MIA Paris Saclay

INRAE

AgroParisTech

université PARIS-SACLAY

Some works on the modeling of a collection of networks by stochastic block models

Applications in ecology

Sophie Donnet. MIA PARIS-SACLAY, INRAE.

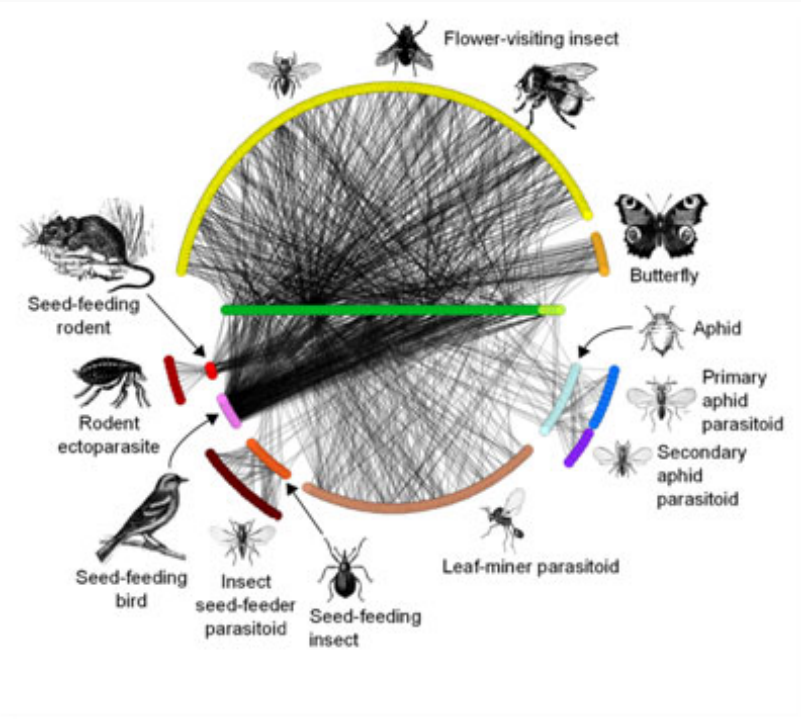
Machine Learning for Life Sciences, 15-17 Nov 2022 Montpellier (France)

About the interest for networks

Fundamental tools in various fields : molecular biology, sociology, ecology ...

In Ecology

- **Vertices** : species (plants or animals)
- **Edges** : predation, pollination, competition...
- **Example of objective** : characterizing the structure of the network because it conditions their robustness to the disappearance of species.



[Pocock et al., 2012]

About the interest for networks

In sociology

- **Vertices** : individuals or organizations
- **Edges** : advice, competition, ...
- **Example of objective** : characterizing the role of individuals in the network, link their role to covariates

Statistical learning

- Qualify the role of nodes : centrality
- Resume the network into a “small“ number of indicators, give a mesoscopic view of its organisation, a summary image.
- Highlight/stress the variability of connection behavior : all the nodes do not play the same role.
- Identify communities
- Huge literature

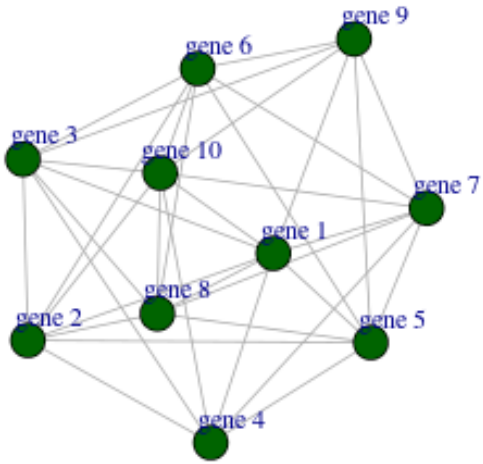
Possible approaches

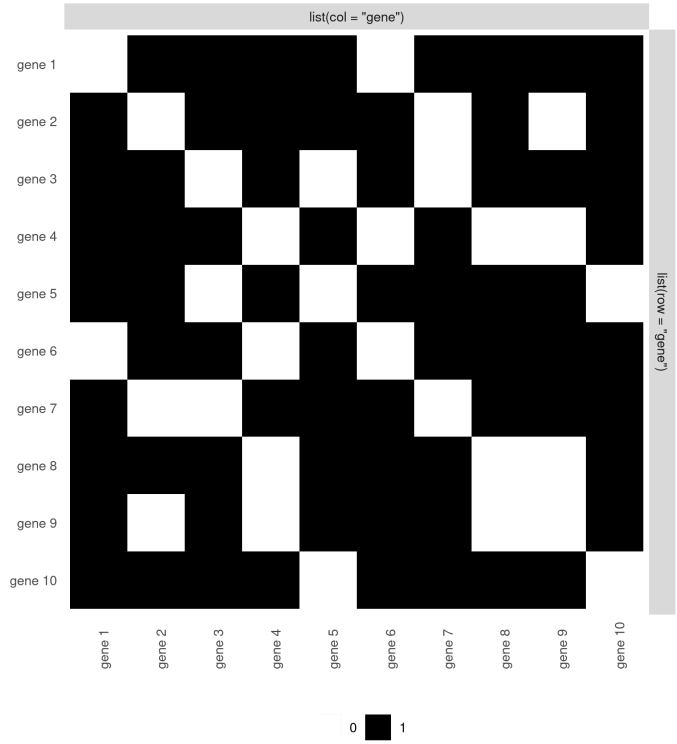
- Classical metrics detecting pre-specified patterns (e.g. modularity, centrality, nestedness...)
- Machine learning tools : autoencoder
- **Probabilistic latent variable models : represent nodes in a smaller space :** . In particular Stochastic Block Models

Some of the people involved

- **MIA Paris Saclay** : J. Aubert, P. Barbillon, J Chiquet, S. Donnet, N. Jouvin J.B. Léger, M.L. Martin-Magniette + students
- **INRAE** : M. Mariadassou (MAIAGE), N. Peyrard (MIAT), N. Verzelen (MISTEA), I. Sanchez (MISTEA), B. Cloez (MISTEA) + students
- **Outside** : C Matias, T. Rebafka, S. Robin, F. Villers (Paris Sorbonne), P. Latouche (Clermont), F. Picard (CNRS), V. Miele (CNRS)...
- Applied collaborators : ANR Econet, Resodiv...

Adjacency matrix Representation





- For any pair of nodes (i, j) (individuals for instance),

$$Y_{ij} = \begin{cases} 1 & \text{if there is an interaction between } i \text{ and } j \\ 0 & \text{otherwise.} \end{cases}$$

- Sometimes $Y_{ij} \in \mathbb{R}$ or $\mathbb{N}$, weighted graph
- Square matrix, symetric or not

A first probabilistic model

- Context : our adjacency Y is the realization of a stochastic process.
- Aim : Propose a stochastic process is able to mimic heterogeneity in the connections.

Erdos Renyi model

(if $Y_{ij} \in \{0, 1\}$)

$$\forall (i, j), \qquad Y_{ij} \sim \mathcal{B}\text{ern}(p)$$

- Homogeneity of the connections : all the nodes play the same role
- No hubs, no community, no nestedness

Stochastic Block Model

- Aim : introduce heterogeneity in the connections
- Tool : introduce blocks of nodes gathering entities that interact roughly similarly in the network

Stochastic Block Model [Snijders, 2001]

Let (Y_{ij}) be an n adjacency matrix

Latent variables

- The nodes $i = 1, \dots, n$ are partitionned into Q clusters
- $Z_i = q$ if node i belongs to cluster (block) q
- Z_i independant variables

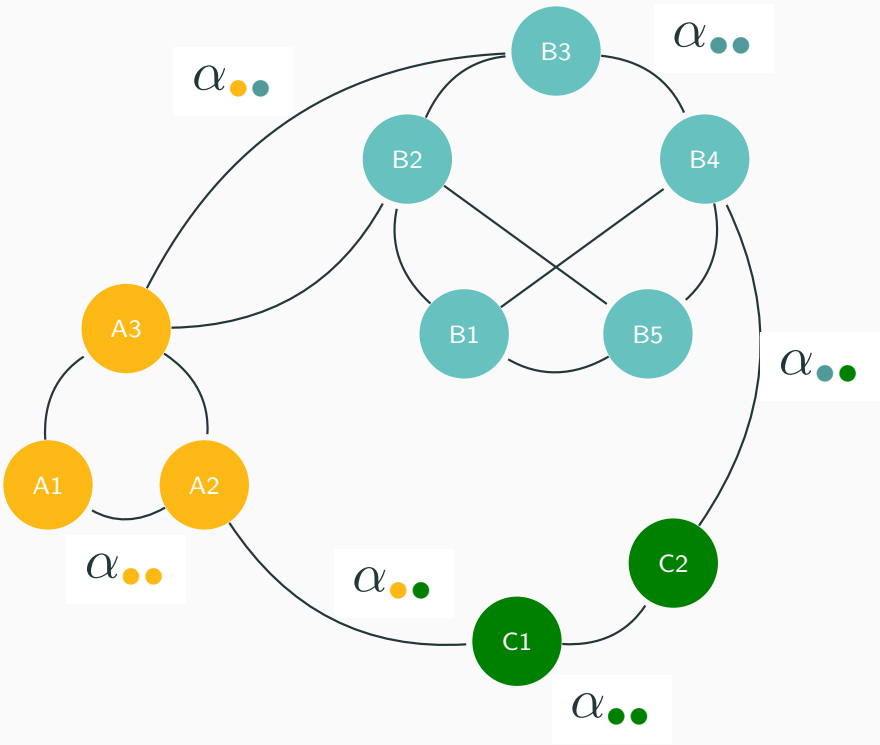
$$\mathbb{P}(Z_i = q) = \pi_q$$

Conditionally to $(Z_i)_{i=1,\dots,n}$...

(Y_{ij}) independant and

$$Y_{ij} | Z_i = q, Z_j = r \sim Bern(\alpha_{qr})$$

Stochastic Block Model : illustration



Parameters

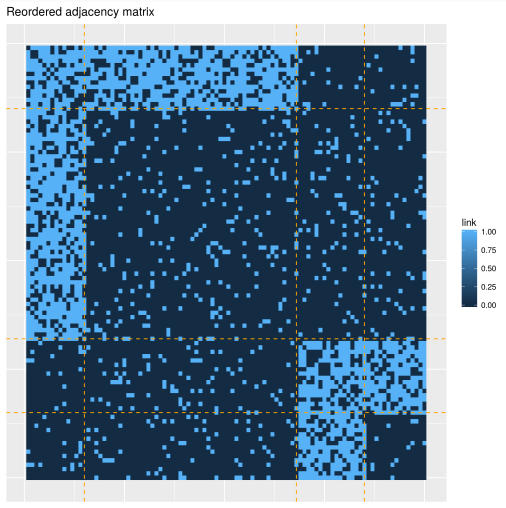
Let n nodes divided into 3 clusters

- $\{\bullet, \bullet, \bullet\}$ clusters
- $\pi_\bullet = \mathbb{P}(i \in \bullet),\ i = 1, \ldots, n$
- $\alpha_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

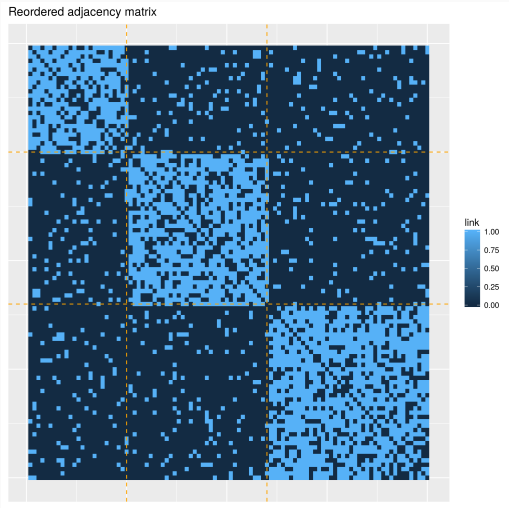
$$\mathbf{Y} \sim \mathrm{SBM}_n(Q, \pi, \alpha)$$

SBM : A great generative model

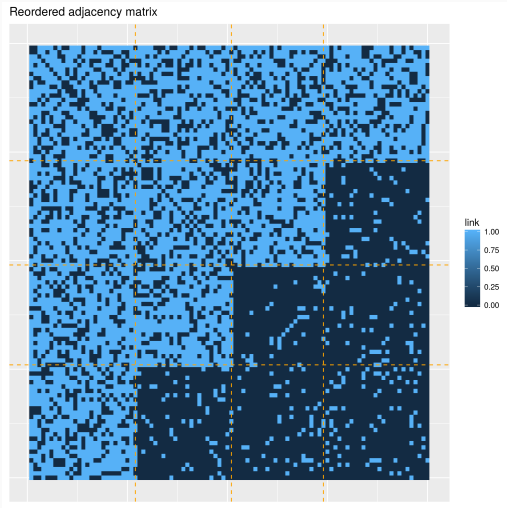
- Generative model : easy to simulate
- Combination of modularity, nestedness, etc...



Hubs



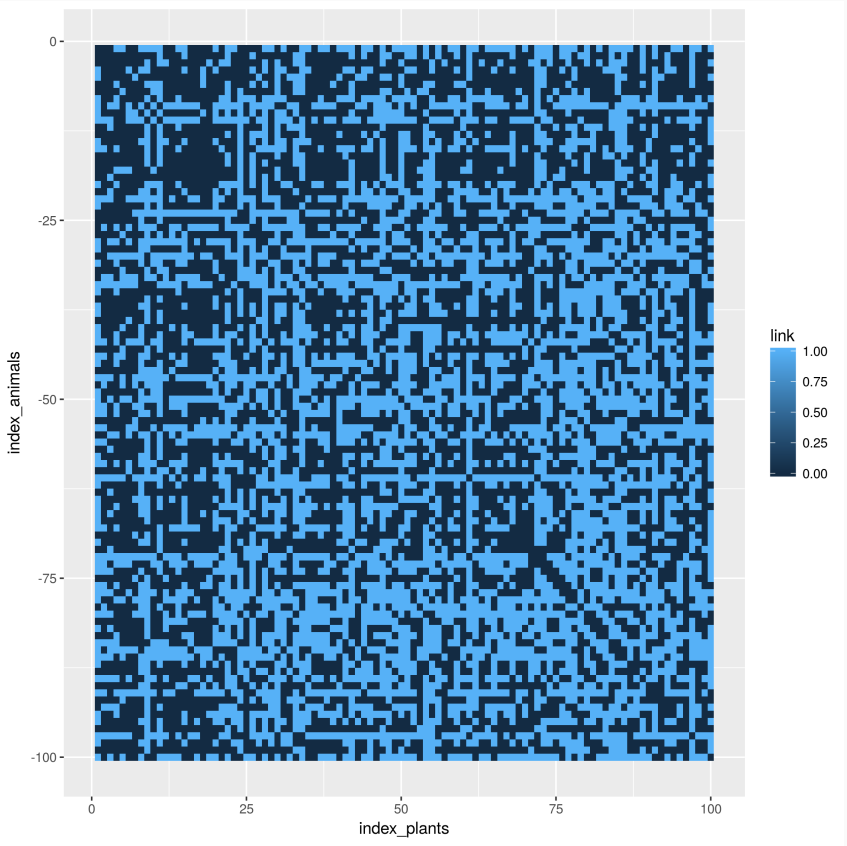
Comunities



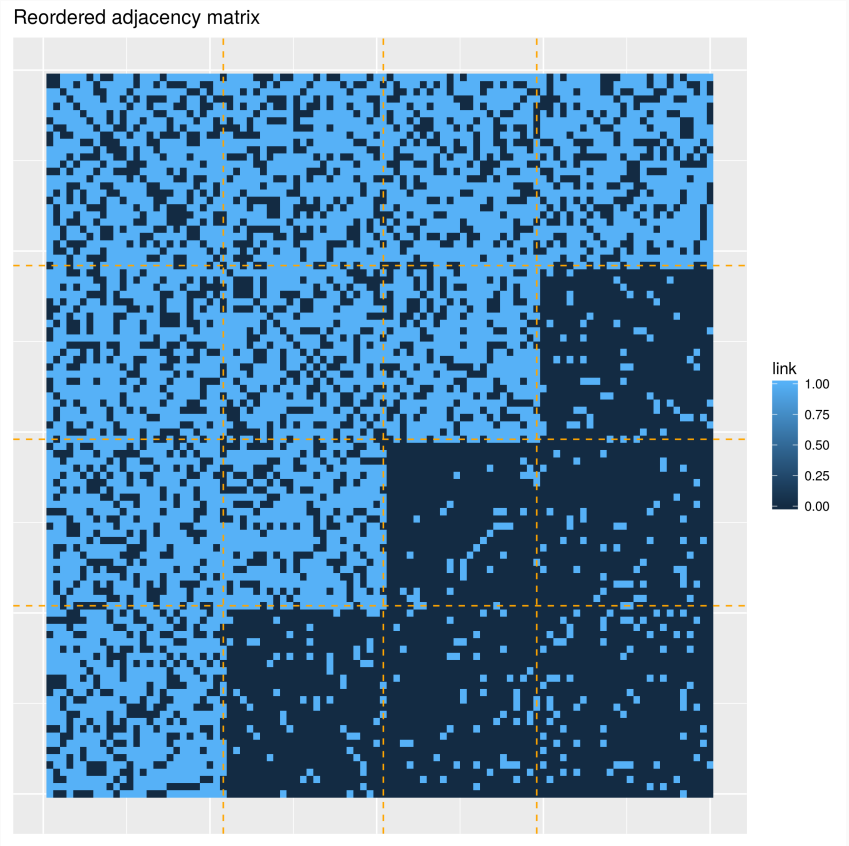
Nested

Statistical inference

From...



to



- Selection of the number of clusters Q for SBM
- Estimation of the parameters $\theta_Q = (\pi, \boldsymbol{\alpha})$ for a given number of clusters Q
- Clustering $\widehat{\mathbf{Z}}$

Likelihood for SBM

Complete likelihood (Y) et (Z)

$$\begin{aligned}
\ell_c(\mathbf{Y}, \mathbf{Z}; \theta) &= p(\mathbf{Y}|\mathbf{Z}; \boldsymbol{\alpha}) p(\mathbf{Z}; \boldsymbol{\pi}) \\
&= \prod_{i,j} f_{\alpha_{Z_i, Z_j}}(Y_{ij}) \times \prod_i \pi_{Z_i} \\
&= \prod_{i,j} \alpha_{Z_i, Z_j}^{Y_{ij}} (1 - \alpha_{Z_i, Z_j})^{1 - Y_{ij}} \prod_i \pi_{Z_i}
\end{aligned}$$

Marginal likelihood (Y)

$$\log \ell(\mathbf{Y}; \theta) = \log \sum_{\mathbf{Z} \in \mathcal{Z}} \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) \,. \tag{1}$$

Marginal likelihood : remark

$$\log \ell(\mathbf{Y}; \theta) = \log \sum_{\mathbf{Z} \in \mathcal{Z}} \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)\,.$$

Remark

$\mathcal{Z} = \{1, \ldots, Q\}^n \Rightarrow$ when Q and n increase, impossible to compute.

Standard tool to maximize the likelihood when latent variables involved : EM algorithm.

From EM to variational EM

Standard EM

At iteration (t) :

- **Step E** : compute

$$\mathcal{Q}(\theta|\theta^{(t-1)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}\left[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)\right]$$

- **Step M** :

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

Limitations of standard EM

- Step E requires the computation of $\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}\left[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)\right]$
- However, once conditioned by par $\mathbf{Y}$, the $\mathbf{Z}$ are not independent anymore : complex distribution if Q and n big.

Variational EM : maximization of a lower bound

Idea : replace the complicated distribution $[\mathbf{Z}|\mathbf{Y}, \theta]$ by a simpler one.

Let $\mathcal{R}_{\mathbf{Y},\tau}$ be any distribution on $\mathbf{Z}$

Central identity

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

Note that :

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) = \log \ell(\mathbf{Y};\theta) \Leftrightarrow \mathcal{R}_{\mathbf{Y},\tau} = p(\cdot|\mathbf{Y};\theta)$$

Variational EM

- Maximization of $\log \ell(\mathbf{Y}; \theta)$ w.r.t. θ replaced by maximization of the lower bound $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau})$ w.r.t. τ and θ.
- **Benefit** : we choose $\mathcal{R}_{\mathbf{Y},\tau}$ such that the maximization calculus can be done explicitly
  - In our case : mean field approximation : neglect dependencies between the (Z_i)

$$P_{\mathcal{R}_{\mathbf{Y},\tau}}(Z_i = q) = \tau_{iq}$$

Variational EM

Algorithm

At iteration (t), given the current value $(\theta^{(t-1)}, \mathcal{R}_{\mathbf{Y},\tau^{(t-1)}})$,

- **Step 1** Maximization w.r.t. τ

$$
\begin{aligned}
\tau^{(t)} &= \arg\max_{\tau \in \mathcal{T}} \mathcal{I}_{\theta^{(t-1)}}(\mathcal{R}_{\mathbf{Y},\tau}) \\
&= \arg\max_{\tau \in \mathcal{T}} \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}} \left[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta^{(t-1)}) \right] + \mathcal{H}\left(\mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z})\right) \\
&= \arg\max_{\tau \in \mathcal{T}} \log \ell(\mathbf{Y}; \theta^{(t-1)}) - \mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y}; \theta^{(t-1)})] \\
&= \arg\min_{\tau \in \mathcal{T}} \mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y}; \theta^{(t-1)})]
\end{aligned}
$$

Variational EM

Algorithm

- **Step 2** Maximization w.r.t. θ

$$\begin{aligned}
\theta^{(t)} &= \arg\max_{\theta} \mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau^{(t)}}) \\
&= \arg\max_{\theta} \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau^{(t)}}}\left[\log \ell_c(\mathbf{Y},\mathbf{Z};\theta)\right] + \mathcal{H}\left(\mathcal{R}_{\mathbf{Y},\tau^{(t)}}(\mathbf{Z})\right) \\
&= \arg\max_{\theta} \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau^{(t)}}}\left[\log \ell_c(\mathbf{Y},\mathbf{Z};\theta)\right]
\end{aligned}$$

In practice

- Really fast
- Strongly depends on the initial values

A penalized likelihood criterion

- Selection of the number of clusters Q
- Integrated Classification Likelihood (ICL) [Biernacki et al., 2000]

$$ICL(\mathcal{M}_{\mathbf{K}}) = \log \ell_c(\mathbf{Y}, \hat{\mathbf{Z}}; \hat{\theta}_Q) - \text{pen}(\mathcal{M}_{\mathbf{K}}) \tag{2}$$

where

$$\hat{Z}_i = \underset{q \in \{1,\dots,Q\}}{\arg\max} \ \hat{\tau}_{iq}. \tag{3}$$

- Integrated Complete Likelihood (ICL)

$$ICL(\mathcal{M}_{\mathbf{K}}) = \mathbb{E}_{p(\cdot|\mathbf{Y},\hat{\theta}_Q)}[\log \ell_c(\mathbf{Y}, \widehat{\mathbf{Z}}; \widehat{\theta}_Q)] - \text{pen}(\mathcal{M}_{\mathbf{K}}) \tag{4}$$

where

$$\text{pen}(\mathcal{M}_{\mathbf{K}}) = \frac{1}{2} \left\{ \underbrace{(Q-1)\log(n)}_{\text{Clust.}} + \underbrace{Q^2 \log(n^2 - n)}_{\text{Conn.}} \right\}$$

Advantages of ICL

- its capacity to outline the clustering structure in networks
- Involves a trade-off between goodness of fit and model complexity
- ICL values : goodness of fit AND clustering sharpness.

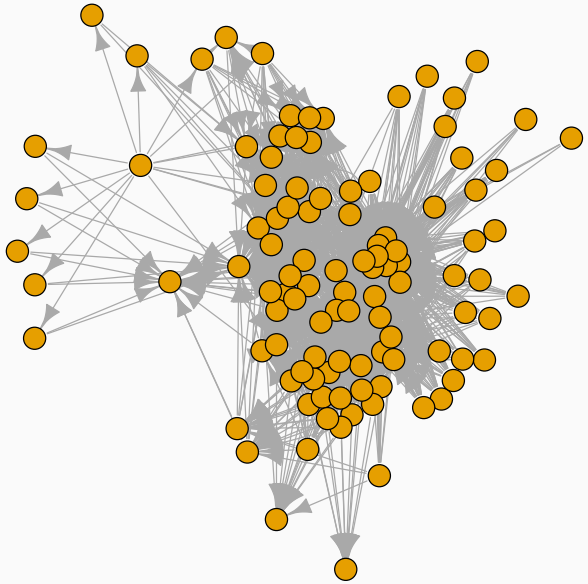
▸ ICL versus BIC

Implementation and theorical guaranties

- Implementation
  - Going trough the models and initiate VEM at the same time
  - Stepwise procedure
  - R-packages : blockmodels and sbm
    https://grosssbm.github.io/sbm/
- Theorical properties
  - Identifiability and a first consistency result by [Celisse et al., 2012]
  - Consistency of the posterior distribution of the latent variables
    [Mariadassou and Matias, 2015]
  - Consistency and properties of the variational estimators
    [Bickel et al., 2013]

▸ Extension to bipartite and multipartite

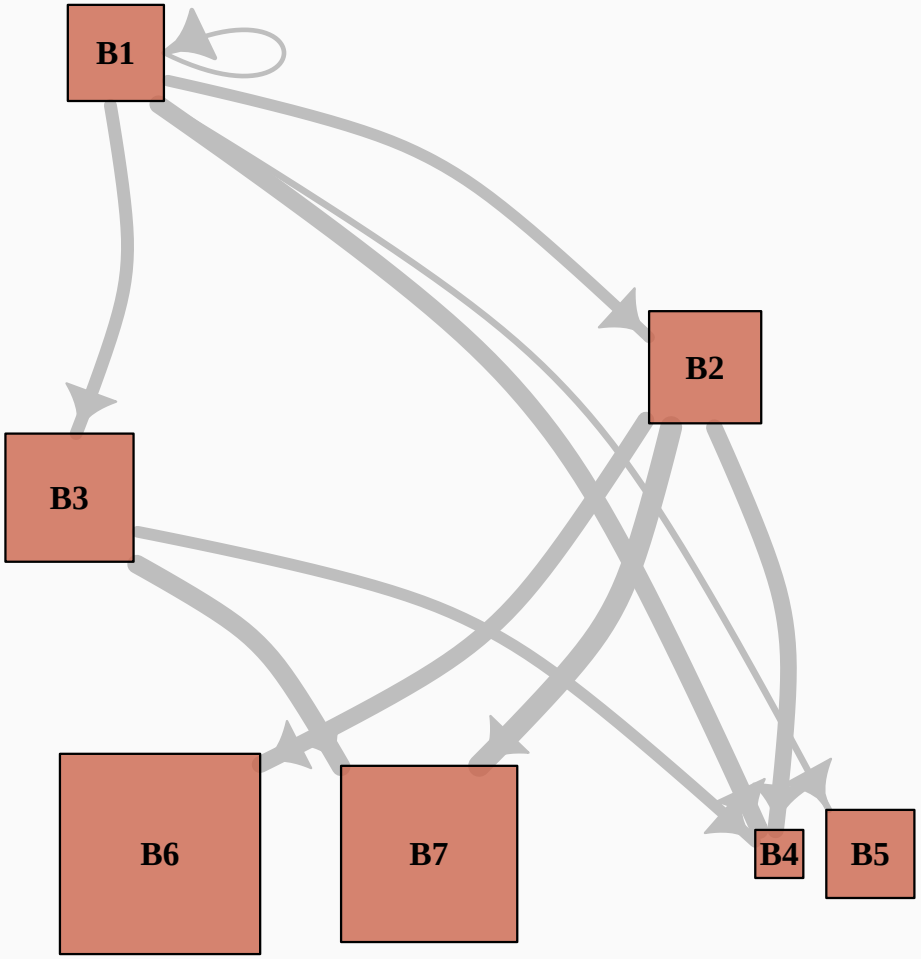
Application on a Chilean foodweb

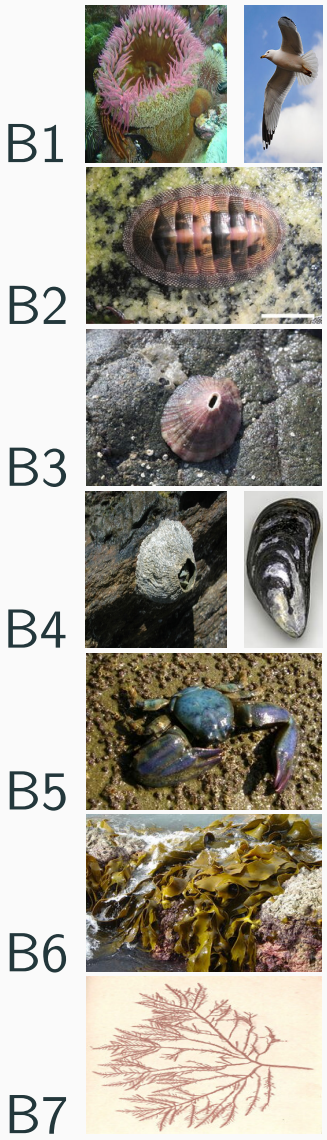


- Intertidal zone of the Chilean Pacific coast
- 106 animal or plant species, sessile or mobile
- 1362 trophic interactions
- [Kéfi et al., 2016]

Application on Chilean

7 blocs





Application on Chilean foodweb

- Schematic representation (inspired by [Picard et al., 2009])
- Left : each vertex is a block and the thickness of the edges represents the probability of interactions between each block (above the 0.1 threshold, for clarity)
- Right : type of species representative of each block. From top to bottom : anemone and gull (B1), chiton (B2), *Fissurella* (B3), *Balanus* and mussel (B4), crab (B5), *Laminariale* (B6) and red algae (B7)

Studying the blocks

- B1 :
  - gather the "super-predators" (top of the trophic chain) which have no predators except some rare trophic links between them
  - wide taxonomic variability, including diverse species such as the anemone or the gull
- ...
- B6 and B7 contain basal algal species, including brown algae and red algae respectively, and which are resources for various mollusks.

Finally

- SBM allows to summarize the complexity induced by the observation of more than a thousand interactions.
- The interpretation of its parameters (the probabilities of interactions between each block) allows a synthetic description of the ecosystem,
- Interpretation of the blocks with exogenous information such as taxonomy and ecological traits.
- To go really further : use this representation to compute a robustness to species disappearance [Chabert-Liddell et al., 2022]

1. Introduction

2. Stochastic block models for a single network

3. Extensions

4. Collection of networks

5. Take home message

Other emission distribution

- For count data

$$Y_{ij}|Z_i = k, Z_j = \ell \sim \mathcal{F}(\alpha_{k\ell})$$

  where $\mathcal{F}$ can be a Poisson, a zero inflated Poisson, a negative binomial distribution [Mariadassou et al., 2010]

- If I have covariates on the each pair of nodes ?

$$Y_{ij}|Z_i = k, Z_i j = \ell \sim \mathcal{F}(\alpha_{kl} + x_{ij}\beta)$$

  The blocks explain the residual structure once the covariates avec been taken into account.

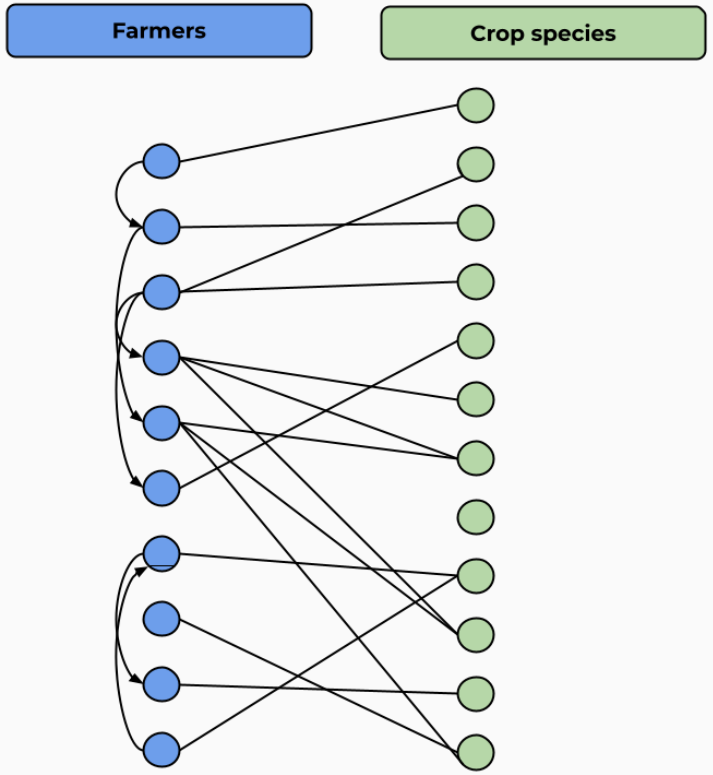
- See the vignette (fungus - tree) of the R Package sbm
- Take into account the way the network was sampled (snow ball effect, etc...) : missSBM + [Tabouy et al., 2020]

Bipartite Networks

If the nodes are divided into two groups

- Plant-pollinators
- Fungi-trees
- Drugs - Side effects
- Crop Species- Farmers

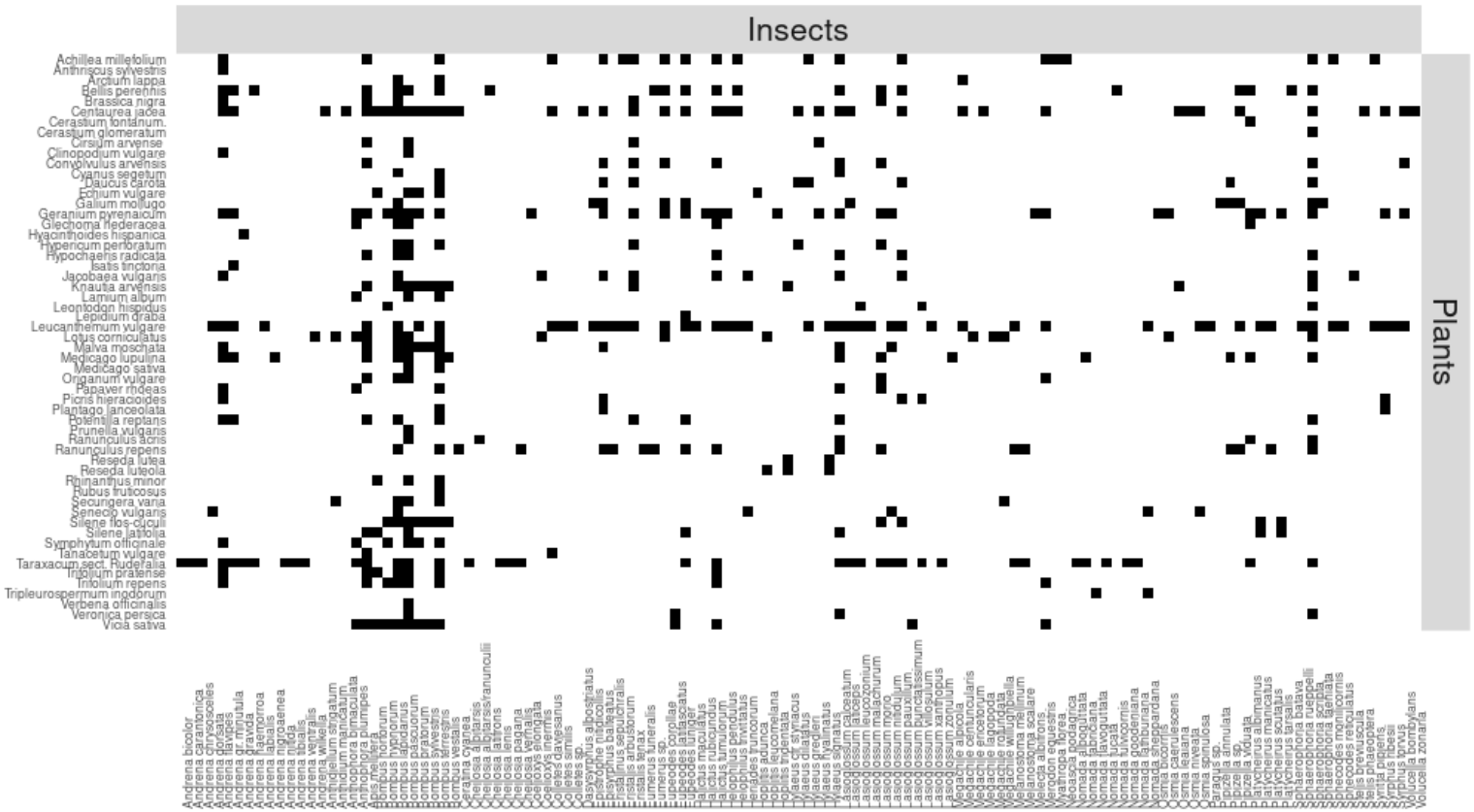
Interactions only between elements of the two groups.



Plants/Pollinators

From ....

Data from [Fisogni et al., 2020]



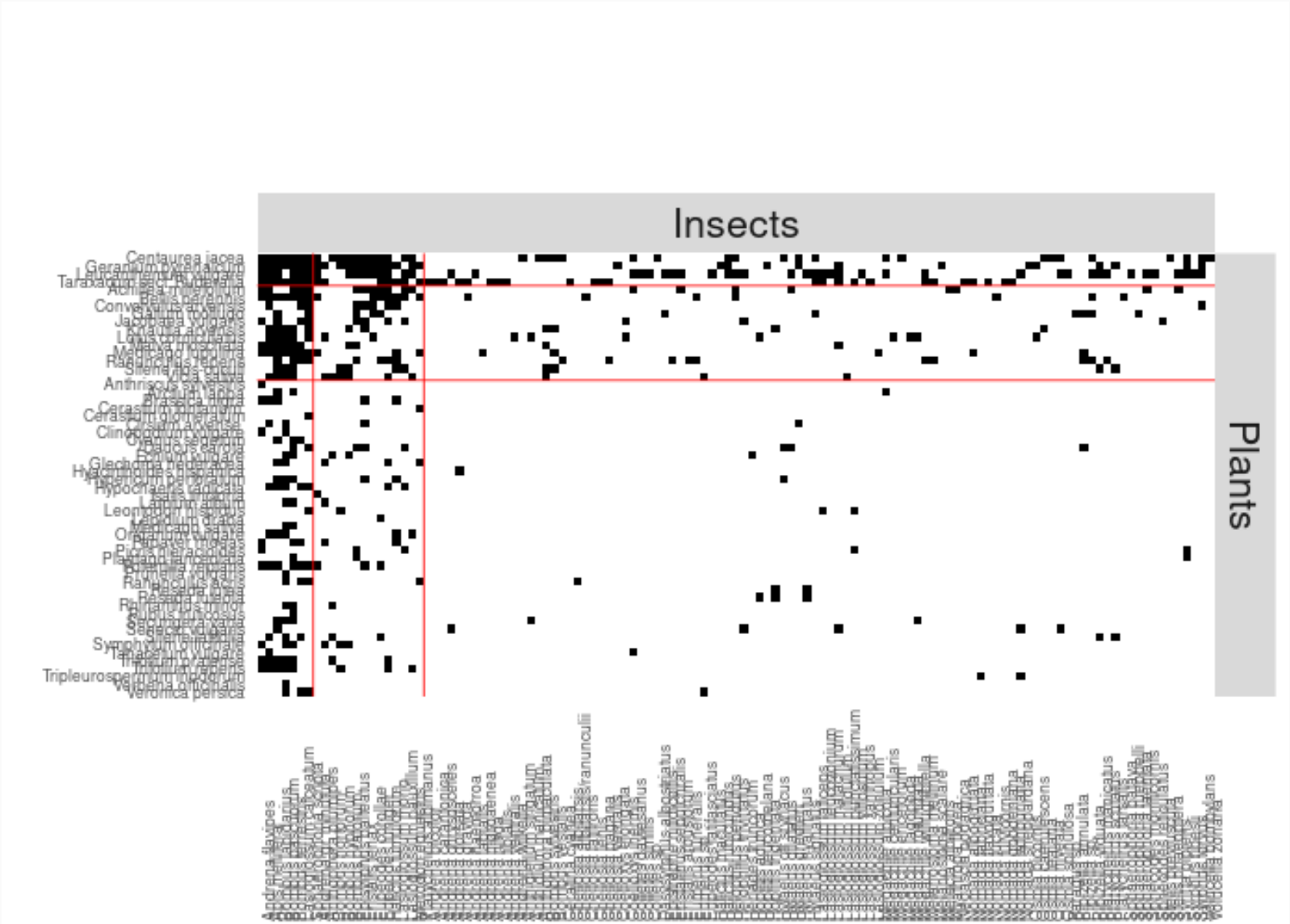
Plants/Pollinators

Introduce blocks of nodes of each type : biclustering

$$
\begin{aligned}
Y_{ij}|U_i = k, V_j &= \ell \sim \mathcal{F}(\alpha_{k\ell}) \\
P(U_i = k) &= \pi_k^U \\
P(V_i = \ell) &= \pi_\ell^V
\end{aligned}
$$

Plants/Pollinators

... to

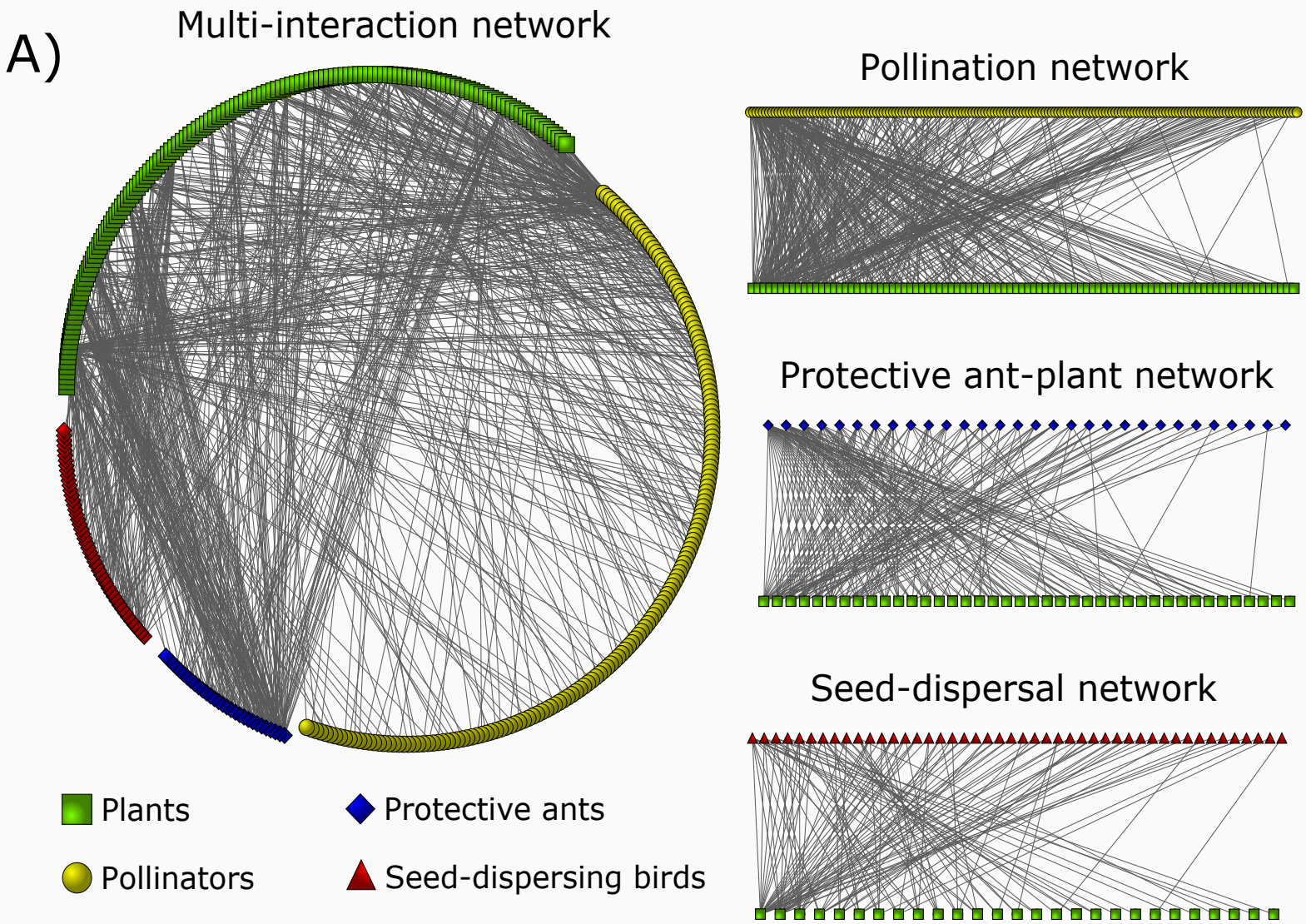


Multipartite networks

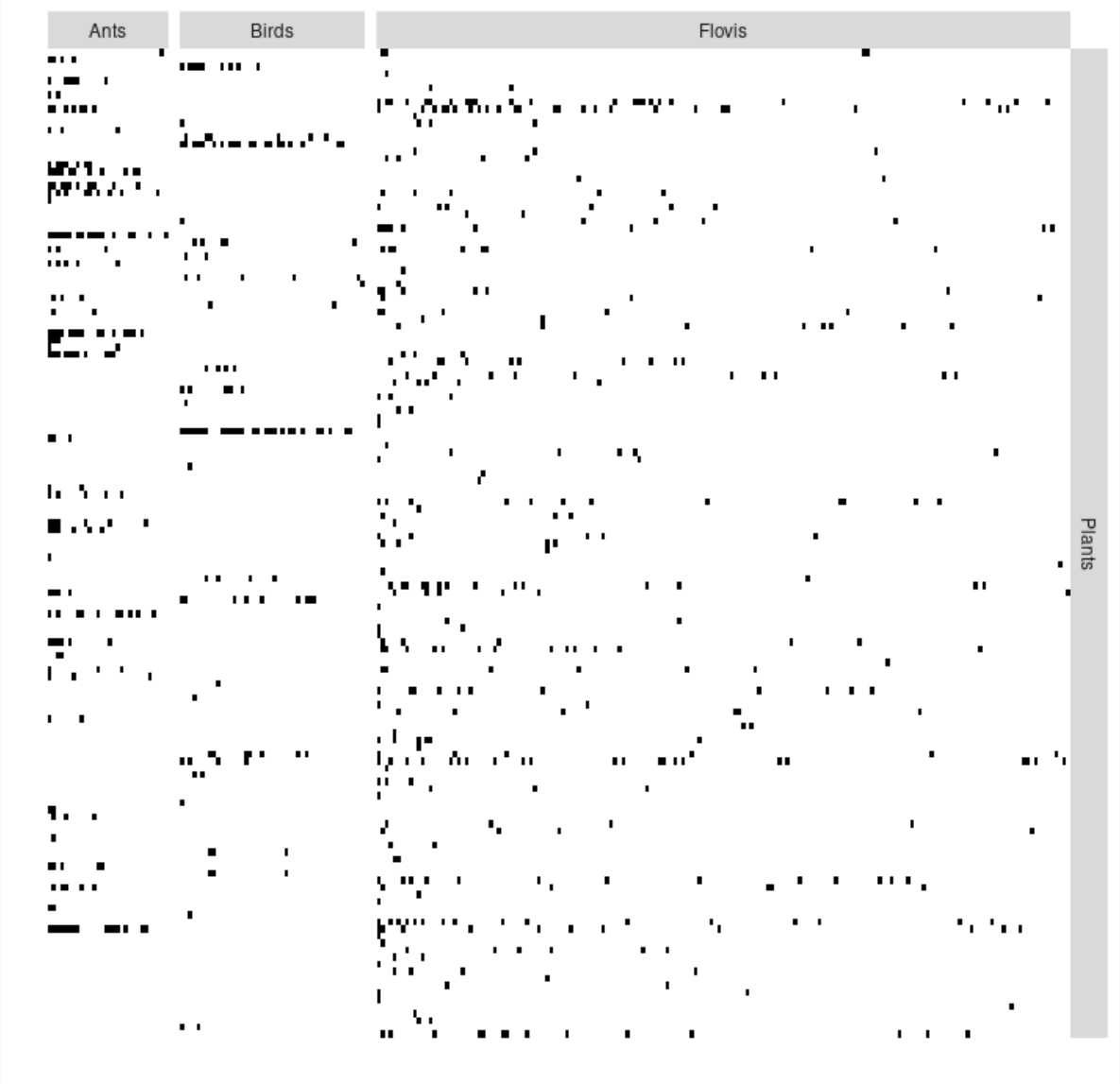
- Divided into 4 functional groups : plants, pollinators, ants, seed-dispersing birds
- Edge between plant i and animal j = an individual of animal specie j has been observed at least once in interaction (pollination, protection, eating seeds) with a plant of specie i.
- Observations made along the Mexican Coast by Wesley Dattilo (INECOL, Xalapa, Mexico)

Dattilo's multipartite network

Submitted to Proceedings of the Royal Society B: For Review Only



Multipartite matrix in ecology



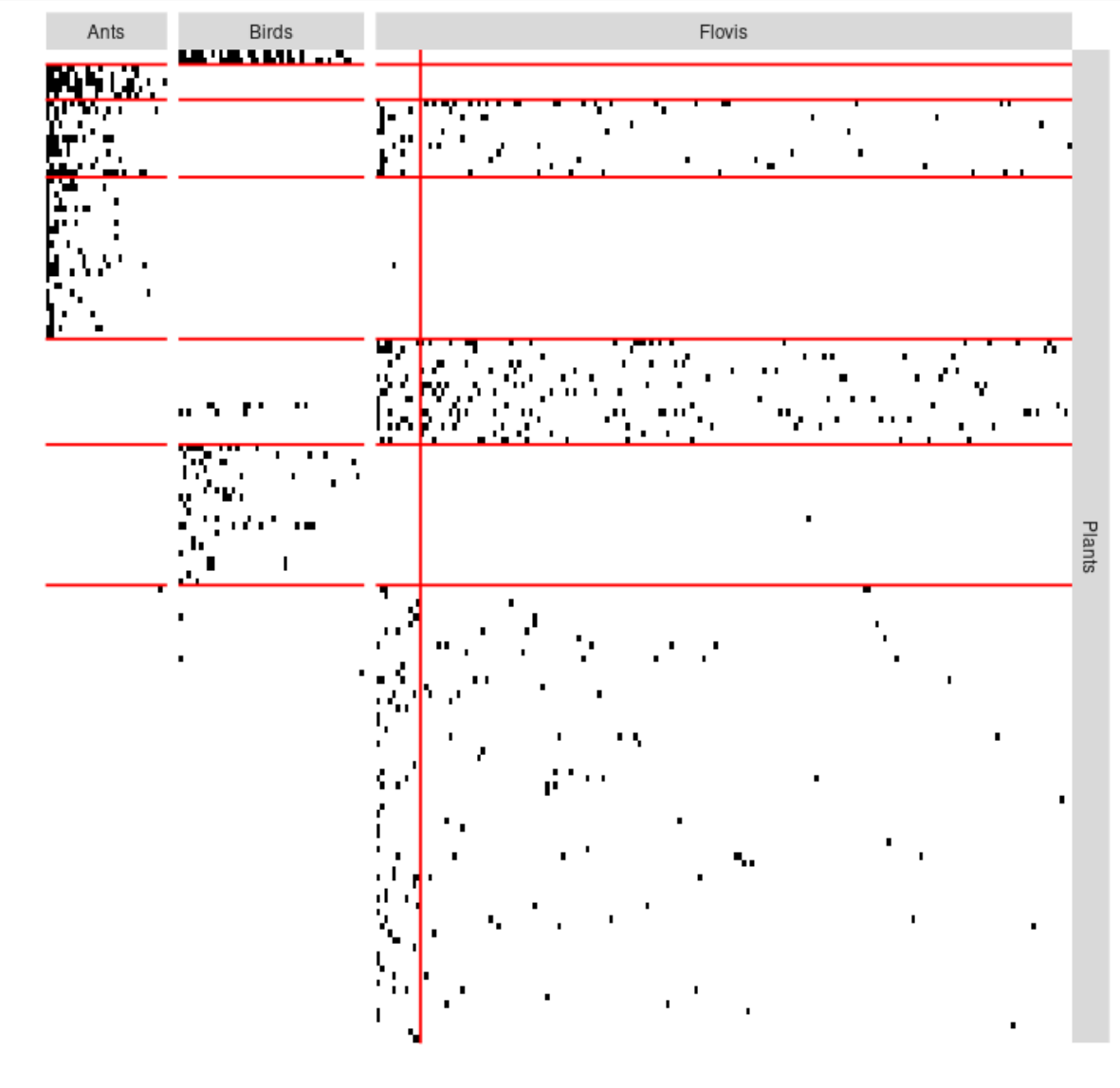
[Bar-Hen et al., 2018]

Results : a mesoscopic view of the network

With our model and model selection
(a few minutes)

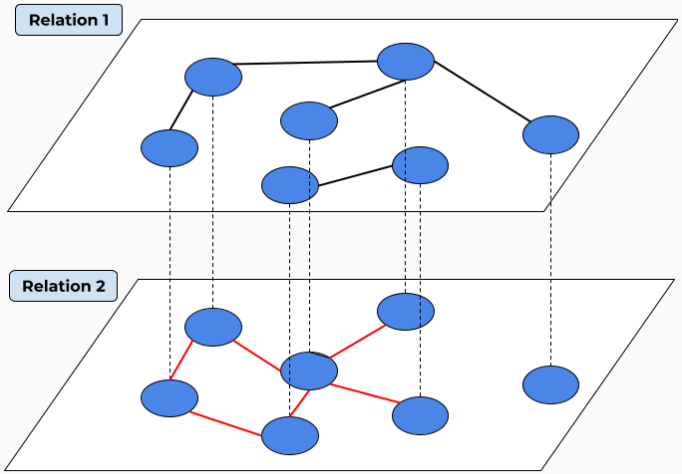
- 7 blocks of plants
- 2 blocks of flower visitors
  (pollinators)
- 1 block of birds
- 2 blocks of ants

Re-ordered matrix



Multiplex networks

I want to study mutualism and competition (or advices and competition at the same time) at the same time ?



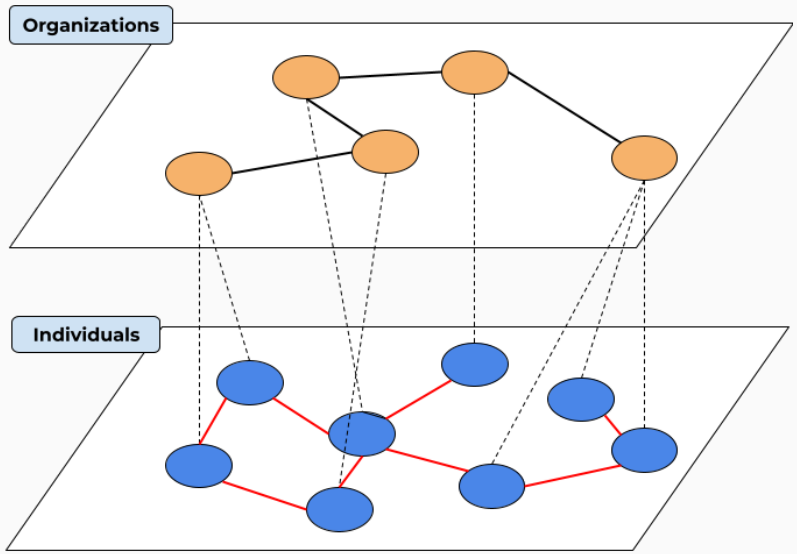
Multiplex network [Barbillon et al., 2016][Kéfi et al., 2016]

Test dependency between the two levels

Other extensions

In the recent years, interest for analyzing jointly a collection of networks (or multilayer networks). See [Bianconi, 2018, Pilosof et al., 2017]

- **Time (or space) -varying** networks : same entities observed interacting along time [Matias and Miele, 2017]
- **Multilevel** networks : organizations and individuals [Chabert-Liddell et al., 2021]



Package sbm

Comprehensive R package available on CRAN and Github gathering several block models and there in references with vignettes.

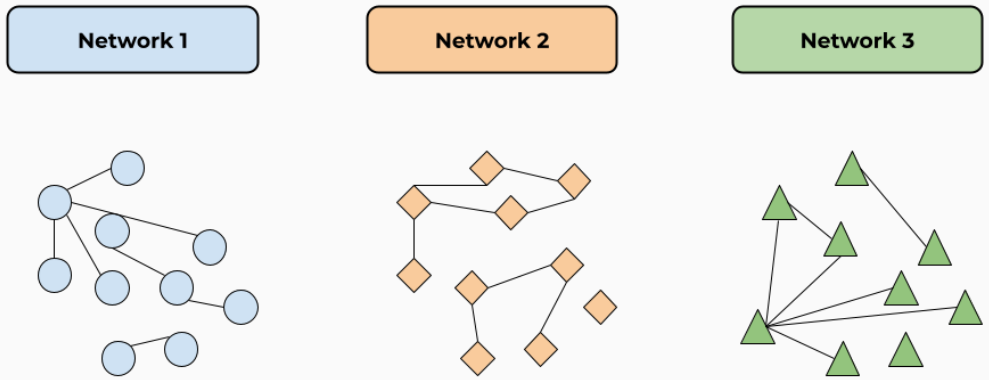
https://grosssbm.github.io/sbm/

Developed by J. Chiquet, P. Barbillon, S. Donnet, J.B. Léger, I. Sanchez, etc...

Collection of networks : consensus in the structure

Objectives

Looking for commun patterns in networks involving non-common sets of nodes

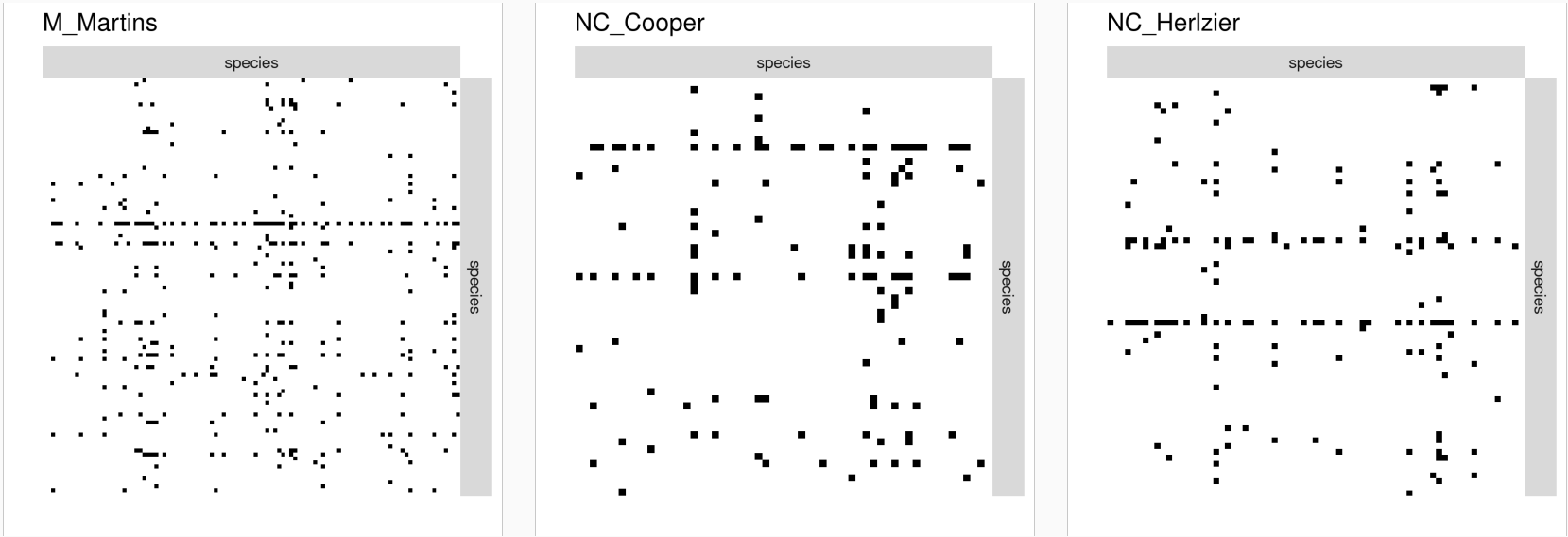


Applications

- Compare the structure of ecological networks
- Compare sociological networks : advices between lawyers, researchers or priests

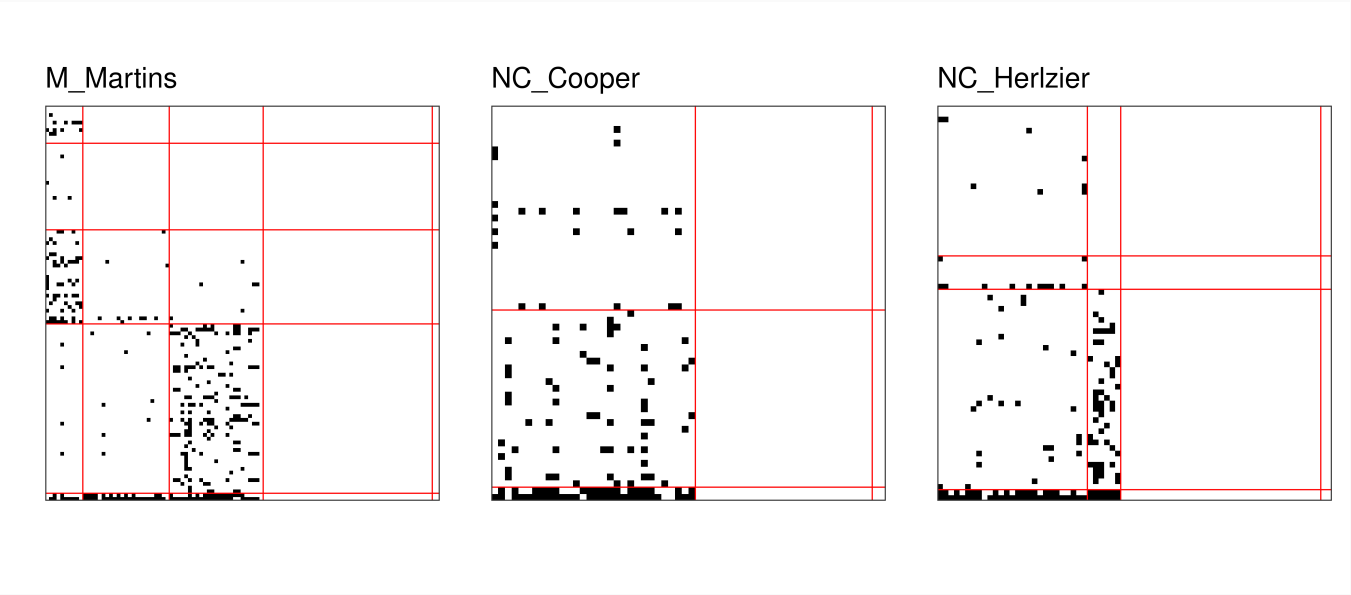
Three foodwebs

- Pine-firest stream food webs issued from Maine, North-Caroline and Nez-Zealand [Thompson and Townsend, 2003]
- Involve respectively 105, 58 and 71 species.
- $Y_{ij} = 1$ if i is eaten by j. Directed relation



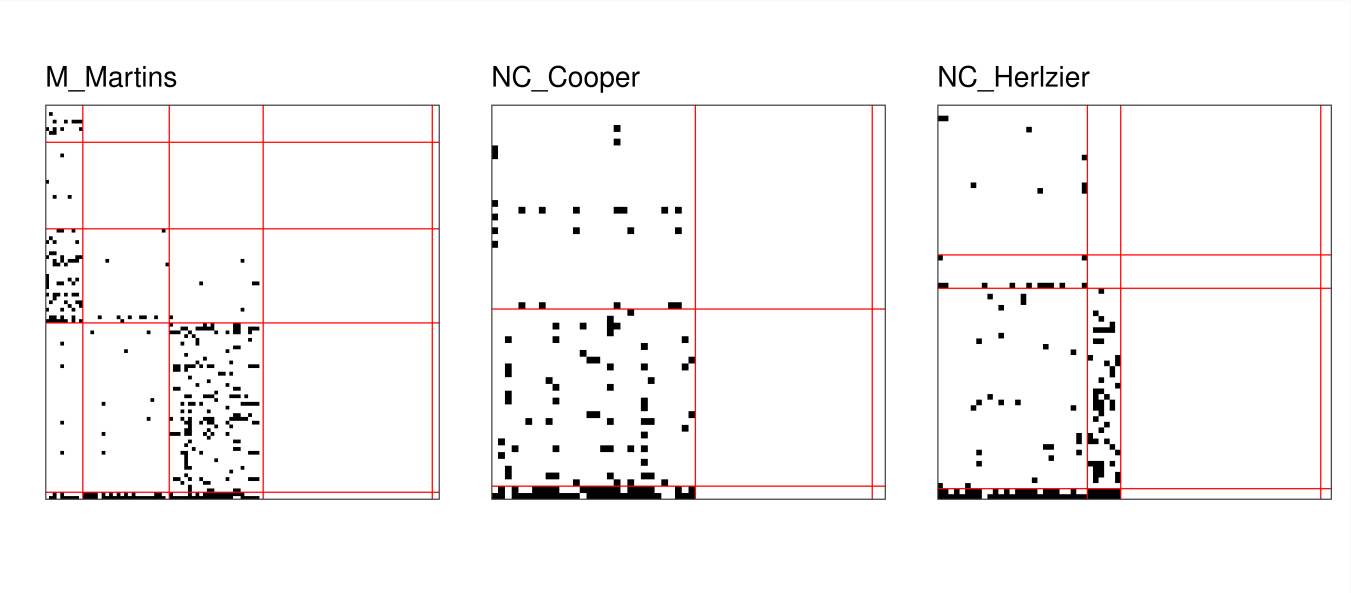
- Look for similarities and differences between network structures.

Separate SBMs



- Fitted SBM on each separately
- Reordered the matrices following the blocks
- Label the blocks following the average out-degrees order

Separate SBMs



- Two bottom groups in each matrix are basal species : eaten by many species and not eating anybody.
-
  - **Martins** : has a separation into 5 blocks, the third one is a medium trophic level, which preys on basal species and is highly preyed by species of the 1st block.
  - **Cooper**. Higher trophic levels grouped together in the same block (lack of statistical power).
  - **Herlzier** : higher trophic level is separated into 2 blocks determined on how much they prey on the less preyed basal block.

Towards a joint modeling of the networks

- Need to model jointly the networks
- Identify the groups playing the same role through out the networks, with an unsupervised strategy.
- Let $(\mathbf{Y}^m)_{m=1,\dots,M}$ denote the collection of networks each involving n_m nodes.
- $(\mathbf{Y}^m)$ independent.
- $$\mathbf{Y}^m \sim \mathrm{SBM}_{n_m}(Q^m, \boldsymbol{\pi}^m, \boldsymbol{\alpha}^m)$$
- Conditions on the parameters $(\boldsymbol{\pi}^m)_{m=1,\dots,M}$ and $(\boldsymbol{\alpha}^m)_{m=1,\dots,M}$

First naive model

iid-colSBM

$$\mathbf{Y}^m \sim \mathrm{SBM}_{n_m}(Q, \boldsymbol{\pi}, \boldsymbol{\alpha})$$

with $\pi_q > 0 \ \forall q \in \{1, \ldots, Q\}$ and $\sum_{q=1}^{Q} \pi_q = 1$.

- $(Q-1) + Q^2$ unknown parameters, M clustering
- Too strict to be applied to the Thomson's dataset

A first relaxed model : π-colSBM

Same structure of connection α, specific proportions of blocks in each network

π-colSBM

$$\mathbf{Y}^m \sim \text{SBM}_{n_m}(Q, \boldsymbol{\pi}^m, \boldsymbol{\alpha})$$

On the block proportions

- $\pi_q^m \geq 0$
- If $\pi_q^m = 0$ then block q is not represented in network m

π-colSBM : different proportions

$M = 2$ networks

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{12} & \alpha_{22} & \alpha_{23} \\ \alpha_{13} & \alpha_{23} & \alpha_{33} \end{pmatrix} \qquad \begin{array}{l} \pi^1 = [.25, .25, .50] \\ \pi^2 = [.20, .50, .30] \end{array}.$$

- Same connection structure between blocks
- Different block proportions
- $2 \times (3 - 1) + 3^2 = 15$ parameters.

π-colSBM : nested structures

$\pi_q^m \geq 0$

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{12} & \alpha_{22} & \alpha_{23} \\ \alpha_{13} & \alpha_{23} & \alpha_{33} \end{pmatrix} \qquad \begin{array}{l} \pi^1 = [.25, .25, .50] \\ \pi^2 = [.40,\ 0\ , .60] \end{array}.$$

- Blocks 1 and 3 are represented in the two networks while block 2 only exists in network 1.
- $3 - 1 + 3 - 2 + 3^2 = 14$ parameters

π-colSBM : partially nested structures

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \cdot \\ \alpha_{31} & \cdot & \alpha_{33} \end{pmatrix} \qquad \begin{array}{l} \pi^1 = [.25, .75, \ 0\ ] \\ \pi^2 = [.40, \ 0\ , .60] \end{array}.$$

- The two networks share block 1 (for instance super predators or basal species)
- The remaining nodes of each network not equivalent in terms of connectivity.
- Blocks 2 and 3 never interact because their elements do not belong to the same network and so α_{23} and α_{32} are not required to define the model.
- $(2-1)+(2-1)+7 = 11$ parameters.

Number of parameters

Let S be the support $M \times Q$ matrix such that

$$S_{mq} = \begin{cases} 1 & \text{if } \pi_q^m > 0 \\ 0 & \text{otherwise .} \end{cases}$$

Then,

$$Nb(\pi\text{-}colSBM) = \sum_{m=1}^{M} \left(\sum_{q=1}^{Q} S_{qm} - 1 \right) + \sum_{q,r=1}^{Q} \mathbf{1}_{(S'S)_{qr} > 0}$$

Varying density model : δ-colSBM

δ-colSBM

$$\mathbf{Y}^m \sim \text{SBM}_{n_m}(Q, \boldsymbol{\pi}, \delta^m \boldsymbol{\alpha})$$

with $\pi_q > 0$,

- M networks exhibit similar intra- and inter blocks connectivity patterns but with proper densities.
- δ^m be a density parameter, specific to each network. $\delta^1 = 1$.
- Mimics differences of effort sampling or abundances
- $(Q-1) + Q^2 + (M-1)$ parameters.

Varying density and block proportion model

$\delta\pi$-colSBM

$$\mathbf{Y}^m \sim \text{SBM}_{n_m}(Q, \boldsymbol{\pi}^m, \delta^m \boldsymbol{\alpha})$$

with $\pi_q^m \geq 0$

- Most flexible model
- $Nb(\pi\text{-}colSBM) + (M - 1)$ parameters.

Summary

M independent networks.

$$\mathbf{Y}^m \sim \text{SBM}(Q^m, \boldsymbol{\pi}^m, \boldsymbol{\alpha}^m)$$

Model name	Block prop.	Connexion param.	Nb of param.
iid-colSBM	$\pi_q^m = \pi_q,\ \pi_q > 0$	$\alpha_{qr}^m = \alpha_{qr}$	$(Q-1) + Q^2$
π-colSBM	$\pi_q^m,\ \pi_q^m \geq 0$	$\alpha_{qr}^m = \alpha_{qr}$	$\leq M(Q-1) + Q^2$
δ-colSBM	$\pi_q^m = \pi_q,\ \pi_q > 0$	$\alpha_{qr}^m = \delta^m \alpha_{qr}$	$(Q-1) + Q^2 + (M-1)$
$\delta\pi$-colSBM	$\pi_q^m,\ \pi_q^m \geq 0$	$\alpha_{qr}^m = \delta^m \alpha_{qr}$	$\leq M(Q-1) + Q^2 + M - 1$
sep-SBM	$\pi_q^m,\ \pi_q^m > 0$	α_{qr}^m	$\sum_{m=1}^M (Q_m - 1) + Q_m^2$

Identifiability

Demonstrated for the most complex SBM, upto label switching of the blocks and permutation of the networks, under light conditions.

For π-colSBM, let us define $\mathcal{Q}_m = \{q \in \{1, \dots, Q\} | \pi_q^m > 0\}$.

1. $\forall m : n_m \geq 2|\mathcal{Q}_m|$
2. $(\alpha \cdot \pi^m)_q \neq (\alpha \cdot \pi^m)_r$ for all $(q \neq r) \in \mathcal{Q}_m^2$
3. $\forall q = 1, \dots, Q, \quad \exists m : q \in \mathcal{Q}_m$
4. Each diagonal entry of α is unique

Inference

VEM algorithm

- Direct extension of VEM previously described for *iid*-coISBM and π-coISBM
- Less obvious with $\delta_m \alpha$: M step not explicit.

Model selection

ICL can be directly extended for *iid*-colSBM and the δ-colSBM

$$\begin{aligned} ICL(Q) &= \mathcal{I}(\hat{\tau}, \hat{\boldsymbol{\theta}}) - \frac{Q-1}{2} \log\left(\sum_{m \in \mathcal{M}} n_m\right) \\ &\quad - \frac{1}{2}\left(\frac{Q(Q+1)}{2} + \nu(\delta)\right) \log\left(\sum_{m \in \mathcal{M}} \frac{n_m(n_m-1)}{2}\right), \quad (5) \end{aligned}$$

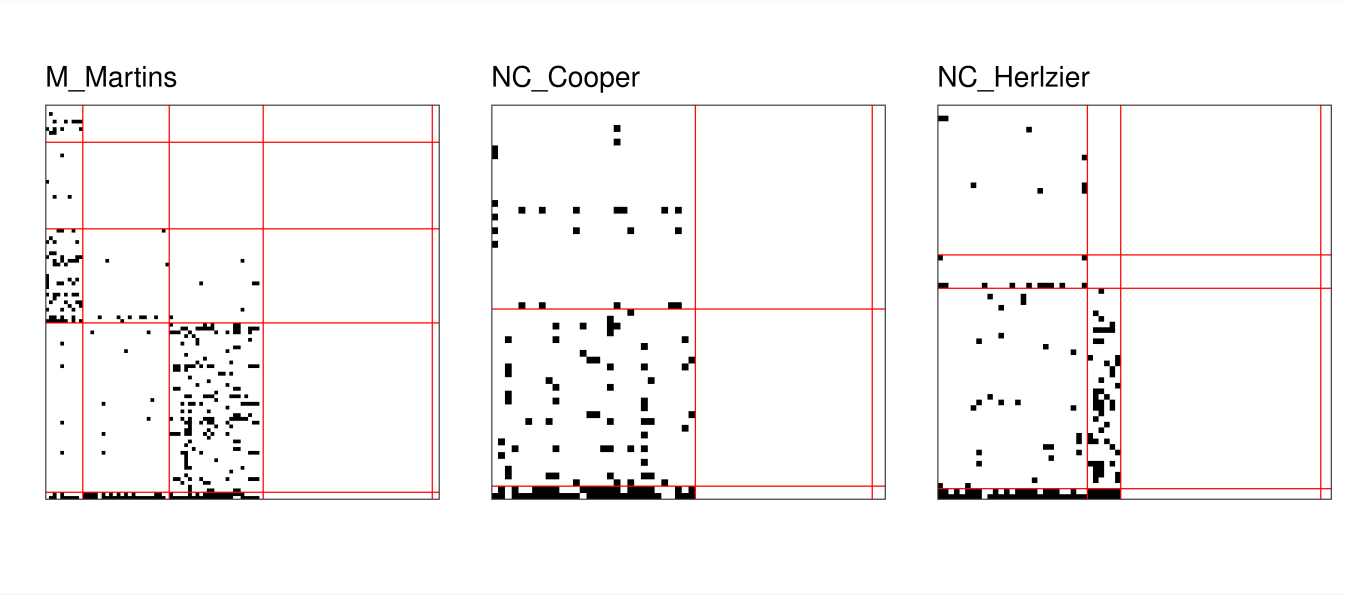
where $\nu(\delta) = M - 1$ for $\delta colSBM$ and 0 otherwise.

Model selection

- For *iid*-colSBM and the δ-colSBM
- π_q^m possibly null. Asymptotic approximation do not hold
- Each couple (Q, S) defines a model.

$$\begin{aligned} ICL(Q,S) &= \mathcal{I}(\hat{\tau}, \hat{\boldsymbol{\theta}}) - \sum_{m=1}^{M} \frac{|\mathcal{Q}_m| - 1}{2} \log(n_m) - \\ &\quad \frac{1}{2} \left(\sum_{q,r=1}^{Q} \mathbf{1}_{(S'S)_{qr}>0} + \nu(\delta) \right) \log \left(\sum_{m=1}^{M} \frac{n_m(n_m-1)}{2} \right) \end{aligned} \quad (6)$$

Application on the foodwebs



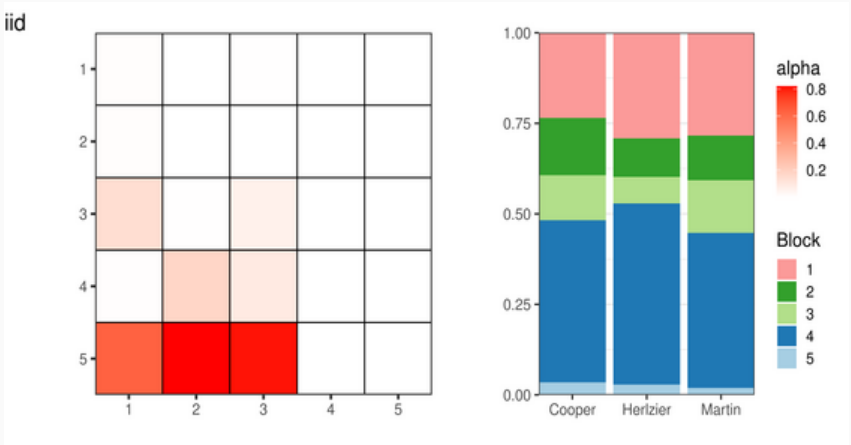
Separate sbm

Model choice

Model	ICL
sepSBM	-2080
iid-colSBM	-1966
π-colSBM	-1982
δ-colSBM	-1969
$\delta\pi$-colSBM	-1989

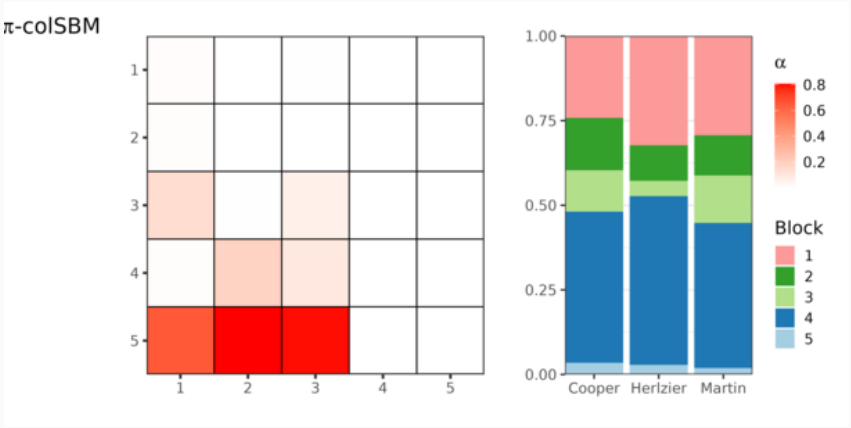
- Reject sepSBM : commun structure in the networks

iid-colSBM : the prefered model



- Makes 5 blocks
- Block 3 (light green) is a small block of intermediate trophic level species with some within block predation.
- The higher trophic level is divided into 2 more blocks,
  - block 2 (dark green) only preys on the 2 basal blocks
  - block 1 (pink) preys on the intermediate block 3 level but only on the most connected basal species block.

π-colSBM



- Also 5 blocks.
- There are no empty blocks
- the block proportions are roughly corresponding to the ones of iid-colSBM .
- Flexibility of the π-*colSBM* of little use compared to the iid-colSBM on this collection.

Conclusion

- The three networks do share a commun structure.
- We can identify the species playing the same role across networks (ecosytems)
- Other results
  - Quality of prediction when missing data.
  - Application in sociology : advices between lawyers, researchers or priests
  - Clustering of networks. Application on a database of 80 networks.

Work in progress

- Develop a wide variety of models
- Very active research field in our group
- Various extensions in progress
  - Taking into account the incertitude of reconstruction of the networks (data from metagenomics)
  - Extension to large multilayer networks such as interactome
  - Looking for tools to compare networks : plant health submitted to combination of stress

References i

Bar-Hen, A., Barbillon, P., and Donnet, S. (2018).

Block models for multipartite networks. applications in ecology and ethnobiology.

arXiv preprint arXiv :1807.10138.

Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2016).

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.

Bianconi, G. (2018).

***Multilayer Networks : Structure and Function*.**

Oxford University Press.

Bickel, P., Choi, D., Chang, X., Zhang, H., et al. (2013).

Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels.

The Annals of Statistics, 41(4) :1922–1943.

Biernacki, C., Celeux, G., and 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.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012).

Consistency of maximum-likelihood and variational estimators in the stochastic block model.

Electronic Journal of Statistics, 6 :1847–1899.

Chabert-Liddell, S.-C., Barbillon, P., Donnet, S., and Lazega, E. (2021).

A stochastic block model approach for the analysis of multilevel networks : An application to the sociology of organizations.

Computational Statistics & Data Analysis, 158 :107179.

References ii

Chabert-Liddell, S., Barbillon, P., and Donnet, S. (2022).

Impact of the mesoscale structure of a bipartite ecological interaction network on its robustness through a probabilistic modeling.

Environmetrics, 33(2).

Fisogni, A., Hautekèete, N., Piquot, Y., Brun, M., Vanappelghem, C., Ohlmann, M., and Massol, F. (2020).

Modifications of the plant-pollinator network structure and species' roles along a gradient of urbanization.

Kéfi, S., Miele, V., Wieters, E. A., Navarrete, S. A., and Berlow, E. L. (2016).

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.

Mariadassou, M. and Matias, C. (2015).

Convergence of the groups posterior distribution in latent or stochastic block models.

Bernoulli, 21(1) :537–573.

Mariadassou, M., Robin, S., and Vacher, C. (2010).

Uncovering latent structure in valued graphs : A variational approach.

The Annals of Applied Statistics, 4(2) :715–742.

Matias, C. and Miele, V. (2017).

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.

Picard, F., Miele, V., Daudin, J.-J., Cottret, L., and Robin, S. (2009).

Deciphering the connectivity structure of biological networks using mixnet.

BMC Bioinformatics, 10(6) :S17.

References iii

Pilosof, S., Porter, M., Pascual, M., and Kéfi, S. (2017).
The multilayer nature of ecological networks.
Nature Ecology & Evolution, 1(4).

Pocock, M. J. O., Evans, D. M., and Memmott, J. (2012).
The robustness and restoration of a network of ecological networks.
Science, 335(6071) :973–977.

Snijders, T. A. (2001).
The statistical evaluation of social network dynamics.
Sociological methodology, 31(1) :361–395.

Tabouy, T., Barbillon, P., and Chiquet, J. (2020).
Variational inference for stochastic block models from sampled data.
Journal of the American Statistical Association, 115(529) :455–466.

Thompson, R. M. and Townsend, C. R. (2003).
Impacts on stream food webs of native and exotic forest : An intercontinental comparison.
Ecology, 84(1) :145–161.

Vissault, S., Cazelles, K., Bergeron, G., Mercier, B., Violet, C., Gravel, D., and Poisot, T. (2020).
rmangal : An R package to interact with Mangal database.
R package version 2.0.2.

To go further : partitionning a collection of networks

- If the networks in a collection do not have the same connectivity structure, we aim to partition them accordingly.
- Finding a partition $\mathcal{G} = (\mathcal{M}_g)_{g=1,\dots,G}$ of $\{1, \dots, M\}$.
  such that

$$\forall g \in \{1, \dots, G\}, \quad \forall m \in \mathcal{M}_g, \quad \mathbf{Y}^m \sim \text{SBM}(K^g, \boldsymbol{\pi}^m, \boldsymbol{\alpha}^g)$$

  networks belonging to the subcollection $\mathcal{M}_g$ share the same mesoscale structure given by π-*colSBM*.

Scoring a partition

- To any partition $\mathcal{G}$ we associate the following score :

$$\mathrm{Sc}(\mathcal{G}) = \sum_{g=1}^{G} \text{BIC-L}((\mathbf{Y}^m)_{m \in \mathcal{M}_g}, \widehat{K^g}).$$

- Best partition $\mathcal{G}$ is chosen as follows :

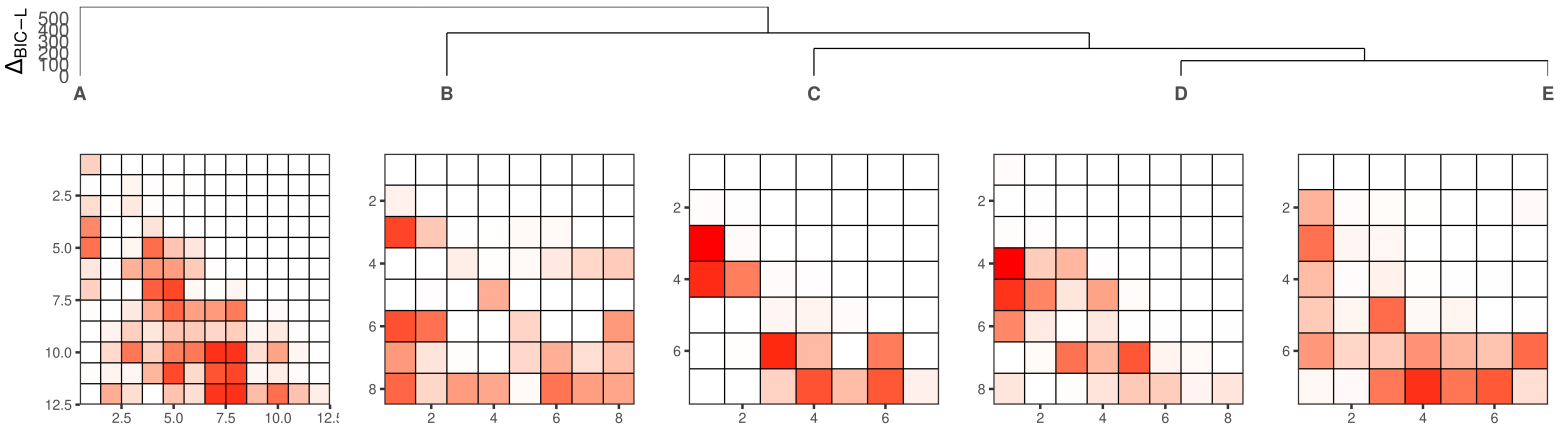
$$\mathcal{G}^* = \arg\max_{\mathcal{G}} \mathrm{Sc}(\mathcal{G}).$$

Partition of the networks from the Mangal database

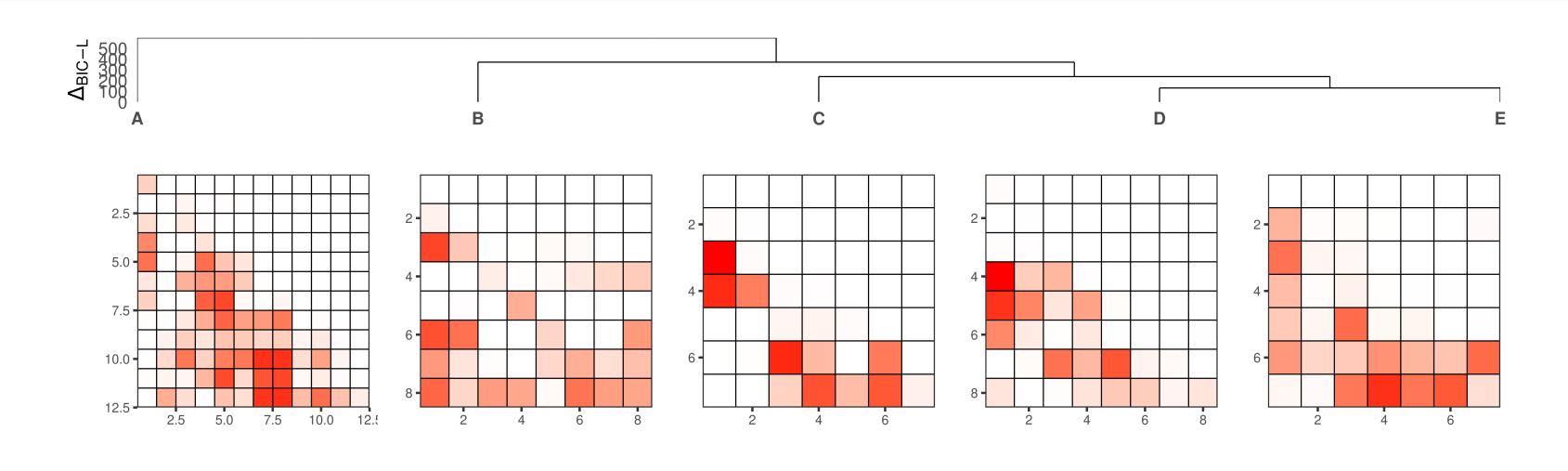
- 67 networks issued from the Mangal database belonging to 33 datasets. [Vissault et al., 2020]
- predation networks which are all directed networks with more than 30 species,
- number of species ranges from 31 to 106 (3395 in total) by network
- Density ranging from .01 to .32 (14934 total predation links).

Aim use our model to propose partition of the networks into group of networks with common mesoscale structure.

Partition on the networks from the Mangal database



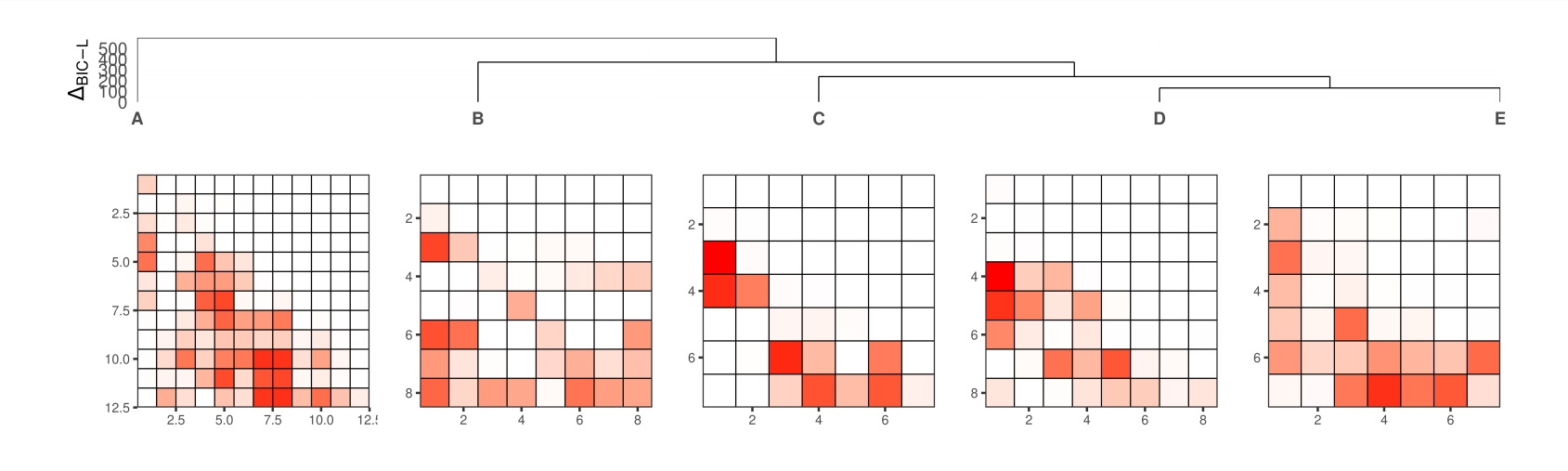
Partition on the networks from the Mangal database



Groupe A

- 7 networks and 12 blocks are required to describe this group of networks
- 5 networks are issued from the same dataset (id : 80).
- These 5 networks populate the 12 blocks, while the other 2 networks only populate parts of them.
- Average density is about 0.18
- Blocks 1 to 3 represent the higher trophic levels, blocks 4 to 8 the intermediate ones and block 9 to 12 the lower ones.

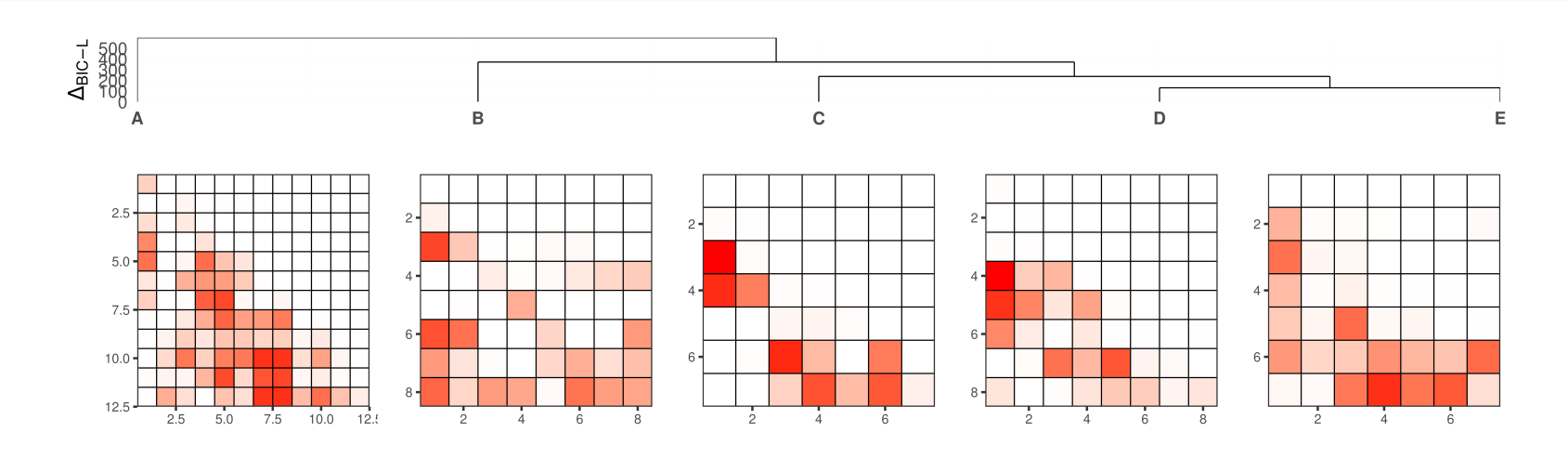
Partition on the networks from the Mangal database



Group B : structure with 8 blocks

- 26 networks with heterogeneous size and density.
- Issued from various datasets
- Most networks populate only parts of the 8 blocks
- Block 4 is represented in only 5 networks where it is either an intermediate or a bottom trophic level.
- Species from top trophic levels prey on basal species.

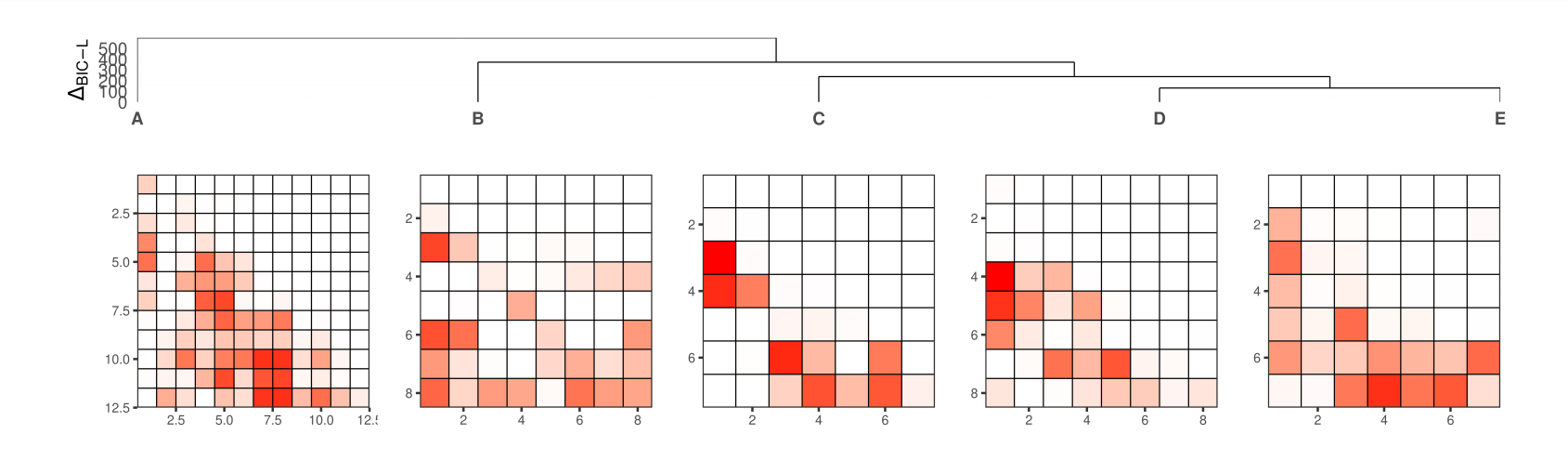
Partition on the networks from the Mangal database



Group C : structure with 7 blocks

- 6 networks with density ranging from .06 to .11.
- All networks are represented in 5 or 6 of the 7 blocks, including the first three blocks.
- 3 of the 5 networks of dataset 48 (diff. collecting sites).
- Top trophic level divided into 2 blocks, species from those blocks preying only on intermediate trophic level species.

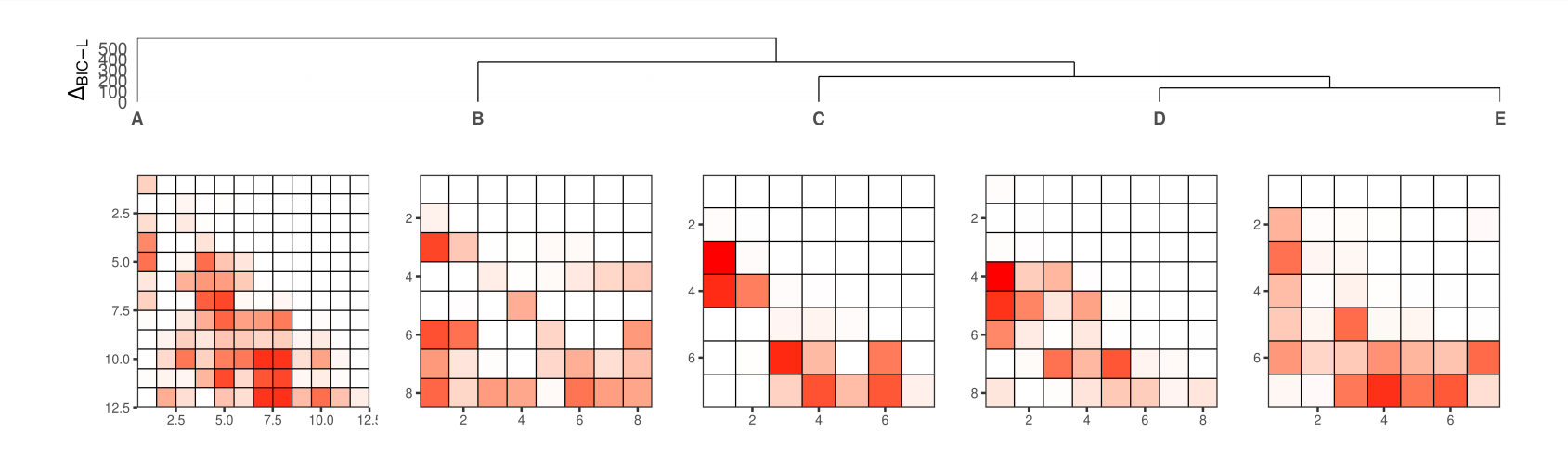
Partition on the networks from the Mangal database



Group D : structure with 7 blocks

- 23 networks.
- The 10 networks from dataset 157 (stream food webs from New Zealand) are divided between groups **B** and **D** based on the type of ecosystem. The data from group **B** were collected in creeks, while the one from group **D** were collected on streams.

Partition on the networks from the Mangal database



Group E : structure with 7 blocks

Comments on the ICL versus BIC

Conjecture

$$BIC(\mathcal{M}_{\mathbf{K}}) \quad = \quad \log \ell(\mathbf{Y}; \hat{\theta}, \mathcal{M}_{\mathbf{K}}) - \text{pen}(\mathcal{M}_{\mathbf{K}})$$

with the same penalty

- Under this conjecture

$$\begin{aligned} ICL(\mathcal{M}_{\mathbf{K}}) &= BIC(\mathcal{M}_{\mathbf{K}}) + \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{Y}; \hat{\theta}_K) \log p(\mathbf{Z}|\mathbf{Y}; \hat{\theta}_K) \\ &= BIC(\mathcal{M}_{\mathbf{K}}) - \mathcal{H}(p(\cdot|\mathbf{Y}; \hat{\theta}_K)) \end{aligned}$$

- As a consequence, because of the entropy, ICL will encourage clustering with well-separated groups
- $$\widehat{ICL}(\mathcal{M}) = BIC(\mathcal{M}) + \sum_{\mathbf{Z}} \mathcal{R}_{\mathbf{Y}}(\mathbf{Z}, \widehat{\tau}) \log \mathcal{R}_{\mathbf{Y}, \widehat{\tau}}(\mathbf{Z}) - \mathbf{KL}[\mathcal{R}_{\mathbf{Y}, \widehat{\tau}}, p(\cdot|\mathbf{Y}; \widehat{\theta})] \, .$$

Back to the presentation

About bipartite networks

- Example : plant / pollinators
- Bi-clustering
- Same principle
  - K_1 blocks of plants, K_2 blocks of pollinators
  - 2 sets of latent clustering variables $(Z_i^{plant})_{i=1,\dots,n_1}, (Z_j^{poll})_{j=1,\dots,n_2} \dots$
  - Conditionnally to latent variables : (Y_{ij}) independent and

$$Y_{ij} | Z_i^{plant} = k, Z_j^{poll} = \ell \sim \mathcal{B}ern(\alpha_{k\ell})$$

About multipartite networks

- Involving more than two functional groups (Q)
- For instance plants : pollinators, seed-dipersal birds, ants...
  - Q-clustering : Q sets of latent clustering variables $(Z_i^q)_{i=1,\dots,n_q}$...
  - K_q blocks in each functional group
  - Conditionnally to latent variables : $(Y_{ij}^{qq'})$ independent and

$$Y_{ij}^{qq'} | Z_i^q = k, Z_j^{q'} = \ell \sim \mathcal{B}ern(\alpha_{k\ell}^{qq'})$$

- Inference, model selection procedure : adapted
- Package sbm