Block models for multipartite networks Applications in ecology and ethnobiology

Avner Bar-Hen *[1], P. Barbillon †[2], and S. Donnet ‡[3]

[1] *CNAM, 75003, Paris, France*
[2] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*
[3] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France*

July 27, 2018

Abstract

Modeling relations between individuals is a classical question in social sciences, ecology, etc. In order to uncover a latent structure in the data, a popular approach consists in clustering individuals according to the observed patterns of interactions. To do so, Stochastic block models (SBM) and Latent Block models (LBM) are standard tools for clustering the individuals with respect to their comportment in a unique network. However, when adopting an integrative point of view, individuals are not involved in a unique network but are part of several networks, resulting into a potentially complex multipartite network. In this contribution, we propose a stochastic block model able to handle multipartite networks, thus supplying a clustering of the individuals based on their connection behavior in more than one network. Our model is an extension of the latent block models (LBM) and stochastic block model (SBM). The parameters –such as the marginal probabilities of assignment to blocks and the matrix of probabilities of connections between blocks– are estimated through a variational Expectation-Maximization procedure. The numbers of blocks are chosen with the Integrated Completed Likelihood criterion, a penalized likelihood criterion. The pertinence of our methodology is illustrated on two datasets issued from ecology and ethnobiology.

Keywords Networks, Latent Block Models, Stochastic Block Models, Variational EM, Model Selection, Ecology, Ethnobiology

*avner@cnam.fr
†pierre.barbillon@agroparistech.fr
‡sophie.donnet@agroparistech.fr

1 Introduction

Networks have become fundamental tools in various fields, such as ecological theory or sociology to name but a few. Statistical analysis aims at studying the structure of these networks and allows the discovery as well as the representation of clusters or communities [see Matias and Robin, 2014, for a review]. The recent years have witnessed a growing interest for complex networks such as multiplex networks –when several types of relations are simultaneously studied on a common set of individuals– [Kéfi et al., 2016, Barbillon et al., 2016] or time evolving networks [Matias and Miele, 2017]. In this paper, we are interested in the modeling and inference of multipartite networks.

Multipartite networks arise when the individuals (nodes) at stake can be in advance partitioned into groups defined by their nature. In what follows, these groups will be referred to as *functional groups*. As a first example, let us consider interactions in ecology where individuals are living species. Functional groups may be species of plants, pollinators, herbivores, ants, seed dispersal birds, etc. A number of studies in ecology have suggested that analyzing the interactions between pairs of functional groups (e.g. plants/pollinators or plants/ants...) separately does not reveal the potential of this information, and considering simultaneously the various interaction networks could contribute to a better understanding of the processes at stake [see Pocock et al., 2012, Kéfi et al., 2016, Dáttilo et al., 2016, for instance]. In an ecosystemic approach, the network resulting from the observation of the interactions between plants on the one hand and pollinators, ants or seed dispersal birds (for instance) on the other hand, is typically a multipartite network.

Ethnobiology is the scientific study of the relations between environment and people. One of the problematic at stake for instance in Thomas and Caillon [2016] is to understand how social relations between individuals (here seed circulation) may structure and guaranty biodiversity in the cultivated crop species. In this case, two functional groups are involved, namely farmers and crop species. The interactions between farmers are seed circulations represented by a simple network. The inventories of the crop species grown by each farmers are also available, inducing a bipartite network between the same farmers and the crop species. When considered simultaneously, the two networks then constitute a multipartite network.

These two datasets issued respectively from ecology and ethnobiology will be treated in this paper. However, our work can be applied to many other application fields. As an example, marketing can also supply multipartite networks data when individuals are connected through social networks but can also be described by their on-line purchases. An application in pharmaco-sciences was also considered by Robert [2017].

Multipartite networks require the development of specific statistical tools. Some extensions of standard descriptive tools –such as community detection– have been proposed in the literature (see for instance Yang and Leskovec [2012]

or Gaskó et al. [2017]). When aiming at clustering similar unities based on their connectivity patterns without any a priori hypothesis about the patterns to be found (e.g. modularity, centrality, hierarchy), probabilistic mixture models have proved their efficiency. In particular, when a unique network is at stake, Stochastic block models (SBM) [Snijders and Nowicki, 1997] and Latent Block models (LBM) [Govaert and Nadif, 2003] are standard tools for clustering the individuals according to the observed patterns of interactions.

In this paper, we extend block models to multipartite networks. Our generalization encompasses SBM and LBM, thus handling simultaneously interactions between individuals belonging to the same functional group as well as interactions between individuals from different functional groups. This joint modeling of several networks allows to study structured interactions among individuals of a network as well as the impact of a part of the individuals on the other interactions. The joint modeling will provide clusterings of individuals in each functional group on the basis of all the interactions they are implied in.

Ability of proposing a joint model for various (simple or bipartite) networks is an important step for understanding the structure of communities within a complex (eco)system.

The paper is organized as follows. Section 2 is dedicated to the introduction of notations supplying a flexible tool to describe multipartite networks. We also give a quick description of our two datasets and illustrate the notations on these specific cases. The block model for multipartite networks is described in Section 3 while its maximum likelihood inference is presented in Section 4. The likelihood function is maximized through a variational version of the Expectation-Maximization algorithm and the model selection task is performed using an adapted penalized likelihood criterion. Finally, the statistical analyses of the two datasets with discussion are presented in Section 5. Perspectives are discussed in the last section.

2 Notations and data

We first introduce some notations and illustrate them on our two motivating examples.

A collection of networks Assume that Q functional groups of individuals are at stake; for any $q = 1 \dots Q$, let n_q be the number of individuals (or equivalently individuals, individuals or individuals) in the q-th functional group.

A multipartite network can be seen as a collection of networks: each network may be simple i.e. describing the relations inside a given functional groups or bipartite i.e. describing the relations between individuals belonging to two different functional groups. We index the collection of networks by pairs (q, q') (q and q' in $[\![1, Q]\!]$). If we deal with a simple network then $q = q'$. The set $\mathcal{E}$ denotes the list of pairs of functional groups for which we observe an interaction network.

For any $(q, q') \in \mathcal{E}$, the interaction network is encoded in a matrix $X^{qq'}$ such that $X_{ij}^{qq'} = 1$ if there is an edge from unit i of functional group q to unit i' of functional group q', 0 otherwise. If $q \neq q'$, $X^{qq'}$ is said to be an *incidence matrix*. X^{qq} is an *adjacency matrix*: it is symmetric if the relation inside the functional group q is undirected, non-symmetric otherwise. Then, the multipartite network we consider is denoted by $\boldsymbol{X} = \left\{ \left(X^{qq'} \right), (q, q') \in \mathcal{E} \right\}$.

Remark 1 *For the sake of simplicity, in this paper, we only present the model and the inference methodology for binary interactions (presence or absence encoded by* $0/1$*). However, as discussed in Section 6, our work can be extended to multipartite valued interactions or multipartite multiplex interaction networks.*

Illustration of the datasets • **Example 1** is in ecology and is issued from Dáttilo et al. [2016]. The authors are interested in studying jointly the mutualistic relations between plants and pollinators, plants and ants, and plants and frugivorous birds. It then results into 4 functional groups, namely plants ($q = 1$), pollinators ($q = 2$), ants ($q = 3$) and birds ($q = 4$). Using the notations previously introduced, we get $\mathcal{E} = \{(1, 2), (1, 3), (1, 4)\}$ and each interaction matrix $X^{qq'}$ is an incidence matrix corresponding to a bipartite network where $X_{ii'}^{1q'} = 1$ if the plant species i has been observed at least once in a mutualistic interaction with the animal species i' of functional group q' during the observation period, 0 otherwise.

• **Example 2** is issued from Thomas and Caillon [2016] and Thomas et al. [2015]. In this dataset, we observe on the one hand, seed circulation between farmers –resulting into a non-symmetric adjacency matrix – and on the other hand the crop species grown by the farmers, resulting into an incidence matrix. Adopting the notations previously introduced, we set $q = 1$ for the farmers, $q = 2$ for crop species, and $\mathcal{E} = \{(1, 1), (1, 2)\}$. $X_{ii'}^{11} = 1$ if farmer i gives seeds to farmer i' (oriented relation), 0 otherwise and $X_{ij}^{12} = 1$ if farmer i cultivates crop species j, 0 otherwise.

More details about these datasets are given in Section 5.

3 A block model for multipartite network

In order to account for heterogeneity among individuals, we propose a mixture model explicitly describing the way edges connect nodes in the various networks. As will be discussed hereafter, our model is an extension and a combination of the SBM and the LBM.

Assume that for each functional group q ($q \in [\![1, Q]\!]$), the n_q individuals are divided into K_q clusters (or equivalently blocks). $\forall q \in [\![1, Q]\!]$ and $\forall i \in [\![1, n_q]\!]$, let Z_i^q be the latent random variable such that $Z_i^q = k$ if individual i of functional group q belongs to cluster k. The random variables Z_i^q's are assumed to be independent and their distributions are such that: $\forall k \in [\![1, K_q]\!], \forall q \in [\![1, Q]\!], \forall i \in$

$[\![1, n_q]\!]$:

$$\mathbb{P}(Z_i^q = k) = \pi_k^q, \tag{1}$$

with $\sum_{k=1}^{K_q} \pi_k^q = 1$, $\forall q \in [\![1, Q]\!]$. In what follows, $\boldsymbol{Z}$ denotes the set of latent variables:

$$\boldsymbol{Z} = (Z_i^q)_{i \in [\![1,n_q]\!], q \in [\![1,Q]\!]}.$$

Now, we set a mixture model on the connections $\boldsymbol{X}$ in the following way. For any $(q, q') \in \mathcal{E}$, conditionally on the latent variables, the entries of the matrices $\left(X_{ii'}^{qq'}\right)$ are assumed to be independent and distributed as follows: $\forall (i, i') \in \mathcal{S}^{qq'}$,

$$X_{ii'}^{qq'} | \{Z_i^q = k, Z_{i'}^{q'} = k'\} \sim_{ind} \mathcal{B}ern(\alpha_{kk'}^{qq'}) \tag{2}$$

meaning that the probability of connection from i of functional group q to i' of functional group q' only depends on the clusters to which they belong. The sets $\mathcal{S}^{qq'}$ are additional notations which are necessary to handle particular cases of adjacency matrices (when $q = q'$). As stressed before, if $(q, q) \in \mathcal{E}$, the corresponding adjacency matrix X^{qq} may be symmetric or not (depending if the relation at stake is directed or not). Moreover, the diagonal can be assumed to be zero or not depending if loop interactions may occur or not. In order to be able to handle all these situations, for any $(q, q) \in \mathcal{E}$, we define $\mathcal{S}^{qq}$ as the set of possible edges:

- $\mathcal{S}^{qq} = [\![1, n_q]\!]^2$ if X^{qq} represents a directed network with possible self-loops.
- $\mathcal{S}^{qq} = \{(i, i') \in [\![1, n_q]\!]^2 | i \neq i'\}$ if X^{qq} represents a directed network without self-loop.
- $\mathcal{S}^{qq} = \{(i, i') \in [\![1, n_q]\!]^2 | i \leq i'\}$ if X^{qq} represents a non-directed network with self-loops.
- $\mathcal{S}^{qq} = \{(i, i') \in [\![1, n_q]\!]^2 | i < i'\}$ if X^{qq} represents a non-directed network without self-loop.

For any $(q, q') \in \mathcal{E}$ such that $q \neq q'$, we set $\mathcal{S}^{qq'} = [\![1, n_q]\!] \times [\![1, n_{q'}]\!]$. With this definition, we encode any interaction network by $X^{qq'} := (X_{i,i'}^{qq'})_{(i,i') \in \mathcal{S}^{qq'}}$.

In what follows, we will refer to this block model for multipartite networks as Multipartite Block Model (MBM).

Comments on the conditional dependencies

• As discussed previously, our model is a generalization of SBM and LBM. Indeed, Equations (1)-(2) exactly define the SBM if $\mathcal{E} = \{(1, 1)\}$ and the LBM if $\mathcal{E} = \{(1, 2)\}$. Our extension sets that the latent structures $\mathbf{Z}$ are shared among the matrices i.e. if a functional group q is at stake in several interaction

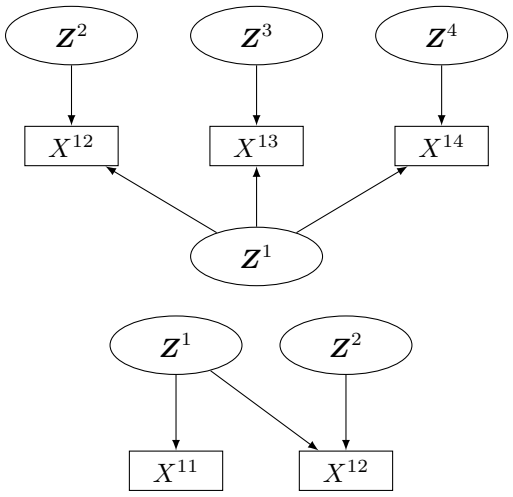


Figure 1: DAG for Example 1 (top) and Example 2 (bottom)

matrices the same latent random variables $\mathbf{Z}^q = (Z_i^q)_{i \in [\![1,n_q]\!]}$ impact the distributions of the corresponding interaction matrices. In other words, the clusters gather individuals sharing the same properties of connection in the collection of networks.

• In terms of probabilistic dependence, conditionally on the clustering variables $\boldsymbol{Z}$, the quantities $\left(X_{ii'}^{qq'}\right)$ are independent. However, $\boldsymbol{Z}$ being latent, their marginalization introduces dependence between the entries of each matrix $X^{qq'}$. Moreover, the marginalization with respect to the latent variables $\boldsymbol{Z}$ may introduce a probabilistic dependence between the observed matrices $\left(X^{qq'}\right)_{(q,q') \in \mathcal{E}}$. Indeed, all the matrices involving a given functional group q are then dependent when marginalizing. Obviously, if each functional group appears in only one element of $\mathcal{E}$, the MBM reduces to independent SBMs or LBMs.

• As a consequence of the previous comment, the clustering variables $(Z_i^q)_{q \in [\![1,Q]\!], i \in [\![1,n_q]\!]}$ are dependent once conditioned by the observations $\boldsymbol{X}$.

• We illustrate the various probabilistic dependences described above by giving the probabilistic directed acyclic graph (DAG) corresponding to Examples 1 and 2 in Figure 1.

In the following section, we develop an adapted version of the Variational Expectation Maximization (VEM) algorithm to maximize the likelihood function. The estimated clusters $\widehat{\boldsymbol{Z}} = \left(\widehat{Z}_i^q\right)_{q \in [\![1,Q]\!], i \in [\![1,n_q]\!]}$ will be a by-product of the inference method. We also propose a penalized likelihood criterion to select the numbers of clusters $\boldsymbol{K} = (K_1, \ldots, K_Q)$.

4 Parameter inference and model selection

Unknown parameters For a given vector $\boldsymbol{K} = (K_1, \ldots, K_Q)$, let $\theta_{\boldsymbol{K}} = (\boldsymbol{\alpha}, \boldsymbol{\pi})$ be the unknown parameters of the model defined by Equations (1) and (2) where

$$\boldsymbol{\pi} = (\pi_k^q)_{k \in [\![1, K_q]\!], q \in [\![1, Q]\!]} \in \mathcal{T}_{\boldsymbol{K}}$$

with $\mathcal{T}_{\boldsymbol{K}} \quad \left(= \otimes_{q=1}^{Q} \mathcal{T}_{K_q} \right)$ and $\forall K \in \mathbb{N}^*$,

$$\mathcal{T}_K = \left\{ (w_1, \ldots, w_K) \in [0,1]^K | \sum_{k=1}^{K} w_k = 1 \right\}.$$

$\boldsymbol{\alpha}$ is the vector of the connection parameters:

$$\boldsymbol{\alpha} = \left(\alpha_{kk'}^{qq'} \right)_{(k,k') \in \mathcal{A}^{qq'}, (q,q') \in \mathcal{E}}.$$

If $q \neq q'$, $\mathcal{A}^{qq'} = [\![1, K_q]\!] \times [\![1, K_{q'}]\!]$. Otherwise, the definition of $\mathcal{A}^{qq}$ depends on the nature of the interaction expressed in X^{qq}. If the interaction is directed, then $\mathcal{A}^{qq} = [\![1, K_q]\!]^2$; on the contrary, if the relation is not directed then $\mathcal{A}^{qq'} = \{(k, k') \in [\![1, K_q]\!] |\ k \leq k'\}^2$. As a consequence $\boldsymbol{\alpha} \in \mathcal{O}_{\boldsymbol{K}} = [0,1]^{\sum_{(q,q') \in \mathcal{E}} |\mathcal{A}^{qq'}|}$ where $|\cdot|$ stands for the cardinal.

Subsection 4.1 is dedicated to the calculus of the likelihood function. This likelihood is maximized in Subsection 4.2 for a fixed $\boldsymbol{K}$. Subsection 4.3 deals with the choice of $\boldsymbol{K}$.

4.1 Complete and marginal likelihoods

In this section, the numbers of clusters $\boldsymbol{K}$ is fixed. For the sake of clarity, we prefer the notation θ over $\theta_{\boldsymbol{K}}$ when no confusion is possible.

Let $\ell(\boldsymbol{X}; \theta)$ denote the likelihood of the observations $\boldsymbol{X}$ for parameter θ. Equations (1) and (2) allow to write explicitly the so-called complete likelihood $\ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta) = f(\boldsymbol{X}|\boldsymbol{Z}; \boldsymbol{\alpha}) f(\boldsymbol{Z}; \boldsymbol{\pi})$:

$$\begin{aligned} \log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta) = & \textstyle\sum_{q,i,k} \mathbb{1}_{Z_i^q = k} \log(\pi_k^q) + \\ & \sum_{(q,q') \in \mathcal{E}} \sum_{(i,i') \in \mathcal{S}^{qq'}} \sum_{(k,k') \in [\![1, K_q]\!]^2} \mathbb{1}_{Z_i^q = k} \mathbb{1}_{Z_{i'}^{q'} = k'} b(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'}) \end{aligned} \tag{3}$$

with

$$b(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'}) = X_{ii'}^{qq'} \log(\alpha_{kk'}^{qq'}) + (1 - X_{ii'}^{qq'}) \log(1 - \alpha_{kk'}^{qq'}). \tag{4}$$

$\boldsymbol{Z}$ being latent variables, the log-likelihood of the observed data $\log \ell(\boldsymbol{X}; \theta)$ is obtained by integrating the complete likelihood over all the possible values of $\boldsymbol{Z}$.

$$\log \ell(\boldsymbol{X}; \theta) = \log \sum_{\boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}} \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta). \tag{5}$$

However, $\boldsymbol{\mathcal{Z}} = \otimes_{q=1}^{Q} [\![1, K_q]\!]^{n_q}$, which implies that, when Q and the K_q's increase this summation becomes intractable. In this context, the variational version of the EM algorithm has proved to be a powerful tool for maximum likelihood inference [see Govaert and Nadif, 2008, Daudin et al., 2008].

The next section describes a variational version of the EM algorithm adapted to our model. The details are postponed to Appendix A.

4.2 Variational EM for maximum likelihood

The EM algorithm [Dempster et al., 1977] applies when the observed data can be enhanced by latent variables. The EM alternates between performing an expectation (E) step computing

$$Q(\theta|\theta^{(t-1)}) = \mathbb{E}[\log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta)|\boldsymbol{X}; \theta^{(t-1)}]$$

and a maximization (M) step computing

$$\theta^{(t)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(t-1)}).$$

The (E)-step requires to compute the conditional distribution $\mathbb{P}(\boldsymbol{Z}|\boldsymbol{X}; \theta)$ for any θ. However, as stressed in Section 3, the (Z_i^q) are not independent when conditioned by the observations $\boldsymbol{X}$, making this calculation unfeasible in a reasonable computational time when $\boldsymbol{\mathcal{Z}}$ is large.

The variational version of the EM algorithm replaces the complex distribution $\mathbb{P}(\boldsymbol{Z}|\boldsymbol{X}; \theta)$ by an optimized simpler version and maximizes a lower bound of the observed likelihood. More precisely, let $\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}$ be any probability distribution on $\boldsymbol{Z}$, we define $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$ as:

$$\begin{aligned}
\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}) &= \log \ell(\boldsymbol{X}; \theta) - \mathbf{KL}[\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}, \mathbb{P}(\cdot|\boldsymbol{X}; \theta)] && (6)\\
&= \mathbb{E}_{\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}} [\log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta)] + \mathcal{H}(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}) && (7)\\
&\leq \log \ell(\boldsymbol{X}; \theta) && (8)
\end{aligned}$$

where $\mathbf{KL}$ is the Kullback-Leibler divergence and $\mathcal{H}(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$ is the entropy of $\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}$. The inequality in (8) derives from the positivity of the $\mathbf{KL}$ divergence and the equality $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}) = \log \ell(\boldsymbol{X}; \theta)$ holds iff $\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}(\boldsymbol{Z}) = \mathbb{P}(\boldsymbol{Z}|\boldsymbol{X}; \theta)$.

The principle of the VEM algorithm is to choose $\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}$ in a family of distributions $\mathcal{P}$ such that the conditional expectation $\mathbb{E}_{\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}} [\log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta)]$ can be computed explicitly. As a result, iteration (t) of VEM consists in the two following steps:

- **M Step** $\theta^{(t)} = \arg\max_\theta \mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau}^{(t-1)},\boldsymbol{X}})$,
- **VE Step**

$$\begin{aligned}
\boldsymbol{\tau}^{(t)} &= \arg\max_{\tau} \mathbb{E}_{\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}} [\log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta)] + \mathcal{H}(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})\\
&= \arg\min_{\tau} \mathbf{KL}[\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}, \mathbb{P}(\cdot|\boldsymbol{X}; \theta^{(t)})]\,.
\end{aligned}$$

The variational EM generates a sequence $(\theta^{(t)}, \boldsymbol{\tau}^{(t)})_{t\geq 0}$ increasing the lower bound $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$ of the likelihood $\log \ell(\boldsymbol{X};\theta)$.

The key point of the procedure is the choice of $\mathcal{P}$, making the calculus tractable but rich enough to be a good approximation of the true conditional distribution $\mathbb{P}(\boldsymbol{Z}|\boldsymbol{X};\theta)$. Following Govaert and Nadif [2008] and Daudin et al. [2008], we adopt the mean-field strategy [Jaakkola, 2000] and chose $\mathcal{P}$ as:

$$\mathcal{P} = \left\{ \mathcal{R}_{\boldsymbol{\tau}} | \mathcal{R}_{\boldsymbol{\tau}}(\boldsymbol{Z}) = \prod_{q=1}^{Q} \prod_{i=1}^{n_q} h_{K_q}(Z_i^q; \boldsymbol{\tau}_i^q) \right\},$$

where $h_{K_q}(\cdot;\xi)$ is the density of a 1 trial - multinomial distribution of parameter $\xi \in \mathcal{T}_{K_q}$, i.e. $\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}$ is such that the latent variables $\boldsymbol{Z}$ are independent and $\mathbb{P}_{\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}}(Z_i^q = k) = \tau_{ik}^q$ with

$$\sum_{k=1}^{K_q} \tau_{ik}^q = 1, \qquad \forall q \in [\![1, Q]\!], \forall i \in [\![1, n_q]\!].$$

From this particular choice of $\mathcal{P}$, we derive the following VEM algorithm.

Variational EM algorithm for MBM

At iteration (t), given the current state $\boldsymbol{\tau}^{(t-1)}$,

- **M Step** $\forall (q,q') \in \mathcal{E}, \quad \forall (k,k') \in \mathcal{A}^{qq'}$

$$\alpha_{kk'}^{qq'(t)} = \frac{\sum_{(i,i')\in\mathcal{S}^{qq'}} X_{ii'}^{qq'} \tau_{ik}^{q(t-1)} \tau_{i'k'}^{q'(t-1)}}{\sum_{(i,i')\in\mathcal{S}^{qq'}} \tau_{ik}^{q(t-1)} \tau_{i'k'}^{q'(t-1)}}$$

  and $\forall q \in [\![1,Q]\!], \forall k \in [\![1, K_q]\!]$:

$$\pi_k^{q(t)} = \sum_{i=1}^{n_q} \tau_{ik}^{q(t-1)} / n_q \,.$$

- **VE Step** Get $\boldsymbol{\tau}^{(t)}$ solving the following sytem: $\forall q \in [\![1,Q]\!], \forall k \in [\![1, K_q]\!], \forall i \in [\![1, n_q]\!]$,

$$\begin{aligned} 0 \quad = \quad & -(1 + \log \tau_{ik}^q) + \log \pi_k^{q(t)} + \left[\sum_{q'\in\mathcal{E}_q} \sum_{i'=1}^{n_{q'}} \sum_{k'=1}^{K'_q} b\left(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'(t)}\right) \tau_{i'k'}^{q'} \right] \\ & + \mathbb{1}_{(q,q)\in\mathcal{E}} \sum_{j\in\mathcal{S}_i^{qq}} \sum_{k'=1}^{K_q} b\left(X_{ij}^{qq}, \alpha_{kk'}^{qq(t)}\right) \tau_{jk'}^q + \mathbb{1}_{(q,q)\in\mathcal{E}} \mathbb{1}_{(i,i)\in\mathcal{S}^{qq}} b\left(X_{ii}^{qq}, \alpha_{kk}^{qq(t)}\right) \\ & + \lambda_i^q \end{aligned}$$

  with

- $\sum_{k=1}^{K_q} \tau_{ik}^q = 1$,
- $\mathcal{E}_q = \{q' \in [\![1, Q]\!] | q' \neq q \text{ and } (q, q') \in \mathcal{E}\}$,
- $\forall (q, q) \in \mathcal{E}, \forall i \in [\![1, n_q]\!]$,

$$\mathcal{S}_i^{qq} = \{i' \in [\![1, n_q]\!] \mid i' \neq i,\ (i, i') \in \mathcal{S}^{qq}\}\ ,$$

and the $(\lambda_j^{q'})_{1 \leq q' \leq Q, 1 \leq j \leq n_{q'}}$ are the Lagrange multipliers.

The details of the calculus are given in Appendix A. We denote by $\widehat{\theta}$ and $\widehat{\boldsymbol{\tau}}$ the resulting estimates. The estimated clustering is the MAP : $\forall q \in [\![1, Q]\!], \forall i \in [\![1, n_q]\!]$,

$$\hat{Z}_i^q = \underset{k \in [\![1, K_q]\!]}{\arg\max}\ \widehat{\tau}_{ik}^q.$$

In practice, the algorithm may converge towards a local optimum. As a consequence, the initialization $\boldsymbol{\tau}^{(0)}$ has a strong impact on the estimates. Obviously, a completely random initialization will lead to a poor optimization. One can derive a $\boldsymbol{\tau}^{(0)}$ from a first clustering of the individuals in each functional group by performing a Hierarchical Cluster Analysis, based on the Manhattan distances between all the rows or columns of the interaction matrices related to the functional group. More sophisticated initializations will be discussed hereafter, in Subsection 4.4.

4.3 Penalized likelihood criterion

In practice, the numbers of clusters $\boldsymbol{K} = (K_1, \ldots, K_Q)$ are unknown and have to be estimated. For standard LBM Biernacki et al. [2000] propose to use the Integrated Completed Likelihood criterion (ICL). In this paper, we adapt it to our MBM. In what follows, $\mathcal{M}_{\boldsymbol{K}}$ refers to the model defined by Equations (1) and (2) with $\boldsymbol{K} = (K_1, \ldots, K_Q)$. ICL is a Bayesian model selection criterion derived as follows.

First assume that the $\boldsymbol{Z}$ are observed and set a prior distribution on $\theta_{\boldsymbol{K}}$ denoted $\pi_{\mathcal{M}_{\boldsymbol{K}}}(\theta)$. Bayesian model selection is based on the integrated complete likelihood:

$$\log m_c(\boldsymbol{X}, \boldsymbol{Z}; \mathcal{M}_{\boldsymbol{K}}) = \log \int_{\Theta_{\boldsymbol{K}}} \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta_{\boldsymbol{K}}) d\pi_{\mathcal{M}_{\boldsymbol{K}}}(\theta_{\boldsymbol{K}})$$

Considering the following prior distributions on θ:

$$\alpha_{kk'}^{qq'} \sim \mathcal{B}(a, a) \quad \text{and} \quad (\pi_1^q, \ldots \pi_{K_Q}^q) \sim \mathcal{D}ir(b, \ldots, b)$$

we can supply an explicit expression for $\log m_c(\boldsymbol{X}, \boldsymbol{Z}; \mathcal{M}_{\boldsymbol{K}})$. From this explicit expression, we are able to derive its asymptotic approximation, in the same spirit as the standard Bayesian Information Criterion (BIC):

$$\log m_c(\boldsymbol{X}, \boldsymbol{Z}; \mathcal{M}_{\boldsymbol{K}})$$
$$\sim_{n_1 \ldots n_Q \to \infty} \max_{\theta_{\boldsymbol{K}} \in \Theta_{\boldsymbol{K}}} \log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta_{\boldsymbol{K}}) - \text{pen}(\mathcal{M}_{\boldsymbol{K}})$$

with

$$\begin{aligned} \text{pen}(\mathcal{M}_{\boldsymbol{K}}) &= \tfrac{1}{2} \Big\{ \textstyle\sum_{q=1}^{Q} (K_q - 1) \log(n_q) \\ &\quad + \Big(\textstyle\sum_{(q,q') \in \mathcal{E}} |\mathcal{A}^{qq'}| \Big) \log \Big(\textstyle\sum_{(q,q') \in \mathcal{E}} |\mathcal{S}^{qq'}| \Big) \Big\} . \end{aligned} \tag{9}$$

The proof of this result is given in Appendix B.
Similarly to LBM or SBM, in our model, $\boldsymbol{Z}$ is not observed. Two strategies can be adopted: either the $\boldsymbol{Z}$ are imputed using their maximum a posteriori value [Biernacki et al., 2000], either the $\boldsymbol{Z}$ are integrated out [see Daudin et al., 2008] with the conditional distribution $\mathbb{P}(\boldsymbol{Z}|\boldsymbol{X}; \hat{\theta})$. We adopt the last strategy and derive the following model selection criterion:

$$\text{ICL}(\mathcal{M}_{\boldsymbol{K}}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X}; \widehat{\theta}_{\boldsymbol{K}}} \Big[\log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \hat{\theta}_{\boldsymbol{K}}) \Big] - \text{pen}(\mathcal{M}_{\boldsymbol{K}})$$

where $\text{pen}(\mathcal{M}_{\boldsymbol{K}})$ has been defined in Equation (9). Finally, $\mathbb{P}(\boldsymbol{Z}|\boldsymbol{X}; \widehat{\theta}, \mathcal{M}_{\boldsymbol{K}})$ being non-explicit, it is replaced in practice by $\mathcal{R}_{\boldsymbol{X}, \widehat{\tau}}$, leading to the following penalized selection criterion:

$$\widehat{\text{ICL}}(\mathcal{M}_{\boldsymbol{K}}) = \mathbb{E}_{\mathcal{R}_{\boldsymbol{X}, \widehat{\tau}}} \Big[\log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \widehat{\theta}_{\boldsymbol{K}}) \Big] - \text{pen}(\mathcal{M}_{\boldsymbol{K}}) . \tag{10}$$

Remark 2 *ICL is known to automatically encourage clustering configurations with well separated clusters. Its capacity to outline the clustering structure in the data has been tested in the literature, either in mixture models [Baudry et al., 2008], LBM [Keribin et al., 2014] or SBM [Mariadassou et al., 2010].*

4.4 Practical algorithm

The practical choice of the model and the estimation of its parameters is a computational intensive task. Indeed, assume that each number of cluster K_q can vary from $K_{q,\text{inf}}$ and $K_{q,\text{sup}}$, then, ideally, we should compare $\prod_{q=1}^{Q} (K_{q,\text{sup}} - K_{q,\text{inf}})$ models. For each model, the variational EM algorithm has to be run starting from a large number of initializations (due to its sensibility to its starting point), resulting into an unreasonable computational cost. Instead, we propose to adopt a stepwise strategy, resulting into a faster exploration of the model space and judicious initializations of the VEM algorithm. The procedure we suggest is the following one.
Starting from a model $\mathcal{M}^{(0)} = \mathcal{M}(K_1^{(0)}, \ldots, K_Q^{(0)})$. The m-th iteration is written as follows.

Model selection strategy for MBM

Given a current model $\mathcal{M}^{(m)} = \mathcal{M}(K_1^{(m)}, \ldots, K_Q^{(m)})$

- **Split proposals**:
  For any $q \in [\![1, Q]\!]$ such that $K_q^{(m)} < K_{q,\sup}$, consider the model $\mathcal{M}(K_1^{(m)}, \ldots, K_q^{(m)} + 1, \ldots, K_Q^{(m)})$
  - Propose $K_q^{(m)}$ initializations by splitting any of the $K_q^{(m)}$ clusters into two clusters.
  - From each of the $K_q^{(m)}$ initialization points, run the VEM algorithm and keep the better estimate, i.e. the one which maximizes the lower bound $\mathcal{I}$:
    $$(\hat{\theta}^{(q,m,+)}, \hat{\tau}^{(q,m,+)})$$
  - Compute the corresponding ICL.
- **Merge proposals**:
  For any $q \in [\![1, Q]\!]$ such that $K_q^{(m)} > K_{q,\inf}$, consider the model $\mathcal{M}(K_1^{(m)}, \ldots, K_q^{(m)} - 1, \ldots, K_Q^{(m)})$.
  - Propose $\frac{K_q^{(m)}(K_q^{(m)}-1)}{2}$ initializations by merging any pairs of clusters among the $K_q^{(m)}$ clusters.
  - From each initialization point, run the VEM algorithm and keep the better estimates:
    $$(\hat{\theta}^{(q,m,-)}, \hat{\tau}^{(q,m,-)})$$
  - Compute the corresponding ICL.
- Let us define

$$\begin{aligned}\mathbb{M}^{(m)} \quad = \quad & \bigcup_{q \in [\![1,\ldots Q]\!]} \Big\{ \mathcal{M}(K_1^{(m)}, \ldots, K_q^{(m)} + 1, \ldots, K_Q^{(m)}) \\ & \mathcal{M}(K_1^{(m)}, \ldots, K_q^{(m)} - 1, \ldots, K_Q^{(m)}) \Big\} \cup \mathcal{M}^{(m)} .\end{aligned}$$

  Set $\mathcal{M}^{(m+1)} = \underset{\mathcal{M} \in \mathbb{M}^{(m)}}{\arg\max}\, ICL(\mathcal{M})$.

  If $\mathcal{M}^{(m+1)} \neq \mathcal{M}^{(m)}$ iterate, otherwise stop.

This algorithm is implemented in the R-package GREMLIN, available on GitHub. Note that the tasks at each iteration can be easily parallelized.

5 Applications

We now apply our methodology on our two examples respectively in ecology and ethnobiology.

5.1 Ecology: interactions plants / animals

Dataset The dataset –compiled and conducted by Dáttilo et al. [2016] at Centro de Investigaciones Costeras La Mancha (CICOLMA), located on the central coast of the Gulf of Mexico, Veracruz, Mexico– involves three general types of plant-animal mutualistic interaction: pollination, seed dispersal by frugivorous birds, and protective mutualisms between ants and plants with extrafloral nectaries. The dataset –which is one of the largest compiled so far with respect to species richness, number of interactions and sampling effort– includes $n_1 = 141$ plant species, $n_2 = 173$ pollinator species, $n_3 = 46$ frugivorous bird species and $n_4 = 30$ ant species. The dataset includes 753 observed interactions of which 55% are plant-pollinator interactions, 17% are plant-birds interactions and 28% are plant-ant interactions. The dataset is plotted in Dáttilo et al. [2016] using two alternative representations: either a unique plot involving all the individuals where the color of a node refers to the functional group, or three bipartite networks, each of them involving the same functional group "plants". These two alternative network representations highlight the fact that there is a need for a mesoscopic representation of the three networks.

Inference We run the procedure described in Subsection 4.4 starting from several automatically chosen initial points $\boldsymbol{K}^{(0)}$, with numbers of clusters bounded between 1 and 10. The computational time on an `Intel® Xeon(R) CPU E5-1650 v3 @ 3.50GHz x12` using 6 cores is less than 10 minutes. The ICL criterion selects 7 clusters of plants, 2 clusters of pollinators, 1 cluster of birds and 2 clusters of ants. The estimated parameters are reported in Table 1. The resulting mesoscopic view of the multipartite network is plotted in Figure 2

Discussion From Figure 2, we deduce that the plants of clusters 7 and 2 only interact with ants (noting that the plants of cluster 7 attract more ant species belonging to cluster 1). The plants of clusters 3 and 6 are only in interaction with birds, the difference between the two clusters being due to the strength of the connection. Our ecosystemic approach highlights the central role played by the pollinators. The difference between the two clusters of pollinators derives from the existence of the cluster 1 of plants.

In order to illustrate the contribution of our method, we also analyze each bipartite network separately (using an LBM) and compare the results in terms of clustering. The clusterings are compared through the Adjusted Rand Index (ARI): if $ARI = 1$ then the clusterings are equal (up to a label switching transformation). The ARIs are given in Table 2.

Using standard LBM, we obtain 2 clusters of ants, 1 cluster of birds and 3 clusters of pollinators. The clusterings of ants and birds are not modified

			Pollinators		Ants		Birds
			1	2	1	2	1
		$\hat{\pi}_k^q$	0.06	0.94	0.1	0.9	1
Plants	1	0.4675	0.0957	0.0075	0	0.0006*	0.0013*
	2	0.1606	0.0042*	0	0.5431	0.0589	0
	3	0.1351	0	0.0003*	0	0	0.0753
	4	0.0784	0.1652	0.0343	0.6620	0.1542	0
	5	0.1061	0.1918	0.0638	0	0	0.0163
	6	0.0142	0	0	0	0	0.5108
	7	0.0380	0	0	0.8492	0.3565	0

Table 1: Estimated parameters for MBM: $\hat{\pi}_k^q$ are in the first row and column, other rows and columns contain the estimates $\widehat{\alpha}_{kk'}^{qq'}$. $\widehat{\alpha}_{kk'}^{qq'}$ identified by * are not plotted in Figure 2.

by the ecosystemic approach, their ARI being equal to 1. The clustering of the pollinators is slightly modified, going from 3 clusters to 2 clusters but the additional block only contains few individuals, thus leading to a high ARI. Since the plants functional group is involved in the three bipartite networks, we obtain 3 clusterings when analyzing them separately. These three clusterings are –as expected– very different from our MBM clustering (the ARIs being respectively equal to 0.118, 0.415 and 0.163, see Table 2). When aiming at proposing a clustering taking into account the 3 bipartite networks, one may adopt a naive strategy by combining (by intersection) these three clusterings. We then obtain 12 blocks of plants and the ARI with the MBM clustering is 0.617. However, this number of clusters (12) is too large with respect to the model selection criterion. Our MBM clustering is a parsimonious trade-off.

Table 2: Comparison of clusterings when the networks are jointly modeled by the MBM (denoted Full) and when the networks are considered apart as bipartite networks. Inter denotes the clustering obtained by intersecting the three clusterings on plants for each bipartite network. The selected number of clusters (in parenthesis) and the ARIs are provided.

	Full/Flovis	Full/Ants	Full/Birds	Full/Inter
Plants	(7/3) 0.118	(7/3) 0.415	(7/3) 0.163	(7/12) 0.617
Flovis	(2/3) 0.997			
Ants		(2/2) 1.000		
Birds			(1/1) 1.000	

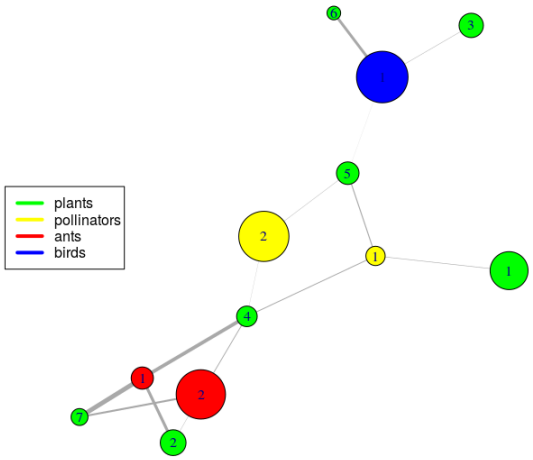


Figure 2: Mesoscopic view of the interconnected network. The size of the nodes are proportional to the size of the clusters and the width of the edges are proportional to the probability of connection between/within clusters. Edges corresponding to probabilities of connection lower than 0.01 are not plotted.

5.2 Seed circulation and crop species inventory

Dataset Seed circulation among farmers is a key process that shapes crop diversity [Coomes et al., 2015, Pautasso et al., 2013]. With the courtesy of Sophie Caillon, we analyze data on seed circulation and inventory data. Data on seed circulation within a community of first-generation migrants (30 farmers) were collected by a field survey in the island of Vanua Lava in the South Pacific archipelago nation of Vanuatu. A farmer is considered as a giver for another farmer if he/she has given at least one crop since they arrived in the new settlement site in Vanua Lava. It results in a connected and directed network of seed circulation. Besides the circulation network, inventory data for each farmer were also collected. They consist in the list of crop landraces they grow. This was aggregated at the species level, leading to 37 crop species. These inventory data can be seen as a bipartite network. The seed circulation data were analyzed in Thomas and Caillon [2016] and the inventory data were analyzed in a meta-analysis in Thomas et al. [2015]. On the basis on the joint modeling we propose in this paper, we aim to provide a clustering on farmers and crop species on the basis of the seed circulation network and the inventory bipartite

network. The two functional groups at stakes are the farmers and the species.

Inference The MBM is inferred by the proposed VEM algorithm and the numbers of clusters are selected by the ICL criterion. Three clusters of farmers and two clusters of crop species were selected. The inferred model is displayed as a mesoscopic view in Figure 3. The clusters were renumbered to make them correspond to the probability of connection: the larger cluster number, the larger marginal probability of connection for the members of the cluster.

Discussion The discovered clusters are straightforwardly interpretable: Cluster 3 gathers farmers who circulate seeds within the cluster and give to the two other clusters, Cluster 2 circulates seeds within the cluster contrary to Cluster 1 who only receives from Cluster 3; the two clusters of crop species are Cluster 2 with more common crop species and Cluster 1 with other species. Clusters 3 and 2 of farmers grow crop species from Clusters 1 and 2 whereas Cluster 1 of farmers grows mainly crop species from Cluster 2. It turns out that Cluster 3 gathers mainly the first migrants and Cluster 1 the last migrants. The pattern of connection is then explained by the fact that first migrants helped the others to settle by providing seeds. Moreover, the first migrants had more time to collect more crop species to grow. Table 3 provides the comparison between the obtained clustering (MBM) and a stand-alone clustering on farmers resulting from an SBM on the circulation network and the clusterings on farmers and crop species from an LBM on the inventory network. The clusterings on crop species remain quite close while the clusterings on farmers are different since the MBM shall make a trade off between the circulation and the inventory for farmers. To ease the comparison between clusterings on farmers, the same renumbering rule was applied for all the different clusterings so that the larger cluster numbers correspond to larger marginal probability of connection. Figure 4 is an alluvial plot which compares the three obtained clusterings of farmers. It shows how the trade-off is made between the two stand-alone clusterings in the MBM clustering. It appears quite obvious that Cluster 1 given by the MBM gathers only farmers from Cluster 1 in the seed circulation network and from Clusters 1 and 2 from the inventory data since this cluster aggregates farmers with fewer connections and who grow less crop species than the others. The same kind of observation can be made for Cluster 3 given by the MBM which aggregates farmers who were in the cluster of the most connected farmers and in the two clusters of farmers who grow more seed.

6 Discussion

In this paper, we proposed an extension of LBM and SBM which can handle multipartite networks, resulting into the so-called MBM. Multipartite networks encompass a lot of various situations such as the two examples dealt with in the paper. Besides, MBM can be also useful for many other contexts with different multipartite structures. Several extensions can be thought of with no additional

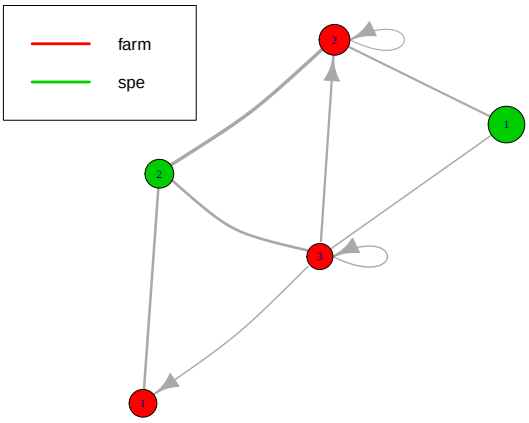


Figure 3: Summarized network provided by the MBM. Nodes correspond to the clusters detected by the MBM: clusters of farmers are in red and clusters of crop species are in green. Size of a node is proportional to the number of farmers or crop species belonging to the considered cluster. The width of the edges are proportional to the probability of connection between/within clusters. The probability of connection below 0.2 are not plotted.

significant difficulty. For instance, one may be interested in considering not only binary interactions but also valued interactions ($X_{ii'}^{qq'} \in \mathbb{N}$ or $\mathbb{R}$) or multiplex interactions ($X_{ii'}^{qq'} \in \{0,1\}^d$). These extensions raise no major difficulties since they only require to modify the function $b(X_{ii'}^{qq'}, \alpha_{qq'}^{kk'})$. Covariates can also be taken into account by writing $\mathbb{P}(X_{ii'}^{qq'} = 1 | Z_i^q = k, Z_{i'}^{q'} = k') = \phi\left(\alpha_{qq'}^{kk'} + \mathbf{y}_{ii'}\beta\right)$ where $\mathbf{y}_{ii'}$ are the covariates.

The main limiting factor is the size of networks. Indeed, the inference algorithm that we propose is suitable for networks up to 1000 nodes in order to keep computational time reasonable. If willing at studying larger networks, one should develop adapted inference algorithms.

In a more general perspective, the study of ecological or sociological interactions supplies a wide variety of complex networks such as multilevel networks or multi-layer networks (see for instance Pilosof et al. [2016] or Lazega and Snijders [2015]). Some of them can be treated as multipartite networks and then by a MBM. The others require the development of suited models which might also rely on a latent variable modeling. They will be the subject of future works.

Table 3: Comparison of clusterings when the networks are jointly modeled by the MBM (denoted Full) and when the networks are considered apart as a circulation network for farmer and bipartite network for indiviuals and crop species. Inter denotes the clustering obtained by intersecting the two clusterings on farmers. The selected number of clusters (in parenthesis) and the ARI are provided.

	Full/Exc	Full/Inv	Full/Inter
Ind	(2/3)	(3/3)	(3/5)
	0.297	0.105	0.273
Spe		(2/2)	
		0.891	

Acknowledgements

We thank Sophie Caillon (CEFE) for sharing the seed circulation and inventory data, Sophie Caillon and Mathieu Thomas (CIRAD) for related discussion concerning the analysis. Support of the ongoing research collaboration through the MIRES network was provided by the Institut National de la Recherche Agronomique (INRA).

A. VEM details

Let us recall the quantities at stake in VEM. $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$ is a lower bound of the log-likelihood defined as:

$$\begin{aligned} \mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}) &= \log \ell(\boldsymbol{X};\theta) - \mathbf{KL}[\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}, p(\cdot|\boldsymbol{X};\theta)] && (11)\\ &= \mathbb{E}_{\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}}\left[\log \ell_c(\boldsymbol{X},\boldsymbol{Z};\theta)\right] + \mathcal{H}(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}) && (12)\end{aligned}$$

with –in our case– $\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}} \in \mathcal{P}$ where

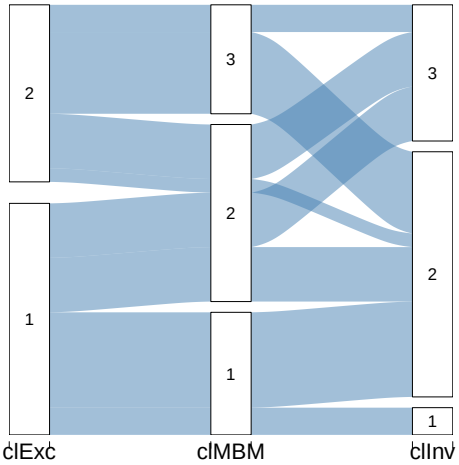
$$\mathcal{P} = \left\{ \mathcal{R}_{\boldsymbol{\tau}} \mid \forall \boldsymbol{Z} \in \boldsymbol{\mathcal{Z}}, \mathcal{R}_{\boldsymbol{\tau}}(\boldsymbol{Z}) = \prod_{q=1}^{Q} \prod_{i=1}^{n_q} h_{K_q}(Z_i^q; \boldsymbol{\tau}_i^q) \right\},$$

and $h_{K_q}(\cdot;\xi)$ is the density of a 1 trial - multinomial distribution of parameter $\xi \in \mathcal{T}_{K_q}$, i.e. $\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}$ is such that the latent variables $\boldsymbol{Z}$ are independent and $\mathbb{P}_{\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}}(Z_i^q = k) = \tau_{ik}^q$ with

$$\sum_{k=1}^{K_q} \tau_{ik}^q = 1, \qquad \forall q \in [\![1, Q]\!], \forall i \in [\![1, n_q]\!].$$

VEM is an alternate optimization of $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$ with respect to $\boldsymbol{\tau}$ and θ. We will now detail these two steps in the particular case of our block model for binary multipartite networks.

Figure 4: Alluvial plot comparing the clustering on farmers obtained from an SBM on the circulation network (clExc), an LBM on the inventory network (clInv) and the MBM (clMBM) on both networks. The cluster numbers are related with the probability of connection, the larger cluster number, the larger marginal probability of connection (between farmers for clExc, between farmers and crop species for clInv).



Explicit expression for $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$.

Using the expression of the complete log-likelihood given in Equations (3) and (12), we obtain:

$$\begin{aligned}\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}}) &= -\sum_{q,i,k} \tau_{ik}^q \log \tau_{ik}^q + \sum_{q,i,k} \tau_{ik}^q \log(\pi_k^q) \\ &+ \sum_{(q,q')} \sum_{(i,i')} \sum_{(k,k')} \mathbb{E}\left[\mathbb{1}_{Z_i^q=k} \mathbb{1}_{Z_{i'}^{q'}=k'}\right] b(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'})\end{aligned} \tag{13}$$

with

$$b(X_{ii'}^{qq'}, \alpha_{kk'}^{qq'}) = X_{ii'}^{qq'} \log(\alpha_{kk'}^{qq'}) + (1 - X_{ii'}^{qq'}) \log(1 - \alpha_{kk'}^{qq'}) .$$

$\mathbb{E}\left[\mathbb{1}_{Z_i^q=k} \mathbb{1}_{Z_{i'}^{q'}=k'}\right]$ has to be carefully calculated, when $i = i'$. To that purpose, let us introduce the following notations :

- $\forall q,\ \mathcal{E}_q = \{q' \in [\![1, Q]\!] \mid q' \neq q \text{ and } (q, q') \in \mathcal{E}\}$. $\mathcal{E}_q$ is the set of incidence matrices involving the functional group q.
- $\forall (q, q') \in \mathcal{E}, \forall i \in [\![1, n_q]\!]$ we define :

$$\mathcal{S}_i^{qq} = \{i' \in [\![1, n_q]\!] \mid i' \neq i,\ (i, i') \in \mathcal{S}^{qq}\} .$$

Using these notations we detail the expression of $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$.

$$
\begin{aligned}
&\sum_{(q,q')}\sum_{(i,i')}\sum_{(k,k')}\mathbb{E}\left[\mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_{i'}^{q'}=k'}\right]b(X_{ii'}^{qq'},\alpha_{kk'}^{qq'}) = \sum_{q}\sum_{q'\in\mathcal{E}_q}\sum_{(i,i')}\sum_{(k,k')}\tau_{ik}^{q}\tau_{i'k'}^{q'}b(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})\\
&+\mathbb{1}_{(q,q)\in\mathcal{E}}\sum_{(i,i')}\sum_{(k,k')}\mathbb{E}\left[\mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_{i'}^{q}=k'}\right]b(X_{ii'}^{qq},\alpha_{kk'}^{qq})\\
=\;&\sum_{q}\sum_{q'\in\mathcal{E}_q}\sum_{(i,i')}\sum_{(k,k')}\tau_{ik}^{q}\tau_{i'k'}^{q'}b(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})+\mathbb{1}_{(q,q)\in\mathcal{E}}\sum_{i}\sum_{i'\in\mathcal{S}_i^{qq}}\sum_{(k,k')}\mathbb{E}\left[\mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_{i'}^{q}=k'}\right]b(X_{ii'}^{qq},\alpha_{kk'}^{qq})\\
&+\mathbb{1}_{(q,q)\in\mathcal{E}}\sum_{i}\sum_{(k,k')}\mathbb{E}\left[\underbrace{\mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_i^q=k'}}_{=0\text{ if }k\neq k'}\right]b(X_{ii}^{qq},\alpha_{kk'}^{qq})\\
=\;&\sum_{q}\sum_{q'\in\mathcal{E}_q}\sum_{(i,i')}\sum_{(k,k')}\tau_{ik}^{q}\tau_{i'k'}^{q'}b(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})+\mathbb{1}_{(q,q)\in\mathcal{E}}\sum_{i}\sum_{i'\in\mathcal{S}_i^{qq}}\sum_{(k,k')}\tau_{ik}^{q}\tau_{i'k'}^{q}b(X_{ii'}^{qq},\alpha_{kk'}^{qq})\\
&+\mathbb{1}_{(q,q)\in\mathcal{E}}\sum_{i\mid(i,i)\in\mathcal{S}^{qq}}\sum_{k}\mathbb{E}\underbrace{\left[\mathbb{1}_{Z_i^q=k}^2\right]}_{=\mathbb{1}_{Z_i^q=k}}b(X_{ii}^{qq},\alpha_{kk}^{qq})\,.
\end{aligned}
$$

As a consequence, we get:

$$
\begin{aligned}
&\textstyle\sum_{(q,q')}\sum_{(i,i')}\sum_{(k,k')}\mathbb{E}\left[\mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_{i'}^{q'}=k'}\right]b(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})\\
=\;&\textstyle\sum_{q}\sum_{q'\in\mathcal{E}_q}\sum_{(i,i')}\sum_{(k,k')}\tau_{ik}^{q}\tau_{i'k'}^{q'}b(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})\\
&\textstyle+\mathbb{1}_{(q,q)\in\mathcal{E}}\sum_{i}\sum_{i'\in\mathcal{S}_i^{qq}}\sum_{(k,k')}\tau_{ik}^{q}\tau_{i'k'}^{q}b(X_{ii'}^{qq},\alpha_{kk'}^{qq})\\
&\textstyle+\mathbb{1}_{(q,q)\in\mathcal{E}}\sum_{i\mid(i,i)\in\mathcal{S}^{qq}}\sum_{k}\tau_{ik}^{q}b(X_{ii}^{qq},\alpha_{kk}^{qq})\,.
\end{aligned}
\tag{14}
$$

Optimization of $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$ with respect to $\boldsymbol{\tau}$, (θ being fixed)

For a fixed θ, we need to find $\boldsymbol{\tau}$ such that $\forall q\in[\![1,Q]\!]$, $\forall k\in[\![1,K_q]\!]$, $\forall i\in[\![1,n_q]\!]$:

$$
\frac{\partial}{\partial\tau_{ik}^{q}}\left[\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})+\sum_{q'=1}^{Q}\sum_{j=1}^{n_{q'}}\lambda_j^{q'}\left(\sum_{k'=1}^{K_q}\tau_{jk'}-1\right)\right]=0 \tag{15}
$$

where $(\lambda_j^{q'})_{1\leq q'\leq Q,1\leq j\leq n_{q'}}$ are the Lagrange multipliers. Combining Equations (13) and (14), we get:

$$
\begin{aligned}
0\;=\;&-(1+\log\tau_{ik}^{q})+\log\pi_k^q+\left[\sum_{q'\in\mathcal{E}_q}\sum_{i'=1}^{n_{q'}}\sum_{k'=1}^{K_q'}b(X_{ii'}^{qq'},\alpha_{kk'}^{qq'})\tau_{i'k'}^{q'}\right] \qquad (16)\\
&+\mathbb{1}_{(q,q)\in\mathcal{E}}\sum_{j\in\mathcal{S}_i^{qq}}\sum_{k'=1}^{K_q}b(X_{ij}^{qq},\alpha_{kk'}^{qq})\tau_{jk'}^{q}+\mathbb{1}_{(q,q)\in\mathcal{E}}\mathbb{1}_{(i,i)\in\mathcal{S}^{qq}}b(X_{ii}^{qq},\alpha_{kk}^{qq})\\
&+\lambda_i^q\,.
\end{aligned}
$$

This system has no explicit solution but can be solved numerically using a fixed point strategy as in Daudin et al. [2008].

Optimization of $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$ with respect to θ, $\boldsymbol{\tau}$ being fixed.
We have to compute the derivatives of $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{\tau},\boldsymbol{X}})$ with respect to θ, the variational parameters $\boldsymbol{\tau}$ being fixed. We thus obtain: $\forall (q,q') \in \mathcal{E}, \forall (k,k') \in [\![1,K_q]\!] \times [\![1,K_{q'}]\!]$:

$$\alpha_{kk'}^{qq'} = \frac{\sum_{(i,i')\in\mathcal{S}^{qq'}} X_{ii'}^{qq'} \tau_{ik}^q \tau_{i'k'}^{q'}}{\sum_{(i,i')\in\mathcal{S}^{qq'}} \tau_{ik}^q \tau_{i'k'}^{q'}}$$

and $\forall q \in [\![1,Q]\!], \forall k = [\![1,K_q]\!]$:

$$\pi_k^q = \frac{1}{n_q} \sum_{i=1}^{n_q} \tau_{ik}^q \,.$$

B. Derivation of the ICL criterion

Explicit expression of the marginal complete likelihood The prior we consider is the following one:

$$\alpha_{kk'}^{qq'} \sim \mathcal{B}(a,a) \quad \text{and} \quad (\pi_1^q, \dots \pi_{K_q}^q) \sim \mathcal{D}ir(b,\dots,b). \tag{17}$$

We work for a fixed $\boldsymbol{K}$, thus we use the following shortcut : $\theta = \theta_{\boldsymbol{K}}$. By definition,

$$\log m_c(\boldsymbol{X},\boldsymbol{Z};\mathcal{M}_{\boldsymbol{K}}) = \log \int \ell_c(\boldsymbol{X},\boldsymbol{Z};\theta_{\boldsymbol{K}})\pi(\theta_{\boldsymbol{K}};\mathcal{M}_{\boldsymbol{K}})d\theta_{\boldsymbol{K}} \,.$$

The prior on θ being such that $\pi(\theta) = \pi(\boldsymbol{\alpha})\pi(\boldsymbol{\pi})$ we obtain :

$$\log m_c(\boldsymbol{X},\boldsymbol{Z};\mathcal{M}_{\boldsymbol{K}}) \quad = \quad \log \int f(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\alpha})\pi(\boldsymbol{\alpha})d\boldsymbol{\alpha} + \log \int f(\boldsymbol{Z};\boldsymbol{\pi})\pi(\boldsymbol{\pi})d\boldsymbol{\pi} \,.$$

Taking advantage of the conditional independences in the model defined by Equations (1) and (2) combined with the independence of the parameters in the prior distribution, we can decompose $\log m_c$ into the following sum :

$$\begin{aligned}\log m_c(\boldsymbol{X},\boldsymbol{Z};\mathcal{M}_{\boldsymbol{K}}) \quad = \quad & \sum_{(q,q')\in\mathcal{E}} \log \int f(\boldsymbol{X}^{qq'}|\boldsymbol{Z}^q,\boldsymbol{Z}^{q'};(\boldsymbol{\alpha}^{qq'}))\pi(\boldsymbol{\alpha}^{qq'})d\boldsymbol{\alpha}^{qq'} \\ & + \sum_{q=1}^{Q} \log \int f(\boldsymbol{Z}^q;\boldsymbol{\pi}^q)\pi(\boldsymbol{\pi}^q)d\boldsymbol{\pi}^q \,.\end{aligned}$$

Using the fact that $f(\boldsymbol{Z}^q;\boldsymbol{\pi}^q) = \prod_{k=1}^{K_q}(\pi_k^q)^{N_k^q}$ with

$$N_k^q = \sum_{i=1}^{n_q} \mathbb{1}_{Z_i^q=k} \tag{18}$$

and the conjugacy of the Dirichlet prior distribution, we easily deduce the following formula:

$$\int f(\boldsymbol{Z}^q;\boldsymbol{\pi}^q)\pi(\boldsymbol{\pi}^q)d\boldsymbol{\pi}^q = \frac{\Gamma(bK_q)}{\Gamma(b)^{K_q}} \frac{\prod_{k=1}^{K_q}\Gamma(N_k^q+b)}{\Gamma(n_q+bK_q)}$$

where Γ is the Gamma function. Now, we can reformulate $f(\boldsymbol{X}^{qq'}|\boldsymbol{Z}^q,\boldsymbol{Z}^{q'};\boldsymbol{\alpha}^{qq'})$ as:

$$\begin{aligned} f(\boldsymbol{X}^{qq'}|\boldsymbol{Z}^q,\boldsymbol{Z}^{q'};\boldsymbol{\alpha}^{qq'}) &= \prod_{(i,i',k,k')} (\alpha_{kk'}^{qq'})^{X_{ii'}^{qq'}\mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_{i'}^{q'}=k}}(1-\alpha_{kk'}^{qq'})^{(1-X_{ii'}^{qq'})\mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_{i'}^{q'}=k}} \\ &= \prod_{k,k'=1}^{K_q,K_{q'}} (\alpha_{kk'}^{qq'})^{S_{kk'}^{qq'}}(1-\alpha_{kk'}^{qq'})^{N_{kk'}^{qq'}-S_{kk'}^{qq'}} \end{aligned}$$

with

$$\begin{aligned} S_{kk'}^{qq'} &= \textstyle\sum_{(i,i')\in\mathcal{S}^{qq'}} X_{ii'}^{qq'}\mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_{i'}^{q'}=k} \\ N_{kk'}^{qq'} &= \textstyle\sum_{(i,i')\in\mathcal{S}^{qq'}} \mathbb{1}_{Z_i^q=k}\mathbb{1}_{Z_{i'}^{q'}=k}. \end{aligned} \tag{19}$$

With the beta prior distribution on each $\alpha_{kk'}^{qq'}$, we get:

$$\int f(\boldsymbol{X}^{qq'}|\boldsymbol{Z}^q,\boldsymbol{Z}^{q'};(\boldsymbol{\alpha}^{qq'}))\pi(\boldsymbol{\alpha}^{qq'})d\boldsymbol{\alpha}^{qq'} = \prod_{k,k'=1}^{K_q,K_{q'}} \frac{\Gamma(2a)}{\Gamma(a)^2}\frac{\Gamma(a+S_{kk'}^{qq'})\Gamma(a+N_{kk'}^{qq'}-S_{kk'}^{qq'})}{\Gamma(2b+N_{kk'}^{qq'})}.$$

Finally, we obtain:

$$\begin{aligned} \log m_c(\boldsymbol{X},\boldsymbol{Z};\mathcal{M}_{\boldsymbol{K}}) &= \left(\sum_{(q,q')\in\mathcal{E}} |\mathcal{A}^{qq'}|\right)(\log\Gamma(2a) - 2\log\Gamma(a)) \\ &\quad + \sum_{(q,q',k,k')} \log\Gamma(a+S_{kk'}^{qq'}) + \log\Gamma(a+N_{kk'}^{qq'}-S_{kk'}^{qq'}) \\ &\quad - \sum_{(q,q',k,k')} \log\Gamma(2b+N_{kk'}^{qq'}) \\ &\quad + \sum_{q=1}^{Q} \log\Gamma(bK_q) - K_q\log(b) - \log\Gamma(n_q+bK_q) \\ &\quad + \sum_{q=1}^{Q}\left(\sum_{k=1}^{K_q}\log\Gamma(N_k^q+b)\right) \end{aligned}$$

where N_k^q has been defined in Equation (18) and $S_{kk'}^{qq'}$ and $N_{kk'}^{qq'}$ in Equation (19)

Asymptotic approximation Using the same arguments as in Daudin et al. [2008] and Brault [2014], we obtain the following asymptotic approximation. Assume that $\forall q \in [\![1, Q]\!], n_q \to \infty$, then :

$$\log m_c(\boldsymbol{X}, \boldsymbol{Z}; \mathcal{M}_{\boldsymbol{K}}) \quad = \quad \max_{\theta_{\boldsymbol{K}} \in \Theta_{\boldsymbol{K}}} \log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta_{\boldsymbol{K}}) - \text{pen}(\mathcal{M}_{\boldsymbol{K}})$$

where

$$\begin{aligned} \text{pen}(\mathcal{M}_{\boldsymbol{K}}) \quad &= \quad \frac{1}{2} \sum_{q=1}^{Q} (K_q - 1) \log(n_q) \\ &+ \quad \frac{1}{2} \left(\sum_{(q,q') \in \mathcal{S}^{qq'}} |\mathcal{A}^{qq'}| \right) \log \left(\sum_{(q,q') \in \mathcal{S}^{qq'}} |\mathcal{S}^{qq'}| \right) . \end{aligned}$$

The first term comes from the application of the Stirling formula to the Gamma function when approximating $f(\boldsymbol{Z}^q; \boldsymbol{\pi}^q)$. The second term comes from a BIC approximation of the part $f(\boldsymbol{X}^{qq'} | \boldsymbol{Z}^q, \boldsymbol{Z}^{q'}; \boldsymbol{\alpha}^{qq'})$.

References

Pierre Barbillon, Sophie Donnet, Emmanuel Lazega, and Avner Bar-Hen. Stochastic block models for multiplex networks: An application to a multilevel network of researchers. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 2016. ISSN 1467985X.

Jean-Patrick Baudry, Gilles Celeux, and Jean-Michel Marin. Selecting models focussing on the modeller's purpose. In Paula Brito, editor, *COMPSTAT 2008*, pages 337–348. Physica-Verlag HD, 2008. ISBN 978-3-7908-2083-6.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725, Jul 2000. ISSN 0162-8828.

Vincent Brault. *Estimation et sélection de modèle pour le modèle des blocs latents*. PhD thesis, University Paris-Sud, France, 2014. Thèse de doctorat dirigée par Celeux, Gilles Mathématiques Paris 11 2014.

Oliver T Coomes, Shawn J McGuire, Eric Garine, Sophie Caillon, Doyle McKey, Elise Demeulenaere, Devra Jarvis, Guntra Aistara, Adeline Barnaud, Pascal Clouvel, et al. Farmer seed networks make a limited contribution to agriculture ? four common misconceptions. *Food Policy*, 56:41–50, 2015.

Wesley Dáttilo, Nubia Lara-Rodríguez, Pedro Jordano, Paulo R. Guimarães, John N. Thompson, Robert J. Marquis, Lucas P. Medeiros, Raul Ortiz-Pulido, Maria A. Marcos-García, and Victor Rico-Gray. Unravelling Darwin's entangled bank: architecture and robustness of mutualistic networks with multiple interaction types. *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1843), 2016. ISSN 0962-8452.

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

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Jr. R. Stat. Soc. B*, 39:1–38, 1977.

Noémi Gaskó, Florentin Bota, Mihai Suciu, and Rodica Ioana Lung. Community structure detection in multipartite networks: A new fitness measure. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '17, pages 259–265, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4920-8.

G. Govaert and M. Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational. Statistics and Data Analysis*, 52(6):3233–3245, February 2008. ISSN 0167-9473.

Gérard Govaert and Mohamed Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463 – 473, 2003. ISSN 0031-3203. Biometrics.

Tommi S. Jaakkola. Tutorial on variational approximation methods. In *In Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press, 2000.

Christine Keribin, Vincent Brault, Gilles Celeux, and Gérard Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16, 2014. ISSN 0960-3174.

Sonia Kéfi, Vincent Miele, Evie A. Wieters, Sergio A. Navarrete, and Eric L. Berlow. How structured is the entangled bank? the surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience. *PLOS Biology*, 14(8):1–21, 08 2016.

E Lazega and T A B Snijders. *Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications*. Methodos Series. Springer International Publishing, 2015. ISBN 9783319245201.

Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2):715–742, 06 2010.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141, 2017.

Catherine Matias and Stéphane Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings and Surveys*, 47:55–74, 2014.

Marco Pautasso, Guntra Aistara, Adeline Barnaud, Sophie Caillon, Pascal Clouvel, Oliver T Coomes, Marc Delêtre, Elise Demeulenaere, Paola De Santis, Thomas Döring, et al. Seed exchange networks for agrobiodiversity conservation. a review. *Agronomy for sustainable development*, 33(1):151–175, 2013.

Shai Pilosof, Mason A Porter, Mercedes Pascual, and Sonia Kéfi. The Multilayer Nature of Ecological Networks. *ArXiv*, 2016.

Michael JO Pocock, Darren M Evans, and Jane Memmott. The robustness and restoration of a network of ecological networks. *Science*, 335(6071):973–977, 2012.

Valérie Robert. *Coclustering for the analysis of pharmacovigilance large datasets*. Theses, Université Paris Saclay ; Université Paris Sud - Orsay, June 2017.

Tom A. B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification*, 14(1):75–100, 1997. ISSN 0176-4268.

Mathieu Thomas and Sophie Caillon. Effects of farmer social status and plant biocultural value on seed circulation networks in Vanuatu. *Ecology and Society*, 21(2), 2016.

Mathieu Thomas, Nicolas Verzelen, Pierre Barbillon, Oliver T Coomes, Sophie Caillon, Doyle McKey, Marianne Elias, Eric Garine, Christine Raimond, Edmond Dounias, et al. A network-based method to detect patterns of local crop biodiversity: validation at the species and infra-species levels. In *Advances in Ecological Research*, volume 53, pages 259–320. Elsevier, 2015.

Jaewon Yang and Jure Leskovec. Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1170–1175. IEEE, 2012.