Stochastic block models for networks

Applications in ecology

Sophie Donnet for the Econet group, INRAE, MIA Paris-Saclay

April 2024

My collaborators

On the R packages

J. Chiquet

(INRAE)

sbm

P. Barbillon

(AgroParisTech)

sbm

J.B. Léger

(Univ. Tech. Compiègne)

blockmodels

Saint-Clair Chabert-Liddell

(INRAE)

colSBM

Other collaborators

T. Vanrenterghem (INRAE), S. Robin (Sorbonne U.), E. Lazega (Sciences Po), F. Massol (CNRS), S. Kefi (CNRS) + ANR Econet + ANR Pastodiv + GDR Resodiv

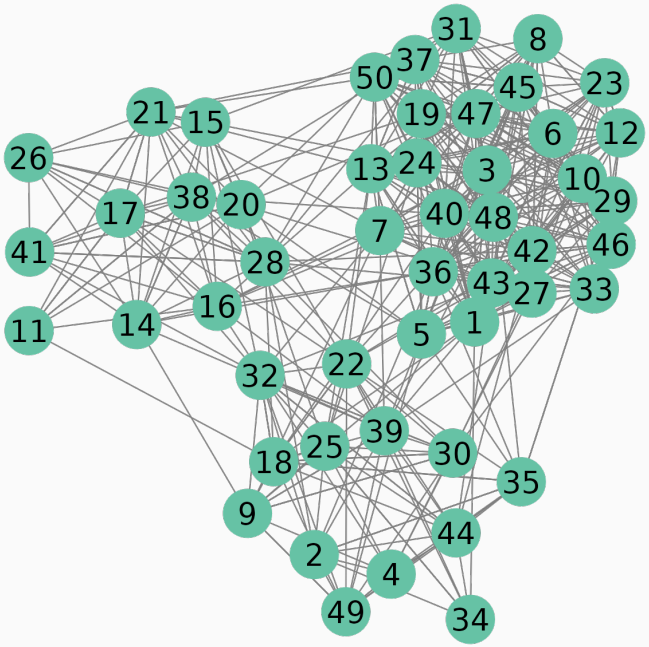
1. Introduction

2. Probabilistic model

3. Inference

Networks

Convenient tools to encode / represent interactions between entities



A network consists in:

- nodes/vertices which represent individuals / species / entities which may interact or not,
- links/edges/connections which stand for an interaction between a pair of nodes / dyads.

Social networks

- Friendship between individuals, 

- Linkedin 
- Twitter 
- Co-publication between researchers
- Advices between lawyers: oriented relation
- Enron email dataset
- Exchanges of seeds between farmers

Networks in ecology

- Ecosystems involve many species
- Interactions between species determine the functioning and evolution of ecosystems
- Several types of interactions

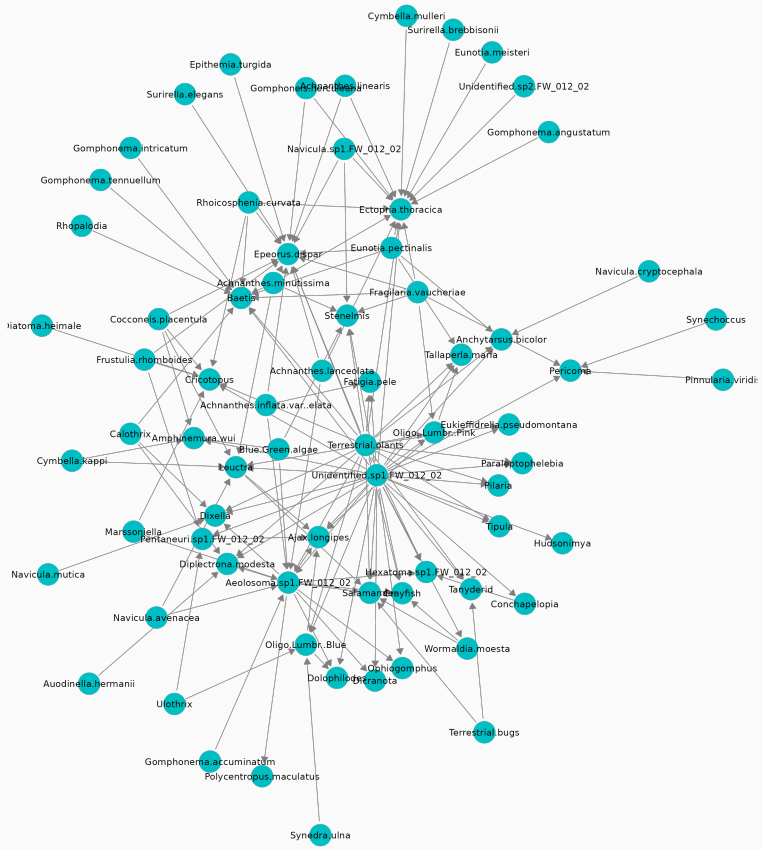
Predation

Parasitism

Pollination

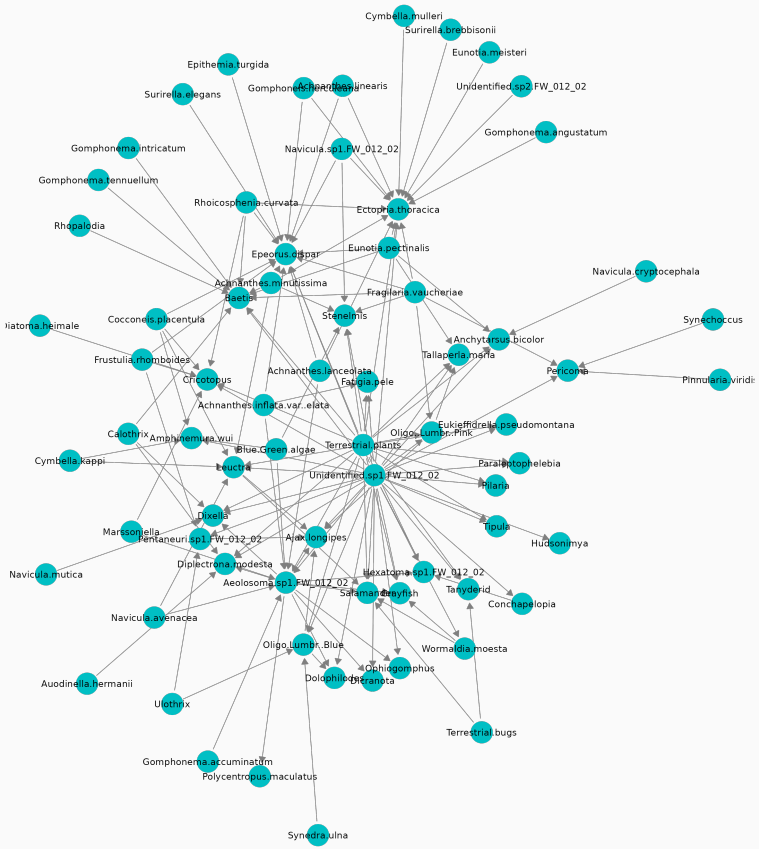
Predation networks: foodwebs

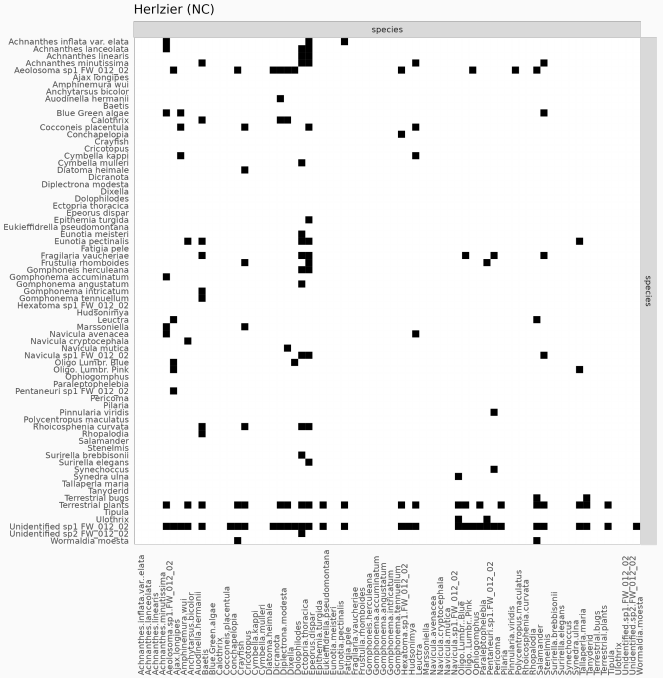
[Thompson and Townsend, 2003] Pine-forest stream foodweb issued from North-Caroline (71 species, 148 interactions)



Foodwebs: adjacency matrix

- $Y = (Y_{ij})_{1 \leq i,j, \leq n} = n \times n$ matrix
- $Y_{ij} = 1$ if i is eaten by j, 0 otherwise

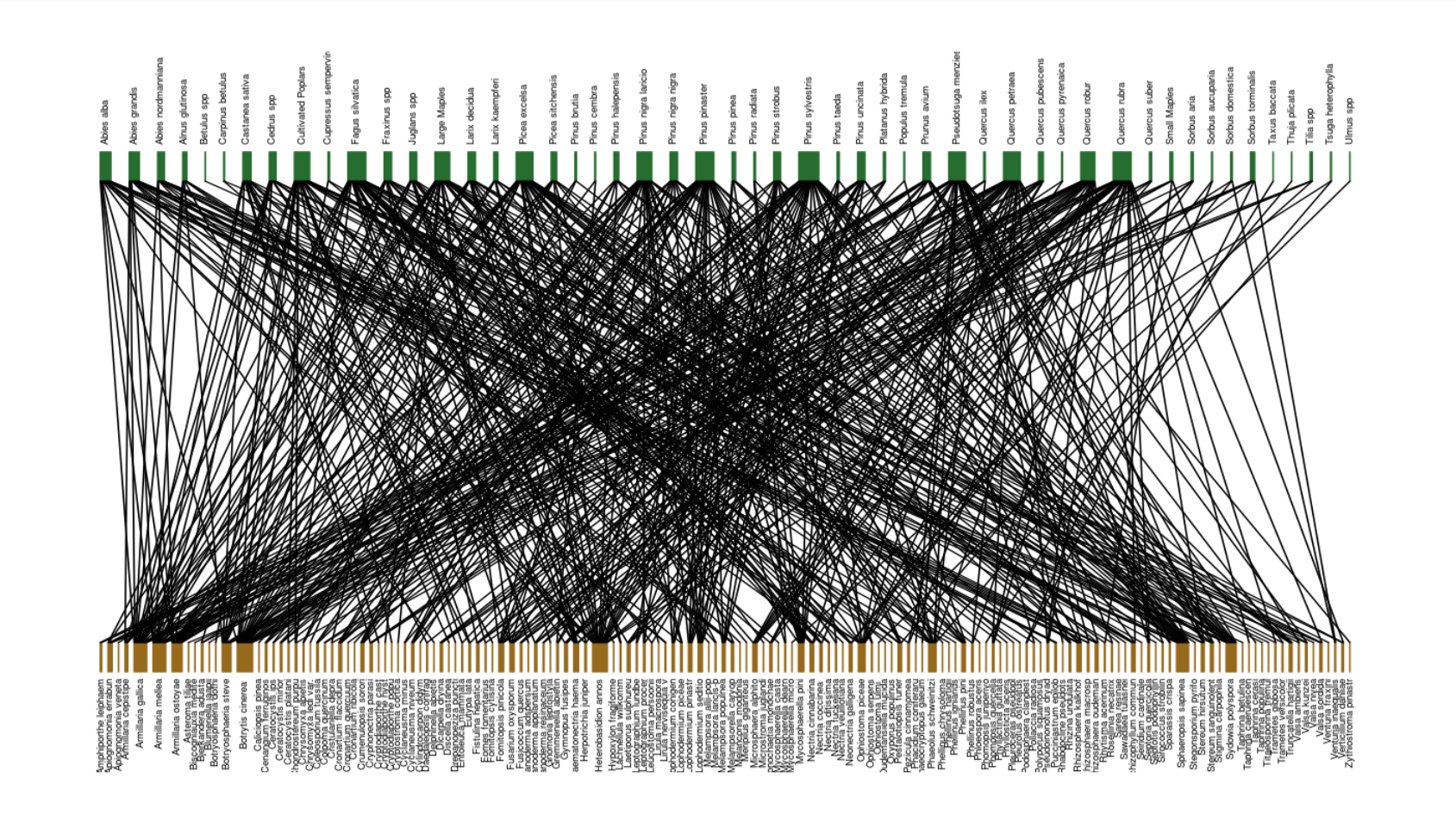




Directed binary relation : Y non symetric and 0/1.

Parasitism : tree-fungus network

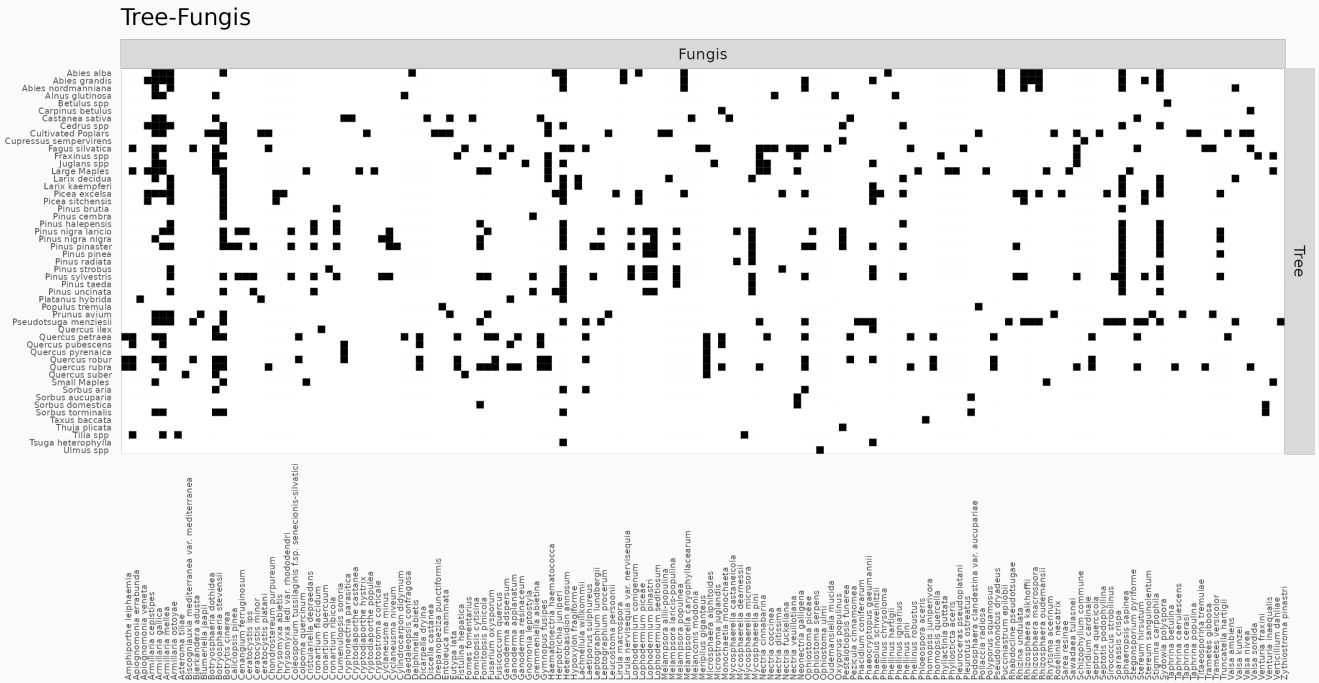
[Vacher et al., 2008] Parasitism relation between $n = 51$ tree species and $p = 154$ fungus species



Nodes of two types: bipartite network

Parasitism : tree-fungus incidence matrix

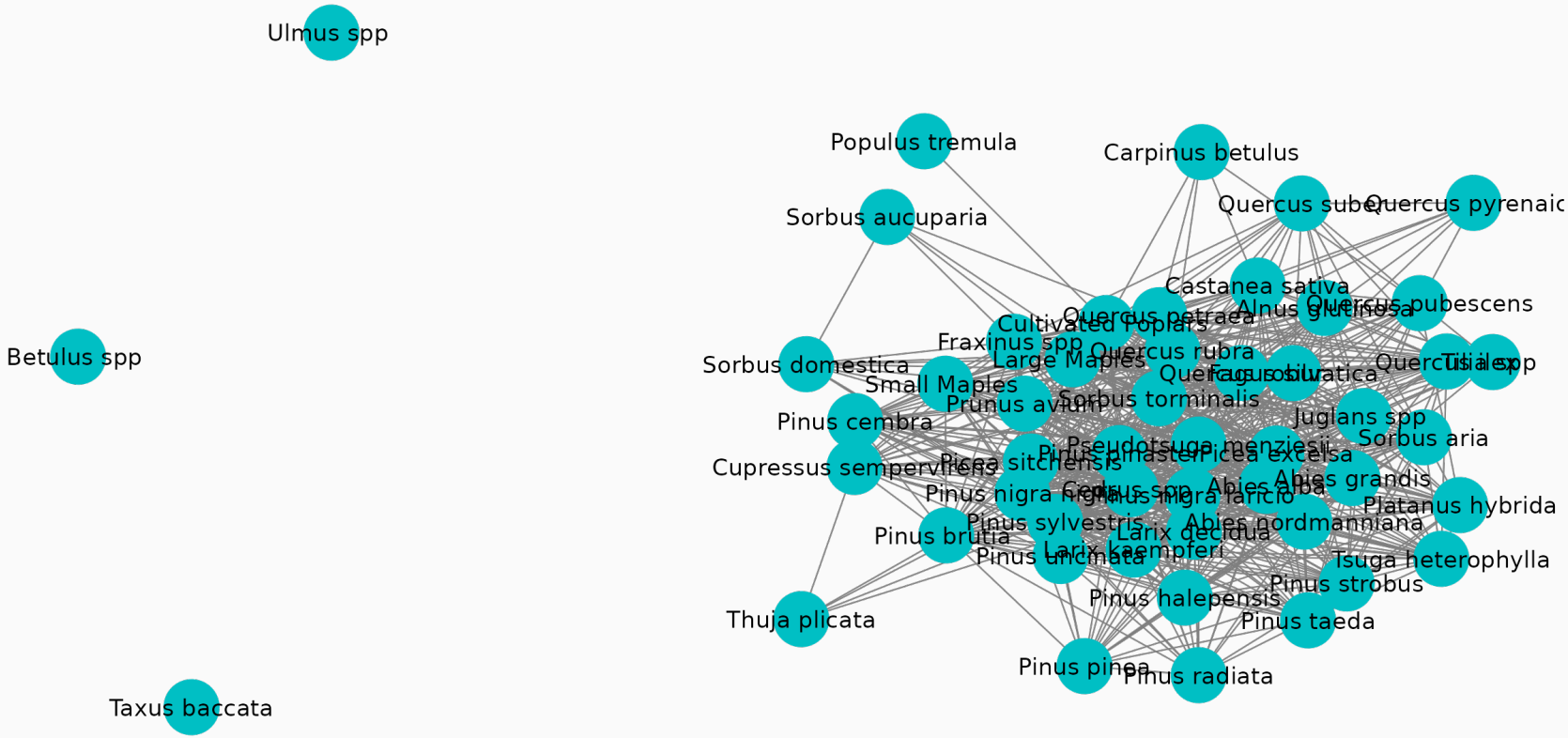
- $Y = (Y_{ij})_{1 \leq i,j, \leq n} = n \times p$ matrix
- $Y_{ij} = 1$ if tree i is parasited by fungus j, 0 otherwise



Binary bipartite network: Y non square and 0/1.

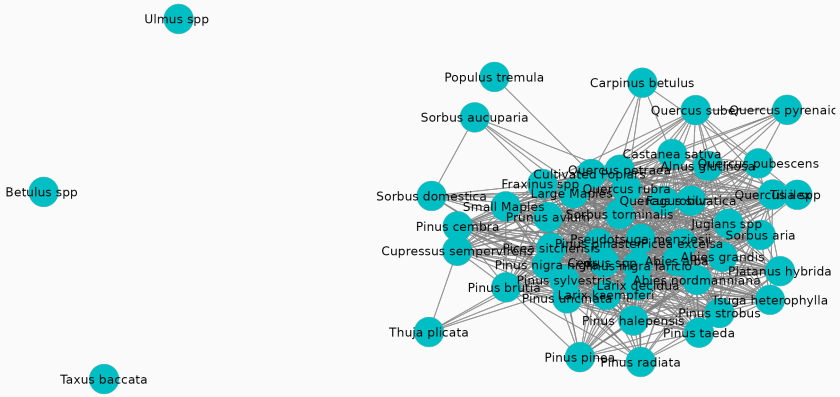
Parasitism : tree-tree network

[Vacher et al., 2008] Number of shared fungus between any pair of the $n = 51$ tree species



Parasitism : weighted adjacency matrix

Y_{ij}: number of shared fungal parasites (fungus hosted by both species)



Weighted non-oriented network: Y symetric and $\in \mathbb{N}$.

Additional information: covariates on pair of trees

For each pair of tree species, 3 distances were also measured:

- taxonomic distance (x^1)
- geographic distance (x^2)
- genetic distance (x^3)

Ecological questions i

→ Ecological aim: caracterize / understand / compare ecosystem organizations.

Foodwebs

- How is organized the network? Can I gather species with similar behavior (trophic levels)?
- Do two given species play the same role in the network?

Fungus-tree networks

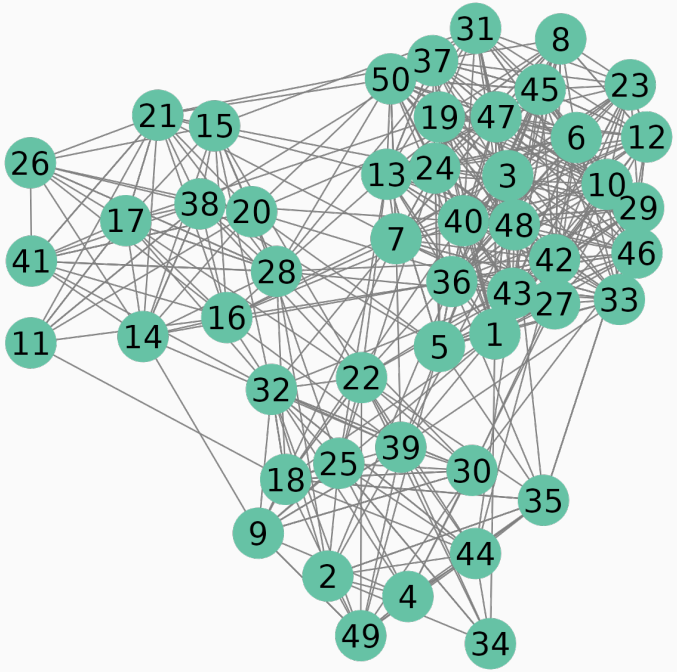
- Can we find groups of trees and fungi that are preferentially associated?

Ecological questions ii

Parasite networks between trees

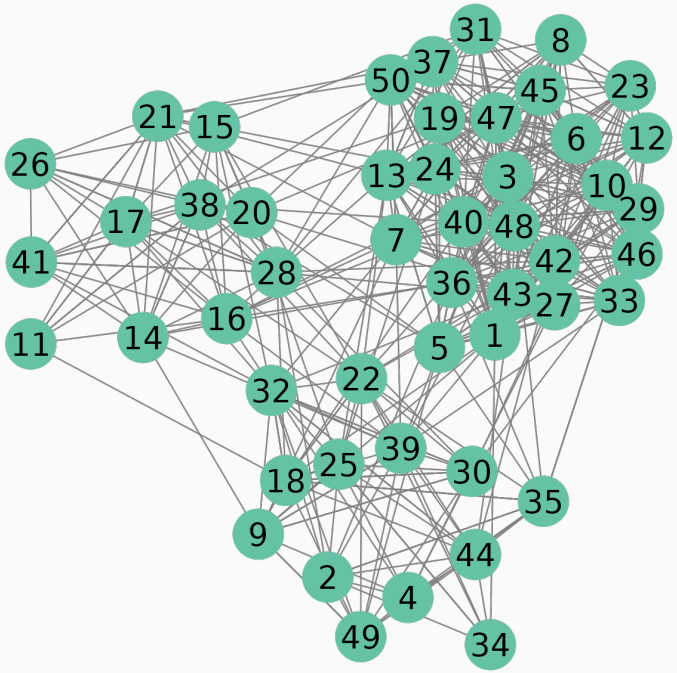
- Do any of the three distances (genetic, geographic or taxonomic) contributes to shape the number of shared parasites?
- Are the covariates sufficient to explain the interactions?

Statistical inference: available data



- the network provided as:
  - an adjacency matrix (for simple network) or an incidence matrix (for bipartite network),
  - a list of pair of nodes / dyads which are linked.
- some additional covariates on nodes, dyads which can account for sampling effort.

Statistical inference: goal



- Unraveling / describing / modeling the network topology.
- Discovering particular structure of interactions between some subsets of nodes.
- Understanding node heterogeneity.
- Not inferring the network !

Probabilistic approach

- Context: our matrix Y is the realization of a stochastic process.
- Aim: Propose a stochastic process is able to mimic heterogeneity in the connections.
- Advantage: benefit from the statistical tools (tests, model selection, etc...)

A first random graph model for network

Erdős-Rényi (1959) Model for n nodes

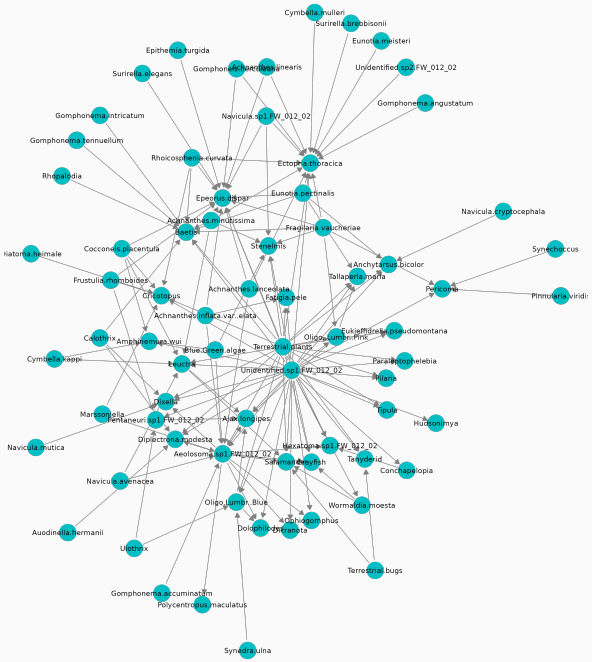
$$\forall 1 \leq i, j \leq n, \quad Y_{ij} \overset{i.i.d.}{\sim} \mathcal{B}\text{ern}(p),$$

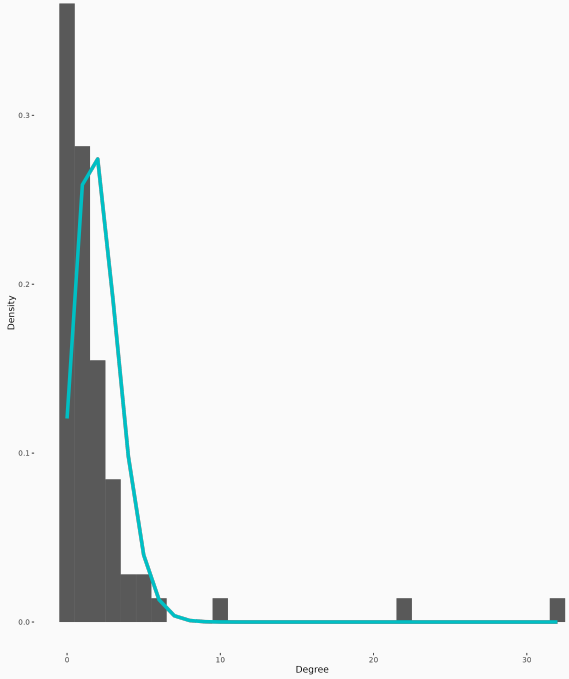
where $p \in [0, 1]$ is the probability for a link to exist.

Consequence

$$\deg(i) \sim_{i.i.d} \mathcal{B}\text{in}(n, p)$$

Confrontation to a real network





Not enough variability in the degree

Limitations of an ER graph to describe real networks

- Homogeneity of the connections
- Degree distribution too concentrated, no high degree nodes,
- All nodes are equivalent (no nestedness...),
- No modularity, no hubs

Stochastic Block Model

[Nowicki and Snijders, 2001] Let (Y_{ij}) be an adjacency matrix

Latent variables

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

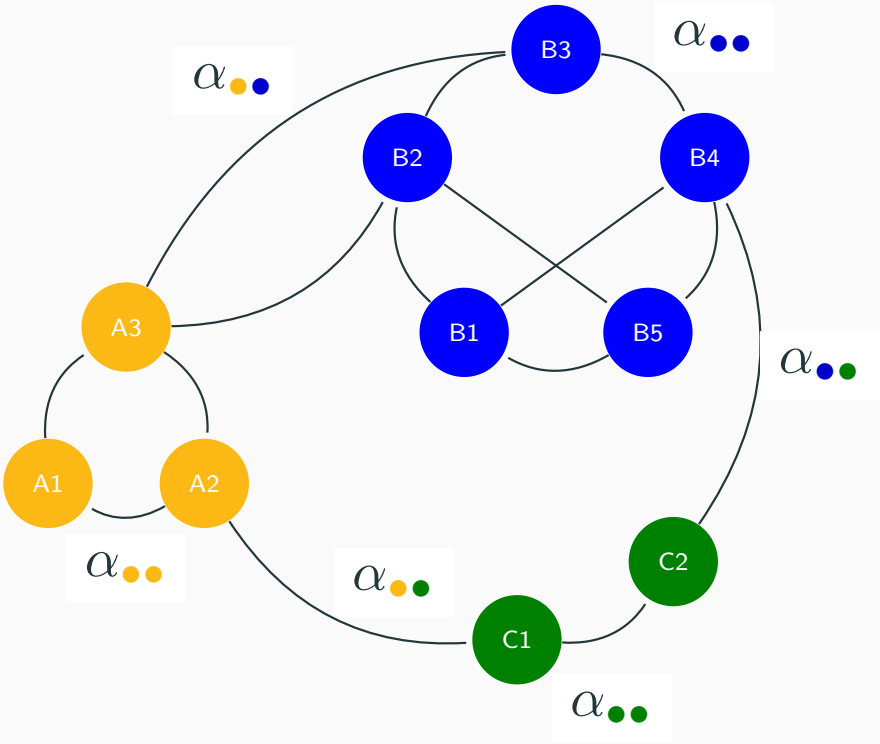
$$\mathbb{P}(Z_i = k) = \pi_k$$

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

(Y_{ij}) independant and

$$Y_{ij}|Z_i, Z_j \sim \mathcal{B}ern(\alpha_{Z_i,Z_j}) \quad \Leftrightarrow \quad P(Y_{ij} = 1|Z_i = k, Z_j = \ell) = \alpha_{k\ell}$$

Stochastic Block Model : illustration



Parameters

Let n nodes divided into 3 clusters

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

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \pi), \quad \forall \bullet \in \mathcal{K},$$
$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{B}(\alpha_{\bullet\bullet})$$

SBM : A great generative model

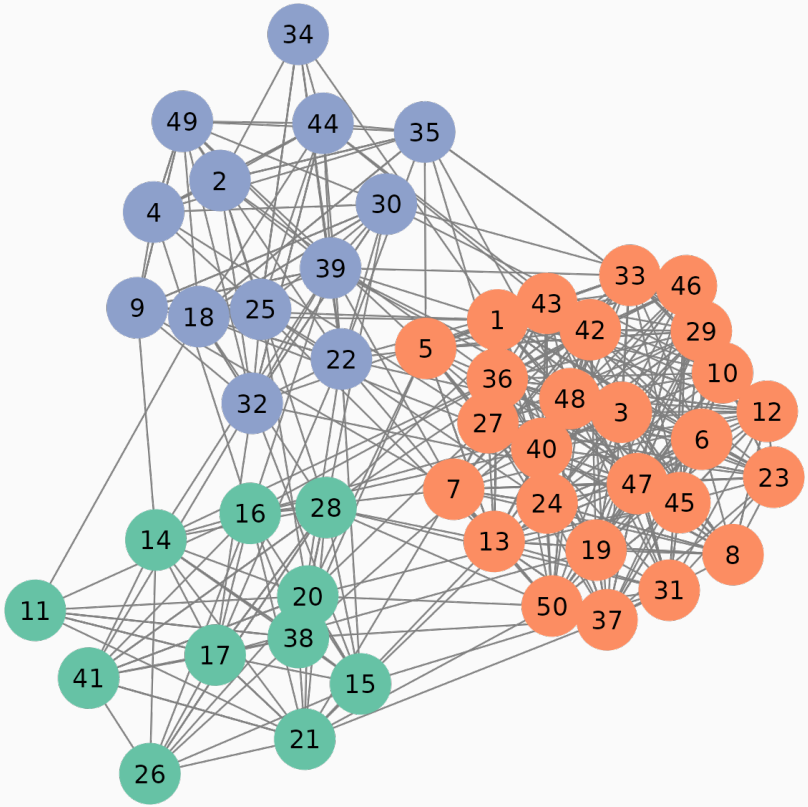
- Generative model : easy to simulate
- No a priori on the type of structure
- Combination of modularity, nestedness, etc...

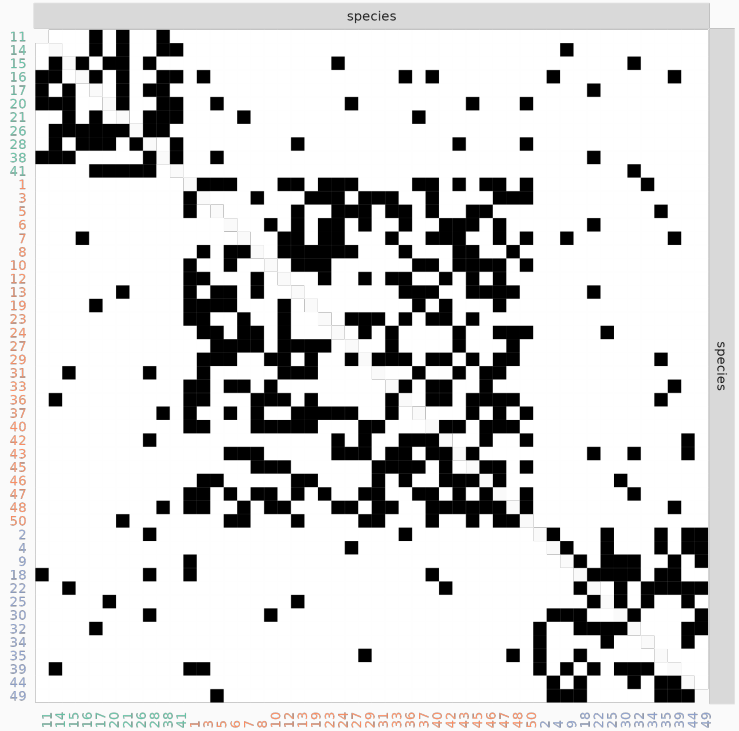
References

- Other ways to model heterogeneity in networks [Matias, Catherine and Robin, Stéphane, 2014]
- Review paper on SBM [Lee and Wilkinson, 2019]

Modelling communities

$$\boldsymbol{\alpha} = \begin{pmatrix} \underline{0.45} & 0.05 & 0.05 \\ 0.05 & \underline{0.45} & 0.05 \\ 0.05 & 0.05 & \underline{0.45} \end{pmatrix} \qquad \pi = (0.25, 0.5, 0.25)$$

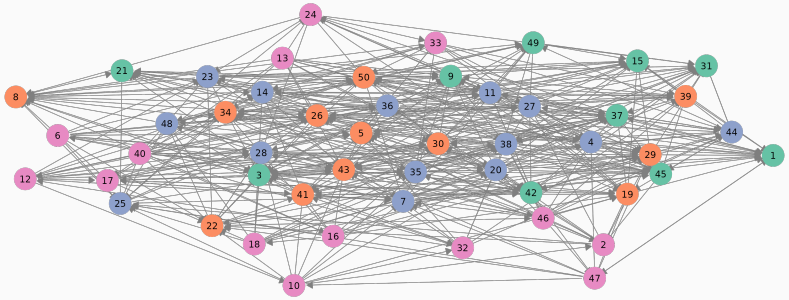


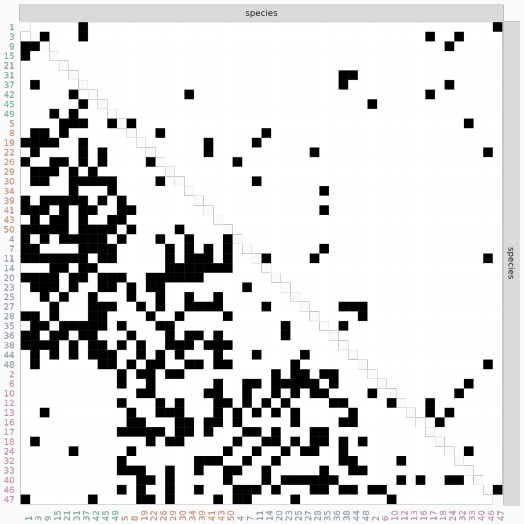


Modelling foodwebs

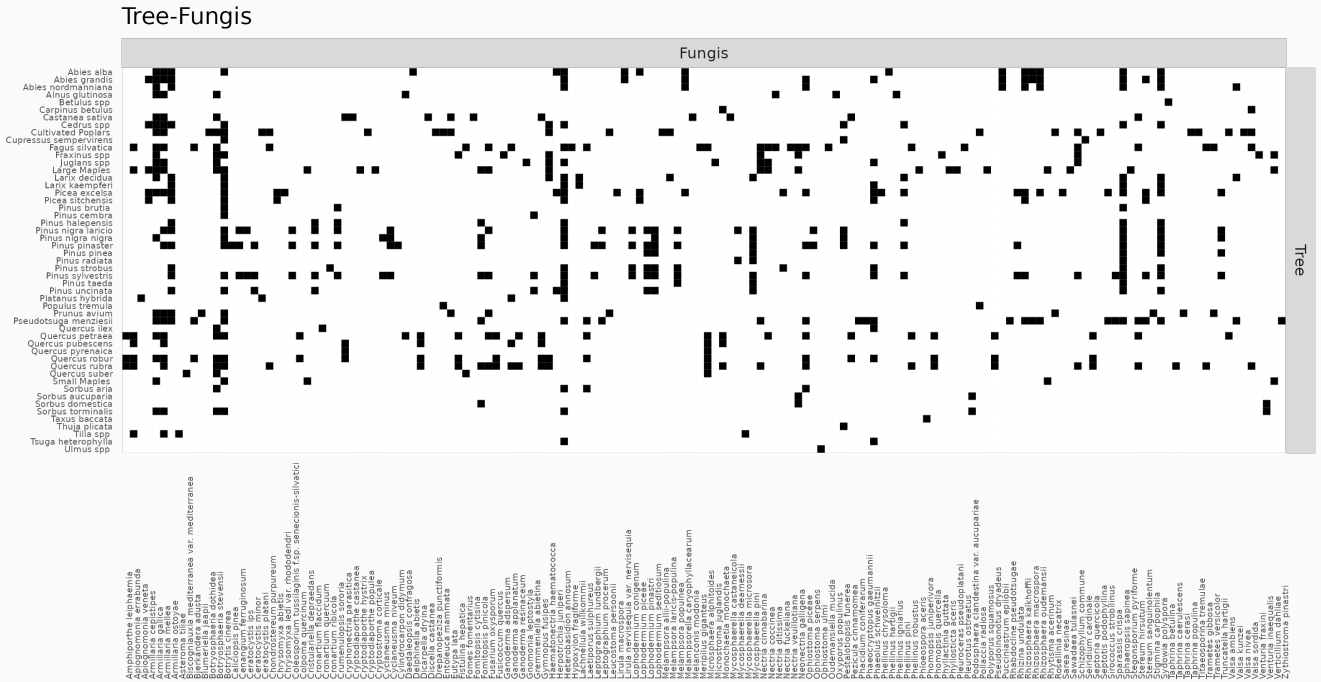
$$\boldsymbol{\alpha} = \begin{pmatrix} 0.10 & 0.02 & 0.02 & 0.02 \\ \underline{0.50} & 0.10 & 0.02 & 0.02 \\ \underline{0.50} & \underline{0.40} & 0.10 & 0.02 \\ 0.02 & \underline{0.40} & \underline{0.40} & 0.10 \end{pmatrix}$$

$$\pi = (0.2, .25, 0.30, 0.25)$$





Probabilistic model for binary bipartite networks



Requires adaptation to bipartite networks: blocks for rows and cols

Probabilistic model for binary bipartite networks

Let Y_{ij} be a bi-partite network. Individuals in row and cols are not the same.

Latent variables : bi-clustering

- Nodes $i = 1, \ldots, n$ partitionned into K clusters, nodes $j = 1, \ldots, p$ partitionned into L clusters
- $$Z_i = k \quad \text{if node } i \text{ belongs to cluster (block) } k$$
  $$W_j = \ell \quad \text{if node } j \text{ belongs to cluster (block) } \ell$$
- $(Z_i)_{i=1,\ldots,n}, (W_j)_{j=1,\ldots,p}$ independent variables

$$\mathbb{P}(Z_i = k) = \pi_k, \quad \mathbb{P}(W_j = \ell) = \rho_\ell$$

Probabilistic model for binary bipartite networks

Conditionally to $(W_i)_{i=1,\dots,n}, (W_j)_{j=1,\dots,p}$...

(Y_{ij}) independent and

$$Y_{ij}|Z_i, W_j \sim \mathcal{B}ern(\alpha_{Z_i,W_j}) \quad \Leftrightarrow \quad \mathbb{P}(Y_{ij} = 1|Z_i = k, W_j = \ell) = \alpha_{k\ell}$$

Also called Latent Block Models [Govaert and Nadif, 2008]

Exercice

Go to the tutorial and simulate networks with your own structures.

Valued-edge networks

Values-edges networks

Information on edges can be something different from presence/absence.
It can be:

1. a count of the number of observed interactions,
2. a quantity interpreted as the interaction strength,

Natural extensions of SBM and LBM

1. Poisson distribution: $Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{P}(\lambda_{\bullet\bullet})$,
2. Gaussian distribution: $Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{N}(\mu_{\bullet\bullet}, \sigma^2)$, [Mariadassou et al., 2010]
3. More generally,

$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{F}(\theta_{\bullet\bullet})$$

Multiplex networks

Several kind of interactions between nodes For instance :

- Love and friendship
- Working relations and friendship
- In ecology : mutualistic and competition

Block model for multiplex networks

$Y_{ij} \in \{0,1\}^Q = (Y_{ij}^a, Y_{ij}^b), \forall w \in \{0,1\}^2$

$$\mathbb{P}((Y_{ij}^a, Y_{ij}^b) = w | Z_i = k, Z_j = \ell) = \alpha_{k\ell}^w$$

[Kéfi et al., 2016], [Barbillon et al., 2017]

In R package: sbm when two relations are at stake.

Remark: a particular case of multiplex network is dynamic network, [Matias and Miele, 2017].

Taking into account covariates

Sometimes covariates are available. They may be on:

- nodes,
- edges,
- both.

1. They can be used a posteriori to explain blocks inferred by SBM.
2. Extension of the SBM which takes into account covariates. Blocks are structure of interaction which is not explained by covariates !

If covariates are sampling conditions, case 2 be may more interesting.

SBM with covariates

- As before : (Y_{ij}) be an adjacency matrix
- Let $x^{ij} \in \mathbb{R}^p$ denote covariates describing the pair (i, j)

Latent variables : as before

- The nodes $i = 1, \ldots, n$ are partitioned into K clusters
- Z_i independent variables

$$\mathbb{P}(Z_i = k) = \pi_k$$

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

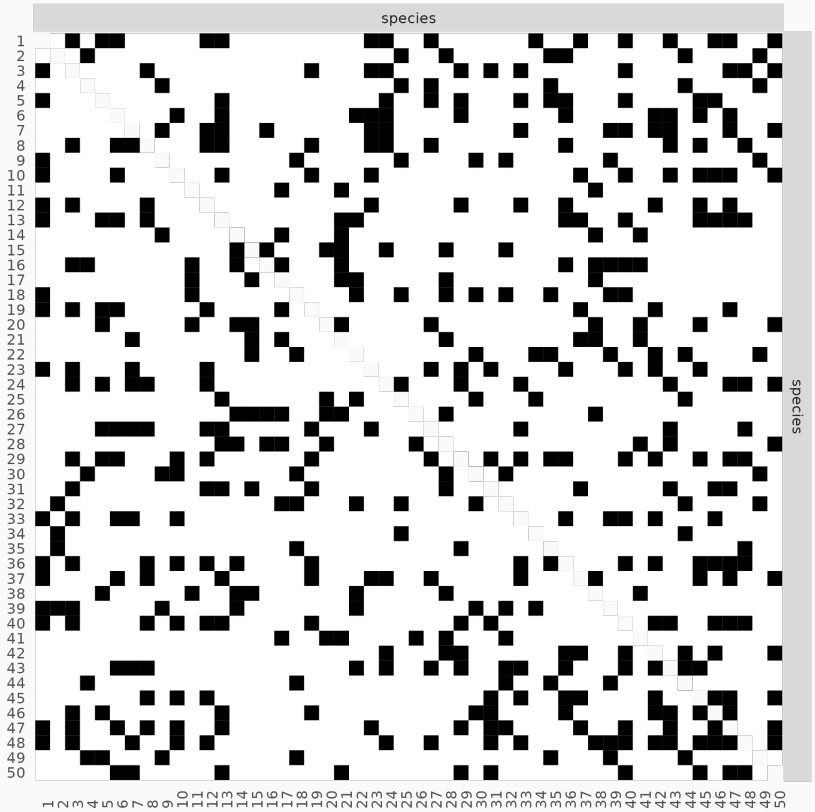
(Y_{ij}) independent and

$$\begin{aligned} Y_{ij}|Z_i, Z_j &\sim \mathcal{B}ern(\text{logit}(\alpha_{Z_i,Z_j} + \theta \cdot x_{ij})) \quad \text{if binary data} \\ Y_{ij}|Z_i, Z_j &\sim \mathcal{P}(\exp(\alpha_{Z_i,Z_j} + \theta \cdot x_{ij})) \quad \text{if counting data} \end{aligned}$$

If $K = 1$: all the connection heterogeneity is explained by the covariates.

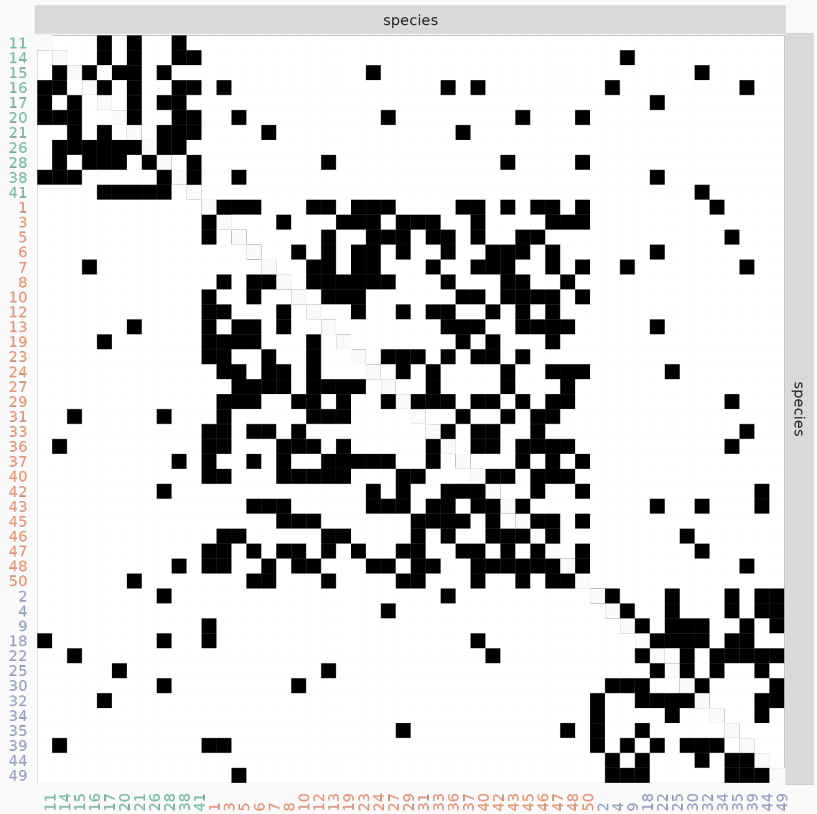
Aim

Going from...



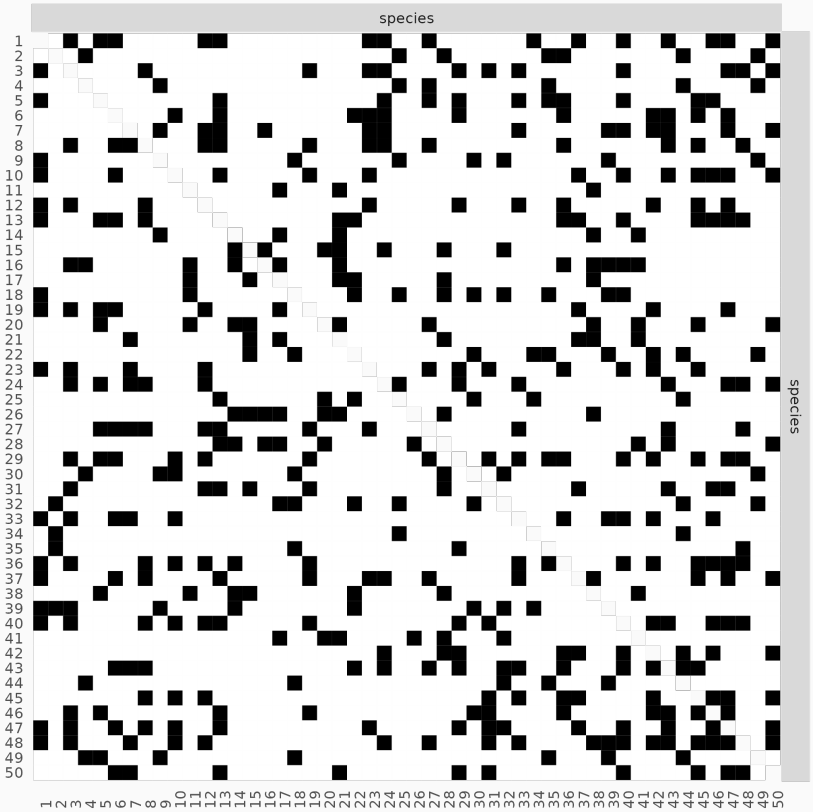
Aim

... to


species
species
11 14 15 16 17 20 21 26 28 38 41 1 3 5 6 7 8 10 12 13 19 23 24 27 29 31 33 36 37 40 42 43 45 46 47 48 50 2 4 9 18 22 25 30 32 34 35 39 44 49

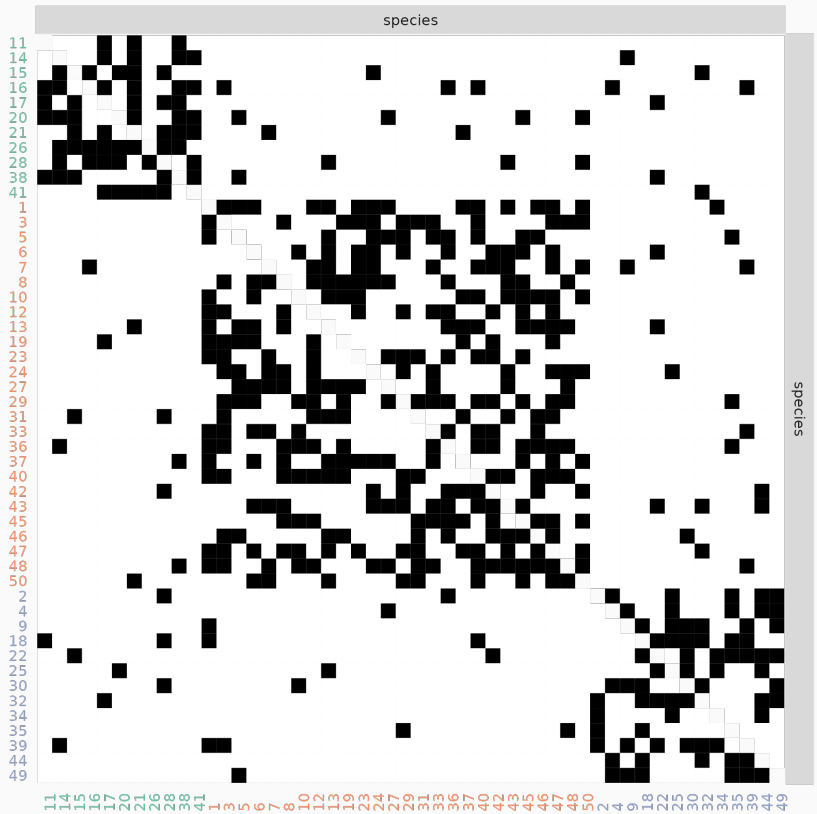
Aim

Going from...



Aim

... to


species
species
11 14 15 16 17 20 21 26 28 38 41 1 3 5 6 7 8 10 12 13 19 23 24 27 29 31 33 36 37 40 42 43 45 46 47 48 50 2 4 9 18 22 25 30 32 34 35 39 44 49

Statistical challlenges

- Selection of the number of clusters
  - K for SBM , K and L for bipartite SBM
- Estimation of the parameters $(\pi, \boldsymbol{\theta})$ for a given number of clusters
- Clustering $\hat{\mathbf{Z}}$

Presented in details for binary SBM.

Likelihood for Bernoulli SBM

Complete likelihood (Y) et (Z)

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

Marginal likelihood (Y)

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

$$\log \ell(\mathbf{Y}; \theta) = \log \sum_{\mathbf{Z} \in \mathcal{Z}} \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) \,. \tag{2}$$

Marginal likelihood : remark

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

Remark

$\mathcal{Z} = \{1, \ldots, K\}^n \Rightarrow$ when K 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

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

- **Step M**:

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

Limitations of standard EM i

Step E requires the computation of $\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)]$

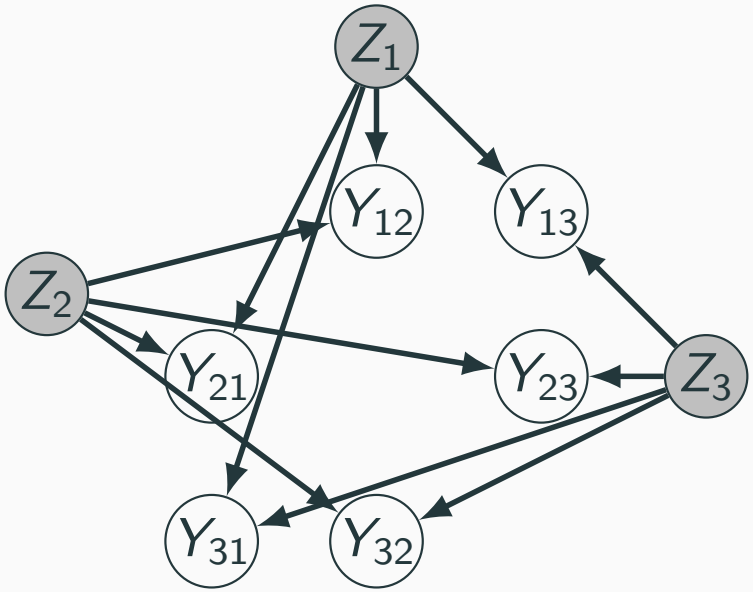
-

$$\begin{aligned}
\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) &= \log\left[\prod_{i \neq j} \alpha_{Z_i, Z_j}^{Y_{ij}} (1 - \alpha_{Z_i, Z_j})^{1 - Y_{ij}}\right] + \log\left[\prod_i \pi_{Z_i}\right] \\
&= \sum_{i \neq j} \sum_{k,\ell=1}^{K} Z_{ik} Z_{j\ell} \left[Y_{ij} \log \alpha_{k\ell} + (1 - Y_{ij}) \log(1 - \alpha_{k\ell})\right] \\
&\quad + \sum_{i,k=1}^{n,K} Z_{ik} \log \pi_k
\end{aligned}$$

with $Z_{ik} = \mathbf{1}_{Z_i = k}$

Limitations of standard EM ii

- However, once conditioned by par $\mathbf{Y}$, the $\mathbf{Z}$ are not independent anymore



$$p(\mathbf{Z}|\mathbf{Y}, \theta^{(t-1)}) \neq \prod_{i=1}^{n} p(Z_i|\mathbf{Y}, \theta^{(t-1)})$$

Variational EM : maximization of a lower bound

Idea : replace the complicated distribution $p(\mathbf{Z}|\mathbf{Y};\theta)$ 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:

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

⏩ skip mathematicals details on VEM

Proof i

By Bayes

$$\begin{aligned} \log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) &= \log p(\mathbf{Z}|\mathbf{Y}; \theta) + \log \ell(\mathbf{Y}; \theta) \\ \log \ell(\mathbf{Y}; \theta) &= \log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) - \log p(\mathbf{Z}|\mathbf{Y}; \theta) \end{aligned}$$

By integration against $\mathcal{R}_{\mathbf{Y},\tau}$:

$$\begin{aligned} \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log \ell(\mathbf{Y}; \theta)] &= \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)] - \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log p(\mathbf{Z}|\mathbf{Y}; \theta)] \\ \log \ell(\mathbf{Y}; \theta) &= \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)] - \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log p(\mathbf{Z}|\mathbf{Y}; \theta)] \end{aligned}$$

Proof ii

As a consequence:

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

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)

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

Variational EM

Algorithm

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

- **Step 1 (VE)** Maximization w.r.t. τ

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

Note that

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

Variational EM

Algorithm

- **Step 2 (M)** Maximization w.r.t. θ

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

Details of the VE-step for binary SBM i

$$\tau^{(t)} = \arg\min_{\tau} \mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y}; \theta^{(t-1)})] = \arg\max_{\tau} \mathcal{I}_{\theta^{(t-1)}}(\mathcal{R}_{\mathbf{Y},\tau}) \,.$$

(we drop out the index $^{(t-1)}$ on θ)

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) \quad = \quad \sum_{\mathbf{Z}} \mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}) \log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) - \sum_{\mathbf{Z}} \mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}) \log \mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}) \,,$$

with

$$\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) \quad = \quad \sum_{i,j=1, i\neq j}^{n} \sum_{k,\ell=1}^{K} Z_{ik} Z_{j\ell} \log p(Y_{ij}|\alpha_{k\ell}) + \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k$$

Details of the VE-step for binary SBM ii

Integration of the $\mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{R}_{\mathbf{Y},\tau}$

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) = \sum_{i,j=1,i\neq j}^{n} \sum_{k,\ell=1}^{K} \tau_{iq}\tau_{j\ell} \log p(Y_{ij}|\alpha_{k\ell}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \log \pi_k$$

Maximization under the constraint: $\forall i = 1 \dots n$, $\sum_{k=1}^{K} \tau_{ik} = 1$.

- Derivatives of

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) + \sum_{i=1}^{n} \lambda_i \left[\sum_{k=1}^{K} \tau_{ik} - 1 \right]$$

with respect to $(\lambda_i)_{i=1\dots n}$ and $(\tau_{ik})_{i=1\dots n, k=1\dots K}$ where λ_i are the Lagrange multipliers,

Details of the VE-step for binary SBM iii

- Leads to collection of equations: for $i = 1 \dots n$ and $k = 1 \dots K$,

$$\sum_{\ell=1}^{K} \sum_{j=1, j \neq i}^{n} \log p(Y_{ij} | \alpha_{k\ell}) \tau_{j\ell} + \log \pi_k - \log \tau_{ik} + 1 + \lambda_i = 0 \,,$$

- Leads to the following fixed point problem:

$$\widehat{\tau}_{ik} = e^{1+\lambda_i} \alpha_k \prod_{j=1, j \neq i}^{n} \prod_{\ell=1}^{K} p(Y_{ij} | \alpha_{k\ell})^{\widehat{\tau}_{j\ell}}, \quad \forall i = 1 \dots n, \forall k = 1 \dots K \,,$$

which has to be solved under the constraints $\forall i = 1 \dots n$, $\sum_{k=1}^{K} \tau_{ik} = 1$. This optimization problem is solved using a standard fixed point algorithm.

Details of the M-step for binary SBM i

$$\theta^{(t)} = \arg\max_{\theta} \mathcal{I}_{\theta^{(t)}}\left(\mathcal{R}_{\mathbf{Y},\tau^{(t)}}\right)$$

under the constraints: $\sum_{k=1}^{k} \pi_k = 1$.

Maximization with respect to π is quite direct:

$$\widehat{\pi}_q = \frac{1}{n}\sum_{i=1}^{n} \widehat{\tau}_{ik}$$

For the Bernoulli SBM:

$$\widehat{\alpha}_{k\ell} = \frac{\sum_{i,j=1,i\neq j}^{n} \widehat{\tau}_{ik}\widehat{\tau}_{j\ell}\, Y_{ij}}{\sum_{i,j=1,i\neq j}^{n} \widehat{\tau}_{ik}\widehat{\tau}_{j\ell}}$$

Details of the M-step for binary SBM ii

If the edge probabilities depend on covariates:

$$\text{logit}(p_{k\ell}) = \alpha_{k\ell} + \beta \cdot x_{ij} \,,$$

then the optimization of $(\alpha_{k\ell})$ and (β) at step M of the VEM is not explicit anymore and one should resort to optimization algorithms such as Newton-Raphson algorithm.

In practice

- Non stochastic algorithm
- Really fast
- Strongly depends on the initial values
- Asymptotic theoretical properties of the variational estimators

Output

For a given number of blocks K:

- $\widehat{\theta}_K$
- $\widehat{\tau}_{ik} \approx P(Z_i = k|\mathbf{Y}, \widehat{\theta}_K)$
- $$\hat{Z}_i = \underset{k \in \{1,\dots,K\}}{\arg\max}\ \hat{\tau}_{ik}$$
- Hopefully : you can reorder the matrix according to the block affectations and see a structure appearing.

Selecting the number of blocks

- The number of parameters increases with K

$$nb(K) = \underbrace{K-1}_{\text{block. prop}} + \underbrace{\frac{K(K+1)}{2}}_{\alpha_{k\ell}}$$

  (for undirected network)
- The fit of the model to the data is better with large K

$$\log \ell(\mathbf{Y}; \widehat{\theta}_K)$$

- Balance between complexity and fit: penalized likelihood criterion

$$\text{Crit}(K) = \log \ell(\mathbf{Y}; \widehat{\theta}_K) - \text{pen}(\mathcal{M}_K)$$

  - Cf : AIC, BIC, ...
  - Penalty is such that we have asymptotic guaranties to select the "true" model

In our model

- Unable to compute the marginal likelihood but able to compute the complete likelihood
- BIC on the complete likelihood

$$\log \ell_c(\mathbf{Y}, \mathbf{Z}; \hat{\theta}_{\mathbf{K}}) - \mathrm{pen}(\mathcal{M}_{\mathbf{K}})$$

- See after for the expression of the penalty term
- But : Z unknown

ICL criterion

- Integrated Classification Likelihood (ICL) [Biernacki et al., 2000]

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

where

$$\hat{Z}_i = \underset{k \in \{1,\dots,K\}}{\arg\max}\ \hat{\tau}_{ik}.$$

- Integrated Complete Likelihood (ICL)

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

Expression of the penalization for SBM

- For directed network

$$pen_{\mathcal{M}} = \frac{1}{2}\left\{(K-1)\log(n) + K^2 \log\left(n^2 - n\right)\right\}$$

- For undirected network

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

Expression of the penalization for bipartite SBM

$$pen_{\mathcal{M}} = -\frac{1}{2}\left\{ \underbrace{(K-1)\log(n) + (L-1)\log(p)}_{\text{Bi-Clust.}} + \underbrace{(KL)\log(np)}_{\text{Connection}} \right\}$$

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.

Algorithm in practice

- Going trough the models and initiate VEM at the same time
- Bounds on K : $\{K_{\text{min}}, \ldots, K_{\text{max}}\}$

Stepwise procedure

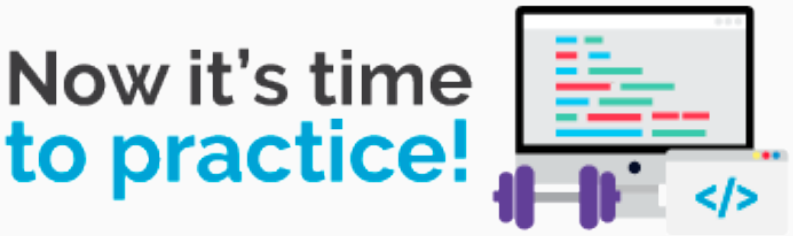
Starting from K

- **Split** : if $K < K_{\text{max}}$
  - Maximize the likelihood (lower bound) of $\mathcal{M}_{K+1}$
  - K initializations of the VEM are proposed : split each cluster into 2 clusters
- **Merge** : If $K > K_{\text{min}}$
  - Maximize the likelihood (lower bound) of model $\mathcal{M}_{K-1}$
  - $\frac{K(K-1)}{2}$ initializations of the VEM are proposed : merging all the possible pairs of clusters

Theoretical properties for SBM

- 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]

Exercice



- Go back to the tutorial and estimate the parameters of your simulated networks.
- Do you obtain the same clusters?
- Decrease the number of nodes. What do you expect?

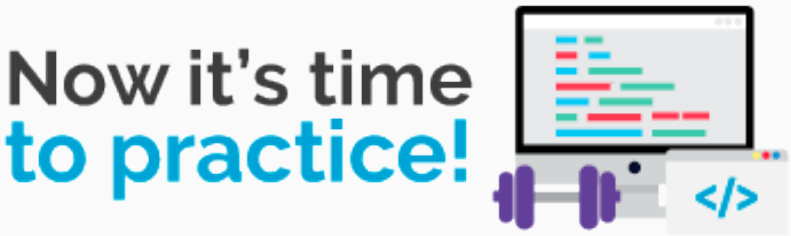
Other extensions

- Time evolving networks Matias
- Multipartite, Multiplexe networks (R-package sbm, Bar-Hen, Barbillon, Donnet)
- Multilevel networks (individuals and organizations) (Chabbert-Liddell)
- Missing data in the network [Tabouy et al., 2019]

Probabilistic model for networks in a nutshell

SBM/LBM

- generative models,
- flexible,
- comprehensive models which can be linked to a lot of classical descriptors.



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

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

Illustration from there

References i

Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2017).

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), 180(1):295–314.

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.

Govaert, G. and Nadif, M. (2008).

Block clustering with bernoulli mixture models: Comparison of different approaches.

Computational. Statistics and Data Analysis, 52(6):3233–3245.

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.

References ii

Lee, C. and Wilkinson, D. J. (2019).

A review of stochastic block models and extensions for graph clustering.

Applied Network Science, 4:122.

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.

Matias, Catherine and Robin, Stéphane (2014).

Modeling heterogeneity in random graphs through latent space models: a selective review*.

ESAIM: Proc., 47:55–74.

Nowicki, K. and Snijders, T. A. B. (2001).

Estimation and prediction for stochastic blockstructures.

Journal of the American Statistical Association, 96(455):1077–1087.

References iii

Tabouy, T., Barbillon, P., and Chiquet, J. (2019).

Variational inference for stochastic block models from sampled data.

Journal of the American Statistical Association, pages 1–23.

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.

Vacher, C., Piou, D., and Desprez-Loustau, M. L. (2008).

Architecture of an antagonistic tree/fungus network: The asymmetric influence of past evolutionary history.

PLoS ONE, 3.