clusters. The dimension of the data sets is given in Table 1 and the performances of semi-supervised algorithms in Figure 5. We observe that Algorithm 1 and *Poisson learning* outperform *Centered similarities* and can still achieve good accuracy even in the presence of noise in labeled data.

**Funding statement.** This research has been done within the project of Inria—Nokia Bell Labs “Distributed Learning and Control for Network Analysis.”

## References

- [1] Abbe E. (2017). Community detection and stochastic block models: Recent developments. *The Journal of Machine Learning Research* 18(1): 6446–6531.
- [2] Adamic L.A. & Glance N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 36–43.
- [3] Avrachenkov K. & Dreveton M. (2019). Almost exact recovery in label spreading. In *International Workshop on Algorithms and Models for the Web-Graph*. Springer, pp. 30–43.
- [4] Avrachenkov K., Kadavankandy A., & Litvak N. (2018). Mean field analysis of personalized PageRank with implications for local graph clustering. *Journal of Statistical Physics* 173(3–4): 895–916.
- [5] Avrachenkov K., Mishenin A., Gonçalves P., & Sokol M. (2012). Generalized optimization framework for graph-based semi-supervised learning. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, pp. 966–974.
- [6] Avrachenkov K.E., Filar J.A., & Howlett P.G. (2013). *Analytic perturbation theory and its applications*. SIAM.
- [7] Banerjee S., Deka P., & Olvera-Cravioto M. (2023). Pagerank nibble on the sparse directed stochastic block model. In *International Workshop on Algorithms and Models for the Web-Graph*. Springer, pp. 147–163.
- [8] Belkin M., Matveeva I., & Niyogi P. (2004). Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory*. Springer, pp. 624–638.
- [9] Ben-David S., Lu T., & Pál D. (2008). Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In *Conference on Learning Theory*.
- [10] Calder J., Cook B., Thorpe M., & Slepcev D. (2020). Poisson learning: Graph based semi-supervised learning at very low label rates. In *International Conference on Machine Learning*. PMLR, pp. 1306–1316.
- [11] Chapelle O., Schölkopf B., & Zien A. (2006). *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press.
- [12] Cozman F.G., Cohen I., & Cirelo M. (2002). Unlabeled data can degrade classification performance of generative classifiers. In *Proceedings of Flairs-02*, pp. 327–331.
- [13] Feige U. & Ofek E. (2005). Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms* 27(2): 251–275.
- [14] Gander W., Golub G.H., & Von Matt U. (1989). A constrained eigenvalue problem. *Linear Algebra and its Applications* 114: 815–839.
- [15] Horn R.A. and Johnson C.R. (2012). *Matrix analysis*. Cambridge University Press.
- [16] Iba Y. (1999). The Nishimori line and Bayesian statistics. *Journal of Physics A: Mathematical and General* 32(21): 3875–3888.
- [17] Kadavankandy A., Avrachenkov K., Cottatellucci L., & Sundaresan R. (2017). The power of side-information in subgraph detection. *IEEE Transactions on Signal Processing* 66(7): 1905–1919.
- [18] Karrer B. & Newman M.E.J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* 83(1): 016107.
- [19] Le C.M., Levina E., & Vershynin R. (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms* 51(3): 538–561.
- [20] LeCun Y., Cortes C., & Burges C.J.C. The MNIST database of handwritten digits. Available at: <https://yann.lecun.com/exdb/mnist/>.
- [21] Mai X. & Couillet R. (2018). A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *Journal of Machine Learning Research* 19(1): 3074–3100.
- [22] Mai X. & Couillet R. (2021). Consistent semi-supervised graph regularization for high dimensional data. *Journal of Machine Learning Research* 22(94): 1–48.
- [23] Nadler B., Srebro N., & Zhou X. (2009). Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data. *Advances in Neural Information Processing Systems* 22: 1330–1338.
- [24] Newman M.E.J. (2013). Spectral methods for community detection and graph partitioning. *Physical Review E* 88(4): 042822.
- [25] Saad H. & Nosratinia A. (2018). Community detection with side information: Exact recovery under the stochastic block model. *IEEE Journal of Selected Topics in Signal Processing* 12(5): 944–958.
- [26] Wagner D. & Wagner F. (1993). Between min cut and graph bisection. In *International Symposium on Mathematical Foundations of Computer Science*. Springer, pp. 744–750.
- [27] Yang J. & Leskovec J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42(1): 181–213.
- [28] Yu Y., Wang T., & Samworth R.J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* 102(2): 315–323.
- [29] Zhou D., Bousquet O., Lal T.N., Weston J., & Schölkopf B. (2004). Learning with local and global consistency. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pp. 16 321–328.
- [30] Zhu X., Ghahramani Z., & Lafferty J.D. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 912–919.

**Appendix A. Derivation of the MAP**

*Proof of Theorem 2.2.* Bayes’ formula gives  $\mathbb{P}(z | A, s) \propto \mathbb{P}(A | z, s) \mathbb{P}(z | s)$ , where the proportionality symbol hides  $\mathbb{P}(A | s)$ -term independent of  $z$ .

The likelihood term can be rewritten as follows:

$$\mathbb{P}(A | z, s) = \mathbb{P}(A | z) \propto \prod_{\substack{i < j \\ z_i = z_j}} \left( \frac{p_{in}}{p_{out}} \frac{1 - \theta_i \theta_j p_{out}}{1 - \theta_i \theta_j p_{in}} \right)^{a_{ij}} \left( \frac{1 - \theta_i \theta_j p_{in}}{1 - \theta_i \theta_j p_{out}} \right),$$

□

where the proportionality hides a constant  $C = \prod_{i < j} \left( \frac{\theta_i \theta_j p_{out}}{1 - \theta_i \theta_j p_{out}} \right)^{a_{ij}} (1 - \theta_i \theta_j p_{out})$  independent of  $z$ . Hence,

$$\begin{aligned} \log \mathbb{P}(A | z, s) &= \log C + \frac{1}{2} \sum_{i,j} 1(z_i \neq z_j) \left( (f_{ij}^{(1)} - f_{ij}^{(0)}) a_{ij} + f_{ij}^{(0)} \right) \\ &= \log C + \frac{1}{2} \sum_{i,j=1}^n \frac{1 - z_i z_j}{2} \left( (f_{ij}^{(1)} - f_{ij}^{(0)}) a_{ij} + f_{ij}^{(0)} \right) \\ &= \log C' - \frac{1}{4} x^T M x \end{aligned} \tag{A.1}$$

for some constant  $C'$  and  $M = (F_1 - F_0) \odot A + F_0$ .

The oracle information, given by the term  $\mathbb{P}(z | s)$ , is equal to

$$\begin{aligned} \mathbb{P}(z | s) &= \prod_{i=1}^n \frac{\mathbb{P}(s_i | z_i)}{\mathbb{P}(s_i)} \mathbb{P}(z_i) \\ &= \left( \frac{\eta_1}{\eta_1 + \eta_0} \right)^{|\{i \in \ell : z_i = s_i\}|} \left( \frac{\eta_0}{\eta_1 + \eta_0} \right)^{|\{i \in \ell : z_i \neq s_i\}|} \left( \frac{1}{2} \right)^n \\ &= \left( \frac{\eta_0}{\eta_1} \right)^{|\{i \in \ell : z_i \neq s_i\}|} \left( \frac{\eta_1}{\eta_1 + \eta_0} \right)^{|\ell|} \left( \frac{1}{2} \right)^n, \end{aligned} \tag{A.2}$$

where we used  $|\{i \in \ell : z_i = s_i\}| + |\{i \in \ell : z_i \neq s_i\}| = |\ell|$  in the last line. Noticing that

$$|\{i \in \ell : z_i \neq s_i\}| = \frac{1}{4} \sum_{i=1}^n ((\mathcal{P}z)_i - s_i)^2 = \frac{1}{4} (\mathcal{P}z - s)^T (\mathcal{P}z - s),$$

yields

$$\log \mathbb{P}(z | s) = -\frac{1}{4} \log \left( \frac{\eta_1}{\eta_0} \right) \cdot \|\mathcal{P}z - s\|^2 + C', \tag{A.3}$$

where  $C'$  is a term independent of  $z$ .

If  $\eta_0 \neq 0$ , the combination of Eqs. (A.1) and (A.3) with Bayes’ formula gives Expression (2.6). If  $\eta_0 = 0$ , then from Eq. (A.2) the term  $\mathbb{P}(z | s)$  is non-zero (and constant) if and only if  $z_i = s_i$  for every labeled node  $i \in [\ell]$ , and we obtain Expression (2.7).

*Proof of Corollary 2.3.* The proof follows from Theorem 2.2 and the fact that  $f_{ij}^{(0)} = \log \frac{1-p_{in}}{1-p_{out}}$  and  $f_{ij}^{(1)} = \log \frac{p_{in}}{p_{out}}$ . □

**Appendix B. Lemmas related to mean-field solution of the secular equation**

**Appendix B.1. Spectral study of a perturbed rank-2 matrix**

**Lemma B.1. (Matrix determinant lemma)** Suppose  $A \in \mathbf{R}^n$  is invertible, and let  $U, V$  be two  $n$  by  $m$  matrices. Then  $\det(A + UV^T) = \det A \det(I_m + V^T A^{-1} U)$ .

*Proof.* We take the determinant of  $\begin{pmatrix} A & -U \\ V^T & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ V^T & I \end{pmatrix} \cdot \begin{pmatrix} I & -A^{-1}U \\ 0 & I + V^T A^{-1}U \end{pmatrix}$  and  $\det \begin{pmatrix} A & -U \\ V^T & I \end{pmatrix} = \det I \det(A + UV^T)$  by the Schur complement formula [15, Sect. 0.8.5]. □

**Proposition B.2.** Let  $M = ZBZ^T$ , where  $B = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$  is a  $2 \times 2$  matrix, and  $Z = \begin{pmatrix} 1_{n/2} & 0_{n/2} \\ 0_{n/2} & 1_{n/2} \end{pmatrix}$  is an  $n \times 2$  matrix. Let  $m$  be an even number. We denote by  $P_{\mathcal{L}}$  the  $n \times n$  diagonal matrix whose first  $\frac{m}{2}$  and last  $\frac{m}{2}$  diagonal elements are ones, all other elements being zeros. Then,  $\det(tI_n + \lambda P_{\mathcal{L}} - M) = t^{n-m-2}(t + \lambda)^{m-2}(t - t_1^+)(t - t_1^-)(t - t_2^+)(t - t_2^-)$  with

$$t_1^{\pm} = \frac{1}{2} \left( \frac{n}{2}(a + b) - \lambda \pm \sqrt{\left( \lambda + \frac{n}{2}(a + b) \right)^2 - 2(a + b)\lambda m} \right),$$

$$t_2^{\pm} = \frac{1}{2} \left( \frac{n}{2}(a - b) - \lambda \pm \sqrt{\left( \lambda + \frac{n}{2}(a - b) \right)^2 - 2(a - b)\lambda m} \right).$$

*Proof.* For now, assume that  $t \neq -\lambda$  and  $t \neq 0$ . Then,  $tI_n + \lambda P_{\mathcal{L}}$  is invertible, and by Lemma B.1,

$$\begin{aligned} \det(tI_n + \lambda P_{\mathcal{L}} - M) &= \det(tI_n + \lambda P_{\mathcal{L}}) \det(I_2 + Z^T (tI_n + \lambda P_{\mathcal{L}})^{-1} (-ZB)) \\ &= (t + \lambda)^m t^{n-m} \det(I_2 - Z^T (tI_n + \lambda P_{\mathcal{L}})^{-1} ZB). \end{aligned} \tag{B.1}$$

Moreover,

$$(tI_n + \lambda P_{\mathcal{L}})^{-1} = \frac{1}{t}(I_n - P_{\mathcal{L}}) + \frac{1}{t + \lambda} P_{\mathcal{L}} = \frac{1}{t} I_n - \frac{\lambda}{t(t + \lambda)} P_{\mathcal{L}}.$$

Therefore, we can write

$$Z^T (tI_n + \lambda P_{\mathcal{L}})^{-1} ZB = \frac{1}{t} Z^T ZB - \frac{\lambda}{t(t + \lambda)} Z^T P_{\mathcal{L}} ZB = \frac{1}{t} \frac{n}{2} B - \frac{\lambda}{t(t + \lambda)} \frac{m}{2} B = xB,$$

where  $x := \frac{n}{2} \frac{1}{t(t + \lambda)} \left( t + \lambda \left( 1 - \frac{m}{n} \right) \right)$ . Thus, a direct computation of the determinant gives

$$\det(I_2 - Z^T (tI_n + \lambda P_{\mathcal{L}})^{-1} ZB) = (1 - x(a + b))(1 - x(a - b)).$$

Going back to Eq. (B.1), we can write

$$\det(tI_n + \lambda P_{\mathcal{L}} - M) = (t + \lambda)^{m-2} t^{n-m-2} P_1(t) P_2(t), \tag{B.2}$$



with  $P_1(t) = t(t + \lambda) - \frac{n}{2}(a + b)(t + \lambda(1 - \frac{m}{n}))$  and  $P_2(t) = t(t + \lambda) - \frac{n}{2}(a - b)(t + \lambda(1 - \frac{m}{n}))$ . Since  $t \in \mathbf{R} \mapsto \det(tI_n + \lambda P_{\mathcal{L}} - M)$  is continuous (even analytic), expression (B.2) is also valid for  $t = 0$  and  $t = -\lambda$  [6]. We end the proof by observing that

$$P_1(t) = (t - t_1^+)(t - t_1^-) \quad \text{and} \quad P_2(t) = (t - t_2^+)(t - t_2^-),$$

where  $t_1^\pm$  and  $t_2^\pm$  are defined in the proposition’s statement. □

**Corollary B.3.** *Let  $A$  be the adjacency matrix of a DC-SBM with  $p_{\text{in}} > p_{\text{out}} > 0$ , and  $s$  be the oracle information. Let  $\lambda, \tau > 0$ , and  $\bar{d}_\tau = \frac{n}{2}(p_{\text{in}} + p_{\text{out}}) - n\tau$ ,  $\bar{\alpha} = \frac{n}{2}(p_{\text{in}} - p_{\text{out}})$ . Let  $A_\tau := A - \tau 1_n 1_n^T$  and  $P_{\mathcal{L}}$  be the diagonal matrix whose element  $(P_{\mathcal{L}})_{ii}$  is 1 if  $s_i \neq 0$ , and 0 otherwise. Then, the spectrum of  $\mathbb{E}\tilde{\mathcal{L}} = -\mathbb{E}A_\tau + \lambda P - \gamma I_n$  is  $\{-\gamma - t_1^\pm, -\gamma - t_2^\pm, -\gamma; -\gamma + \lambda; 0\}$ , where*

$$t_1^\pm = \frac{1}{2} \left( \bar{d}_\tau - \lambda \pm \sqrt{(\lambda + \bar{d}_\tau)^2 - 4\bar{d}_\tau \lambda (\eta_1 + \eta_0)} \right),$$

$$t_2^\pm = \frac{1}{2} \left( \bar{\alpha} - \lambda \pm \sqrt{(\lambda + \bar{\alpha})^2 - 4\bar{\alpha} \lambda (\eta_1 + \eta_0)} \right).$$

*Proof.* Let  $M = \begin{pmatrix} p_{\text{in}} - \tau & p_{\text{out}} - \tau \\ p_{\text{out}} - \tau & p_{\text{in}} - \tau \end{pmatrix}$  and  $Z = \begin{pmatrix} 1_{n/2} & 0_{n/2} \\ 0_{n/2} & 1_{n/2} \end{pmatrix}$ . Then, we notice that  $\mathbb{E}A_\tau = ZMZ^T$  and we can apply Proposition B.2 to compute the characteristic polynomial of  $\mathbb{E}\tilde{\mathcal{L}}$ . For  $x \in \mathbf{R}$ ,  $\det(\mathbb{E}\tilde{\mathcal{L}} - xI_n) = \det((-\gamma - x)I_n - \mathbb{E}A_\tau + \lambda P)$ , whose roots are  $-\gamma - t_1^\pm, -\gamma - t_2^\pm, -\gamma$ , and  $-\gamma + \lambda$ . □

*Appendix B.1.1. Bounds for  $\bar{\gamma}_*$*

**Lemma B.4.** *Let  $\bar{\gamma}_*$  be the solution of Eq. (3.6) for the mean-field model. Then,*

$$-\bar{\alpha}(1 - 2\eta_0) \leq \bar{\gamma}_* \leq -\bar{\alpha}.$$

*Proof.* For  $\lambda \geq 0$ , we denote by  $(\bar{x}_\lambda, \bar{\gamma}_*(\lambda))$  the solution of the system (3.4) on a mean-field DC-SBM. The proof is in two steps. First, let us show that  $\bar{\gamma}_*(0) = -\bar{\alpha}$  and  $\bar{\gamma}_*(\infty) = -\bar{\alpha}(1 - 2\eta_0)$ . For  $\lambda = 0$ , the constrained linear system (3.4) reduces to an eigenvalue problem, and hence  $\bar{\gamma}_*(0)$  equals  $-\alpha$ , the smallest eigenvalue of  $-\mathbb{E}A_\tau$ . Moreover, when  $\lambda = \infty$ , the hard constraint  $x_\ell = \bar{s}_\ell$  is enforced, and the system (3.4) becomes

$$\begin{cases} (-\mathbb{E}A_\tau - \bar{\gamma}_*(\infty)I_n)_{uu} \bar{x}_u &= (\mathbb{E}A_\tau)_{ut} \bar{s}_\ell \\ \bar{x}_u^T \bar{x}_u &= n(1 - \eta_0 - \eta_1) \end{cases}$$

and we verify by hand that  $\bar{\gamma}_*(\infty) = -\bar{\alpha}(1 - 2\eta_0)$  together with  $\bar{x}_u = Z_u$  is indeed the solution.

Second, if we let  $C_\lambda(x) = -x^T \mathbb{E}A_\tau x + \lambda(\bar{s} - Px)^T(\bar{s} - Px)$  be the cost function minimized in (3.1), then from Eq. (3.4) we have  $\bar{\gamma}_*(\lambda_1) - \bar{\gamma}_*(\lambda_2) = C_{\lambda_1}(\bar{x}_1) - C_{\lambda_2}(\bar{x}_2) + \lambda_1 \bar{x}_1^T \bar{s} - \lambda_2 \bar{x}_2^T \bar{s}$ . Since  $\lambda \mapsto C_\lambda(x)$  is increasing, then  $\lambda_1 \leq \lambda_2$  implies  $C_{\lambda_1}(\bar{x}_1) \leq C_{\lambda_2}(\bar{x}_2)$ . Since  $\bar{x}_\lambda^T \bar{s} \geq 0$  (if it was not the case, then  $C_\lambda(-\bar{x}_\lambda) \leq C_\lambda(\bar{x}_\lambda)$ , and hence  $\bar{x}_\lambda \neq \arg \min_{x \in \mathbf{R}^n} C_\lambda(x)$ ), we can conclude that  $\bar{\gamma}_*(0) \leq \bar{\gamma}_*(\lambda)$  and that  $\bar{\gamma}_*(\lambda) \leq \bar{\gamma}_*(\infty)$ . □



Appendix B.1.2. Concentration of  $\gamma_*$

**Proposition B.5.** Let  $\gamma_*$  and  $\bar{\gamma}_*$  be the solutions of Eq. (3.4) for a DC-SBM and the mean-field DC-SBM, respectively. Then

$$|\gamma_* - \bar{\gamma}_*| \leq \left(1 + \frac{(\bar{\alpha} + \lambda)^3}{2\sqrt{\eta_1 + \eta_0}(\eta_1 - \eta_0)\bar{\alpha}^2\lambda}\right) \sqrt{\bar{d}}.$$

*Proof.* The gradient with respect to  $(\bar{\delta}_1, \dots, \bar{\delta}_n, \bar{b}_1, \dots, \bar{b}_n, \gamma)$  of the left-hand-side of Eq. (3.6) is equal to

$$2 \sum_{i=1}^n \frac{\bar{b}_i}{\bar{\delta}_i - \bar{\gamma}} \left[ \frac{\Delta b_i}{\bar{\delta}_i - \bar{\gamma}_*} - \frac{\bar{b}_i \Delta \bar{\delta}_i}{(\bar{\delta}_i - \bar{\gamma}_*)^2} + \frac{\bar{b}_i \Delta \gamma}{(\bar{\delta}_i - \bar{\gamma}_*)^2} \right].$$

Thus, we have

$$\Delta \gamma \sum_{i=1}^n \frac{\bar{b}_i^2}{(\bar{\delta}_i - \bar{\gamma}_*)^3} = \sum_{i=1}^n \frac{\bar{b}_i^2}{(\bar{\delta}_i - \bar{\gamma}_*)^3} \Delta \bar{\delta}_i - \sum_{i=1}^n \frac{\bar{b}_i}{(\bar{\delta}_i - \bar{\gamma}_*)^2} \Delta b_i + o(\Delta \bar{\delta}_i, \Delta b_i).$$

Firstly, we see that for all  $i \in [n]$ ,  $\Delta \bar{\delta}_i = |\delta_i - \bar{\delta}_i| \leq \|A - \mathbb{E}A\| \leq \bar{d}$  by the concentration of the adjacency matrix of a DC-SBM graph. Therefore, using this fact and  $\bar{\gamma}_* \leq \bar{\delta}_1 \leq \bar{\delta}_2 \leq \dots \leq \bar{\delta}_n$ ,

$$\begin{aligned} \Delta \gamma = |\gamma_* - \bar{\gamma}_*| &\leq \max_i |\delta_i - \bar{\delta}_i| + \frac{\max_i \frac{1}{(\bar{\delta}_i - \bar{\gamma}_*)^2} \sum_i |\bar{b}_i| \cdot |b_i - \bar{b}_i|}{\min_i \frac{1}{(\bar{\delta}_i - \bar{\gamma}_*)^3} \sum_i \bar{b}_i^2} \\ &\leq \sqrt{\bar{d}} + \frac{\max_i (\bar{\delta}_i - \bar{\gamma}_*)^3 \sum_i |\bar{b}_i| \cdot |b_i - \bar{b}_i|}{\min_i (\bar{\delta}_i - \bar{\gamma}_*)^2 \sum_i \bar{b}_i^2}. \end{aligned}$$

We notice that  $\min_i |\bar{\delta}_i - \bar{\gamma}_*| = \bar{\delta}_1 - \bar{\gamma}_*$ . By using Lemma B.4 and the expression of  $\bar{\delta}_1$  given in Corollary B.3, we have

$$\min_i |\bar{\delta}_i - \bar{\gamma}_*| \geq \bar{\alpha} + \lambda.$$

Similarly,  $\max_i |\bar{\delta}_i - \bar{\gamma}_*| = \bar{\delta}_n - \bar{\gamma}_* = \bar{\delta}_n - \bar{\delta}_1 + \bar{\delta}_1 - \bar{\gamma}_*$ . Corollary B.3 implies  $\bar{\delta}_n = \lambda$  and  $\bar{\delta}_1 = \frac{1}{2} \left( \lambda - \bar{\alpha} - \sqrt{(\lambda + \bar{\alpha})^2 - 4\bar{\alpha}\lambda(\eta_0 + \eta_1)} \right)$ , thus  $\bar{\delta}_n - \bar{\delta}_1 \leq \bar{\alpha} + \lambda$ . Hence, using Lemma B.4,

$$\max_i |\bar{\delta}_i - \bar{\gamma}_*| \leq \frac{3}{2} (\bar{\alpha} + \lambda).$$

Therefore, we have

$$|\gamma_* - \bar{\gamma}_*| \leq \sqrt{\bar{d}} + \frac{27}{8} (\bar{\alpha} + \lambda) \cdot \frac{\sum_i |\bar{b}_i| \cdot |b_i - \bar{b}_i|}{\sum_i \bar{b}_i^2}. \tag{B.3}$$

The term  $\frac{\sum_i |\bar{b}_i| \cdot |b_i - \bar{b}_i|}{\sum_i \bar{b}_i^2}$  can be bounded as follow. Let  $\mathcal{I} = \{i \in [n] : \bar{b}_i \neq 0\}$ . Then

$$\sum_i |\bar{b}_i| \cdot |b_i - \bar{b}_i| \leq \max_{i \in \mathcal{I}} |b_i - \bar{b}_i| \cdot \sum_{i \in \mathcal{I}} |\bar{b}_i|.$$

Combining the Cauchy–Schwarz inequality

$$|b_i - \bar{b}_i| = \lambda |(Q_{\cdot i} - \bar{Q}_{\cdot i})^T \bar{s}| \leq \lambda \|Q_{\cdot i} - \bar{Q}_{\cdot i}\|_2 \cdot \|\bar{s}\|,$$

with the Davis–Kahan theorem [28]

$$\|Q_{\cdot i} - \bar{Q}_{\cdot i}\|_2 \leq \frac{2^{3/2} \|A - \mathbb{E}A\|}{\min\{\bar{\delta}_i - \bar{\delta}_{i-1}, \bar{\delta}_{i+1} - \bar{\delta}_i\}},$$

$\|\bar{s}\| = \sqrt{(\eta_0 + \eta_1)n}$ , and the concentration of  $A$  toward  $\mathbb{E}A$ , yields

$$\max_{i \in \mathcal{I}} |b_i - \bar{b}_i| \leq \frac{\lambda \sqrt{(\eta_0 + \eta_1)n}}{\min_{i \in \mathcal{I}}\{\bar{\delta}_i - \bar{\delta}_{i-1}, \bar{\delta}_{i+1} - \bar{\delta}_i\}} \cdot 2^{3/2} \sqrt{d}.$$

Using Lemma B.6, we see that  $\mathcal{I} = \{i \in [n] : \delta_i \notin \{0, t_1^-\}\}$ . Combining it with Corollary B.3, gives

$$\begin{aligned} \min_{i \in \mathcal{I}} \{\bar{\delta}_i - \bar{\delta}_{i-1}, \bar{\delta}_{i+1} - \bar{\delta}_i\} &= \lambda + t_2^+ \\ &= \frac{\alpha + \lambda}{2} \left( 1 - \sqrt{1 - 4 \frac{\alpha \lambda}{(\alpha + \lambda)^2} (\eta_0 + \eta_1)} \right) \\ &\geq \frac{\alpha \lambda}{\alpha + \lambda} (\eta_0 + \eta_1), \end{aligned}$$

where we used  $\sqrt{1-x} \leq 1-x/2$ . Therefore,

$$\max_{i \in \mathcal{I}} |b_i - \bar{b}_i| \leq 2^{3/2} \sqrt{\frac{n\bar{d}}{\eta_0 + \eta_1}} \cdot \frac{\alpha + \lambda}{\alpha}.$$

Finally, Lemma B.7 ensures that

$$\frac{\sum_i |\bar{b}_i|}{\sum_i \bar{b}_i^2} \leq \frac{2}{\sqrt{n}(\eta_1 - \eta_0)} \cdot \frac{\lambda + \alpha}{\alpha \lambda} \left( 1 + \frac{2\eta_0 n \sqrt{\eta_1 + \eta_0}}{\lambda} \right).$$

Therefore,

$$\begin{aligned} \frac{\sum_i |\bar{b}_i| \cdot |b_i - \bar{b}_i|}{\sum_i \bar{b}_i^2} &\leq 2^{5/2} \left( \frac{\alpha + \lambda}{\alpha} \right)^2 \frac{\sqrt{d}}{(\eta_1 - \eta_0) \sqrt{\eta_0 + \eta_1}} \left( 1 + \frac{2\eta_0 n \sqrt{\eta_1 + \eta_0}}{\lambda} \right) \\ &\leq 2^3 \left( \frac{\alpha + \lambda}{\alpha} \right)^2 \frac{\sqrt{d}}{(\eta_1 - \eta_0) \sqrt{\eta_0 + \eta_1}}, \end{aligned}$$

where we used the condition  $2\eta_0 n \sqrt{\eta_1 + \eta_0} \ll \lambda$ .

Going back to inequality (B.3), this implies that  $|\gamma_* - \bar{\gamma}_*| \leq \left( 1 + \frac{27}{2^6} \frac{(\alpha + \lambda)^3}{\alpha^2 \lambda} \frac{1}{(\eta_1 - \eta_0) \sqrt{\eta_0 + \eta_1}} \right) \sqrt{d}$ , and this concludes the proof.  $\square$

**Lemma B.6.** *Let  $-\mathbb{E}A_\tau + \lambda P = \bar{Q} \bar{\Delta} \bar{Q}^T$ , where  $\bar{\Delta} = \text{diag}(\bar{\delta}_1, \dots, \bar{\delta}_n)$  and  $\bar{Q}^T \bar{Q} = I_n$ . Denote  $\bar{b} = \lambda \bar{Q}^T s$ . We have  $\bar{b}_1 \geq \sqrt{n} \frac{\lambda(\eta_1 - \eta_0)}{2} \frac{\bar{\alpha}}{\lambda + \bar{\alpha}}$ . Moreover,  $\bar{b}_i = 0$  if  $\bar{\delta}_i = 0$  or if  $\bar{\delta}_i = -t_1^-$ .*

*Proof.* First, from Corollary B.3,  $\bar{\delta}_1 = -t_2^+ = -\frac{1}{2} \left( \bar{\alpha} - \lambda + \sqrt{(\lambda + \bar{\alpha})^2 - 4\bar{\alpha}\lambda(\eta_1 + \eta_0)} \right)$ . By symmetry, the  $i$ th component of the first eigenvector  $\bar{Q}_{\cdot 1}$  (associated with  $\bar{\delta}_1$ ) is equal to

$$\begin{cases} v_1 Z_i & \text{if } i \in [\ell], \\ v_0 Z_i & \text{if } i \notin [\ell], \end{cases}$$

where  $v_1$  and  $v_0$  are to be determined. Thus, the equation  $(-\mathbb{E}A_\tau + \lambda\mathcal{P})\bar{Q}_{\cdot 1} = \bar{\delta}_1\bar{Q}_{\cdot 1}$  leads to

$$\begin{cases} \bar{\alpha}((\eta_1 + \eta_0)v_1 + (1 - \eta_1 - \eta_0)v_0) & = -t_2^+v_0 \\ \bar{\alpha}((\eta_1 + \eta_0)v_1 + (1 - \eta_1 - \eta_0)v_0) + \lambda v_1 & = -t_2^+v_1, \end{cases}$$

which, given the norm constraint  $\|v\|_2 = 1$ , yields

$$\begin{cases} v_1 & = \frac{1}{\sqrt{n}} \frac{t_2^+}{\sqrt{(\eta_1 + \eta_0)(t_2^+)^2 + (1 - \eta_1 - \eta_0)(t_2^+ + \lambda)^2}}, \\ v_0 & = \frac{1}{\sqrt{n}} \frac{t_2^+ + \lambda}{\sqrt{(\eta_1 + \eta_0)(t_2^+)^2 + (1 - \eta_1 - \eta_0)(t_2^+ + \lambda)^2}}. \end{cases}$$

Since  $\bar{b}_1 = \lambda v^T \bar{s} = \lambda(\eta_1 - \eta_0)nv_1$ , we have

$$\frac{\bar{b}_1}{\sqrt{n}} = \lambda(\eta_1 - \eta_0) \frac{t_2^+}{\sqrt{(\eta_1 + \eta_0)(t_2^+)^2 + (1 - \eta_1 - \eta_0)(t_2^+ + \lambda)^2}}.$$

The proof ends by noticing that  $t_2^+ \geq \frac{\bar{\alpha}}{2}$  and  $t_2^+ \leq \bar{\alpha}$ . Indeed,

$$\begin{aligned} \frac{\bar{b}_1}{\sqrt{n}} &\geq \lambda(\eta_1 - \eta_0) \frac{\bar{\alpha}}{2\sqrt{(\eta_1 + \eta_0)\bar{\alpha}^2 + (1 - \eta_1 - \eta_0)(\bar{\alpha} + \lambda)^2}} \\ &\geq \frac{\lambda(\eta_1 - \eta_0)}{2} \frac{\bar{\alpha}}{(\bar{\alpha} + \lambda)\sqrt{(\eta_1 + \eta_0)\left(\frac{\bar{\alpha}}{\bar{\alpha} + \lambda}\right)^2 + 1 - \eta_1 - \eta_0}} \\ &\geq \frac{\lambda(\eta_1 - \eta_0)}{2} \frac{\bar{\alpha}}{\lambda + \bar{\alpha}}. \end{aligned}$$

This proves the first claim of the lemma.

Similarly, by symmetry the  $i$ th component of the eigenvector  $v'$  associated with  $-t_1^-$  equals  $v'_\ell$  if  $i \in \ell$ , and  $v'_u$  otherwise, and therefore  $(v')^T s = 0$ .

Finally, let  $I_0 := \{i \in [n] : \bar{\delta}_i = 0\}$ . By Corollary B.3, we have  $|I_0| = n(1 - \eta_1 - \eta_0) - 2$ . Since 0 is also eigenvalue of order  $n(1 - \eta_0 - \eta_1) - 2$  of the extracted sub-matrix  $(-\mathbb{E}A_\tau + \lambda\mathcal{P})_{u,u} = (-\mathbb{E}A_\tau)_{u,u}$ , we have for all  $k \in I_0$ ,  $\bar{Q}_{ik} = 0$  for every  $i \in [n]$ . Therefore, for  $k \in I_0$ ,  $b_k = \lambda\bar{Q}_{\cdot k}^T s = 0$ .  $\square$

**Lemma B.7.** Let  $-\mathbb{E}A_\tau + \lambda\mathcal{P} = \bar{Q}\bar{\Delta}\bar{Q}^T$ , where  $\bar{\Delta} = \text{diag}(\bar{\delta}_1, \dots, \bar{\delta}_n)$  and  $\bar{Q}^T\bar{Q} = I_n$ . Denote  $\bar{b} = \lambda\bar{Q}^T s$  and let  $\mathcal{I} = \{i \in [n] : \bar{b}_i \neq 0\}$ . We have  $\frac{\sum_{i \in \mathcal{I}} |\bar{b}_i|}{\sum_{i \in \mathcal{I}} |\bar{b}_i|^2} \leq \frac{2}{\sqrt{n}(\eta_1 - \eta_0)} \cdot \frac{\lambda + \alpha}{\alpha\lambda} \left( 1 + \frac{2\eta_0 n \sqrt{\eta_1 + \eta_0}}{\lambda} \right)$ .

*Proof.* Using [Lemma B.6](#), we see that  $\mathcal{I} = \{i \in [n] : \delta_i \notin \{0, t_1^-\}\}$ . Thus,

$$\frac{\sum_{i \in \mathcal{I}} |\bar{b}_i|}{\sum_{i \in \mathcal{I}} |\bar{b}_i|^2} = \frac{|b_1| + \sum_{i: \delta_i = \lambda} |\bar{b}_i|}{|\bar{b}_1|^2 + \sum_{i: \delta_i = \lambda} |\bar{b}_i|^2} \leq \frac{1}{|\bar{b}_1|} + \frac{\sum_{i: \delta_i = \lambda} |\bar{b}_i|}{|\bar{b}_1|^2},$$

where  $\bar{b}_1$  denotes the element of vector  $\bar{b}$  corresponding to eigenvalue  $\delta_1 = -t_2^+$ . By [Lemma B.6](#), we have  $\bar{b}_1 \geq \sqrt{n} \frac{\lambda(\eta_1 - \eta_0)}{2} \frac{\bar{\alpha}}{\lambda + \bar{\alpha}}$ . Hence,

$$\frac{1}{|\bar{b}_1|} \leq \frac{\lambda + \bar{\alpha}}{\bar{\alpha}\lambda} \frac{2}{(\eta_1 - \eta_0)\sqrt{n}}. \tag{B.4}$$

We note that the eigenvalue  $\delta_i = \lambda$  is of multiplicity  $\eta n - 2$ . Let us denote by  $\{v_i\}$  the corresponding  $\eta n - 2$  orthonormal eigenvectors associated with eigenvalue  $\lambda$ . Let  $v_{ij}$  denote the  $j$ th entry of  $v_i$ . We notice from the block structure of  $-\mathbb{E}A_\tau + \lambda \mathcal{P}$  that  $v_{ij} = 0$  if  $j \notin \ell$ . Moreover, if we let  $\tilde{v}_i$  be the restriction of  $v_i$  to  $\ell$ , then  $\tilde{v}_i$  belongs to the kernel of  $(-\mathbb{E}A_\tau)_{\ell\ell}$ . Therefore,  $\sum_{j \in \ell} \tilde{v}_{ij} = 0$ , and

$$\bar{b}_i = \lambda v_i^T s = -2\lambda \sum_{j \in \ell_0} \tilde{v}_{ij},$$

where  $\ell_0 = \{j \in \ell : s_i \neq z_i\}$  is the set of nodes mislabeled by the oracle. Hence,

$$\sum_{i: \delta_i = \lambda} |\bar{b}_i| = 2\lambda \sum_{i: \delta_i = \lambda} \left| \sum_{j \in \ell_0} \tilde{v}_{ij} \right| \leq 2\lambda \sum_{i \in \ell} \sum_{j \in \ell_0} |\tilde{v}_{ij}| \leq 2\lambda \sum_{j \in \ell_0} \sqrt{|\ell|} \sqrt{\sum_{i \in \ell} |\tilde{v}_{ij}|^2} \leq 2\lambda |\ell_0| \sqrt{|\ell|},$$

where the last inequality follows from the fact that the matrix  $(\tilde{v}_{ij})_{i,j}$  is orthogonal. Hence, using  $\bar{b}_1 \geq \sqrt{n} \frac{\lambda(\eta_1 - \eta_0)}{2} \frac{\bar{\alpha}}{\lambda + \bar{\alpha}}$ ,  $|\ell_0| = \eta_0$ , and  $|\ell| = (\eta_0 + \eta_1)n$ , we obtain

$$\frac{\sum_{i: \delta_i = \lambda} |\bar{b}_i|}{|\bar{b}_1|^2} \leq 4\eta_0 \sqrt{n} \frac{\sqrt{\eta_1 + \eta_0}}{\eta_1 - \eta_0} \frac{\lambda + \alpha}{\alpha}.$$

Combining the latter inequality with [\(B.4\)](#) leads to the desired result. □

### Appendix C. Mean-field solution

In this section, we calculate the solution  $\bar{x}$  to the mean-field model and deduce from it the conditions to recover the clusters.

**Proposition C.1.** *Suppose that  $\tau > p_{\text{out}}$ . Then the solution of Eq. (3.5) on the mean-field DC-SBM is the vector  $\bar{x}$  whose element  $\bar{x}_i$  is given by*

$$\bar{x}_i = \begin{cases} C(-1 + (\eta_1 - \eta_0)\bar{\alpha}B)Z_i, & \text{if } i \in \ell \text{ and } s_i \neq Z_i, \\ C(1 + (\eta_1 - \eta_0)\bar{\alpha}B)Z_i, & \text{if } i \in \ell \text{ and } s_i = Z_i, \\ \frac{-\bar{\alpha}C}{\bar{\alpha}(1 - \eta_1 - \eta_0) + \bar{\gamma}_*} (\eta_1 - \eta_0)(1 + (\eta_1 + \eta_0)\bar{\alpha}B)Z_i, & \text{if } i \notin \ell, \end{cases}$$

where  $\bar{\alpha} = \frac{n}{2}(p_{\text{in}} - p_{\text{out}})$ ,  $B = \frac{\bar{\alpha}\bar{\gamma}_*}{\lambda\bar{\alpha}(1 - \eta_1 - \eta_0) + \bar{\gamma}_*(\lambda - \bar{\alpha} - \bar{\gamma}_*)}$  and  $C = \frac{\lambda}{\lambda - \bar{\gamma}_*}$ .

*Proof.* Let  $\bar{x}$  be a solution of Eq. (3.5). By symmetry, we have

$$\bar{x}_i = \begin{cases} x_t Z_i, & \text{if } i \in [\ell] \text{ and } \bar{s}_i = Z_i, \\ x_f Z_i, & \text{if } i \in [\ell] \text{ and } \bar{s}_i = -Z_i, \\ x_0 Z_i, & \text{if } i \notin [\ell], \end{cases}$$

where  $x_t$ ,  $x_f$  and  $x_0$  are unknowns to be determined. Since for every  $i \in [n]$

$$(\mathbb{E}A_\tau \bar{x})_i = \bar{\alpha} (x_0(1 - \eta_1 - \eta_0) + x_t \eta_1 + x_f \eta_0),$$

the linear system composed of the equations  $((-\mathbb{E}A_\tau + \lambda \mathcal{P} - \bar{\gamma}_* I_n) \bar{x})_i = \lambda s_i$  for all  $i \in [n]$  leads to the system

$$\begin{cases} -\bar{\alpha} ((1 - \eta_1 - \eta_0)x_0 + x_t \eta_1 + x_f \eta_0) - \bar{\gamma}_* x_0 & = 0, \\ -\bar{\alpha} ((1 - \eta_1 - \eta_0)x_0 + x_t \eta_1 + x_f \eta_0) - \bar{\gamma}_* x_t + \lambda x_t & = \lambda, \\ -\bar{\alpha} ((1 - \eta_1 - \eta_0)x_0 + x_t \eta_1 + x_f \eta_0) - \bar{\gamma}_* x_f + \lambda x_f & = -\lambda. \end{cases}$$

The rows of the latter system correspond to a node unlabeled by the oracle, correctly labeled and falsely labeled, respectively. This system can be rewritten as follows:

$$\begin{cases} x_0 & = \frac{-\bar{\alpha}}{\bar{\alpha}(1-\eta_1-\eta_0)+\bar{\gamma}_*} (\eta_1 x_t + \eta_0 x_f), \\ \bar{\gamma}_* x_0 + x_t(\lambda - \bar{\gamma}_*) & = \lambda, \\ \bar{\gamma}_* x_0 + x_f(\lambda - \bar{\gamma}_*) & = -\lambda. \end{cases}$$

In particular, we have  $x_t - x_f = \frac{2\lambda}{\lambda - \bar{\gamma}_*}$ . By subsequently eliminating  $x_0$  and  $x_t$  in the equation  $\bar{\gamma}_* x_0 + x_f(\lambda - \bar{\gamma}_*) = -\lambda$ , we find

$$\begin{aligned} x_f &= \frac{\lambda}{\lambda - \bar{\gamma}_*} \left( -1 + \frac{\bar{\alpha} \bar{\gamma}_* (\eta_1 - \eta_0)}{\lambda \bar{\alpha} (1 - \eta_1 - \eta_0) + \lambda \bar{\gamma}_* - \bar{\gamma}_* (\bar{\alpha} + \bar{\gamma}_*)} \right), \\ x_t &= \frac{\lambda}{\lambda - \bar{\gamma}_*} \left( 1 + \frac{\bar{\alpha} \bar{\gamma}_* (\eta_1 - \eta_0)}{\lambda \bar{\alpha} (1 - \eta_1 - \eta_0) + \lambda \bar{\gamma}_* - \bar{\gamma}_* (\bar{\alpha} + \bar{\gamma}_*)} \right), \end{aligned}$$

and finally

$$x_0 = \frac{-\bar{\alpha}}{\bar{\alpha}(1 - \eta_1 - \eta_0) + \bar{\gamma}_*} \cdot \frac{\lambda}{\lambda - \bar{\gamma}_*} \left( 1 + \frac{\bar{\alpha} \bar{\gamma}_* (\eta_1 + \eta_0)}{\lambda \bar{\alpha} (1 - \eta_1 - \eta_0) + \lambda \bar{\gamma}_* - \bar{\gamma}_* (\bar{\alpha} + \bar{\gamma}_*)} \right). \quad \square$$

**Corollary C.2.** *Suppose that  $\tau > p_{\text{out}}$ . Then  $\text{sign}(\bar{x}_i) = \text{sign}(Z_i)$  if*

- node  $i$  is not labeled by the oracle;
- node  $i$  is correctly labeled by the oracle;
- node  $i$  is mislabeled by the oracle and  $\lambda < (1 - 2\eta_0)\bar{\alpha} \frac{\eta_1 - \eta_0}{\eta_1 + \eta_0}$ .

*Proof.* A node  $i$  is correctly classified by decision rule (3.3) if the sign of  $\bar{x}_i$  is equal to the sign of  $Z_i$ . Using Lemma B.4 in Appendix B.1.1, we have  $-\bar{\alpha} \leq \bar{\gamma}_* \leq -\bar{\alpha}(1 - 2\eta_0)$ . Therefore, the quantities  $B$  and  $C$  in Proposition C.1 verify  $C \geq 0$  and  $\frac{1-2\eta_0}{\lambda(\eta_0+\eta_1)} \leq B \leq \frac{1}{\lambda(\eta_0+\eta_1)}$ . The statement then follows from the expression of  $\bar{x}_i$  computed in Proposition C.1.  $\square$

**Appendix D. Cost comparison in the constrained eigenvalue problem**

**Lemma D.1.**

Let  $(\gamma_1, x_1)$  and  $(\gamma_2, x_2)$  be two solutions of the system (3.4), and denote by  $\mathcal{C}(x) = -x^T A_\tau x + \lambda(s - \mathcal{P}x)^T (s - \mathcal{P}x)$  the cost function minimized in (3.1). Then, we have

$$\mathcal{C}(x_1) - \mathcal{C}(x_2) = \frac{\gamma_1 - \gamma_2}{2} \|x_1 - x_2\|^2.$$

*Proof.* Because  $(\gamma_1, x_1)$  and  $(\gamma_2, x_2)$  are solutions of (3.4), it holds that

$$(-A_\tau + \lambda\mathcal{P})x_1 = \gamma_1 x_1 + \lambda s, \tag{D.1}$$

$$(-A_\tau + \lambda\mathcal{P})x_2 = \gamma_2 x_2 + \lambda s, \tag{D.2}$$

as well as  $\|x_1\|^2 = \|x_2\|^2 = n$ . Thus, we notice that

$$\begin{aligned} \mathcal{C}(x_1) &= x_1^T (-A_\tau + \lambda\mathcal{P})x_1 + \lambda s^T s - 2\lambda x_1^T \mathcal{P}s \\ &= -\lambda x_1^T s + \gamma_1 n + \lambda s^T s, \end{aligned}$$

where we used  $\mathcal{P}s = s$  and the fact that  $(\gamma_1, x_1)$  is a solution of the system (3.4). Therefore,

$$\mathcal{C}(x_1) - \mathcal{C}(x_2) = (\gamma_1 - \gamma_2)n + \lambda(x_2 - x_1)^T s.$$

Finally, by multiplying on the left Eq. (D.1) by  $x_2^T$  (resp., Eq. (D.2) by  $x_1^T$ ), we obtain

$$\begin{aligned} \lambda x_2^T s &= x_2^T (-A_\tau + \lambda\mathcal{P})x_1 - \gamma_1 x_2^T x_1, \\ \lambda x_1^T s &= x_1^T (-A_\tau + \lambda\mathcal{P})x_2 - \gamma_2 x_1^T x_2. \end{aligned}$$

Thus,

$$\mathcal{C}(x_1) - \mathcal{C}(x_2) = (\gamma_1 - \gamma_2) \left( n - x_1^T x_2 \right) = \frac{\gamma_1 - \gamma_2}{2} \left( \|x_1\|^2 + \|x_2\|^2 - 2x_1^T x_2 \right) = \frac{\gamma_1 - \gamma_2}{2} \|x_1 - x_2\|^2,$$

where we used the constraints  $\|x_1\|^2 = \|x_2\|^2 = n$ . □