Rapport de stage

Détection et comparaison de structures de réseaux écologiques

Stage de M2 Mathématiques pour les Sciences du vivant

Unité de recherche: Mathématiques et informatique appliquées (AgroParisTech), UMR 518 INRAE
Référent: Camille Coron

par

Louis Lacoste

Encadrement du stage :

Pierre Barbillon — Co-encadrant
Sophie Donnet — Co-encadrante

Table des matières

Remerciements

Je tiens à remercier en premier lieu Sophie Donnet et Pierre Barbillon pour leur encadrement remarquable, leur disponibilité, conseils avisés et leur gentillesse. Saint-Clair Chabert-Liddell pour son accompagnement, ses remarques, ses explications et le temps qu'il m'a consacré. Merci à Julien pour sa sympathie et tout son travail en tant que DU.

Merci à Farida, Christelle et Sébastien pour avoir expliqué et mené les démarches administratives.

Un merci tout particulier à tous les doctorants : Mary, Marina, Emré, Tam, Caroline, Jérémy, Florian, Annaïg, Jules, Hayato, Jeanne, Tanguy, Barbara, Bastien et Armand. Merci à tous les autres stagiaires, particulièrement : Alizée, Taliesin, Antoine, Alexandre, Francois, Vincent, Pierre, Camille et Maxime.

Merci à tous les permanents du 3^{e} étage, parmi lesquels : Christophe, Stéphane et Vincent.

Merci à Liliane, Isabelle, Hugo, Théodore, Éric, Jean-Benoist, Nicolas, Lucia, Tristan, Sarah, Jade et Pierre Gloaguen.

Un grand merci à tous ceux qui ont participé de près ou de loin au bon déroulement de ce stage.

Chapter 1
Introduction

Networks are versatile objects able to represent various types of interactions and bipartite networks are particularly useful in ecological context for interaction between different entities (e.g. plant-pollinator). As the networks grow in size, reliable metrics, models and methods are needed to detect structure and perform analysis. Those methods exist and are pretty robust for single network analyses but we have motivation to consider a collection of network, in order to compare their structure or partition them. For collection of simple networks a colSBM (collection Stochastic Block Model Chabert-Liddell et al., 2024) has been proposed. We adapt this model to the bipartite case with a variational Expectation-Maximization algorithm for inference, a clever parameter space exploration and a BIC-like criterion for model selection. Building on this method we present a partitioning algorithm to gather networks based on their shared structures. We perform simulation studies to assess performance of our models and algorithm. Finally, we apply our clustering algorithm on ecological networks.

A common usage of bipartite graphs in ecology is the representation of ecological interactions like plant-pollinator (Ramos-Jiliberto et al., 2010), birds-seed dispersion, prey-predator or host-parasite (Kaszewska-Gilas et al., 2021). For plant-pollinator interactions, the rows are pollinator species and the columns are plant species, and the intersection is a value, binary if it is a presence/absence or a value if it is an abundance count.

Some interesting results can arise when applying a tool widely used on a particular kind of interactions is used on another kind of interactions. In Desjardins-Proulx et al., 2017 the authors use the *K-nearest neighbour* (KNN) algorithm as a Recommender to predict missing preys for predators in a predator-prey network.

Bipartite graphs are widely used in biology in general, in various fields, among which the previously cited ecological networks, but also in medicine with biomedical networks, biomolecular networks or epidemiological networks. (Pavlopoulos et al., 2018)

There is a need for comparison methods of bipartite networks in literature, and it is being actively developed, e.g. Pichon et al., 2024 use structures at multiple scales (species degree, motif frequency, nestedness ...) to tell apart mutualistic and antagonistic networks.

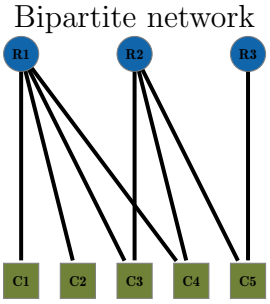
This motivates us to propose a model for structure detection in bipartite collections.

Chapter 2

Structure detection in a collection of bipartite networks

2.1 Formalism of bipartite graphs

Bipartite graphs, denoted as $G = (U, V, E)$ with U and V two disjoint and independent sets of vertices and E the set of edges connecting U vertices to V vertices.



$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Incidence matrix

X is the *incidence matrix* and is the mathematical object on which computations are performed. It is filled with the following rule:

$$\begin{cases} X_{ij} = 0 & \text{if no interaction is observed between species } i \text{ and } j \\ X_{ij} \neq 0 & \text{otherwise} \end{cases}$$

If the network represents binary observations (like presence-absence) then $X_{ij} \in \mathcal{K} = \{0, 1\}, \forall (i, j)$; if the interactions are weighted (like an abundance count), $X_{ij} \in \mathcal{K} = \mathbb{N}, \forall (i, j)$.

This representation can be used to represent various forms of interactions were two kinds of "actors" interact. Those interactions can be binary or valued and a numeric representation is the incidence matrix, in the above example X.

2.2 Latent Block Model

The Latent Block Model (LBM) introduced by Govaert and Nadif, 2010 adapts the Stochastic Block Model (SBM) (Holland et al., 1983; Snijders & Nowicki, 1997) to bipartite graphs.

Note : Please note that we prefer the term "BiSBM" and will use both LBM and BiSBM to designate the Stochastic Block Model adapted to bipartite networks.

This model supposes that:

- Row nodes are members of row blocks and column nodes are members of column blocks.

- The connectivity of two individuals is determined by their block memberships.
- An interaction can only occur between a row and a column node.

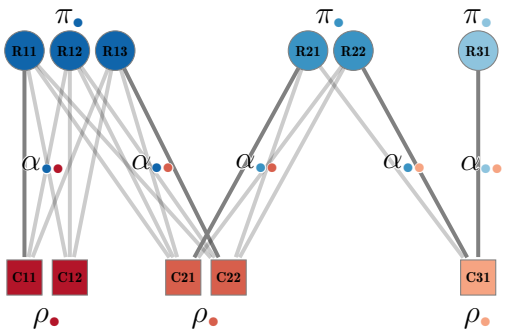


Figure 2.1: An LBM model visualization

- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ *given* blocks in rows
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ *given* blocks in columns

Parameters

- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$ for rows and $\rho_\bullet = \mathbb{P}(W_j = \bullet)$ for columns
- $\alpha_{\bullet\bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$, parameter influencing the probability and value of a link knowing node membership blocks.

On figure 2.1, $\boldsymbol{\pi}$ are the probabilities for a row node to belong to the row block of corresponding color, $\boldsymbol{\rho}$ are the probabilities for a column node to belong to the column block of corresponding color and $\boldsymbol{\alpha}$ is a matrix $Q_1 \times Q_2$ of the connectivity parameters between the row and column blocks. When we talk about the "structure" of the network we are referring to this connectivity matrix.

This model can be used to easily generate bipartite graphs with complex and very varied structures. But when trying to determine the structure of a given network we need to find those parameters and as the row and column block memberships are *latent* i.e., they are not known, they must be inferred.

For this a common approach is to use a *variational* EM algorithm, proposed for SBM in Daudin et al., 2008 and for LBM in Govaert and Nadif, 2005. The groups and required parameters can be inferred by maximizing a lower bound of the likelihood.

2.3 colSBM model, a joint model for a collection of networks

The *colSBM* model introduced by Chabert-Liddell et al., 2024 propose an extension of the SBM model to collections of simple (or unipartite) networks. A

collection is a set of networks which nodes are not in common nor linked between different networks and the interactions have the same valuations.

The model can retrieve the shared structure in a collection, indicate if networks should be grouped in a collection and in a large pool of networks, collections with common structures.

The next step after designing this collection model for unipartite networks was to extend it to the bipartite case.

2.4 Definition of a bipartite collection

We define a collection of bipartite networks as $\boldsymbol{X} = (X^1, \dots X^m, \dots, X^M)$ the collection of incidence matrix. Moreover, all the networks in the collection have the same valuation of the interactions (e.g., they are all binary).

2.5 Separate BiSBM (sep-BiSBM)

A first approach to deal with a collection of networks is to adjust separate BiSBM for each network of the collection.

For network m, let n_1^m (resp. n_2^m) be the number of nodes in row (resp. column) divided into Q_1^m row clusters (resp. Q_2^m column clusters).
Let $Z^m = (Z_1^m, \dots, Z_i^m, \dots, Z_{n_1^m}^m)$ and $W^m = (W_1^m, \dots, W_j^m, \dots, W_{n_2^m}^m)$ be independent latent variables such that $Z_i^m = q$ if row node i of network m belongs to row cluster q ($q \in \{1, \dots, Q_1^m\}$) and $W_j^m = r$ if column node j of network m belong to column block r ($r \in \{1, \dots, Q_2^m\}$). And we have

$$\mathbb{P}(Z_i^m = q) = \pi_q^m, \qquad \mathbb{P}(W_j^m = r) = \rho_r^m \tag{2.1}$$

where $\pi_q^m > 0$, $\rho_r^m > 0$, $\sum_{q=1}^{Q_1^m} \pi_q^m = 1$ and $\sum_{r=1}^{Q_2^m} \rho_r^m = 1$. Given the latent variables Z^m, W^m, the X_{ij}^ms are assumed to be independent and distributed as

$$X_{ij}^m | Z_i^m = q, W_j^m = r \sim \mathcal{F}(.; \alpha_{qr}^m) \tag{2.2}$$

where $\mathcal{F}$ is referred to as the emission distribution. $\mathcal{F}$ is chosen to be the Bernoulli distribution for binary interactions, and the Poisson distribution for weighted interactions such as counts. Let f be the density of the emission distribution, then:

$$\log f(X_{ij}^m; \alpha_{qr}^m) = \begin{cases} X_{ij}^m \log(\alpha_{qr}^m) + (1 - X_{ij}^m) \log(1 - \alpha_{qr}^m) & \text{for Bernoulli emission} \\ -\alpha_{qr}^m + X_{ij}^m \log(\alpha_{qr}^m) - \log(X_{ij}^m!) & \text{for Poisson emission} \end{cases} \tag{2.3}$$

Equations (2.1), (2.2) and (2.3) defines the BiSBM model and we will now use a short notation:

$$X^m \sim \mathcal{F}\text{-BiSBM}_{n_1^m, n_2^m}(Q_1^m, Q_2^m, \boldsymbol{\pi^m}, \boldsymbol{\rho^m}, \boldsymbol{\alpha^m}) \quad \forall m = 1, \dots M \quad (\textit{sep-BiSBM})$$

where $\mathcal{F}$ encodes the emission distribution, n_1^m, n_2^m are the number of row and column nodes, Q_1^m, Q_2^m are the number of row and column blocks in network m, $\boldsymbol{\pi}^m = (\pi_q^m)_{q=1,\dots,Q_1^m}$ and $\boldsymbol{\rho}^m = (\rho_r^m)_{r=1,\dots,Q_2^m}$ are the vectors of their proportions. The $Q_1^m \times Q_2^m$ matrix $\boldsymbol{\alpha}^m = (\alpha_{qr}^m)_{\substack{q=1,\dots,Q_1^m \\ r=1,\dots,Q_2^m}}$ are the connectivity parameters, i.e. the parameters of the emission distribution. $\alpha_{qr}^m \in \mathcal{A}_\mathcal{F}$ where, for the Bernoulli (resp. Poisson) emission distribution, $\mathcal{A}_\mathcal{F} = (0,1)$ (resp. $\mathcal{A}_\mathcal{F} = \mathbb{R}^{*+}$). In this *sep*-BiSBM model each network m is assumed to follow a BiSBM with its own parameters $(\boldsymbol{\pi}^m, \boldsymbol{\rho}^m, \boldsymbol{\alpha}^m)$.

2.6 Definition of the *colBiSBM* models

2.6.1 A collection of iid bipartite SBM

As for *colSBM* this first model is the most constrained. It assumes that all the networks are the independent realizations of the same Q_1-Q_2-BiSBM with identical parameters. The *iid*-colBiSBM is defined as follows:

$$X^m \sim \mathcal{F} - BiSBM_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha}), \quad \forall m = 1, \dots M \qquad (iid\text{-colBiSBM})$$

where $\forall (q,r) \in \{1,\dots,Q_1\} \times \{1,\dots,Q_2\}$, $\alpha_{qr} \in \mathcal{A}_\mathcal{F}$, $\pi_q \in (0,1], \sum_{q=1}^{Q_1} \pi_q = 1$ and $\rho_r \in (0,1], \sum_{r=1}^{Q_2} \rho_r = 1$. This model involves $(Q_1 - 1) + (Q_2 - 1) + Q_1 \times Q_2$ parameters, the two first terms corresponding to block proportions on the row and column and the third term to connectivity parameters.

But the assumption that block proportions are the same among the networks is a strong assumption. In plant-pollinator networks, the proportion of specialist species can differ between networks and thus the model may benefit from not having the same block proportions but sharing a common connectivity structure. The following models relaxes this assumption on either row, column or both.

2.6.2 A collection of bipartite SBM with varying block size on either rows or columns

π-colBiSBM model still assumes that the networks share a common connectivity structure represented by $\boldsymbol{\alpha}$ but that each network has its own row block proportions. For $m \in \{1,\dots,M\}$, the X^m are independent and

$$X^m \sim \mathcal{F} - BiSBM_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi^m}, \boldsymbol{\rho}, \boldsymbol{\alpha}), \quad \forall m = 1, \dots, M \qquad (\pi\text{-colBiSBM})$$

where $\forall (q,r) \in \{1,\dots,Q_1\} \times \{1,\dots,Q_2\}$, $\alpha_{qr} \in \mathcal{A}_\mathcal{F}$, $\pi_q^m \in [0,1], \sum_{q=1}^{Q_1} \pi_q^m = 1, \forall m \in \{1,\dots,M\}$ and $\rho_r \in (0,1], \sum_{r=1}^{Q_2} \rho_r = 1$. This model is more flexible than the iid-colBiSBM as it allows the row block proportions to vary between networks and even to be null ($\pi_q^m \in [0,1]$): if $\pi_q^m = 0$ then the block q is not represented in the network m. The connectivity structure is thus a subset of a large connectivity structure common to all networks. We face the same problems as Chabert-Liddell

et al., 2024 and adapt the support S they define for the π-colSBM to the bipartite case by having S^1 of size $M \times Q_1$ the support for the rows and S^2 of size $M \times Q_2$ the support for the columns. Thus $S^1_{mq} = \mathbb{1}_{\pi^m_q > 0}$ and $S^2_{mr} = \mathbb{1}_{\rho^m_r > 0}$. In this case, $S^2 = \mathbf{1}$, because there is no freedom on the column dimension.

For a given number of blocks Q_1, Q_2 and matrix S^1 (S^2 being in this case the matrix full of ones), the number of parameters is:

$$\text{NP}(\pi\text{-colBiSBM}) = \sum_{m=1}^{M} \left(\sum_{q=1}^{Q_1} S^1_{mq} - 1 \right) + (Q_2 - 1) + \sum_{\substack{q=1,\dots,Q_1 \\ r=1,\dots,Q_2}} \mathbb{1}_{(S^{1\prime} S^2)_{qr} > 0}$$

The first term corresponds to the non-null block proportions in each network. The third quantity accounts for the fact that some blocks may never be represented simultaneously in any network, so the corresponding connection parameters α_{qr} are not useful for defining the model.

ρ-colBiSBM model still assumes that the networks share a common connectivity structure represented by $\boldsymbol{\alpha}$ but that each network has its own column block proportions. For $m \in \{1, \dots, M\}$, the X^m are independent and

$$X^m \sim \mathcal{F} - BiSBM_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi}, \boldsymbol{\rho^m}, \boldsymbol{\alpha}), \quad \forall m = 1, \dots, M \quad (\rho\text{-colBiSBM})$$

where $\forall (q, r) \in \{1, \dots, Q_1\} \times \{1, \dots, Q_2\}$, $\alpha_{qr} \in \mathcal{A}_{\mathcal{F}}$, $\pi_q \in (0, 1]$, $\sum_{q=1}^{Q_1} \pi_q = 1$ and $\rho^m_r \in [0, 1]$, $\sum_{r=1}^{Q_2} \rho^m_r = 1$. This model is more flexible than the iid-colBiSBM as it allows proportions to vary between networks and even to be null ($\rho^m_r \in [0, 1]$): if $\rho^m_r = 0$ then the column block r is not represented in the network m.

"Mirroring" the formulas for the π-colBiSBM we relax the constraints on the column dimension.

For a given number of blocks Q_1, Q_2 and matrix S^2 (S^1 being in this case the matrix full of ones), the number of parameters is:

$$\text{NP}(\rho\text{-colBiSBM}) = (Q_1 - 1) + \sum_{m=1}^{M} \left(\sum_{r=1}^{Q_2} S^2_{mr} - 1 \right) + \sum_{\substack{q=1,\dots,Q_1 \\ r=1,\dots,Q_2}} \mathbb{1}_{(S^{1\prime} S^2)_{qr} > 0}$$

$\pi\rho$-colBiSBM model still assumes that the networks share a common connectivity structure represented by $\boldsymbol{\alpha}$ but that each network has its own row and column block proportions, it is the least constrained model. For $m \in \{1, \dots, M\}$, the X^m are independent and

$$X^m \sim \mathcal{F} - BiSBM_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi^m}, \boldsymbol{\rho^m}, \boldsymbol{\alpha}), \quad \forall m = 1, \dots, M \quad (\pi\rho\text{-colBiSBM})$$

where $\forall (q, r) \in \{1, \dots, Q_1\} \times \{1, \dots, Q_2\}$, $\alpha_{qr} \in \mathcal{A}_{\mathcal{F}}$, $\pi^m_q \in [0, 1]$, $\sum_{q=1}^{Q_1} \pi^m_q = 1, \forall m \in \{1, \dots, M\}$ and $\rho^m_r \in [0, 1]$, $\sum_{r=1}^{Q_2} \rho^m_r = 1$.

For a given number of blocks Q_1, Q_2 and matrices S^1, S^2, the number of parameters is:

$$\text{NP}(\pi\rho\text{-colBiSBM}) = \sum_{m=1}^{M} \left(\sum_{q=1}^{Q_1} S^1_{mq} - 1 \right) + \sum_{m=1}^{M} \left(\sum_{r=1}^{Q_2} S^2_{mr} - 1 \right) + \sum_{\substack{q=1,\dots,Q_1 \\ r=1,\dots,Q_2}} \mathbb{1}_{(S^{1\prime} S^2)_{qr} > 0}$$

2.7 Variational estimation of the parameters

In practice, the estimation of the likelihood is not tractable. Following the classical approach defined in Daudin et al., 2008 we use a variational version of the Expectation Maximization (VEM) algorithm.

We maximize a variational lower bound of the log-likelihood of the observed data, the so-called Evidence Lower Bound (or ELBO), by approximating $p(\boldsymbol{Z}, \boldsymbol{W}|\boldsymbol{X}; \boldsymbol{\theta})$ with a distribution on $\boldsymbol{Z}$ and $\boldsymbol{W}$ named $\mathcal{R}$ defined as $\mathcal{R} = \otimes_{m=1}^{M} \mathcal{R}_m$.

The lower bound is defined as:

$$\mathcal{J}(\mathcal{R}; \boldsymbol{\theta}) \coloneqq \sum_{m=1}^{M} \left(\mathbb{E}_{\mathcal{R}_m}[\ell(X^m, Z^m, W^m; \boldsymbol{\theta})] + \mathcal{H}(\mathcal{R}_m) \right) \leq \ell(\boldsymbol{X}; \boldsymbol{\theta})$$

$(Z_i^m)_{i=1\ldots n_1^m}$ and $(W_j^m)_{j=1\ldots n_2^m}$ are redefined using the *one-hot encoded* conversion (i.e., $Z_i^m = q \rightarrow Z_{iq}^m = 1$ and $W_j^m = r \rightarrow W_{jr}^m = 1$).

When $\mathcal{R}_m$ is issued from the set of the factorizable distributions, we denote $\tau_{iq}^{1,m} = \mathbb{P}_{\mathcal{R}_m}(Z_{iq}^m = 1|X_{i\bullet}^m)$ and $\tau_{jr}^{2,m} = \mathbb{P}_{\mathcal{R}_m}(W_{jr}^m = 1|X_{\bullet j}^m)$, thus we have: $\mathbb{P}_{\mathcal{R}_m}(Z_{iq}^m = 1, W_{jr}^m = 1|X^m) = \mathbb{P}_{\mathcal{R}_m}(Z_{iq}^m = 1|X_{i\bullet}^m) \times \mathbb{P}_{\mathcal{R}_m}(W_{jr}^m = 1|X_{\bullet j}^m) = \tau_{iq}^{1,m} \times \tau_{jr}^{2,m}$.

The formula for the entropy per network is thus:

$$\mathcal{H}(\mathcal{R}_m) = -\sum_{i=1}^{n_1^m} \tau_{iq}^{1,m} \log \tau_{iq}^{1,m} - \sum_{j=1}^{n_2^m} \tau_{jr}^{2,m} \log \tau_{jr}^{2,m}$$

And the expectation of the completed log-likelihood under the $\mathcal{R}_m$ variational distribution for network m is:

$$\begin{aligned}\mathbb{E}_{\mathcal{R}_m}[\ell(X^m, Z^m, W^m; \boldsymbol{\theta})] = &\sum_{i=1}^{n_1^m} \sum_{j=1}^{n_2^m} \sum_{q \in \mathcal{Q}_1^m} \sum_{r \in \mathcal{Q}_2^m} \tau_{iq}^{1,m} \tau_{jr}^{2,m} \log f(X_{ij}^m; \alpha_{qr}) \\ &+ \sum_{i=1}^{n_1^m} \sum_{q \in \mathcal{Q}_1^m} \tau_{iq}^{1,m} \log {\pi_q}^m + \sum_{j=1}^{n_2^m} \sum_{r \in \mathcal{Q}_2^m} \tau_{jr}^{2,m} \log {\rho_r}^m\end{aligned}$$

with $\mathcal{Q}_1^m = \{q \in \{1 \ldots, Q_1\} | \pi_q^m > 0\}$ and $\mathcal{Q}_2^m = \{r \in \{1 \ldots, Q_2\} | \rho_r^m > 0\}$ And thus the lower bound becomes:

$$\begin{aligned}\mathcal{J}(\boldsymbol{\tau}; \boldsymbol{\theta}) \coloneqq \sum_{m=1}^{M} \Bigg(&\sum_{i=1}^{n_1^m} \sum_{j=1}^{n_2^m} \sum_{q \in \mathcal{Q}_1^m} \sum_{r \in \mathcal{Q}_2^m} \tau_{iq}^{1,m} \tau_{jr}^{2,m} \log f(X_{ij}^m; \alpha_{qr}) \\ &+ \sum_{i=1}^{n_1^m} \sum_{q \in \mathcal{Q}_1^m} \tau_{iq}^{1,m} \log {\pi_q}^m + \sum_{j=1}^{n_2^m} \sum_{r \in \mathcal{Q}_2^m} \tau_{jr}^{2,m} \log {\rho_r}^m \\ &- \sum_{i=1}^{n_1^m} \tau_{iq}^{1,m} \log \tau_{iq}^{1,m} - \sum_{j=1}^{n_2^m} \tau_{jr}^{2,m} \log \tau_{jr}^{2,m} \Bigg)\end{aligned}$$

where we identify the variational distribution $\mathcal{R}$ with its parameter $\boldsymbol{\tau}$.

The VEM algorithm alternates between two steps, the variational E step and the M step. The E steps consists in optimizing $\mathcal{J}(\boldsymbol{\tau};\boldsymbol{\theta})$ for a current value of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\tau}$. And the M step consists of maximizing $\mathcal{J}(\boldsymbol{\tau};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and for a given variational distribution $\boldsymbol{\tau}$.

2.7.1 Variational E step

At this step we maximize with respect to the variational distribution $\boldsymbol{\tau}$:

$$\widehat{\boldsymbol{\tau}}^{(t+1)} = \arg\max_{\boldsymbol{\tau}} \mathcal{J}(\boldsymbol{\tau}, \widehat{\boldsymbol{\theta}}^{(t)}).$$

And we obtain the following formulae for the $\boldsymbol{\tau^m}$:

$$\begin{cases} \widehat{\tau}_{iq}^{1,m} \propto \widehat{\pi}_q^{m(t)} \prod_{j=1}^{n_2^m} \prod_{r\in\mathcal{Q}_2^m} f(X_{ij}^m; \widehat{\alpha}_{qr}^{(t)})^{\widehat{\tau}_{jr}^{2,m(t+1)}} & \forall i = 1,\dots,n_1^m, q \in \mathcal{Q}_1^m \\ \widehat{\tau}_{jr}^{2,m} \propto \widehat{\rho}_r^{m(t)} \prod_{i=1}^{n_1^m} \prod_{q\in\mathcal{Q}_1^m} f(X_{ij}^m; \widehat{\alpha}_{qr}^{(t)})^{\widehat{\tau}_{iq}^{1,m(t+1)}} & \forall j = 1,\dots,n_2^m, r \in \mathcal{Q}_2^m \end{cases}$$

which are used to update iteratively the values by a fixed point algorithm with only one step.

2.7.2 M step of the algorithm

At iteration (t) the M-step maximizes the variational bound with respect to the model parameters $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{J}(\widehat{\boldsymbol{\tau}}^{(t+1)}, \boldsymbol{\theta})$$

The following quantities are involved in the obtained formulae:

$$e_{qr}^m = \sum_{i=1}^{n_1^m}\sum_{j=1}^{n_2^m} \tau_{iq}^{1,m}\tau_{jr}^{2,m} X_{ij}^m, \quad n_{qr}^m = \sum_{i=1}^{n_1^m}\sum_{j=1}^{n_2^m} \tau_{iq}^{1,m}\tau_{jr}^{2,m}, \quad n_q^{1,m} = \sum_{i=1}^{n_1^m} \tau_{iq}^{1,m}, \quad n_r^{2,m} = \sum_{j=1}^{n_2^m} \tau_{jr}^{2,m}$$

The block proportions, in free mixture models, $(\pi_q^m)_{q\in\mathcal{Q}_1^m}, (\rho_r^m)_{r\in\mathcal{Q}_2^m}$ are estimated as

$$\widehat{\pi}_q^m = \frac{n_q^{1,m}}{n_1^m} \qquad \text{for } \pi\text{-colBiSBM and } \pi\rho\text{-colBiSBM}$$

$$\widehat{\rho}_r^m = \frac{n_r^{2,m}}{n_2^m} \qquad \text{for } \rho\text{-colBiSBM and } \pi\rho\text{-colBiSBM}$$

while on the other hand,

$$\widehat{\pi}_q = \frac{\sum_{m=1}^M n_q^{1,m}}{\sum_{m=1}^M n_1^m} \qquad \text{for } iid\text{-colBiSBM and } \rho\text{-colBiSBM}$$

$$\widehat{\rho}_r = \frac{\sum_{m=1}^M n_r^{2,m}}{\sum_{m=1}^M n_2^m} \qquad \text{for } iid\text{-colBiSBM and } \pi\text{-colBiSBM}$$

the parameters take into account all the networks at the same time. The connectivity parameters α_{qr} for all models are estimated as the ratio of the number of observed interactions between row block q and column block r among all networks over the number of possible interactions:

$$\widehat{\alpha}_{qr} = \frac{\sum_{m=1}^{M} e_{qr}^{m}}{\sum_{m=1}^{M} n_{qr}^{m}}$$

Please note that those formulae can vary with the emission distribution used.

2.8 Model selection

The section 2.7 explains how we estimate the parameters of the model for *fixed* number of blocks Q_1 and Q_2. But as they are in general not known we need to explore the latent space to find the *best* values. As discussed in Chabert-Liddell et al., 2024, the algorithmic aspect becomes complex when dealing with the bipartite case. Due to the latent space being $\mathbb{N}^2$, conducting a complete exploration of the latent space is practically infeasible. Therefore, in addition to adapting the existing formulas, our contribution to addressing this challenge involved making significant choices, which are outlined below.

The below procedures are implemented in the *colSBM* package, available on https://github.com/Chabert-Liddell/colSBM.

2.8.1 The *Bayesian Information Criterion like* (BIC-L) criterion for model selection

To select the best number of blocks we need a criterion to measure adequacy between our model and data. The ELBO might seem a good criterion at first but as for the likelihood, the more complex the model, the higher it gets. And thus a good criterion should make a *trade-off* between fitting to data and model complexity.

The Integrated Classified Likelihood (ICL) is a well-established tool in the SBM and LBM domains for selecting the appropriate number of blocks. It was introduced by Daudin et al., 2008; Biernacki et al., 2000. The ICL is derived from an asymptotic approximation of the marginal complete likelihood. In this approach, the model parameters are integrated out using a prior distribution, resulting in a penalized likelihood criterion. By employing the ICL, one can effectively determine the optimal number of blocks for the given problem in a systematic manner. We obtain the following expression

$$\text{ICL} = \max_{\theta} \mathbb{E}_{\widehat{\mathcal{R}}}[\ell(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}; \boldsymbol{\theta})] - \frac{1}{2}\text{pen}$$

with pen the penalties.
Using the formula $\mathbb{E}_{\widehat{\mathcal{R}}}[\ell(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}; \boldsymbol{\theta})] \approx \ell(\boldsymbol{X}; \boldsymbol{\theta}) - \mathcal{H}(\widehat{\mathcal{R}})$, it becomes clearer, as highlighted in the existing literature, that the Integrated Classified Likelihood

(ICL) gives preference to well-separated blocks by imposing a penalty on the entropy of node grouping. However, the objective of our study extends beyond grouping nodes into coherent blocks. We also aim to assess the similarity of connectivity patterns across different networks. Consequently, we aim to permit models that offer more flexible node grouping by not penalizing on entropy.

This leads us to formulate a BIC-like criterion in the following manner:

$$\text{BIC-L} = \max_{\boldsymbol{\theta}} \mathbb{E}_{\widehat{\mathcal{R}}}[\ell(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}; \boldsymbol{\theta})] + \mathcal{H}(\widehat{\mathcal{R}}) - \frac{1}{2}\text{pen} = \max_{\boldsymbol{\theta}} \mathcal{J}(\widehat{\mathcal{R}}, \boldsymbol{\theta}) - \frac{1}{2}\text{pen}$$

We provide below the expression for the penalties for the 4 models that we propose.

***iid*-colBiSBM** For the $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$:

$$\text{pen}_{\pi}(Q_1) = (Q_1 - 1)\log(\sum_{m=1}^{M} n_1^m), \quad \text{pen}_{\rho}(Q_2) = (Q_2 - 1)\log(\sum_{m=1}^{M} n_2^m)$$

For the $\boldsymbol{\alpha}$:

$$\text{pen}_{\alpha}(Q_1, Q_2) = Q_1 \times Q_2 \log(N_M)$$

with

$$N_M = \sum_{m=1}^{M} n_1^m \times n_2^m$$

And thus the BIC-L formula is the following:

$$\text{BIC-L}(\boldsymbol{X}, Q_1, Q_2) = \max_{\theta} \mathcal{J}(\hat{\mathcal{R}}, \boldsymbol{\theta}) - \frac{1}{2}[\text{pen}_{\pi}(Q_1) + \text{pen}_{\rho}(Q_2) + \text{pen}_{\alpha}(Q_1, Q_2)]$$

$\boldsymbol{\pi\rho}$-colBiSBM The support penalties are

$$\text{pen}_{S_1}(Q_1) = -2\log p_{Q_1}(S_1), \qquad \text{pen}_{S_2}(Q_2) = -2\log p_{Q_2}(S_2)$$

with

$$\log p_{Q_1}(S_1) = -M\log(Q_1) - \textstyle\sum_{m=1}^{M} \log\binom{Q_1}{Q_1^{(m)}},$$
$$\log p_{Q_2}(S_2) = -M\log(Q_2) - \textstyle\sum_{m=1}^{M} \log\binom{Q_2}{Q_2^{(m)}}.$$

And penalties for the $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$ are

$$\text{pen}_{\pi}(Q_1, S_1) = \sum_{m=1}^{M} (Q_1^{(m)} - 1)\log n_1^m, \ \text{pen}_{\rho}(Q_2, S_2) = \sum_{m=1}^{M} (Q_2^{(m)} - 1)\log n_2^m.$$

Penalties for the $\boldsymbol{\alpha}$

$$\text{pen}_{\alpha}(Q_1, Q_2, S_1, S_2) = (\sum_{q=1}^{Q_1}\sum_{r=1}^{Q_2} \mathbb{1}_{(S_1)'S_2>0})\log(N_M).$$

And the corresponding BIC-L formula,

$$\begin{aligned}\text{BIC-L}(\boldsymbol{X}, Q_1, Q_2) = \max_{S_1,S_2}[& \max_{\theta_{S_1,S_2}\in\Theta_{S_1,S_2}} \mathcal{J}(\hat{\mathcal{R}}, \theta_{S_1,S_2}) \\ & -\frac{1}{2}(\text{pen}_\pi(Q_1, S_1) + \text{pen}_\rho(Q_2, S_2) \\ & \quad + \text{pen}_\alpha(Q_1, Q_2, S_1, S_2) \\ & \quad + \text{pen}_{S_1}(Q_1) + \text{pen}_{S_2}(Q_2))]\end{aligned}$$

2.8.2 Initialization and pairing of the models

The row (resp. column) block memberships are the labels of row (resp. column) nodes corresponding to the group to which they were assigned based on their connection patterns. This adds another layer of complexity to the model selection as we need to find the best Q_1, Q_2 and the best memberships for each vertex.

First to combine the information from the M networks we fit a LBM model for each network at the two points $Q = (1, 2)$ and $Q = (2, 1)$. Using the previously described VEM algorithm we obtain for each network its parameters $(\boldsymbol{\rho}, \boldsymbol{\pi}, \boldsymbol{\alpha})$. We then compute the marginal laws for each dimension, for each network. Then we order the network blocks by the probabilities obtained in decreasing order.

For the memberships on the columns: $col\ order_m = order\,(\pi_m \times \alpha_m)$.

For the memberships on the rows: $row\ order_m = order\,(\rho_m \times\ {}^t(\alpha_m))$.

Using this order we relabel the memberships for the M fitted collection of a single network. We then use the M memberships to compute first $\boldsymbol{\tau}$ to fit a collection containing the M networks.

2.8.3 Greedy exploration to find an estimation of the mode

Using the previously fitted models for $Q = (1, 2)$ and $Q = (2, 1)$ we choose to perform a greedy exploration from each of those points to find a first mode.

Meaning that for a given $Q = (Q_1, Q_2)$ we will compute all the possible memberships for the points $Q \in \{(Q_1 + 1, Q_2), (Q_1, Q_2 + 1), (Q_1 - 1, Q_2), (Q_1, Q_2 - 1)\}$, fit the corresponding models and choose the one that maximizes the BIC-L as the next point from which to repeat the procedure. We repeat the procedure until the BIC-L stops increasing 2 times in a row.

Let us denote the neighborhood in the latent space of a point Q by $\mathcal{N}(Q) = Q + (1, 0), (0, 1), (-1, 0), (0, -1)$, the four neighbors of Q in the grid.

Algorithm 1: Greedy Exploration for Mode Estimation

Input : Fitted models for $Q = (1, 2)$ and $Q = (2, 1)$
Output: Estimation of the mode using greedy exploration

for $Q_{start} \in \{(1, 2), (2, 1)\}$ **do**
  Initialize BIC-$L_{max} \leftarrow$ BIC-L(Q_{start})
  Initialize *consecutive_count* as 0
  $Q_{curr} \leftarrow Q_{start}$
  while *consecutive_count* < 2 **do**
    Fit models in $\mathcal{N}(Q_{curr})$;
    $Q \leftarrow \arg\max_{Q \in \mathcal{N}(Q_{curr})}$ BIC-L(Q)
    BIC-$L_{curr} \leftarrow \max_{Q \in \mathcal{N}(Q_{curr})}$ BIC-L(Q)
    if *BIC-*$L_{curr} >$ *BIC-*L_{max} **then**
      BIC-$L_{max} \leftarrow$ BIC-L_{curr}
      consecutive_count $\leftarrow 0$
    end
    else
      consecutive_count $\leftarrow$ *consecutive_count* $+ 1$
    end
  end
end

Output: Estimation of the mode using greedy exploration

When this first estimation of the BIC-L mode has been find we apply the moving window on it.

2.8.4 Moving window to update the block memberships and the BIC-L

The *moving window* is used to update the block memberships on rows and columns and fit new models with those changes. To define the window, we use a center point and a *depth*, giving us the bottom left corner $(Q_{1,center} - depth, Q_{2,center} - depth)$ and the top right corner of the window $(Q_{1,center} + depth, Q_{2,center} + depth)$. All the points in this square will be updated and contribute to the update of the others. This procedure is repeated until convergence of the BIC-L.

The figure 2.2 illustrates the procedure. It consists of two alternating steps:

- the *forward pass*: repeatedly computing the possible splits to fit the current model.
- the *backward pass*: computing the possible merges to fit the current model.

Forward pass The forward pass consists for a model at (Q_1, Q_2) to compute the possible splits from the block memberships of its "predecessors". The predecessors are the point at the left $(Q_1 - 1, Q_2)$ and below $(Q_1, Q_2 - 1)$ the current model (if they exist). To update the current model, we take its predecessors block memberships and try to split one of the blocks in two. Then the current model is fitted

Algorithm 2: Moving Window Procedure

Input : Center point $(Q_{1,\text{center}}, Q_{2,\text{center}})$, depth
Output: Best model with maximum BIC-L in the window

Define bottom left corner $(Q_{1,\text{center}} - \text{depth}, Q_{2,\text{center}} - \text{depth})$
Define top right corner $(Q_{1,\text{center}} + \text{depth}, Q_{2,\text{center}} + \text{depth})$

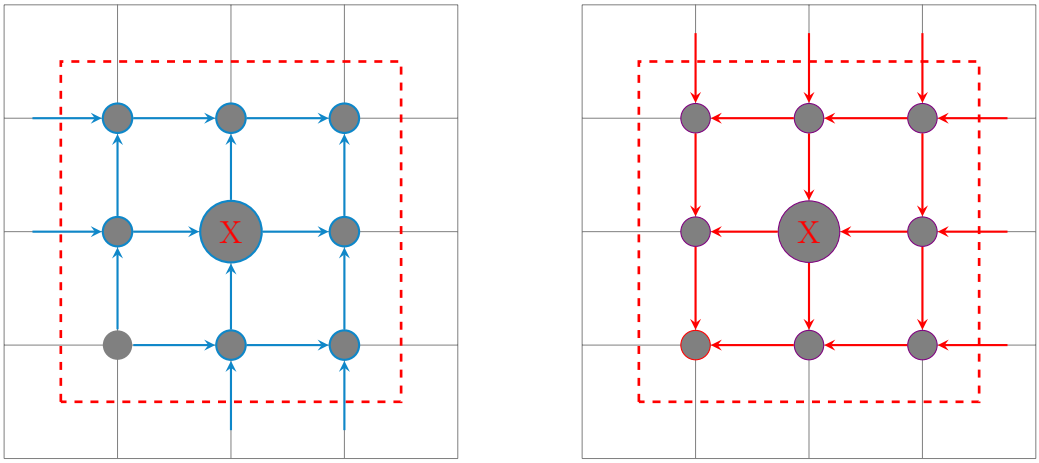
while *not converged* **do**
  Forward pass:
  for $Q_1 \in [Q_{1,center} - depth; Q_{1,center} + depth]$ **do**
    for $Q_2 \in [Q_{2,center} - depth; Q_{2,center} + depth]$ **do**
      Compute possible splits from predecessors $(Q_1 - 1, Q_2)$ and $(Q_1, Q_2 - 1)$
      Among the model generated from the splits choose the best in regard of the BIC-L
    end
  end
  Backward pass:
  for $Q_1 \in [Q_{1,center} + depth; Q_{1,center} - depth]$ **do**
    for $Q_2 \in [Q_{2,center} + depth; Q_{2,center} - depth]$ **do**
      Compute possible merges from predecessors $(Q_1 + 1, Q_2)$ and $(Q_1, Q_2 + 1)$
      Among the model generated from the merges choose the best in regard of the BIC-L
    end
  end
  Choose the mode as the one that maximizes the BIC-L
end

Output: Best model with maximum BIC-L in the window

using this clustering as a starting clustering. Once all the possible splits are fitted, they are compared, keeping the best, in the sense of the BIC-L. If a model was already present it is also compared and the best is chosen as the model for this round at (Q_1, Q_2).
The procedure then repeats for the point at (Q_1+1, Q_2) until it reaches $(Q_{1,center} + depth, Q_2)$ from which it repeats from $(Q_{1,center} - depth, Q_2 + 1)$. This repeats until computing the best model for $(Q_{1,center} + depth, Q_{2,center} + depth)$.

Note on the initialization: The forward pass starts from the point $(Q_{1,center} + depth, Q_{2,center} + depth)$, so this points needs to have at least a model fitted. In the best case, the greedy exploration will have visited this point. But if the point has not been visited, a model will be fitted from a spectral initialization (i.e the block memberships is computed by using a spectral clustering). From this point, the next model will have at least one predecessor and the procedure can iterate.

Backward pass The backward pass consists for a model at (Q_1, Q_2) to compute the possible merges from the block memberships of its "predecessors". The predecessors are the point at the right $(Q_1 + 1, Q_2)$ and on top $(Q_1, Q_2 + 1)$ of the



(a) Visualisation of a forward pass of moving window

(b) Visualisation of a backward pass of moving window

Figure 2.2: Moving window procedure, the center node marked with an X is the mode of BIC-L

current model (if the predecessors exist). To update the current model, we take its predecessors block memberships and try to merge two blocks in one. Then the current model is fitted using this clustering as a starting clustering. Once all the possible merges are fitted, they are compared, keeping the best, in the sense of the BIC-L. If a model was already present it is also compared and the best is chosen as the model for this round at (Q_1, Q_2).
The procedure then repeats for the point at $(Q_1 - 1, Q_2)$ until it reaches $(Q_{1,center} - depth, Q_2)$ from which it repeats from $(Q_{1,center} - depth, Q_2 - 1)$. This repeats until computing the best model for $(Q_{1,center} - depth, Q_{2,center} - depth)$. *Note on the initialization:* The backward pass starts from $(Q_{1,center} + depth, Q_{2,center} + depth)$, we know it was initialized at least by the forward pass, no special case here.

At the end of the moving window pass, the model of max BIC-L is the new best fit and the procedure repeats until convergence.

2.9 Networks clustering

As in Chabert-Liddell et al., 2024 we use a recursive algorithm to determine the best clustering of the given networks. The procedure being the same, we will present it briefly and focus on adjustments.

When networks in a collection do not share the same mesoscale connectivity structure we want to be able to partition them correctly. For this we perform a clustering of networks.

The process of clustering a collection of networks involves discovering a partition $\mathcal{G} = (\mathcal{M}_g)_{g=1,\dots,G}$ of $\{1, \dots, M\}$. Given $\mathcal{G}$ we set the following model on $\boldsymbol{X}$:

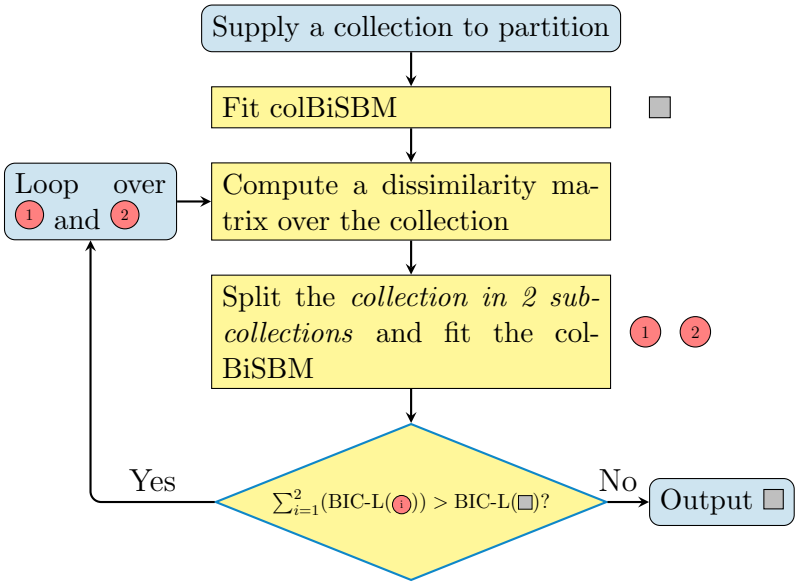


Figure 2.3: Network clustering procedure

$$\forall g \in \{1, \dots, G\},\ \forall m \in \mathcal{M}_g,\ X^m \sim \mathcal{F}\text{-}BiSBM(Q_1^g, Q_2^g, \boldsymbol{\pi^m}, \boldsymbol{\rho^m}, \boldsymbol{\alpha}^g)$$

And we defined the score of a given partition $\mathcal{G}$:

$$Sc(\mathcal{G}) = \sum_{g=1}^{G} \max_{Q^g=1,\dots,Q_{\max}} \text{BIC-L}((X^m)_{m\in\mathcal{M}_g}, Q_1^g, Q_2^g)$$

Thus the score consists of the sum of the BIC-L of the sub-collections for the partition $\mathcal{G}$.

2.9.1 Dissimilarity between two networks

The parameters for the dissimilarity are defined as follow:

$$\widetilde{n}_{qr}^m = \sum_{i=1}^{n_1^m}\sum_{j=1}^{n_2^m} \widehat{\tau}_{iq}^{1,m}\widehat{\tau}_{jr}^{2,m}, \qquad \widetilde{\alpha}_{qr}^m = \frac{\sum_{i=1}^{n_1^m}\sum_{j=1}^{n_2^m} \widehat{\tau}_{iq}^{1,m}\widehat{\tau}_{jr}^{2,m} X_{ij}^m}{\widetilde{n}_{qr}^m},$$

$$\widetilde{\pi}_q^m = \frac{\sum_{i=1}^{n_1^m} \widehat{\tau}_{iq}^{1,m}}{n_1^m}, \qquad \widetilde{\rho}_r^m = \frac{\sum_{j=1}^{n_2^m} \widehat{\tau}_{jr}^{2,m}}{n_2^m}.$$

And the pairwise dissimilarity for networks $(m, m') \in \mathcal{M}^2$ is then:

$$D_{\mathcal{M}}(m, m') = \sum_{q=1}^{Q_1}\sum_{r=1}^{Q_2} \max(\widetilde{\pi}_q^m, \widetilde{\pi}_q^{m'}) \left(\widetilde{\alpha}_{qr}^m - \widetilde{\alpha}_{qr}^{m'}\right)^2 \max(\widetilde{\rho}_r^m, \widetilde{\rho}_r^{m'})$$

The above figure, 2.3, shows a condensed explanation of the network clustering algorithm.

The idea is to adjust the colBiSBM model over the full collection of M networks and then compute the dissimilarity matrix between all networks of the collection. We obtain the collection $\mathcal{G} = \{\mathcal{M}\}$ the trivial partition in a unique group.

Then using the *Kmeans* we split the collection in two sub-collections with the dissimilarity matrix. The two sub-collections are fitted and we compute the score of this new partition $\mathcal{G}^* = \{G_1, G_2\}$. If $Sc(\mathcal{G}^*) > Sc(\mathcal{G})$, we repeat the same procedure on G_1 and G_2. Else we return $\mathcal{G}$. We illustrate our capacity to perform a partition of a collection for all colBiSBM models in 3.4.

2.10 Model identifiability

The goal here is to prove that if $\ell(\boldsymbol{X}; \boldsymbol{\theta}) = \ell(\boldsymbol{X}; \boldsymbol{\theta}')$ for any collection $\boldsymbol{X}$ then $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

Following the proof proposed by Chabert-Liddell et al., 2024, that adapted it to the collection case and Keribin et al., 2015 that extended the result of Celisse et al., 2012 to the LBM Bernoulli model, we obtain the following result of identifiability[1] for the *iid*-colBiSBM:

Properties 1 (Identifiability of *iid*-colBiSBM)**.** *The parameters* $(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ *are identifiable up to a label switching of the blocks if those conditions are achieved:*

(1.1) $\exists m^* \in \{1, \dots, M\} : n^1_{m^*} \geq 2Q_2 - 1$ *and* $n^2_{m^*} \geq 2Q_1 - 1$.

(1.2) $\forall 1 \leq q \leq Q_1, \pi_q > 0$ *and the coordinates of vector* $\boldsymbol{\rho} X^{m^*T}$ *are distinct (where* X^{m^*T} *is the transpose of* X^{m^*}*).*

(1.3) $\forall 1 \leq r \leq Q_2, \rho_r > 0$ *and the coordinates of vector* $\boldsymbol{\pi} X^{m^*}$ *are distinct.*

[1] The proof is in appendix. A.1

Chapter 3
Simulation studies

The simulations below are meant to test the capacities of our models. We assess the inference capacities of the algorithm and method, the model selection performances and the clustering capacities.

Reproducibility All the codes used to obtain data and to perform the analysis can be found on the report repository at https://gitea.polarolouis.fr/polarolouis/mia-stage-2024.

3.1 Efficiency of the inference

The goal here is to assess the quality of the inference procedure.

Simulation settings For this simulation the data is simulated with $M = 2, n_1^m = 120,\ n_2^m = 120,\ Q_1 = Q_2 = 4$, $\boldsymbol{\alpha}, \boldsymbol{\pi}$ and $\boldsymbol{\rho}$ are set as follows:

$$\boldsymbol{\alpha} = .25 + \begin{pmatrix} 3\epsilon_\alpha & 2\epsilon_\alpha & \epsilon_\alpha & -\epsilon_\alpha \\ 2\epsilon_\alpha & 2\epsilon_\alpha & -\epsilon_\alpha & \epsilon_\alpha \\ \epsilon_\alpha & -\epsilon_\alpha & \epsilon_\alpha & 2\epsilon_\alpha \\ -\epsilon_\alpha & \epsilon_\alpha & 2\epsilon_\alpha & 0 \end{pmatrix},$$

$$\boldsymbol{\pi}^1 = \sigma_1 \begin{pmatrix} 0.2 & 0.4 & 0.4 & 0 \end{pmatrix}, \qquad \boldsymbol{\pi}^2 = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix},$$
$$\boldsymbol{\rho}^1 = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}, \qquad \boldsymbol{\rho}^2 = \sigma_2 \begin{pmatrix} 0 & 0.33 & 0.33 & 0.33 \end{pmatrix},$$

with ϵ_α taking nine equally spaced values ranging from 0 to 0.24. For each value of ϵ_α, 108 datasets (X_1, X_2) are simulated, resulting in $9 \times 108 = 972$ datasets. More precisely, for each dataset, we pick uniformly at random two permutations of $\{1, \dots, 4\}$ (σ_1, σ_2) with the constraint that $\sigma_1(4) \neq \sigma_2(1)$. This ensures that each of the two networks have a non-empty block that is empty in the other one. Then the networks are simulated with $\mathcal{B}\text{ern-}BiSBM_{120,120}(Q_1 = 4, Q_2 = 4, \boldsymbol{\alpha}, \boldsymbol{\pi}^m, \boldsymbol{\rho}^m)$ with the previous parameters. Each network has 2 blocks in common and their connectivity structures encompass a mix of core-periphery, assortative community and dis-assortative community structures, depending on which 3 of the 4 blocks are selected for each network. ϵ_α represents the strength of these structures, the larger, the easier it is to tell apart one block from another. The true model of all the simulation is a $\pi\rho$-colBiSBM.

Inference We want to measure the quality of the inference procedure, for this we use the inference described in the section 2.7.

Quality indicators To assess the quality of the inference, we will use the following indicators:

- First, for each dataset, we put in competition π-colBiSBM with *sep-BiSBM*, *iid*-colBiSBM, ρ-colBiSBM, $\pi\rho$-colBiSBM respectively. To do so, for each dataset, we compute the BIC-L of each model π-colBiSBM is preferred to *sep-BiSBM* (resp. *iid*-colBiSBM, ρ-colBiSBM, $\pi\rho$-colBiSBM) if its BIC-L is greater.

- When considering our colBiSBM models we compare $\widehat{Q_1}$, $\widehat{Q_2}$ to their true values. ($Q_1 = 4$ and $Q_2 = 4$)

- Finally, we assess the quality of the node grouping by computing the Adjusted Rand Index (Hubert & Arabie, 1985), ARI $= 0$ for a random grouping, ARI $= 1$ for a perfect match between groupings[1]. For each network, for the π-colBiSBM, ρ-colBiSBM, $\pi\rho$-colBiSBM we compare the inferred block memberships to the real ones by computing the mean of the ARI per dimension over the two networks

$$\overline{\text{ARI}}_d = \frac{1}{2}\big(\text{ARI}(\widehat{\boldsymbol{Z}_d^1}, \boldsymbol{Z}_d^1) + \text{ARI}(\widehat{\boldsymbol{Z}_d^2}, \boldsymbol{Z}_d^2)\big),$$

  where d is the dimension (i.e., rows, $d = 1$, or columns, $d = 2$) of the block memberships. And we compute the ARI of the whole set of nodes to account for block pairing between networks

$$\text{ARI}_d = \text{ARI}\big((\widehat{\boldsymbol{Z}_d^1}, \widehat{\boldsymbol{Z}_d^2}), (\boldsymbol{Z}_d^1, \boldsymbol{Z}_d^2)\big).$$

  The purpose of this metric is to verify that the block labels found in one network match the block labels in the second network.

All these quality indicators are averaged over the 108 datasets. The results are provided in the tables B.2 to B.5. Each line corresponds to the 108 datasets for a given value of ϵ_α. Graphical representation of some results are shown on figures 3.1 and 3.2.

[1]Please note that even if Rand Index can only yield values between 0 and 1, ARI can return negative values if the RI is less than the expected value. This indicates a structure in grouping discordance.

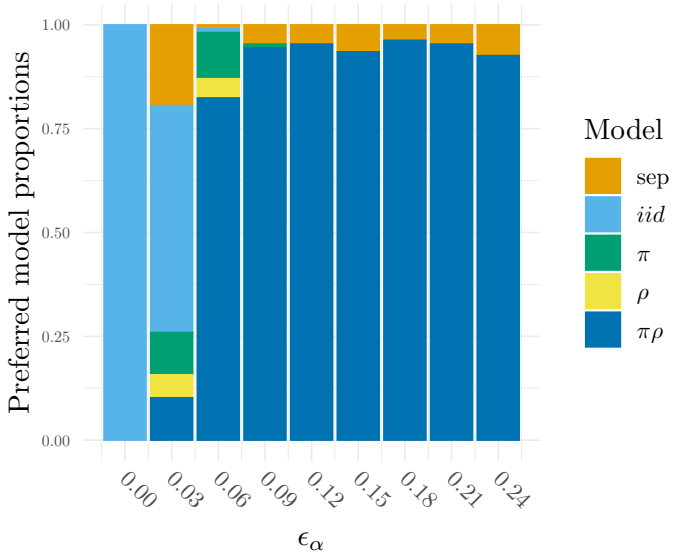


Figure 3.1: Preferred model proportions over all datasets in function of ϵ_α

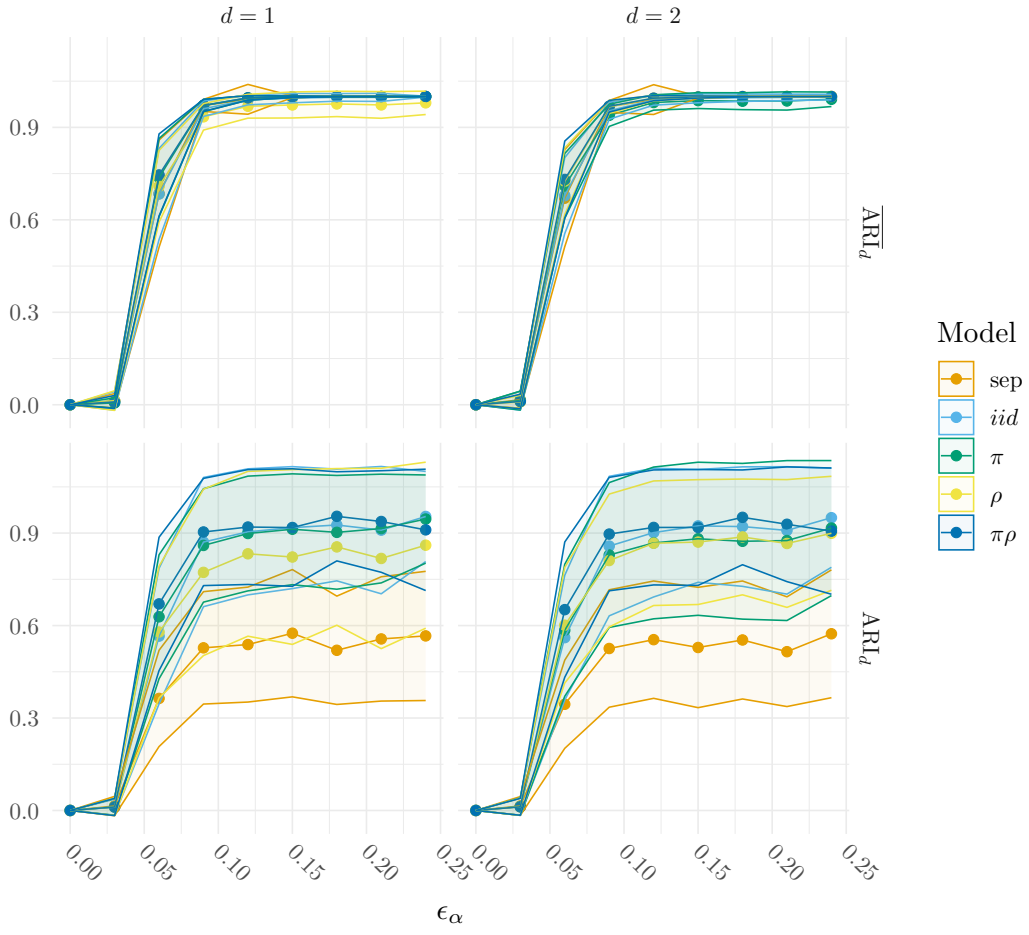


Figure 3.2: Plot of the ARI quality indicators in function of ϵ_α

Results For the model comparison, when ϵ_α is small ($\epsilon_\alpha \in [0, .03]$), the simulation model is close to an Erdős-Reńyi network (Erdős & Rényi, 1959), and it is very hard to find any structure beyond the one of a single block on each dimension.

On the figure 3.1 one can see that from $\epsilon_\alpha = 0.06$ around 75% of the time the $\pi\rho$-colBiSBM model (i.e., the correct one) is selected.

The figure 3.2 shows that for $\epsilon_\alpha \geq 0.09$, all the models, even the sep, have a $\overline{\text{ARI}}$ around 0.94. This indicates that the models are able to assign correct nodes group memberships and thus that the inference works correctly.

An interesting result we can read in the tables is that our models outperform the *sep-BiSBM* when considering the ARI on the whole set of nodes (ARI_d). This means that our models are able to recover the block pairing *between the networks* in addition to recovering the blocks and their parameters.

3.2 Capacity to distinguish $\pi\rho$-colBiSBM from *iid*-colBiSBM and other models

The idea of this model selection simulations is to assess how the model select the correct colBiSBM model among the possible ones: $iid, \pi, \rho, \pi\rho$. This difference being based on the row and col block proportions.

Simulation settings For this task we choose the same simulation settings as Chabert-Liddell et al., 2024.
Namely, $n_1^m = 90, n_2^m = 90, Q_1 = Q_2 = 3$, $\boldsymbol{\alpha}, \boldsymbol{\pi}$ and $\boldsymbol{\rho}$ are set as follows:

$$\boldsymbol{\alpha} = .25 + \begin{pmatrix} 3\epsilon_\alpha & 2\epsilon_\alpha & \epsilon_\alpha \\ 2\epsilon_\alpha & 2\epsilon_\alpha & -\epsilon_\alpha \\ \epsilon_\alpha & -\epsilon_\alpha & \epsilon_\alpha \end{pmatrix}, \quad \begin{aligned} \boldsymbol{\pi}^1 &= \left(\tfrac{1}{3}, \quad \tfrac{1}{3}, \quad \tfrac{1}{3}\right), \quad \boldsymbol{\pi}^2 = \sigma\left(\tfrac{1}{3} - \epsilon_\pi, \quad \tfrac{1}{3}, \quad \tfrac{1}{3} + \epsilon_\pi\right), \\ \boldsymbol{\rho}^1 &= \left(\tfrac{1}{3}, \quad \tfrac{1}{3}, \quad \tfrac{1}{3}\right), \quad \boldsymbol{\rho}^2 = \sigma\left(\tfrac{1}{3} - \epsilon_\rho, \quad \tfrac{1}{3}, \quad \tfrac{1}{3} + \epsilon_\rho\right), \end{aligned}$$

with $\epsilon_\alpha = 0.16$, ϵ_π and ϵ_ρ taking 9 values equally spaced in $[0, .28]$.
We simulate 324 different collections for each value of ϵ_π and ϵ_ρ.

$\pi\rho$-colBiSBM, π-colBiSBM, ρ-colBiSBM, *iid*-colBiSBM and *sep-BiSBM* are put in competition and the model with the greater BIC-L is selected as the *preferred model*.

When $\epsilon_\pi = 0$, $\boldsymbol{\pi}^1 = \boldsymbol{\pi}^2$, $\epsilon_\rho = 0$ and $\boldsymbol{\rho}^1 = \boldsymbol{\rho}^2$, the generated collection is an *iid*-colBiSBM. When $\epsilon_\pi > 0$ or $\boldsymbol{\pi}^1 \neq \boldsymbol{\pi}^2$, the model is a π-colBiSBM. When $\epsilon_\rho > 0$ or $\boldsymbol{\rho}^1 \neq \boldsymbol{\rho}^2$, the model is a ρ-colBiSBM. Finally, when $\epsilon_\pi > 0$ or $\boldsymbol{\pi}^1 \neq \boldsymbol{\pi}^2$ and $\epsilon_\rho > 0$ or $\boldsymbol{\rho}^1 \neq \boldsymbol{\rho}^2$, the model is a $\pi\rho$-colBiSBM.

Results On the figure 3.3 and table B.6, one can see that there is a turning point around $\epsilon_\pi = 0.2$ (resp. $\epsilon_\rho = 0.2$), before which *iid*-colBiSBM and ρ-colBiSBM (resp. π-colBiSBM) are selected very often and after 0.2 the π-colBiSBM (resp. ρ-colBiSBM) and $\pi\rho$-colBiSBM gets more and more selected. Moreover, the number of blocks are correctly detected in most of the case. These two results highlight our capacity to recover the simulated structure.

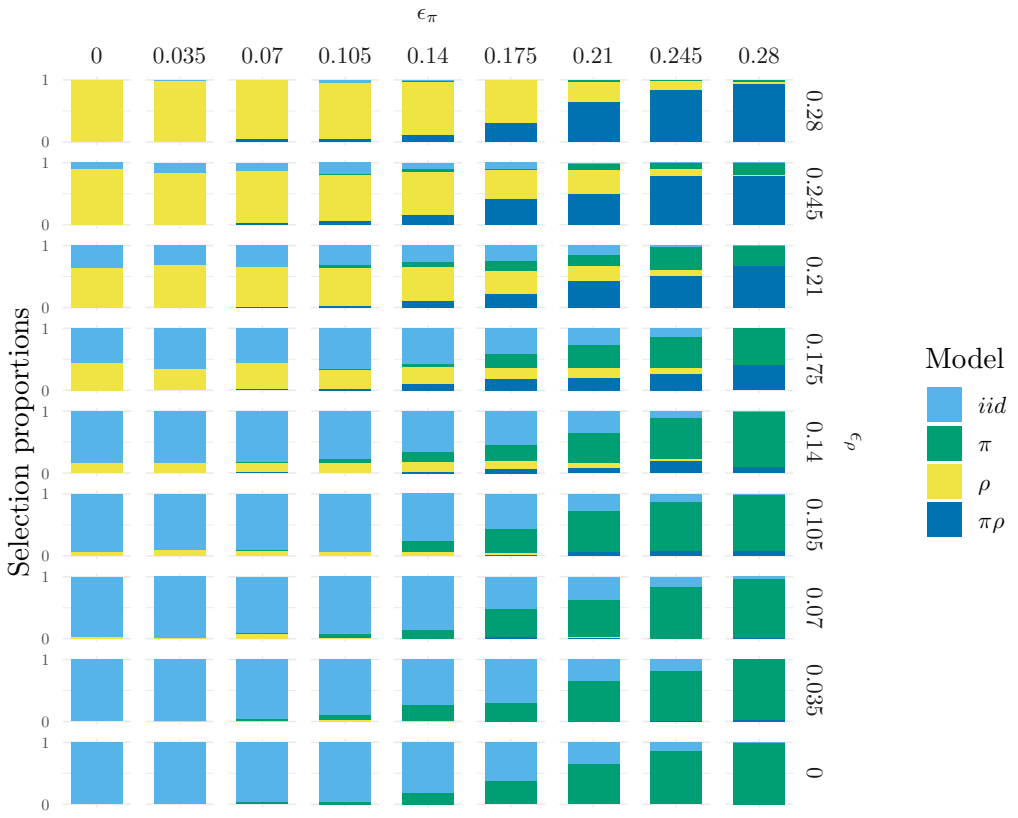


Figure 3.3: Plot of model selection proportions over the different datasets in function of ϵ_π and ϵ_ρ

As ϵ_π and ϵ_ρ need to be above 0.2 to see $\pi\rho$ model being preferred this may indicate the need of a strong difference between blocks to select this model.

3.3 Information transfer between networks

One of the motivation for collections of networks is *information transfer* between the networks, allowing robustness to missing data and enabling the finding of finer structures in small networks with the help of bigger ones.

Simulation settings We want to compare the performance of retrieving the nodes blocks with missing edges (that are labeled as `NA` in the incidence matrix).

For this purpose we generate collections of networks with the following parameters:

$$\boldsymbol{\pi}^m = \begin{cases} \boldsymbol{\pi} = (0.5, 0.3, 0.2) & \text{for } iid \\ \sigma_1^m(\boldsymbol{\pi}) & \text{for } \pi \text{ and } \pi\rho \end{cases}, \boldsymbol{\rho}^m = \begin{cases} \boldsymbol{\rho} = (0.5, 0.3, 0.2) & \text{for } iid \\ \sigma_2^m(\boldsymbol{\rho}) & \text{for } \rho \text{ and } \pi\rho, \end{cases}$$

for the block proportions, and two different structures with the corresponding $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha}^{modular} = \begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.2 & 0.05 \\ 0.05 & 0.05 & 0.8 \end{pmatrix}, \boldsymbol{\alpha}^{nested} = \begin{pmatrix} 0.9 & 0.65 & 0.1 \\ 0.35 & 0.15 & 0.05 \\ 0.1 & 0.05 & 0.05 \end{pmatrix},$$

where $\boldsymbol{\alpha}^{modular}$ represents networks where there are look-a-like communities, which tends to interact preferentially within the community and less with the other communities. And $\boldsymbol{\alpha}^{nested}$ represents a common structure detected in ecology with generalist and specialist species and a "nested" structure.

The collections contain two networks ($M = 2$) of size $n_1^{m=1} = n_2^{m=1} = 20$ and $n_1^{m=2} = n_2^{m=2} = 120$. One collection is generated for each colBiSBM model. And the nodes block memberships (i.e., the row and column blocks they belong to) are saved.

Per colBiSBM model, 10 collections are generated and their results are averaged.

In the network $m = 1$ (i.e., the smaller one) a proportion of the edges $p_{\texttt{NA}}$ see their values replaced by `NA`s, the "forgotten" values are stored.

Test procedure A LBM is fitted on the first network, and the predicted block memberships are saved, along with the predicted links using the inferred parameters. This will serve as a baseline to see if the use of the collection benefits the predictions.

A colBiSBM model is then fitted (with a model matching the dataset considered) and we store the same predictions.

Quality metrics To benchmark the performance we use the *Area Under the Curve* (AUC) for predicted versus real link values and the ARI for predicted versus real block memberships.

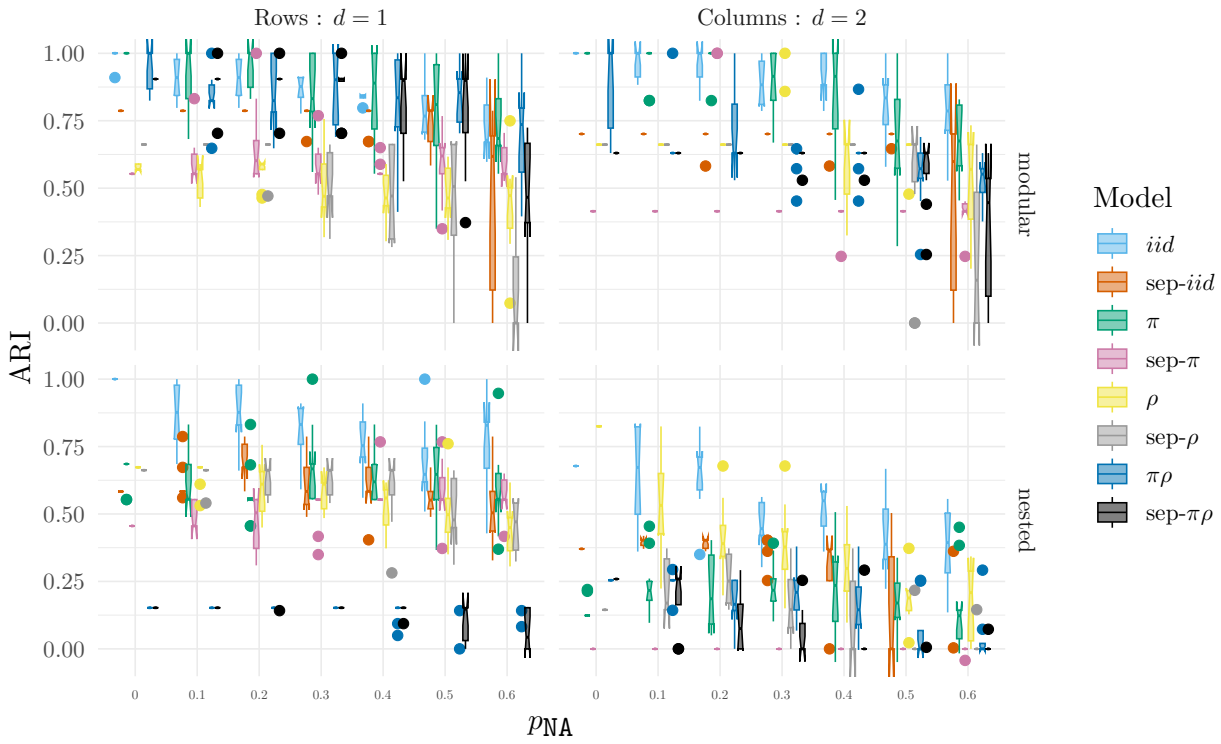


Figure 3.4: ARI in function of $p_{\mathtt{NA}}$, the proportion of missing links for various colBiSBM models and their LBM counterparts

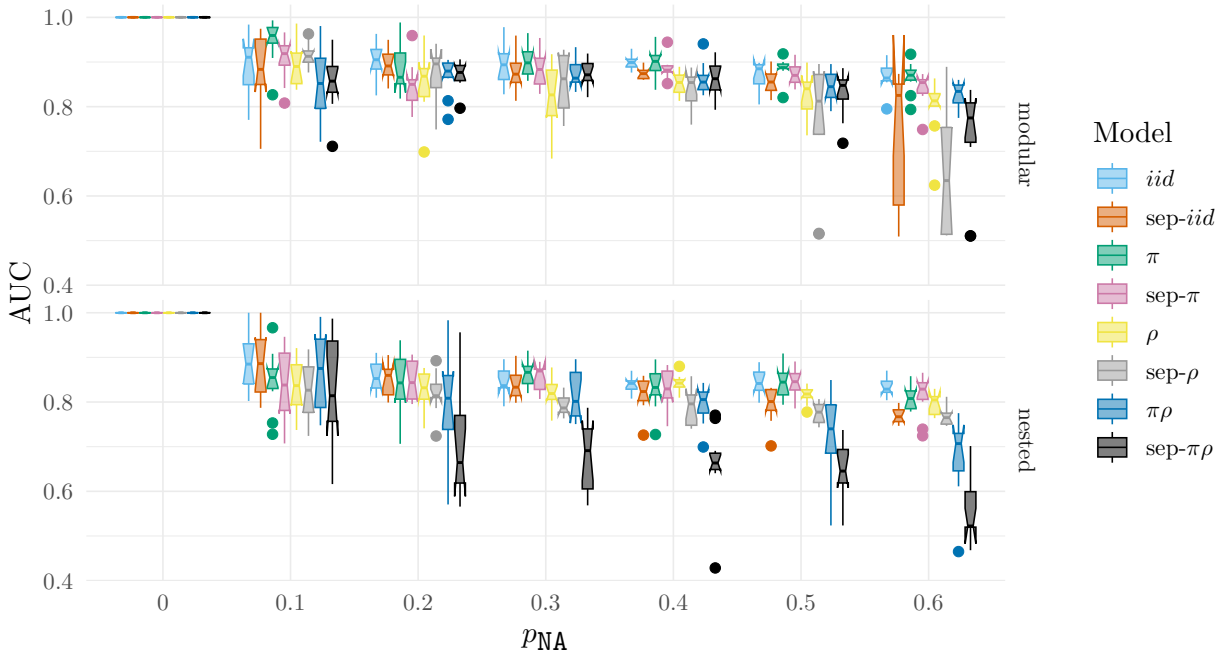


Figure 3.5: AUC in function of $p_{\mathtt{NA}}$, the proportion of missing links for various colBiSBM models and their LBM counterparts

Results Figures 3.5 and 3.4 show a box plots named "sep-*model*" that corresponds to the results given by a LBM fitted on data generated with the corresponding *model*. We will compare the results for one model box plot to the corresponding sep-model box plot, serving as a baseline.

For the figure 3.4, our models almost always do at least as good as the sep counterpart. The *iid* model is the only one for which the sep performs better on the columns block memberships. The nested structure seems to complexify the block membership attribution with only ARI less than 0.75

For the figure 3.5, in almost all cases and for almost all models the differences are not significant but our models seems to perform marginally better and are only a few times under their LBM counterpart. This indicates that information is transferred from the bigger network when estimating the parameters and predicting link values.

On the differences between nested and modular structures, the latter shows a smaller variance in the AUC with our models predictions contained between 0.7 and 0.9. Whereas for the nested structure, *iid* and π models are in quite similar value ranges with small variances but ρ and $\pi\rho$ present smaller values and larger variances.

An explanation for the cases in which our models return lower values than expected could be to look for in our simulation parameters. They may, combined with the ρ model be a difficult case for the estimation. As we currently do not have identifiability results this is just an hypothesis.

3.4 Network clustering of simulated networks

Simulation settings For all models we simulate $M = 9$ networks with $\forall m \in \{1 \dots M\}, n_1^m = n_2^m = 75$ with $Q_1 = Q_2 = 3$.
For the simulations the proportions are the following:
$\boldsymbol{\pi}^1 = (0.2, 0.3, 0.5)$, $\boldsymbol{\rho}^1 = (0.2, 0.3, 0.5)$ and for all $m = 2, \dots, 9$

$$\boldsymbol{\pi}^m = \begin{cases} \boldsymbol{\pi}^1 & \text{for } iid \\ \sigma_1^m(\boldsymbol{\pi}^1) & \text{for } \pi \text{ and } \pi\rho \end{cases}, \boldsymbol{\rho}^m = \begin{cases} \boldsymbol{\rho}^1 & \text{for } iid \\ \sigma_2^m(\boldsymbol{\rho}^1) & \text{for } \rho \text{ and } \pi\rho \end{cases}$$

where σ_1^m and σ_2^m are permutations of $\{1, 2, 3\}$ proper to network m and $\sigma_1(\pi) = (\pi_{\sigma_1(i)})_{i=\{1,\dots,3\}}$ and $\sigma_2(\rho) = (\rho_{\sigma_2(i)})_{i=\{1,\dots,3\}}$.

The networks are divided into 3 sub-collections of 3 networks with connectivity parameters as follows:

$$\boldsymbol{\alpha}^{as} = .3 + \begin{pmatrix} \epsilon & -\frac{\epsilon}{2} & -\frac{\epsilon}{2} \\ -\frac{\epsilon}{2} & \epsilon & -\frac{\epsilon}{2} \\ -\frac{\epsilon}{2} & -\frac{\epsilon}{2} & \epsilon \end{pmatrix}, \quad \boldsymbol{\alpha}^{dis} = .3 + \begin{pmatrix} -\frac{\epsilon}{2} & \epsilon & \epsilon \\ \epsilon & -\frac{\epsilon}{2} & \epsilon \\ \epsilon & \epsilon & -\frac{\epsilon}{2} \end{pmatrix}, \boldsymbol{\alpha}^{cp} = .3 + \begin{pmatrix} \frac{3\epsilon}{2} & \epsilon & \frac{\epsilon}{2} \\ \epsilon & \frac{\epsilon}{2} & 0 \\ \frac{\epsilon}{2} & 0 & -\frac{\epsilon}{2} \end{pmatrix}$$

with $\epsilon \in [.1, .4]$. $\boldsymbol{\alpha}^{as}$ represents a classical assortative community structure, while $\boldsymbol{\alpha}^{cp}$ is a layered core-periphery structure with block 2 acting as a semi-core. Finally, $\boldsymbol{\alpha}^{dis}$ is a dis-assortative community structure with stronger connections between blocks than within blocks. If $\epsilon = 0$, the three matrices are equal and the 9 networks have the same connection structure. Increasing ϵ differentiates the 3 sub-collections of networks.

Results The evaluation of our method involves a comparison between the resulting partition of the network collection and the simulated partition using the ARI index. As the value of ϵ increases, our ability to distinguish between the networks improves, and this distinction becomes nearly perfect in all setups of the colBiSBM.

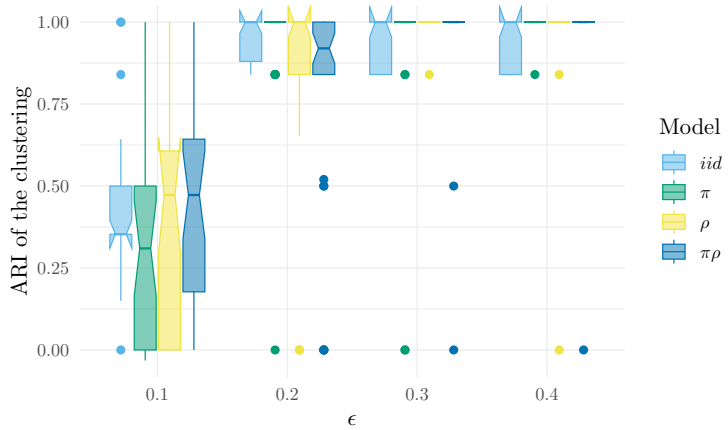


Figure 3.6: ARI obtained for the clustering with the different models in function of ϵ

Chapter 4

Applications on ecological networks

4.1 Clustering of Baldock et al., 2019, 2011

Baldock et al., 2019 study the diversity, robustness and impact of the type of environment on the ecological aspect of plant-pollinator networks in four major english cities. The networks are presented in figures 4.1a to 4.1d

Baldock et al., 2011 aim to study the daily temporal structure in a savanna pollination network. The data was collected in 2004 and in two sites, *Turkana Boma* (TB) and *Junction* (JN) in Kenya. We will not look for a temporal structure but only use the full networks merged in one due to their small sizes. The network obtained is presented in figure 4.2.

In the following results, the row nodes represent the plants and the column nodes represent the pollinators (mostly insects in this case).

Note: those networks were extracted from a bigger dataset from Doré et al., 2021. The full dataset was also clusterized but issues arose that are discussed in sub-section 5.1.2.

(a) Bristol (b) Edinburgh

(c) Leeds (d) Reading

Figure 4.1: English networks from Baldock et al., 2019

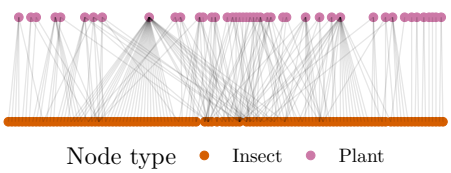
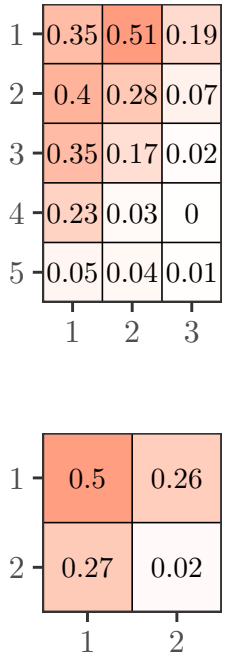
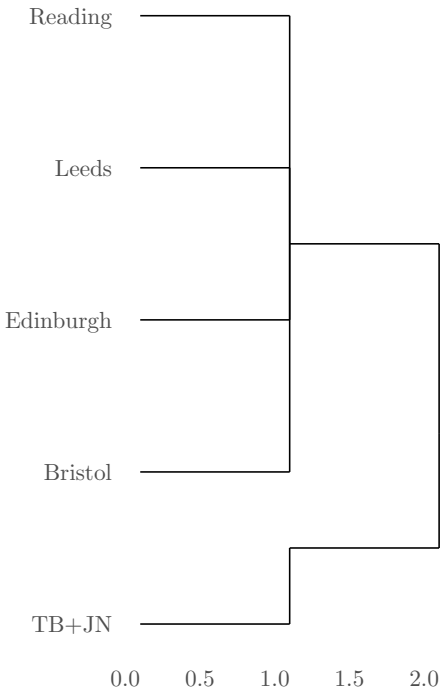


Figure 4.2: African network from Baldock et al., 2011

We applied our clustering method on those 6 networks, using the four models. Interesting results arose from *iid* and $\pi\rho$ models, which are presented below in figures 4.4 and 4.5.



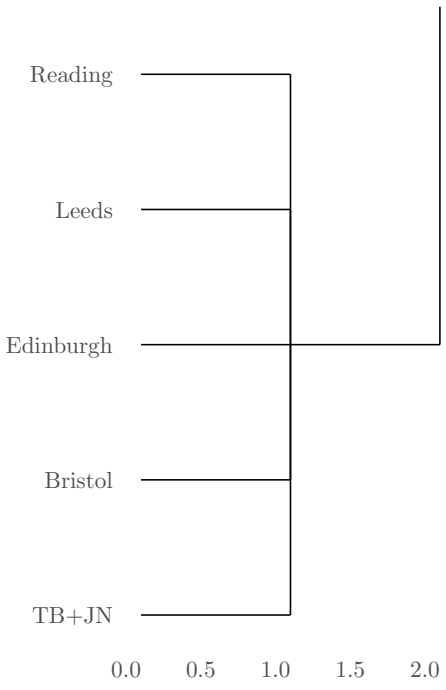
(a) $\boldsymbol{\alpha}$ structures of the collections identified



(b) Tree of splits

Figure 4.4: Results for *iid* clustering

(a) $\boldsymbol{\alpha}$ structure of the collection identified



(b) Tree of splits

Figure 4.5: Results for $\pi\rho$ clustering

Results The main thing one can see when comparing the two clusterings is that while *iid* do not find a common structure, $\pi\rho$ manage to find one.

When comparing figures 4.3a, 4.4a and with the figures C.1, C.2 in appendix which show the block proportions we can deduce the following differences between the two models.

The TB+JN network doesn't have the 3^{rd} column block and the first and second column blocks coincides with the English networks. But interestingly, the 3^{rd} column block is the larger one in the English networks.

For the row blocks, where with the *iid* model only two groups where detected in the African network, it is refined with the $\pi\rho$ detecting 4 blocks in its row nodes. The African network does not have the first row block, a few row nodes in row blocks 2 and 3, more in the 4^{th} block and the majority in the 5^{th}. This contrasts, once again, with the English networks that have the majority of their row nodes in the 4^{th} block.

Two interesting points can be made with those results, firstly the $\pi\rho$ collection model allowed the detection of a finer structure in the African network regarding the row block memberships. And secondly, whereas the English networks are really similar, the African network while presenting the same blocks fills them differently. Those observations invite us to look at the species that are in those different groups.

Chapter 5

Conclusions and future work

5.1 Conclusion

5.1.1 State of work at the end of the internship

At the end of this internship we now have:

- A model capable to find structure in collections of bipartite network. Enabling the possibility to bring together networks that may have not seem evident to put together.
- A clustering method that is able to partition a collection into collections that are similar in their structures.
- All the described methods implemented into an `R` package.

5.1.2 Difficulties encountered

Local optima While using our clustering on data from Doré et al., 2021 we obtained quite interesting results but investigating further, we noticed that the clustering on such big collections ($M = 123$) was not fully reproducible. It depends a lot on the random generator seed and as there is no possibility to merge back collections the clustering dendrograms do not stabilize on large collections. This, currently, prevents us to clusterize large collections. We suspect that our model selection method gets stuck on local optima of the BIC-L.

Large penalties with free mixture models We observed while testing clustering with the different models that the π, ρ and $\pi\rho$ model, with their increased number of parameters for block memberships parameters tends to give smaller BIC-L criterion values while having a higher Evidence Lower Bound than the *iid*. This arises because of the penalties on the block memberships and supports that increase significantly and exceeds the gain on the ELBO and the diminution of the connectivity parameters.

5.2 Future work

Fixing local optima We are currently investigating the procedure and code to see if reducing or escaping seed dependance is possible and would allow escaping the local optima.

Identifiability As stated in section 2.10, we only have identifiability for the *iid*-colBiSBM and we will work on establishing identifiability for π, ρ and $\pi\rho$ models which are the most challenging with regard to identifiability.

Finding a trade-off between *iid* and $\pi\rho$ An idea to tackle the problem of large penalties with π, ρ and $\pi\rho$ could be to suppose that the block memberships for network m are themselves the realizations of random variables and thus introduce sort of a mixed effect model. This may allow a self-penalization that could keep the flexibility intended in these models.

More ecological applications We have leads to apply the models on interesting cases, among them are the following:

- Networks that are spaced along an altitude gradient, which could be accounted for in the dissimilarity measure for instance.
- Collection of different sorts of interactions (plants-seed dispersors, host-parasites, ...)

Turning this work into an article We will work in the upcoming months to write an article about the method.

Comparison to other graphs clustering methods Recent work have been comparing `colSBM` (Chabert-Liddell et al., 2024) and `graphclust` (Rebafka, 2023) assessing various capabilities of the models and particularly focusing on networks clustering. We will reproduce and adapt the analysis to test other simulation settings that were not considered in this work.

Bibliography

Baldock, K. C. R., Goddard, M. A., Hicks, D. M., Kunin, W. E., Mitschunas, N., Morse, H., Osgathorpe, L. M., Potts, S. G., Robertson, K. M., Scott, A. V., Staniczenko, P. P. A., Stone, G. N., Vaughan, I. P., & Memmott, J. (2019). A systems approach reveals urban pollinator hotspots and conservation opportunities. *Nat Ecol Evol*, *3*(3), 363–373. https://doi.org/10.1038/s41559-018-0769-y

Baldock, K. C. R., Memmott, J., Ruiz-Guajardo, J. C., Roze, D., & Stone, G. N. (2011). Daily temporal structure in African savanna flower visitation networks and consequences for network sampling. *Ecology*, *92*(3), 687–698. https://doi.org/10.1890/10-1110.1

Chabert-Liddell, S.-C., Barbillon, P., & Donnet, S. (2024). Learning common structures in a collection of networks. An application to food webs. *The Annals of Applied Statistics*, *18*(2), 1213–1235. https://doi.org/10.1214/23-AOAS1831

Ramos-Jiliberto, R., Domínguez, D., Espinoza, C., López, G., Valdovinos, F. S., Bustamante, R. O., & Medel, R. (2010). Topological change of Andean plant–pollinator networks along an altitudinal gradient. *Ecological Complexity*, *7*(1), 86–90. https://doi.org/10.1016/j.ecocom.2009.06.001

Kaszewska-Gilas, K., Kosicki, J. Z., Hromada, M., & Skoracki, M. (2021). Global Studies of the Host-Parasite Relationships between Ectoparasitic Mites of the Family Syringophilidae and Birds of the Order Columbiformes. *Animals*, *11*(12), 3392. https://doi.org/10.3390/ani11123392

Desjardins-Proulx, P., Laigle, I., Poisot, T., & Gravel, D. (2017). Ecological interactions and the Netflix problem. *PeerJ*, *5*, e3644. https://doi.org/10.7717/peerj.3644

Pavlopoulos, G. A., Kontou, P. I., Pavlopoulou, A., Bouyioukos, C., Markou, E., & Bagos, P. G. (2018). Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience*, *7*(4), giy014. https://doi.org/10.1093/gigascience/giy014

Pichon, B., Le Goff, R., Morlon, H., & Perez-Lamarque, B. (2024). Telling mutualistic and antagonistic ecological networks apart by learning their multiscale structure. *Methods in Ecology and Evolution*, *15*(6), 1113–1128. https://doi.org/10.1111/2041-210X.14328

Govaert, G., & Nadif, M. (2010). Latent Block Model for Contingency Table. *Communications in Statistics - Theory and Methods*, *39*(3), 416–425. https://doi.org/10.1080/03610920903140197

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, *5*(2), 109–137. https://doi.org/10.1016/0378-8733(83)90021-7

Snijders, T. A., & Nowicki, K. (1997). Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *J. of Classification*, *14*(1), 75–100. https://doi.org/10.1007/s003579900004

Daudin, J.-J., Picard, F., & Robin, S. (2008). A mixture model for random graphs. *Stat Comput*, *18*(2), 173–183. https://doi.org/10.1007/s11222-007-9046-7

Govaert, G., & Nadif, M. (2005). An EM algorithm for the block mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(4), 643–647. https://doi.org/10.1109/TPAMI.2005.69

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725. https://doi.org/10.1109/34.865189

Keribin, C., Brault, V., Celeux, G., & Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Stat Comput*, *25*(6), 1201–1216. https://doi.org/10.1007/s11222-014-9472-2

Celisse, A., Daudin, J.-J., & Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, *6*, 1847–1899. https://doi.org/10.1214/12-EJS729

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Erdős, P., & Rényi, A. (1959). On random graphs. I. *Publ. Math. Debrecen*, *6*(3-4), 290–297. https://doi.org/10.5486/PMD.1959.6.3-4.12

Doré, M., Fontaine, C., & Thébault, E. (2021). Relative effects of anthropogenic pressures, climate, and sampling design on the structure of pollination networks at the global scale. *Global Change Biology*, *27*(6), 1266–1280. https://doi.org/10.1111/gcb.15474

Rebafka, T. (2023, November 6). *Model-based clustering of multiple networks with a hierarchical algorithm*. arXiv: 2211.02314 [math, stat]. https://doi.org/10.48550/arXiv.2211.02314

Appendix A

Supplementary for Structure detection in bipartite collection

A.1 Proof of the idenfiability result

We recall the following

Properties 1 (Identifiability of *iid*-colBiSBM)**.** *The parameters* $(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ *are identifiable up to a label switching of the blocks if those conditions are achieved:*

(1.1) $\exists m^* \in \{1, \dots, M\} : n^1_{m^*} \geq 2Q_2 - 1$ *and* $n^2_{m^*} \geq 2Q_1 - 1$.

(1.2) $\forall 1 \leq q \leq Q_1, \pi_q > 0$ *and the coordinates of vector* $\boldsymbol{\rho} X^{m^*T}$ *are distinct (where* X^{m^*T} *is the transpose of* X^{m^*}*).*

(1.3) $\forall 1 \leq r \leq Q_2, \rho_r > 0$ *and the coordinates of vector* $\boldsymbol{\pi} X^{m^*}$ *are distinct.*

Proof. Following the tracks of Chabert-Liddell et al., 2024 we derive the result in Properties 1.

Keribin et al., 2015 building on Celisse et al., 2012, proved that the parameters $(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ of the $\mathcal{F}$-BiSBM$_{n_1^m, n_2^m}(Q_1^m, Q_2^m, \boldsymbol{\pi^m}, \boldsymbol{\rho^m}, \boldsymbol{\alpha^m})$ are identifiable from the observation of network X^m when $\mathcal{F}$ is the Bernoulli distribution and the following conditions are met:

1. $n_1^m \geq 2Q_2^m - 1$ and $n_2^m \geq 2Q_1^m - 1$.
2. $\forall 1 \leq q \leq Q_1^m, \pi_q^m > 0$ and the coordinates of vector $\boldsymbol{\rho^m} X^{m^*T}$ are distinct (where X^{m^*T} is the transpose of X^{m^*}).
3. $\forall 1 \leq r \leq Q_2^m, \rho_r^m > 0$ and the coordinates of vector $\boldsymbol{\pi^m} X^{m^*}$ are distinct.

Under the *iid*-colBiSBM model, for all $m = 1 \dots M$, $X^m \sim \mathcal{F}$-BiSBM$_{n_1^m, n_2^m}(Q_1, Q_2, \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$. This means that following Keribin et al., 2015, the identifiability of $\boldsymbol{\alpha}$, $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ is obtained from the distribution of X^{m^*} under assumptions (1.1), (1.2) and (1.3). □

Appendix B

Supplementary for Simulation studies

Below are the supplementary material for the Simulation studies.

B.1 Supplementary for Efficiency of the inference

The tables B.2 to B.5 show detailed results for the inference of the model detailed in this section.

Table B.1: Inference results for *sep*

(a) Quality metrics for *sep*-BiSBM

ϵ_α	$\overline{\text{ARI}}_1$	$\overline{\text{ARI}}_2$	ARI_1	ARI_2
0.00	0	0	0	0
0.03	0.01	0.02	0.01	0.01
0.06	0.68 ± 0.02	0.67 ± 0.02	0.36 ± 0.02	0.34 ± 0.01
0.09	0.97	0.97	0.53 ± 0.02	0.53 ± 0.02
0.12	0.99	0.99	0.54 ± 0.02	0.55 ± 0.02
0.15	1	1	0.58 ± 0.02	0.53 ± 0.02
0.18	1	1	0.52 ± 0.02	0.55 ± 0.02
0.21	1	1	0.56 ± 0.02	0.52 ± 0.02
0.24	1	1	0.57 ± 0.02	0.57 ± 0.02

Table B.2: Inference results for *iid*

(a) Quality metrics for *iid*-colBiSBM

ϵ_α	$\overline{\text{ARI}}_1$	$\overline{\text{ARI}}_2$	ARI_1	ARI_2
0.00	0	0	0	0
0.03	0.01	0.01	0.01	0.01
0.06	0.68 ± 0.01	0.68 ± 0.01	0.57 ± 0.02	0.56 ± 0.02
0.09	0.96	0.95	0.87 ± 0.02	0.86 ± 0.02
0.12	0.99	0.99	0.9 ± 0.02	0.9 ± 0.02
0.15	0.99	0.99	0.92 ± 0.02	0.92 ± 0.02
0.18	1	1	0.93 ± 0.02	0.92 ± 0.02
0.21	1	1	0.91 ± 0.02	0.91 ± 0.02
0.24	1	1	0.95 ± 0.01	0.95 ± 0.02

(b) Bloc recovery for *iid*-colBiSBM

ϵ_α	$\mathbb{1}_{\widehat{Q_1}<Q_1}$	$\mathbb{1}_{\widehat{Q_1}=Q_1}$	$\mathbb{1}_{\widehat{Q_1}>Q_1}$	$\mathbb{1}_{\widehat{Q_2}<Q_2}$	$\mathbb{1}_{\widehat{Q_2}=Q_2}$	$\mathbb{1}_{\widehat{Q_2}>Q_2}$
0.00	1	0	0	1	0	0
0.03	1	0	0	1	0	0
0.06	0.31 ± 0.04	0.69 ± 0.04	0	0.45 ± 0.05	0.55 ± 0.05	0
0.09	0	0.96 ± 0.02	0.04 ± 0.02	0	0.98 ± 0.01	0.02 ± 0.01
0.12	0	0.88 ± 0.03	0.12 ± 0.03	0	0.94 ± 0.02	0.06 ± 0.02
0.15	0	0.94 ± 0.02	0.06 ± 0.02	0	0.91 ± 0.03	0.09 ± 0.03
0.18	0	0.92 ± 0.03	0.08 ± 0.03	0	0.94 ± 0.02	0.06 ± 0.02
0.21	0	0.88 ± 0.03	0.12 ± 0.03	0	0.88 ± 0.03	0.12 ± 0.03
0.24	0	0.94 ± 0.02	0.06 ± 0.02	0	0.97 ± 0.02	0.03 ± 0.02

Table B.3: Inference results for π

(a) Quality metrics for π-colBiSBM

ϵ_α	$\overline{\text{ARI}}_1$	$\overline{\text{ARI}}_2$	ARI_1	ARI_2
0.00	0	0	0	0
0.03	0.01	0.01	0.01	0.01
0.06	0.74 ± 0.01	0.71 ± 0.01	0.63 ± 0.02	0.58 ± 0.02
0.09	0.97	0.94	0.86 ± 0.02	0.83 ± 0.02
0.12	1	0.98	0.9 ± 0.02	0.87 ± 0.02
0.15	1	0.99	0.91 ± 0.02	0.88 ± 0.02
0.18	1	0.98	0.9 ± 0.02	0.87 ± 0.02
0.21	1	0.99	0.91 ± 0.02	0.88 ± 0.02
0.24	1	0.99	0.95 ± 0.01	0.92 ± 0.02

(b) Bloc recovery for π-colBiSBM

ϵ_α	$\mathbb{1}_{\widehat{Q_1}<Q_1}$	$\mathbb{1}_{\widehat{Q_1}=Q_1}$	$\mathbb{1}_{\widehat{Q_1}>Q_1}$	$\mathbb{1}_{\widehat{Q_2}<Q_2}$	$\mathbb{1}_{\widehat{Q_2}=Q_2}$	$\mathbb{1}_{\widehat{Q_2}>Q_2}$
0.00	1	0	0	1	0	0
0.03	1	0	0	1	0	0
0.06	0.09 ± 0.03	0.64 ± 0.05	0.27 ± 0.04	0.36 ± 0.05	0.64 ± 0.05	0
0.09	0	0.63 ± 0.05	0.37 ± 0.05	0	1	0
0.12	0	0.76 ± 0.04	0.24 ± 0.04	0	1	0
0.15	0	0.8 ± 0.04	0.2 ± 0.04	0	1	0
0.18	0	0.77 ± 0.04	0.23 ± 0.04	0	1	0
0.21	0	0.81 ± 0.04	0.19 ± 0.04	0	1	0
0.24	0	0.87 ± 0.03	0.13 ± 0.03	0	1	0

Table B.4: Inference results for ρ

(a) Quality metrics for ρ-colBiSBM

ϵ_α	$\overline{\text{ARI}}_1$	$\overline{\text{ARI}}_2$	ARI_1	ARI_2
0.00	0	0	0	0
0.03	0.01	0.01	0.01	0.01
0.06	0.71 ± 0.01	0.73 ± 0.01	0.58 ± 0.02	0.6 ± 0.02
0.09	0.93	0.97	0.77 ± 0.03	0.81 ± 0.02
0.12	0.97	0.99	0.83 ± 0.03	0.87 ± 0.02
0.15	0.97	1	0.82 ± 0.03	0.87 ± 0.02
0.18	0.98	1	0.86 ± 0.02	0.89 ± 0.02
0.21	0.97	1	0.82 ± 0.03	0.87 ± 0.02
0.24	0.98	1	0.86 ± 0.03	0.9 ± 0.02

(b) Bloc recovery for ρ-colBiSBM

ϵ_α	$\mathbb{1}_{\widehat{Q_1}<Q_1}$	$\mathbb{1}_{\widehat{Q_1}=Q_1}$	$\mathbb{1}_{\widehat{Q_1}>Q_1}$	$\mathbb{1}_{\widehat{Q_2}<Q_2}$	$\mathbb{1}_{\widehat{Q_2}=Q_2}$	$\mathbb{1}_{\widehat{Q_2}>Q_2}$
0.00	1	0	0	1	0	0
0.03	1	0	0	1	0	0
0.06	0.29 ± 0.04	0.71 ± 0.04	0	0.06 ± 0.02	0.63 ± 0.05	0.31 ± 0.04
0.09	0	1	0	0	0.56 ± 0.05	0.44 ± 0.05
0.12	0	1	0	0	0.67 ± 0.05	0.33 ± 0.05
0.15	0	1	0	0	0.7 ± 0.04	0.3 ± 0.04
0.18	0	1	0	0	0.71 ± 0.04	0.29 ± 0.04
0.21	0	1	0	0	0.69 ± 0.04	0.31 ± 0.04
0.24	0	1	0	0	0.76 ± 0.04	0.24 ± 0.04

Table B.5: Inference results for $\pi\rho$

(a) Quality metrics for $\pi\rho$-colBiSBM

ϵ_α	$\overline{\text{ARI}}_1$	$\overline{\text{ARI}}_2$	ARI_1	ARI_2
0.00	0	0	0	0
0.03	0.01	0.01	0.01	0.01
0.06	0.75 ± 0.01	0.73 ± 0.01	0.67 ± 0.02	0.65 ± 0.02
0.09	0.97	0.97	0.9 ± 0.02	0.9 ± 0.02
0.12	1	0.99	0.92 ± 0.02	0.92 ± 0.02
0.15	1	1	0.92 ± 0.02	0.92 ± 0.02
0.18	1	1	0.95 ± 0.01	0.95 ± 0.01
0.21	1	1	0.94 ± 0.02	0.93 ± 0.02
0.24	1	1	0.91 ± 0.02	0.91 ± 0.02

(b) Bloc recovery for $\pi\rho$-colBiSBM

ϵ_α	$\mathbb{1}_{\widehat{Q_1}<Q_1}$	$\mathbb{1}_{\widehat{Q_1}=Q_1}$	$\mathbb{1}_{\widehat{Q_1}>Q_1}$	$\mathbb{1}_{\widehat{Q_2}<Q_2}$	$\mathbb{1}_{\widehat{Q_2}=Q_2}$	$\mathbb{1}_{\widehat{Q_2}>Q_2}$
0.00	1	0	0	1	0	0
0.03	1	0	0	1	0	0
0.06	0.1 ± 0.03	0.88 ± 0.03	0.02 ± 0.01	0.11 ± 0.03	0.89 ± 0.03	0
0.09	0	0.92 ± 0.03	0.08 ± 0.03	0	0.88 ± 0.03	0.12 ± 0.03
0.12	0	0.9 ± 0.03	0.1 ± 0.03	0	0.89 ± 0.03	0.11 ± 0.03
0.15	0	0.87 ± 0.03	0.13 ± 0.03	0	0.9 ± 0.03	0.1 ± 0.03
0.18	0	0.94 ± 0.02	0.06 ± 0.02	0	0.93 ± 0.03	0.07 ± 0.03
0.21	0	0.93 ± 0.03	0.07 ± 0.03	0	0.88 ± 0.03	0.12 ± 0.03
0.24	0	0.87 ± 0.03	0.13 ± 0.03	0	0.86 ± 0.03	0.14 ± 0.03

B.2 Supplementary for Capacity to distinguish models

The table B.6 present the results discussed in section 3.2 For the block number recovery part, the *minimum* values are in **bold** as they indicate conditions for which all the different models did not recovered the correct structure.

For the model proportion part of the table, the *maximum* values are in **bold** and highlight the model that was selected the most among the conditions.

Please note that blank space indicates that among all conditions the corresponding model was not selected at all.

Table B.6: Filtered block recovery and model selection proportions

		Block number recovery								Model selection proportions			
		iid		π		ρ		$\pi\rho$					
ϵ_π	ϵ_ρ	$\mathbb{1}_{\widehat{Q1}_{iid}=3}$	$\mathbb{1}_{\widehat{Q2}_{iid}=3}$	$\mathbb{1}_{\widehat{Q1}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q2}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q1}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q2}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q1}_{\pi\rho}=3}$	$\mathbb{1}_{\widehat{Q2}_{\pi\rho}=3}$	*iid*	π	ρ	$\pi\rho$
	0.000	1	1	1	1	1	1	1	1	**1**			
	0.035	1	1	1	1	1	1	1	1	**1**			
	0.070	1	1	1	1	1	1	1	1	**0.981**	0.019		
	0.105	1	1	1	1	1	1	1	1	**0.944**	0.056		
0.000	0.140	1	1	1	1	1	1	1	1	**0.833**	0.157	0.009	
	0.175	1	1	1	1	1	1	1	1	**0.556**	0.444		
	0.210	1	1	1	1	1	1	1	1	0.361	**0.639**		
	0.245	1	1	1	1	1	1	1	1	0.102	**0.898**		
	0.280	1	**0.991**	1	1	1	1	1	1	0.009	**0.991**		
	0.000	1	1	1	1	1	1	1	1	**1**			
	0.035	1	1	1	1	1	1	1	1	**0.991**	0.009		
	0.070	1	1	1	1	1	1	1	1	**0.991**	0.009		
	0.105	1	1	1	1	1	1	1	1	**0.917**	0.083		
	0.140	1	1	1	1	1	1	1	1	**0.843**	0.157		
	0.175	1	1	1	1	1	1	1	1	**0.657**	0.333	0.009	
	0.210	1	1	1	1	1	1	1	1	0.315	**0.685**		

Table B.6: Filtered block recovery and model selection proportions *(continued)*

0.035		Block number recovery								Model selection proportions			
		iid		π		ρ		$\pi\rho$					
ϵ_π	ϵ_ρ	$\mathbb{1}_{\widehat{Q_{1}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi\rho}=3}$	*iid*	π	ρ	$\pi\rho$
	0.245	1	1	1	1	1	1	1	1	0.176	**0.824**		
	0.280	1	1	1	1	1	1	1	1	0.028	**0.972**		
	0.000	1	1	1	1	1	1	1	1	**0.972**		0.028	
	0.035	1	1	1	1	1	1	1	1	**0.963**	0.009	0.028	
	0.070	1	1	1	1	1	1	1	1	**0.917**	0.065	0.019	
	0.105	1	1	1	1	1	1	1	1	**0.917**	0.065	0.019	
0.070	0.140	1	1	1	1	1	1	1	1	**0.824**	0.157	0.009	0.009
	0.175	1	1	1	1	1	1	1	1	**0.565**	0.426		0.009
	0.210	1	1	1	1	1	1	1	1	0.343	**0.639**	0.009	0.009
	0.245	1	1	1	1	1	1	**0.991**	1	0.139	**0.843**		0.019
	0.280	1	1	1	1	**0.991**	1	1	1		**0.963**		0.037
	0.000	1	1	1	1	1	1	1	1	**0.963**		0.037	
	0.035	1	1	1	1	1	1	1	1	**0.898**	0.019	0.083	
	0.070	1	1	1	1	1	1	1	1	**0.926**	0.009	0.065	
	0.105	1	1	1	1	1	1	1	1	**0.935**	0.056	0.009	
	0.140	1	1	1	1	1	1	1	1	**0.778**	0.157	0.065	

Table B.6: Filtered block recovery and model selection proportions *(continued)*

		Block number recovery								Model selection proportions			
		iid		π		ρ		$\pi\rho$					
ϵ_π	ϵ_ρ	$\mathbb{1}_{\widehat{Q_{1}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi\rho}=3}$	*iid*	π	ρ	$\pi\rho$
0.105	0.175	1	1	1	1	1	1	1	1	**0.657**	0.296	0.028	0.019
	0.210	1	1	1	1	1	1	1	1	0.324	**0.611**	0.037	0.028
	0.245	**0.991**	1	1	1	1	1	1	1	0.185	**0.750**	0.009	0.056
	0.280	1	1	1	1	1	1	1	1	0.046	**0.917**		0.037
	0.000	1	1	1	1	1	1	1	1	**0.815**		0.185	
	0.035	1	1	1	1	1	1	1	1	**0.741**	0.009	0.250	
	0.070	1	1	1	1	1	1	1	1	**0.861**		0.139	
	0.105	1	1	1	1	1	1	1	1	**0.759**	0.056	0.185	
0.140	0.140	1	1	1	1	1	1	1	1	**0.667**	0.157	0.157	0.019
	0.175	1	1	1	1	1	1	1	1	**0.583**	0.269	0.046	0.102
	0.210	1	1	1	1	1	1	1	1	0.278	**0.546**	0.065	0.111
	0.245	1	1	1	1	1	1	1	1	0.111	**0.704**	0.037	0.148
	0.280	1	**0.991**	1	1	**0.991**	1	1	1	0.028	**0.852**	0.009	0.111
	0.000	1	1	1	1	1	1	1	1	**0.630**		0.370	
	0.035	1	1	1	1	1	1	1	1	**0.694**		0.306	
	0.070	1	1	1	1	1	1	1	1	**0.528**	0.009	0.444	0.019

Table B.6: Filtered block recovery and model selection proportions *(continued)*

		Block number recovery								Model selection proportions			
		iid		π		ρ		$\pi\rho$					
ϵ_π	ϵ_ρ	$\mathbb{1}_{\widehat{Q_{1}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi\rho}=3}$	*iid*	π	ρ	$\pi\rho$
	0.105	1	1	1	1	1	1	1	1	**0.574**	0.028	0.389	0.009
0.175	0.140	1	**0.991**	1	1	1	1	1	1	**0.556**	0.130	0.250	0.065
	0.175	1	1	1	1	1	1	**0.991**	1	**0.417**	0.167	0.231	0.185
	0.210	1	1	1	1	1	1	1	1	0.250	**0.380**	0.157	0.213
	0.245	1	1	1	1	**0.991**	1	1	1	0.102	**0.463**	0.019	0.417
	0.280	1	1	1	1	1	1	1	1		**0.694**	0.009	0.296
	0.000	1	1	1	1	1	1	1	1	0.361		**0.639**	
	0.035	1	1	1	1	1	1	1	1	0.343		**0.657**	
	0.070	1	1	1	1	1	1	1	1	0.370	0.009	**0.611**	0.009
	0.105	**0.991**	**0.991**	1	1	1	1	1	1	0.287	0.009	**0.648**	0.056
0.210	0.140	1	1	1	1	1	1	1	1	0.361	0.083	**0.481**	0.074
	0.175	**0.991**	1	1	1	1	1	1	1	0.278	0.167	**0.361**	0.194
	0.210	1	1	1	1	1	1	1	1	0.157	0.231	0.176	**0.435**
	0.245	1	1	1	1	1	1	1	1	0.028	0.380	0.093	**0.500**
	0.280	1	**0.991**	1	**0.991**	1	1	1	1		0.333	0.037	**0.630**
	0.000	1	1	1	1	1	1	1	1	0.148		**0.852**	

Table B.6: Filtered block recovery and model selection proportions *(continued)*

		Block number recovery								Model selection proportions			
		iid		π		ρ		$\pi\rho$					
ϵ_π	ϵ_ρ	$\mathbb{1}_{\widehat{Q_{1}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi\rho}=3}$	*iid*	π	ρ	$\pi\rho$
	0.035	**0.991**	1	1	1	1	1	1	1	0.194		**0.796**	0.009
	0.070	1	1	1	1	1	1	1	1	0.176		**0.824**	
	0.105	1	1	1	1	1	1	1	1	0.139	0.009	**0.778**	0.074
0.245	0.140	1	1	1	1	1	1	1	1	0.120	0.037	**0.657**	0.185
	0.175	1	1	1	1	**0.991**	1	1	1	0.139	0.093	**0.509**	0.259
	0.210	**0.991**	1	1	1	1	1	1	1	0.028	0.093	0.370	**0.509**
	0.245	1	1	1	1	1	1	1	1	0.009	0.102	0.102	**0.787**
	0.280	**0.991**	1	1	1	**0.991**	1	1	1	0.009	0.148	0.019	**0.824**
	0.000	**0.991**	1	1	1	1	1	**0.991**	1	0.009		**0.991**	
	0.035	**0.981**	1	1	1	1	1	1	1			**0.981**	0.019
	0.070	**0.991**	1	1	1	1	1	1	1	0.037		**0.954**	0.009
	0.105	**0.991**	**0.991**	1	**0.991**	1	1	1	**0.991**	0.028	0.009	**0.889**	0.074
	0.140	**0.991**	1	1	1	1	1	**0.991**	1	0.019		**0.889**	0.093
	0.175	**0.991**	1	1	1	1	1	1	1			**0.602**	0.398
	0.210	0.981	1	1	1	1	1	**0.972**	1	0.009		0.324	**0.667**
	0.245	1	1	1	1	1	1	**0.991**	1	0.009	0.009	0.194	**0.787**

Table B.6: Filtered block recovery and model selection proportions *(continued)*

0.280

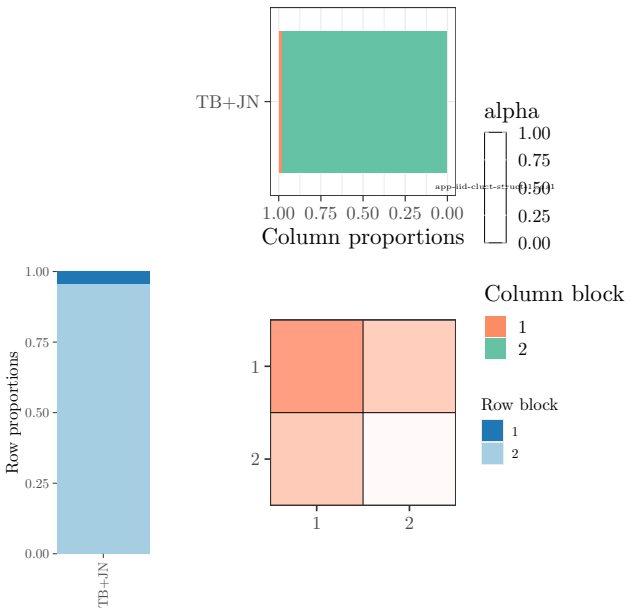
		Block number recovery								Model selection proportions			
		iid		π		ρ		$\pi\rho$					
ϵ_π	ϵ_ρ	$\mathbb{1}_{\widehat{Q_{1}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{iid}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\rho}=3}$	$\mathbb{1}_{\widehat{Q_{1}}_{\pi\rho}=3}$	$\mathbb{1}_{\widehat{Q_{2}}_{\pi\rho}=3}$	*iid*	π	ρ	$\pi\rho$
	0.280	1	**0.991**	1	1	1	1	**0.991**	1		0.019	0.046	**0.935**

Appendix C

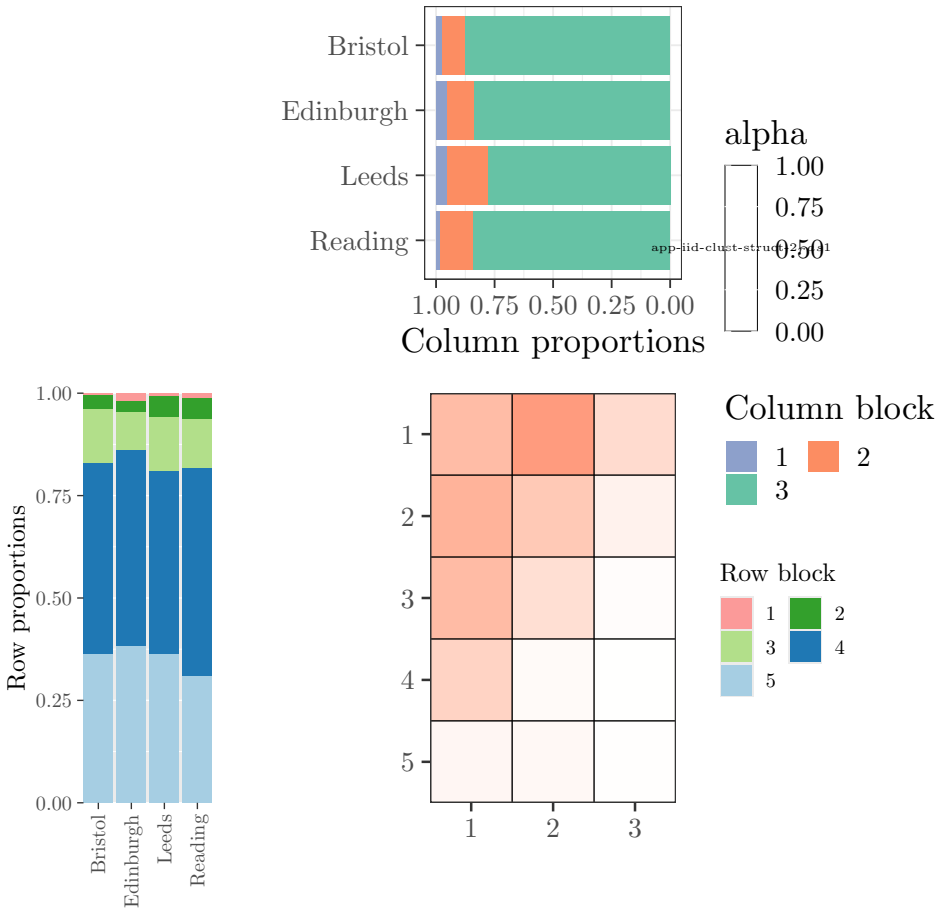
Supplementary for Applications on ecological networks

C.1 Additional information on Clustering of Baldock et al., 2019, 2011

Due to report size limitations we included these plots here as they are not crucial to understand what is going on in the section 4.1. Yet they are useful to confirm the explanation given.



(a) Small collection structure



(b) English networks collection structure

Figure C.1: Structure and mixture proportions for *iid* clustering

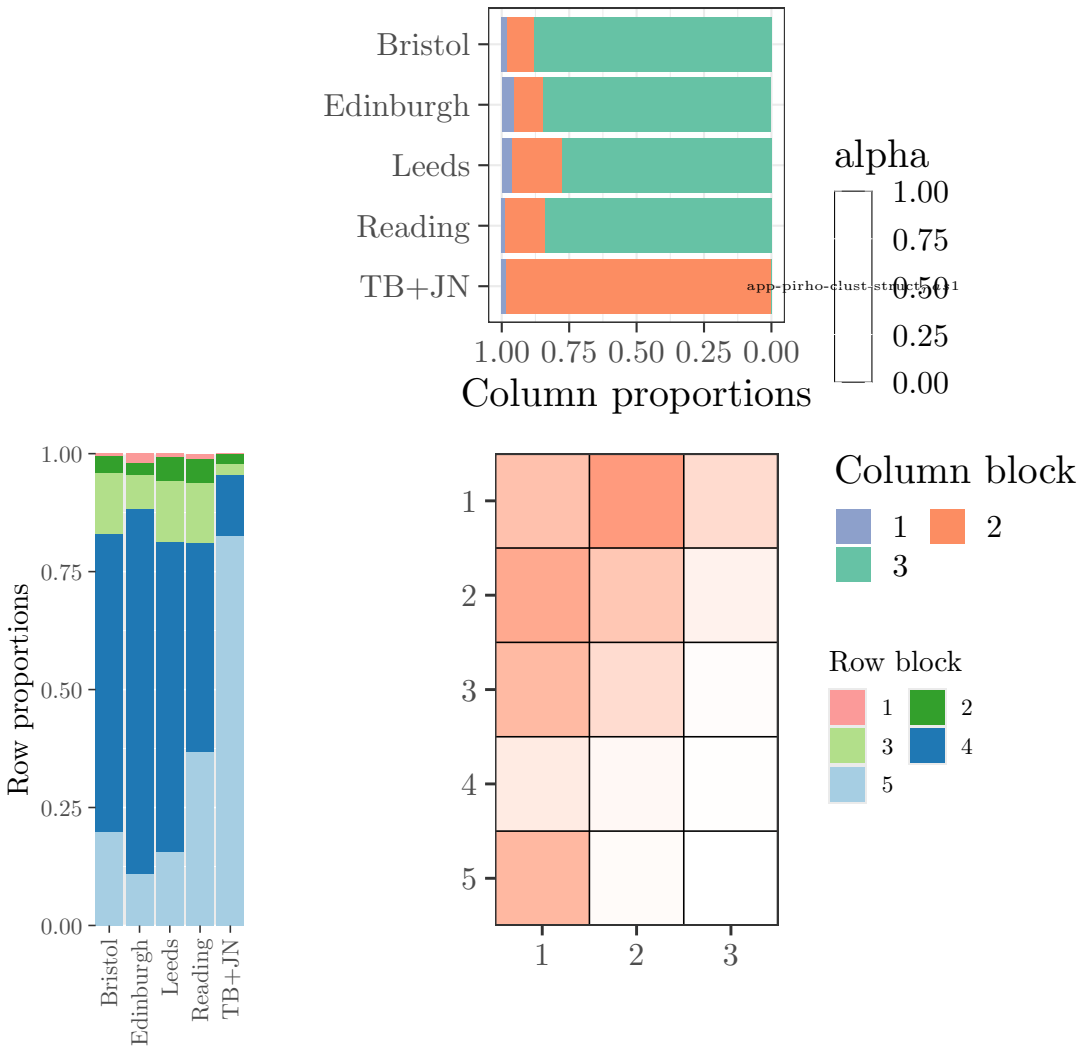


Figure C.2: Structure and mixture proportions for $\pi\rho$ clustering