Inférence Variationnelle du Modèle à Blocs Stochastiques (SBM) avec covariables en présence de données manquantes

Timothée Tabouy [1], Pierre Barbillon [1], Julien Chiquet [1]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*
prenom.nom@agroparistech.fr

Résumé. Le modèle à blocs stochastiques ou *Stochastic Block Model* (SBM) (Nowicki and Snijders, 2001) est un modèle de graphe aléatoire généralisant le modèle d'Erdős-Reyni (Erdős and Renyi, 1959) à l'aide d'une structure latente sur les nœuds. L'utilisation de variables latentes dans le SBM permet de modéliser un large spectre de topologies de réseau, en particulier les graphes en affiliation, en étoile ou bipartite. L'inférence de ces modèles repose sur des modifications de l'algorithme EM (Expectation Maximization), comme par exemple l'approche EM variationnelle (Daudin et al., 2008) ou Bayésienne variationnelle (Latouche et al., 2012). Dans ces approches, le réseau est toujours considéré comme parfaitement observé, alors que de nombreux cas d'application (en particulier en sociologie) suggèrent que son observation est partielle et guidée par une stratégie d'échantillonnage dépendant du réseau lui-même, par exemple centrée sur les nœuds.

Dans un prédécent travail (Tabouy et al., 2019) nous avions constaté qu'un échantillonnage partiel du réseau peut induire un biais d'estimation dans le modèle SBM. Notre objectif était alors la modélisation de la stratégie d'échantillonnage utilisée et son intégration dans les procédure d'inférence. S'appuyant sur la théorie des données manquantes développée par Rubin (1976), nous avons adapté les définitions de Missing At Random (MAR) et Not MAR (NMAR) aux cas de modèles à variables latentes.

Nous proposons dans cette présentation de montrer que la prise en compte de covariables dans la modélisation peut induire un changement de nature (MAR ou NMAR) des données manquantes. Ce constat a pour but de faire le lien entre les différentes natures des données manquantes suivant la modélisation et de la prendre en compte dans l'inférence. En effet nous montrerons que différents modèles sans et avec covariables, couplés avec une loi d'échantillonnage donnant lieu à des donnés manquantes de natures différentes aboutissent au même modèle.

Mots-clés. Modèle à blocs stochastiques · données manquantes · covariables

Abstract. The stochastic block model (SBM) (Nowicki and Snijders, 2001) is a random graph model generalizing the Erdős-Reyni model (Erdős and Renyi, 1959) by means of a latent structure on the nodes. The use of latent variables in the SBM allows us to model a broad variety of network topologies, in particular affiliation networks, star networks or bipartite networks. The inference of such models is based on modifications of the EM (Expectation Maximization) algorithm, such as the variational EM (Daudin et al., 2008) or the variational Bayes algorithm (Latouche et al., 2012). In these approaches, the network is always considered to be perfectly observed, whereas many cases of application (particularly in sociology) suggest that its observation is partial and guided by a sampling strategy depending on the network itself.

In a previous work (Tabouy et al., 2019) we found that partial sampling of the network can induce estimation bias in the SBM model. Our objective was then to model the sampling strategy used and integrate it into the inference procedures. Based on the theory of missing data developed by Rubin (1976), we have adapted the Missing At Random (MAR) and Not MAR (NMAR) definitions to cases of latent variable models.

In this presentation, we propose to show that taking into account covariates in the modelling can induce a change in the nature (MAR or NMAR) of the missing data. The purpose of this observation is to make the link between the different types of missing data according to the model and to take it into account in the inference. Indeed, we will show that different models without and with covariates, coupled with a sampling law giving rise to missing data of different natures, result in the same model.

Keywords. Stochastic block Model · missing data · covariates

Notations

- $Y \in \mathbb{M}_n(\mathbb{R})$ the adjacency matrix,
- Y^{o} the observed part of the adjacency matrix,

- $X_i \in \mathbb{R}^N$ some covariates of node i,
- $X = [X_1, .., X_n] \in \mathbb{M}_{N\times n}(\mathbb{R})$ the matrix of covariates,
- $R \in \mathbb{M}_n(\mathbb{R})$ the sampling matrix : $R_{ij} = 1$ si Y_{ij} sampled, 0 otherwise,
- $V \in \{0,1\}^n$ the sampling vector indicating which node is sampled ($V_i = 1$) or not ($V_i = 0$),
- $Q \in \mathbb{R}$ the number of blocks,
- $z \in [\![1, Q]\!]^n$ the vector of block memberships,
- $Z_i \in \{0,1\}^Q$ the vector such that $(Z_i)_q = Z_{iq} = \mathbb{1}_{z_i=q}$.

1 SBM with covariates

We will consider two different structures in which covariates impact the SBM. In the former model, covariates influence directly latent variables and the sampling. In the latter model covariates influence also the sampling and the distribution of edges. Conditionnal dependencies of these models are represented with a DAG in Figure 1.

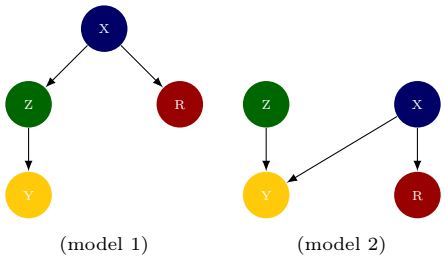


Figure 1: DAGs of relationships between Y, Z, R and X considered in the framework of missing data for SBM with covariates. The systematic edge between Z and Y is part of the SBM.

Model 1 Writting $\alpha_\cdot = (\alpha_{\cdot 1}, ..., \alpha_{\cdot Q}) \in [0,1]^Q$ we define

$$\alpha_{iq} = \frac{e^{\beta_q^t X_i \mathbb{1}_{q\neq Q}}}{1 + \sum_{k=1}^{Q-1} \beta_k^t X_i}, \quad \forall (i,q) \in [\![1,n]\!] \times [\![1,Q]\!],$$

with $\beta_q \in \mathbb{R}^N$ for all $q \in [\![1, Q-1]\!]$ and $\beta_Q = 0$. Then we have

$$\begin{aligned} Z_i \mid X_i &\sim^{\text{iid}} \mathcal{M}(1, \alpha_i), \quad \forall i \in [\![1,n]\!], \\ Y_{ij} \mid \{Z_i, Z_j\} &\sim^{\text{ind}} \mathcal{B}(\pi_{z_i z_j}), \quad \forall (i,j) \in [\![1,n]\!]^2, \end{aligned}$$

with $\pi_{q\ell} \in [0,1]$.

Model 2

$$\begin{aligned} Z_i &\sim^{\text{iid}} \mathcal{M}(1, \alpha), \quad \forall i \in [\![1,n]\!], \\ Y_{ij} \mid \{Z_i, Z_j, X_i, X_j\} &\sim^{\text{ind}} \mathcal{B}(g(\gamma_{z_i z_j} + \beta^t \phi(X_i, X_j))), \quad \forall (i,j) \in [\![1,n]\!]^2, \end{aligned}$$

where $\gamma_{q\ell} \in \mathbb{R}$, $\beta \in \mathbb{R}^m$, $\alpha \in [0,1]^Q$, $g(x) = (1+e^{-x})^{-1}$ and $\phi(\cdot,\cdot) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^m$ an arbitrary "similarity" measure vector such that $\phi(X_i, X_j) = \phi(X_j, X_i)$.

2 Sampling according to the covariates

In this section we define two sampling designs, one dyad-centered and the other node-centered. In both sampling, the probability to observe a dyad (resp. node) depends on the value of the covariate. If the covariates have no effect, then the probability to observe a dyad (resp. node) is independant of the dyad (resp. node).

Definition 1 (Dyad-centered sampling)**.** *Let $\alpha \in \mathbb{R}$, $\kappa \in \mathbb{R}^p$ and $\psi(\cdot,\cdot) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^p$. The probability to observe a dyad is*

$$\mathbb{P}(R_{ij} = 1|X) = g(\alpha + \kappa^t \psi(X_i, X_j)).$$

Definition 2 (Node-centered sampling)**.** *Let $\nu \in \mathbb{R}$ and $\eta \in \mathbb{R}^N$. The probability to observe all dyads corresponding to a node is*

$$\mathbb{P}(V_i = 1|X) = g(\nu + \eta^t X_i).$$

3 Statistical inference

On the basis of Figure 1, the type of missingness for SBM is defined as follows:

$$\text{Sampling design for SBM is } \begin{cases} \text{MCAR} & \text{if } R \perp\!\!\!\perp (Y^{\text{m}}, Z, Y^{\text{o}}) \mid X, \\ \text{MAR} & \text{if } R \perp\!\!\!\perp (Y^{\text{m}}, Z) \mid (Y^{\text{o}}, X), \\ \text{NMAR} & \text{otherwise.} \end{cases} \tag{1}$$

Proposition 1. *From* (1)*, if the sampling is MAR or MCAR then maximizing $p_{\theta,\psi}(Y^{\text{o}}, R)$ or $p_\theta(Y^{\text{o}})$ in θ is equivalent.*

In the light of (1), sampling designs defined in definition 1 and 2 are missing completely at random (MCAR) conditionnaly to covariates.

3.1 Background on variational approximation in SBM

In the presence of latent variables, the EM algorithm (Dempster et al., 1977) is the natural choice. However, it requires the evaluation of the conditional mean $\mathbb{E}_{Z|Y^{\text{o}},X}[\log p_\theta(Y^{\text{o}}, Z|X)]$ which is intractable for SBM since all latent variables Z_i depend conditionally on Y^{o}. The variational approach circumvents this limitation by maximizing a lower bound of the log-likelihood based on an approximation $\tilde{p}_\tau$ of the true conditional distribution $p_\theta(Z|Y^{\text{o}}, X)$,

$$\begin{aligned} \log p_\theta(Y^{\text{o}}|X) \geq J_{\tau,\theta}(Y^{\text{o}}|X) &\triangleq \log(p_\theta(Y^{\text{o}}|X)) - \text{KL}[\tilde{p}_\tau(Z)||p_\theta(Z|Y^{\text{o}}, X)], \\ &= \mathbb{E}_{\tilde{p}_\tau}[\log(p_\theta(Y^{\text{o}}, Z|X))] - \mathbb{E}_{\tilde{p}_\tau}[\log \tilde{p}_\tau(Z)], \end{aligned}$$

where τ are some variational parameters and KL the Kullback-Leibler divergence. The approximated distribution is chosen so that the integration over the latent variables simplifies by factorization. In SBM, the prior distribution of $Z_i = (Z_{i1}, \ldots, Z_{iQ})$ is the multinomial distribution, thus the natural variational counterpart to $p_\theta(Z|Y^{\text{o}}, X)$ is

$$\tilde{p}_\tau(Z) = \prod_{i \in \mathcal{N}} m(Z_i; \tau_i), \tag{2}$$

where $\tau_i = (\tau_{i1}, \ldots, \tau_{iQ})$, and $m(\cdot; \tau_i)$ is the multinomial probability density function with parameters τ_i. The full set of variational parameters is denoted by $\tau = \{\tau_1, \ldots, \tau_n\}$.

3.2 Inference of Model 1

In the MAR case, inference is conducted on the observed part of the adjacency matrix. We start by recalling the complete likelihood $p_\theta(Y^{\text{o}}, Z|X)$ which has an explicit form contrary to the likelihood of the observed data $p_\theta(Y^{\text{o}}|X)$.

The complete log-likelihood restricted to the observed variables is

$$\log p_\theta(Y^{\mathrm{o}}, Z|X) = \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \sum_{q,\ell} Z_{iq} Z_{j\ell} \log b(Y_{ij}, \pi_{q\ell}) + \sum_{i\in\mathcal{N}^{\mathrm{o}}} \sum_{q} Z_{iq} \log(\alpha_{iq}), \tag{3}$$

with $b(x,\pi) = \pi^x(1-\pi)^{1-x}$ the Bernoulli probability density function.

By Equation (3) and approximation (2), the close form of the lower bound is

$$J_{\tau,\theta}(Y^{\mathrm{o}}|X) = \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \sum_{q,\ell} \tau_{iq}\tau_{j\ell} \log b(Y_{ij}, \pi_{q\ell}) + \sum_{i\in\mathcal{N}^{\mathrm{o}}} \sum_{q} \tau_{iq} \log(\alpha_{iq}/\tau_{iq}). \tag{4}$$

The two maximization problems are solved as stated in the following proposition, straightforwardly derived from Daudin et al. (2008).

Proposition 2. *Consider the lower bound $J_{\tau,\theta}(Y^{\mathrm{o}}|X)$ given by (4).*

1. *The parameters $\theta = (\alpha, \pi)$ maximizing $J_\theta(Y^{\mathrm{o}}|X)$ when τ is held fixed are*

$$\hat{\alpha}_{iq} = \arg\max_{\alpha} J_{\tau,\theta}(Y^{\mathrm{o}}|X), \qquad \hat{\pi}_{q\ell} = \frac{\sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \hat{\tau}_{iq}\hat{\tau}_{j\ell} Y_{ij}}{\sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \hat{\tau}_{iq}\hat{\tau}_{j\ell}}.$$

2. *The variational parameters τ maximizing $J_\tau(Y^{\mathrm{o}}|X)$ when θ is held fixed are obtained thanks to the following fixed point relation:*

$$\hat{\tau}_{iq} \propto \alpha_{iq} \left(\prod_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \prod_{\ell\in\mathcal{Q}} b(Y_{ij}; \pi_{q\ell})^{\hat{\tau}_{j\ell}} \right).$$

3.3 Inference of Model 2

The complete log-likelihood for Model 2 is

$$\begin{aligned} \log p_\theta(Y^{\mathrm{o}}, Z|X) &= \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} Y_{ij}\, g\left({Z_i}^t \gamma Z_j + \beta^t \phi(X_i, X_j)\right) + (1 - Y_{ij}) \log\left(1 - \mathrm{logistic}\left({Z_i}^t \gamma Z_j + \beta^t \phi(X_i, X_j)\right)\right) \\ &\quad + \sum_{i\in\mathcal{N}^{\mathrm{o}}} Z_i^t \log(\alpha), \end{aligned} \tag{5}$$

with $g(x) = \log(\mathrm{logistic}(x)) = -\log(1 + e^{-x})$. As a consequences the variational lower bound is given by

$$J_{\tau,\theta}(Y^{\mathrm{o}}|X) = \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \sum_{q,\ell} \tau_{iq}\tau_{j\ell} \Big\{ (Y_{ij} - 1)(\gamma_{q\ell} + \beta^t \phi(X_i, X_j)) + g(\gamma_{q\ell} + \beta^t \phi(X_i, X_j)) \Big\} \tag{6}$$

$$+ \sum_{i,q} \tau_{iq} \log\left(\frac{\alpha_q}{\tau_{iq}}\right). \tag{7}$$

Proposition 3. *Consider the maximization of the lower bound (7).*

1. *The parameters γ and β maximizing $J_{\tau,\theta}(Y^{\mathrm{o}}|X)$ when all other parameters are held fixed are*

$$(\hat{\gamma}, \hat{\beta}) = \arg\max_{(\gamma,\beta)} J_{\tau,\theta}(Y^{\mathrm{o}}|X).$$

with

$$\frac{\partial J_{\tau,\theta}(Y^{\mathrm{o}}|X)}{\partial \gamma_{q\ell}} = \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \tau_{iq}\tau_{j\ell} \left\{ Y_{ij} - 1 + \frac{e^{-(\gamma_{q\ell} + \beta^t \phi(X_i, X_j))}}{1 + e^{-(\gamma_{q\ell} + \beta^t \phi(X_i, X_j))}} \right\}, \tag{8}$$

$$\frac{\partial J_{\tau,\theta}(Y^{\mathrm{o}}|X)}{\partial \beta_k} = \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \sum_{q,\ell} \tau_{iq}\tau_{j\ell} \left(\phi(X_i, X_j)\right)_k \left\{ Y_{ij} - 1 + \frac{e^{-(\gamma_{q\ell} + \beta^t \phi(X_i, X_j))}}{1 + e^{-(\gamma_{q\ell} + \beta^t \phi(X_i, X_j))}} \right\}. \tag{9}$$

2. *The optimal τ in $J_{\tau,\theta}(Y^{\mathrm{o}}|X)$ when all other parameters are held fixed verify*

$$\hat{\tau}_{iq} \propto \alpha_q \prod_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \prod_{\ell} \exp\Big\{ (Y_{ij} - 1)(\gamma_{q\ell} + \beta^t \phi(X_i, X_j)) + g(\gamma_{q\ell} + \beta^t \phi(X_i, X_j)) \Big\}^{\hat{\tau}_{j\ell}}. \tag{10}$$

3.4 Model selection

The Integrated Classification Likelihood criterion was introduced by Biernacki et al. (2000) for mixture models, where the likelihood – and thus BIC – is usually intractable. Daudin et al. (2008) adapted a variational ICL in the context of SBM. Under MAR condition, this criterion requires a slight adaptation stated in the following proposition.

Proposition 4 (Model selection)**.** *For an SBM with Q blocks and for $\hat{\theta} = \arg\max \log p_\theta(Y^o, Z)$, the ICL criterion is given by*

$$\mathrm{ICL}(Q) = -2\mathbb{E}_{\tilde{p}_\tau}\left[\log p_{\hat{\theta}}(Y^o, Z; Q, X)\right] + pen_Q,$$

and

$$pen_Q = \begin{cases} \left(\frac{Q(Q+1)}{2}\right)\log \mathrm{card}\,(\mathcal{D}^o) + N(Q-1)\log \mathrm{card}\,(\mathcal{N}^o) & \textit{(in model 1)}, \\ \left(\frac{Q(Q+1)}{2} + N\right)\log \mathrm{card}\,(\mathcal{D}^o) + (Q-1)\log \mathrm{card}\,(\mathcal{N}^o) & \textit{(in model 2)}. \end{cases} \tag{11}$$

Note that a dyad is only counted once since we work with symmetric networks. The number of blocks chosen is the one associated to the lowest ICL.

4 Duality: with or without covariates?

We have seen that missing data can be generated by a sampling strategy satisfying MAR condition conditionnaly to observation of covariates. However, if the covariates are unknown and if the data are generated by a SBM defined as in section 1, then the sampling strategy that generated the observed data is not MAR anymore but NMAR. In this section we will explain this phenomenon in mathematical terms through an example.

Definition 3 (Block-dyad sampling)**.** *Let $\rho = (\rho_{q\ell})_{1\leq q,\ell\leq Q} \in \mathcal{M}_Q([0,1])$. Then, in Block-on-dyad sampling the conditionnal distribution of $R|Z$ is given by*

$$R_{ij}|Z_i, Z_j \sim^{ind} \mathcal{B}(\rho_{Z_i Z_j}). \tag{12}$$

Proposition 5. *Let's define the category-specific covariates matrix $X \in \mathcal{M}_{1\times n}(\{1, ...Q\})$ for $Q > 1$. The model is generated as follow*

$$\begin{aligned} X_i \quad &\sim^{iid} \quad \mathcal{M}(1,\nu), \quad \forall i \in [\![1,n]\!], \\ \mathbb{P}(Z_i = a|X_i = a) = \delta \quad &\textit{and} \quad \mathbb{P}(Z_i \neq a|X_i = a) = \frac{1-\delta}{Q-1}, \quad \forall i \in [\![1,n]\!]. \end{aligned}$$

The probabilities to observe a dyad are given by

$$p_{q\ell} = \mathbb{P}(R_{ij} = 1|X_i = q, X_j = \ell).$$

Then, this modelisation corresponds to a classical SBM with sampling design the Block-on-dyad sampling, which is NMAR, with parameters

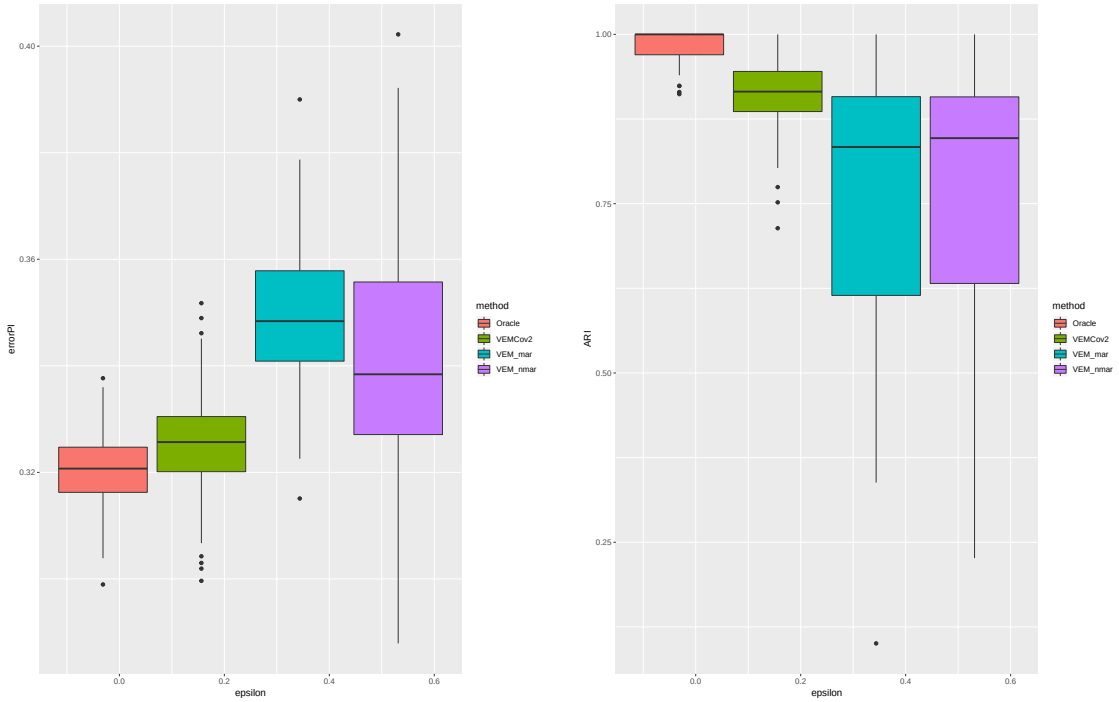
$$\begin{aligned} \rho_{q\ell} &= \mathbb{P}(R_{ij} = 1|Z_i = q, Z_j = \ell), \\ &= \frac{\delta^2 p_{q\ell}\nu_q\nu_\ell + \left(\frac{1-\delta}{Q-1}\right)^2 \sum_{\substack{a\neq q \\ b\neq \ell}} p_{ab}\nu_a\nu_b + 2\delta\left(\frac{1-\delta}{Q-1}\right)\sum_{\substack{a\neq q \\ b=\ell}} p_{ab}\nu_a\nu_b}{(\delta\nu_q + \frac{1-\delta}{Q-1}\sum_{c\neq q}\nu_c)(\delta\nu_\ell + \frac{1-\delta}{Q-1}\sum_{c\neq \ell}\nu_c)}. \end{aligned}$$

Remark 1. $\delta = 1 \Rightarrow \rho_{q\ell} = p_{q\ell}$.

Figure 2 illustrate Proposition 5 with ...

References

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(7):719–725, July 2000. ISSN 0162-8828. doi: 10.1109/34.865189. URL `http://dx.doi.org/10.1109/34.865189`.



Methods

Figure 2

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Stat. comp.*, 18(2):173–183, 2008.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B-Met.*, 39(1):1–38, 1977.

P. Erdős and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

P. Latouche, É. Birmelé, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Stat. Modelling*, 12(1):93–115, 2012.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Soc.*, 96(455):1077–1087, September 2001.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

T. Tabouy, P. Barbillon, and J. Chiquet. Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, 0(ja):1–20, 2019. doi: 10.1080/01621459.2018.1562934. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.2018.1562934`.