Stochastic block models for a collection of networks

Applications in ecology and sociology

Sophie Donnet, INRAE, MIA Paris-Saclay

Aug. 2023

Collaborators

Joint work with

P. Barbillon

(AgroParisTech)

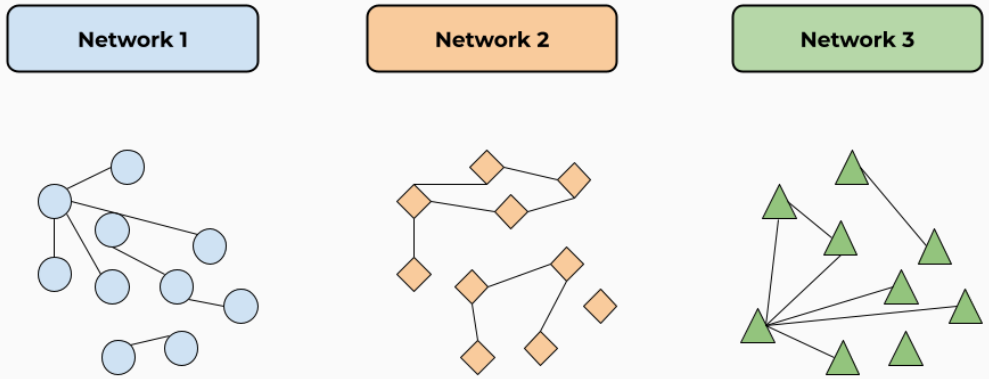
S.C. Chabert-Liddell

(INRAE)

Collection of networks : consensus in the structure

Objectives

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

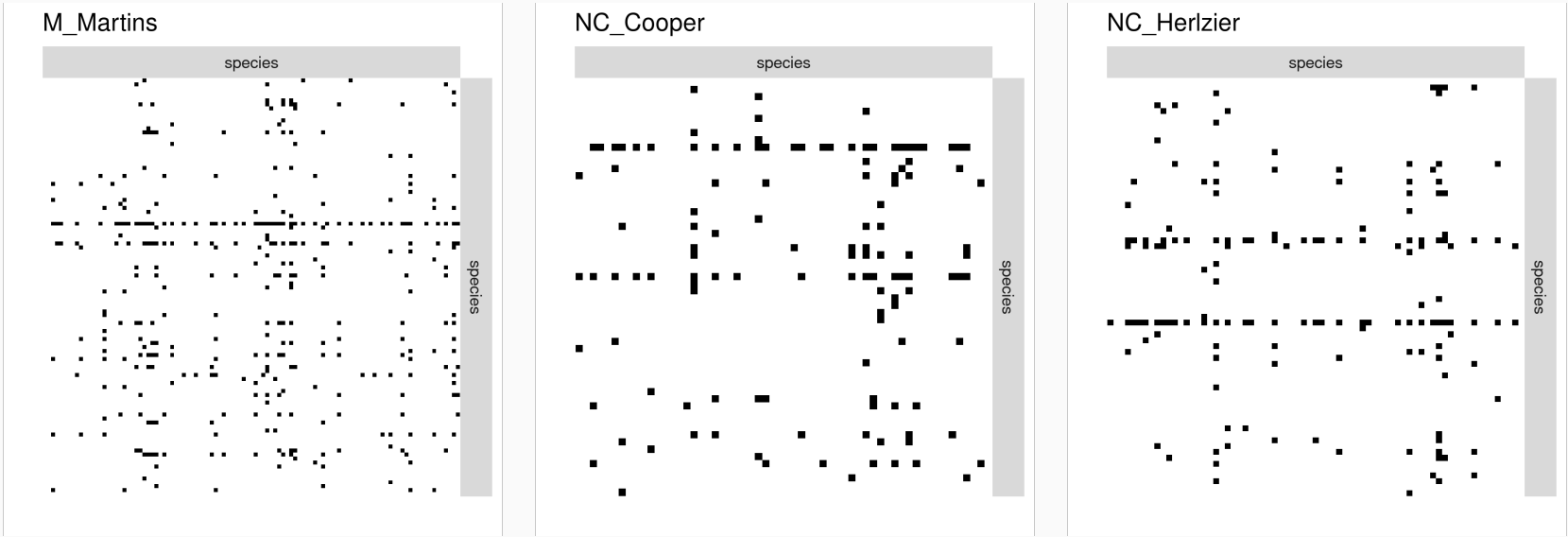


Applications

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

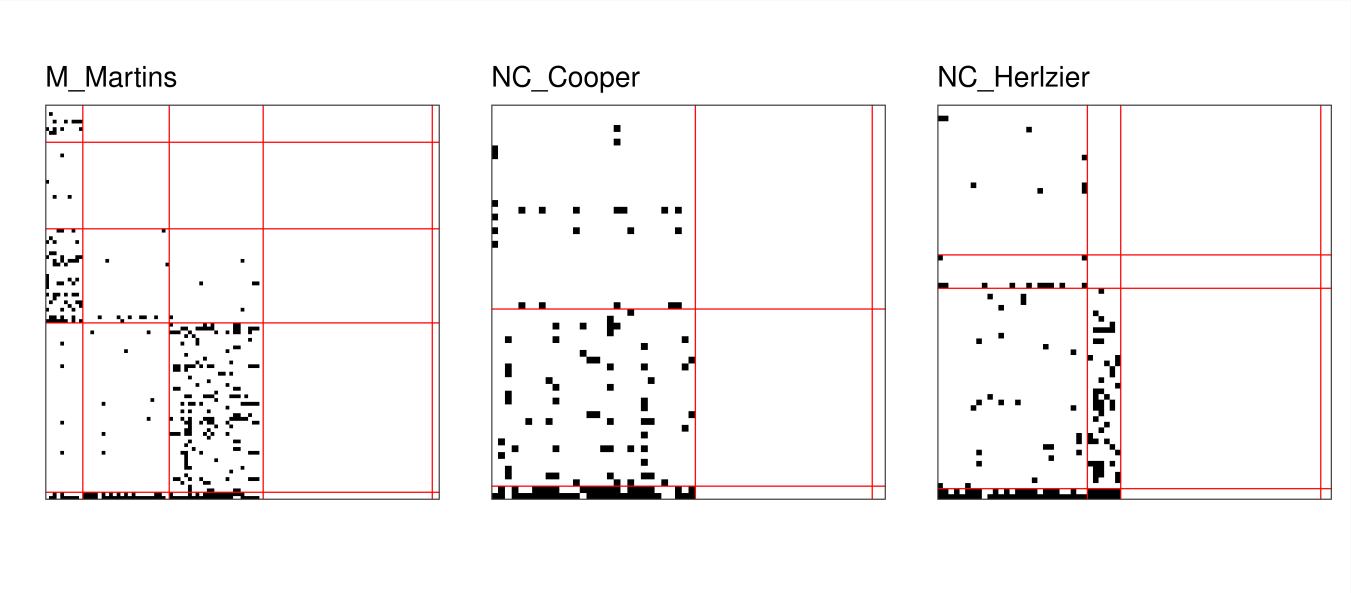
Three foodwebs

- 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

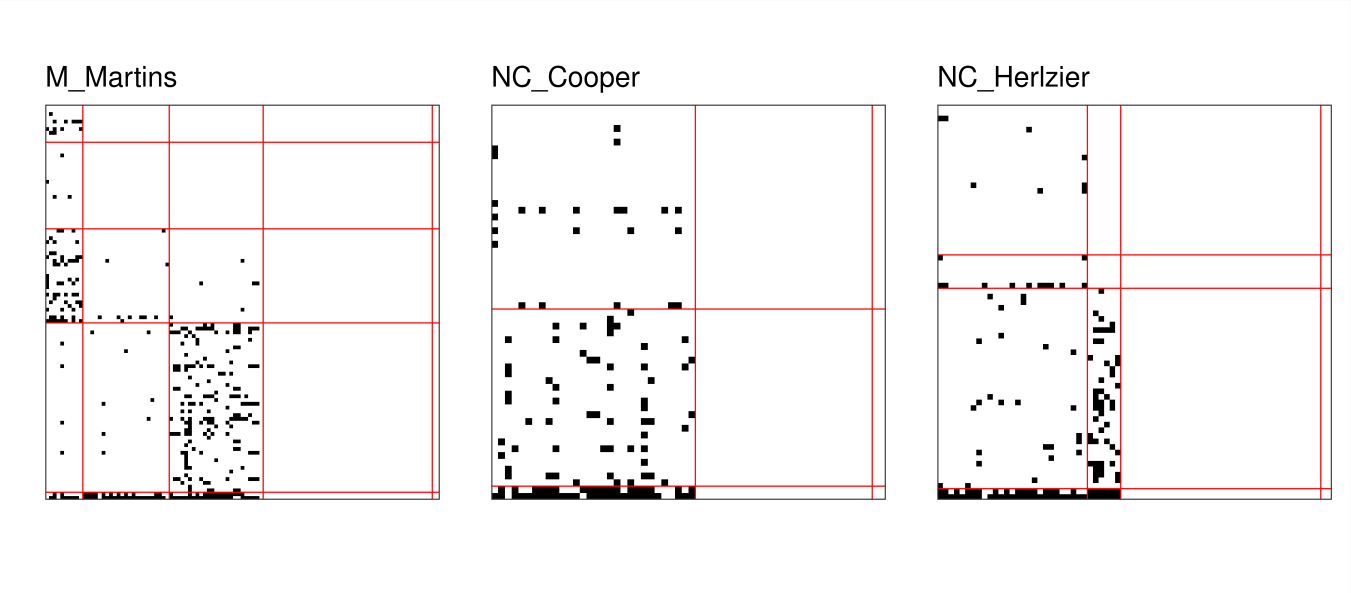


- Fitted SBM on each separately
- Reordered the matrices following the blocks
- Label the blocks following the average out-degrees order

Interpretation :

- In row : is eaten by...
- In col : eats...

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

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

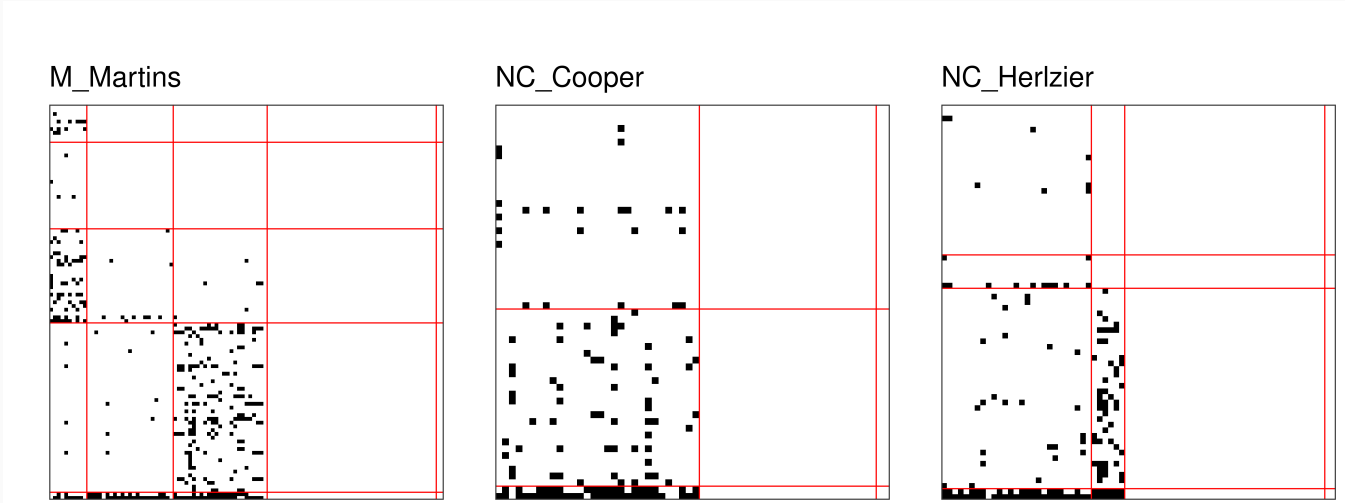
where $\nu(\delta) = M - 1$ for $\delta colSBM$ and 0 otherwise.

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.

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

Application on the foodwebs



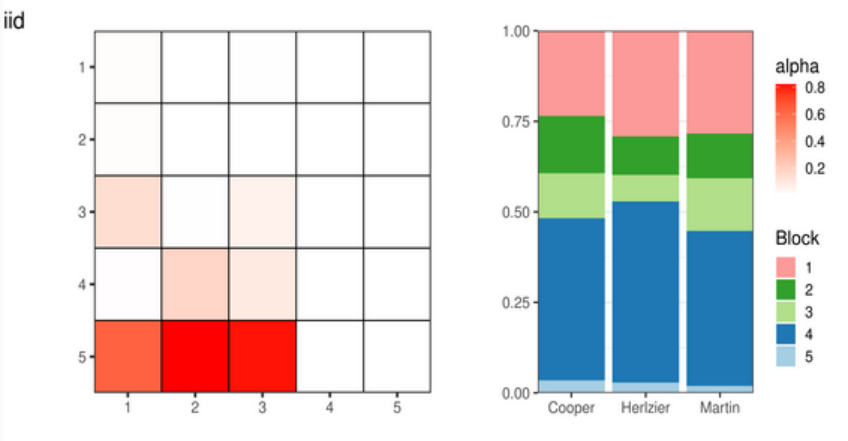
Separate sbm

Model choice

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

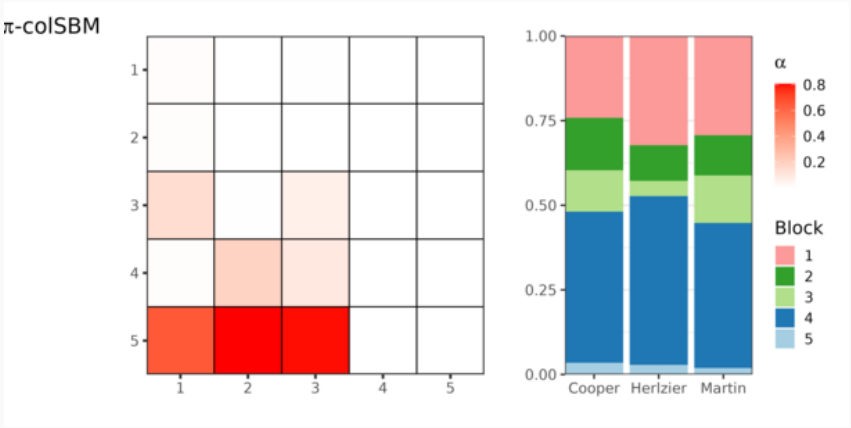
- Reject sepSBM : commun structure in the networks

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

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 :

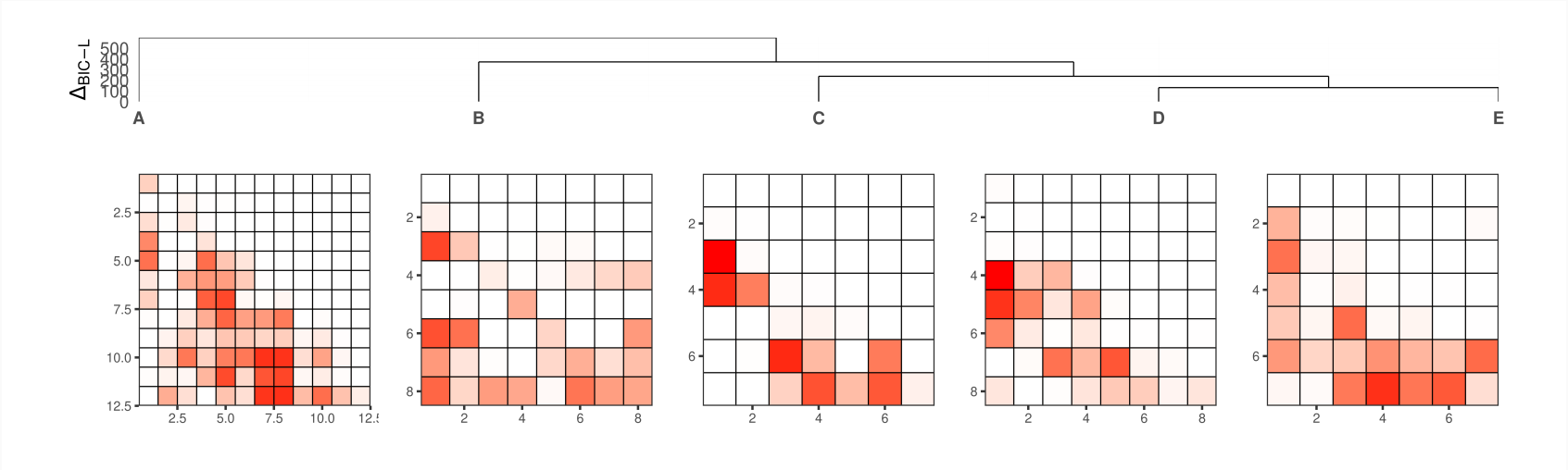
$$\mathcal{G}^* = \arg\max_{\mathcal{G}} \mathrm{Sc}(\mathcal{G}).$$

Partition of the networks from the Mangal database

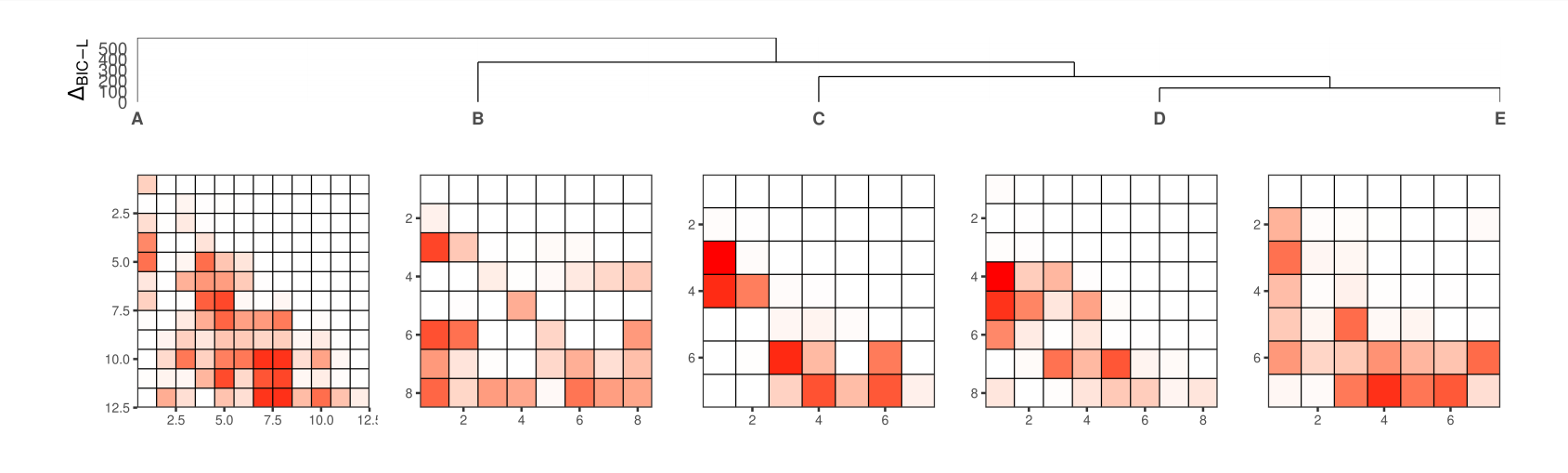
- 67 networks issued from the Mangal database belonging to 33 datasets. [Vissault et al., 2020]
- predation networks which are all directed networks with more than 30 species,
- number of species ranges from 31 to 106 (3395 in total) by network
- Density ranging from .01 to .32 (14934 total predation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



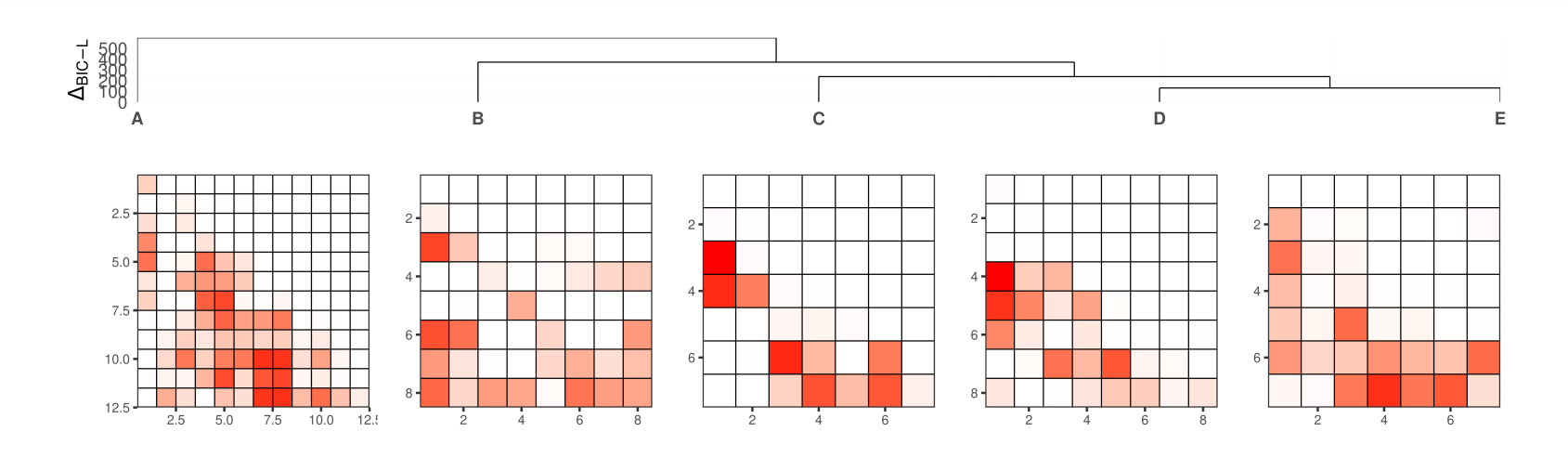
Partition on the networks from the Mangal database



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.

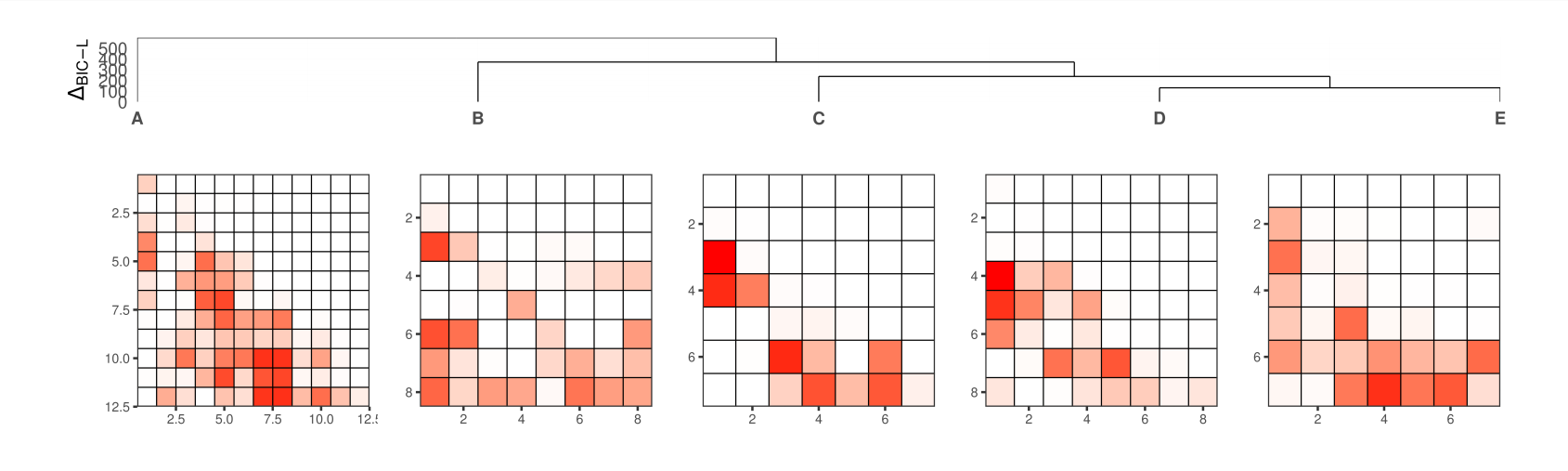
Partition on the networks from the Mangal database



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.

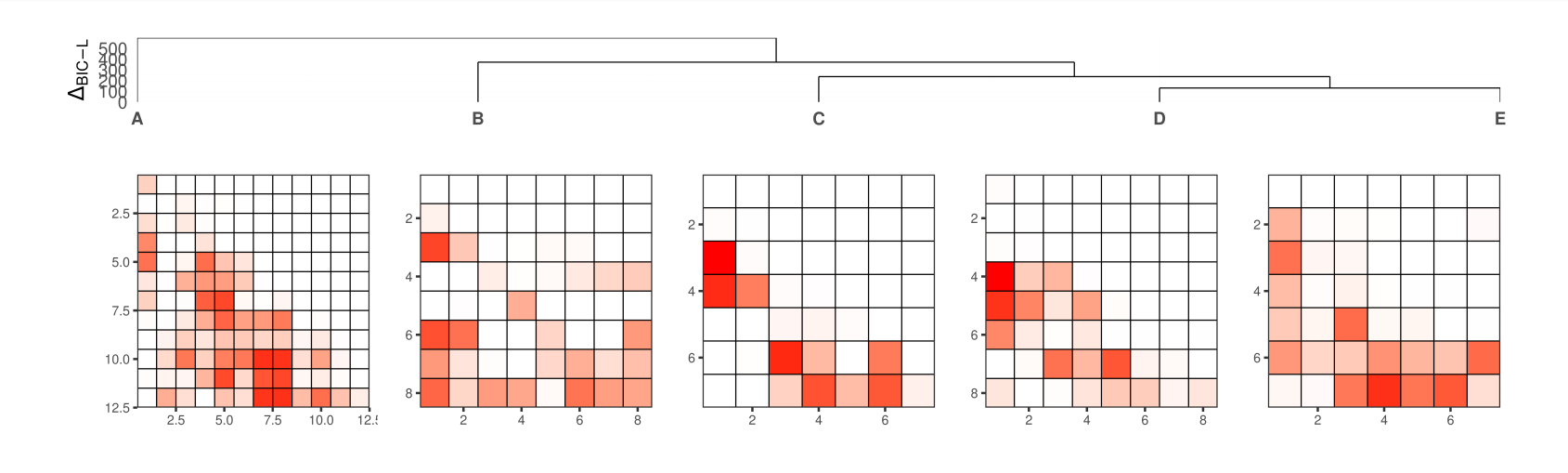
Partition on the networks from the Mangal database



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.

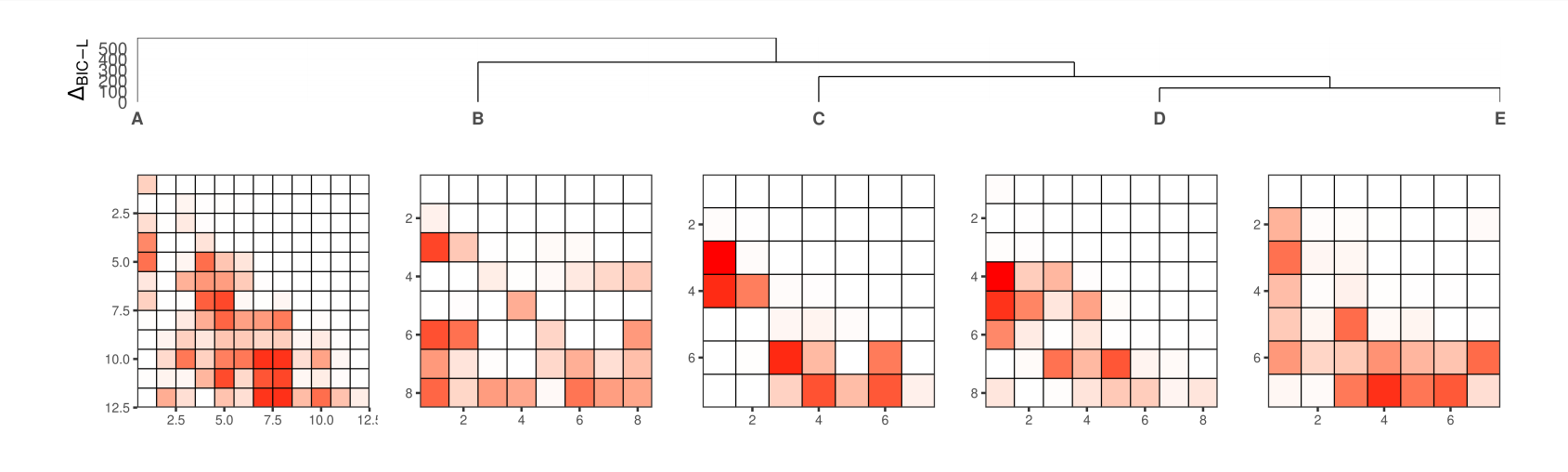
Partition on the networks from the Mangal database



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