Optimality of variational inference for stochastic block model with missing links

Anonymous Author(s)
Affiliation
Address
email

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.

***Keywords*—** Network models, link prediction, variational methods, maximum likelihood

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, [32]. Recently, variational methods ([25, 45]) 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 [31] and sampled data [44]. Behind the estimation of global model parameters, variational inference has been successfully applied to the problem of community detection, see, e.g., [2, 49, 23, 41]. In particular, the authors of [49] show that an iterative Batch Coordinate Ascent Variational Inference algorithm attains the statistical optimality for community detection problem under an assortative stochastic block model.

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 [15] shows that, for the problem of estimating the matrix of connection probabilities, the least square estimator is minimax optimal and [16] provides optimal rate for Bayes estimation. For the more challenging case of sparse graphs, the minimax optimal rates have been first obtained in [26] building on the restricted least square estimator. In [14], 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 computationally expensive to be used in practice. Many other approaches have been proposed, for example, spectral clustering [36, 20, 42], modularity maximization [39, 7], belief propagation [13], neighbourhood smoothing [50], convex relaxation of k-means clustering [18] and of likelihood maximization [3], and universal singular value thresholding [10, 27, 47]. 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.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

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 [44] using the closeness of maximum likelihood and maximum variational likelihood and on the results that show the minimax optimality of the maximum likelihood estimator [17].

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, [29, 48, 21, 19]. 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 in practice that incomplete observation of the network structure may considerably affect the accuracy of inference methods [28] 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 [33, 51], 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 connected but are likely to be connected can be recommended as promising friends.

1.1 Constribution 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 any positive integer d, we denote by $[d]$ the set $\{1, ..., d\}$. 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 observations of 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. For simplicity, 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, it is assumed that the variables $(\boldsymbol{A}_{ij})_{1\le i<j\le n}$ are independent Bernoulli random variables of parameter $(\boldsymbol{\Theta}^*_{ij})_{1\le i<j\le n}$, where $\boldsymbol{\Theta}^* = (\boldsymbol{\Theta}^*_{ij})_{1\le i<j\le n}$ 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:

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

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 that we consider in the present paper.

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 citations 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., [24]). 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 (this model is sometimes called the conditional stochastic block model), or as a latent variable. In this case, 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 and $\alpha_a^* n$ is the expected size of 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. This implies that 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. Thus, the conditional stochastic block model can be written as:

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

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 $\boldsymbol{Q}^*$. 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 z^*, see, for example, [35, 9, 1, 38].

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 $\boldsymbol{Q}^*$. For any two label functions z, z', we write $z \sim z'$ if there exists a permutation σ of $\{1, ..., k\}$ such that $\left(z(\sigma(a))\right)_{a\le k} = \left(z(a)\right)_{a\le k}$.

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 the sampling matrix $\boldsymbol{X}$ is random and independent from the adjacency matrix $\boldsymbol{A}$ and its expectation $\boldsymbol{\Theta}^*$. For any $1 \le i < j \le n$, 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 $p \gg \log(n)/n$.

2.2 Conditional maximum likelihood estimator

In the conditional stochastic block model, 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

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

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

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

It is interesting to note that, for a fixed label function z, maximizing the likelihood or minimizing the least square criterion defined as $\mathcal{C}_{\boldsymbol{X}}\left(\boldsymbol{A}; z, \boldsymbol{Q}\right) = \sum_{i<j} \boldsymbol{X}_{ij}\left(\boldsymbol{A}_{ij} - \boldsymbol{Q}_{z(i),z(j)}\right)^2$ 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. [17].

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 $\forall i < j$,

$$0 < \gamma_n \leq \mathbf{\Theta}^*_{ij} \leq \rho_n < 1. \tag{2}$$

This assumption ensures that the loss associated to the maximum likelihood estimator is Lipschitz continuous. See, for example, [6] and [46], 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, $\widehat{\mathbf{\Theta}}$, is based on the maximization of the likelihood among block constant matrices with entries in $[\gamma_n, \rho_n]$:

$$\begin{aligned} &\widehat{\mathbf{\Theta}}_{i<j} = \widehat{\boldsymbol{Q}}_{\widehat{z}(i)\widehat{z}(j)},\ \widehat{\mathbf{\Theta}}_{ii} = 0 \\ &(\widehat{\boldsymbol{Q}}, \widehat{z}) \in \underset{\boldsymbol{Q}\in[\gamma_n,\rho_n]^{k\times k}_{\text{sym}}, z\in\mathcal{Z}_{n,k}}{\arg\max} \mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A}; z, \boldsymbol{Q}). \end{aligned} \tag{3}$$

In (3), γ_n and ρ_n are assumed to be known (see [17] 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 [30] or a likelihood-based model selection method [46]. The following result provides the upper bound on the estimation risk of the maximum likelihood estimator:

Theorem 1 (Corollary 2 in [17])**.** *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 $1 - 9\exp\left(-C\rho_n\left(k^2 + n\log(k)\right)\right)$,*

$$\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_2^2 \leq C'\left(\frac{\rho_n^2}{((1-\rho_n)^2 \wedge \gamma_n^2)}\right)\frac{\rho_n\left(k^2 + n\log(k)\right)}{p}. \tag{4}$$

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, [26, 14]).

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 (3) 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 [44] used variational methods to approximate the maximum likelihood estimator in networks with missing observations. The main idea of this method is to attach a distribution with free parameters to the latent variables. These parameters are fitted to get a distribution that is close to the true posterior. By choosing a variational distribution simpler than the true posterior, the optimization problem can be solved in an approximate way.

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

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

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 $\mathbb{P}\left(\cdot|\boldsymbol{X}\odot\boldsymbol{A}, \alpha, \boldsymbol{Q}\right)$ 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 $\mathbb{P}\left(\cdot|\boldsymbol{X}\odot\boldsymbol{A}, \alpha, \boldsymbol{Q}\right)$ 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 $\tau^i = \left(\tau_1^i, ..., \tau_k^i\right)$, and $\tau = \left(\tau^1, ..., \tau^n\right)$. Then, the variational estimator is defined as

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

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 $(\alpha, \boldsymbol{Q})$, $KL\left(\mathbb{P}_\tau(\cdot)||\mathbb{P}\left(\cdot|\boldsymbol{X}\odot\boldsymbol{A},\alpha,\boldsymbol{Q}\right)\right)\geq 0$, 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 [44] can be used to iteratively compute 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 :

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

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

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

The EM algorithm is not guaranteed to converge to a global maximum. To circumvent this problem, the algorithm should be initialized with care, by using, for example, a clustering step. This solution is implemented in the package `missSBM`.

Statistical guarantees for the variational estimator obtained in [12, 6, 34] 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 $\widehat{z}^{VAR}$:

$$\forall\, i\leq n,\ \widehat{z}^{VAR}(i) \triangleq \underset{a\leq k}{\arg\max}\left(\widehat{\tau}^{VAR}\right)_a^i. \tag{5}$$

Once we have estimated the community labels using (5), we replace the estimator $\widehat{\boldsymbol{Q}}^{VAR}$ of the matrix of connection probabilities by the empirical mean estimator:

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

and define $\widehat{\boldsymbol{\Theta}}^{VAR}$ as

$$\widehat{\boldsymbol{\Theta}}_{i\neq j}^{VAR} = \widehat{\boldsymbol{Q}}_{\widehat{z}^{VAR}(i),\widehat{z}^{VAR}(j)}^{ML-VAR},\ \widehat{\boldsymbol{\Theta}}_{ii}^{VAR} = 0. \tag{6}$$

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

$$(\widehat{\boldsymbol{Q}},\widehat{z}) \in \underset{\boldsymbol{Q}\in\mathcal{Q},z\in\mathcal{Z}_{n,k}}{\arg\max}\ \mathcal{L}_{\boldsymbol{X}}(\boldsymbol{A};z,\boldsymbol{Q}) \tag{7}$$

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 $\left(\widehat{\boldsymbol{Q}}_{\sigma(a),\sigma(b)}^{ML-VAR}\right)_{a,b\leq k} = \left(\widehat{\boldsymbol{Q}}_{a,b}\right)_{a,b\leq k}$. 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 a positive constant c 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. Assumption A.3 ensures the identifiability of stochastic block model parameters. 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. Note that under full observation of the network, the estimator of the matrix of connection probabilities $\widehat{\boldsymbol{Q}}^{ML-VAR}$ simply becomes :

$$\widehat{\boldsymbol{Q}}^{ML-VAR}_{ab} \triangleq \frac{\sum\limits_{i\in(\widehat{z}^{VAR})^{-1}(a), j\in(\widehat{z}^{VAR})^{-1}(b)} \boldsymbol{A}_{ij}}{n_{ab}(\widehat{z}^{VAR})},$$

where for any $(a,b) \in \{1,..,k\}$, we define

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

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 $\sum\limits_{a,b=1}^{k} \alpha^0_a \alpha^0_b \boldsymbol{Q}^0_{ab} = 1$

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 [17]:

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*

$$\mathbb{P}\left(\left\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}^{VAR}\right\|_2^2 \leq C_{\boldsymbol{Q}^0}\rho_n\left(k^2 + n\log(k)\right)\right) \underset{n\to\infty}{\to} 1. \tag{8}$$

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.

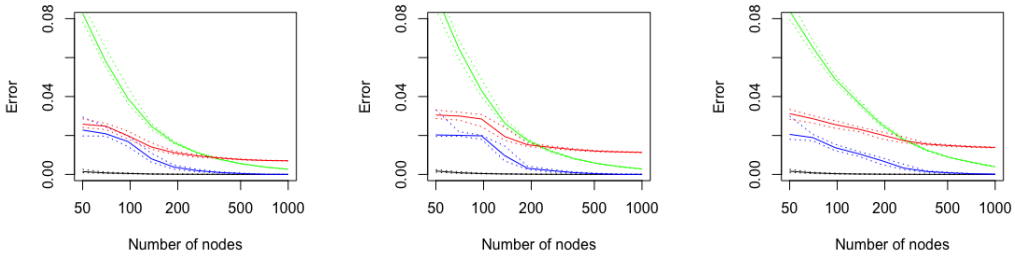
4 Numerical Results

4.1 Synthetic data

In this section we provide a simulation study of the performances of the maximum likelihood estimator defined in (6), and compare it to the variational estimator defined in [44] and implemented in the package `missSBM`, as well as to the Universal Singular Value Thresholding estimator introduced in [22] and implemented in the package `softImpute`. 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 (6) 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 experiments are conducted for a number of nodes n increasing from 50 to 1000. The simulation is repeated 100 times for each configuration $(\alpha^*, \boldsymbol{Q}^*, n)$. The results are reported in Figure 1.



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

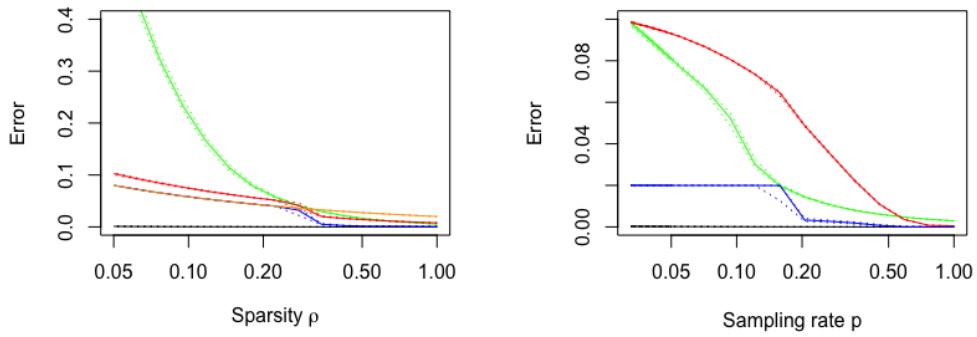
Figure 1: **Dense stochastic block model**: Error of connection probabilities estimation as a function of the number of nodes (left : assortative SBM with balanced communities; middle : disassortative SBM with balanced communities; right : mixed SBM with unbalanced communities). For each model, we compare the variational approximation to the maximum likelihood estimator (in blue) to that of `missSBM` (in red), that of `softImpute` (in green), and that of the oracle estimator with knowledge of the label z^* (in black). The full lines indicate the median of the mean squared error of the estimators over 100 repetitions, while the dashed lines indicate its 25% and 75% quantiles.

The variational approximation to the maximum likelihood estimator defined in (6) 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 `missSBM`. We underline however that the primary focus of the missSBM method is to infer the parameters $(\alpha^*, \boldsymbol{Q}^*)$.

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 of connection probabilities $\boldsymbol{\Theta}^*$ divided by ρ^2. The results are reported in Figure 2a.

As the network sparsity increases, the clustering of the nodes becomes more difficult. The normalized error of the estimator $\widehat{\boldsymbol{\Theta}}^{VAR}$ 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, [26])). 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



(a) Robustness against sparsity.

(b) Robustness against missing observations.

Figure 2: **Robustness against sparsity and missing observations**: Error of connection probabilities estimation as a function of the sparsity parameter ρ (left) and of the sampling rate p (right). For each model, 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). The full lines indicate the median respectively of the mean squared error divided by the sparsity parameter ρ (left) and of the mean squared error (right) of the estimators over 100 repetitions, while the dashed lines indicate its 25% and 75% quantiles.

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 of connection probabilities $\mathbf{\Theta}^*$. The results are reported in Figure 2b.

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 [43]. 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 (6) 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 $\widehat{\boldsymbol{\Theta}}_{ij}^{naive} = d/n$ 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).

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: Normalized squared distance between the observed adjacency matrix for the network on interaction on Day 2, and its predicted value.

The variational approximation to the maximum likelihood estimator predicts most accurately the interactions on Day 2. It is closely followed by the variational 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 $\widehat{\boldsymbol{\Theta}}^{VAR}$ is 1.4% smaller than that of $\widehat{\boldsymbol{\Theta}}^{missSBM}$, and 12.4% smaller than that of $\widehat{\boldsymbol{\Theta}}^{softImpute}$. In addition, the precision-recall curve presented in the Appendix indicates that no estimator is better across all sensitivity levels. Interestingly, the naive persistent estimator obtains a high error, which suggests a certain versatility in the children's social behaviour.

4.2.2 Network of co-authorship

Finally, we use variational approximation to the maximum likelihood estimator to predict unobserved links in a network of co-authorship between scientists working on network analysis, first analysed in [40]. 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 rather 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 likelihood estimator

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.

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.

Broader Impact

This paper proposes a new method for network analysis which is theoretically optimal, and has empirical advantages compared to current algorithms. It can be useful to researchers and practitioners with applications, e.g., in biology [37], ethnology [44] and ecology [4]. We acknowledge that improvements in network analysis and link prediction can have negative downstream societal impacts such as contributing to community-based polarization of opinion, or improving online surveillance.

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] 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.

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

[4] Avner Bar-Hen, Pierre Barbillon, and Sophie Donnet. Block models for generalized multipartite networks: Applications in ecology and ethnobiology. Statistical Modelling, 0(0):1471082X20963254, 0.

[5] O. Benyahia, C. Largeron, and B. Jeudy. Community detection in dynamic graphs with missing edges. In 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, 07 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, 06 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, Dec 2011.

[14] 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, January 2016.

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

[16] 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.

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

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

[19] 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.

[20] 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.

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

[22] 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.

[23] Jake M. Hofman and Chris H. Wiggins. Bayesian approach to network modularity. Phys. Rev. Lett., 100:258701, Jun 2008.

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

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

[26] 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.

[27] Olga Klopp and Nicolas Verzelen. Optimal graphon estimation in cut distance. Working Papers 2017-42, Center for Research in Economics and Statistics, December 2017.

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

[29] 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, September 2012.

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

[31] 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.

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

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

[34] 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.

[35] 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.

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

[37] Leonardo Morelli, Valentina Giansanti, and Davide Cittaro. Nested stochastic block models applied to the analysis of single cell data. bioRxiv, 2021.

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

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

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

[41] 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.

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

[43] 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, 08 2011.

[44] 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.

[45] 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.

[46] 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.

[47] 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, 10–15 Jul 2018.

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

[49] 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.

[50] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighbourhood smoothing. Biometrika, 104(4):771–783, 09 2017.

[51] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See, e.g., discussions on the experimental results in Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the Broader Impact Section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results?[Yes]
   (b) Did you include complete proofs of all theoretical results?[Yes] Proofs are provided in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Codes and datasets for reproducting the numerical experiments are provided in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Details of the experiments are provided in the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We include the 25% and 75% quantiles of the error in our graphs.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The running time of the experiments remains small (a few seconds are required to compute the estimators for a graph with a thousand node on a laptop with a 2.3 GHz Intel Core i5 dual core processor).
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] Those questions have already been discussed in the original papers presenting the datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Those questions have already been discussed in the original papers presenting the datasets.
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]