Statistical learning of collections of networks

Application to ecology and sociology

Saint-Clair Chabert-Liddell
Supervized by S. Donnet and P. Barbillon

17 March 2022

PhD defense

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

Introduction

A Stochastic Block Model for multilevel networks

Robustness of bipartite ecological interaction networks

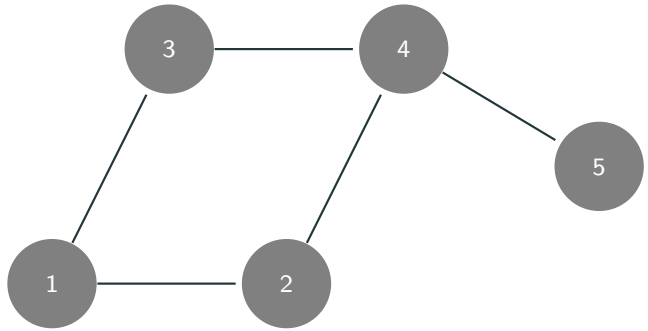
Finding common structures in a collection of networks

Introduction

3 basic types of networks

Simple undirected networks

- Networks with 1 type of nodes and interactions
- Undirected: Reciprocal interaction between nodes
  - Collaboration networks...

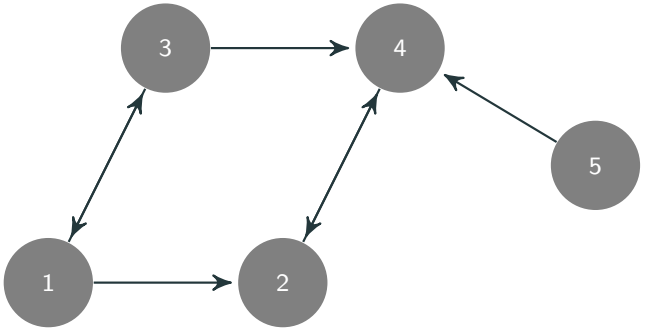


$$X = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

3 basic types of networks

Simple directed networks

- Networks with 1 type of nodes and interactions
- Directed: Interaction from one node to another
  - Ecology: Food webs...
  - Sociology: Advice networks...

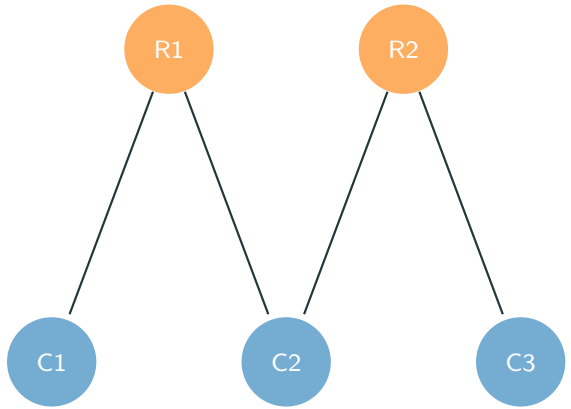


$$X = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

3 basic types of networks

Bipartite networks

- Networks with 2 types of nodes and 1 type of interaction
- Interaction between nodes of different types
  - Ecological interaction networks
    - Mutualistic (plant-pollinator, seed-dispersal. . .)
    - Antagonistic (host-parasite, herbivory . . .)
  - Social sciences
    - Contingency tables (seed-owner)
    - Affiliation networks

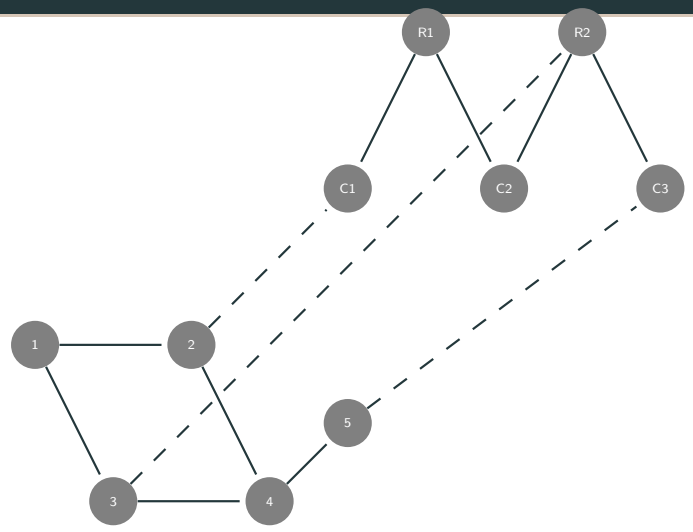


$$X = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

Collection of networks

Multilayer networks
Collection of networks

- Different types of interactions
- Linked through their nodes
  - Multiplex or temporal networks
  - Multipartite networks (ecosystem...)
  - Multilevel networks (socio-economic networks)



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

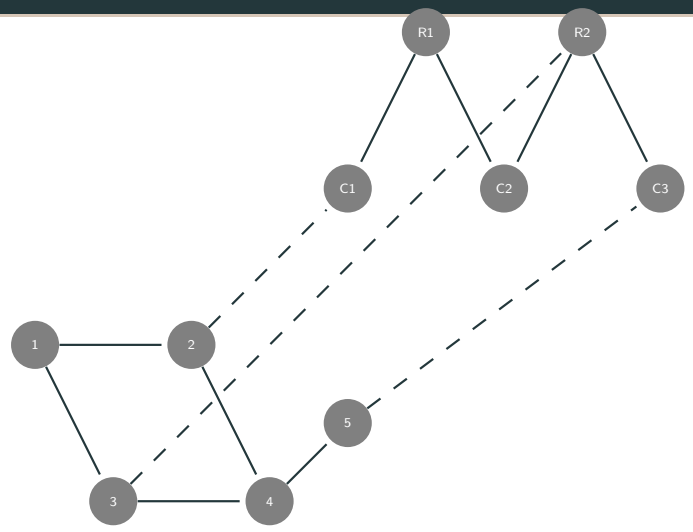
$$X^B = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$X^{AB} = ?$$

Collection of networks

Multilayer networks
Collection of networks

- Different types of interactions
- Linked through their nodes
  - Multiplex or temporal networks
  - Multipartite networks (ecosystem...)
  - Multilevel networks (socio-economic networks)



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

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

$$X^{AB} = ?$$

Collection of the same basic type

- Collection of bipartite networks (mutualistic, antagonistic...)
- Collection of simple networks (advice, food webs...)

Statistical Learning

Data

- A network X or a collection of networks $(X^1, \ldots, X^m, \ldots, X^M)$

Objectives

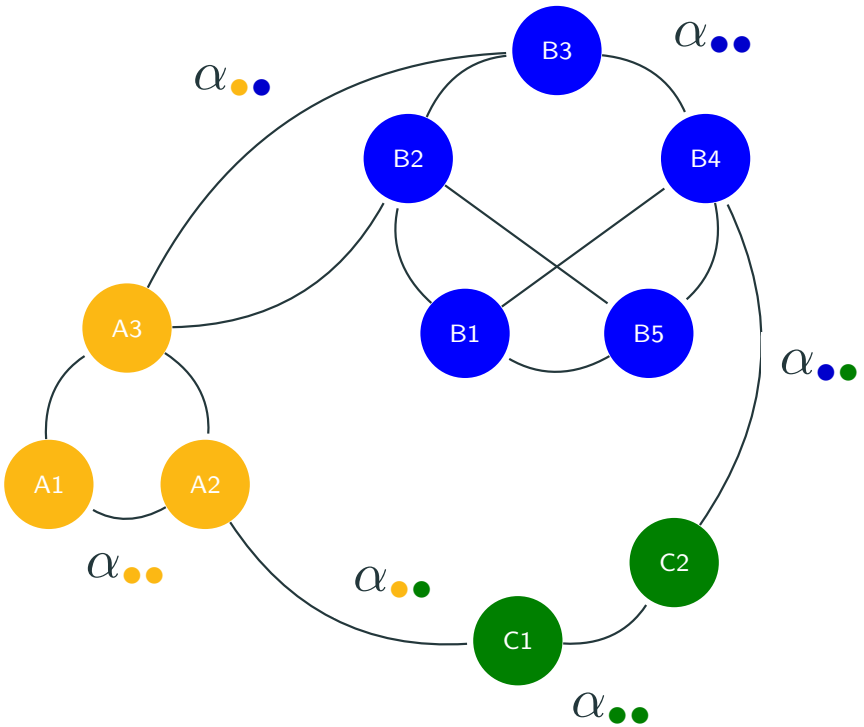
- Understand the structure/topology of the network
  - Heterogeneity in the connection
  - Group nodes with similar behavior (ecologically equivalent species...)
  - Unravel mesoscale structure (communities, core-periphery...)
- Predict missing interactions under an incomplete sampling

Method

- Probabilistic approach
  - Latent space model
    - Stochastic Block Model

Stochastic Block Models

Stochastic Block Model (SBM)



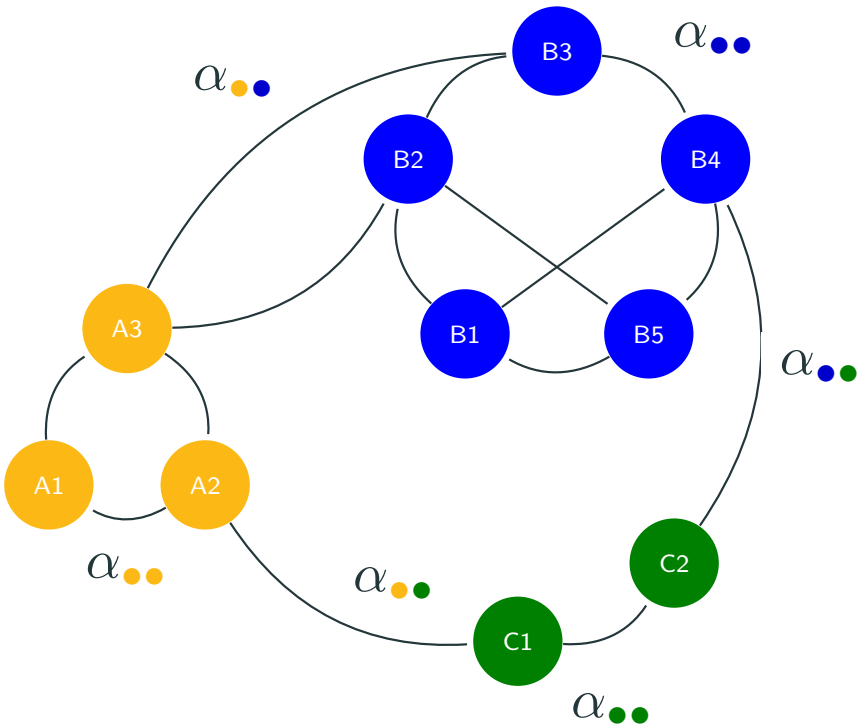
Mixture model for graph

n nodes into Q blocks

- Latent variable model
  $\mathbf{Z} = \{Z_1, \dots Z_n\} \in \{1, \dots, Q\}^n$
- $\mathbb{P}(Z_i = q) = \pi_q$

Snijders and Nowicki (1997)

Stochastic Block Model (SBM)



Mixture model for graph

n nodes into Q blocks

- Latent variable model
  $\mathbf{Z} = \{Z_1, \dots Z_n\} \in \{1, \dots, Q\}^n$
- $\mathbb{P}(Z_i = q) = \pi_q$
- $\mathbb{P}(X_{ii'} = 1 | Z_i = q, Z_{i'} = r) = \alpha_{qr}$

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

Snijders and Nowicki (1997)

Bipartite Stochastic Block Model

Bipartite Stochastic Block Model (biSBM)

- n_r row nodes into Q_r blocks and n_c column nodes into Q_c blocks
- $\mathbb{P}(Z_i = k) = \pi_k$ and $\mathbb{P}(W_j = q) = \rho_q$
- $\mathbb{P}(X_{ij} = 1 | Z_i = k, W_j = q) = \alpha_{kq}$

Notation: $X \sim \text{biSBM}_{n_r, n_c}(Q_r, Q_c, \pi, \rho, \alpha)$

Govaert and Nadif (2003)

Maximum likelihood inference

For fixed Q,

Objective Clustering of nodes $\mathbf{Z}$ and estimates of $\boldsymbol{\theta} = \{\pi, \alpha\}$

Method Maximum likelihood of the observed data

Problem Integrating complete likelihood on $\mathbf{Z}$ not tractable

$$\sum_{q_1,\ldots,q_n=1}^{Q} \mathcal{L}_\alpha(X|Z_1 = q_1, \ldots Z_n = q_n)\mathbb{P}_\pi(Z_1 = q_1, \ldots Z_n = q_n)$$

sum of Q^n terms

Daudin et al. (2008)

Maximum likelihood inference

For fixed Q,

Objective Clustering of nodes $\mathbf{Z}$ and estimates of $\boldsymbol{\theta} = \{\pi, \alpha\}$

Method Maximum likelihood of the observed data

Problem Integrating complete likelihood on $\mathbf{Z}$ not tractable

$$\sum_{q_1,\ldots,q_n=1}^{Q} \mathcal{L}_\alpha(X|Z_1 = q_1, \ldots Z_n = q_n)\mathbb{P}_\pi(Z_1 = q_1, \ldots Z_n = q_n)$$

sum of Q^n terms

Solution EM algorithm

Problem $\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{Z}|X)$ also not tractable

Solution Variational approach of the EM algorithm

Daudin et al. (2008)

Variational EM

Maximize a lower bound of the observed data log-likelihood

$$\begin{aligned}\ell_{\boldsymbol{\theta}}(X) &\geq \ell_{\boldsymbol{\theta}}(X) - KL(\mathcal{R}(\mathbf{Z})\|\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{Z}|X)) \\ &= \mathbb{E}_{\mathcal{R}}[\ell_{\boldsymbol{\theta}}(X,\mathbf{Z})] + \mathcal{H}(\mathcal{R}(\mathbf{Z})) \\ &= \mathcal{J}_{\boldsymbol{\theta}}(\mathcal{R}(\mathbf{Z}))\end{aligned}$$

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

$\mathcal{H}$ is the entropy

VEM algorithm

2–step iterative algorithm

VE Maximize $\mathcal{J}_{\boldsymbol{\theta}}(\mathcal{R}(\mathbf{Z}))$ w.r.t. $\mathcal{R}(\mathbf{Z})$

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

Choosing Q, model selection for SBM

Integrated Classified Likelihood (ICL)

- Penalized criterion for choosing the number of blocks
- Favors well separated blocks
  Asymptotic approximation of $\log \int_\theta \mathcal{L}_\theta(X, \mathbf{Z}) p(\theta) \mathrm{d}\theta$

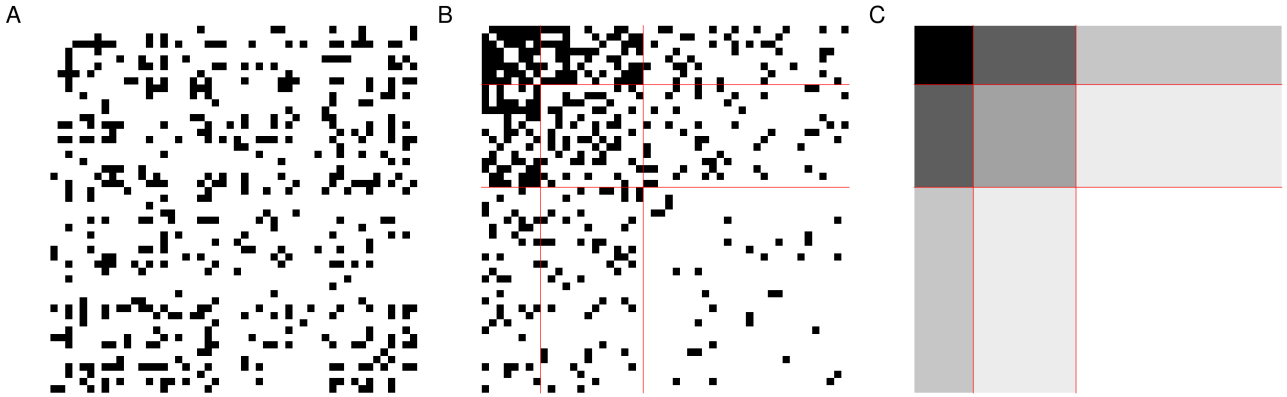
$$ICL(Q, \hat{\mathbf{Z}}) = \max_\theta \ell_\theta(X, \hat{\mathbf{Z}}) - \underbrace{\frac{1}{2} \frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2}}_{\alpha} - \underbrace{\frac{Q-1}{2} \log n}_{\pi}$$

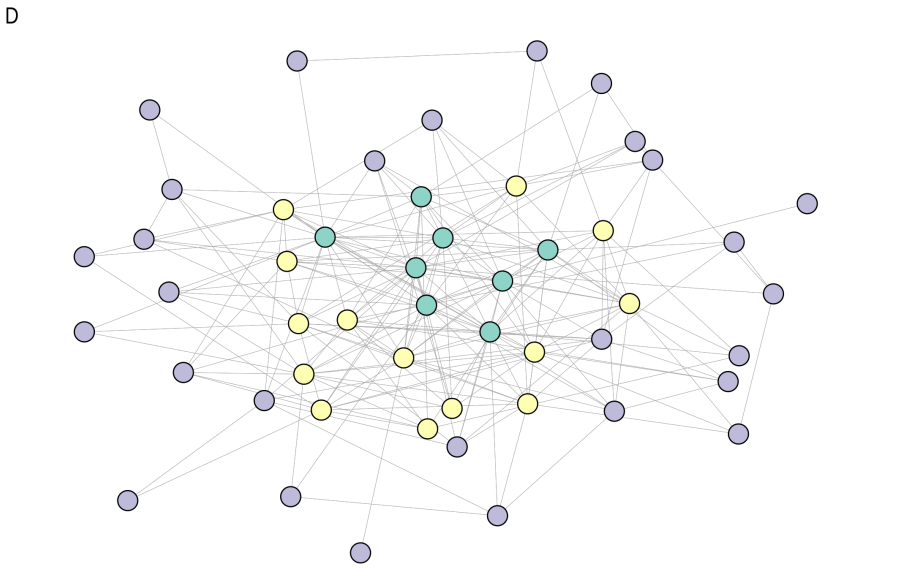
- Exact version available for some models[4]

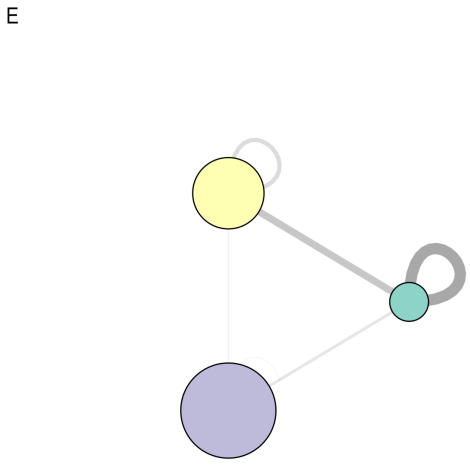
[4] Côme and Latouche (2015)
Biernacki et al. (2000)

Vizualisation of SBM (Core-periphery structure)

Simulated $X \sim \text{SBM}_{50}(3, \pi, \alpha)$ where $\pi = [.2, .3, .5]$ $\alpha = \begin{bmatrix} .8 & .5 & .2 \\ .5 & .3 & .1 \\ .2 & .1 & .05 \end{bmatrix}$



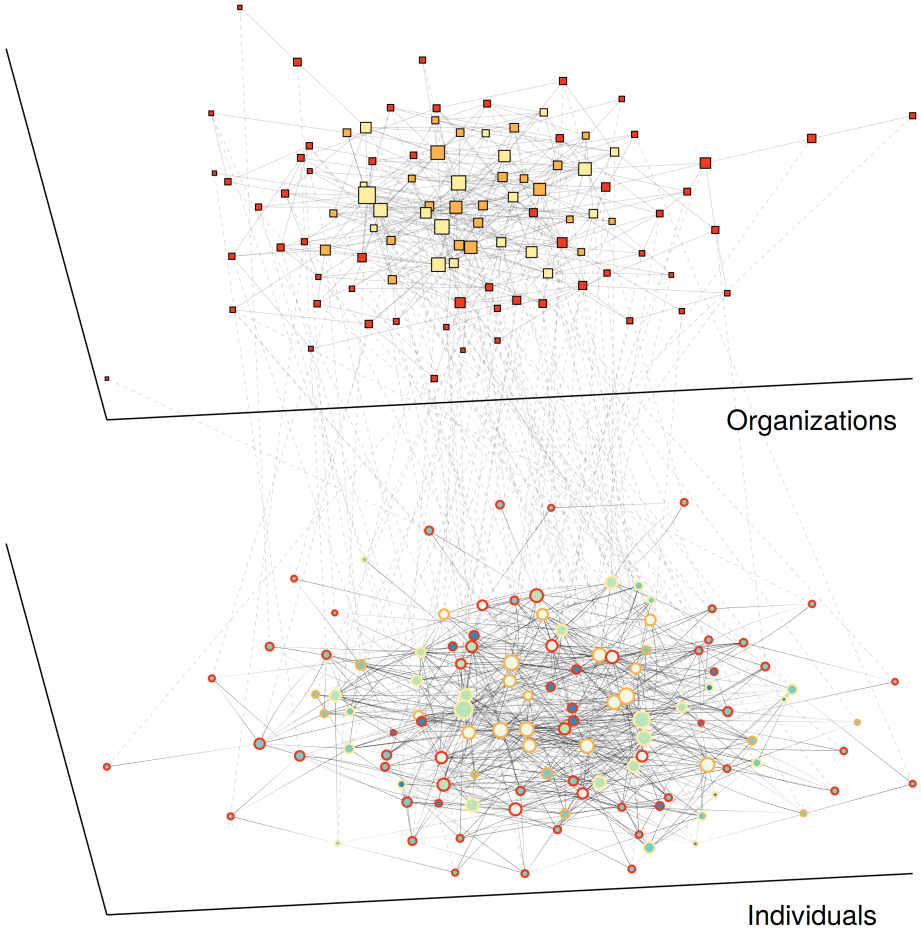




A Stochastic Block Model for multilevel networks

Motivation Dataset

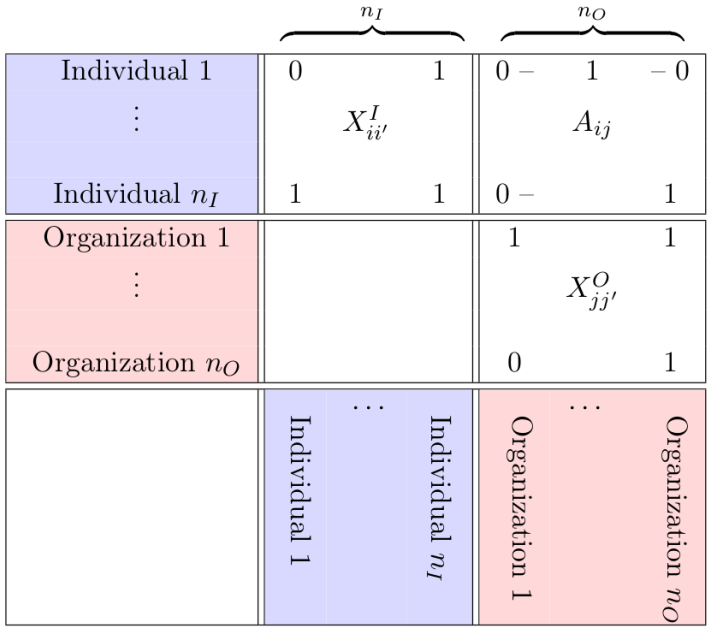
Economic and social networks in a television trade fair [6].



- Economic network: 109 companies signing deals (undirected interactions)
- Represented on the trade fair by representatives
- Social network: 128 representatives sharing advice (directed interactions)

[6] Brailly (2016)
Lazega et al. (2007)

Objective of this work



Data

X^I Interactions between individuals

X^O Interactions between organizations

A Affiliations of the individuals to the organizations

$A_{ij} = 1$ if i is affiliated to j

Only one affiliation per individual

Objectives

- Evaluate the influence of the inter-organizational level on the inter-individual level

Method

- Joint probabilistic model on $\mathbf{X} = \{X^I, X^O\}$ given A

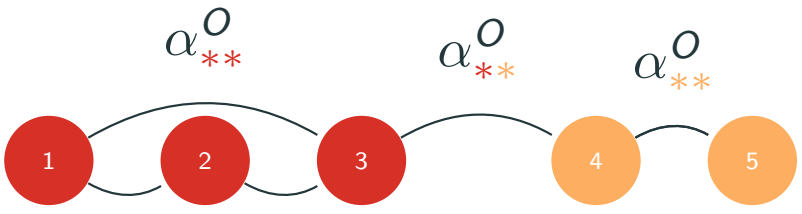
A new SBM model dedicated to multilevel networks

Inter-organizational Level

- n_O organizations into Q_O blocks
- Latent variables are independent
- $Z_j^O = \ell \Leftrightarrow j \in \ell, \ell \in \{1, \dots, Q_O\}$

$$\mathbb{P}(Z_j^O = \ell) = \pi_\ell^O$$

A new SBM model dedicated to multilevel networks



Inter-organizational Level

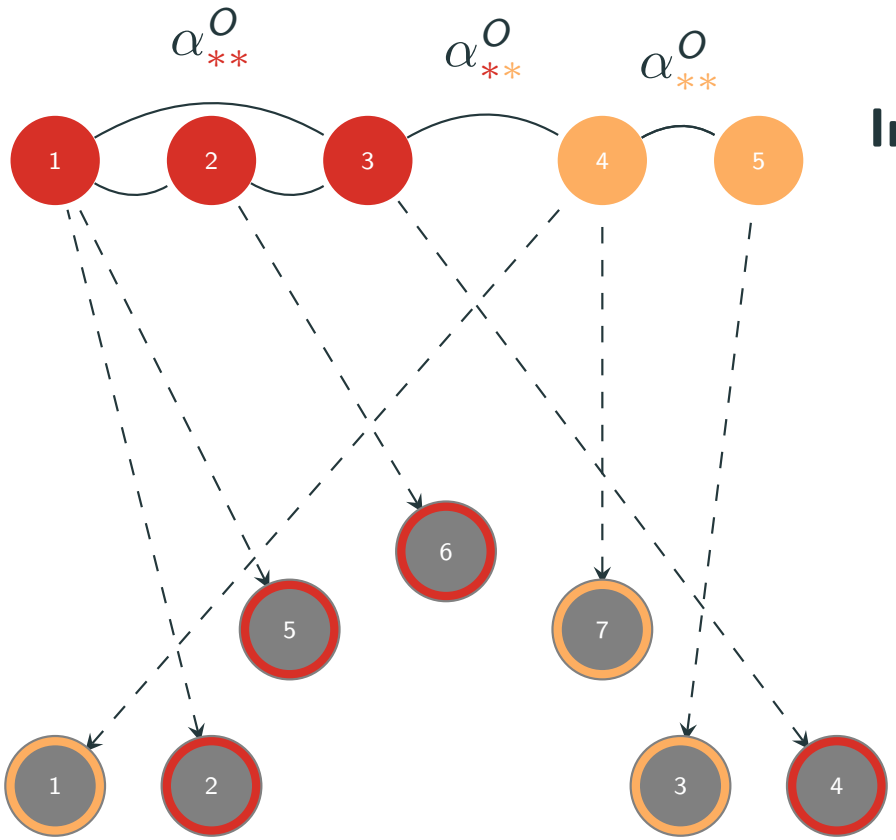
- n_O organizations into Q_O blocks
- Latent variables are independent
- $Z_j^O = \ell \Leftrightarrow j \in \ell, \ell \in \{1, \dots, Q_O\}$

$$\mathbb{P}(Z_j^O = \ell) = \pi_\ell^O$$

- Connections are independent given the latent variables

$$\mathbb{P}(X_{jj'}^O = 1 | Z_j^O = \ell, Z_{j'}^O = \ell') = \alpha_{\ell\ell'}^O$$

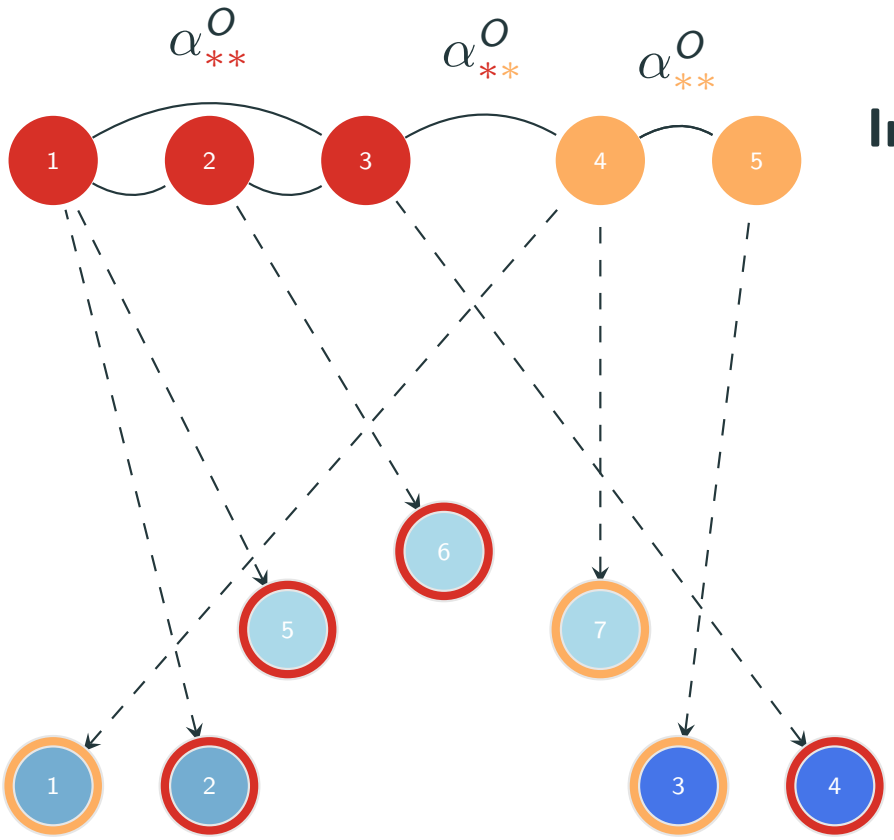
A new SBM model dedicated to multilevel networks



Inter-individual Level

- n_I individuals into Q_I blocks
- The block of an individual depends on the block of her/his organization through A

A new SBM model dedicated to multilevel networks

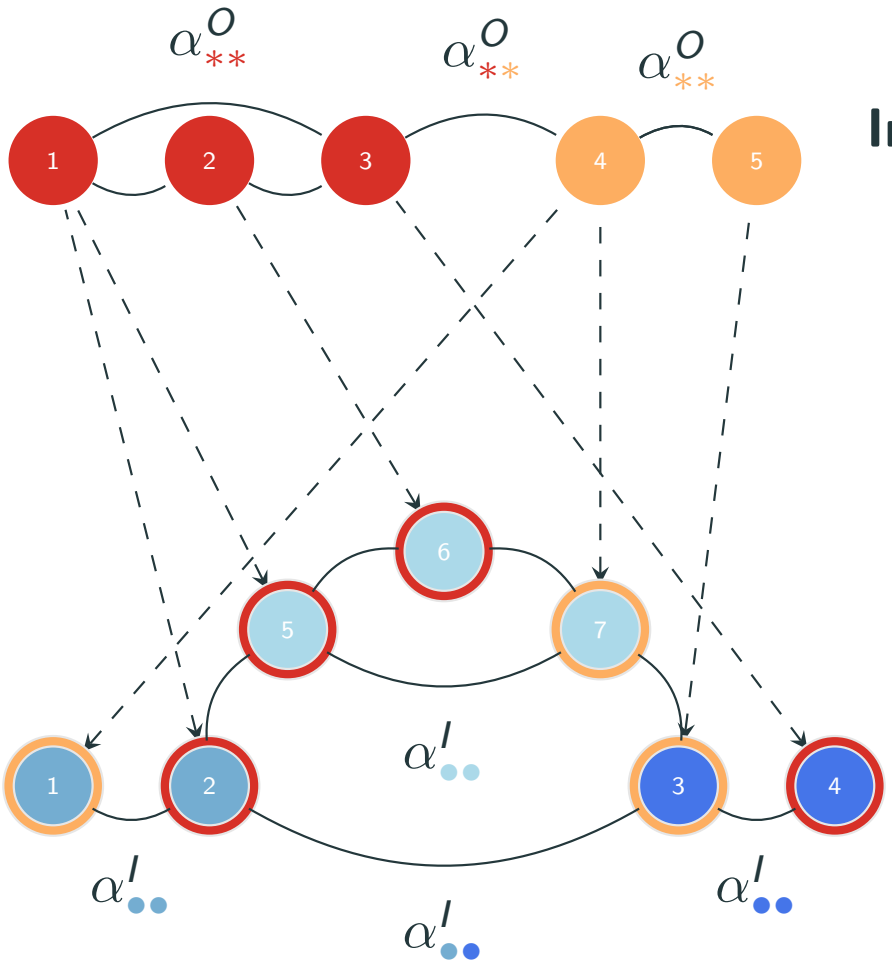


Inter-individual Level

- n_I individuals into Q_I blocks
- The block of an individual depends on the block of her/his organization through A
- $Z_i^I = k \Leftrightarrow i \in k, k \in \{1, \ldots, Q_I\}$

$$\mathbb{P}(Z_i^I = k | A_i = j, Z_j^O = \ell) = \gamma_{k\ell}$$

A new SBM model dedicated to multilevel networks



Inter-individual Level

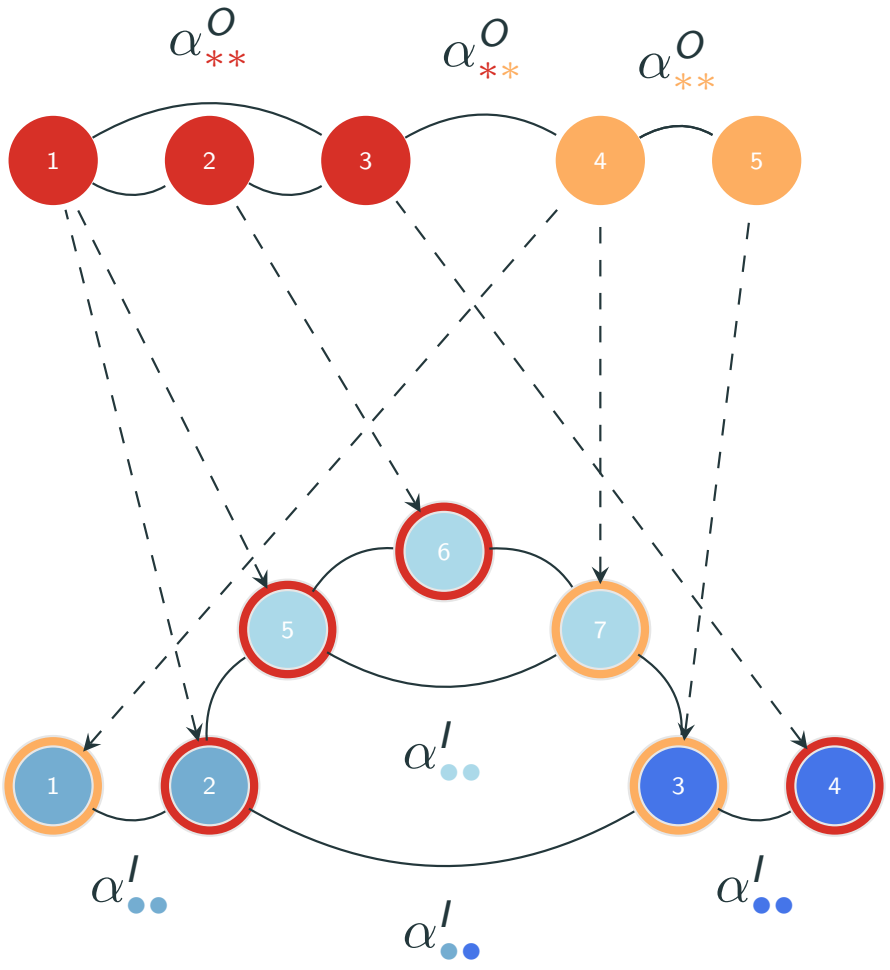
- n_I individuals into Q_I blocks
- The block of an individual depends on the block of her/his organization through A
- $Z_i^I = k \Leftrightarrow i \in k, k \in \{1, \dots, Q_I\}$

$$\mathbb{P}(Z_i^I = k | A_i = j, Z_j^O = \ell) = \gamma_{k\ell}$$

- Connections are independent given the latent variables

$$\mathbb{P}(X_{ii'}^I = 1 | Z_i^I = k, Z_{i'}^I = k') = \alpha_{kk'}^I$$

Independence between levels



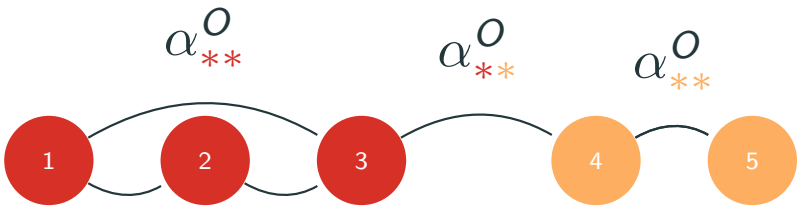
- Each column of γ is a probability vector

$$\gamma_{kl} = \mathbb{P}(Z_i^I = k | A_i = j, Z_j^O = \ell)$$

- If $\gamma_{kl} = \gamma_{kl'} \quad \forall k, \ell, \ell'$

$$\mathcal{L}(X^I, X^O | A) = \mathcal{L}(X^I)\mathcal{L}(X^O)$$

Independence between levels

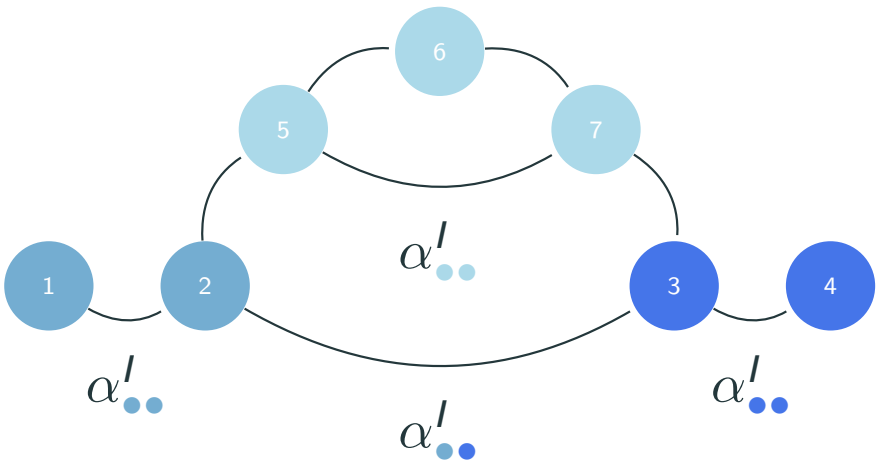


- Each column of γ is a probability vector

$$\gamma_{kl} = \mathbb{P}(Z_i^I = k | A_i = j, Z_j^O = \ell)$$

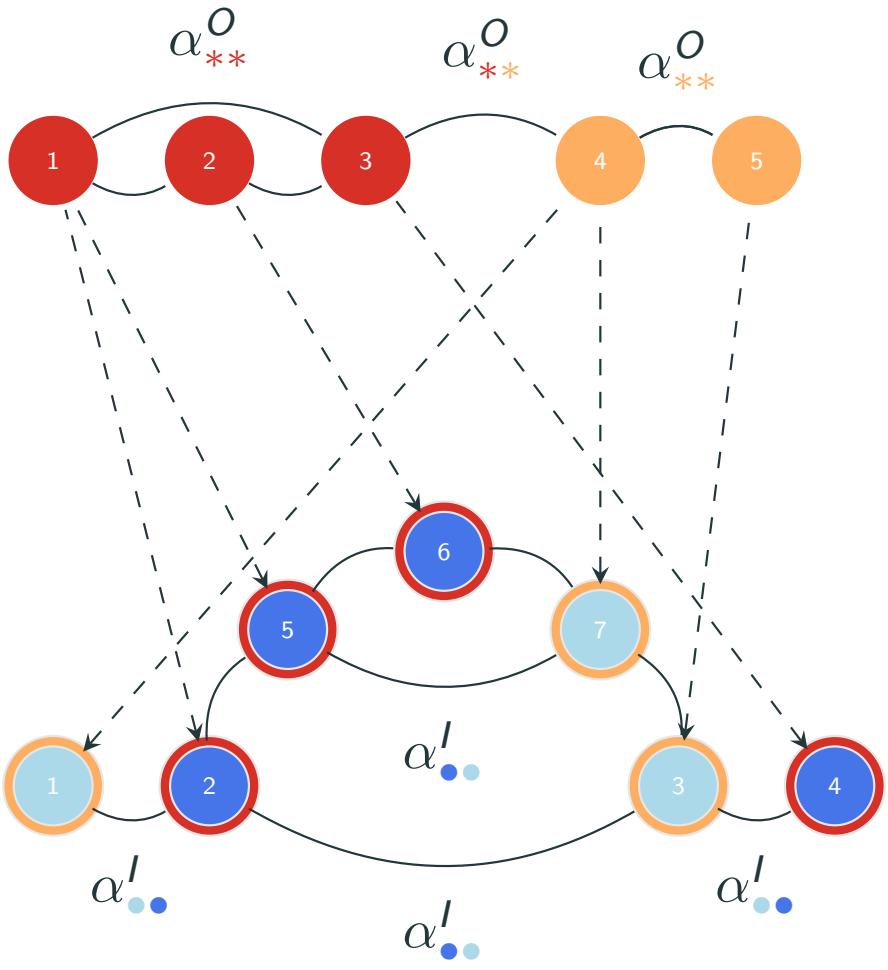
- If $\gamma_{kl} = \gamma_{kl'} \quad \forall k, \ell, \ell'$

$$\mathcal{L}(X^I, X^O | A) = \mathcal{L}(X^I)\mathcal{L}(X^O)$$



- Each level of the multilevel network is a SBM with $\pi^I = \gamma_{\cdot 1}$
- Organizational structure has no influence on the connections of individuals

Strong dependence between levels



- Each column of γ is a probability vector

$$\gamma_{k\ell} = \mathbb{P}(Z_i^I = k | A_i = j, Z_j^O = \ell)$$

- If $\forall \ell, \exists k, \quad \gamma_{k\ell} \approx 1$
- Blocks of individuals are determined by blocks of organizations
- ! Does not mean that the connection patterns are the same

Results

Mathematical results

Identifiability Under weak hypotheses

- On parameters
- On the number of nodes to number of blocks ratio

Results

Mathematical results

Identifiability Under weak hypotheses

- On parameters
- On the number of nodes to number of blocks ratio

Algorithmic results

Inference Variational EM

Model Selection Selecting the number of clusters (Q_I, Q_O)

- ICL criterion
- Step-wise procedure to navigate between models of different sizes

Results

Mathematical results

Identifiability Under weak hypotheses

- On parameters
- On the number of nodes to number of blocks ratio

Algorithmic results

Inference Variational EM

Model Selection Selecting the number of clusters (Q_I, Q_O)

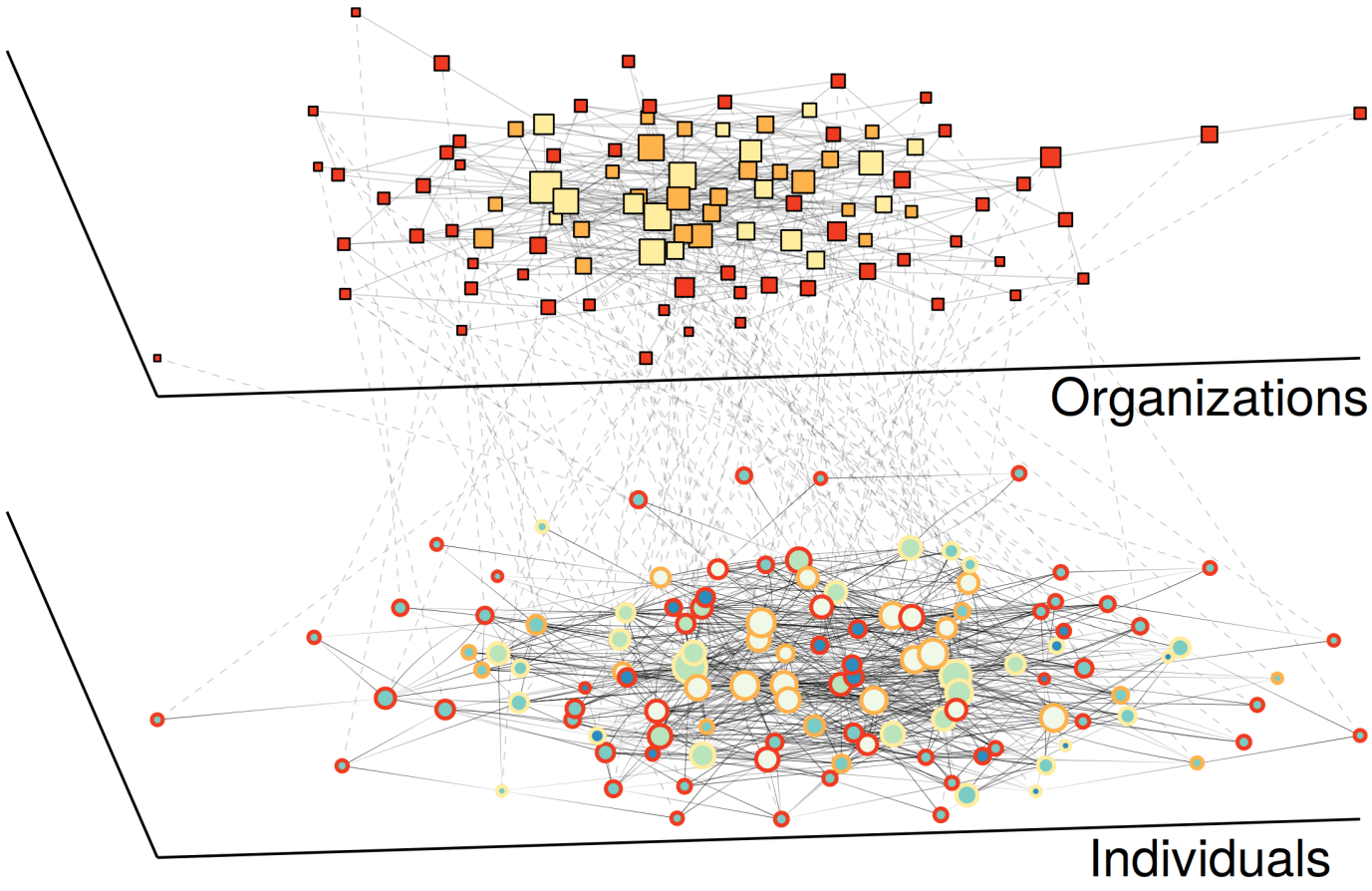
- ICL criterion
- Step-wise procedure to navigate between models of different sizes

Independence Between the two levels $\mathcal{L}(X^I, X^O|A) = \mathcal{L}(X^I)\mathcal{L}(X^O)$

- Condition on γ parameter
- ICL to state on independence

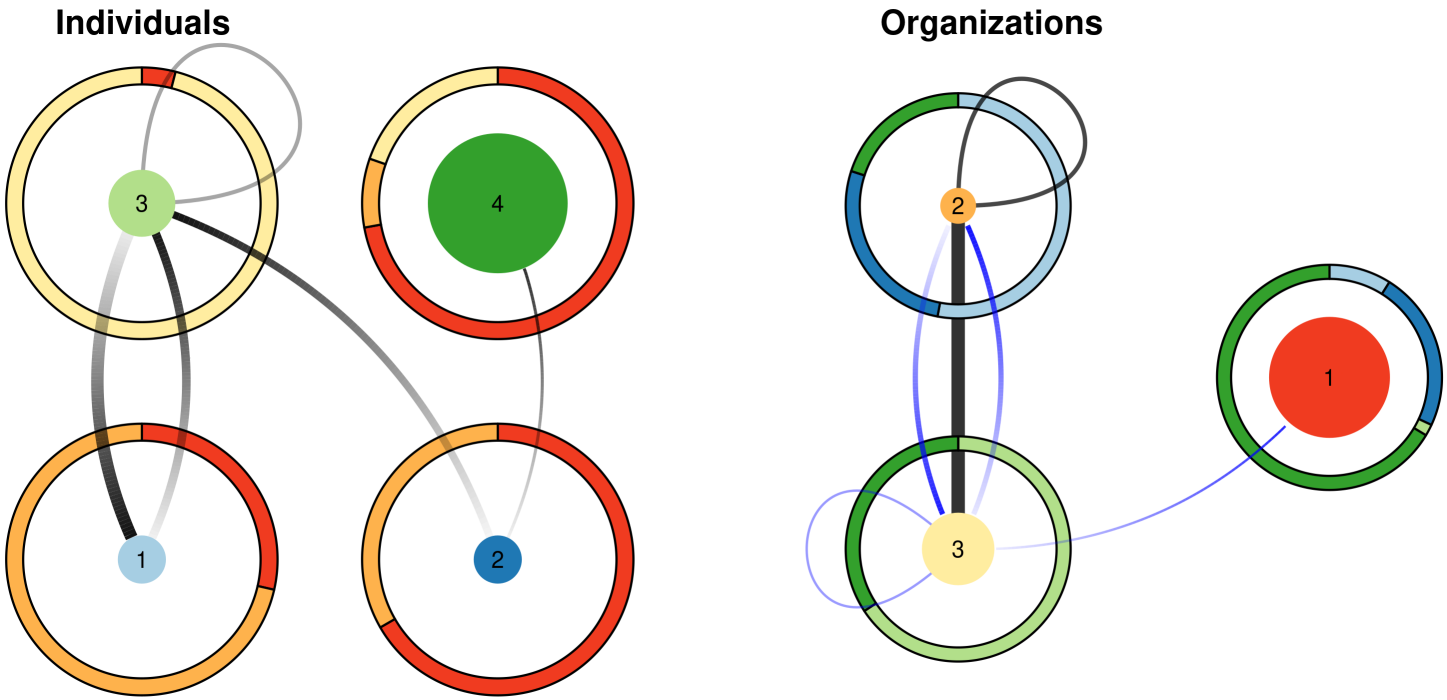
Application to a Television Program Trade Fair Dataset[8]

128 representatives (buyers and sellers) with directed interactions (advice) and 109 companies with undirected interactions (deal).



[8] Brailly (2016)

Dataset analysis



- Levels are interdependent, $(Q^I, Q^O) = (4, 3)$
- Core-periphery structure on X^O
- Mainly inter-block connections for X^I (except block 3, sub-group of sellers)
- Intra-block connection between individuals do not replicate the intra-block connections of their organizations (block 2 and 3)

Additional results

Numerical studies

- Simulation based on α^I and γ
- Strong dependence helps blocks recovery
- ICL good but conservative at detecting interdependence betwen levels
- Recovery of the blocks when both X^I and X^O are hard to infer
- Show improvement on prediction of missing links between individuals on the TV program data set compared to a single level SBM
  - X^O helps predicting links on X^I

Model extension

- To more than 2 levels
- To any number of affiliations (including none)

Diffusion

- S-C. C-L, P. Barbillon, S. Donnet et E. Lazega, (2021) A Stochastic block model approach for the anlysis of multilevel networks. *Computational Statistics & Data Analysis*, 158:107179
- `MLVSBM` available on CRAN and at `https://chabert-liddell.github.io/MLVSBM/`
  - Simulation and inference of multilevel networks
  - Handling of missing data on X^I and/or X^O
  - Prediction on missing dyads, missing links and spurious links
  - Extension to multi-affiliation datasets

Robustness of bipartite ecological interaction networks

Motivation & framework for robustness

Data

- Bipartite Ecological Interaction Networks $X \in \{0,1\}^{n_r \times n_c}$
  - Mutualistic: *Pollination*, Seed-Dispersal...
  - Antagonistic: Host-Parasite...

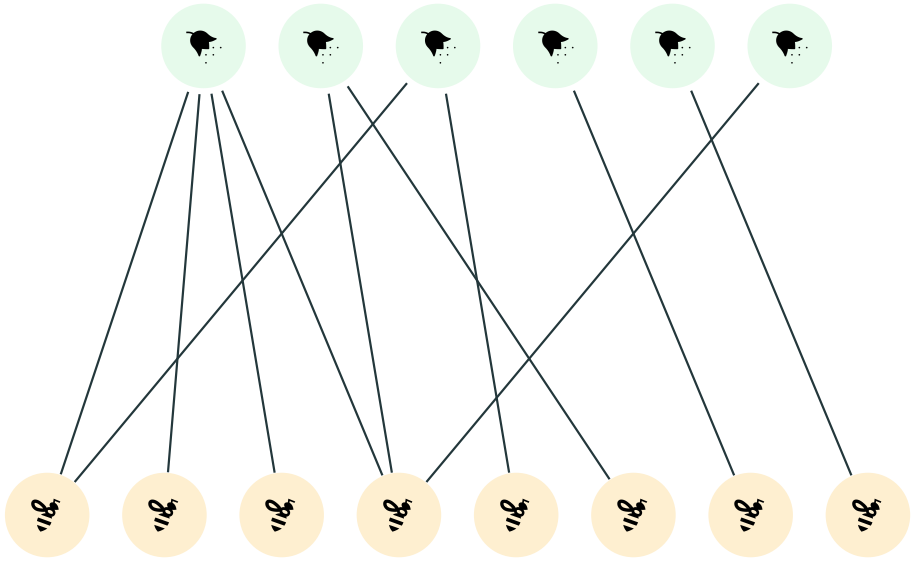
Objective

- Quantifying the impact of species loss on ecosystems

Method

- Counting the number of disconnected species in a network
- Extinction model
  - Primary extinctions sequence on row species (plants)
  - Secondary extinctions on column species (pollinators) with no connection

What is robustness?

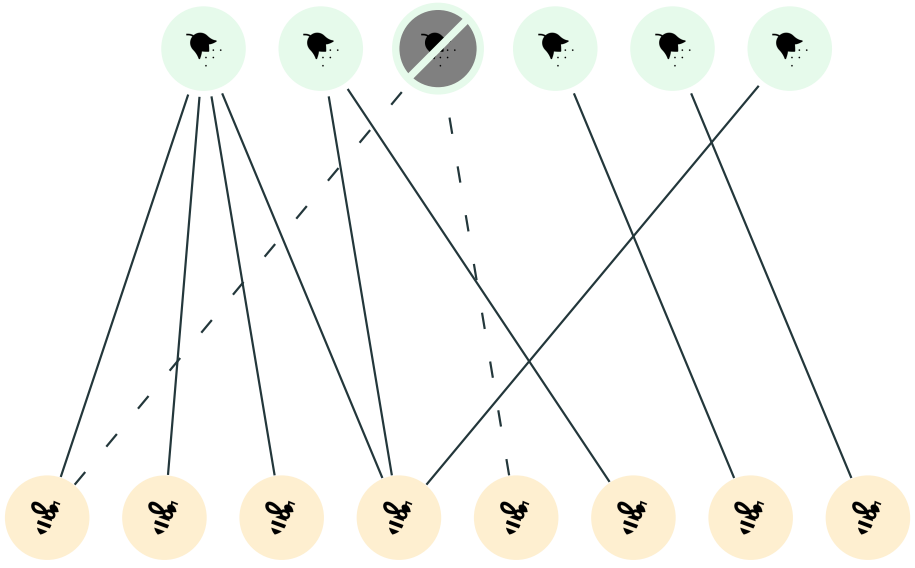


$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

s a sequence of plant extinctions

Illustration from `https://icons8.com/`

What is robustness?

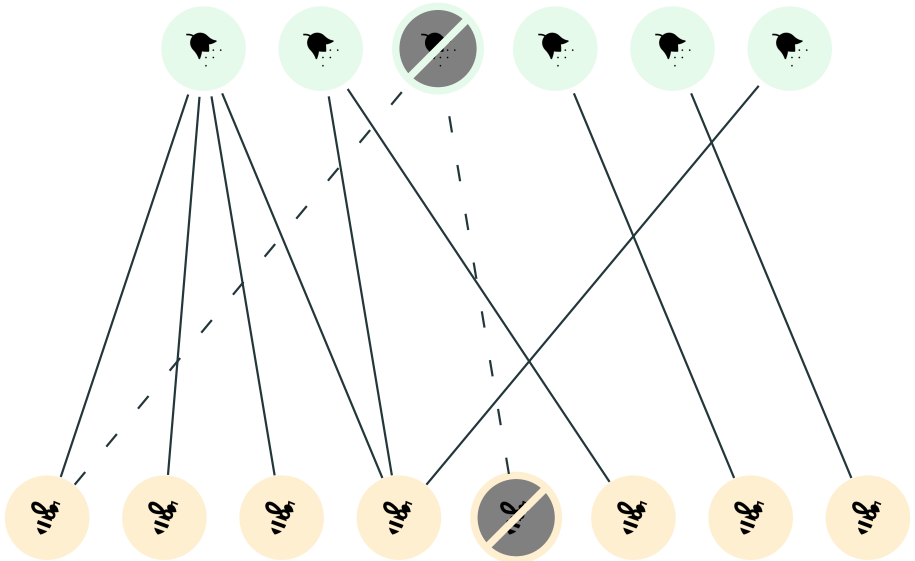


$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

s a sequence of plant extinctions

Illustration from `https://icons8.com/`

What is robustness?



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

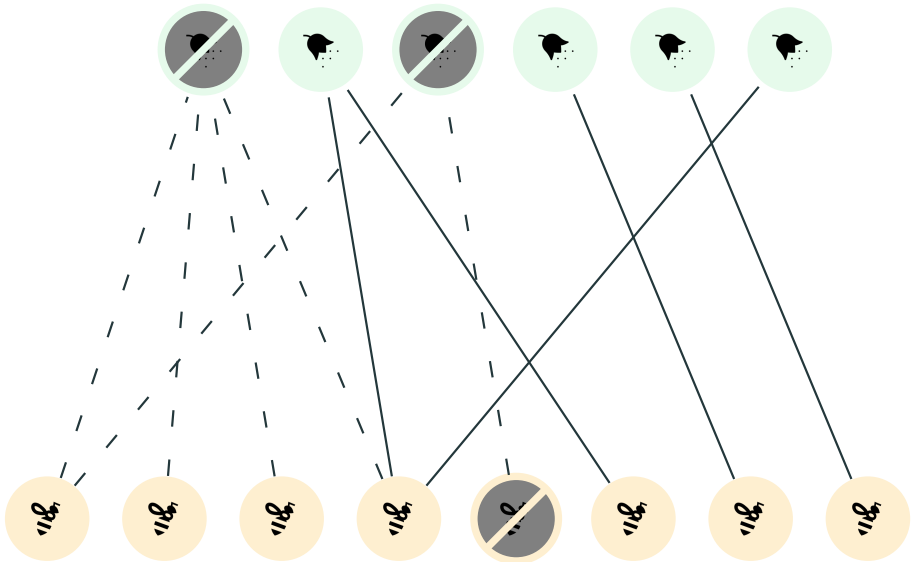
s a sequence of plant extinctions

After m primary extinctions, the proportion of remaining pollinators:

$$R(X, s, m) = 1 - \frac{1}{n_c} \sum_{j=1}^{n_c} \mathbf{1}_{\{\sum_{i=m+1}^{n_r} X_{s(i)j} = 0\}}$$

Illustration from `https://icons8.com/`

What is robustness?



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

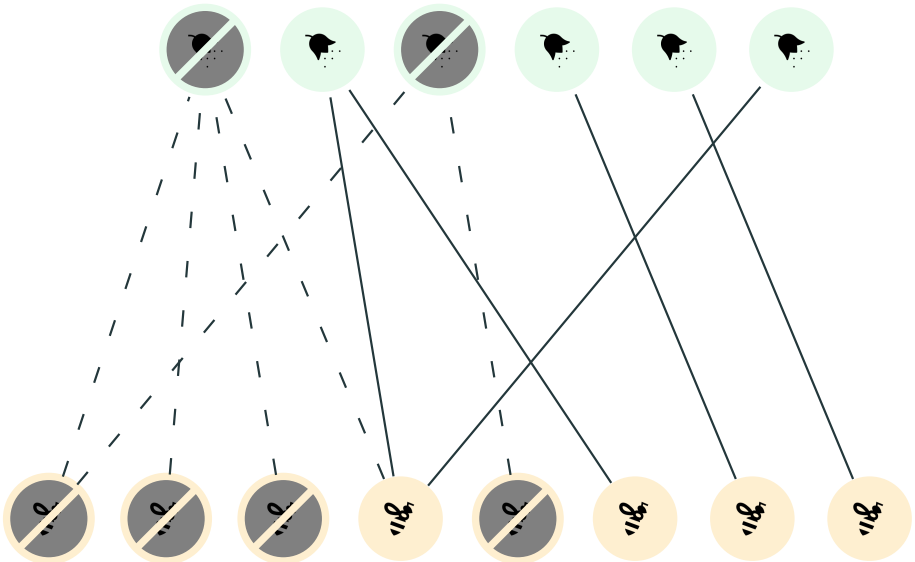
s a sequence of plant extinctions

After m primary extinctions, the proportion of remaining pollinators:

$$R(X, s, m) = 1 - \frac{1}{n_c} \sum_{j=1}^{n_c} \mathbf{1}_{\{\sum_{i=m+1}^{n_r} X_{s(i)j} = 0\}}$$

Illustration from `https://icons8.com/`

What is robustness?



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

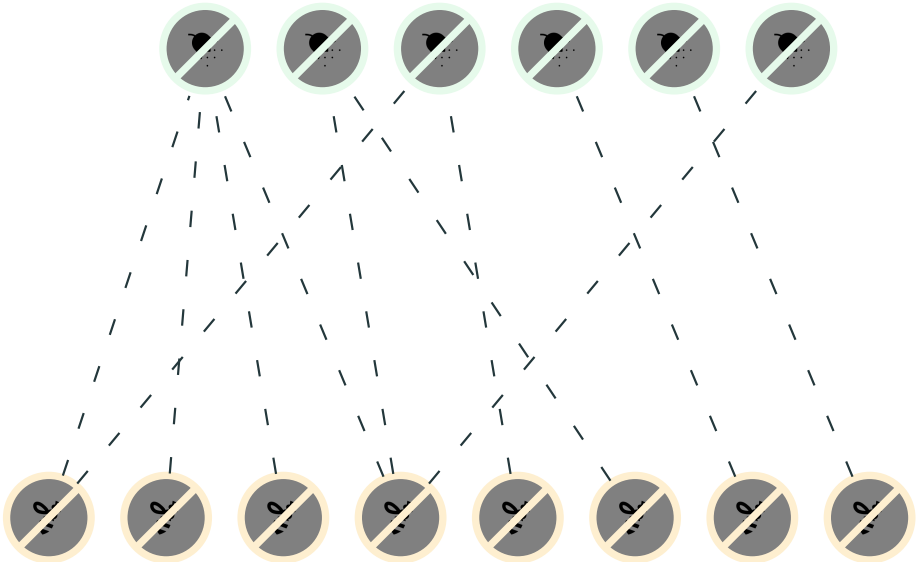
s a sequence of plant extinctions

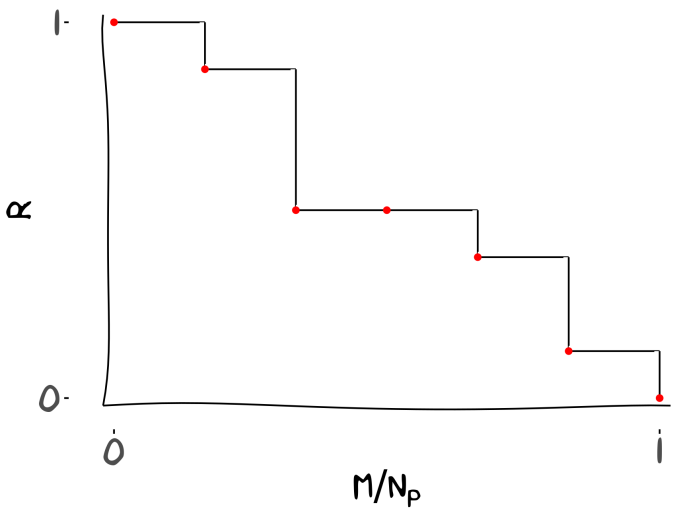
After m primary extinctions, the proportion of remaining pollinators:

$$R(X, s, m) = 1 - \frac{1}{n_c} \sum_{j=1}^{n_c} \mathbf{1}_{\{\sum_{i=m+1}^{n_r} X_{s(i)j} = 0\}}$$

Illustration from `https://icons8.com/`

What is robustness?





s a sequence of plant extinctions

After m primary extinctions, the proportion of remaining pollinators:

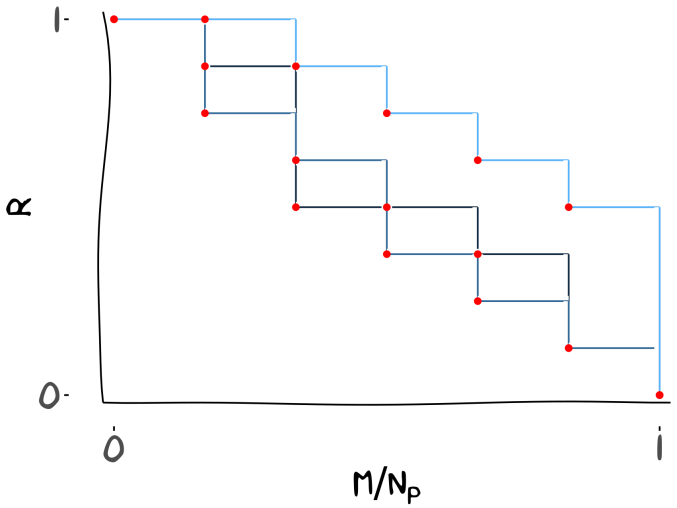
$$R(X, s, m) = 1 - \frac{1}{n_c} \sum_{j=1}^{n_c} \mathbf{1}_{\{\sum_{i=m+1}^{n_r} X_{s(i)j} = 0\}}$$

Illustration from `https://icons8.com/`

What is robustness?

s the realization of r.v. $S \sim \mathbb{S}$

- $\mathbb{S}$ uniform on all plants extinction sequences
- $\mathbb{S}$ by decreasing or increasing degree sequences



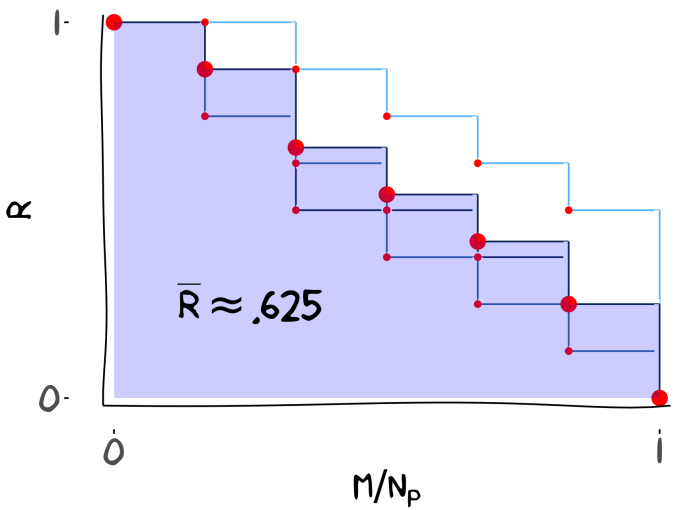
robustness function: average over the sequences of plants extinction sequences

$$m \mapsto R_{\mathbb{S}}(X, m) = \mathbb{E}_{\mathbb{S}}[R(X, S, m)].$$

What is robustness?

s the realization of r.v. $S \sim \mathbb{S}$

- $\mathbb{S}$ uniform on all plants extinction sequences
- $\mathbb{S}$ by decreasing or increasing degree sequences



robustness function: average over the sequences of plants extinction sequences

$$m \mapsto R_{\mathbb{S}}(X, m) = \mathbb{E}_{\mathbb{S}}\left[R(X, S, m)\right].$$

robustness statistic: the area under the curve

$$\overline{R}_{\mathbb{S}}(X) = \frac{1}{n_r} \sum_{m=0}^{n_r} R_{\mathbb{S}}(X, m)$$

How the topology of X is related to the robustness?

Network Model and Robustness

Our proposition:

Model X with a distribution $\mathbb{X}$

Models encompass some of the *network topology*.

- density, number of species, degree sequence, mesoscale structure...

$(\mathbb{X}, \mathbb{S})$ joint distribution over the network and the plant extinctions.

Robustness function under a network model:

$$R_{\mathbb{X},\mathbb{S}}(m) = \mathbb{E}_{(\mathbb{X},\mathbb{S})}[R(X, S, m)]$$

Robustness for biSBM

- Species from the same block are ecologically equivalent and exchangeable
- Exchangeable species are the same for biSBM Robustness
- Computation becomes tractable
- Analytical form to derive properties

For $\mathbb{S} = \mathbb{U}$ uniform on all row species and
$X \sim \text{biSBM}_{n_r, n_c}(Q_r, Q_c, \pi, \rho, \alpha)$:

$$R_{\pi,\rho,\alpha,n,\mathbb{U}}(m) = 1 - \sum_{q=1}^{Q_c} \rho_q \left(1 - \sum_{k=1}^{Q_r} \pi_k \alpha_{kq}\right)^{n_r - m}$$

Robustness for biSBM

For $\mathbb{S} = \mathbb{U}$ uniform on all row species and
$X \sim \text{biSBM}_{n_r,n_c}(Q_r, Q_c, \pi, \rho, \alpha)$:

$$R_{\pi,\rho,\alpha,n,\mathbb{U}}(m) \quad = \quad 1 - \sum_{q=1}^{Q_c} \rho_q \left(1 - \sum_{k=1}^{Q_r} \pi_k \alpha_{kq}\right)^{n_r - m}$$

- Variance also available in closed form
- Upper bound of robustness for given number of species and density
- Set of parameters which reach the upper bound and minimize the variance
- Robustness is an increasing function of the density and the number of plants

Other distributions

Extinction sequences distribution

- Extinction sequences which depend on the latent blocks
- Mimic targeted attack or extinction of ecologically equivalent group of species

Other distributions

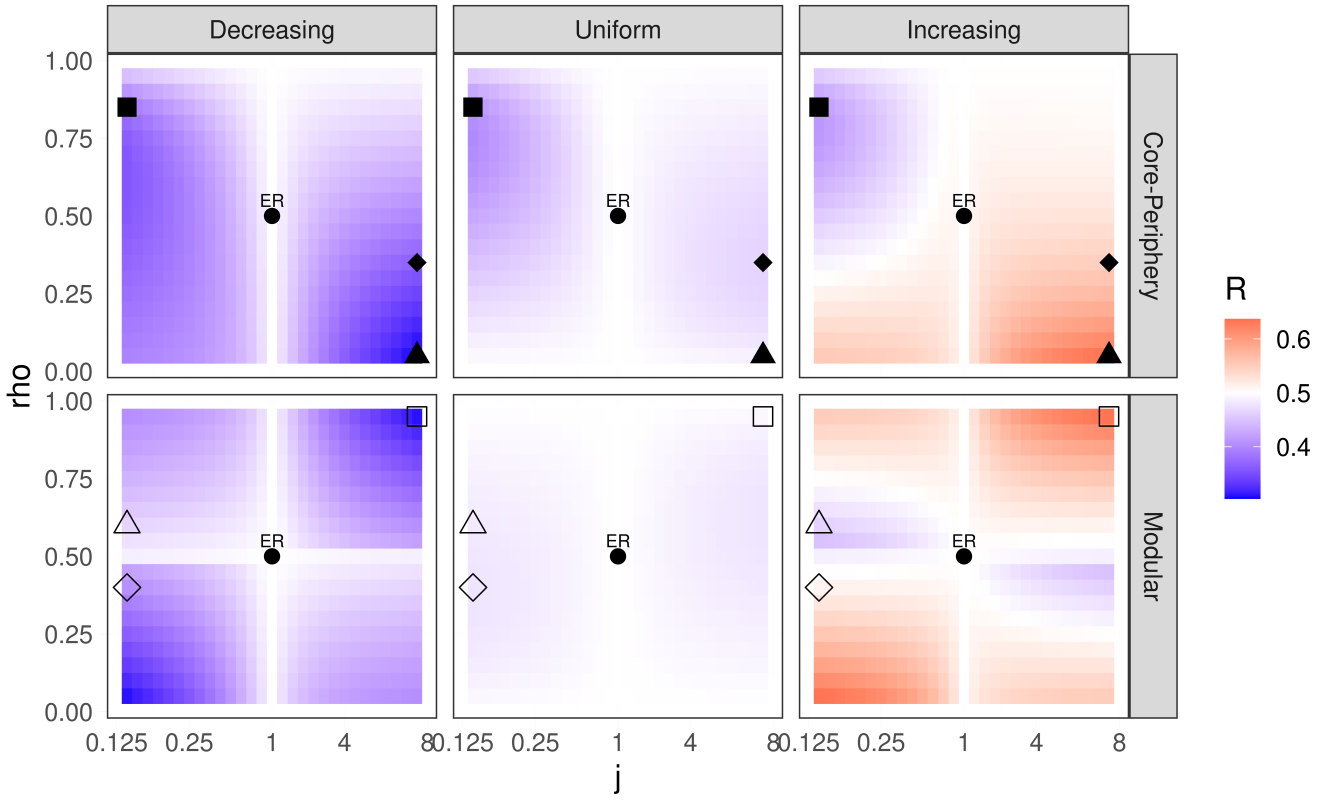
Extinction sequences distribution

- Extinction sequences which depend on the latent blocks
- Mimic targeted attack or extinction of ecologically equivalent group of species

Network distribution

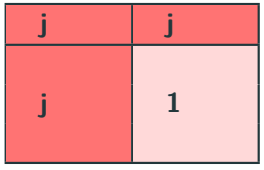
- Model with number of species and density (Erdős-Rényi)
- Model with the degree distribution of species (EDD)
- Model with both degree distribution and mesoscale structure (DCbiSBM)

Analysis of robustness and mesoscale structure

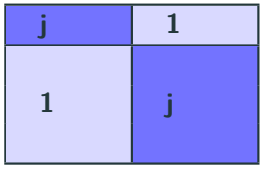


Fixed number of species and density

Core-periphery:



Modular:



Diffusion and additional work

- S-C. C-L, P. Barbillon et S. Donnet (2022), Estimating the robustness of a bipartite ecological networks through a probabilistic modeling *Environmetrics*, 33(2), e2709.
- `robber` available on `cran`
  `https://chabert-liddell.github.io/robber/`

Additional work

- Ability of different models to agree with the classical robustness
- biSBM allows through rescaling of the parameters to:

  Predict by computing the robustness of networks with incomplete sampling effort

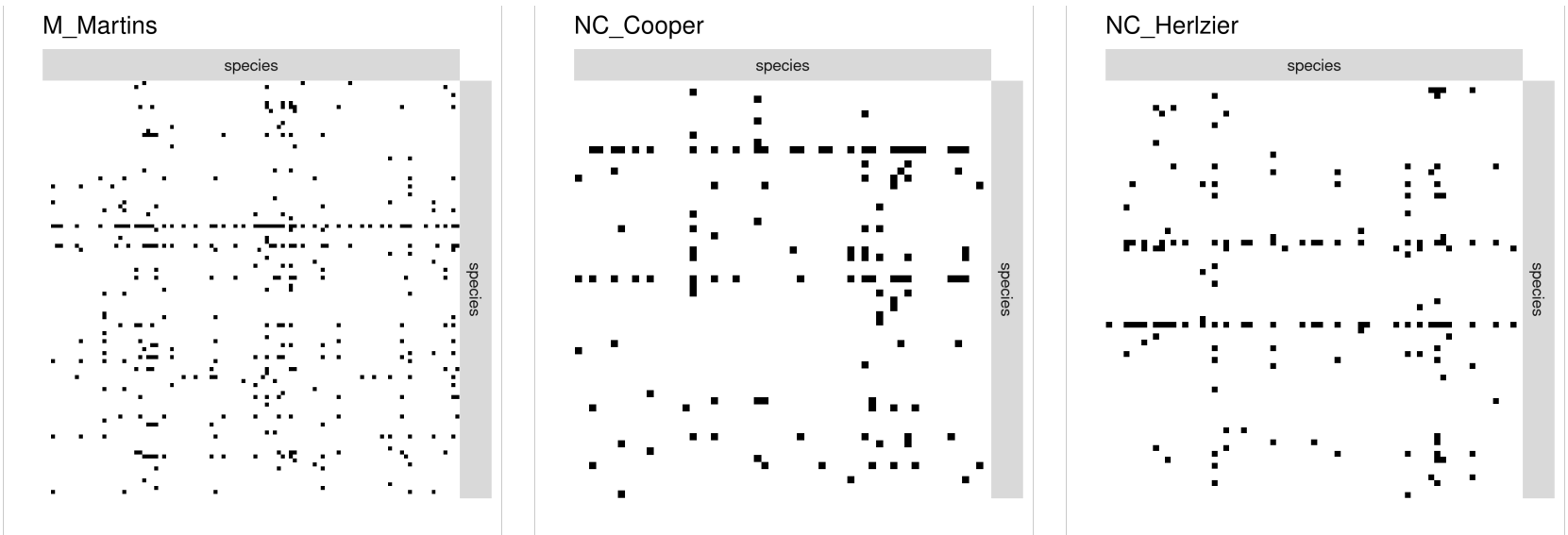
  Compare robustness in a collection of networks of different number of species and density

Finding common structures in a collection of networks

Motivation

Data

- Collection $\mathbf{X} = (\dots, X^m, \dots)_{m \in \mathcal{M}}$, $M = |\mathcal{M}|$ networks
- Same type:
  - Simple and directed: *Food webs*, Advice networks



Motivation

Data

- Collection $\mathbf{X} = (\dots, X^m, \dots)_{m \in \mathcal{M}}$, $M = |\mathcal{M}|$ networks
- Same type:
  - Simple and directed: *Food webs*, Advice networks

Objectives

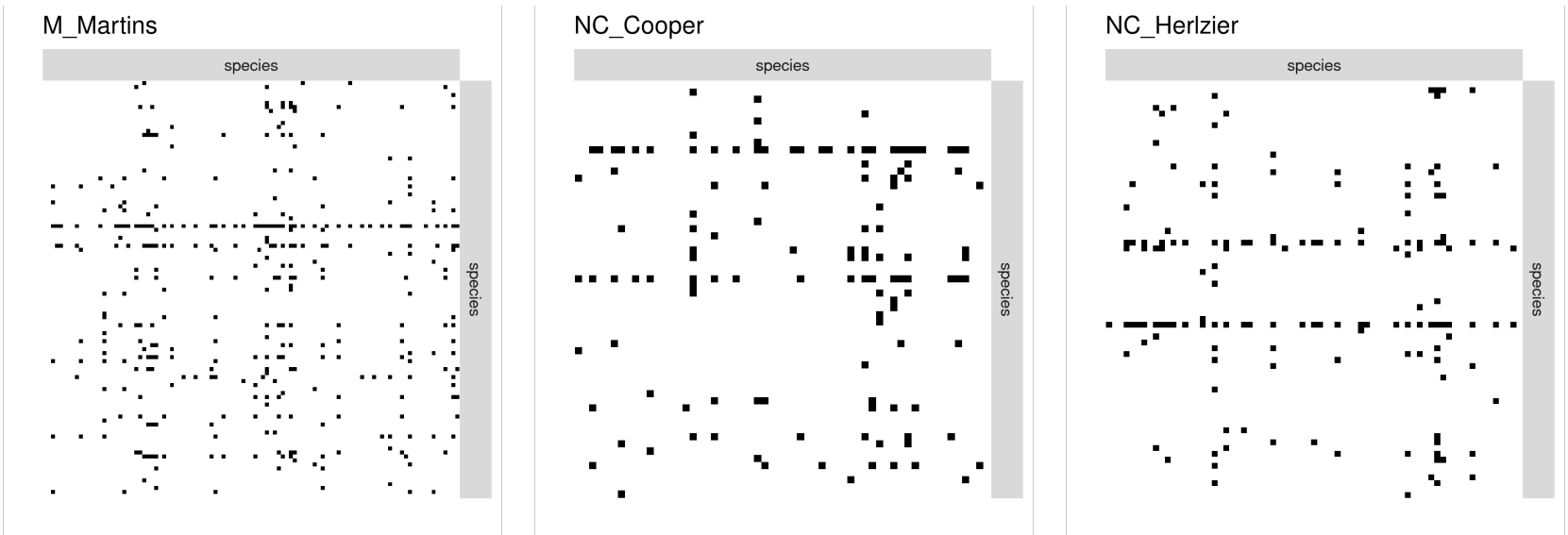
- Find common connectivity structures if relevant
- Identify the nodes playing the same ecological (social) roles
- Partition networks by connectivity structures

Method

- Joint modeling of the collection with *Stochastic Block Model*

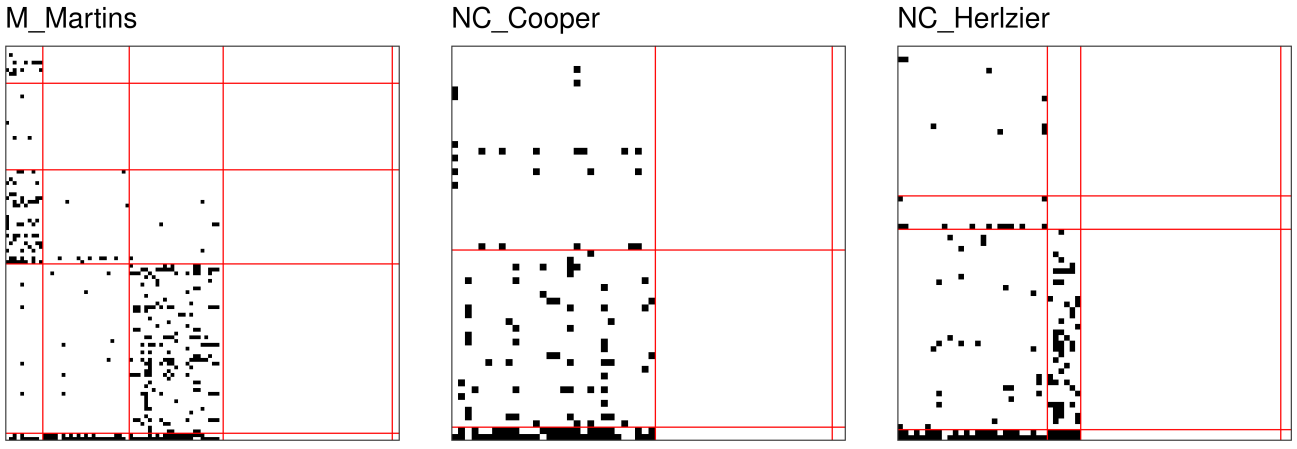
Three food webs

- Pine-forest stream food webs issued from Maine and North-Carolina (Thompson and Townsend, 2003)
- Involve respectively 105, 58 and 71 species.



- Look for similarities and differences between network structures.

Separated SBMs

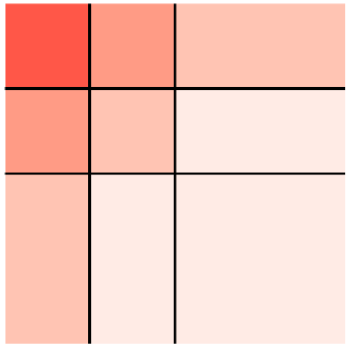


- Fitted SBM on each separately
- Reordered the matrices following the blocks
- Label the blocks following the average out-degrees order
- Bottom two groups: basal species (eaten by many species and not eating anybody)

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.
- (X^m) independent.
- $$X^m \sim \mathrm{SBM}_{n_m}(Q_m, \pi^m, \alpha^m)$$
- Conditions on the parameters $(\pi^m)_{m \in \mathcal{M}}$ and $(\alpha^m)_{m \in \mathcal{M}}$

First naive model

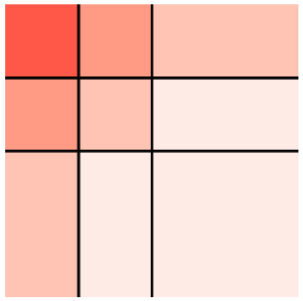


iid-colSBM

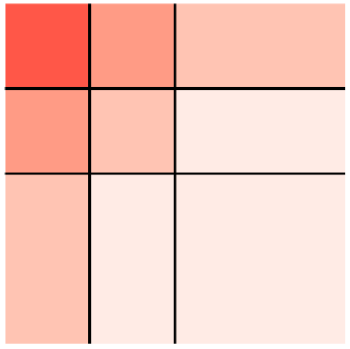
$$X^m \sim \mathrm{SBM}_{n_m}(Q, \pi, \alpha)$$

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

- Same blocks proportions
- Same connectivity structure



First naive model

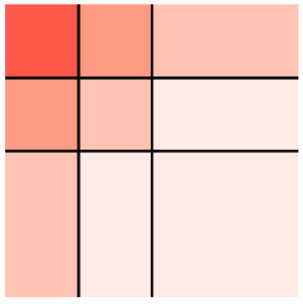


iid-colSBM

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

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

- Same blocks proportions
- Same connectivity structure

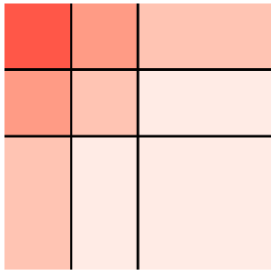


- i.i.d. assumption too strict for most datasets, 2 new relaxations:
  - Free proportion of blocks between networks
  - Density varies between networks

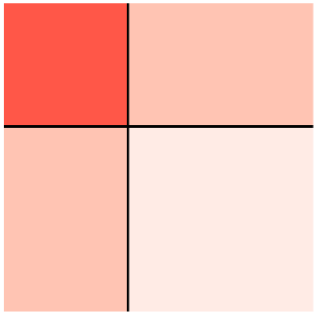
A first relaxed model : π-colSBM

π-colSBM

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



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



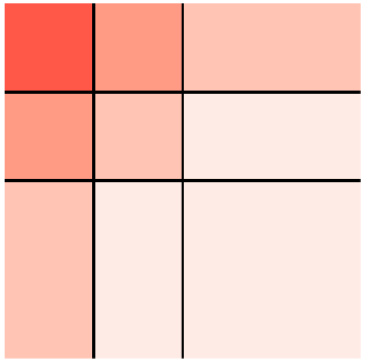
On the block proportions

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

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

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

Varying density model: δ-colSBM

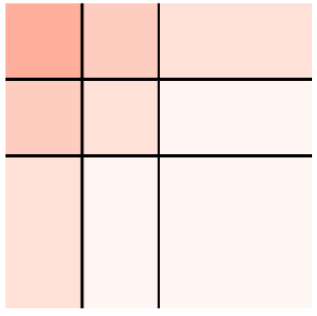


δ-colSBM

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

with $\pi_q > 0$.

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



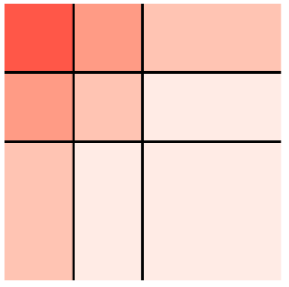
- Mimics differences of effort sampling or abundances

Varying density and block proportion model: $\delta\pi$-colSBM

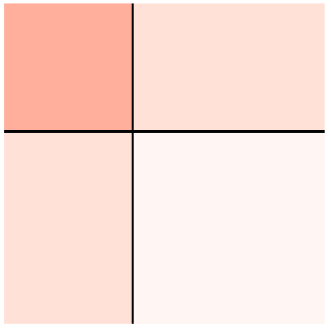
$\delta\pi$-colSBM

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

with $\pi_q^m \geq 0$



- Same connectivity structure α
- Specific proportions of blocks in each network
- Network specific density density parameter. $\delta^1 = 1$



  - Most flexible model

Results

Mathematical results

Identifiability Already proven for separated SBMs (Celisse et al., 2012)

- Proven for all 4 colSBMs
  - Trivial for *iid*-colSBM and δ-colSBM
  - More demanding for π-colSBM and $\delta\pi$-colSBM because of empty blocks (unknown support S)

Algorithmic results

Variation EM For fixed Q, support S

- Introduce stochasticy in the V-EM algorithm to avoid local maximum (VE–step are independent for each network)
- (δ–$\delta\pi$)colSBM: M–Step not explicit for Bernoulli model

Algorithmic results

Variation EM For fixed Q, support S

- Introduce stochasticy in the V-EM algorithm to avoid local maximum (VE–step are independent for each network)
- (δ–$\delta\pi$)colSBM: M–Step not explicit for Bernoulli model

Model selection Choosing Q

- BIC like criterion to not penalize the entropy of fuzzy clustering
- Adapted to allow for empty blocks

$$BIC\text{-}L(\mathbf{X}, Q) = \mathcal{J}(\hat{\mathcal{R}}(\mathbf{Z}), \hat{\boldsymbol{\theta}}) - \text{pen}_{\text{colSBM}}$$

- Forward-backward procedure to navigate between model
- Threshold on π^m to find support S for a given Q

Partitioning a collection of networks

- BIC-L to assess relevance of common structure (to choose between coISBM and separated SBMs)
- Different networks of the collection share different structures
- Group M networks sharing the same structure into one of G clusters

$$X^m \sim \text{SBM}_{n_m}(Q^g, \pi^m, \alpha^g), \quad g \in \{1, \ldots, G\} \qquad \text{(for } \pi\text{–coISBM)}$$

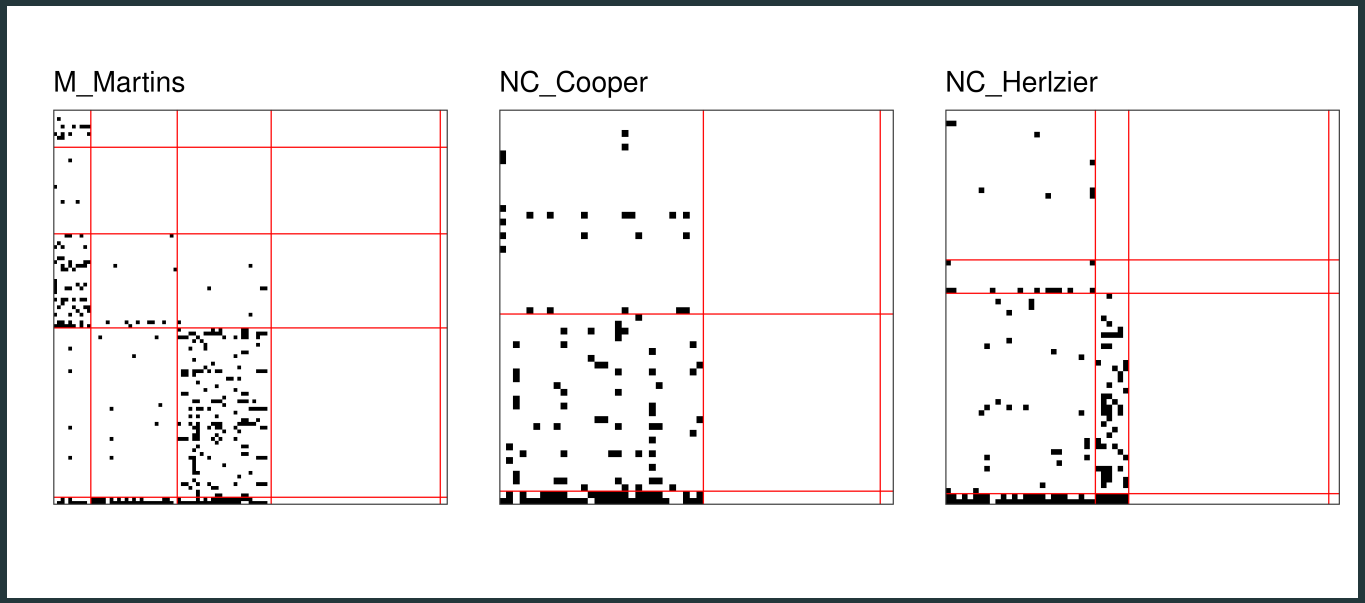
- Find the partition with the highest *BIC-L*
- Recursive partitioning to cluster the networks of the collection

Numerical results

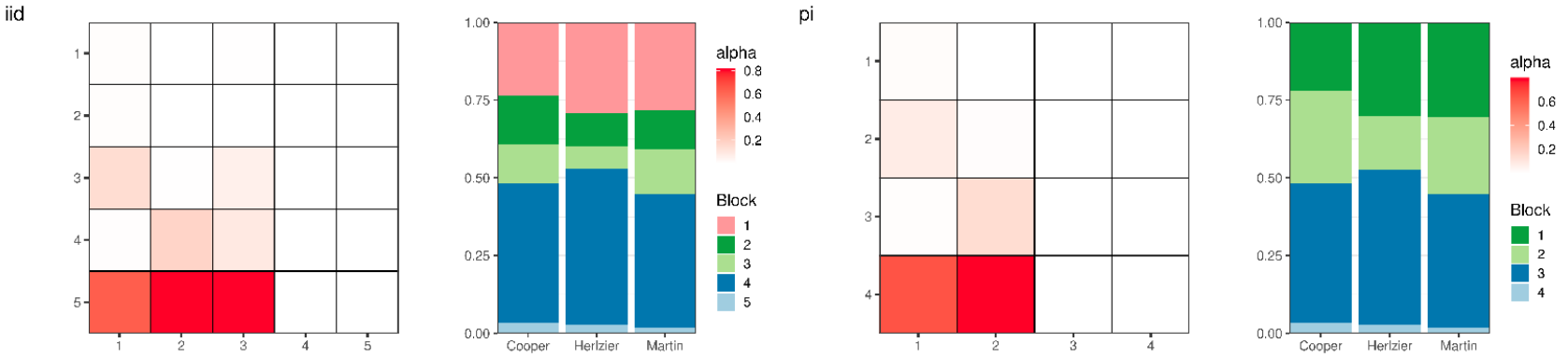
Numerical studies

- To test procedures on the ability to recover:
  - Connectivity parameter (α)
  - Number of blocks Q and support S
  - Block memberships (ARI)
  - True model (SBM vs πcolSBM vs *iid*colSBM)
  - Partition of networks
- Ability to find finer block structures than separated SBMs

Application on the stream food webs



colSBMs on stream food webs

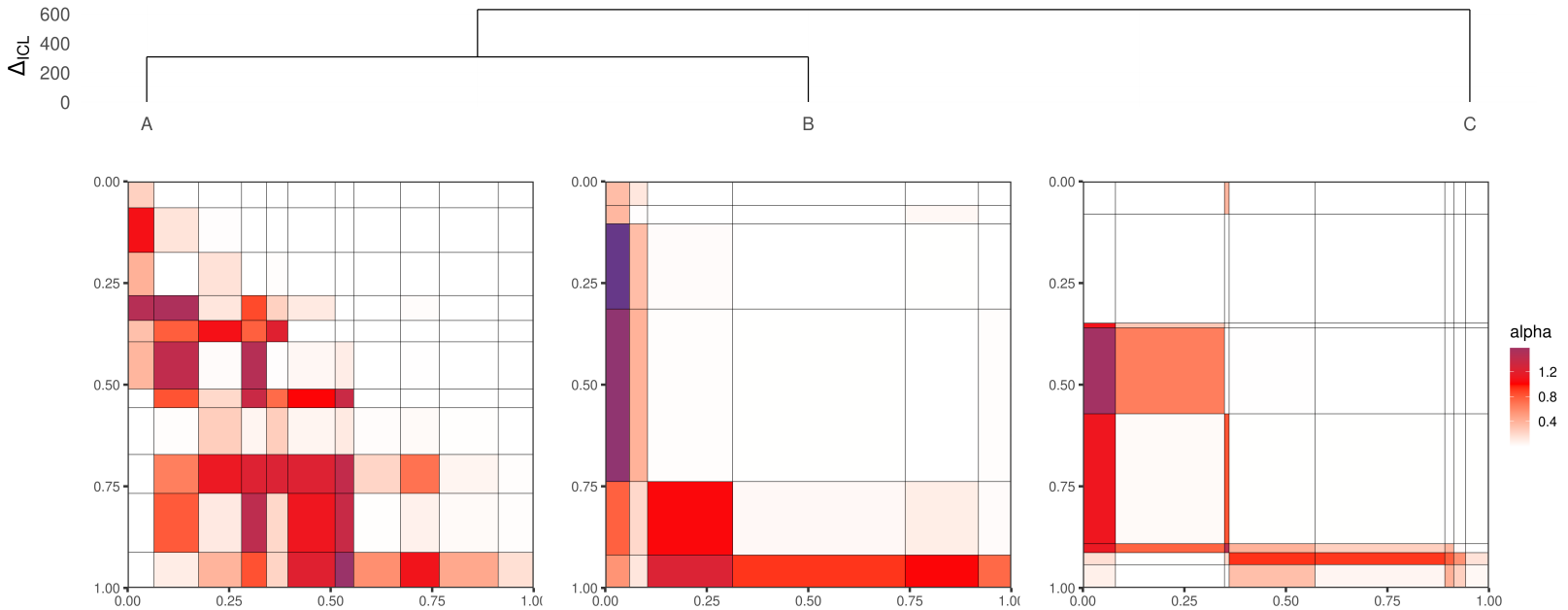


BIC-L: sepSBM: -2080, iid-colSBM: -1966 (left), π-colSBM: -1982 (right)

- Reject separated SBMs
- iid-colSBM : preferred model. Make 5 blocks
- π-colSBM: block proportion quite similar. Make no use of its flexibility

Partition of food webs (δcolSBM)

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



$$|\mathcal{M}_A| = 8, Q^A = 11 \quad |\mathcal{M}_B| = 28, Q^B = 6 \quad |\mathcal{M}_C| = 31, Q^C = 8$$

Diffusion

- arXiv available soon
- `colSBM` available on github
  `https://chabert-liddell.github.io/colSBM/`
  - Simulation and inference of collection of simple networks (directed and undirected)
  - Handle missing data
  - Prediction on missing dyads, missing links and spurious links

Conclusion

Conclusion

3 original contributions

Multilevel Modeling the dependence between levels

Robustness Considering model encompassing topology of the networks

Collection Joint modeling to detect common structures and clusterize the networks by their structure

Prediction of missing interactions

- For networks with incomplete sampling effort
- Simulate missing information from observed networks
  - Useful to assess effectiveness of procedures and pertinence of joint modeling
  - Can be used to quantify the transmission of information between levels/networks

Perspectives

- Modeling of networks
  - Extend colSBM to bipartite and multipartite networks
  - Deal with covariates on nodes, edges and *networks*
- Summary statistics
  - Apply the method used for the robustness to other common statistics: modularity, nestedness, reciprocity...
  - Compare network structures under different models through common statistics
- Improving estimation and/or rescaling of SBM parameters
  - For network issued from incomplete sampling effort
  - For the comparison of observed networks

Thank you for your attention!

Bibliography

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.

Brailly, J. (2016). Dynamics of networks in trade fairsa multilevel relational approach to the cooperation among competitors. *Journal of Economic Geography*, 16(6):1279–1301.

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., 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):e2709.

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.

Côme, E. and Latouche, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, 15(6):564–589.

References ii

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

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473.

Lazega, E., Jourda, M.-T., Mounier, L., and Stofer, R. (2007). Des poissons et des mares: l'analyse de réseaux multi-niveaux. *Revue française de sociologie*, 48(1):93–131.

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.

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.