Learning common structures in a collection of networks

Do the networks share commmon structures?

Saint-Clair Chabert-Liddell
Joint work with S. Donnet and P. Barbillon

12 December 2021

Netbio

Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris

Summary

Stochastic Block Model

Modeling a Collection of Networks

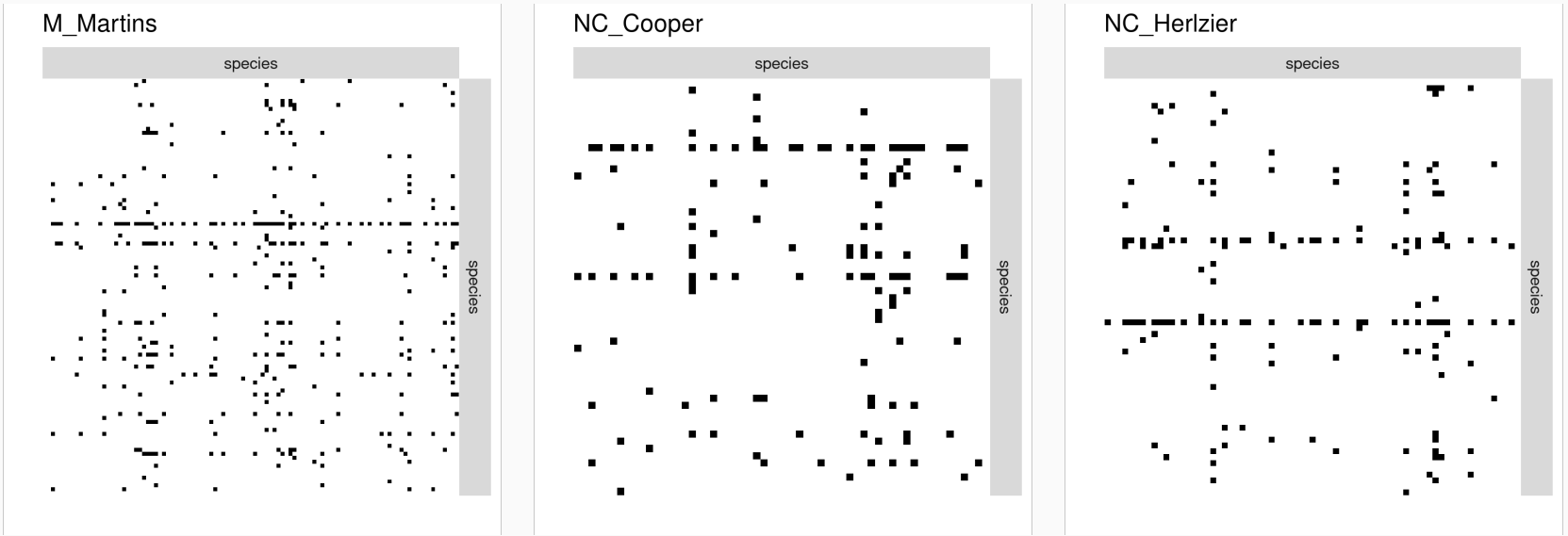
Inference, Model Selection and Partition of Networks

Applications to food webs

Motivation

Data

- Collection $\mathbf{X} = \{\ldots, X^m, \ldots\}$, $m \in \mathcal{M}$ of $M = |\mathcal{M}|$ networks
- Same type:
  - Simple, Bipartite...
  - Undirected, Directed: *Food web*, Advice network
- Same value type:
  - Binary (Bernoulli), Count (Poisson)...



Motivation

Data

- Collection $\mathbf{X} = \{\ldots, X^m, \ldots\}$, $m \in \mathcal{M}$ of $M = |\mathcal{M}|$ networks
- Same type:
  - Simple, Bipartite...
  - Undirected, Directed: *Food web*, Advice network
- Same value type:
  - Binary (Bernoulli), Count (Poisson)...

Objective Find a common connectivity structure

Question Is the common structure relevant?

Objective Partition networks by connectivity structures

Method Joint modeling with *Stochastic Block Model (SBM)*

Stochastic Block Model

Stochastic Block Model (Snijders and Nowicki, 1997)

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

Latent variables

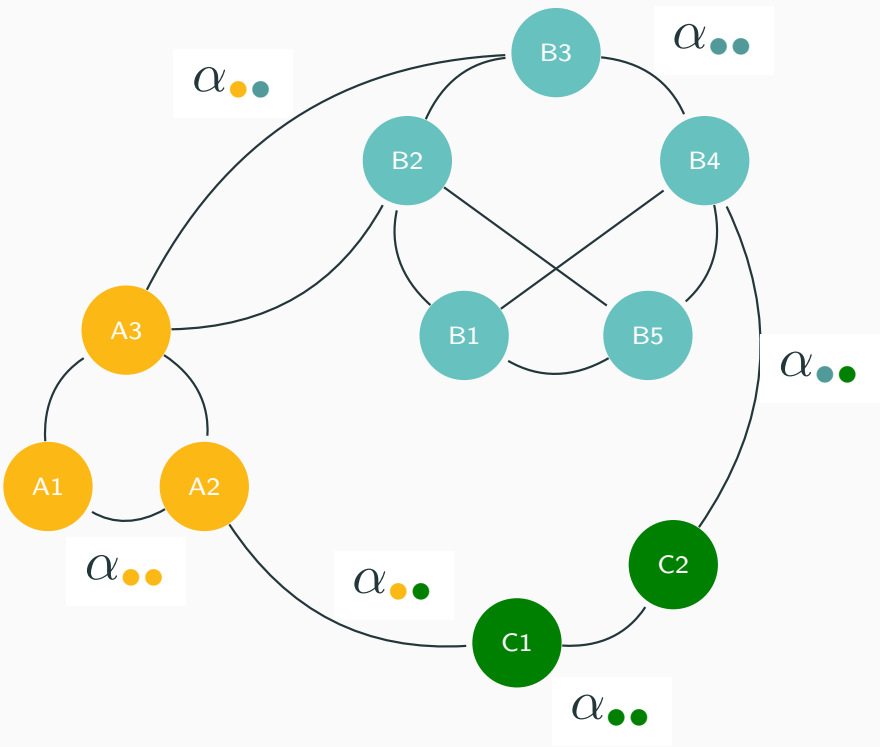
- The nodes $i = 1, \dots, n$ are partitionned into Q clusters
- $Z_i = q$ if node i belongs to cluster (block) q
- Z_i independant variables

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

Conditionally to $(Z_i)_{i=1,\dots,n}$**...**
(X_{ij}) independant and

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

Stochastic Block Model : illustration



Parameters

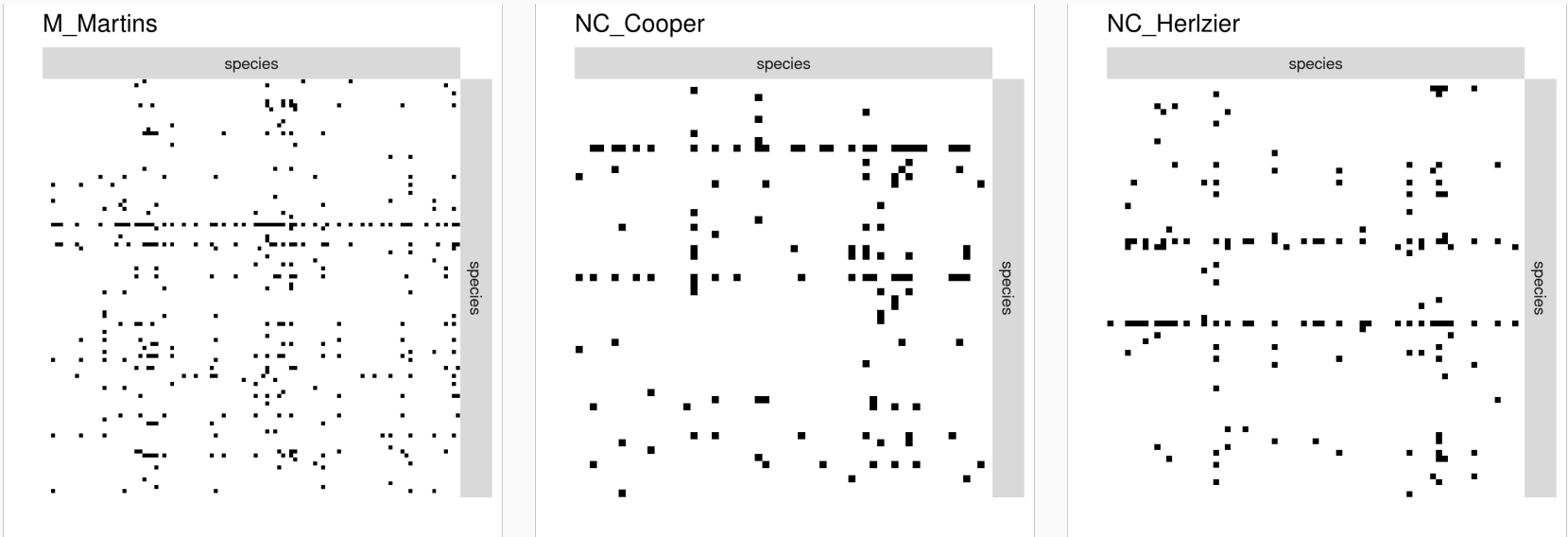
Let n nodes divided into 3 clusters

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

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

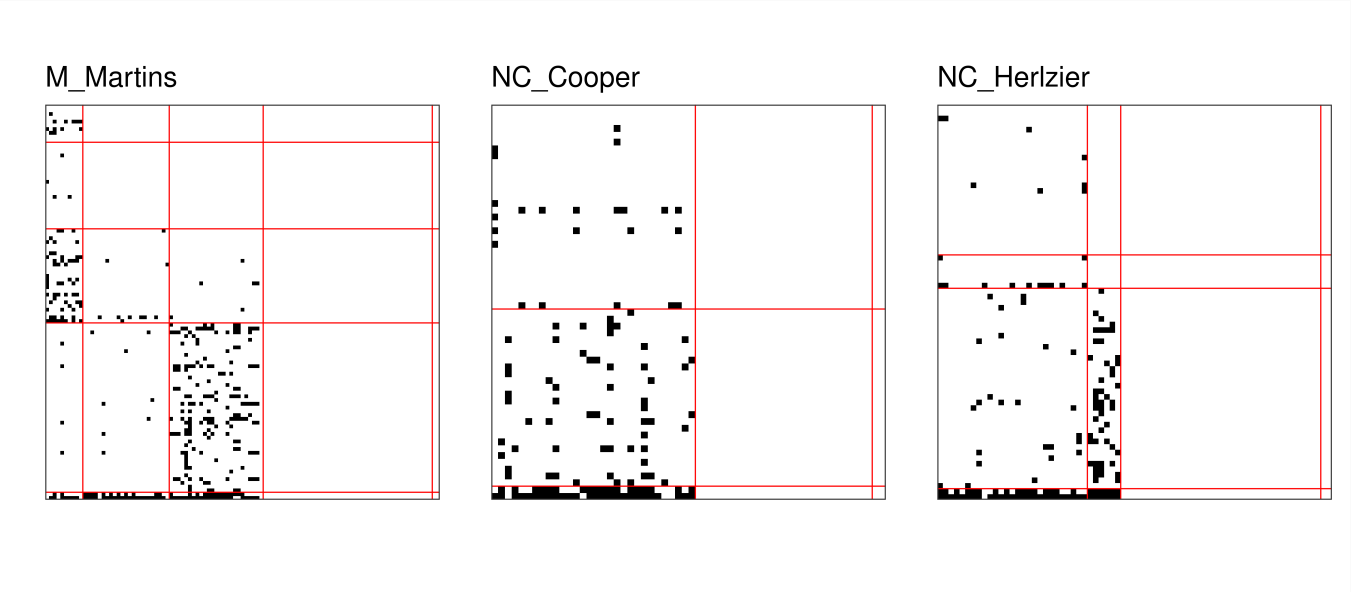
Three food webs

- Pine-forest stream food webs issued from Maine and North-Carolina (Thompson and Townsend, 2003)
- Involve respectively 105, 58 and 71 species.
- $X_{ij} = 1$ if i is eaten by j. Directed relation



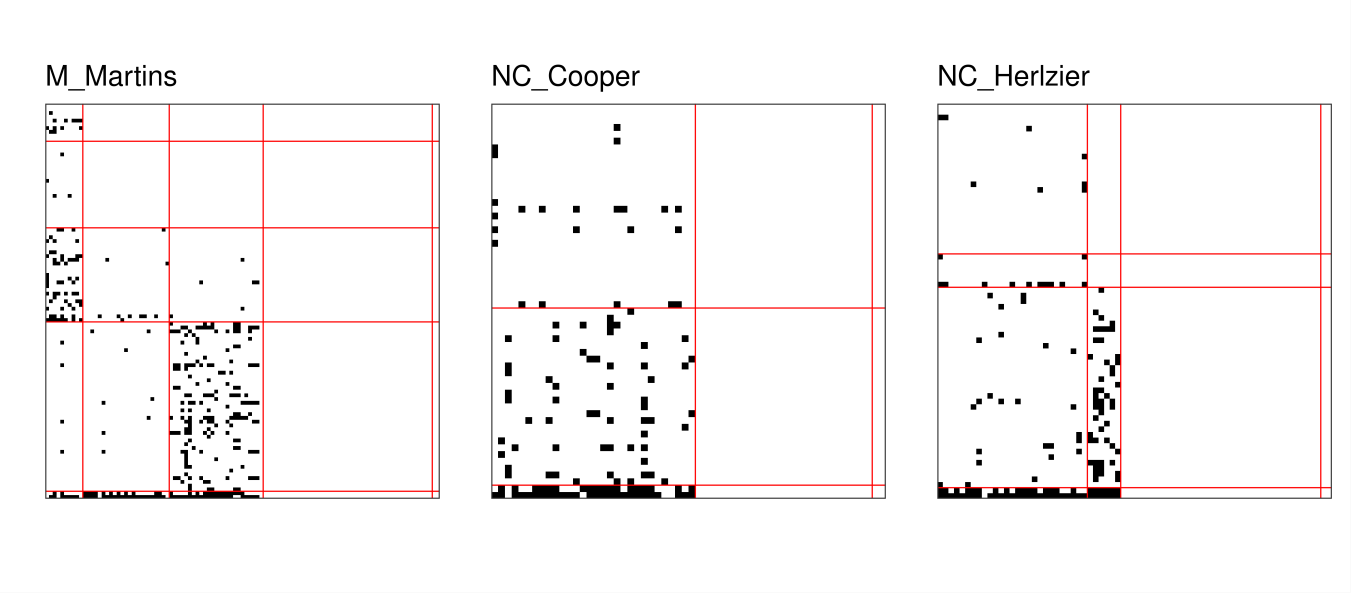
- Look for similarities and differences between network structures.

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**: 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.

Modeling a Collection of Networks

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 $(X^m)_{m=1,\dots,M}$ denote the collection of networks each involving n_m nodes.
- (X^m) independent.
- $$X^m \sim \mathrm{SBM}_{n_m}(Q_m, \pi^m, \alpha^m)$$
- Conditions on the parameters $(\pi^m)_{m=1,\dots,M}$ and $(\alpha^m)_{m=1,\dots,M}$

First naive model

iid-colSBM

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

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

- Same blocks proportions
- Same connectivity structure

- $(Q-1) + Q^2$ unknown parameters, M clustering

First naive model

iid-colSBM

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

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

- Same blocks proportions
- Same connectivity structure

- $(Q-1) + Q^2$ unknown parameters, M clustering
- i.i.d. assumption too strict for most datasets, 2 new mechanisms:
  - Free proportion of blocks between networks
  - Density varies between networks

A first relaxed model : π-colSBM

π-colSBM

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

- Same structure of connection α
- Specific proportions of blocks in each network

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

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

- 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,

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

Varying density model : δ-colSBM

δ-colSBM

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

with $\pi_q > 0$.

- Similar intra- and inter blocks connectivity patterns
- Network specific density density parameter. $\delta^1 = 1$

- Mimics differences of effort sampling or abundances
- $(Q-1) + Q^2 + (M-1)$ parameters.

Varying density and block proportion model

$\delta\pi$-colSBM

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

with $\pi_q^m \geq 0$

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

Summary

M independent networks.

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

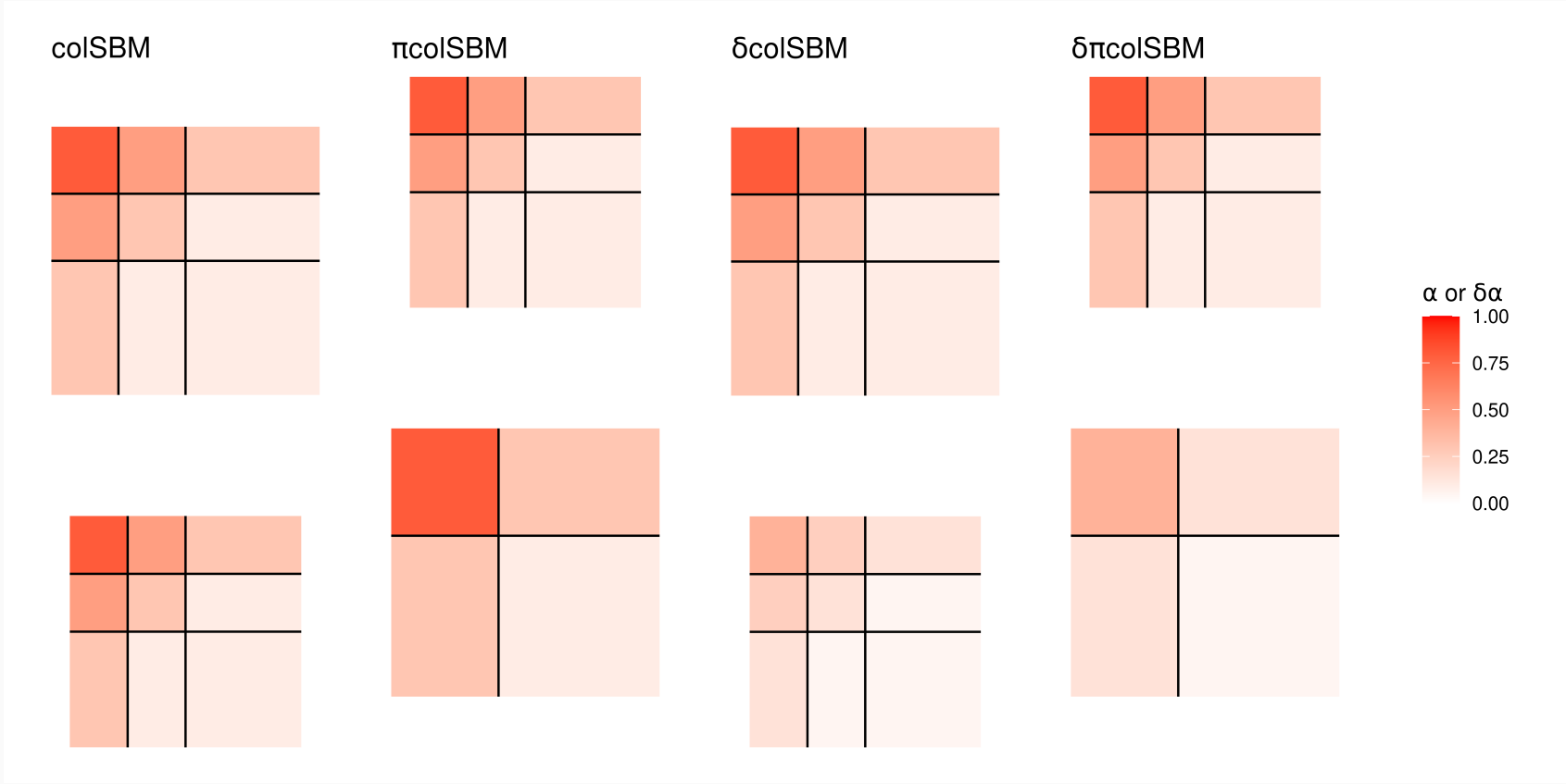
Model name	Block prop.	Connexion param.	Nb of param.
iid-colSBM	$\pi_q^m = \pi_q,\ \pi_q > 0$	$\alpha_{qr}^m = \alpha_{qr}$	$(Q-1) + Q^2$
π-colSBM	$\pi_q^m,\ \pi_q^m \geq 0$	$\alpha_{qr}^m = \alpha_{qr}$	$\leq M(Q-1) + Q^2$
δ-colSBM	$\pi_q^m = \pi_q,\ \pi_q > 0$	$\alpha_{qr}^m = \delta^m \alpha_{qr}$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

where $Q_m = \sum_{m=1}^M S_{mq}$

Summary

M independent networks.

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



Identifiability

Proven for separated SBMs (Celisse et al., 2012)

Demonstrated for all colSBMs, 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, Model Selection and Partition of Networks

Maximum Likelihood Inference

For fixed Q, support S, $\boldsymbol{\theta} = \{\alpha, \pi, \delta\}$:

Objective Joint clustering of $\mathrm{Z} = \{Z^1, \ldots, Z^M\}$ and estimates of $\boldsymbol{\theta}$

Method Maximum likelihood of the observed data

Idea Compute complete likelihood and integrate on Z

Problem Intractable, sum of $\prod_{m \in \mathcal{M}} |\mathcal{Q}_m|^{n_m}$ terms

Solution EM algorithm

Problem $\mathcal{L}(\mathrm{Z}|\mathrm{X})$ also intractable

Solution Variational approach of the EM algorithm

Daudin et al. (2008)

Variational bound

$$\begin{aligned}\ell(\mathsf{X};\boldsymbol{\theta}) &\geq \sum_{m\in\mathcal{M}} \ell(X^m;\boldsymbol{\theta}) - D_{KL}(\mathcal{R}(Z^m)\|p(Z^m|X^m)) \\ &= \sum_{m\in\mathcal{M}} (\mathbb{E}_{\mathcal{R}}[\ell(X^m,Z^m;\boldsymbol{\theta})] + \mathcal{H}(\mathcal{R}(Z^m))) =: \mathcal{J}(\mathcal{R}(\mathsf{Z}),\boldsymbol{\theta}).\end{aligned}$$

$\mathcal{R}(\mathsf{Z})$ is a mean-field approximation of $\mathsf{Z}|\mathsf{X}$

$\mathcal{H}$ is the entropy

Variational EM

V-EM algorithm
2 steps iterative algorithm, for each $m \in \mathcal{M}$:

VE Maximize $\mathcal{J}(\mathcal{R}(Z^m), \boldsymbol{\theta})$ w.r.t. $\mathcal{R}(\mathrm{Z})$

M Maximize $\mathcal{J}(\mathcal{R}(\mathrm{Z}), \boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$

- *VE*-step are independent for each network
- Introduce stochasticy in the V-EM algorithm
- (δ–$\delta\pi$)colSBM: *M*–Step not explicit for Bernoulli model
- *M*-step explicit for Poisson model, very good when:
  - networks have few interactions by nodes
  - Goal is the clustering of nodes

Model selection

Penalized model-based criterion

- To choose Q or S
- To determine if common structure is relevant
- Based on Integrated Classification Likelihood (ICL)
- Modified to not penalize fuzzy clustering
- Adapted to allow for empty blocks
- Straightforward *iid*-colSBM and the δ-colSBM

$$\textit{BIC-L}(Q, S) = \mathcal{J}(\hat{\tau}, \hat{\boldsymbol{\theta}}) - \textit{pen}_{\text{colSBM}}$$

Biernacki et al. (2000)

Penalty for (π-$\delta\pi$)colSBMs

- π_q^m possibly null. Asymptotic approximation do not hold
- Each couple (Q, S) defines a model
- Penalty on the size of the model space

$$
\begin{aligned}
pen_{\pi\text{colSBM}} &= \underbrace{\frac{1}{2}\sum_{m=1}^{M}(Q_m - 1)\log(n_m)}_{pen_{\pi}} \\
&+ \underbrace{\frac{1}{2}\left(\sum_{q,r=1}^{Q} 1_{(S'S)_{qr}>0} + \nu(\delta)\right)\log\left(\sum_{m=1}^{M} n_m(n_m-1)\right)}_{pen_{(\alpha,\delta)}} \\
&+ \underbrace{\sum_{m=1}^{M}\log\binom{Q}{Q_m} + M\log(Q)}_{pen_{(Q,S)}},
\end{aligned}
$$

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

Relevance of the joint modeling

Common structure is relevant if:

$$\sum_{m=1}^{M} \max_{Q_m} BIC\text{-}L_{SBM}(Q_m) < \max_{(Q,S)} BIC\text{-}L_{\text{colSBM}}(Q, S)$$

Partition of networks

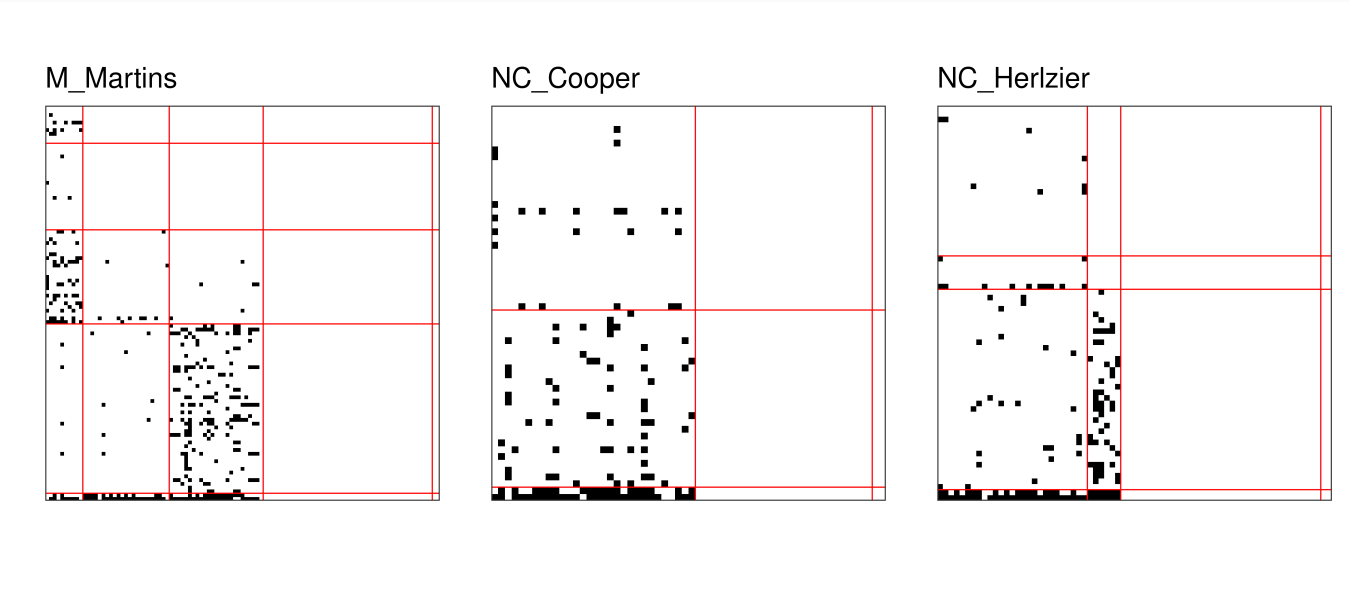
- Some networks may share common connectivity structure
- Group networks sharing the same structure
- Find the partition with the highest *BIC-L*

$\mathcal{G}$ a partition of $\mathcal{M}$ in G groups $\mathcal{M}_1, \ldots, \mathcal{M}_G$.

$$\mathcal{G}^* = \arg \max_{\mathcal{G} \in \mathcal{P}(\mathcal{M})} \sum_{g=1}^{G} \max_{(Q_g, S_g)} \textit{BIC-L}(Q_g, S_g | \mathcal{M}_g)$$

Applications to food webs

Application on the stream food webs



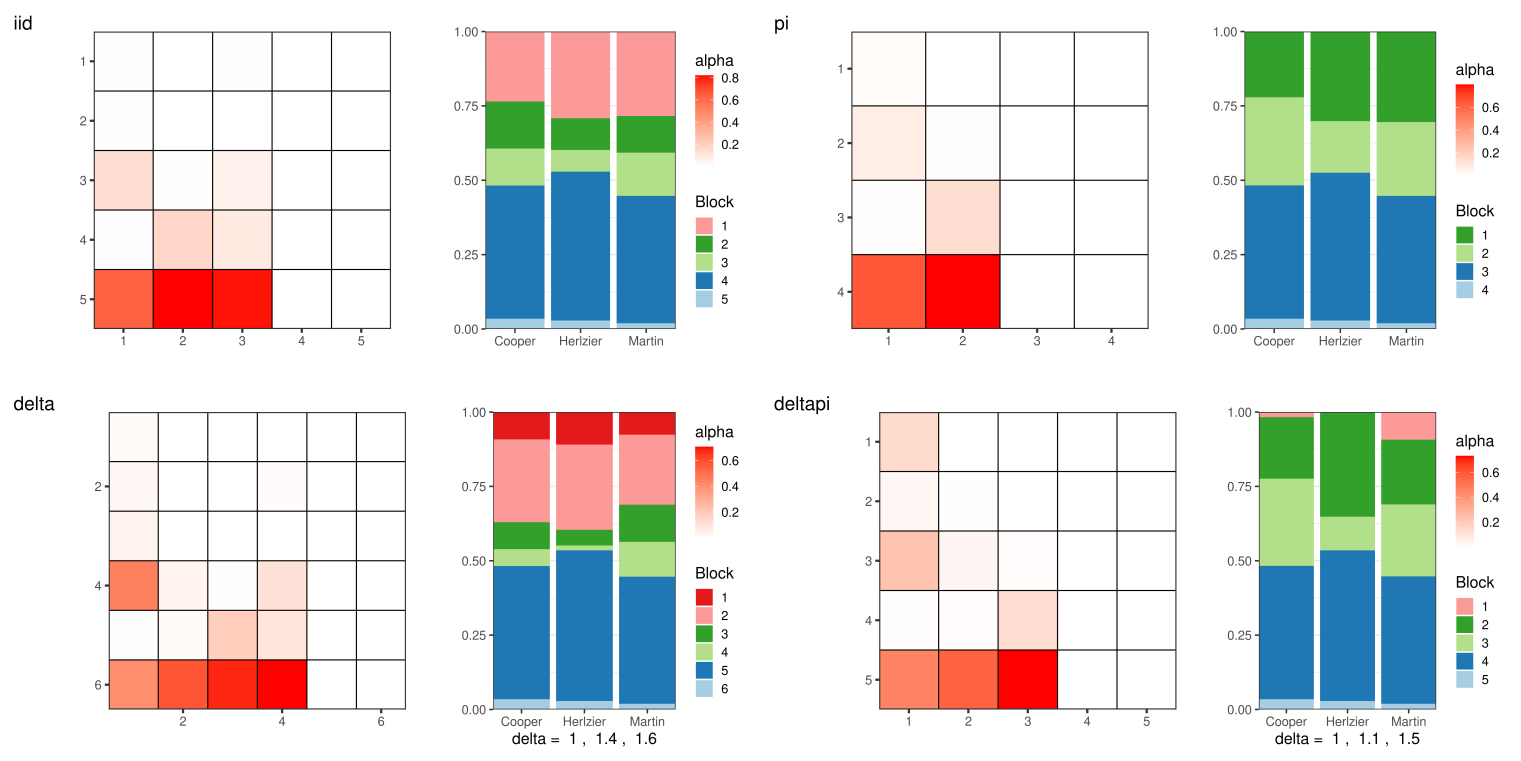
Separate sbm

Model choice

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

- Reject sepSBM : common structure in the networks

colSBMs on stream food webs

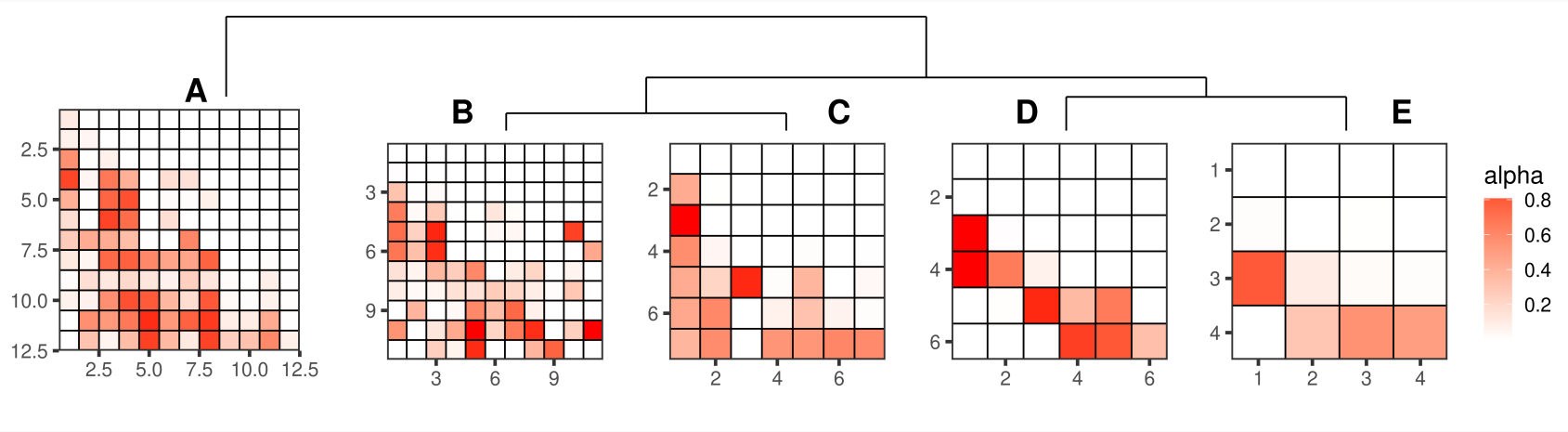


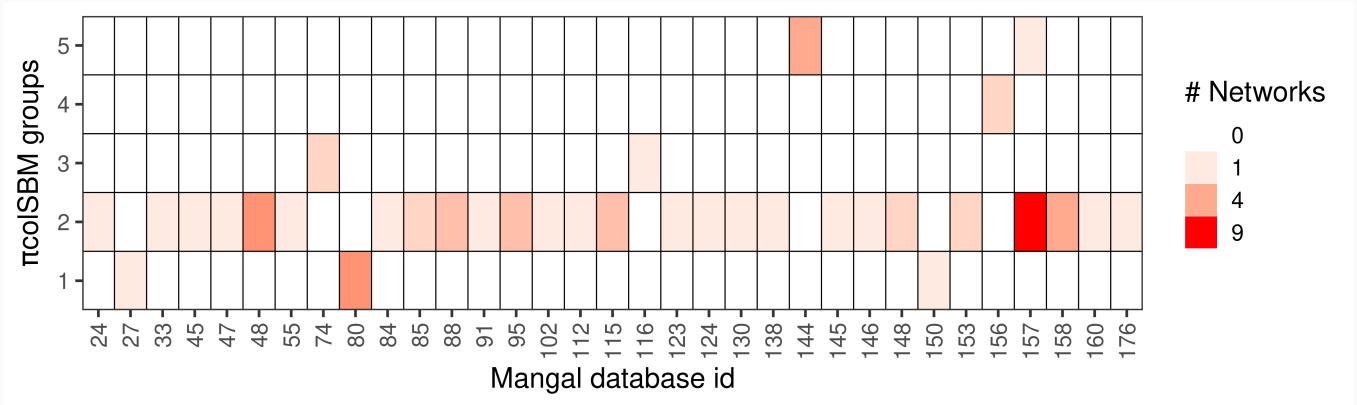
Top left : iid-colSBM (-1966). Top right: π-colSBM (-1982) Bottom-left: δ-colSBM (-1969). Bottom-right: $\delta\pi$-colSBM (-1989)

- iid-colSBM : prefered model. Make 5 blocks
- π-colSBM: block proportion quite similar. Make no use of its flexibility

Partition of Predation Networks

- $M = 67$ networks from Mangal database (Vissault et al., 2020)
- 31 to 106 species nodes
- Density range in $[.01, .32]$
- Modeling the collection with πcolSBM





Take Home Message

- Joint modeling of a collection of networks with coISBMs
  - Find a common structure between the different networks
  - Identify blocks between networks
  - Improve prediction of missing data (see arXiv paper soon)
  - Application in sociology: advices between judges, lawyers, priests or researchers
- Extension to other types of networks: bipartite, multipartite...
- Dealing with covariates on nodes, edges and networks
- Effect on common statistics: modularity, nestedness, reciprocity, robustness...

Any questions? `saint-clair.chabert-liddell@inrae.fr`

References

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.

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

References ii

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100.

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.

Partition of networks

All the networks in the collection may not have the same structure.

$$\mathcal{G}^* = \arg \max_{\mathcal{G} \in \mathcal{P}(\mathcal{M})} \sum_{g \in \mathcal{G}} \max_{(Q_g, S_g)} BIC - L(Q_g, S_g | \mathcal{M}_g)$$

Need 2^M partitions to compute all partitions. Too costly if M large.

Dissimilarity

- colSBMs allow to match Z^ms
- Compute dissimilarity matrix using MLE of SBM on colSBMs block:

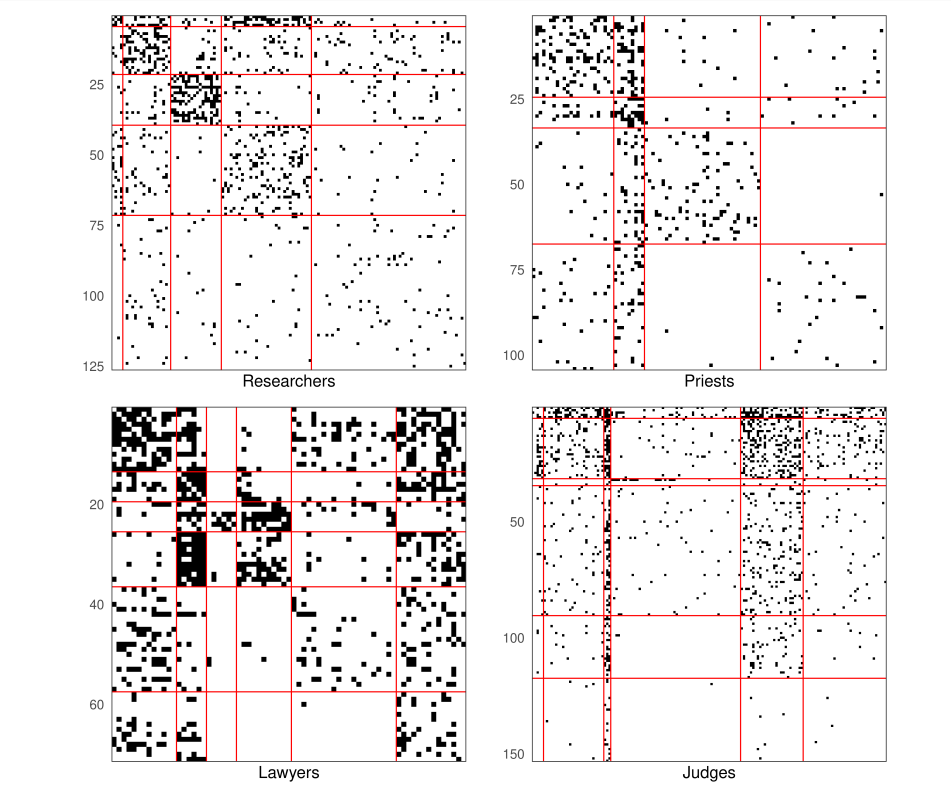
$$D(m, m') = \sum_{q,r \in \mathcal{Q}} \max\left(\hat{\pi}_q^m, \hat{\pi}_q^{m'}\right) \max\left(\hat{\pi}_r^m, \hat{\pi}_r^{m'}\right) \left(\frac{\hat{\alpha}_{qr}^m}{\hat{\delta}^m} - \frac{\hat{\alpha}_{qr}^{m'}}{\hat{\delta}^{m'}}\right)^2$$

- Use clustering algorithm on D (hierarchical clustering, k-medoids. . .)
- Compute $BIC\text{-}L_{colSBM}$ on obtained partition

Application to a Collection of Advice Networks

Application to advice networks (1)

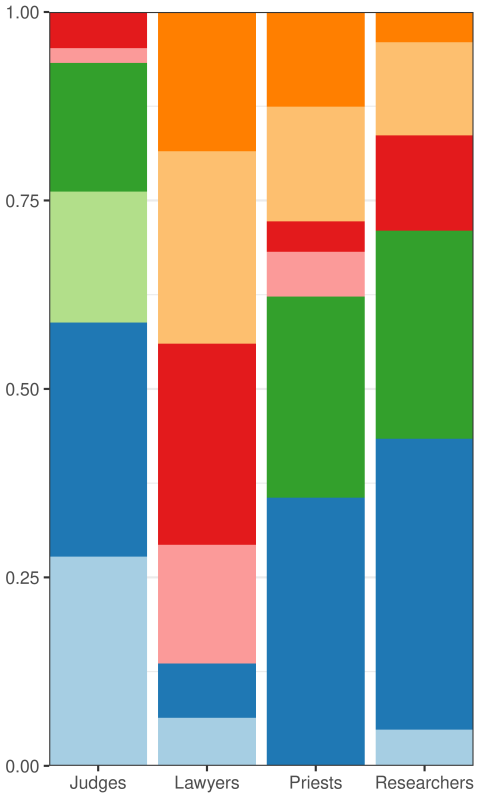
- 4 advice networks [3]
- $(126, 104, 71, 153)$ individuals in $(5, 4, 6, 6)$ SBM Blocks.
- Density: $(.061, .049, .18, .053)$

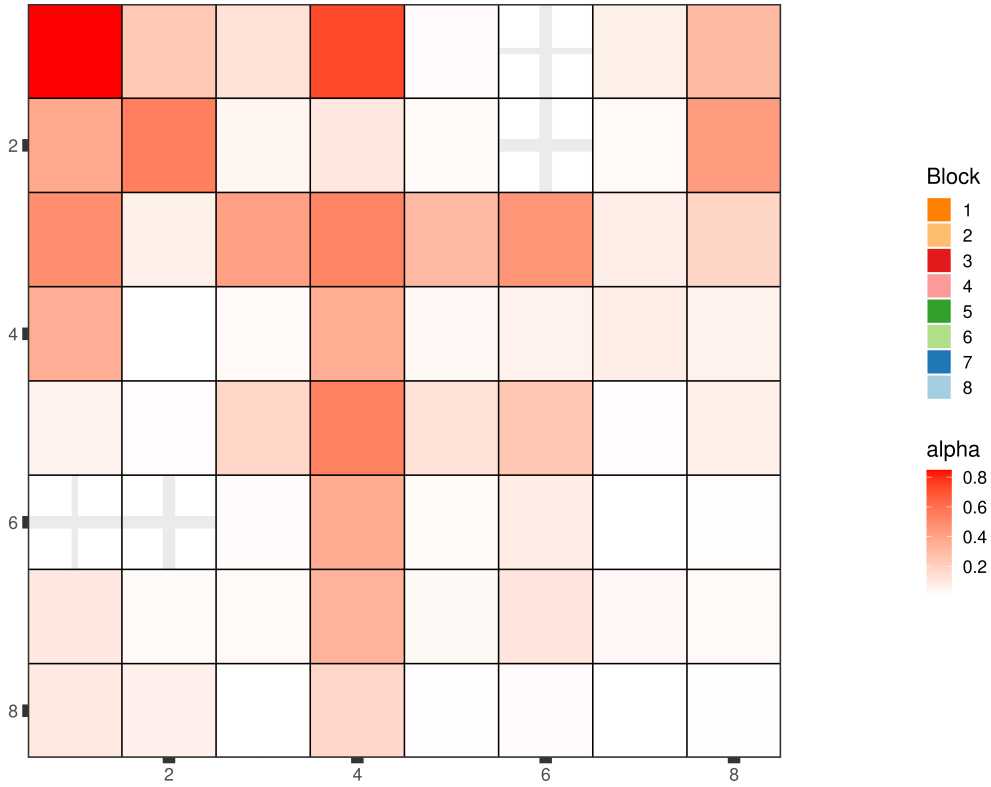


[3] Courtesy of E. Lazega

Application to advice networks (2)

- Modeling 4 networks with $\delta\pi$colSBM
- $ICL_{\delta\pi\text{colSBM}} \approx -11147 > -11209 \approx ICL_{SBM}$
- No good common structure for the other models

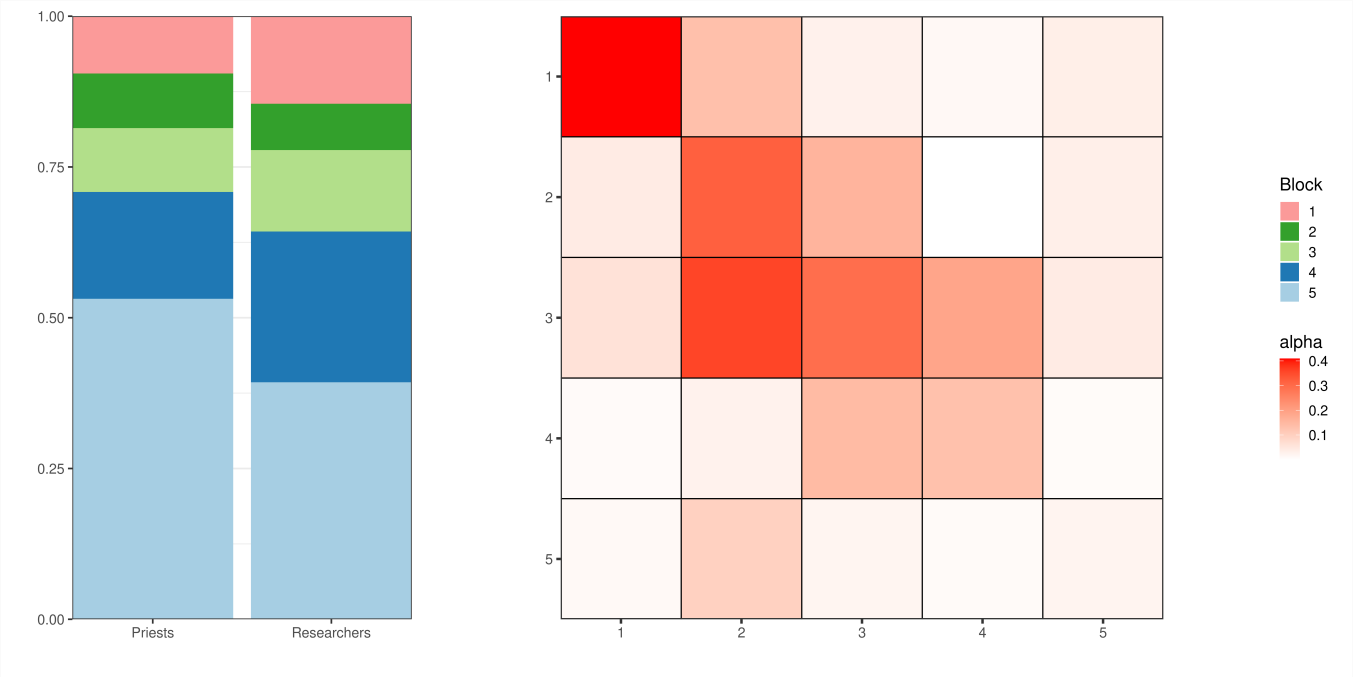




$\hat{\delta} = (1, 0.7, 0.45, .79)$

Application to advice networks (3)

- $\delta\pi$colSBM difficult to analyze
- Other colSBMs: structure of network with judges is different
- Best partition for πcolSBM: Priests-Researchers, Lawyers, Judges ($ICL_{\pi\text{colSBM}} \approx -11177$)



Predicting missing advices

Better prediction of advices between researchers with advice networks?

- Encoding proportion K of entries as `NA`
- Fit δcolSBMs (using Poisson model for inference purpose)
- Using information from different set of networks with δcolSBM
- $$\hat{p}_{ij}^{res} = \sum_{q,r \in \hat{\mathcal{Q}}_{res}} \hat{\mathbb{P}}_{\mathcal{R}}(Z_{iq}^{res} = 1)\hat{\mathbb{P}}_{\mathcal{R}}(Z_{jr}^{res} = 1)\hat{\delta}^{res}\hat{\alpha}_{qr}$$
- ROC AUC to judge quality of prediction

Predicting missing advices

Better prediction of advices between researchers with advice networks?

- Baseline is black dot (researchers on their own)
- Researchers, Lawyers information very insightful when K small
- Judges always bad except for large K