Optimality of variational inference for stochastic block model with missing links

Solenne Gaucher, Olga Klopp

Optimality of variational inference for stochastic block model with missing links

Solenne Gaucher *[1] and Olga Klopp †[2,3]

[1]Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay
[2]ESSEC Business School
[3]CREST, ENSAE

Abstract

Variational methods are extremely popular in the analysis of network data. Statistical guarantees obtained for these methods typically provide asymptotic normality for the problem of estimation of global model parameters under the stochastic block model. In the present work, we consider the case of networks with missing links that is important in application and show that the variational approximation to the maximum likelihood estimator converges at the minimax rate. This provides the first minimax optimal and tractable estimator for the problem of parameter estimation for the stochastic block model with missing links. We complement our results with numerical studies of simulated and real networks, which confirm the advantages of this estimator over current methods.

1 Introduction

The analysis of network data poses both computational and theoretical challenges. Most results obtained in the literature concentrate on the stochastic block model (SBM) which is known to be a good proxy for more general models, such as the inhomogeneous random graph model, [34]. Recently, variational methods ([27, 47]) have attracted considerable attention as they offer computationally tractable algorithms often combined with theoretical guarantees. Theoretical results that one can find for such variational methods provide asymptotic normality rates for parameter estimates of stochastic block data. For example, consistency has been shown for profile likelihood maximization [7] and variational approximation to the maximum likelihood estimator [12], [6]. These results have been extended to the case of dynamic stochastic block model [33] and sampled data [46]. These work focus on parameter estimation, as in [42] and [51], who establish the minimax optimality of variational methods in a large class of models (which does however not include the stochastic block model). Variational inference has also been successfully applied to the problem of community detection, see, e.g., [3, 52, 25, 43]. In particular, the authors of [52] show that an iterative Batch Coordinate Ascent Variational Inference algorithm designed for the two-parameters, assortative stochastic block model achieves statistical optimality for community detection problem. Note that this algorithm cannot be extended to the more general stochastic block model considered here.

In parallel with this line of work, the problem of statistical estimation of model parameters, in particular, the question of minimax optimal convergence rates, has been actively studied in the statistical community. In the case of dense graphs, a pioneering paper [16] shows that, for the problem of estimating the matrix of connection probabilities, the least square estimator is minimax optimal and [17] provides optimal rate for Bayes estimation. For the more challenging case of sparse graphs, the minimax optimal rates have been first obtained in [28] building on the restricted least square estimator. In [15], the authors consider the least square estimator in the setting when observations about the presence or absence of an edge are missing independently at random with the same probability p. Unfortunately, least square estimation is too

*solenne.gaucher@math.u-psud.fr
†kloppolga@math.cnrs.fr

computationally expensive to be used in practice. Many other approaches have been proposed, for example, spectral clustering [38, 21, 44], modularity maximization [40, 7], belief propagation [13], neighbourhood smoothing [53], convex relaxation of k-means clustering [19] and of likelihood maximization [4], and universal singular value thresholding [10, 29, 49]. These approaches are computationally tractable but show sub-optimal statistical performances. So the question of possible computational gap when no polynomial time algorithm can achieve minimax optimal rate of convergence has been raised.

The present work goes in these two directions. We study the statistical properties of the mean field variational Bayes method and show that it achieves the optimal statistical accuracy. In particular, these results close the open question on the possible existence of a computational gap for the problem of global parameter estimation. We built our analysis on the approach developed in [12], [6] and [46] using the closeness of maximum likelihood and maximum variational likelihood and on the results that show the minimax optimality of the maximum likelihood estimator [18].

In the present paper, we deal with settings where the network is not fully observed, a common problem when studying real life networks. In many applications the network has missing data as detecting interactions can require significant experimental effort, see, [31, 50, 23, 20]. For example, in biology graphs are used to model interactions between proteins. Discovery of these interactions can be costly and time-consuming [8]. On the other hand, the size of some networks from social media or genome sequencing may be so large that only subsamples of the data are considered [5]. It has been observed that incomplete observation of the network structure may considerably affect the accuracy of inference methods [30] and missing data must be taken into account while analyzing networks data. A popular approach consists in considering the edges with uncertain status as non-existing. In the present paper, we use a different framework by considering such edges as missing and introducing a separated data missing mechanism. A natural application of our method is link prediction [35, 54], the task of predicting whether two nodes in a network are connected. Our approach allows to deduce the pairs of nodes that are most likely to interact based on the known interactions in the network. Behind inference of the networks structure, our algorithms can be used to predict the links that may appear in the future if we consider networks evolving over the time. For example, in a social network, two users that are not yet connected but are likely to be connected can be recommended as promising friends.

1.1 Contribution and outline

The paper is organized as follows. After summarizing notations, we introduce our model and the maximum likelihood estimator for the stochastic block model with missing observations in Section 2. In Section 3, we introduce the mean field variational Bayes method and present a new estimator which combines the labels obtained using the variational method and the empirical mean for estimation of connection probabilities. In Section 3.2, we show that our estimator is minimax optimal for dense stochastic block models with missing observations as well as for sparse stochastic block models. Finally, in Section 4 we provide an extensive numerical study both on synthetic and real-life data which shows clear advantages of our estimator over current methods.

1.2 Notations

We provide here a summary of the notations used throughout the paper. For all $d \in \mathbb{N}_*$, we denote by $[d]$ the set $\{1, ..., d\}$. For $z : [k] \to [n]$ and all $(a, b) \in [d] \times [d]$, we abuse notations and denote $z^{-1}(a, b) = \{(i, j) : z(i) = a, z(j) = b, i \neq j\}$. For any two label functions z, z', we write $z \sim z'$ if there exists a permutation σ of $\{1, ..., k\}$ such that $(z(\sigma(a)))_{a \leq k} = (z(a))_{a \leq k}$. For any set $\mathcal{S}$, we denote by $|\mathcal{S}|$ its cardinality. For any matrix $\boldsymbol{A}$, we denote by $\boldsymbol{A}_{ij}$ its entry on row i and column j. If $\boldsymbol{A} \in [0, 1]^{n \times n}$ and $\boldsymbol{A}$ is symmetric, we write $\boldsymbol{A} \in [0, 1]^{n \times n}_{\text{sym}}$. We denote by $\boldsymbol{A} \odot \boldsymbol{B}$ the Hadamard product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The Frobenius norm of a matrix $\boldsymbol{A}$ is denoted by $\|\boldsymbol{A}\|_2 = \sqrt{\sum_{i,j} A_{ij}^2}$. We denote by C and C' positive constants that can vary from line to line. These are absolute constants unless otherwise mentioned. For any two positive sequences $(a_n)_{n \in \mathbb{N}}$, $(b_n)_{n \in \mathbb{N}}$, we write $a_n = \omega(b_n)$ if $a_n / b_n \to \infty$.

2 Maximum likelihood estimation in the stochastic block model with missing links

2.1 Network model and missing data scheme

In the simplest situation, a network can be represented as undirected, unweighted graph with n nodes indexed from 1 to n. Then, the network can be encoded by its *adjacency matrix* $\boldsymbol{A} = (A_{ij})$. The adjacency matrix is a $n \times n$ symmetric matrix such that for any $i < j$, $\boldsymbol{A}_{ij} = 1$ if there exists an edge between node i and node j, $\boldsymbol{A}_{ij} = 0$ otherwise. We consider that there is no edge linking a node to itself, so $\boldsymbol{A}_{ii} = 0$ for any i. A common approach in network data analysis is to assume that the observations are random variables drawn from a probability distribution over the space of adjacency matrices. More precisely, for $i < j$ the variables $\boldsymbol{A}_{ij}$ are assumed to be independent Bernoulli random variables of parameter $\boldsymbol{\Theta}^*_{ij}$, where $\boldsymbol{\Theta}^* = (\boldsymbol{\Theta}^*_{ij})_{1 \leq i < j \leq n}$ is a $n \times n$ symmetric matrix with zero diagonal entries. The matrix $\boldsymbol{\Theta}^*$ corresponds to the matrix of probabilities of observing an edge between nodes i and j. This model is known as the *inhomogeneous random graph* model:

$$\forall 1 \leq i < j \leq n,\ \boldsymbol{A}_{ij} | \boldsymbol{\Theta}^*_{ij} \overset{ind.}{\sim} \text{Bernoulli}\left(\boldsymbol{\Theta}^*_{ij}\right). \tag{1}$$

Our focus is on the problem of estimation of the generative matrix $\boldsymbol{\Theta}^*$ which determines the overall structure of the network. This question is of particular interest for the task of link prediction.

Many of real-life networks are characterized by block structure. Loosely speaking, the block structure means that the nodes of the network are partitioned into groups called blocks, and that the distribution of the connections between nodes depends on the blocks to which the nodes belong. For example, when considering citation networks, where two articles are linked if one is cited by the other, it amounts to saying that the probability that two articles are linked only depends on their topic. Similarly, if one considers students of a school in a social network, it is a reasonable assumption to say that the probability that two students are linked only depends on their cohorts.

A very popular model that formalizes this idea is the stochastic block model (see, e.g., [26]). In this model, nodes are classified into k communities: each node i is associated with a community $z^*(i)$, where $z^* : [n] \to [k]$ is called the label function. This label function can either be treated as a parameter to estimate, or as a latent variable. In this last case, it is assumed that the indexes follow a multinomial distribution: $\forall i$, $z^*(i) \overset{i.i.d}{\sim} \text{Multinomial}(1; \alpha^*)$ where $\forall a \in [k]$, α_a is the probability that node i belongs to the community a. Given this label function, the probability that there exists an edge between nodes i and j depends only on the communities of i and j. Thus, the matrix of connection probabilities $\boldsymbol{\Theta}^*$ can be factorized as follows: $\boldsymbol{\Theta}^*_{ij} = \boldsymbol{Q}^*_{z^*(i)z^*(j)}$, with $\boldsymbol{Q}^*$ a $k \times k$ symmetric matrix such that $\boldsymbol{Q}^*_{ab}$ is the probability that there exists an edge between a given member of the community a and a given member of the community b. The conditional stochastic block model can be written as:

$$\begin{gathered} \exists \boldsymbol{Q}^* \in [0,1]^{k \times k}_{\text{sym}},\ \exists z^* : [n] \to [k] \\ \forall 1 \leq i < j \leq n,\ \boldsymbol{A}_{ij} | \left(\boldsymbol{Q}^*, z^*\right) \overset{ind.}{\sim} \text{Bernoulli}\left(\boldsymbol{Q}^*_{z^*(i)z^*(j)}\right),\ \boldsymbol{A}_{ii} = 0. \end{gathered} \tag{2}$$

Assuming that the network follows the stochastic block model, the problem of estimating the matrix of connection probabilities reduces to estimating the label function z^* and the matrix of probabilities of connections between communities $\boldsymbol{Q}^*$. Note that the conditional stochastic block model is at best identifiable up to a simultaneous permutation of the communities and of the rows and columns of the parameters $\boldsymbol{Q}^*$.

The stochastic block model has attracted considerable interest from the learning community. An important line of work has focused on the problem of estimation of the latent variables z^*, see, for example, [37, 9, 1, 39]. The best understood framework is the binary, balanced, symmetric, assortative block model. In this simpler model, the two communities have the same size, the same probability of intra-community connection ($\boldsymbol{Q}^*_{11} = \boldsymbol{Q}^*_{22} = p$), and nodes are assumed to be more connected with nodes of the same community ($p > q = \boldsymbol{Q}^*_{12}$). Much work has been done on the precise characterisation of the conditions on p, q that allow for strong recovery of z^*, i.e. to estimate z^* exactly with high probability. Closest to model (2) is perhaps the setting considered in [14]. In this work, the authors consider the related problem of community recovery in the binary block model [22],[2], and provide tight bounds on the recovery threshold for the balanced, two communities stochastic block model with missing observations. They propose a computationally efficient

algorithm for estimating z^* in regime where strong recovery is possible; this, however requires prior knowledge of the parameter $\boldsymbol{Q}^*$.

Missing observations scheme Usually, when working with network data, not all the edges are observed. To account for this situation we introduce $\boldsymbol{X} \in \{0,1\}^{n\times n}_{sym}$ the known sampling matrix where $\boldsymbol{X}_{ij} = 1$ if $\boldsymbol{A}_{ij}$ is observed and $\boldsymbol{X}_{ij} = 0$ otherwise. We assume that $\boldsymbol{X}$ is random and independent from the adjacency matrix $\boldsymbol{A}$ and its expectation $\boldsymbol{\Theta}^*$. For any $1 \leq i < j \leq n$, its entries $\boldsymbol{X}_{ij}$ are mutually independent and $\boldsymbol{X}_{ij} \overset{ind.}{\sim} \text{Bernoulli}(p)$ for some sampling rate $p \to 0$ such that $p = \omega\left(\log(n)/n\right)$ when $n \to \infty$.

2.2 Conditional maximum likelihood estimator

The log-likelihood of the parameters $(z, \boldsymbol{Q})$ with respect to the adjacency matrix $\boldsymbol{A}$ and the sampling matrix $\boldsymbol{X}$ is given by

$$\begin{aligned}\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z, \boldsymbol{Q}) &= \sum_{1\leq i<j\leq n} \boldsymbol{X}_{ij}\left(\boldsymbol{A}_{ij}\log(\boldsymbol{Q}_{z(i)z(j)}) + (1-\boldsymbol{A}_{ij})\log(1-\boldsymbol{Q}_{z(i)z(j)})\right)\\ &= \sum_{a\leq b}\log(\boldsymbol{Q}_{ab}) \sum_{(i,j)\in z^{-1}(a,b)} \boldsymbol{X}_{ij}\boldsymbol{A}_{ij} + \sum_{a\leq b}\log(1-\boldsymbol{Q}_{ab}) \sum_{(i,j)\in z^{-1}(a,b)} \boldsymbol{X}_{ij}(1-\boldsymbol{A}_{ij}).\end{aligned}$$

Let us denote by $\mathcal{Z}_{n,k}$ the set of all label functions $z : [n] \to [k]$. For a given label function $z \in \mathcal{Z}_{n,k}$, the log-likelihood is maximized by taking

$$\boldsymbol{Q}_{ab} = \frac{\sum_{(i,j)\in z^{-1}(a,b)} \boldsymbol{X}_{ij}\boldsymbol{A}_{ij}}{\sum_{(i,j)\in z^{-1}(a,b)} \boldsymbol{X}_{ij}}.$$

It is interesting to note that, for a fixed label function z, maximizing the likelihood or minimizing the least square criterion defined as $\mathcal{C}_{\boldsymbol{X}}(\boldsymbol{A}; z, \boldsymbol{Q}) = \sum_{i<j} \boldsymbol{X}_{ij}\left(\boldsymbol{A}_{ij} - \boldsymbol{Q}_{z(i),z(j)}\right)^2$ yields the same estimator for the matrix $\boldsymbol{Q}$. The main difference between these two methods is rooted in the label functions selected by the two criteria, see, e.g. [18].

To bound the risk of the maximum likelihood estimator, it is usual to assume that there exists sequences ρ_n and γ_n such that $\forall i < j$,

$$0 < \gamma_n \leq \boldsymbol{\Theta}^*_{ij} \leq \rho_n < 1. \tag{3}$$

This assumption ensures that the loss associated to the maximum likelihood estimator is Lipschitz continuous. See, for example, [6] and [48], where the authors assume that the adjacency matrix is generated by an homogeneous stochastic block model for which the matrix $\boldsymbol{Q}^*/\rho_n$ has entries bounded away from 0.

The restricted maximum likelihood estimator, $\widehat{\boldsymbol{\Theta}}$, is based on the maximization of the likelihood among block constant matrices with entries in $[\gamma_n, \rho_n]$:

$$\begin{aligned}&\widehat{\boldsymbol{\Theta}}_{i<j} = \widehat{\boldsymbol{Q}}_{\widehat{z}(i)\widehat{z}(j)},\ \widehat{\boldsymbol{\Theta}}_{ii} = 0\\ &(\widehat{\boldsymbol{Q}}, \widehat{z}) \in \underset{\boldsymbol{Q}\in[\gamma_n,\rho_n]^{k\times k}_{\text{sym}}, z\in\mathcal{Z}_{n,k}}{\arg\max} \mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z, \boldsymbol{Q}).\end{aligned} \tag{4}$$

In (4), γ_n and ρ_n are assumed to be known (see [18] for a discussion on how to estimate these parameters). Note that the Expectation-Maximization algorithm used in practice to obtain the variational approximation to the maximum likelihood estimator does not require the knowledge of these parameters. We also assume that k is known and that it can depend on the number of nodes n; it can be chosen using a network cross-validation method [11], a sequential goodness-of-fit testing procedure [32] or a likelihood-based model selection method [48]. The following result provides the upper bound on the estimation risk of the maximum likelihood estimator:

Theorem 1 (Corollary 2 in [18])**.** *Assume that $\boldsymbol{A}$ is drawn according to the conditional stochastic block model and $\rho_n = \omega(n^{-1})$. Then, there exists absolute constants $C, C' > 0$ such that, with probability at least $1 - 9\exp\left(-C\rho_n\left(k^2 + n\log(k)\right)\right)$,*

$$\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}\|_2^2 \leq C'\left(\frac{\rho_n^2}{((1-\rho_n)^2\wedge\gamma_n^2)}\right)\frac{\rho_n\left(k^2 + n\log(k)\right)}{p}. \tag{5}$$

When all network entries are observed, we have $p = 1$. Note that this results implies that, when $\rho_n = O(\gamma_n)$, the maximum likelihood estimator is minimax optimal (see, [28, 15] for a statement of the lower bound).

3 Variational approximation to the maximum likelihood estimator

3.1 Definition of the estimator

The optimization of the log-likelihood function $\mathcal{L}_{\boldsymbol{X}}$ requires a search over the set of k^n labels. As a consequence, the maximum likelihood estimator defined in (4) is computationally intractable. Celisse et al. [12] and Bickel et al. [6] are the first to study a variational approximation to this estimator. More recently, the authors of [46] used variational methods to approximate the maximum likelihood estimator in networks with missing observations. We start by formally introducing the variational approximation to the maximum likelihood estimator. We consider a stochastic block model with random labels with parameters $(\alpha, \boldsymbol{Q})$. For this model, the likelihood of the observed adjacency matrix $\boldsymbol{A}$ and sampling matrix $\boldsymbol{X}$ is given by

$$l_{\boldsymbol{X}}(\boldsymbol{A}; \alpha, \boldsymbol{Q}) = \sum_{z \in \mathcal{Z}_{n,k}} \left(\prod_{i \leq n} \alpha_{z(i)} \right) \exp\left(\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z, \boldsymbol{Q}) \right).$$

Note that the maximization of $l_{\boldsymbol{X}}$ still requires to evaluate the expectation of the label function z for given parameters $(\alpha, \boldsymbol{Q})$ by summing over k^n possible labels. To circumvent this problem, one can use the mean-field approximation, which amounts to approximating the posterior distribution $\mathbb{P}(\cdot | \boldsymbol{X} \odot \boldsymbol{A}, \alpha, \boldsymbol{Q})$ by a product distribution. To ensure that this product distribution remains close to the posterior distribution, the objective function is penalized by the Kullback-Leibler divergence of the two distributions. More precisely, the posterior distribution $\mathbb{P}(\cdot | \boldsymbol{X} \odot \boldsymbol{A}, \alpha, \boldsymbol{Q})$ is approximated by a multinomial distribution denoted $\mathbb{P}_\tau$, such that $\mathbb{P}_\tau(z) = \prod_{1 \leq i \leq n} m(z | \tau^i)$, where $m(\cdot | \tau^i)$ is the density of the multinomial distribution with parameter $\tau^i = (\tau_1^i, ..., \tau_k^i)$, and $\tau = (\tau^1, ..., \tau^n)$. Then, the variational estimator is defined as

$$\begin{aligned} \left(\widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}, \widehat{\tau}^{VAR} \right) &= \underset{\alpha \in \mathcal{A}, \boldsymbol{Q} \in \mathcal{Q}, \tau \in \mathcal{T}}{\arg\max} \mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A}; \tau, \alpha, \boldsymbol{Q}) \qquad (6) \\ \text{for} \quad \mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A}; \tau, \alpha, \boldsymbol{Q}) &= \log\left(l_{\boldsymbol{X}}(\boldsymbol{A}; \alpha, \boldsymbol{Q}) \right) - KL\left(\mathbb{P}_\tau(\cdot) || \mathbb{P}(\cdot | \boldsymbol{X} \odot \boldsymbol{A}, \alpha, \boldsymbol{Q}) \right) \end{aligned}$$

where $\mathcal{A}$, $\mathcal{Q}$ and $\mathcal{T}$ are the respective parameter spaces for the parameters α, $\boldsymbol{Q}$ and τ, KL denotes the Kullback-Leibler divergence between two distributions, and $\boldsymbol{X} \odot \boldsymbol{A}$ denotes the observed entries of $\boldsymbol{A}$. Since for any parameter $(\alpha, \boldsymbol{Q})$, $KL\left(\mathbb{P}_\tau(\cdot) || \mathbb{P}(\cdot | \boldsymbol{X} \odot \boldsymbol{A}, \alpha, \boldsymbol{Q}) \right) \geq 0$, we see that $\exp\left(\mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A}; \tau, \alpha, \boldsymbol{Q}) \right)$ provides a lower bound on $l_{\boldsymbol{X}}(\boldsymbol{A}; \alpha, \boldsymbol{Q})$.

The expectation - maximization (EM) algorithm derived in [46] can be used to iteratively approximate the variational estimator. This algorithm alternates between the following two steps :

- Estimation Step: given parameters $(\alpha, \boldsymbol{Q})$, the variational parameter τ maximizing $\mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A}; \tau, \alpha, \boldsymbol{Q})$ is given by the fixed point equation :

$$\tau_a^i = c_i \alpha_a \prod_{j \neq i : \boldsymbol{X}_{ij} = 1} \prod_{b \leq k} \left(\boldsymbol{Q}_{ab}^{\boldsymbol{A}_{ij}} (1 - \boldsymbol{Q}_{ab})^{1 - \boldsymbol{A}_{ij}} \right)^{\tau_b^j} \qquad \text{where } c_i \text{ is a normalizing constant;}$$

- Maximization Step: given parameter τ, the parameters $(\alpha, \boldsymbol{Q})$ maximizing $\mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A}; \tau, \alpha, \boldsymbol{Q})$ are given by

$$\alpha_a = \frac{\sum_i \tau_a^i}{n} \ , \ \boldsymbol{Q}_{ab} = \frac{\sum_{i \neq j} \boldsymbol{X}_{ij} \tau_a^i \tau_b^j \boldsymbol{A}_{ij}}{\sum_{i \neq j} \boldsymbol{X}_{ij} \tau_a^i \tau_b^j}.$$

Since this algorithm is not guaranteed to converge to a global maximum, it should be initialized with care, by using, for example, a first clustering step. This solution is implemented in the package `missSBM`.

Statistical guarantees for the variational estimator obtained in [12, 6, 36] establish that maximizing $\max_{\tau\in\mathcal{T}} \mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A};\tau,\alpha,\boldsymbol{Q})$ is equivalent to maximizing $l_{\boldsymbol{X}}(\boldsymbol{A};\alpha,\boldsymbol{Q})$, and that the estimator obtained by maximizing $l_{\boldsymbol{X}}(\boldsymbol{A};\alpha,\boldsymbol{Q})$ converges to the true parameters $(\alpha^*,\boldsymbol{Q}^*)$. This in turn implies that $(\widehat{\alpha}^{VAR},\widehat{\boldsymbol{Q}}^{VAR})$ also converges to $(\alpha^*,\boldsymbol{Q}^*)$. Note that these results do not provide guarantees on the recovery of the true labels z^* or on the matrix of connection probabilities $\boldsymbol{\Theta}^*$. In order to estimate $\boldsymbol{\Theta}^*$, we first define the label estimator $\widehat{z}^{VAR}$ using the minimizer of the objective function (6):

$$\forall\ i\leq n,\ \widehat{z}^{VAR}(i) \triangleq \underset{a\leq k}{\arg\max}\left(\widehat{\tau}^{VAR}\right)_a^i. \tag{7}$$

Once we have estimated the community labels using (7), we replace the estimator $\widehat{\boldsymbol{Q}}^{VAR}$ of the matrix of connection probabilities by the empirical mean estimator:

$$\forall a<k \text{ and } b<k, \widehat{\boldsymbol{Q}}^{ML-VAR}_{ab} \triangleq \frac{\sum_{(i,j)\in(\widehat{z}^{VAR})^{-1}(a,b)}\boldsymbol{X}_{ij}\boldsymbol{A}_{ij}}{\sum_{(i,j)\in(\widehat{z}^{VAR})^{-1}(a,b)}\boldsymbol{X}_{ij}}$$

$$\text{and define } \widehat{\boldsymbol{\Theta}}^{VAR} \text{ as } \quad \widehat{\boldsymbol{\Theta}}^{VAR}_{i\neq j} = \widehat{\boldsymbol{Q}}^{ML-VAR}_{\widehat{z}^{VAR}(i),\widehat{z}^{VAR}(j)},\ \widehat{\boldsymbol{\Theta}}^{VAR}_{ii}=0. \tag{8}$$

We will show respectively in Theorems 2 and 3 that this new estimator $\left(\widehat{z}^{VAR},\widehat{\boldsymbol{Q}}^{ML-VAR}\right)$ is minimax optimal for dense networks with missing observations as well as for sparse networks. The simulation study provided in Section 4 reveals that this estimator also has good performances in practice.

3.2 Convergence rates of variational approximation to the maximum likelihood estimator

In this section, we show the asymptotic equivalence of $\widehat{z}^{VAR}$ and $\widehat{z}$, where

$$(\widehat{\boldsymbol{Q}},\widehat{z}) \in \underset{\boldsymbol{Q}\in\mathcal{Q},z\in\mathcal{Z}_{n,k}}{\arg\max}\ \mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A};z,\boldsymbol{Q}) \tag{9}$$

is the maximum likelihood estimator. More precisely, we show that, with large probability, there exists a permutation σ of $\{1,...,k\}$ such that $\left(z^{VAR}(\sigma(a))\right)_{a\leq k} = (z(a))_{a\leq k}$ and $\left(\widehat{\boldsymbol{Q}}^{ML-VAR}_{\sigma(a),\sigma(b)}\right)_{a,b\leq k} = \left(\widehat{\boldsymbol{Q}}_{a,b}\right)_{a,b\leq k}$. When this hold, the tractable estimator $\left(\widehat{z}^{VAR},\widehat{\boldsymbol{Q}}^{ML-VAR}\right)$ is minimax optimal. These results are established under the following assumptions:

A.1 There exists $c>0$ and a compact interval $C_{\boldsymbol{Q}}\subset(0,1)$ such that $\mathcal{A}\subset[c,1-c]$ and $\mathcal{Q}\subset C_{\boldsymbol{Q}}^{k\times k}$;

A.2 The true parameters α^* and $\boldsymbol{Q}^*$ lie respectively in the interior of $\mathcal{A}$ and $\mathcal{Q}$;

A.3 The coordinates of $\alpha^*\boldsymbol{Q}^*$ are pairwise distinct.

Note that Assumption A.2 and A.3 are standard. Assumption A.2 requires that the true parameters lie in the interior of the parameter space, which is classical in parametric estimation. In the most simple case, the parameters α^* and $\boldsymbol{Q}^*$ lie respectively in the interior of sets $\mathcal{A}$ and $\mathcal{Q}$ of the form $\mathcal{A}=[c,1-c]$, $\mathcal{Q}=[c',1-c']^{k\times k}_{sym}$, for some $c,c'\in(0,1/2)$. Assumption A.3 ensures the identifiability of stochastic block model parameters. Then, under the assumption that $p=\omega\left(n/\log(n)\right)$, strong recovery of the labels is possible. Assumption A.1 is more restrictive, as it implies that the network is dense. This assumption will be relaxed in Theorem 3, where we consider sparse stochastic block models such that $\boldsymbol{Q}^*=\rho_n\boldsymbol{Q}^0$ for some fixed $\boldsymbol{Q}^0$ and some decreasing, sparsity inducing sequence ρ_n.

The following Theorem shows the minimax optimality of the tractable estimator $\widehat{\boldsymbol{\Theta}}^{VAR}$ under assumptions A.1 - A.3.

Theorem 2. *Assume that $\boldsymbol{A}$ is generated from a stochastic block model with parameters $(\alpha^*,\boldsymbol{Q}^*)$ satisfying assumptions A.1 - A.3. Then, $\mathbb{P}\left(\widehat{z}^{VAR}\sim\widehat{z}\right)\to 1$ when $n\to\infty$. Moreover, there exists a constant $C_{\boldsymbol{Q}^*}>0$ depending on $\boldsymbol{Q}^*$ such that*

$$\mathbb{P}\left(\left\|\boldsymbol{\Theta}^*-\widehat{\boldsymbol{\Theta}}^{VAR}\right\|_2^2 \leq \frac{C_{\boldsymbol{Q}^*}\left(k^2+n\log(k)\right)}{p}\right) \underset{n\to\infty}{\to} 1.$$

Let us now discuss the extension of Theorem 2 to the case of sparse networks. To avoid technicalities, we will consider the case when the network is fully observed. We will also assume that the proportions of different communities are held constant, while the probabilities of connections between communities may decreases at rate ρ_n. That is, the parameters $(\alpha^*, \boldsymbol{Q}^*)$ verify the following assumptions:

A.4 $\alpha^* = \alpha^0$ for some fixed α^0 such that $\alpha^0_a > 0$ for any $a \in \{1, ..., k\}$

A.5 $\boldsymbol{Q}^* = \rho_n \boldsymbol{Q}^0$ for some fixed $\boldsymbol{Q}^0 \in (0,1)^{k \times k}$ such that $\sum_{a,b=1}^{k} \alpha^0_a \alpha^0_b \boldsymbol{Q}^0_{ab} = 1$

Assumption A.5 relaxes Assumption A.1 and allows us consider sparse networks. The normalization constraint $\sum_{1 \leq a,b \leq k} \alpha^0_a \alpha^0_b \boldsymbol{Q}^0_{ab} = 1$ ensure the identifiability of the parameters $(\boldsymbol{Q}^0, \rho_n)$ (see [6]). In the following, we denote by $\mathcal{Q}$ the set of parameters $(\alpha, \boldsymbol{Q})$ verifying Assumptions A.4 and A.5.

The following theorem provides the analogous of Theorem 2 in the case of fully observed sparse networks. It is obtained by combining Propositions 2 and 3 in [18]:

Theorem 3. *Assume that $\boldsymbol{A}$ is fully observed, and is generated from a stochastic block model with parameters $(\alpha^*, \boldsymbol{Q}^*)$ satisfying Assumptions A.4 and A.5, such that $\boldsymbol{Q}^0$ has no identical columns and the sparsity inducing sequence ρ_n satisfies $\rho_n \gg \log(n)/n$. Then, $\mathbb{P}\left(\widehat{z}^{VAR} \sim \widehat{z}\right) \to 1$ when $n \to \infty$. Moreover, there exists a constant $C_{\boldsymbol{Q}^0} > 0$ depending on $\boldsymbol{Q}^0$ such that*

$$\mathbb{P}\left(\left\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}^{VAR}\right\|_2^2 \leq C_{\boldsymbol{Q}^0} \rho_n \left(k^2 + n \log(k)\right)\right) \underset{n \to \infty}{\to} 1. \tag{10}$$

Theorems 2 and 3 establish that the variational estimator $\widehat{\boldsymbol{\Theta}}^{VAR}$ is minimax optimal for both the estimation of dense networks with observations missing uniformly at random, and sparse networks. For proofs and discussion see Appendix B.1.

4 Numerical Results

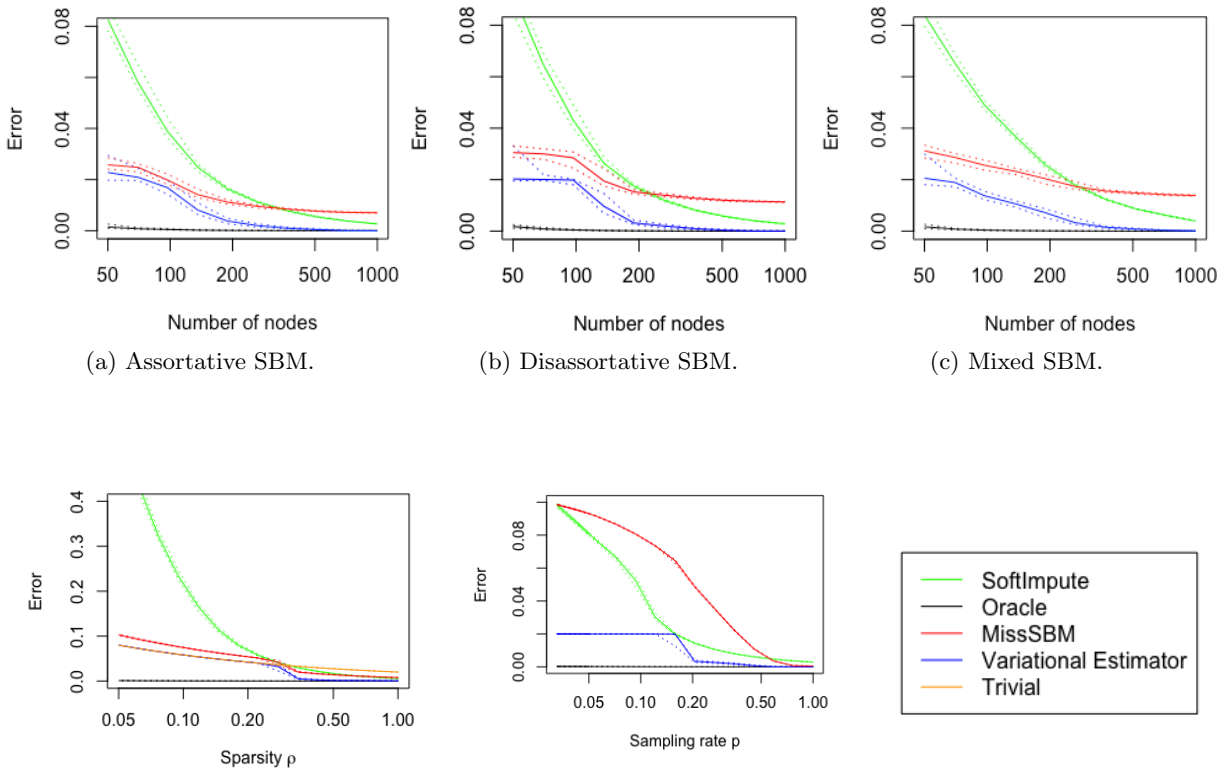
4.1 Synthetic data

In this section we provide a simulation study of the performances of the maximum likelihood estimator defined in (8), and compare it to the variational estimator defined in [46] and implemented in the package `missSBM`, as well as to the Universal Singular Value Thresholding estimator introduced in [24] and implemented in the package `softImpute`. The results are reported in Figure 1. Thorough descriptions of the simulation protocols are provided in the Appendix.

Dense stochastic block model First, we evaluate the empirical performances of the variational approximation of the maximum likelihood estimator defined in (8) on dense stochastic block models. We estimate the matrix of probabilities of connections, and we compare our estimator with the estimator given by the methods missSBM and softImpute. The quality of the inference is assessed by computing the squared Frobenius distance between the estimators and the true matrix of connection probabilities $\boldsymbol{\Theta}^*$.

We consider three types of three-communities stochastic block model. The first model, given by $(\alpha^{assort.}, \boldsymbol{Q}^{assort.})$, provides a simple assortative network, where individuals are more connected with people from their communities than with other individuals. On the contrary, the second model, given by $(\alpha^{disassort.}, \boldsymbol{Q}^{disassort.})$, is disassortative: individuals are more connected with individuals from outside of their communities. Both the assortative and disassortative models have balanced communities. The third model considered, given by $(\alpha^{mix.}, \boldsymbol{Q}^{mix.})$, exhibits neither assortativity nor disassortativity, and the communities are unbalanced. We introduce missing data by observing each entry of the adjacency matrix independently with probability 0.5.

The variational approximation to the maximum likelihood estimator defined in (8) outperforms the softImpute method across all models and all number of nodes. Its error is equivalent to that of the oracle estimator with hindsight knowledge of the true label function z^* when the network is a few hundred nodes large. Interestingly, our estimator also outperforms the variational estimator implement in the package



(a) Assortative SBM.

(b) Disassortative SBM.

(c) Mixed SBM.

(d) Robustness against sparsity.

(e) Robustness against missing observations.

Figure 1: Top : Error of connection probabilities estimation as a function of the number of nodes (top left : assortative SBM with balanced communities; top middle : disassortative SBM with balanced communities; top right : mixed SBM with unbalanced communities) or of the sparsity parameter ρ (bottom left) and of the sampling rate p (bottom right). We compare the variational approximation to the maximum likelihood estimator (in blue) to that of `missSBM` (in red), that of `softImpute` (in green), that of the oracle estimator with knowledge of the label z^* (in black), and that of the trivial estimator with entries equal to the empirical average degree divided by the number of nodes (orange, bottom only). The full lines indicate the median respectively of the mean squared error (top and bottom right) and of the mean squared error divided by the sparsity parameter ρ (bottom left) of the estimators over 100 repetitions, while the dashed lines indicate its 25% and 75% quantiles.

`missSBM`. We underline however that the primary focus of the missSBM method is to infer the parameters $(\alpha^*, \boldsymbol{Q}^*)$.

Additional experiments illustrating the strong consistency of the variational estimator canbe found in Appendix B.2.

Sparse stochastic block model Next, we investigate the behaviour of our estimator on increasingly sparse networks. We consider a three-communities assortative stochastic block model of 500 nodes with balanced communities, and 50% missing values. The probabilities of connections are given by $\boldsymbol{Q}^* = \rho \boldsymbol{Q}^0$, where ρ is a parameter controlling the sparsity, which ranges from 0.05 to 1. We compare the performance of the variational approximation to the maximum likelihood estimator to that of the methods softImpute and missSBM. We also compare these estimators to the trivial estimator with all entries equal to the average

degree divided by the number of nodes. The error is measured as the squared Frobenius distance between the estimator and the matrix $\mathbf{\Theta}^*$ divided by ρ^2.

As the network sparsity increases, the clustering of the nodes becomes more difficult. The normalized error of the estimator $\widehat{\mathbf{\Theta}}^{VAR}$ increases up to a threshold corresponding to the normalized error of the trivial estimator with all entries equal to the empirical degree, divided by the number of nodes. Note that when considering very sparse networks, with $\rho \ll \log(n)/n$, it is known that the trivial estimator with entries equal to the empirical mean degree is minimax optimal (see, eg, [28])). Thus, the estimator enjoys relatively low error rates in both high and low signal regime. By contrast, the normalized error of the softImpute method diverges as the network becomes increasingly sparse.

Stochastic block model with missing observations To conclude our simulation study, we evaluate the robustness of the methods against missing observations. We consider a three-communities assortative stochastic block model with balanced communities and 500 nodes. We increase the proportion of missing observations, and we compare the performance of the variational approximation to the maximum likelihood estimator to that of the methods softImpute and missSBM. The error is measured as the squared Frobenius distance between the estimator and the matrix $\mathbf{\Theta}^*$.

As the sampling rate p decreases, the clustering becomes impossible and the error rate of the estimator $\widehat{\mathbf{\Theta}}^{VAR}$ increases up to that of the trivial estimator obtained by averaging the observed entries of the adjacency matrix. By contrast, the methods softImpute and missSBM lack robustness against missing observations, and their error diverges as the number of missing observations increases.

4.2 Analysis of real networks

4.2.1 Prediction of interactions within a elementary school

We apply our algorithm to analyze a network of interactions within a French elementary school collected by the authors of [45]. The network records durations of physical interactions occurring within a primary school between 222 children divided into 10 classes and their 10 teachers over the course of two consecutive days; this dataset was collected using a system of sensors worn by the participants. We consider that an interaction has occurred if the corresponding duration is greater than one minute. If an interaction of less than one minute is observed, we consider that this observation may be erroneous, and treat the corresponding data as missing. By doing so, we remove respectively 11 and 13% of the observations on Day 1 and Day 2.

The graphs of interactions recorded during Day 1 and Day 2 can be considered as two outcomes of the same random network model characterized by the matrix of connection probabilities $\mathbf{\Theta}^*$. In this spirit, we use the observations collected on Day 1 estimate the matrix $\mathbf{\Theta}^*$, and evaluate those estimators on the network of interactions corresponding to Day 2. We note that the network of interactions for Day 1 has rather homogeneous degrees (the maximum degree is 41 and the minimum degree is 5, while the mean degree is 20). Moreover, it exhibits a strong community structure. Therefore, we expect the networks of interactions to be well approximated by a stochastic block model.

We compare the performance in terms of link prediction of the estimator $\widehat{\mathbf{\Theta}}^{VAR}$ defined in (8) to that of the method missSBM, and that of the method softImpute. In this last method, we set the penalty to 0, and we choose the rank of the estimator to be equal to the number of communities, which is estimated according to the Integrated Likelihood Criterion. We also compare these methods to the naive persistent estimator $\widehat{\mathbf{\Theta}}^{naive}$ given by $\widehat{\mathbf{\Theta}}^{naive}_{ij} = 1$ if an interaction between i and j has been recorded on Day 1, $\widehat{\mathbf{\Theta}}^{naive} = 0$ if no such interaction has been recorded, and $\widehat{\mathbf{\Theta}}^{naive}_{ij} = d/n$ if the information is missing, where d is the average degree of the graph for Day 1. Table 1 present the error of the different estimators, measured as the squared Frobenius distance between the adjacency matrix of Day 2 and its predicted value, divided by the squared Frobenius norm of the adjacency matrix of Day 2 (i.e, the error of the trivial null estimator).

The variational method predicts most accurately the interactions on Day 2. It is closely followed by the estimator provided by the package missSBM. By contrast to the simulation study, the reduction in error when using the new estimator is moderate : the error of $\widehat{\mathbf{\Theta}}^{VAR}$ is respectively 1.4% and 12.4% smaller than that of $\widehat{\mathbf{\Theta}}^{missSBM}$ and $\widehat{\mathbf{\Theta}}^{softImpute}$. In addition, the precision-recall curve presented in the Appendix indicates

Estimator	$\widehat{\boldsymbol{\Theta}}^{VAR}$	$\widehat{\boldsymbol{\Theta}}^{missSBM}$	$\widehat{\boldsymbol{\Theta}}^{SVT}$	$\widehat{\boldsymbol{\Theta}}^{naive}$
$\|\boldsymbol{X} \odot (\boldsymbol{A} - \widehat{\boldsymbol{\Theta}})\|_2^2 / \|\boldsymbol{X} \odot \boldsymbol{A}\|_2^2$	0.312	0.317	0.357	0.541

Table 1: Link prediction error on the network of interactions within a primary school.

that no estimator is better across all sensitivity levels. Interestingly, the naive estimator obtains a high error, which suggests a certain versatility in the children's behaviour.

4.2.2 Network of co-authorship

Finally, we use variational approximation to predict unobserved links in a network of co-authorship between scientists working on network analysis, first analysed in [41]. We discard the smallest connected components (with less than 5 nodes), and we obtain a network of 892 nodes. By contrast to the network of interaction in an elementary school, the network of co-authorship is quite sparse, and presents heterogeneous degrees: the average number of collaborators is 5, while the maximum and minimum number of collaborators are respectively 37 and 1.

In order to obtain unbiased estimates of the error of the estimators $\widehat{\boldsymbol{\Theta}}^{VAR}$, softImpute, and missSBM, we introduce 50% of missing values in the dataset. We train the three estimators on the observed entries of the adjacency matrix, and we use the unobserved entries to evaluate their imputation error. Table 2 present the mean imputation error of the different estimators over 100 random samplings, measured in term of squared Frobenius error and normalized by the squared Frobenius norm of the adjacency matrix of the remaining entries (i.e, the error of the null estimator). Here again, the variational approximation to the maximum

Estimator	$\widehat{\boldsymbol{\Theta}}^{VAR}$	$\widehat{\boldsymbol{\Theta}}^{missSBM}$	$\widehat{\boldsymbol{\Theta}}^{SVT}$
$\|(\mathbf{1} - \boldsymbol{X}) \odot (\boldsymbol{A} - \hat{\boldsymbol{\Theta}})\|_2^2 / \|(\mathbf{1} - \boldsymbol{X}) \odot \boldsymbol{A}\|_2^2$	0.857	0.869	0.894

Table 2: Imputation error of the estimators on the network of co-authorship.

likelihood estimator obtains the best performance. The precision-recall curves of these methods, included in the Appendix, indicates that this new estimator is preferable across almost all sensitivity levels. We underline however that the errors in term of Frobenius norm of the three estimators are close, and relatively high. This comes as no surprise, as the high sparsity of the network causes the link prediction problem to be difficult.

5 Conclusion

In this work, we have introduced a new tractable estimator based on variational approximation of the maximum likelihood estimator. We show that it enjoys the same convergence rates as the maximum likelihood estimator, and that it is therefore minimax optimal. Our simulation studies reveal the advantages of our estimator over current methods. In particular, they highlight its robustness against network sparsity and missing observations. Our results pave the way for analysing variational approximations of more general structured network models such as the latent block model.

Acknowledgments

We thank the anonymous referees for their helpfull comments.

References

[1] E. Abbe and C. Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, pages 670–688, 2015.

[2] Emmanuel Abbe, Afonso S. Bandeira, Annina Bracher, and Amit Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. IEEE Transactions on Network Science and Engineering, 1(1):10–22, 2014.

[3] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9(65):1981–2014, 2008.

[4] Arash A. Amini and Elizaveta Levina. On semidefinite relaxations for the block model. The Annals of Statistics, 46(1):149 – 179, 2018.

[5] O. Benyahia, C. Largeron, and B. Jeudy. Community detection in dynamic graphs with missing edges. 2017 11th International Conference on Research Challenges in Information Science (RCIS), pages 372–381, 2017.

[6] Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. The Annals of Statistics, 41(4):1922 – 1943, 2013.

[7] Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. Proceedings of the National Academy of Sciences, 106(50):21068–21073, 2009.

[8] Kevin Bleakley, Gérard Biau, and Jean-Philippe Vert. Supervised reconstruction of biological networks with local models. Bioinformatics, 23(13):i57–i65, 2007.

[9] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Nonbacktracking spectrum of random graphs: Community detection and nonregular Ramanujan graphs. The Annals of Probability, 46(1):1 – 71, 2018.

[10] Sourav Chatterjee. Matrix estimation by Universal Singular Value Thresholding. The Annals of Statistics, 43(1):177 – 214, 2015.

[11] Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. Journal of the American Statistical Association, 113:241 – 251, 2014.

[12] Jean-Jacques Daudin, Franck Picard, and Stéphane Robin. A mixture model for random graph. Statistics and Computing, 18:173–183, 2008.

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

[14] Souvik Dhara, Julia Gaudio, Elchanan Mossel, and Colin Sandon. Spectral recovery of binary censored block models. arXiv, 2021.

[15] Chao Gao, Yu Lu, Zongming Ma, and Harrison H. Zhou. Optimal estimation and completion of matrices with biclustering structures. Journal of Machine Learning Research, 17(1):5602–5630, 2016.

[16] Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-optimal graphon estimation. The Annals of Statistics, 43(6):2624 – 2652, 2015.

[17] Chao Gao, Aad W. van der Vaart, and Harrison H. Zhou. A general framework for Bayes structured linear models. The Annals of Statistics, 48(5):2848 – 2878, 2020.

[18] Solenne Gaucher and Olga Klopp. Maximum likelihood estimation of sparse networks with missing observations. Journal of Statistical Planning and Inference, 215:299–329, 2021.

[19] Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed K-means. Mathematical Statistics and Learning, 1(3):317–374, 2018.

[20] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences, 106(52):22073–22078, 2009.

[21] L. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 11(9):1074–1085, 1992.

[22] Bruce Hajek, Yihong Wu, and Jiaming Xu. Exact recovery threshold in the binary censored block model. In 2015 IEEE Information Theory Workshop - Fall (ITW), pages 99–103, 2015.

[23] Mark S. Handcock and Krista J. Gile. Modeling social networks from sampled data. The Annals of Applied Statistics, 4(1), 2010.

[24] Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. Journal of Machine Learning Research, 16(104):3367–3402, 2015.

[25] Jake M. Hofman and Chris H. Wiggins. Bayesian approach to network modularity. Physical Review Letters, 100:258701, 2008.

[26] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. Social Networks, 5(2):109 – 137, 1983.

[27] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. Machine Learning, 37(2):183–233, 1999.

[28] Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. The Annals of Statistics, 45(1):316 – 354, 2017.

[29] Olga Klopp and Nicolas Verzelen. Optimal graphon estimation in cut distance. Probability Theory and Related Fields, 174(3):1033–1090, 2019.

[30] Gueorgi Kossinets. Effects of missing data in social networks. Social Networks, 28(3):247–268, 2006.

[31] Meghana Kshirsagar, Jaime Carbonell, and Judith Klein-Seetharaman. Techniques to cope with missing data in host–pathogen protein interaction prediction. Bioinformatics, 28(18):i466–i472, 2012.

[32] Jing Lei. A goodness-of-fit test for stochastic block models. The Annals of Statistics, 44(1):401 – 424, 2016.

[33] Léa Longepierre and Catherine Matias. Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model. Electronic Journal of Statistics, 13(2):4157 – 4223, 2019.

[34] L. Lovász. Large Networks and Graph Limits. American Mathematical Society colloquium publications. American Mathematical Society, 2012.

[35] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 390(6):1150–1170, 2011.

[36] Mahendra Mariadassou and Timothée Tabouy. Consistency and asymptotic normality of stochastic block models estimators from sampled data. Electronic Journal of Statistics, 14(2):3672 – 3704, 2020.

[37] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing, STOC '14, page 694–703, New York, NY, USA, 2014. Association for Computing Machinery.

[38] F. McSherry. Spectral partitioning of random graphs. In Proceedings 42nd IEEE Symposium on Foundations of Computer Science, pages 529–537, 2001.

[39] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. Electronic Journal of Probability, 21(none):1 – 24, 2016.

[40] M. E. J. Newman. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23):8577–8582, 2006.

[41] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. Physical review E, 74(3):036104, 2006.

[42] Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. In Amos Storkey and Fernando Perez-Cruz, editors, Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 1579–1588. PMLR, 2018.

[43] Zahra S. Razaee, Arash A. Amini, and Jingyi Jessica Li. Matched bipartite block model with covariates. Journal of Machine Learning Research, 20(34):1–44, 2019.

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

[45] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. PLOS ONE, 6(8):1–13, 2011.

[46] Timothée Tabouy, Pierre Barbillon, and Julien Chiquet. Variational inference for stochastic block models from sampled data. Journal of the American Statistical Association, 115(529):455–466, 2020.

[47] M.J. Wainwright and M.I. Jordan. Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends in Machine Learning, 1(1-2):1–305, 2008.

[48] Y. X. Rachel Wang and Peter J. Bickel. Likelihood-based model selection for stochastic block models. The Annals of Statistics, 45(2):500 – 528, 2017.

[49] Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 5433–5442. PMLR, 2018.

[50] Bowen Yan and Steve Gregory. Finding missing edges in networks based on their community structure. Physical review. E, 85:056112, 2012.

[51] Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. The Annals of Statistics, 48(2):886 – 905, 2020.

[52] Anderson Y. Zhang and Harrison H. Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. The Annals of Statistics, 48(5):2575 – 2598, 2020.

[53] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighbourhood smoothing. Biometrika, 104(4):771–783, 2017.

[54] Yunpeng Zhao, Yun-Jhong Wu, Elizaveta Levina, and Ji Zhu. Link prediction for partially observed networks. Journal of Computational and Graphical Statistics, 26(3):725–733, 2017.

A Proofs

In this section, we prove Theorem 2. This proof follows to some extend that of Theorem 3, so we underline the main differences. Because of missing links, we introduce new techniques to compare the restricted and unrestricted maximum likelihood estimators. We also need to establish the strong consistency of the maximum likelihood estimator for the conditional SBM (in the full observation setting, this result is a direct consequence of [7]). Similarly, the proof of Theorem 3 relies heavily on the fact that the likelihood function at the parameters and the profile likelihood function at the parameters are asymptotically equivalent, which is a direct consequence of Lemma 3 [6]. This result does not hold under missing observations, and we develop new arguments to prove the strong consistency of the variational estimate of the labels.

A.1 Proof of Theorem 2

To prove Theorem 2, we first show that $\mathbb{P}\left(\cdot|\boldsymbol{X}\odot\boldsymbol{A},\widehat{\alpha}^{Var},\widehat{\boldsymbol{Q}}^{Var}\right)$, i.e. the posterior distribution of z at the variational estimator $(\widehat{\alpha}^{Var},\widehat{\boldsymbol{Q}}^{Var})$, concentrates around $\delta_{z'}$, the dirac distribution at some label function z' such that $z'\sim z^*$:

$$\mathbb{P}\left(z'|\boldsymbol{X}\odot\boldsymbol{A},\widehat{\alpha}^{Var},\widehat{\boldsymbol{Q}}^{Var}\right)=1-o_p(1). \tag{11}$$

Then, we show that it implies the concentration of the estimator $\widehat{z}^{Var}$:

$$\mathbb{P}\left(\widehat{z}^{Var}=z'|\boldsymbol{X}\odot\boldsymbol{A}\right)=1-o_p(1). \tag{12}$$

Since $\mathbb{P}\left(\widehat{z}^{Var}=z'|\boldsymbol{X}\odot\boldsymbol{A}\right)$ is bounded, this also implies that it converges to 1 in expectation :

$$\mathbb{P}\left(\widehat{z}^{Var}=z'\right)\to 1. \tag{13}$$

Finally, we show that with probability going to one,

$$\mathbb{P}\left(\widehat{z}\sim z^*\right)\to 1. \tag{14}$$

Combing Equations (12) and (14), we prove the first part of Theorem 2 :

$$\mathbb{P}\left(\widehat{z}\sim\widehat{z}^{VAR}\right)\to 1. \tag{15}$$

To establish the second part of Theorem 2, we show that the maximum likelihood estimator defined in (9) is equal to the restricted maximum estimator (4). Theorem 3 then follows from Theorem 1.

Define $c_{min}=\min_{a,b}\boldsymbol{Q}^*_{a,b}$ and $c_{max}=\max_{a,b}\boldsymbol{Q}^*_{a,b}$. Theorem 1 implies that for some absolute constant $C>0$,

$$\mathbb{P}\left(\left\|\boldsymbol{\Theta}^*-\widehat{\boldsymbol{\Theta}}^r\right\|_2^2\leq C(c_{max}/c_{min})^2\left(k^2+n\log(k)\right)\right)\to 1,$$

where the restricted maximum likelihood estimator $\widehat{\boldsymbol{\Theta}}^r$ is defined as

$$\begin{aligned}&\widehat{\boldsymbol{\Theta}}^r_{i<j}=\widehat{\boldsymbol{Q}}^r_{\widehat{z}^r(i)\widehat{z}^r(j)},\ \widehat{\boldsymbol{\Theta}}^r_{ii}=0\\&(\widehat{\boldsymbol{Q}}^r,\widehat{z}^r)\in\underset{\boldsymbol{Q}\in[c_{min}/2,2c_{max}]^{k\times k}_{\text{sym}},z\in\mathcal{Z}_{n,k}}{\arg\max}\sum_{i\neq j}\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}_{ij},\boldsymbol{Q}_{z(i)z(j)}).\end{aligned}$$

Now, Equation (15) implies that with probability going to one, the variational estimator of the probabilities of connections $\widehat{\boldsymbol{\Theta}}^{VAR}$ is equal to the maximum likelihood estimator $\widehat{\boldsymbol{\Theta}}$ given by

$$\begin{aligned}&\widehat{\boldsymbol{\Theta}}_{i<j}=\widehat{\boldsymbol{Q}}_{\widehat{z}(i)\widehat{z}(j)},\ \widehat{\boldsymbol{\Theta}}_{ii}=0\\&\text{for }(\widehat{\boldsymbol{Q}},\widehat{z})\in\underset{\boldsymbol{Q}\in\mathcal{Q},z\in\mathcal{Z}_{n,k}}{\arg\min}\sum_{i\neq j}\mathcal{K}(\boldsymbol{A}_{ij},\boldsymbol{Q}_{z(i)z(j)}).\end{aligned}$$

Thus, it is enough to show that $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}^r$ with large probability to prove the second part of Theorem 3. To do so, we show that

$$\mathbb{P}\left(Q(\widehat{z}) \in [c_{min}/2, 2c_{max}]^{k\times k}\right) \to 1. \tag{16}$$

Equation (16) implies that with probability going to 1, the maximum likelihood estimator of the probabilities of connections between nodes coincides $\widehat{\boldsymbol{\Theta}}$ with the restricted maximum likelihood estimator $\widehat{\boldsymbol{\Theta}}^r$. This concludes the proof of Theorem 3.

Proof of Equation (11)

For any $z \in \mathcal{Z}_{n,k}$ and $(\alpha, \boldsymbol{Q}) \in \mathcal{Q}$, let $l'_{\boldsymbol{X}}(\boldsymbol{A}, z; \alpha, \boldsymbol{Q}) = \left(\prod_{i\leq n} \alpha_{z(i)}\right) \exp\left(\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z, \boldsymbol{Q})\right)$ be the profile likelihood of the parameters $(z, \boldsymbol{Q})$. Then,

$$l'_{\boldsymbol{X}}(\boldsymbol{A}, z; \alpha, \boldsymbol{Q}) \leq \sup_{\tau\in\mathcal{T}} \exp\left(\mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A}; \tau, \alpha, \boldsymbol{Q})\right) \leq l_{\boldsymbol{X}}(\boldsymbol{A}; \alpha, \boldsymbol{Q}). \tag{17}$$

Let $z' = \arg\max_{z:z\sim z^*} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)$. By definition of $l_{\boldsymbol{X}}$,

$$l_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = \sum_{z\sim z'} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) + \sum_{z\not\sim z'} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right). \tag{18}$$

On the one hand, we bound the sum $\sum_{z\sim z'} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)$ using the following result, proven in [36] :

Proposition 1 (Proposition 6.11 in [36])**.** *For any $(\alpha, \boldsymbol{Q}) \in \mathcal{Q}$,*

$$\frac{\sum_{z\sim z^*} l'_{\boldsymbol{X}}(\boldsymbol{A}, z; \alpha, \boldsymbol{Q})}{l'_{\boldsymbol{X}}(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} = \#Sym(\alpha, \boldsymbol{Q}) \max_{z'\sim z^*} \frac{l'_{\boldsymbol{X}}(\boldsymbol{A}, z'; \alpha, \boldsymbol{Q})}{l'_{\boldsymbol{X}}(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} (1 + o_p(1))$$

where the $o_p(1)$ is uniform in $(\alpha, \boldsymbol{Q})$ and

$$Sym(\alpha, \boldsymbol{Q}) = \left\{\sigma \in \mathcal{S}_k : \left(\alpha_{\sigma(a)}\right)_{a\leq k} = (\alpha_a)_{a\leq k} \text{ and } \left(\boldsymbol{Q}_{\sigma(a),\sigma(b)}\right)_{a,b\leq k} = \left(\boldsymbol{Q}_{a,b}\right)_{a,b\leq k}\right\}$$

for $\mathcal{S}_k$ the set of permutations of $[k]$.

Now, with probability going to one, $(\widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR})$ exhibits no symmetry, i.e. $\#Sym(\widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}) = 1$ (see Section B.11 in [36] for a proof of this result). Then, Proposition 1 implies that

$$\sum_{z\sim z'} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)(1 + o_p(1))$$

which in turn implies

$$\sum_{z\sim z'} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) + l_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) o_p(1). \tag{19}$$

On the other hand, we bound the term $\sum_{z\not\sim z'} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)$ by combining the two following propositions from [36] :

Proposition 2 (Proposition 6.8 in [36])**.** *Let $(t_n)_{n\in\mathbb{N}}$ be a positive sequence such that $t_n \to 0$ and $pnt_n/\sqrt{\log(n)} \to +\infty$. Then, on an event of probability going to 1 and for n large enough,*

$$\sup_{(\alpha,\boldsymbol{Q})\in\mathcal{Q}} \sum_{z\notin S(z^*,t_n)} l'_{\boldsymbol{X}}(\boldsymbol{A}, z; \alpha, \boldsymbol{Q}) = o_p\left(l'_{\boldsymbol{X}}(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)\right)$$

where $S(z^*, t_n) = \{z \in \mathcal{Z}_{n,k} : \exists z' \sim z, \sum |z^*_i - z'_i| \leq nt_n\}$.

Proposition 3 (Proposition 6.10 in [36]). *There exists a positive constant C such that*

$$\sup_{(\alpha,\boldsymbol{Q})\in\mathcal{Q}} \sum_{z\in S(z^*,C), z\not\sim z^*} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \alpha, \boldsymbol{Q}\right) = o_p\left(l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*\right)\right).$$

Combining Propositions 2 and 3, we find that on a event of probability going to 1,

$$\sum_{z\not\sim z^*} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*\right) o_p(1).$$

Now, we use the definition of the variational estimator and Equation (17), and find that

$$l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*\right) \leq \sup_{\tau\in\mathcal{T}} \exp\left(\mathcal{J}_{\boldsymbol{X}}\left(\boldsymbol{A}; \tau, \alpha^*, \boldsymbol{Q}^*\right)\right) \leq \exp\left(\mathcal{J}_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\tau}^{VAR}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) \leq l_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right).$$

Thus,

$$\sum_{z\not\sim z^*} l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = l_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) o_p(1). \tag{20}$$

Combining Equations (18), (19) and (20), we find that

$$l_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) + l_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) o_p(1).$$

Dividing both sides by $l_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)$, we find that

$$\mathbb{P}\left(z'|\boldsymbol{X}\odot\boldsymbol{A}, \widehat{\alpha}^{Var}, \widehat{\boldsymbol{Q}}^{Var}\right) = \frac{l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l_{\boldsymbol{X}}\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)} = 1 + o_p(1)$$

which proves Equation (11).

Proof of Equation (12)
By definition of $\mathcal{J}_{\boldsymbol{X}}$,

$$KL\left(\mathbb{P}_{\widehat{\tau}^{VAR}}(\cdot)||\mathbb{P}\left(\cdot|\boldsymbol{X}\odot\boldsymbol{A}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) = \log\left(l_{\boldsymbol{X}}(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR})\right) - \mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A}; \widehat{\tau}^{VAR}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}).$$

Equation (17) implies that

$$\mathcal{J}_{\boldsymbol{X}}(\boldsymbol{A}; \widehat{\tau}^{VAR}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}) \geq \log\left(l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right)$$

so

$$KL\left(\mathbb{P}_{\widehat{\tau}^{VAR}}(\cdot)||\mathbb{P}\left(\cdot|\boldsymbol{X}\odot\boldsymbol{A}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) \leq \log\left(l_{\boldsymbol{X}}(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR})\right) - \log\left(l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right).$$

Note that Equation (11) implies

$$\log\left(l_{\boldsymbol{X}}(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR})\right) - \log\left(l'_{\boldsymbol{X}}\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) = o_p(1).$$

Now, using Pinsker's inequality, we see that

$$\left|\mathbb{P}_{\widehat{\tau}^{VAR}}(z') - \mathbb{P}\left(z'|\boldsymbol{X}\odot\boldsymbol{A}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right| = o_p(1).$$

We use Equation (11) and the definition of $\widehat{z}^{(VAR)}$ to conclude the proof of Equation (12).

Proof of Equation (14)

For $z \in \mathcal{Z}_{n,k}$, define

$$\begin{aligned}\Lambda(z) &= max_{\boldsymbol{Q}\in\mathcal{Q}}\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z, \boldsymbol{Q}) - \mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z^*, \boldsymbol{Q}^*) \quad \text{and} \\ \widetilde{\Lambda}(z) &= max_{\boldsymbol{Q}\in\mathcal{Q}}\mathbb{E}\left[\mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z, \boldsymbol{Q}) - \mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z^*, \boldsymbol{Q}^*)\middle| z^*\right].\end{aligned}$$

Moreover, for $z \in \mathcal{Z}_{n,k}$ and $(\alpha, \boldsymbol{Q})$, define

$$\|z - z^*\|_{\sim,0} = \min_{z':z'\sim z^*} \|z' - z^*\|_0$$

where $\|z' - z^*\|_0$ is the Hamming distance between the label functions z' and z^*.

To prove Equation (14), we will use the following results.

Proposition 4 (Equation (B.1) in [36]). *There exists a constant $c > 0$ such that on an event of probability going to one, for all positive sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \to 0$ and $pnt_n/\sqrt{\log(n)} \to +\infty$, $\forall z \notin S(z^*, t_n)$,*

$$\tilde{\Lambda}(z) \leq -\frac{3cpn^2 t_n \delta(\boldsymbol{Q}^*)}{4}$$

where and $\delta(\boldsymbol{Q}) = \min_{a,a'} \max_c KL\left(\boldsymbol{Q}_{ac}, \boldsymbol{Q}_{a'c}\right)$ and $S(z^*, t_n) = \left\{z \in \mathcal{Z}_{n,k} : \|z - z^*\|_{\sim,0} \leq nt_n\right\}$.

Proposition 5 (Proposition 6.7 in [36]). *There exists a constant $C_{\mathcal{Q}} > 0$ depending on $\mathcal{Q}$ such that for any sequence $(\epsilon_n)_{n\in\mathbb{N}}$ with $\epsilon_n < C_{\mathcal{Q}}$ and $\epsilon_n \geq k^2/(\sqrt{8}n)$,*

$$\sup_{z\in\mathcal{Z}_{n,k}} \left(\Lambda(z) - \tilde{\Lambda}(z)\right) = O_p(\epsilon_n n^2).$$

We choose $\epsilon_n = 3\delta(\boldsymbol{Q}^*)\log(n)/(8n)$. Then, Proposition 5 implies that there exists a constant $C > 0$ such that with probability going to 1, $\sup_{z\in\mathcal{Z}_{n,k}} \left(\Lambda(z) - \tilde{\Lambda}(z)\right) \leq C\epsilon_n n^2$. Moreover, we choose $t_n = 2C\log(n)/(cnp)$ and note that under the assumption $p \gg \log(n)/n$, $t_n \to 0$. Then, Propositions 4 and 5 imply that with probability going to one

$$\begin{aligned}\sup_{z\notin S(z^*,t_n)} \Lambda(z) &\leq \sup_{z\notin S(z^*,t_n)} \widetilde{\Lambda}(z) + \sup_{z\notin S(z^*,t_n)} \left(\Lambda(z) - \widetilde{\Lambda}(z)\right) \\ &\leq -\frac{3Cpn^2 t_n \delta(\boldsymbol{Q}^*)}{4} + \frac{3Cpn^2 t_n \delta(\boldsymbol{Q}^*)}{8} \\ &\leq -\frac{3Cn\log(n)\delta(\boldsymbol{Q}^*)}{8}.\end{aligned}$$

This implies in particular that

$$\mathbb{P}\left(\sup_{z\notin S(z^*,t_n)} \Lambda(z) < 0\right) \to 1. \tag{21}$$

We show a similar result for label functions z that are close to z^*. To do so, we use the following result.

Proposition 6 (Proposition 6.5 in [36]). *There exists a positive constant C such that on an event of probability going to 1, for all $z \in S(z^*, C)$,*

$$\tilde{\Lambda}(z) \leq -\frac{3cpn^2\delta(\boldsymbol{Q}^*)\|z - z^*\|_{\sim,0}}{4}.$$

We use Proposition 4, where we choose $\epsilon_n = k^2/n$. Then, there exists a constant $C' > 0$ such that with probability going to 1, $\sup_{z\in\mathcal{Z}_{n,k}} \left(\Lambda(z) - \widetilde{\Lambda}(z)\right) \leq C'nk^2$. Now, Proposition 6 implies that with probability going to 1,

$$\begin{aligned}\sup_{z\in S(z^*,C),z\not\sim z^*}\Lambda(z) &\leq \sup_{z\in S(z^*,C),z\not\sim z^*}\widetilde{\Lambda}(z)+\sup_{z\in S(z^*,C),z\not\sim z^*}\left(\Lambda(z)-\widetilde{\Lambda}(z)\right)\\ &\leq -\frac{3cpn^2\delta(\boldsymbol{Q}^*)}{4}+C'nk^2\\ &\leq nk^2\left(C'-\frac{3cpn\delta(\boldsymbol{Q}^*)}{8k^2}\right).\end{aligned}$$

Since $pn\to+\infty$, this implies that

$$\mathbb{P}\left(\sup_{z\in S(z^*,C),z\not\sim z^*}\Lambda(z)<0\right)\to 1. \tag{22}$$

Finally, since $t_n\to 0$, for n large enough $\mathcal{Z}_{n,k}=S(z^*,C)\cup\overline{S(z^*,t_n)}$. Thus, Equations (21) and (23) imply that

$$\mathbb{P}\left(\sup_{z\not\sim z^*}\Lambda(z)<0\right)\to 1. \tag{23}$$

Now, $\Lambda(z^*)=0$. Thus, with probability going to 1, $\arg\max\Lambda(z)\sim z^*$, so $\widehat{z}\sim z^*$.

Proof of Equation (16)

To prove Equation (16), we use Bernstein's inequality, which we recall here for sake of completeness :

Theorem 4 (Bernstein's inequality)**.** *Let $X_1,...,X_n$ be independent centered random variables. Assume that for any $i\in[n]$, $|X_i|\leq M$ almost surely, then*

$$\mathbb{P}\left(\left|\sum_{1\leq i\leq n}X_i\right|\geq\sqrt{2t\sum_{1\leq i\leq n}\mathbb{E}[X_i^2]}+\frac{2M}{3}t\right)\leq 2e^{-t}.$$

For $z\in\mathcal{Z}_{n,k}$ and $(a,b)\in[k]^2$, define

$$n_{ab}(z)=\begin{cases}|(z)^{-1}(a)|\times|(z)^{-1}(b)| & \text{if } a\neq b\\ |(z)^{-1}(a)|\times\left(|(z)^{-1}(a)|-1\right) & \text{otherwise}\end{cases}$$

and

$$n_{ab}^{\boldsymbol{X}}(z)=\sum_{\substack{i\in z^{-1}(a),j\in z^{-1}(b)\\ i\neq j}}\boldsymbol{X}_{ij}$$

the number of entries and of observed entries of the adjacency matrix between nodes of the communities a and b, and $\boldsymbol{Q}(z)=(\boldsymbol{Q}(z)_{ab})$ such that $\boldsymbol{Q}(z)_{ab}=\left(\sum_{i\in z^{-1}(a),j\in z^{-1}(b)}\boldsymbol{X}_{ij}\boldsymbol{A}_{ij}\right)/n_{ab}^{\boldsymbol{X}}(z)$. With these notations, we note that $\widehat{\boldsymbol{Q}}=\boldsymbol{Q}(\widehat{z})$.

Note that $|(z^*)^{-1}(a)|$ is a sum of n independent Bernoulli random variables with mean α_a. Using Bernstein's inequality 4, we find that for any a,

$$\mathbb{P}\left(n\alpha_a-|(z^*)^{-1}(a)|\geq 0.5n\alpha_a\right)\leq 2e^{-n\alpha_a/16}.$$

Thus,

$$\mathbb{P}\left(\min_a|(z^*)^{-1}(a)|\leq 0.5n\min_a\alpha_a\right)\leq 2ke^{-n\min_a\alpha_a/16}.$$

Therefore, the event $\Omega=\left\{\min_{a,b}n_{a,b}(z^*)\geq n^2\min_a(\alpha_a)^2/5\right\}$ holds with probability going to 1.

Similarly, note that conditionally on z^*, $n_{ab}^{\boldsymbol{X}}(z^*)$ is a sum of $n_{ab}(z^*)$ independent Bernoulli variables with parameter p. Then, for any two $(a,b)\in[k]^2$, Bernstein's inequality 4 implies that

$$\mathbb{P}\left(|pn_{ab}(z^*)-n_{ab}^{\boldsymbol{X}}(z^*)|\geq 0.5pn_{ab}(z^*)\middle|z^*\right)\leq 2e^{-pn_{ab}(z^*)/16}.$$

Thus,

$$\mathbb{P}\left(\min_{a,b} n_{ab}^{\boldsymbol{X}}(z^*) \leq 0.5p \min_{a,b} n_{ab}(z^*) \big| z^*\right) \leq 2ke^{-p\min_{a,b} n_{ab}(z^*)/16}.$$

This implies that

$$\mathbb{P}\left(\min_{a,b} n_{ab}^{\boldsymbol{X}}(z^*) \leq 0.1n^2 p \min_a \alpha_a^2 \big| \Omega\right) \leq 2ke^{-pn^2 \min_a \alpha_a/80}.$$

Since $p \gg \log(n)/n$, the event $\Omega' = \{\forall (a,b) \in [k]^2, n_{ab}^{\boldsymbol{X}}(z^*) \geq 0.1n^2 p \min_a \alpha_a^2\}$ holds with probability going to 1.

Now, we show that on the event Ω', with large probability, $\boldsymbol{Q}(z^*) \in [c_{min}/2, 2c_{max}]^{k\times k}$. Recall that for any a, b, conditionally on z^* and $\boldsymbol{X}$, $n_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}(z^*)_{ab}$ is a sum of $n_{ab}^{\boldsymbol{X}}(z^*)$ independent Bernoulli random variables with mean $\boldsymbol{Q}_{ab}^*$. Then, Bernstein's inequality implies that for any $t > 0$

$$\mathbb{P}\left(\left|n_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}(z^*)_{ab} - n_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}_{ab}^*\right| \geq \sqrt{2tn_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}_{ab}^*} + \frac{2t}{3}\Big| z^*, \boldsymbol{X}\right) \leq 2e^{-t}.$$

Choosing $t = n_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}_{ab}^*/16$ yields

$$\mathbb{P}\left(\left|n_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}(z^*)_{ab} - n_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}_{ab}^*\right| \geq 0.5n_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}_{ab}^*\Big| z^*, \boldsymbol{X}\right) \leq 2e^{-n_{ab}^{\boldsymbol{X}}(z^*)\boldsymbol{Q}_{ab}^*/16}.$$

On the event Ω', this implies that

$$\mathbb{P}\left(\left|\boldsymbol{Q}(z^*)_{ab} - \boldsymbol{Q}_{ab}^*\right| \geq 0.5\boldsymbol{Q}_{ab}^*\Big|\Omega'\right) \leq 2e^{-n^2\boldsymbol{Q}_{ab}^*(\min_a \alpha_a)^2/160}.$$

A union bound yields

$$\mathbb{P}\left(\boldsymbol{Q}(z^*) \notin [c_{min}/2, 2c_{max}]^{k\times k}\Big|\Omega'\right) \leq 2k^2 e^{-n^2 \min_{a,b} \boldsymbol{Q}_{ab}^*(\min_a \alpha_a)^2/160}.$$

Since $\mathbb{P}(\Omega') \to 1$, this shows that

$$\mathbb{P}\left(\boldsymbol{Q}(z^*) \in [c_{min}/2, 2c_{max}]^{k\times k}\right) \to 1.$$

Now, Equation (14) shows that with probability going to 1, $\widehat{z} \sim z^*$. Thus,

$$\mathbb{P}\left(\boldsymbol{Q}(\widehat{z}) \in [c_{min}/2, 2c_{max}]^{k\times k}\right) \to 1.$$

A.2 Proof of Theorem 3

In the case of fully observed network, we alleviate notations and write

$$\begin{aligned}
\mathcal{L}(\boldsymbol{A}; z, \boldsymbol{Q}) &= \sum_{i\neq j} \boldsymbol{A}_{ij} \log\left(\boldsymbol{Q}_{z(i),z(j)}\right) + (1-\boldsymbol{A}_{ij}) \log\left(1-\boldsymbol{Q}_{z(i),z(j)}\right), \\
l(\boldsymbol{A}; \alpha, \boldsymbol{Q}) &= \sum_{z\in\mathcal{Z}_{n,k}} \left(\prod_i \alpha_{z(i)}\right) \exp\left(\mathcal{L}(\boldsymbol{A}; z, \boldsymbol{Q})\right), \\
\text{and } \mathcal{J}(\boldsymbol{A}; \tau, \alpha, \boldsymbol{Q}) &= \log\left(l(\boldsymbol{A}; \alpha, \boldsymbol{Q})\right) - KL\left(\mathbb{P}_\tau(\cdot) || \mathbb{P}(\cdot|\boldsymbol{A}, \alpha, \boldsymbol{Q})\right).
\end{aligned}$$

For any $z \in \mathcal{Z}_{n,k}$ and $(\alpha, \boldsymbol{Q}) \in \mathcal{Q}$, we denote

$$l'(\boldsymbol{A}, z; \alpha, \boldsymbol{Q}) = \left(\prod_{i\leq n} \alpha_{z(i)}\right) \exp\left(\mathcal{L}(\boldsymbol{A}; z, \boldsymbol{Q})\right)$$

the likelihood of the parameters $(\alpha, \boldsymbol{Q})$ and the label function z. Then, the likelihood of the stochastic block model with parameters $(\alpha, \boldsymbol{Q})$ is given by $l(\boldsymbol{A}; \alpha, \boldsymbol{Q}) = \sum_{z\in\mathcal{Z}_{n,k}} l'(\boldsymbol{A}, z; \alpha, \boldsymbol{Q})$. Note that the likelihood

functions $l(\boldsymbol{A};\alpha,\boldsymbol{Q})$ and $l'(\boldsymbol{A},z;\alpha,\boldsymbol{Q})$ provide lower and upper bounds on the variational objective function $\mathcal{J}(\boldsymbol{A};\tau,\alpha,\boldsymbol{Q})$: for any parameter $(\alpha,\boldsymbol{Q})$ and any label function $z \in \mathcal{Z}_{n,k}$,

$$l'(\boldsymbol{A},z;\alpha,\boldsymbol{Q}) \leq \sup_{\tau\in\mathcal{T}} \exp\left(\mathcal{J}(\boldsymbol{A};\tau,\alpha,\boldsymbol{Q})\right) \leq l(\boldsymbol{A};\alpha,\boldsymbol{Q}). \tag{24}$$

To prove Proposition 3, we first show that $\mathbb{P}\left(\cdot|\boldsymbol{A},\widehat{\alpha}^{Var},\widehat{\boldsymbol{Q}}^{Var}\right)$, i.e. the posterior distribution of z at the variational estimator $(\widehat{\alpha}^{Var},\widehat{\boldsymbol{Q}}^{Var})$, concentrates around $\delta_{z'}$, the dirac distribution at the label function $z' = \arg\max_{z:z\sim z^*} l'\left(\boldsymbol{A},z;\widehat{\alpha}^{VAR},\widehat{\boldsymbol{Q}}^{VAR}\right)$:

$$\mathbb{P}\left(z'|\boldsymbol{A},\widehat{\alpha}^{Var},\widehat{\boldsymbol{Q}}^{Var}\right) = 1 - o_p(1). \tag{25}$$

Then, we show that it implies the concentration of the estimator $\widehat{z}^{Var}$:

$$\mathbb{P}\left(\widehat{z}^{Var} = z'|\boldsymbol{A}\right) = 1 - o_p(1). \tag{26}$$

Together (25) and (26) imply $\mathbb{P}\left(\widehat{z}^{Var} \sim z^*|\boldsymbol{A}\right) = 1 - o_p(1)$. Since the random variable $\mathbb{P}\left(\widehat{z}^{Var} \sim z^*|\boldsymbol{A}\right)$ is bounded, Equation (26) also implies that it converges to 1 in expectation. Finally, we show that with probability going to one, the maximum likelihood estimator of the label function is equal to the true label function (up to permutation):

$$\mathbb{P}\left(\widehat{z} \sim z^*\right) = 1 - o_p(1) \tag{27}$$

which concludes the proof of the first part of Theorem 3.

To prove the second part of Theorem 3, we show that the maximum likelihood estimator studied in Proposition 3 is equal to the restricted maximum estimator studied in Theorem 1. More precisely, define $c_{min} = \min_{a,b} \boldsymbol{Q}^0_{a,b}$ and $c_{max} = \max_{a,b} \boldsymbol{Q}^0_{a,b}$. Theorem 1 implies that for some absolute constant $C > 0$,

$$\mathbb{P}\left(\left\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}^r\right\|_2^2 \leq C(c_{max}/c_{min})^2\rho_n\left(k^2 + n\log(k)\right)\right) \to 1,$$

where the restricted maximum likelihood estimator $\widehat{\boldsymbol{\Theta}}^r$ is defined as

$$\widehat{\boldsymbol{\Theta}}^r_{i<j} = \widehat{\boldsymbol{Q}}^r_{\widehat{z}^r(i)\widehat{z}^r(j)},\ \widehat{\boldsymbol{\Theta}}^r_{ii} = 0$$
$$(\widehat{\boldsymbol{Q}}^r,\widehat{z}^r) \in \underset{\boldsymbol{Q}\in[c_{min}\rho_n/2,2c_{max}\rho_n]^{k\times k}_{\text{sym}},z\in\mathcal{Z}_{n,k}}{\arg\min} \sum_{i\neq j} \mathcal{K}(\boldsymbol{A}_{ij},\boldsymbol{Q}_{z(i)z(j)}).$$

One the other hand, Proposition 3 implies that with probability going to one, the variational estimator of the probabilities of connections $\widehat{\boldsymbol{\Theta}}^{VAR}$ is equal to the maximum likelihood estimator $\widehat{\boldsymbol{\Theta}}$ given by

$$\widehat{\boldsymbol{\Theta}}_{i<j} = \widehat{\boldsymbol{Q}}_{\widehat{z}(i)\widehat{z}(j)},\ \widehat{\boldsymbol{\Theta}}_{ii} = 0$$
$$\text{for } (\widehat{\boldsymbol{Q}},\widehat{z}) \in \underset{\boldsymbol{Q}\in\mathcal{Q},z\in\mathcal{Z}_{n,k}}{\arg\min} \sum_{i\neq j} \mathcal{K}(\boldsymbol{A}_{ij},\boldsymbol{Q}_{z(i)z(j)}).$$

We show that

$$\mathbb{P}\left(\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}^r\right) \to 1, \tag{28}$$

which concludes the proof of Theorem 3.

Proof of Equation (25)

The proof of Equation (25) relies on results proven in [6], which we recall for the sake of completeness. For any two parameters $(\alpha,\boldsymbol{Q})$ and $(\alpha',\boldsymbol{Q}')$ in $\mathcal{Q}$, we say that $(\alpha',\boldsymbol{Q}') \in \mathcal{S}_{\alpha,\boldsymbol{Q}}$ if there exists a permutation σ of $\{1,...,k\}$ such that for any $(a,b) \in \{1,...,k\}^2$, $\boldsymbol{Q}'_{\sigma(a),\sigma(b)} = \boldsymbol{Q}_{a,b}$ and $\alpha'_{\sigma(a)} = \alpha_a$.

Theorem 5 (Theorem 1 in [6]). *Let (z^*, A) be generated from a stochastic block model with parameters $(\alpha^*, \boldsymbol{Q}^*) \in \mathcal{Q}$ such that $\boldsymbol{Q}^0$ has no identical columns and $\rho_n \gg \log(n)/n$. Then, for any $(\alpha, \boldsymbol{Q}) \in \mathcal{Q}$,*

$$\frac{l(\boldsymbol{A}; \alpha, \boldsymbol{Q})}{l(\boldsymbol{A}; \alpha^*, \boldsymbol{Q}^*)} = \max_{(\alpha', \boldsymbol{Q}') \in \mathcal{S}_{\alpha, \boldsymbol{Q}}} \frac{l'(\boldsymbol{A}, z^*; \alpha', \boldsymbol{Q}')}{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} \left(1 + \epsilon_n\left(\left(\alpha', \boldsymbol{Q}'\right), k\right)\right) + \epsilon_n\left(\left(\alpha', \boldsymbol{Q}'\right), k\right)$$

where $\sup_{(\alpha, \boldsymbol{Q}) \in \mathcal{Q}} \epsilon_n((\alpha, \boldsymbol{Q}), k) = o_p(1)$.

Proposition 7 (Lemma 3 in [6]). *Let (z^*, A) be generated from a stochastic block model with parameters $(\alpha^*, \boldsymbol{Q}^*) \in \mathcal{Q}$ such that $\boldsymbol{Q}^0$ has no identical columns and $\rho_n \gg \log(n)/n$. Then,*

$$\frac{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)}{l(\boldsymbol{A}; \alpha^*, \boldsymbol{Q}^*)} = 1 + o_p(1).$$

Recall that $z' = \arg\max_{z: z \sim z^*} l'\left(\boldsymbol{A}, z^*; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)$. By definition of l and l',

$$\sum_{z \neq z'} l'\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) - l'\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right).$$

Thus

$$\frac{\sum_{z \neq z'} l'\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} = \frac{l(\boldsymbol{A}; \alpha^*, \boldsymbol{Q}^*)}{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} \times \frac{l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l(\boldsymbol{A}; \alpha^*, \boldsymbol{Q}^*)} - \frac{l'\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} \quad (29)$$

Using Proposition 7, we have that

$$\frac{l(\boldsymbol{A}; \alpha^*, \boldsymbol{Q}^*)}{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} = 1 + o_p(1). \quad (30)$$

Moreover, we note that

$$\begin{aligned} \max_{(\alpha', \boldsymbol{Q}') \in \mathcal{S}_{\widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}}} l'(\boldsymbol{A}, z^*; \alpha', \boldsymbol{Q}') &= \max_{z \sim z^*} l'\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) \\ &= l'\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) \end{aligned}$$

by the definition of z'. Then, applying Theorem 5, we get that

$$\frac{l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l(\boldsymbol{A}; \alpha^*, \boldsymbol{Q}^*)} = \frac{l'\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} (1 + o_p(1)) + o_p(1). \quad (31)$$

Combining Equations (29), (30) and (31), we obtain that

$$\frac{\sum_{z \neq z'} l'\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} = \frac{l'\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*)} o_p(1) + o_p(1).$$

Thus,

$$\sum_{z \neq z'} l'\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = \max\left\{l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*), l'\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right\} o_p(1). \quad (32)$$

On the one hand, using Equation (24) and the definition of $(\widehat{\tau}^{VAR}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR})$, we find that

$$\begin{aligned} l'(\boldsymbol{A}, z^*; \alpha^*, \boldsymbol{Q}^*) &\leq \sup_{\tau \in \mathcal{T}} \exp\left(\mathcal{J}(\boldsymbol{A}; \tau, \alpha^*, \boldsymbol{Q}^*)\right) \\ &\leq \exp\left(\mathcal{J}\left(\boldsymbol{A}; \widehat{\tau}^{VAR}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) \\ &\leq l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right). \end{aligned}$$

Also, by the definition of l and l', we have that $l'\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) \leq \left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)$. Thus, Equation (32) implies

$$\sum_{z \neq z'} l'\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) = l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) o_p(1). \tag{33}$$

Now, we can conclude the proof of Equation (25) by noticing that

$$\begin{aligned}
\mathbb{P}\left(z' | \boldsymbol{A}, \widehat{\alpha}^{Var}, \widehat{\boldsymbol{Q}}^{Var}\right) &= \frac{l'\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)} \\
&= 1 - \frac{\sum\limits_{z \neq z'} l'\left(\boldsymbol{A}, z; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}{l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)}
\end{aligned}$$

and using Equation (33).

Proof of Equation (26) By the definition of $\mathcal{J}\left(\boldsymbol{A}; \tau, \alpha, \boldsymbol{Q}\right)$, we have that

$$KL\left(\mathbb{P}_{\widehat{\tau}^{VAR}}(\cdot) || \mathbb{P}\left(\cdot | \boldsymbol{A}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) = \log\left(l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) - \mathcal{J}\left(\boldsymbol{A}; \widehat{\tau}^{VAR}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right).$$

Equation (24) implies that $\mathcal{J}\left(\boldsymbol{A}; \widehat{\tau}^{VAR}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right) \geq \log\left(l\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right)$, so

$$KL\left(\mathbb{P}_{\widehat{\tau}^{VAR}}(\cdot) || \mathbb{P}\left(\cdot | \boldsymbol{A}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) \leq \log\left(l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) - \log\left(l\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right).$$

Note that Equation (25) implies

$$\log\left(l\left(\boldsymbol{A}; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) - \log\left(l\left(\boldsymbol{A}, z'; \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right) = o_p(1).$$

Now, using Pinsker's inequality, we see that

$$\left|\mathbb{P}_{\widehat{\tau}^{VAR}}(z') - \mathbb{P}\left(z' | \boldsymbol{A}, \widehat{\alpha}^{VAR}, \widehat{\boldsymbol{Q}}^{VAR}\right)\right| = o_p(1).$$

We use Equation (25) and the definition of $\widehat{z}^{(VAR)}$ to conclude the proof of Equation (26).

Proof of Equation (27)
Equation (27) is proven in [7]. In this work, the authors define the profile likelihood modularity $\mathcal{Q}_{LM}(A, z)$ of a label function $z \in \mathcal{Z}_{n,k}$ as

$$\mathcal{Q}_{LM}(A, z) = \frac{1}{2}\sum_{a,b} n_{ab}\left(\frac{\boldsymbol{O}_{ab}}{n_{ab}}\log\left(\frac{\boldsymbol{O}_{ab}}{n_{ab}}\right) + \left(1 - \frac{\boldsymbol{O}_{ab}}{n_{ab}}\right)\log\left(1 - \frac{\boldsymbol{O}_{ab}}{n_{ab}}\right)\right).$$

for $\boldsymbol{O}_{ab} = \sum\limits_{i \in z^{-1}(a), j \in z^{-1}(b)} \boldsymbol{A}_{ij}$ and

$$n_{ab} = \begin{cases} |z^{-1}(a)| \times |z^{-1}(b)| & \text{if } a \neq b \\ |z^{-1}(a)| \times \left(|z^{-1}(a)| - 1\right) & \text{otherwise} \end{cases}$$

For $\hat{z}^{LM} = \arg\max_{z \in \mathcal{Z}_{n,k}} \mathcal{Q}_{LM}(A, z)$, the authors of [7] prove that under the assumptions of Proposition 3, with probability going to 1, $\hat{z}^{LM} \sim z^*$. Since maximizing $\mathcal{Q}_{LM}(A, z)$ is equivalent to maximizing $\max_{\boldsymbol{Q}} \mathcal{L}\left(\boldsymbol{A}; \boldsymbol{Q}, z\right)$, this implies that $\widehat{z} \sim z^*$ with probability going to 1.

Proof of Equation (28) To do so, we show that with large probability, $\boldsymbol{Q}(\widehat{z}) \in [c_{min}\rho_n/2, 2c_{max}\rho_n]^{k\times k}]$. We define

$$n_{ab}(z) = \begin{cases} |z^{-1}(a)| \times |z^{-1}(b)| & \text{if } a \neq b \\ |z^{-1}(a)| \times \left(|z^{-1}(a)| - 1\right) & \text{otherwise} \end{cases}$$

for $z \in \mathcal{Z}_{n,k}$, and $\boldsymbol{Q}(z) = (\boldsymbol{Q}(z)_{ab})$ such that $\boldsymbol{Q}(z)_{ab} = \left(\sum_{i\in z^{-1}(a), j\in z^{-1}(b)} \boldsymbol{A}_{ij}\right)/n_{ab}(z)$. With these notations, we note that $\widehat{\boldsymbol{Q}} = \boldsymbol{Q}(\widehat{z})$.

Recall that $|(z^*)^{-1}(a)|$ is a sum of n independent Bernoulli random variables with mean α_a^0. Using Bernstein's inequality 4, we find that for any a,

$$\mathbb{P}\left(n\alpha_a^0 - |(z^*)^{-1}(a)| \geq 0.5n\alpha_a^0\right) \leq 2e^{-n\alpha_a^0/16}.$$

Thus,

$$\mathbb{P}\left(\min_a |(z^*)^{-1}(a)| \leq 0.5n \min_a \alpha_a^0\right) \leq 2ke^{-n\min_a \alpha_a^0/16}.$$

Therefore, the event $\Omega = \left\{\min_{a,b} n_{a,b}(z^*) \geq n^2 \min_a(\alpha_a^0)^2/5\right\}$ holds with probability going to 1.

Now, we show that on the event Ω, with large probability, $\boldsymbol{Q}(z^*) \in [c_{min}\rho_n/2, 2c_{max}\rho_n]^{k\times k}$. Recall that for any a, b, conditionally on z^*, $n_{ab}(z^*)\boldsymbol{Q}(z^*)_{ab}$ is a sum of $n_{ab}(z^*)$ independent Bernoulli random variables with mean $\rho_n\boldsymbol{Q}^0_{ab}$. Then, Bernstein's inequality 4 implies that for any $t > 0$

$$\mathbb{P}\left(\left|n_{ab}(z^*)\boldsymbol{Q}(z^*)_{ab} - n_{ab}(z^*)\rho_n\boldsymbol{Q}^0_{ab}\right| \geq \sqrt{2tn_{ab}(z^*)\rho_n\boldsymbol{Q}^0_{ab}} + \frac{2t}{3}\right) \leq 2e^{-t}.$$

Choosing $t = n_{ab}(z^*)\rho_n\boldsymbol{Q}^0_{ab}/16$ yields

$$\mathbb{P}\left(\left|n_{ab}(z^*)\boldsymbol{Q}(z^*)_{ab} - n_{ab}(z^*)\rho_n\boldsymbol{Q}^0_{ab}\right| \geq 0.5n_{ab}(z^*)\rho_n\boldsymbol{Q}^0_{ab}\right) \leq 2e^{-n_{ab}(z^*)\rho_n\boldsymbol{Q}^0_{ab}/16}.$$

On the event Ω, this implies that

$$\mathbb{P}\left(\left|n_{ab}(z^*)\boldsymbol{Q}(z^*)_{ab} - n_{ab}(z^*)\rho_n\boldsymbol{Q}^0_{ab}\right| \geq 0.5n_{ab}(z^*)\rho_n\boldsymbol{Q}^0_{ab}\right) \leq 2e^{-n^2\rho_n\boldsymbol{Q}^0_{ab}(\min_a \alpha_a^0)^2/80}.$$

A union bound yields

$$\mathbb{P}\left(\boldsymbol{Q}(z^*) \notin [c_{min}\rho_n/2, 2c_{max}\rho_n]^{k\times k}\right) \leq 2k^2e^{-n^2\rho_n \min_{a,b}\boldsymbol{Q}^0_{ab}(\min_a \alpha_a^0)^2/80}$$

on the event Ω. Since $\mathbb{P}(\Omega) \to 1$ and $n^2\rho_n \to +\infty$, this shows that

$$\mathbb{P}\left(\boldsymbol{Q}(z^*) \in [c_{min}\rho_n/2, 2c_{max}\rho_n]^{k\times k}\right) \to 1.$$

Now, Equation (27) shows that with probability going to 1, $\widehat{z} \sim z^*$. Thus, $Q(\widehat{z}) \in [c_{min}\rho_n/2, 2c_{max}\rho_n]^{k\times k}$ with probability going to one, and the maximum likelihood estimator of the probabilities of connections between nodes coincides with the restricted maximum likelihood estimator. This concludes the proof of Equation (28).

B Further informations on the numerical experiments

B.1 Simulation protocol

In this section, we provide details on the simulation protocol for Section 4.1. The numerical experiments where conducted using R version 4.0.3, the package softImpute version 1.4.1, and the package missSBM version 0.3.0.

Dense stochastic block model The parameters used for the simulations are the following : $\alpha^{assort.} = \alpha^{disassort.} = (1/3, 1/3, 1/3)$, $\alpha^{mix.} = (0.1, 0.3, 0.6)$, and

$$\boldsymbol{Q}^{assort.} = \begin{pmatrix} 0.5 & 0.2 & 0.2 \\ 0.2 & 0.5 & 0.2 \\ 0.2 & 0.2 & 0.5 \end{pmatrix}, \ \boldsymbol{Q}^{disassort.} = \begin{pmatrix} 0.2 & 0.5 & 0.5 \\ 0.5 & 0.2 & 0.5 \\ 0.5 & 0.5 & 0.2 \end{pmatrix}, \ \boldsymbol{Q}^{mix.} = \begin{pmatrix} 0.1 & 0.5 & 0.3 \\ 0.5 & 0.2 & 0.4 \\ 0.3 & 0.4 & 0.6 \end{pmatrix}.$$

For each model and each number of nodes, we simulate 100 networks. For each networks, entries of the adjacency matrix are observed independently from one another with probability 1/2. Then, the matrix of connection probabilities $\boldsymbol{\Theta}^*$ is estimated using each method (variational approximation to the maximum likelihood estimator, missSBM, and softImpute). The oracle estimator is obtained as

$$\forall a < k \text{ and } b < k, \widehat{\boldsymbol{Q}}^*_{ab} \triangleq \frac{\sum_{i \in (z^*)^{-1}(a), j \in (z^*)^{-1}(b), i \neq j} \boldsymbol{X}_{ij} \boldsymbol{A}_{ij}}{\sum_{i \in (z^*)^{-1}(a), j \in (z^*)^{-1}(b), i \neq j} \boldsymbol{X}_{ij}}$$

Sparse stochastic block model The parameters $(\alpha, \boldsymbol{Q})$ of the stochastic block model are given by $\alpha = (1/3, 1/3, 1/3)$, and

$$\boldsymbol{Q} = \rho \begin{pmatrix} 0.5 & 0.2 & 0.2 \\ 0.2 & 0.5 & 0.2 \\ 0.2 & 0.2 & 0.5 \end{pmatrix}$$

for ρ ranging between 0.05 and 1. For each sparsity, we simulate 100 networks with 500 nodes. For each networks, entries of the adjacency matrix are observed independently from one another with probability 1/2. Then, the matrix of connection probabilities $\boldsymbol{\Theta}^*$ is estimated using each method (variational approximation to the maximum likelihood estimator, missSBM, softImpute, the oracle estimator and the naive estimator).

Stochastic block model with missing observations The parameters $(\alpha, \boldsymbol{Q})$ of the stochastic block model are given by
$\alpha = (1/3, 1/3, 1/3)$, and

$$\boldsymbol{Q} = \begin{pmatrix} 0.5 & 0.2 & 0.2 \\ 0.2 & 0.5 & 0.2 \\ 0.2 & 0.2 & 0.5 \end{pmatrix}$$

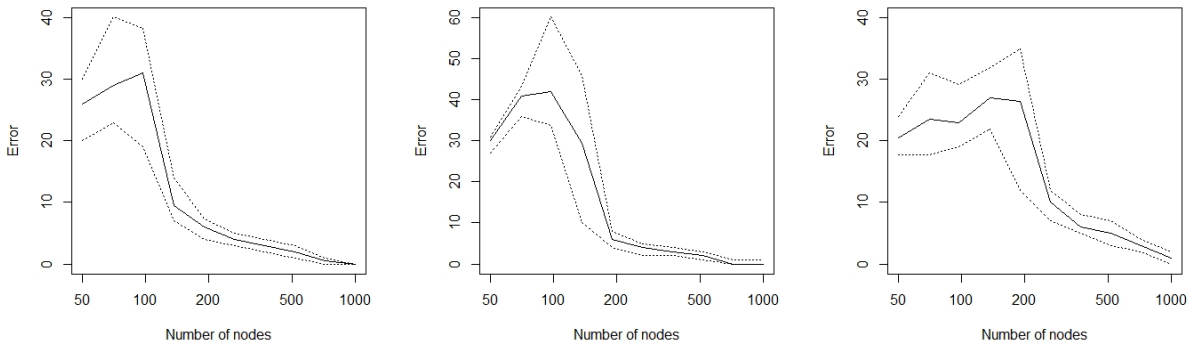
The proportion of observed entries p varies between 0.02 and 1. For each p, we simulate 100 networks with 500 nodes. For each networks, entries of the adjacency matrix are observed independently from one another with probability p. Then, the matrix of connection probabilities $\boldsymbol{\Theta}^*$ is estimated using each method (variational approximation to the maximum likelihood estimator, missSBM, softImpute, the oracle estimator and the naive estimator).

B.2 Empirical strong consistency of the variational estimator

We illustrate the empirical strong consistency of the variational estimator. Using the parameters chosen for simulating dense stochastic block models, we compute the number of misclassified nodes, defined as

$$\min_{z \sim \hat{z}} \left\{ \sum_i \mathbb{1}\left\{ z^*(i) \neq z(i) \right\} \right\}.$$

The total classification error for the assortative, dissasortartive and mixed models are presented in Figure 2. These simulations confirm that the variational estimator achieves strong recovery of the labels, even in unbalanced setting when neither assortative or disassortative behaviour are observed.



(a) Assortative SBM. (b) Disassortative SBM. (c) Mixed SBM.

Figure 2: Number of nodes misclassified by the variational estimator in the assortative SBM with balanced communities (left), in the disassortative SBM with balanced communities (middle), and in the mixed SBM with unbalanced communities (right). The full lines indicate the median of the number of misclassified nodes over 100 repetitions, while the dashed lines indicate its 25% and 75% quantiles.

B.3 Prediction of interactions within an elementary school

To compare the errors in term of link prediction of the methods missSBM and softImpute with that of our estimator, we plot the precision-recall curves of these estimators. More precisely, for any estimator $\widehat{\mathbf{\Theta}}$ of the matrix of connection probabilities $\mathbf{\Theta}^*$, and all thresholds $t \in [0,1]$, one can define the link-prediction estimator $\widehat{A}$ as follows : $\widehat{A}_{ij} = 1$ if and only if $\widehat{\mathbf{\Theta}}_{ij} \geq t$, that is, we predict that there exists a link between nodes i and j is the estimated probability that these nodes are connected is larger than the threshold t. The recall-precision curves obtained by varying this threshold is presented in Figure 3. We also represent the mean precision-recall curve of the baseline estimator obtained by predicting edges independently at random with an increasing probability.

The three methods used for link prediction obtain quite similar precision-recall curves. No single method is better across all sensitivity levels.

B.4 Prediction of collaboration in the co-authorship network

Similarly, we plot the precision-recall curves of the link-prediction methods obtained by using our new estimator, missSBM and softImpute. We also represent the mean precision-recall curve of the baseline estimator obtained by predicting edges independently at random with an increasing probability. The recall-precision curves is presented in Figure 4.

The precision-recall curve of the variational approximation to the maximum likelihood estimator is equivalent to or better than the other estimators across all sensitivity levels.

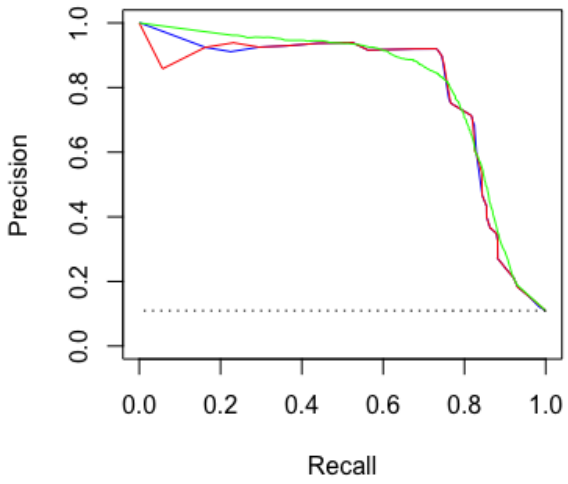


Figure 3: **Precision-recall curves for link prediction in the network of interactions within a school**: Precision-recall curves of the estimator obtained using `missSBM` (in red), of the estimator obtained using `softImpute` (in green), and of the variational approximation to the maximum likelihood estimator (in blue). The dotted black line represents the precision of the baseline estimator.

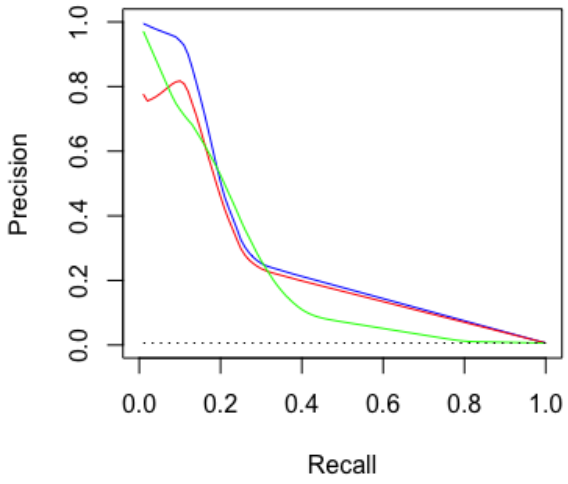


Figure 4: **Precision-recall curves for link prediction in the network co-authorship**: Precision-recall curves of the estimator obtained using `missSBM` (in red), of the estimator obtained using `softImpute` (in green), and of the variational approximation to the maximum likelihood estimator (in blue). The dotted black line represents the precision of the baseline estimator.