# Introduction to block models for ecological and sociological networks

Sophie Donnet

Janvier 2018. Journée Statistiques / Apprentissage. IHES

# Network data

Social networks can account for

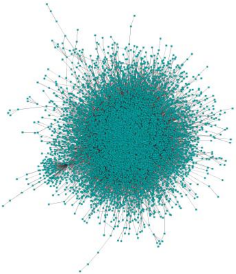
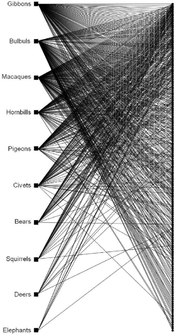
- Advice networks
- Competitors networks
- Friendship
- Copublications,
- Seed exchange...

Ecological networks can account for

- Food web,
- Co-existence networks,
- Host-parasite interactions,
- Plant-pollinator interactions,

# Bipartite or interaction networks

Networks may be or not bipartite : Interactions between nodes belonging to the same or to different functional group(s).

| **Interaction** | **Bipartite** |
|---|---|
|  |  |

# Terminology

A network consists in :

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

A network may be

- directed / oriented (e.g. food web...),
- symmetric / undirected (e.g. coexistence network),
- with or without loops.

These distinctions only make sense for simple networks (not bipartite).

# Available data and goal

## 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.

## Goals

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

# Example in ecology

High number of interaction types between plants and animal species co-exist within the natural environment : pollinisation, protective ants , seed-dispersing birds, herbivory...



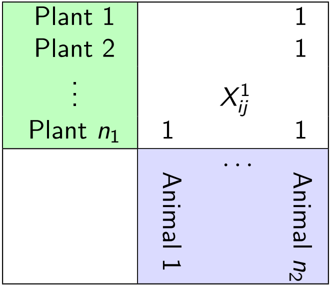



# Example in ecology

- Interactions play a key role in structuring biodiversity.
- Network tools intensively used to understand the structure of these ecological interaction network.

# Ecological network here : an incidence matrix

$$X_{ij} = \begin{cases} 1 & \text{if animal } j \text{ has been observed on plant } i \\ 0 & \text{otherwise} \end{cases}$$



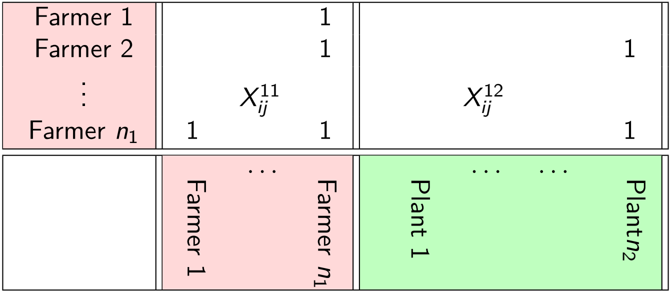

Here : $X_{ij} \in \{0, 1\}$ to avoid sampling issues. But we can have $X_{ij} \in \mathbb{N}$ or $X_{ij} \in \mathbb{R}$

# Example 2 : in ethnobiology (MIRES)

- Relations of seed exchange between farmers : adjcency matrix
- Inventory of the cultivated plants for each farmer of the network

# Example 2 : in ethnobiology

- 2 functional groups :
  - Famers : groupe 1
  - Plants : groupe 2
- Interactions :
  - Farmers / Farmers
  - Farmers / Plants
- 1 non-symmetric adjacency matrix and one incidence matrix



# Goals

## But

Identify sub-groups of the individuals at stake saring the same connexion behavior

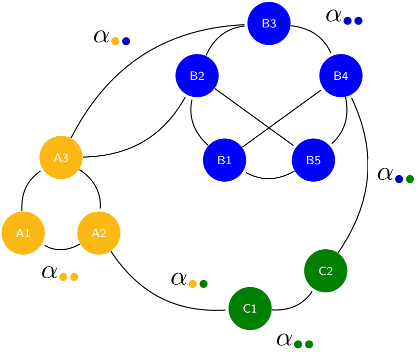
## Existing solutions : Descriptives statistics

- Modularities
  - to detect communities : sub-groups of individuals that are more connected inside their groups than outside.
- Nestedness
- Indegrees, outdegrees
- Distances ....

## Here : probabilistic modeling approach

Stochastic block models (SBM) and latent block models (LBM) relying on the introduction of latent variables

# Stochastic Block Model




## Latent variables

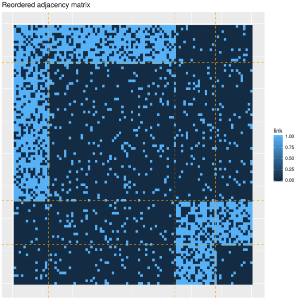
Let $n$ nodes divided into $K$ classes

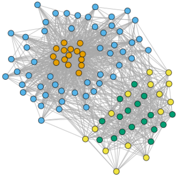
- $\forall i = 1 \dots n$, $Z_i \in \{1, \dots, K\}$ latent variable
- $\pi_k = \mathbb{P}(Z_i = k), \forall i, \forall k$
- $\sum_{k=1}^{K} \pi_k = 1$
- i.i.d. variables

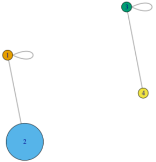
## Emission distribution

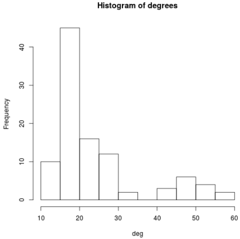
$$X_{ij} \mid \{Z_i = k, Z_j = \ell\} \sim^{\text{ind}} \mathcal{F}(\alpha_{k,\ell})$$

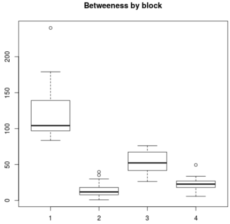
# A very flexible generative model : networks with hubs



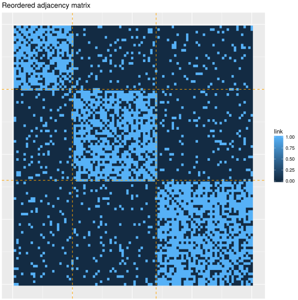



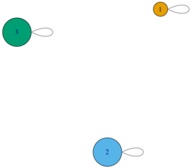


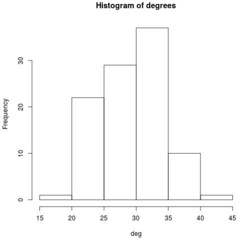



# A very flexible generative model : community network

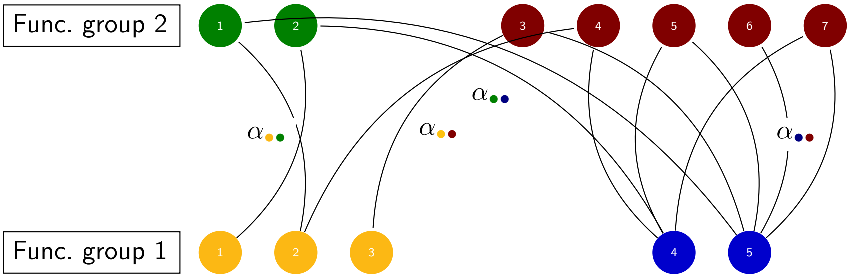


# Latent block models for incidence matrix

Two functional groups of sizes $n_1$ and $n_2$




## Latent variables

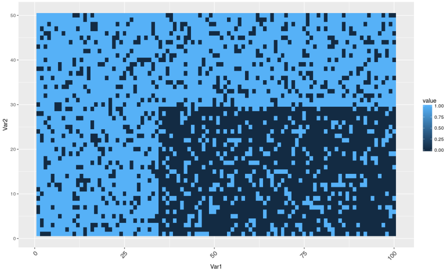
Let the $n_1$ nodes divided into $K_1$ clusters and $n_2$ notes divided into $K_2$ clusters

- $q = 1, 2, \forall i = 1 \dots n_q, \; Z_i^q \in \{1, \dots, K_q\}$ latent variables, i.i.d.
- $\mathbb{P}(Z_i^q = k) = \pi_k^q, \forall i, \forall k$
- $\sum_{k=1}^{K} \pi_k^q = 1, \; q = 1, 2$

# Latent block models for incidence matrix

## Emission distribution

$$X_{ij} \mid \{Z_i^1 = k, Z_j^2 = \ell\} \sim^{\text{ind}} \mathcal{F}(\alpha^{12}_{k,\ell})$$



# Dependencies between the entries $X_{ij}$

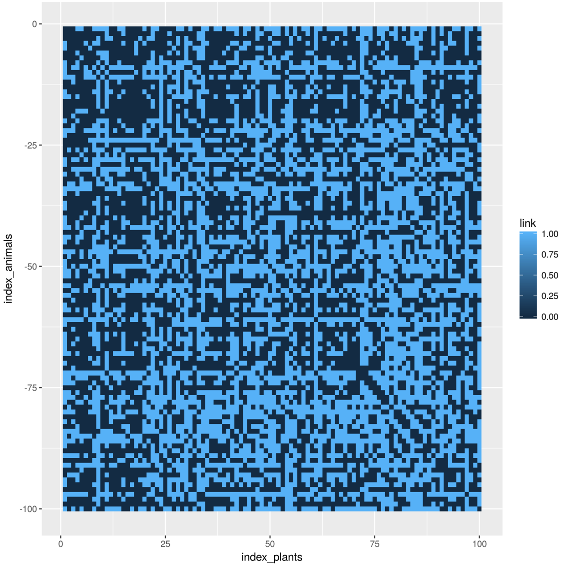
- Independence of the $X_{ij}$ given the latent variables $\boldsymbol{Z}$ (SBM) or $\boldsymbol{Z}^1, \boldsymbol{Z}^2$ (LBM)
- But $\boldsymbol{Z}^q$ have to be integrated $\Rightarrow$ dependence between the $X_{ij}$
- **Consequences** on $\boldsymbol{Z}^q|\boldsymbol{X}$
  - Conditionally on $\boldsymbol{X}$, $(\boldsymbol{Z}^q_i)_{i=1\dots n_q}$ non independent.
  - Complex distribution
  - Different from a classical mixture model

# Covariates

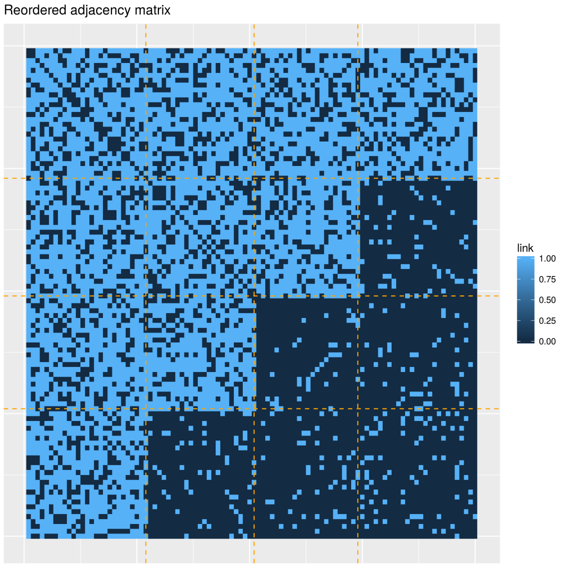
In the SBM model, we can take into account covariates

- Covariables on pairs $(i, j)$
- $X_{i,j} | Z_i = k, Z_j = l \sim \mathcal{F}(\alpha_{kl} + \mathbf{x}_{ij}\beta)$
- If only one group then all the connexions are explained by the covariates.
- If $K > 1$, the covariates explain only a part of the phenomenon.

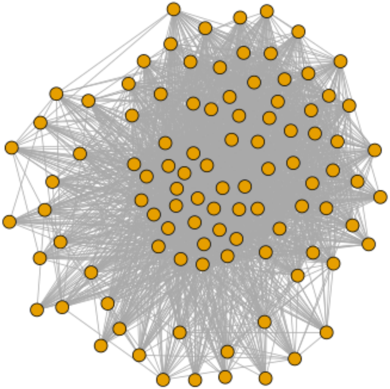
# Statistical Inference



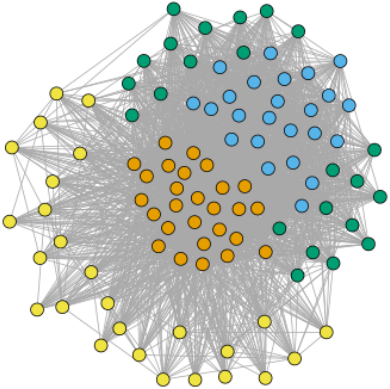
# Statistical Inference



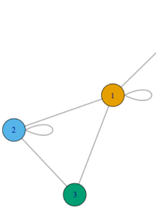
# Statistical Inference



# Statistical Inference



# Statistical Inference




## Goals

- Choose the number(s) of clusters $K$ (or $K_1$ and $K_2$)
- Estimate the parameters $\alpha$ and $\pi$ of the corresponding model
- Estimate the $\boldsymbol{Z}$

# Completed likelihood of $(\boldsymbol{X})$ et $(\boldsymbol{Z})$

$$\ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta) = p(\boldsymbol{X}|\boldsymbol{Z}; \boldsymbol{\alpha}) p(\boldsymbol{Z}; \boldsymbol{\pi})$$

## SBM (*Bernoulli case*)

$$\ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta) = \prod_{i=1}^{n} \prod_{j, j \neq i}^{n} (\alpha_{Z_i, Z_j})^{X_{ij}} (1 - \alpha_{Z_i, Z_j})^{1 - X_{ij}} \times \prod_{i=1}^{n} \pi_{Z_i}.$$

## LBM (*Bernoulli case*)

$$\begin{aligned} \ell_c(\boldsymbol{X}, \boldsymbol{Z}^1, \boldsymbol{Z}^2; \theta) &= \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} (\alpha_{Z_i^1, Z_j^2})^{X_{ij}} (1 - \alpha_{Z_i^1, Z_j^2})^{1 - X_{ij}} \\ &\quad \times \prod_{i=1}^{n_1} \pi^1_{Z_i^1} \cdot \prod_{j=1}^{n_2} \pi^2_{Z_j^2} \end{aligned}$$

# Likelihood of the observed matrix ($\boldsymbol{X}$)

## Likelihood of the observations ($\boldsymbol{X}$)

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

## Remark

$\mathcal{Z} = \otimes_{q=1,2}\{1,\dots,K_q\}^{n_q} \Rightarrow$ when $n$ (or $(n_1, n_2)$) or $K$ or ($K_1$ and $K_2$) increase, impossible to calculate

# Maximization of the likelihood

## Standard EM algorithm

At iteration$(t)$ :

- **Step E** : compute

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

- **Step M** :

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

# Limitations of the standard EM

- Step $E$ requires the computation of $\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\theta}\left[\ell_c(\boldsymbol{X},\boldsymbol{Z};\theta^{(t-1)})\right]$
- However, once conditioned by $\boldsymbol{X}$, the $\boldsymbol{Z}$ are not independant : complexe distribution if $K$ or $K_1, K_2$ are big

# Variational EM : maximisation of a lower bound

- Since $\boldsymbol{Z}|\boldsymbol{X}$ is too complexe, we will replace it by a simpler one
- Let $\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}$ be any probability distribution on $\boldsymbol{Z}$.

## Central identity

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

## Remark

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

# Variational EM

- Maximisation of $\log \ell(\boldsymbol{X};\theta)$ en $\theta$ replaced by the maximisation of the lower bound $\mathcal{I}_\theta(\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}})$ as $\tau$ and $\theta$.
- **Advantage** : we choose $\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}$ such that the maximization and expectation calculus are explicit
  - In our case, mean field approximation : neglect dependences between the $(Z_i^q)$

$$P_{\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}}}(Z_i^q = k) = \tau_{ik}^q$$

# Variational EM

## Algorithm

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

- **Step 1** Maximization in $\tau$

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

- **Step 2** Maximization in $\theta$

$$
\begin{aligned}
\theta^{(t)} &= \arg\max_{\theta} \mathcal{I}_{\theta}(\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}^{(t)}}) \\
&= \arg\max_{\theta} \mathcal{I}_{\theta}(\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}^{(t)}}) \\
&= \arg\max_{\theta} \mathbb{E}_{\mathcal{R}_{\boldsymbol{X},\boldsymbol{\tau}^{(t)}}}[\ell_c(\boldsymbol{X},\boldsymbol{Z};\theta)]
\end{aligned}
$$

# In practice

- Reaches (local) maxima really fast
- Depends strongly on the initial values

# Penalized likelihood criteria

- Selection on the number of clusters $K$ or $K_1, K_2$
- Bayesien Information Criteria : Laplace approximation of the marginal likelihood $m_c(\boldsymbol{X}, \boldsymbol{Z}; \mathcal{M})$ where the parameters $\theta$ have been integrated out with a prior distribution.

## BIC criteria if the latent variables $\boldsymbol{Z}$ are observed

$\mathcal{M} = \mathcal{M}_K$ or $\mathcal{M} = \mathcal{M}_{K_1,K_2}$

$$\log m_c(\boldsymbol{X}, \boldsymbol{Z}; \mathcal{M}) \approx_{n,n_1,n_2 \to \infty} \max_{\theta} \log \ell_c(\boldsymbol{X}, \boldsymbol{Z}; \theta, \mathcal{M}) + pen_{\mathcal{M}}$$

# BIC penalty if the $\boldsymbol{Z}$ are observed I

## Penalty for SBM (without covariates)

$$pen_{\mathcal{M}} = -\frac{1}{2}\left\{\underbrace{(K-1)\log(n)}_{\text{Estimation of } \pi} + \underbrace{\mathcal{K}\log(\mathcal{N})}_{\text{Estimation of } \boldsymbol{\alpha}}\right\}$$

with

$$\begin{aligned} \mathcal{K} &= K^2 && \text{if adjacency matrix } X \text{ is not symmetric} \\ &= \tfrac{K(K+1)}{2} && \text{if } X \text{ is symmetric.} \end{aligned}$$

and

$$\mathcal{N} = \begin{cases} n^2 - n & \text{if} \quad X \text{ non symmetric} \\ \frac{n^2-n}{2} & \text{if} \quad X^{qq} \text{ symmetric} \end{cases}$$

# BIC penalty if the $\boldsymbol{Z}$ are observed II

## Penalty for LBM

$$pen_{\mathcal{M}} = -\frac{1}{2}\left\{(K_1 - 1)\log(n_1) + (K_2 - 1)\log(n_2) + \mathcal{K}\log\left(\mathcal{N}\right)\right\}$$

with

$$\mathcal{K} = K_1 K_2 \quad \text{and} \quad \mathcal{N} = n_1 n_2$$

# Penalized criteria when the $\boldsymbol{Z}$ are non-observed : ICL

- Imputation of the $\boldsymbol{Z}$ by the maximum a posteriori [Biernacki et al., 2000]
  - $\widehat{\boldsymbol{Z}} = \arg\max_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X};\widehat{\theta},\mathcal{M}) \approx \arg\max_{\boldsymbol{Z}} \mathcal{R}_{\hat{\tau}}(\boldsymbol{Z}|\boldsymbol{X};\widehat{\theta},\mathcal{M})$
  - $\widehat{ICL}_{\mathcal{M}} = \log \ell_c(\boldsymbol{X},\widehat{\boldsymbol{Z}};\widehat{\theta},\mathcal{M}) + pen_{\mathcal{M}}$
- Integration of the $\boldsymbol{Z}$ [Daudin et al., 2008, Barbillon et al., 2016]
  - $ICL(\mathcal{M}) = E_{\boldsymbol{Z}|\boldsymbol{X};\widehat{\theta}\mathcal{M}}\left[\log \ell_c(\boldsymbol{X},\boldsymbol{Z};\widehat{\theta},\mathcal{M})\right] + pen_{\mathcal{M}}$
  - $p(\boldsymbol{Z}|\boldsymbol{X};\widehat{\theta},\mathcal{M}) \Rightarrow \mathcal{R}_{\boldsymbol{X},\widehat{\boldsymbol{\tau}}}$
  - $\widehat{ICL}(\mathcal{M}) = E_{\mathcal{R}_{\boldsymbol{X},\widehat{\boldsymbol{\tau}}}}[\log \ell_c(\boldsymbol{X},\boldsymbol{Z};\theta,\mathcal{M})] + pen_{\mathcal{M}}$

# Comments on ICL versus BIC

$$
\begin{aligned}
ICL(\mathcal{M}) &= BIC(\mathcal{M}) + \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X};\widehat{\theta}) \log p(\boldsymbol{Z}|\boldsymbol{X};\widehat{\theta}) \\
&\approx BIC(\mathcal{M}) - \mathcal{H}(p(\cdot|\boldsymbol{X};\widehat{\theta}))
\end{aligned}
$$

where $\mathcal{H}$ is the entropy
$\Rightarrow$ As consequence, ICL will prefer clusters with well-separated clusters (due to the entropy term)

## In practice

$$
ICL(\mathcal{M}) = BIC(\mathcal{M}) + \sum_{\boldsymbol{Z}} \mathcal{R}_{\boldsymbol{X}}(\boldsymbol{Z},\widehat{\boldsymbol{\tau}}) \log \mathcal{R}_{\boldsymbol{X},\widehat{\boldsymbol{\tau}}}(\boldsymbol{Z}) - \mathbf{KL}[\mathcal{R}_{\boldsymbol{X},\widehat{\boldsymbol{\tau}}}, p(\cdot|\boldsymbol{X};\widehat{\theta})] \,.
$$

- $\widehat{\boldsymbol{\tau}}$ such that $\mathbf{KL}[\mathcal{R}_{\boldsymbol{X},\widehat{\boldsymbol{\tau}}}, p(\cdot|\boldsymbol{X};\widehat{\theta})] \approx 0$
- $\widehat{ICL}(\mathcal{M}) \approx BIC(\mathcal{M}) - \mathcal{H}(\mathcal{R}_{\boldsymbol{X}}(\cdot,\widehat{\boldsymbol{\tau}}))$ where $\mathcal{H}$ is the entropy

# Estimation and model selection in practice

- Choosing the model ? Initialization of the VEM algorithms ?
- Limits on $K$ or $K_1$, $K_2$ : $\{K_{\min,1}, \dots, K_{\max,2}\}$

## Stepwise procedure

Starting from $K_1, K_2$

- **Split** : For each functional group $q = 1, 2$ such that $K_q < K_{\max,q}$
  - Maximize the likelihood of model $(K_1 + 1, K_2)$ and $(K_1, K_2 + 1)$
  - Respectively $K_1$ and $K_2$ proposed initializations of VEM : split of each cluster into two clusters.
  - At most : $\sum_{q=1}^{Q} K_q$ runs
- **Merge** : For each $q = 1, 2$ such that $K_q > K_{\min,q}$
  - Maximize the likelihood of model $(K_1, \dots, K_2 - 1)$ , $(K_1 - 1, \dots, K_2)$
  - $\frac{K_q(K_q-1)}{2}$ proposed initializations of VEM : merge of all the possible pairs of clusters.
  - At most : $\sum_{q=1}^{Q} \frac{K_q(K_q-1)}{2}$ runs
- Comparison of the models through ICL . If ICL improved, go on, if not stop.

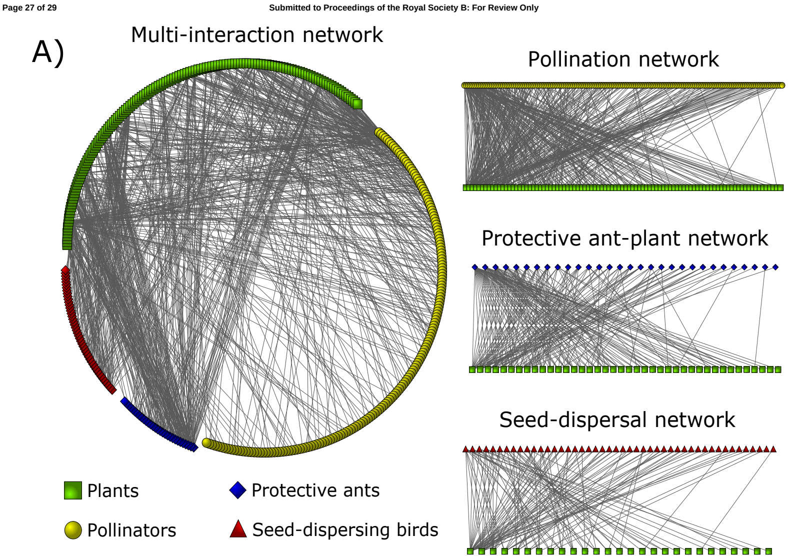
# Algorithm in practice

- Theoretical convergence not established
- In practice, good performances of model selection, stability
- R-package : blockmodels de J.-B. Léger.

# Ecological datasets

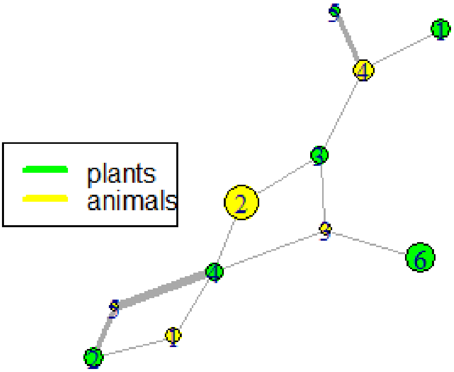
- Weasley Dattilo, Inecol, Jalapa, Mexique [Dáttilo et al., 2016]
-
  - $n_1 = 141$ plant species
  - $n_2 = 249$ animal species ( 30 ants, 46 seed dispersal birds, 173 pollinisators)
- In total 753 observed interactions

# Ecological networks



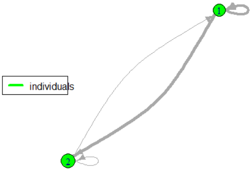
# Results of LBM inference

- 6 groups of plants - 5 groups of animals

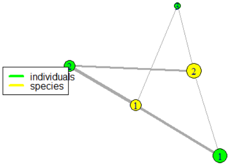


# Agroecology datasets

SBM on seed exchange dataset




LBM on inventory datasets



# Extensions

- Taking into account covariable to explain connection

$$P(X_{ij} = 1 | Z_i = k, Z_j = \ell) = logit(\alpha_{k,\ell} + y'_{ij}\beta)$$

  (S. Ouadah and S. Robin and P. Latouche)
- Networks observed along time : evolution of the blocks along time (See the works of C. Matias, T. Rebafka)
- Nature of the edges
  - multivariate [Barbillon et al., 2016]
  - textual edges (See the work by P. Latouche and colleagues)
- Observing several networks (adjacency and incidence) between several functional groups at the same time : work in progress
- Multilevel networks (interaction between individuals and between organisations where individuals belong to) : in progress
- Theoretical results for the asymptotic behavior of the estimates by Bickel

# References

Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2016).
Stochastic block models for multiplex networks : An application to a multilevel network of researchers.
*Journal of the Royal Statistical Society. Series A : Statistics in Society.*

Biernacki, C., Celeux, G., and Govaert, G. (2000).
Assessing a mixture model for clustering with the integrated completed likelihood.
*Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7) :719–725.

Dáttilo, W., Lara-Rodríguez, N., Jordano, P., Guimarães, P. R., Thompson, J. N., Marquis, R. J., Medeiros, L. P., Ortiz-Pulido, R., Marcos-García, M. A., and Rico-Gray, V. (2016).
Unravelling darwin's entangled bank : architecture and robustness of mutualistic networks with multiple interaction types.
*Proceedings of the Royal Society of London B : Biological Sciences*, 283(1843).

Daudin, J. J., Picard, F., and Robin, S. (2008).
A mixture model for random graphs.