Probabilistic models for ecological networks (SBM) Partie 1

Sophie Donnet, MIA Paris Saclay. INRAE

Ecole de printemps, Chaire MMB, Aussois, juin 2022

Colleagues

- Modèles, articles : Julie Aubert (INRAE), Pierre Barbillon (AgroParisTech), Avner Bar-Hen (CNAM), Saint-Clair Chabert Liddell (INRAE), Emmanuel Lazega (IEP), Vincent Miele (LBBE), Sarah Ouadah (AgroParisTech), Stéphane Robin (Sorbonne Université) ANR Econet
- Packages R : Jean-Benoit Léger (UTC), Julien Chiquet (INRAE)
- Données: Sonia Kefi (ISEM), Corinne Vacher (INRAE)

Network data

Networks arise when one want to study interactions between entities of a (eco)system.

Biological networks at all scales

- **Molecular networks**: gene regulation, proteines interactions,
- **Microbiote**: interactions between micro-organisms, (bacterias, fungi...)
- **Ecological networks** : Food web, Co-existence networks, Host-parasite interactions, Plant-pollinator interactions

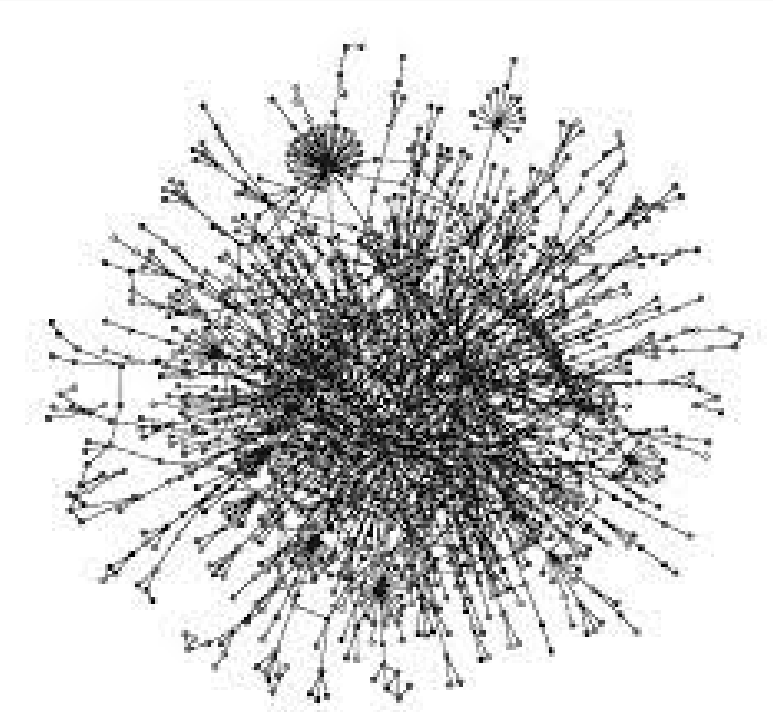
Ecological networks

- Direct observations
- Direct application: allows to modelize the robustness of an ecosystem

Why are networks popular?

- Encodes/summarizes interactions between a large number of entities
- Represent a complex system in a synthetic and generic way
- Network: interesting mathematical object

Statistical goal



- Unraveling / describing / modeling / summarizing the network organization.
- Discovering particular structure of interaction between some subsets of nodes.
- Understanding network heterogeneity.
- Inference of network: out of the scope of this talk

Terminology

A network consists in:

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

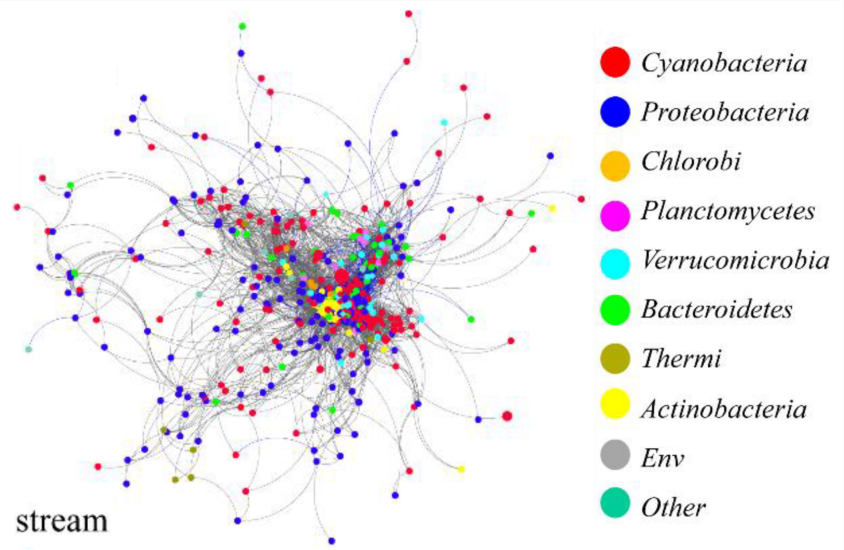
A network may be

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

Bipartite / simple network

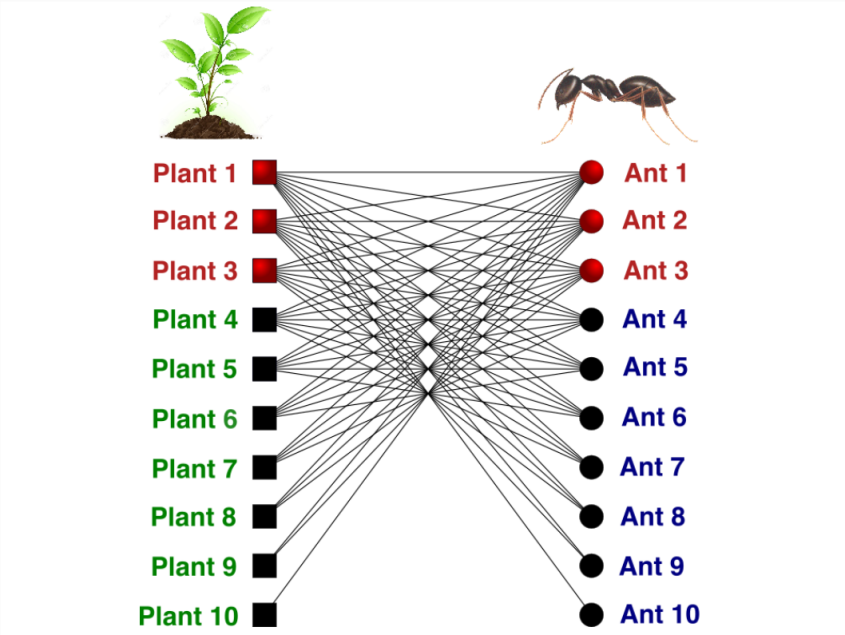
Networks may be or not bipartite: Interactions between nodes belonging to the same or to different functional group(s).

Simple network



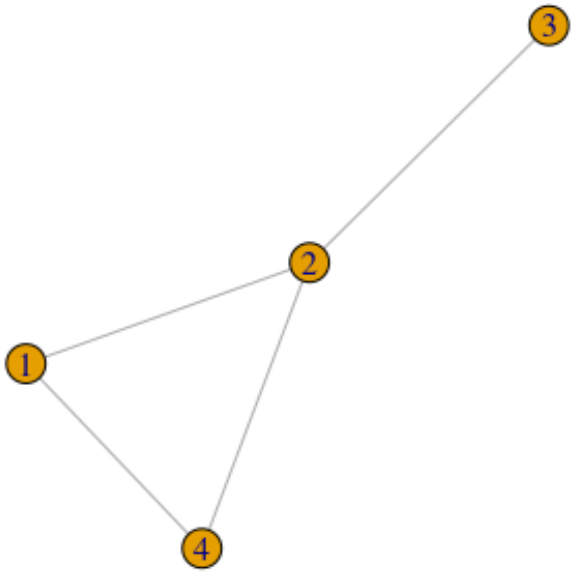
[Liu et al., 2019]

Bipartite network



Network representation and adjacency matrix

For a non-directed network

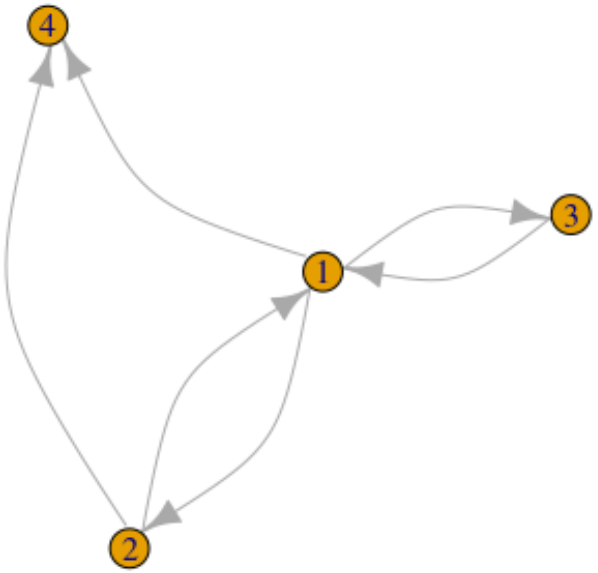


$$Y = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

- n rows and n columns,
- symmetric matrix

Network representation and adjacency matrix

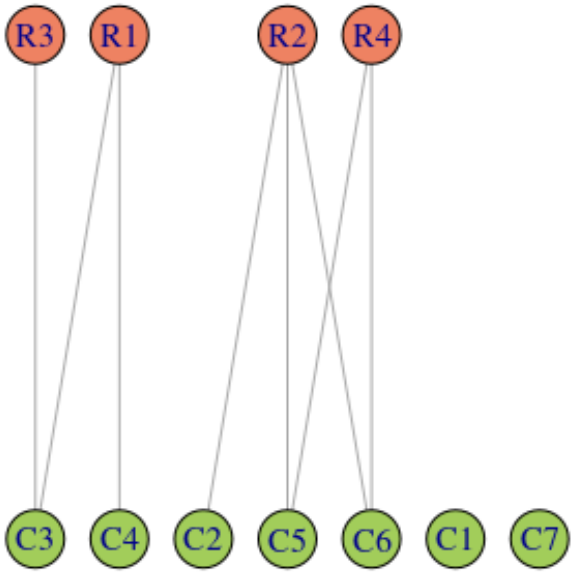
For a directed network



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

- n rows and n columns,
- non symmetric matrix

Bipartite network and incidence matrix



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

- n rows and m columns, rectangular matrix.

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: can account for sampling effort, distances between species.

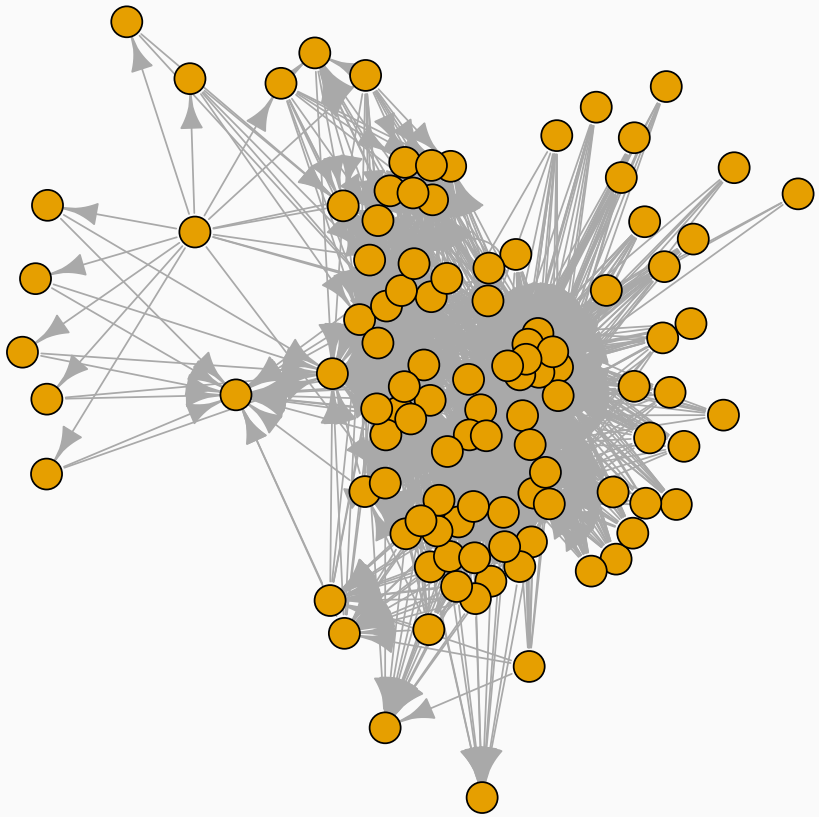
Network analysis

Aim : give a short description of the network, give a hint about its structure, look for heterogeneity in the connections

- Many metrics supplied for simple networks
- Have been extended to bipartite networks
- Metrics on nodes or on the network globally

Example : Chilean foodweb

[Kéfi et al., 2016] [Aubert et al., 2022]



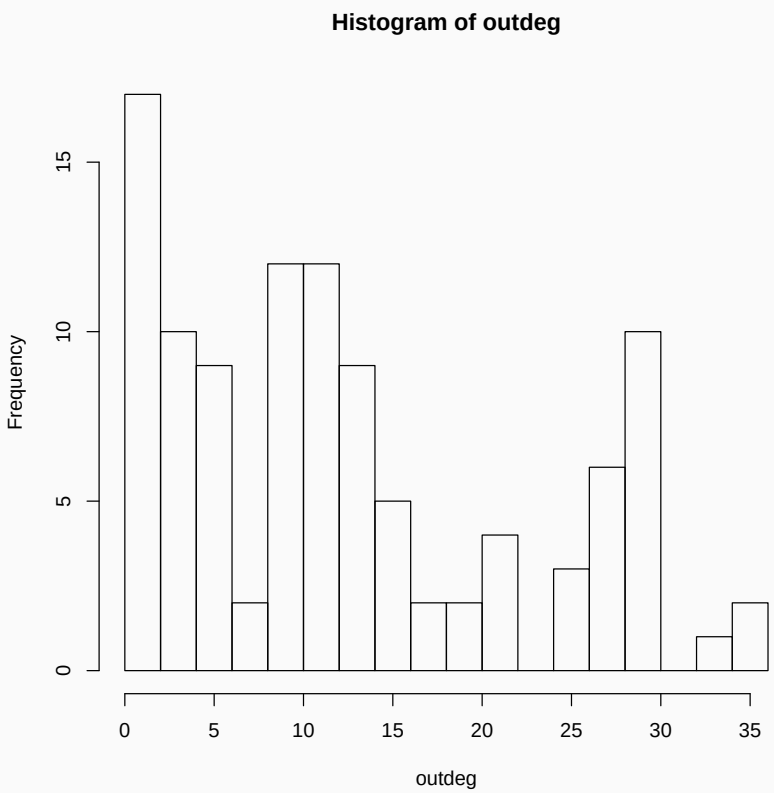
Degree

$$\begin{array}{rclcrcl} \deg(u) & = & \sum_{v \in V} (u \leftrightarrow v), & & \deg(v) & = & \sum_{u \in U} (u \leftrightarrow v) \\ \deg_i & = & \sum_{j=1}^{|V|} Y_{ij} & & \deg_j & = & \sum_{i=1}^{|U|} Y_{ij} \end{array}$$

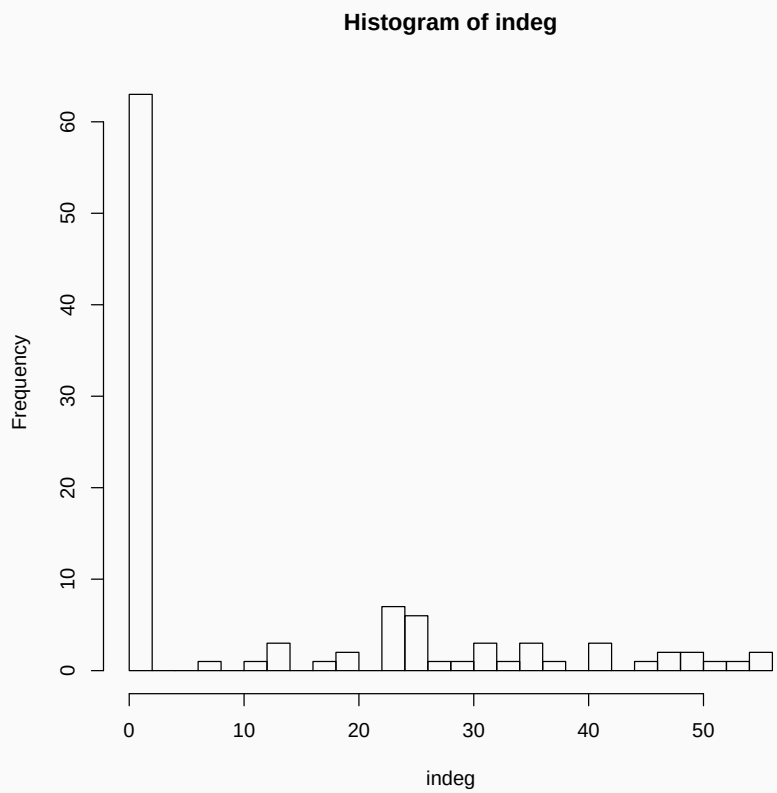
- Nodes with high degree are hubs
- Nodes with null degree are isolated
- If edges are oriented : in- and out- degrees can be computed.

Degrees on Chilean foodweb

Out degree distribution



In degree distribution



Closeness centrality

Property on a node

Definition

Determine whether a node can communicate with other nodes of the network directly or through the short paths.

$$C(u) = \frac{1}{\sum_{w \in U \cup V} d(u, w)}$$

where $d(u, w)$ is the length of the shortest path between u and w (through the network).

Note that, for bipartite networks

- A node $u \in U$ can have a minimum distance of 1 with $v \in V$.
- A node $u \in U$ can have a minimum distance of 2 with $u' \in U$.
- All paths between nodes of the same set are of even length.

Betweenness centrality

Property on a node

Definition

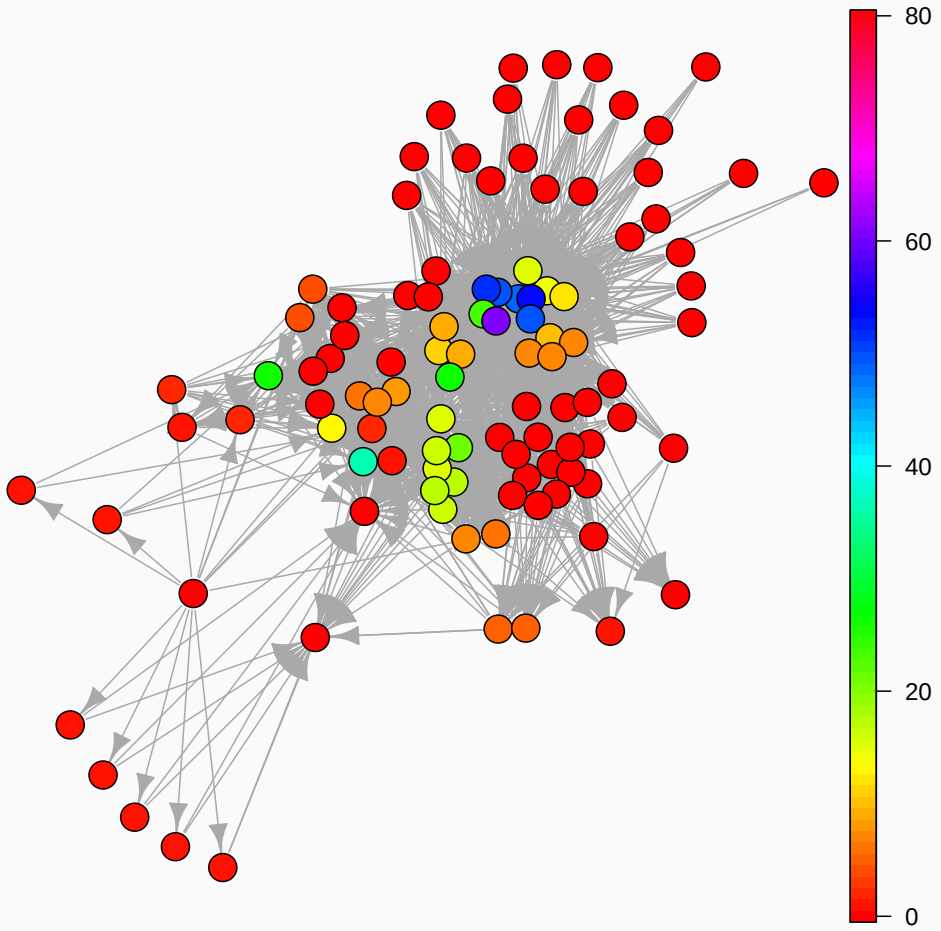
Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.

The betweenness of a vertex v is computed as follows.

- For each pair of vertices (w, w'), compute the shortest paths between them. $\delta_{w,w'}$ is the number of shortest paths between (w, w')
- For each pair of vertices (w, w'), determine the fraction of shortest paths that pass through v : $\frac{\delta_{w,w'}(v)}{\delta_{w,w'}}$
- Sum this fraction over all pairs of vertices (w, w').

$$B(v) = \sum_{w \neq w' \neq v} \frac{\delta_{w,w'}(v)}{\delta_{w,w'}}$$

Betweenness centrality



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	6.604	6.929	59.570

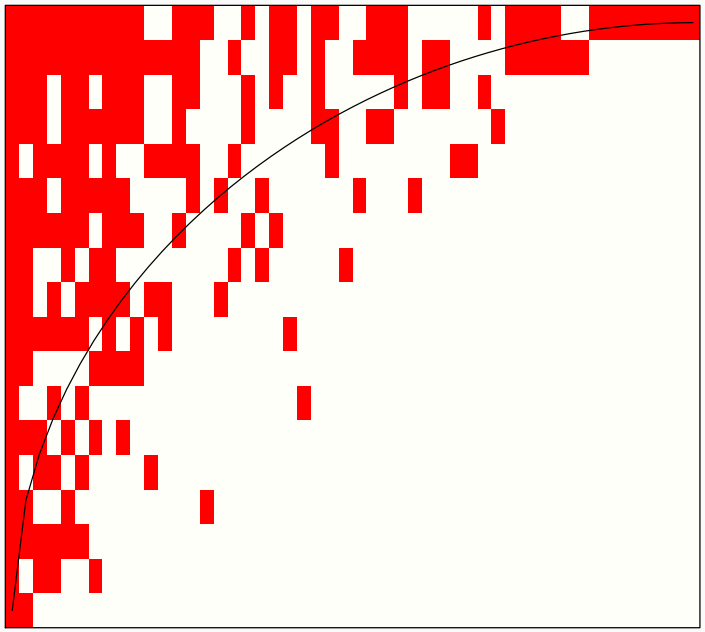
Nestedness

Property on the network

Definition

- Important property in ecology
- Defined as a pattern of interactions in which specialists (e.g. pollinators that visit few plant species) interact with plants that are visited by generalists.
- Mathematically, looking for a reordering of rows and columns such that Y is nested

Nestedness



- more generally used on incidence matrices,
- significance of the nestedness index computed by random permutations of the matrix,
- this food web is found to be nested.

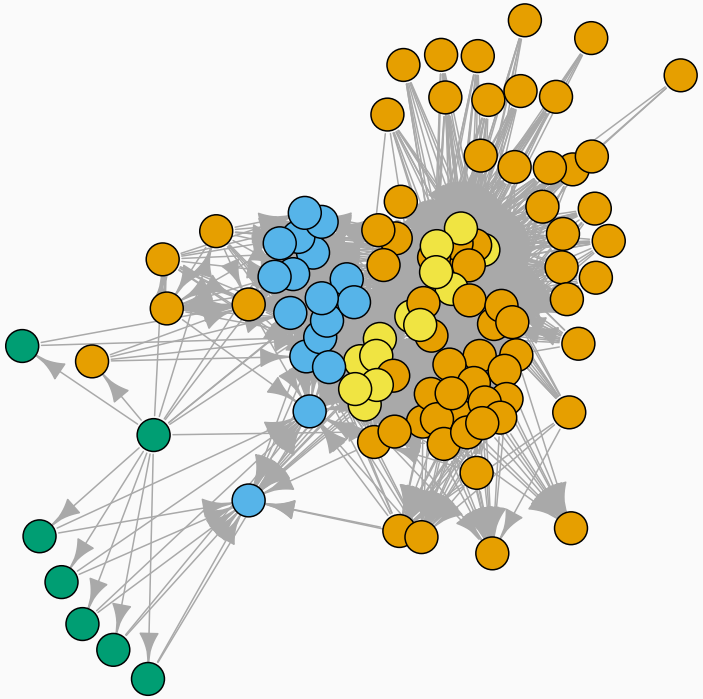
Modularity

Property on the network

Definition

Existence of clusters (blocks, module, communities) where nodes are much more connected than with other clusters

Modularity



1	2	3	4
69	17	7	13
- very low modularity.

Probabilistic approach

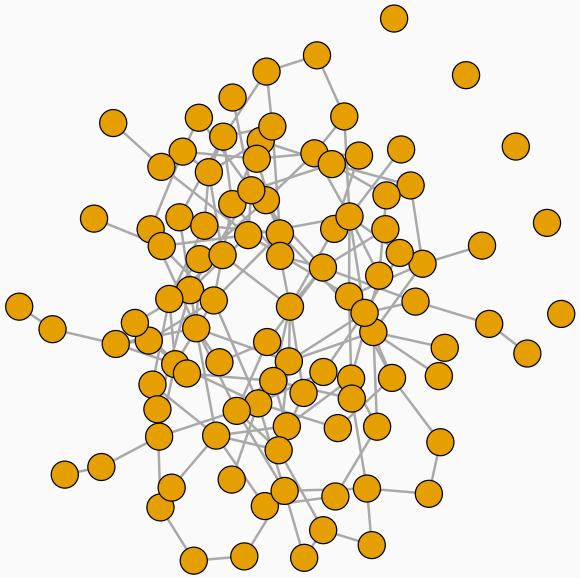
- Context: our matrix $\mathbf{Y} = (Y_{ij})_{i,j=1,\dots,n}$ is the realization of a stochastic process.
- Aims:
  - Propose a stochastic process is able to mimic heterogeneity in the connections and
  - Adjusting its parameters to fit the data $\mathbf{Y}$.
- Advantages:
  - Benefit from the statistical toolkit:
    - Theoretical (asymptotic) properties, Tests, model selection
  - Easy to extend to more complexe networks, to non binary interactions, to partially observed networks etc...

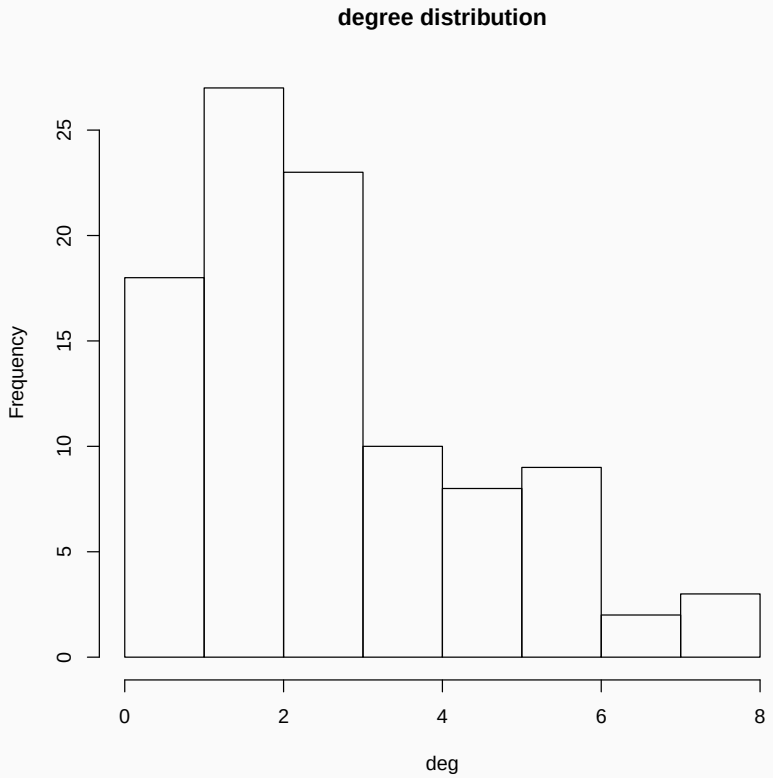
A first random graph model for network: null model

[Erdös and Rényi, 1959] Model for n nodes

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

where $\mathcal{B}ern$ is the Bernoulli distribution and $p \in [0, 1]$ a probability for a link to exist.





Limitations of an ER graph to describe real networks

- Homogeneity of the connections
- Degree distribution too concentrated

$$D_i \sim \mathcal{B}in(n, p)$$

  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

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

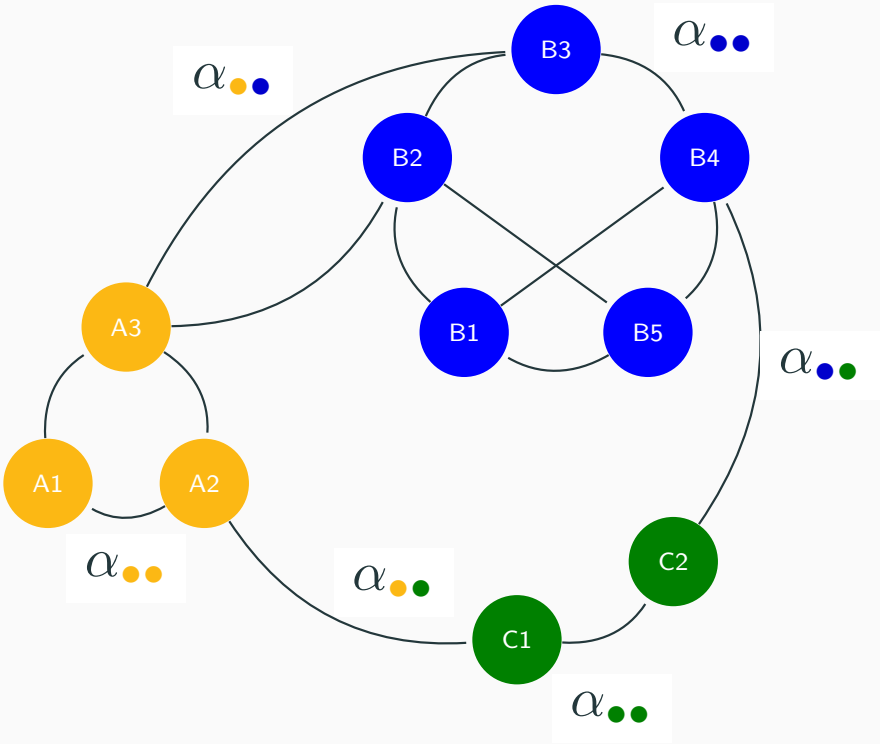
Conditionally to $(Z_i)_{i=1,\ldots,n}$...

(Y_{ij}) independant and

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

$$\mathbf{Y} \sim \mathrm{SBM}_n(K, \boldsymbol{\pi}, \boldsymbol{\alpha})$$

Stochastic Block Model : illustration



Parameters

Let n nodes divided into 3 clusters

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

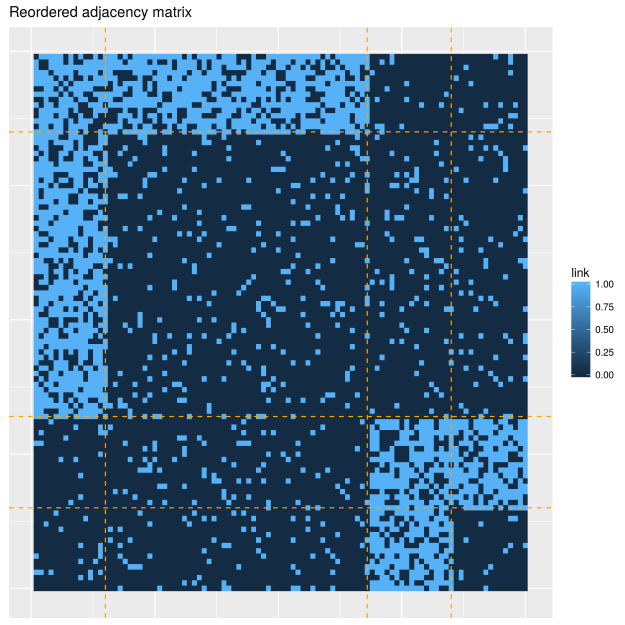
$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \pi), \quad \forall \bullet \in \mathcal{K},$$
$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{B}(\alpha_{\bullet\bullet})$$

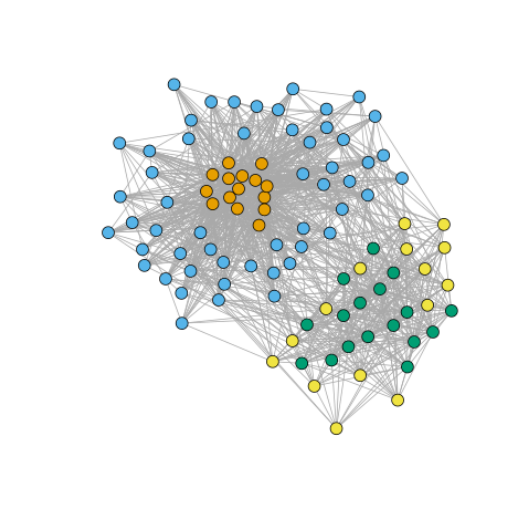
SBM : A great generative model

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

Networks with hubs generated by SBM

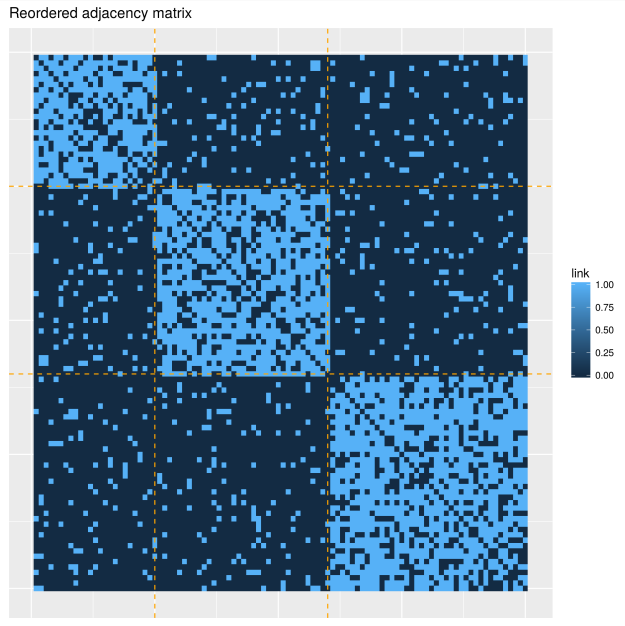
- $\pi = c(.15, .35, .15, .35)$
- $\alpha = \begin{pmatrix} 0.80 & 0.80 & 0.20 & 0.20 \\ 0.80 & 0.20 & 0.20 & 0.20 \\ 0.20 & 0.20 & 0.80 & 0.80 \\ 0.20 & 0.20 & 0.80 & 0.20 \end{pmatrix}$

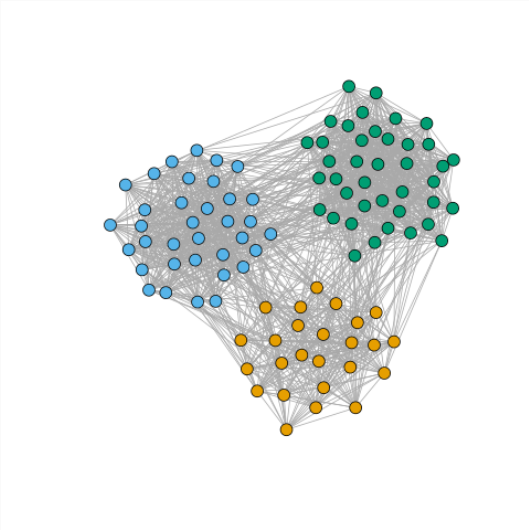




Community network generated by SBM

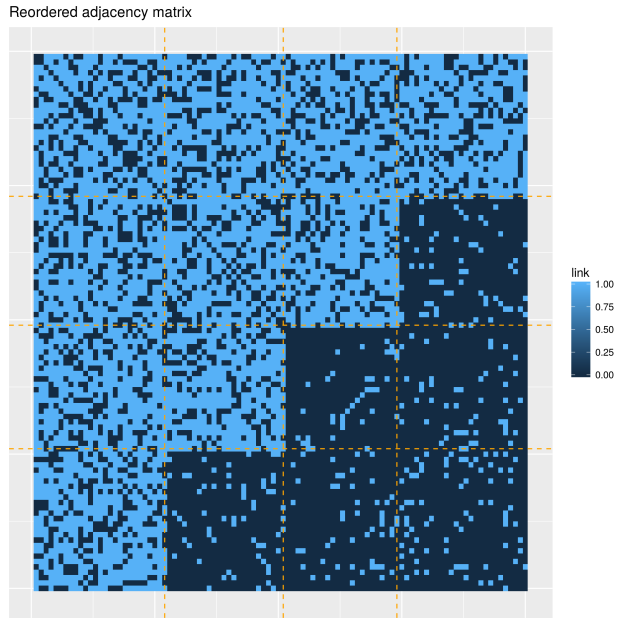
- $\pi = c(0.25, 0.35, 0.40)$
- $\alpha = \begin{pmatrix} 0.80 & 0.20 & 0.20 \\ 0.20 & 0.80 & 0.20 \\ 0.20 & 0.20 & 0.80 \end{pmatrix}$

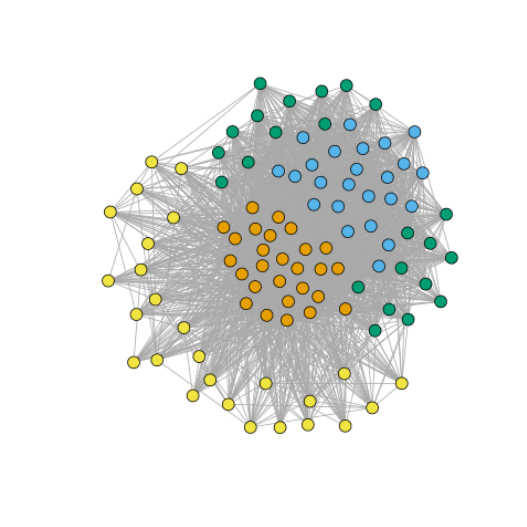




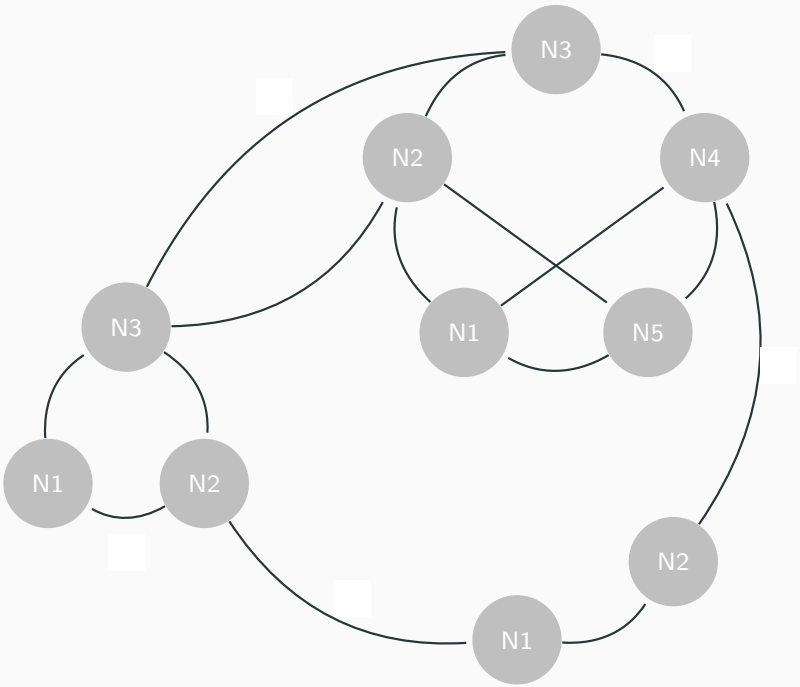
Nestedness generated by SBM

- $\pi = c(.15, .35, .15, .35)$
- $\alpha = \begin{pmatrix} 0.80 & 0.80 & 0.80 & 0.80 \\ 0.80 & 0.80 & 0.80 & 0.20 \\ 0.20 & 0.80 & 0.20 & 0.80 \\ 0.80 & 0.20 & 0.20 & 0.20 \end{pmatrix}$





Statistical inference



Stochastic Block Model

Let n nodes divided into

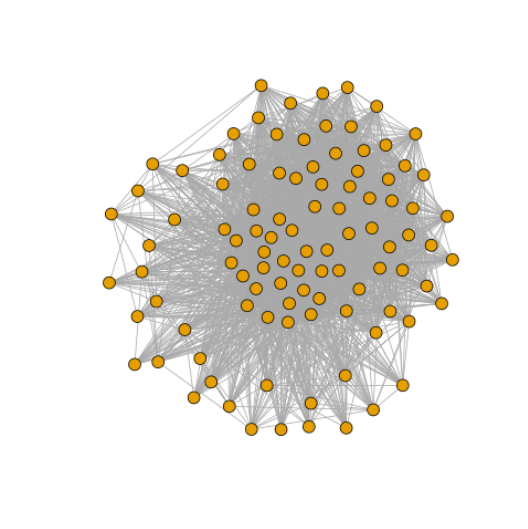
- $\mathcal{K} = \{\bullet, \bullet, \bullet\}$, card$(\mathcal{K})$ known
- $\pi_\bullet = ?$,
- $\alpha_{\bullet\bullet} = ?$

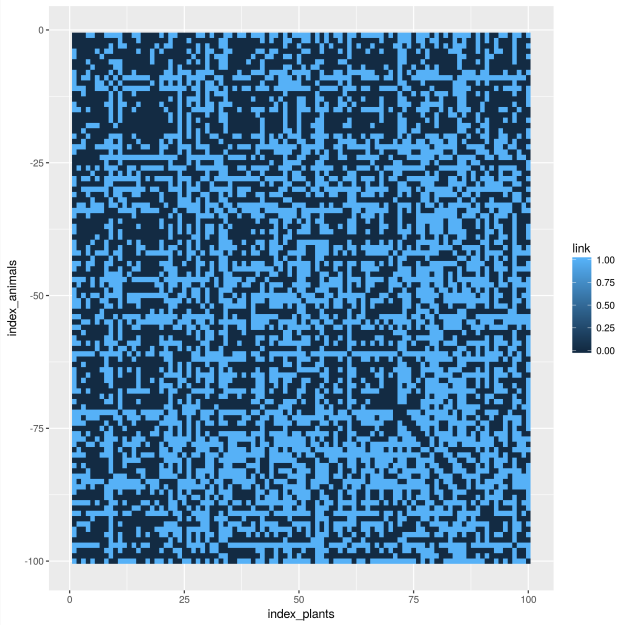
[Nowicki and Snijders, 2001], [Daudin et al., 2008]

R package: blockmodels, sbm

Statistical inference

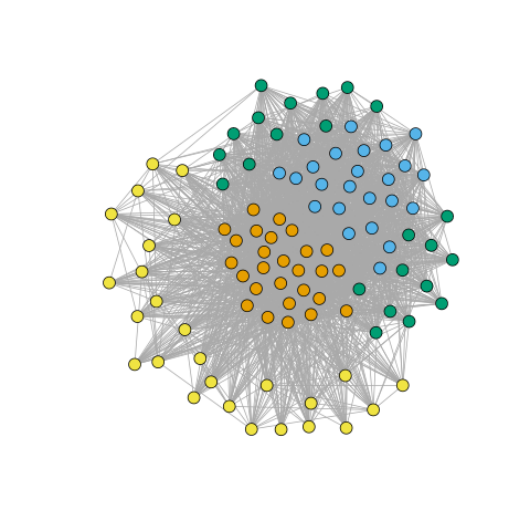
From....

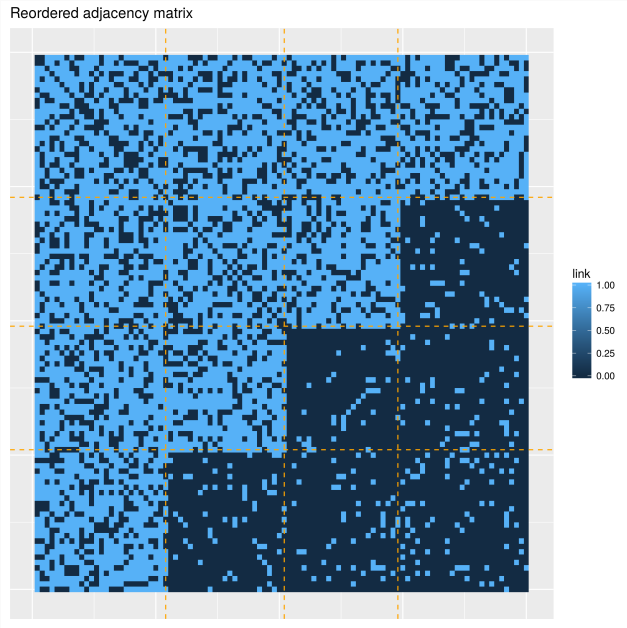




Statistical inference

... to





Statistical inference

Tasks

- For a fixed number of clusters/blocks K
  - Estimate the parameters :
    - Block proportions π
    - probabilities of connexion inside and between blocks α
  - Get the better clustering $\widehat{Z}$
- Find the number of clusters K

Practical implementation + Theoretical results

Extension to bipartite networks

Let $(Y_{ij})_{i,j}$ be a bi-partite network. Individuals in row and cols are not the same (plants - pollinators for instance)

Latent variables : bi-clustering

- Nodes $i = 1, \ldots, n_1$ partitionned into K_1 clusters, nodes $j = 1, \ldots, n_2$ partitionned into K_2 clusters
- $$Z_i^1 = k \quad \text{if node } i \text{ belongs to cluster (block) } k$$
  $$Z_j^2 = \ell \quad \text{if node } j \text{ belongs to cluster (block) } \ell$$
- Z_i^1, Z_j^2 independent variables

$$\mathbb{P}(Z_i^1 = k) = \pi_k^1, \quad \mathbb{P}(Z_j^2 = \ell) = \pi_\ell^2$$

Probabilistic model for binary bipartite networks

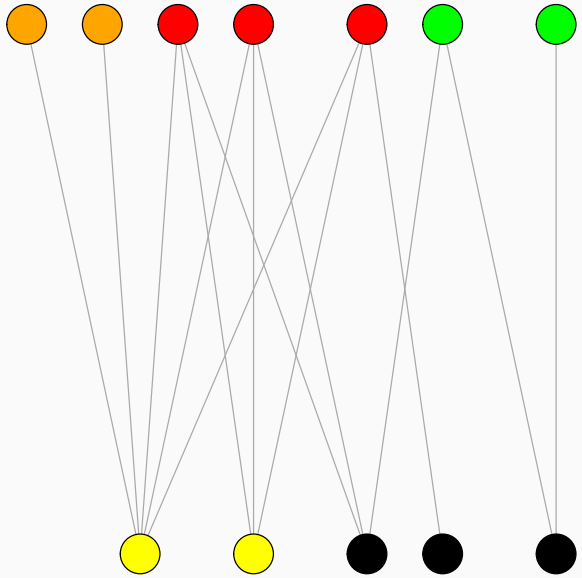
Conditionally to $(Z_i^1)_{i=1,\dots,n_1}, (Z_j^2)_{j=1,\dots,n_2}$ **...**

(Y_{ij}) independent and

$$Y_{ij}|Z_i^1, Z_j^2 \sim \mathcal{B}ern(\alpha_{Z_i^1, Z_j^2}) \quad \Leftrightarrow \quad \mathbb{P}(Y_{ij} = 1|Z_i^1 = k, Z_j^2 = \ell) = \alpha_{k\ell}$$

[Govaert and Nadif, 2008]

Latent Block Model : illustration



Latent Block Model

- n_1 row nodes $\mathcal{K}_1 = \{\bullet, \bullet, \bullet\}$ classes
- $\pi^1_\bullet = \mathbb{P}(i \in \bullet), \bullet \in \mathcal{K}_1, i = 1, \ldots, n$
- n_2 column nodes $\mathcal{K}_2 = \{\bullet, \bullet\}$ classes
- $\pi^2_\bullet = \mathbb{P}(j \in \bullet), \bullet \in \mathcal{K}_2, j = 1, \ldots, m$
- $\alpha_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

$$
\begin{aligned}
Z_i^1 = \mathbf{1}_{\{i \in \bullet\}} &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\pi}^1), \quad \forall \bullet \in \mathcal{Q}_1, \\
Z_j^2 = \mathbf{1}_{\{j \in \bullet\}} &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\pi}^2), \quad \forall \bullet \in \mathcal{Q}_2, \\
Y_{ij} \mid \{i \in \bullet, j \in \bullet\} &\sim^{\text{ind}} \mathcal{B}ern(\alpha_{\bullet\bullet})
\end{aligned}
$$

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: blockmodels, 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

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

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

Statistical Inference

- Selection of the number of clusters K for SBM or K_1, K_2 for LBM
- Estimation of the parameters π, θ_K for a given number of clusters
- Clustering $\hat{\mathbf{Z}}$

Maximum likelihood

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \ell(\mathbf{Y}; \theta) = \arg\max_{\theta \in \Theta} \log \ell(\mathbf{Y}; \theta)$$

Looking for the value of θ such that under my SBM model, the observed data is most probable.

Likelihood for SBM

With latent variables $\mathbf{Z}$

(Marginal) Likelihood ($\mathbf{Y}$)

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

Complete likelihood (Y) et (Z)

For binary directed networks

$$\begin{cases} Y_{ij}|Z_i = k, Z_j = \ell & \sim_{ind} & \mathcal{B}ern(\alpha_{k\ell}) \\ P(Z_i = k) & = & \pi_k \\ (Z_i)_{i=1,\dots n} & & \text{independent} \end{cases}$$

$$\begin{aligned} \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) &= p(\mathbf{Y}|\mathbf{Z}; \alpha) p(\mathbf{Z}; \boldsymbol{\pi}) \\ &= \prod_{i \neq j = 1}^{n} \alpha_{Z_i, Z_j}^{Y_{ij}} (1 - \alpha_{Z_i, Z_j})^{1 - Y_{ij}} \prod_{i=1}^{n} \pi_{Z_i} \\ &= \prod_{(i \neq j) = 1}^{n} \prod_{k=1}^{K} \prod_{\ell=1}^{K} \left(\alpha_{k\ell}^{Y_{ij}} (1 - \alpha_{k\ell})^{1 - Y_{ij}} \right)^{\mathbf{1}_{Z_i = k} \mathbf{1}_{Z_j = \ell}} \times \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k)^{\mathbf{1}_{Z_i = k}} \end{aligned}$$

Marginal likelihood : remarks

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

Remarks

- $\mathcal{Z} = \{1, \ldots, K\}^n \Rightarrow$ when K and n increase, heavy/impossible to compute.
- Because of the $\sum_{\mathbf{Z} \in \mathcal{Z}}$, setting the derivatives with respect to the parameters π, α to 0 will not lead to an explicit solution

Standard tool to maximize the likelihood when latent variables involved : EM algorithm.

EM algorithm [Dempster et al., 1977]

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**:

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

Why EM seems to be a convenient solution?

Reason 1: E-step and M-step produce a sequence $\theta^{(t-1)}$ such that

$$\log \ell(\mathbf{Y}; \theta^{(t-1)}) \leq \log \ell(\mathbf{Y}; \theta^{(t)})$$

⏩ Proof

Reason 2: relies on $\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)$

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

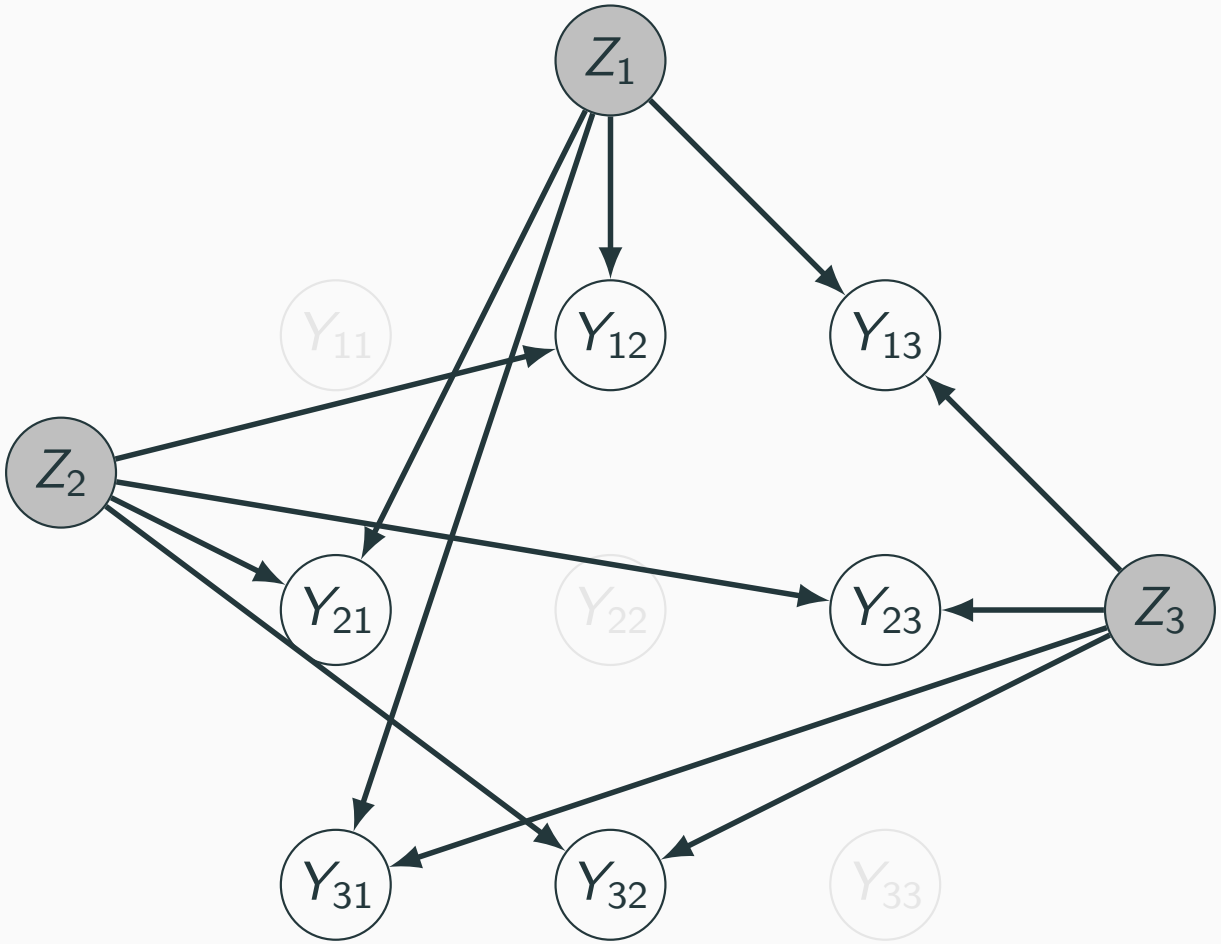
Why EM is NOT a convenient solution?

Reason 2: relies on ~~$\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)$~~ $\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}\left[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)\right]$

$$\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}\left[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)\right]$$
$$= \sum_{(i\neq j)=1,k,\ell=1}^{n,K} \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}\left[\mathbf{1}_{Z_i=k}\mathbf{1}_{Z_j=\ell}\right]\left[Y_{ij}\log\alpha_{k\ell} + (1-Y_{ij})\log(1-\alpha_{kl})\right]$$
$$+ \sum_{i=1,k=1}^{n,K} \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}\left[\mathbf{1}_{Z_i=k}\right]\log\pi_k$$

Problem: Z_i and Z_j are not independent conditionally to $\mathbf{Y}$

DAG



- Step E requires the computation of $\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)]$
- However, once conditioned by par $\mathbf{Y}$, the $\mathbf{Z}$ are not independent anymore: complex distribution if K and n big.

Variational EM : maximization of a lower bound

Idea : replace the complicated distribution $p(\mathbf{Z}|\mathbf{Y};\theta) = [\mathbf{Z}|\mathbf{Y},\theta]$ by a simpler one.

Let $\mathcal{R}_{\mathbf{Y},\tau}$ be any distribution on $\mathbf{Z}$ depending on a parameter τ.

Lower bound

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) \quad = \quad \log \ell(\mathbf{Y};\theta) - \mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y};\theta)] \quad \leq \log \ell(\mathbf{Y};\theta)$$

About the Kullback-Leibler divergence

- $\mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y};\theta)] = \int_{\mathbf{Z}} \mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}) \log \frac{\mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{Y};\theta)} d\mathbf{Z} \geq 0$
- $\mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y};\theta)] \neq \mathbf{KL}[p(\cdot|\mathbf{Y};\theta), \mathcal{R}_{\mathbf{Y},\tau}]$
- $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) = \log \ell(\mathbf{Y};\theta) \Leftrightarrow \mathcal{R}_{\mathbf{Y},\tau} = p(\cdot|\mathbf{Y};\theta)$

Variational EM : central equality

Let $\mathcal{R}_{\mathbf{Y},\tau}$ be any distribution on $\mathbf{Z}$ depending on a parameter τ.

Central equality

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

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

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

Proof ii

As a consequence:

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

So as stated

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) = \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log \ell_c(\mathbf{Y},\mathbf{Z};\theta)] + \mathcal{H}(\mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z})) \leq \log \ell(\mathbf{Y};\theta)$$

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.

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) = \log \ell(\mathbf{Y}; \theta) - \mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y}; \theta)] \quad \leq \log \ell(\mathbf{Y}; \theta)$$

Benefits

- Reformulation $\mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}} [\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)] + \mathcal{H}(\mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}))$
- Choose $\mathcal{R}_{\mathbf{Y},\tau}$ such that the maximization / expectation 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 VE** Maximization w.r.t. τ

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

Variational EM

Algorithm

- **Step M** Maximization w.r.t. θ

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

Lower bound for SBM

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

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

- Integration with $\mathbf{Z} \sim \mathcal{R}_{\mathbf{Y},\tau}$

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

with $\log p(Y_{ij}|\alpha_{k\ell}) = Y_{ij} \log \alpha_{k\ell} + (1 - Y_{ij}) \log(1 - \alpha_{k\ell})$

M-step for 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 $\boldsymbol{\pi}$ is quite direct:

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

For the Bernoulli SBM:

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

M-step for 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.

VE-step for 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)

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,

VE-step for SBM ii

- 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.

In practice

- Really fast
- Strongly depends on the initial values

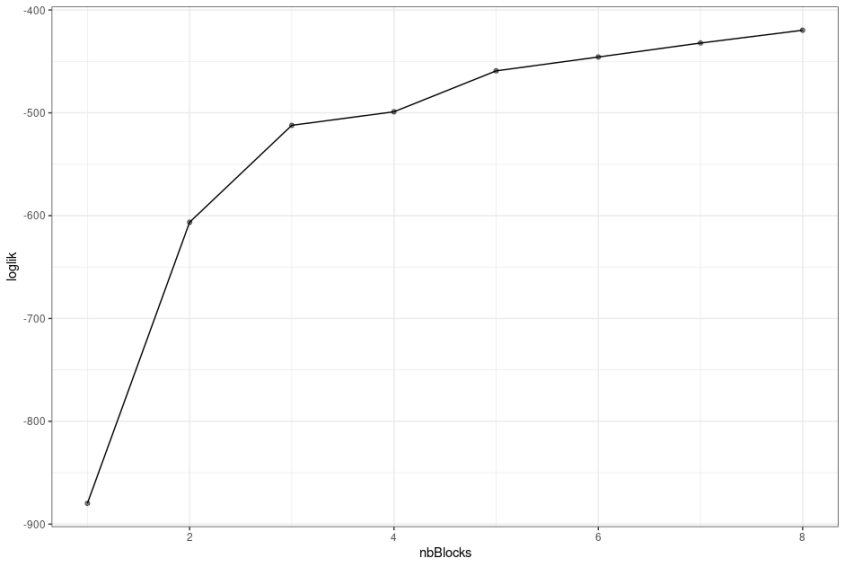
Theoretical properties

- 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]

Model selection objective

Aim: choosing the number of clusters K (or K_1, K_2 in the LBM)

Remark $K \mapsto \log \ell(\mathbf{Y}, \widehat{\theta}_K)$



- Maximized likelihood is not a good criterion
- Occam's razor (philocophical principle): a model with fewer parameters, is to be preferred.

Penalized likelihood

Introduce a penalty term taking into account the number of parameters to estimate AND the size of the data $\mathbf{Y}$

$$\hat{K} = \arg\max_K \log \ell(\mathbf{Y}, \widehat{\theta}_K) - \text{pen}(n, K)$$

when $\text{pen}(n, K)$ has to be chosen.

Bayesian model selection in two slides i

Generally speaking

- Let $\mathcal{M}_K$ be a stochastic model depending on parameters θ_K

$$\mathbf{Y}_n|\theta_K, \mathcal{M}_K \sim \ell_{\mathcal{M}_K}(\cdot|\theta_K)$$

- Prior distributions
  - Prior distribution on $\theta_K|\mathcal{M}_K = p_K(\theta_K)$
  - Prior distribution on $\mathcal{M}_K = p(\mathcal{M}_K) \propto 1$
- Posterior probability

$$p(\mathcal{M}_K|\mathbf{Y}_n) = \frac{m(\mathbf{Y}_n|\mathcal{M}_K)p(K)}{p(\mathbf{Y}_n)} \propto m(\mathbf{Y}|\mathcal{M}_K)p(\mathcal{M}_K) \propto m(\mathbf{Y}_n|\mathcal{M}_K)$$

Bayesian model selection in two slides ii

- Best model a posteriori

  Chose

$$\widehat{\mathcal{M}}_K = \arg\max_{\mathcal{M}_K} p(\mathcal{M}_K|\mathbf{Y}_n) = \arg\max_{\mathcal{M}_K} m(\mathbf{Y}|\mathcal{M}_K)$$

Bayesian Information Criterion

- Relies on $m(\mathbf{Y}_n|\mathcal{M}_K)$ where

$$m(\mathbf{Y}_n|\mathcal{M}_K) = \int_{\theta_K} \ell_{\mathcal{M}_K}(\mathbf{Y}_n|\theta_K) p_K(\theta_K) d\theta_K$$

  θ_K has to be integrated out!

- Asymptotic approximation (Laplace approximation)

$$\log m(\mathbf{Y}_n|\mathcal{M}_K) \underset{n\to\infty}{} \max_{\theta_K} \log \ell_{\mathcal{M}_K}(\mathbf{Y}_n|\theta_K) - \frac{dim(\theta_K)}{2} \log n.$$

- BIC

$$\hat{K} = \arg\max_K \log \ell(\mathbf{Y}, \widehat{\theta}_K) - \frac{dim(\theta_K)}{2} \log n$$

Limitations for SBM

- $\ell_{\mathcal{M}_K}(\mathbf{Y}_n|\theta_K)$: latent variables have been integrated. Too heavy to compute
- Laplace approximation relies on regularity conditions that we do not have here

Towards ICL [Biernacki et al., 2000] i

BIC if $\mathbf{Z}$ is observed.

- Assume that $\mathbf{Y}$ and $\mathbf{Z}$ are distributed as SBM.
- Let $p(\theta_K)$ be a prior distribution on θ_K.

$$\begin{aligned}(\pi_1, \dots \pi_K) &\sim \mathcal{D}ir(b, \dots, b) \\ \alpha_{k\ell} &\sim \mathcal{B}eta(a, c)\end{aligned}$$

Then

$$\begin{aligned}\log m(\mathbf{Y}, \mathbf{Z}|K) &= \log \int_{\theta_K} p(\mathbf{Y}, \mathbf{Z}; \theta_K) p(d\theta_K) \\ &\approx_{n\to\infty} \max_{\theta_K} \log p(\mathbf{Y}, \mathbf{Z}; \theta_K) - \mathrm{pen}(\mathcal{M}_K) \qquad (2)\end{aligned}$$

Towards ICL [Biernacki et al., 2000] ii

where

$$\mathrm{pen}(\mathcal{M}_K) = \frac{1}{2}\left\{(K-1)\log(n) + K^2 \log(n(n-1))\right\} \tag{3}$$

⏩ Proof

From BIC to ICL

- Latent variables $\mathbf{Z}$ are not observed
- Remove $\mathbf{Z}$ in BIC : either maximize or integrate

$$\mathrm{ICL}(\mathcal{M}_K) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{Y},\hat{\theta}_K)}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\theta}}_K)] - \mathrm{pen}(\mathcal{M}_K). \tag{4}$$

Finally

$$\hat{K} = \arg\max_K \mathrm{ICL}(\mathcal{M}_K)$$

[Biernacki et al., 2000]

Remarks

$$\begin{aligned}
\mathrm{ICL}(\mathcal{M}_K) &= \mathbb{E}_{p(\mathbf{Z}|\mathbf{Y},\hat{\boldsymbol{\theta}}_K)}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \hat{\boldsymbol{\theta}}_K)] - \mathrm{pen}(\mathcal{M}_K). \\
&= \ell(\mathbf{Y}; \hat{\boldsymbol{\theta}}_K) - \mathcal{H}\left(p(\mathbf{Z}|\mathbf{Y}, \hat{\boldsymbol{\theta}}_K)\right) - \mathrm{pen}(\mathcal{M}_K) \\
&= \mathrm{BIC}(\mathcal{M}_K) - \mathcal{H}\left(p(\mathbf{Z}|\mathbf{Y}, \hat{\boldsymbol{\theta}}_K)\right)
\end{aligned}$$

As a consequence, because of the entropy, ICL will encourage clustering with well-separated groups

Advantages of the 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.

Penalization term for bipartite networks

$$pen_{\mathcal{M}} = -\frac{1}{2} \left\{ \underbrace{(K_1 - 1)\log(n_1) + (K_2 - 1)\log(n_2)}_{\text{Bi-Clust.}} + \underbrace{(K_1 K_2)\log(n_1 n_2)}_{\text{Connection}} \right\}$$

Algorithm in practice

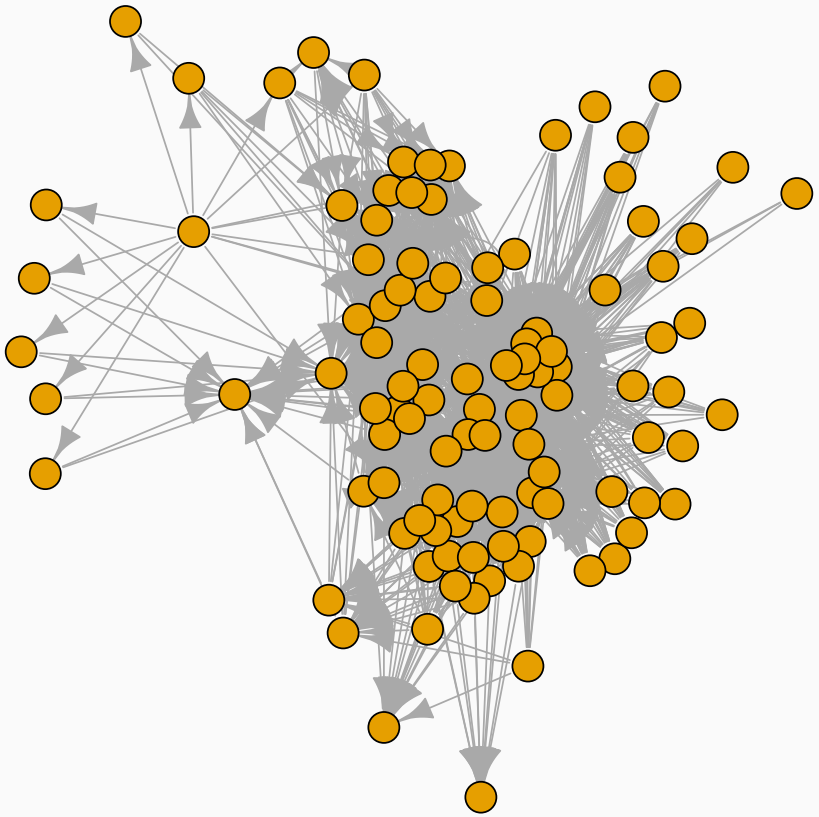
- Going trough the models and initiate VEM at the same time
- Bounds on K : $\{K_{\text{min}}, \ldots, K_{\text{max}}\}$

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

Chilean foodweb

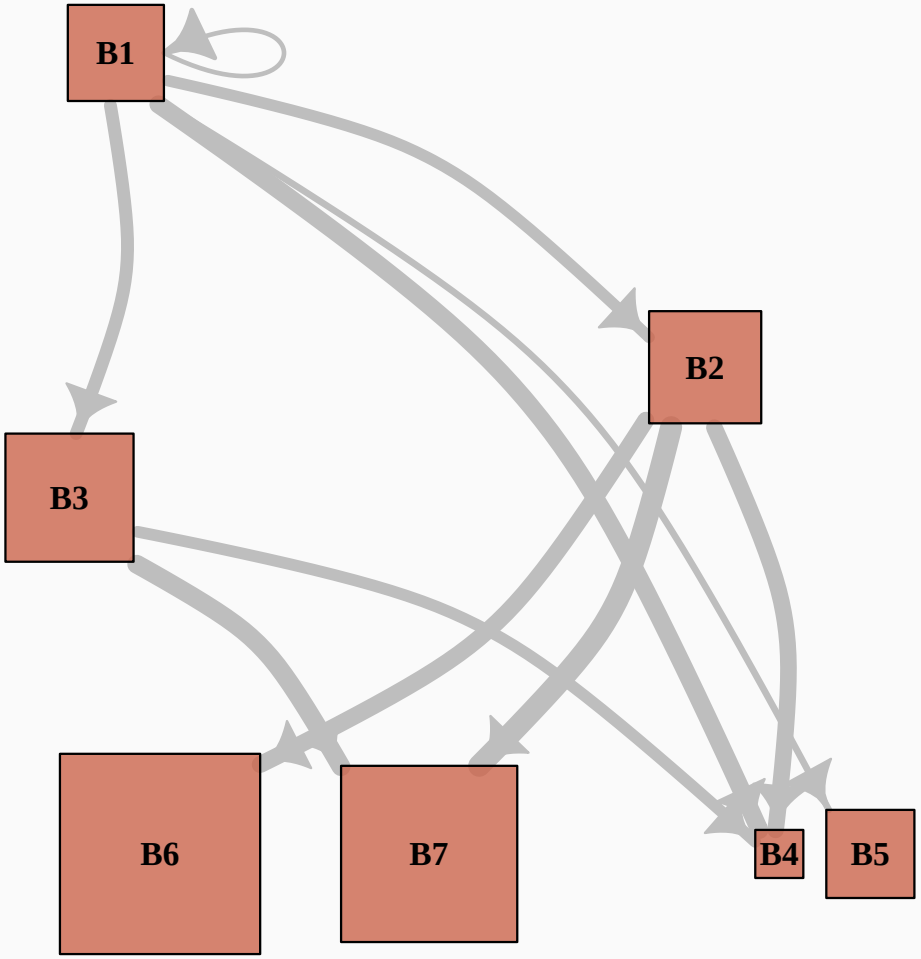


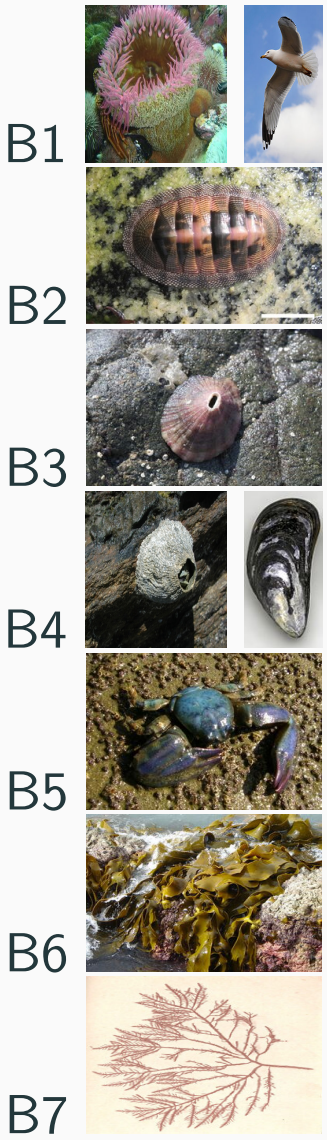
Chilean foodweb

- Intertidal zone of the Chilean Pacific coast
- 106 animal or plant species, sessile or mobile
- 1362 trophic interactions
- [Kéfi et al., 2016]

Application on Chilean

7 blocs





Application on Chilean foodweb

- Schematic representation (inspired by [Picard et al., 2009])
- Left: each vertex is a block and the thickness of the edges represents the probability of interactions between each block (above the 0.1 threshold, for clarity)
- Right: type of species representative of each block. From top to bottom: anemone and gull (B1), chiton (B2), *Fissurella* (B3), *Balanus* and mussel (B4), crab (B5), *Laminariale* (B6) and red algae (B7)

Studying the blocks

- B1:
  - gather the "super-predators" (top of the trophic chain) which have no predators except some rare trophic links between them
  - wide taxonomic variability, including diverse species such as the anemone or the gull
- ...
- B6 and B7 contain basal algal species, including brown algae and red algae respectively, and which are resources for various mollusks.

Finally

- SBM allows to summarize the complexity induced by the observation of more than a thousand interactions.
- The interpretation of its parameters (the probabilities of interactions between each block) allows a synthetic description of the ecosystem,
- Interpretation of the blocks with exogenous information such as taxonomy and ecological traits.

Application on an a tree-fungus bipartite network

R package sbm

Vignette of the sbm R package

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,
- Sampling effort (Emré Anakok, Pierre Barbillon, Colin Fontaine et Elisa Thébault)

Probabilistic model for networks in a nutshell

SBM/LBM

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

References i

Aubert, J., Barbillon, P., Donnet, S., and Miele, V. (2022).

Using Latent Block Models to Detect Structure in Ecological Networks, chapter 6, pages 117–134.

John Wiley and Sons, Ltd.

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.

Brault, V. (2014).

Estimation et sélection de modèle pour le modèle des blocs latents.

PhD thesis, Université Paris Sud-Paris XI.

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.

References ii

Daudin, J.-J., Picard, F., and Robin, S. (2008).

A mixture model for random graphs.

Statistics and computing, 18(2):173–183.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977).

Maximum likelihood from incomplete data via the em algorithm.

Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22.

Erdös, P. and Rényi, A. (1959).

On random graphs i.

Publicationes Mathematicae Debrecen, 6:290.

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.

Liu, Y., Qu, X., Elser, J. J., Peng, W., Zhang, M., Ren, Z., Zhang, H., Zhang, Y., and Yang, H. (2019).

Impact of nutrient and stoichiometry gradients on microbial assemblages in erhai lake and its input streams.

Water, 11(8).

References iii

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.

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

Estimation and prediction for stochastic blockstructures.

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

Picard, F., Miele, V., Daudin, J.-J., Cottret, L., and Robin, S. (2009).

Deciphering the connectivity structure of biological networks using mixnet.

BMC Bioinformatics, 10(6):S17.

About the EM algorithm

- First proposed by [Dempster et al., 1977] for a large class of incomplete data models, including mixture models.
- Based on a decomposition of the incomplete data likelihood.

Proposition (Decomposition of the log-likelihood)

For any θ and θ'

$$\log p_\theta(\mathbf{Y}) = \mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Y}, \mathbf{Z}) | Y\right] - \mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}\right].$$

Proof

It suffices to develop

$$\mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}\right] \quad = \quad \mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Y},\mathbf{Z}) - \log p_\theta(\mathbf{Y})|\mathbf{Y}\right]$$

reminding that $\mathbb{E}_{\theta'}\left[\log p_\theta(\mathbf{Y})|\mathbf{Y}\right] = \log p_\theta(Y)$.

Remarks

1. Decomposition of Slide 106 is convenient bacause makes a connexion between $\log p_\theta(\mathbf{Y})$ (often intractable) and $\log p_\theta(\mathbf{Y}, \mathbf{Z})$ (generally more manageable).
2. if $\theta' = \theta$, the second term is the entropy of the latent variables $\mathbf{Z}$ given the observed $\mathbf{Y}$:

$$\mathcal{H}[p_\theta(\mathbf{Z}|\mathbf{Y})] := -\mathbb{E}_\theta[\log p_\theta(\mathbf{Z}|\mathbf{Y})|\mathbf{Y}]$$

EM Algorithm

$$\widehat{\theta} = \arg\max_{\theta} \log p_{\theta}(\mathbf{Y}).$$

Algorithm (EM)

Repeat until convergence:

Expectation step *(E-step) given the current estimate* θ^h *of* θ*, compute* $p_{\theta^h}(\mathbf{Z}|\mathbf{Y})$*, or at least all the quantities needed to compute* $\mathbb{E}_{\theta^h}[\log p_{\theta}(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}]$*;*

Maximization step *(M-step) update the estimate of* θ *as*

$$\theta^{h+1} = \arg\max_{\theta} \mathbb{E}_{\theta^h}[\log p_{\theta}(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}].$$

Property

Proposition

The log-likelihood of the observed data $\log p_\theta(\mathbf{Y})$ *increases at each step:*

$$\log p_{\theta^{h+1}}(\mathbf{Y}) \geq \log p_{\theta^h}(\mathbf{Y}).$$

Proof i

Because $\theta^{h+1} = \arg\max_\theta \mathbb{E}_{\theta^h}[\log p_\theta(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}]$, we have

$$
\begin{aligned}
0 &\leq \mathbb{E}_{\theta^h}[\log p_{\theta^{h+1}}(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}] - \mathbb{E}_{\theta^h}[\log p_{\theta^h}(\mathbf{Y}, \mathbf{Z})|\mathbf{Y}] && (5)\\
&= \mathbb{E}_{\theta^h}\left[\log \frac{p_{\theta^{h+1}}(\mathbf{Y}, \mathbf{Z})}{p_{\theta^h}(\mathbf{Y}, \mathbf{Z})}|\mathbf{Y}\right] && (6)\\
&\leq \log \mathbb{E}_{\theta^h}\left[\frac{p_{\theta^{h+1}}(\mathbf{Y}, \mathbf{Z})}{p_{\theta^h}(\mathbf{Y}, \mathbf{Z})}|\mathbf{Y}\right] && (7)
\end{aligned}
$$

by Jensen's inequality.

Proof ii

We further develop $\log \mathbb{E}_{\theta^h}\left[p_{\theta^{h+1}}(\mathbf{Y},\mathbf{Z})\,/\,p_{\theta^h}(\mathbf{Y},\mathbf{Z})\,|\mathbf{Y}\right]$ as

$$\begin{aligned}
\log \int \frac{p_{\theta^{h+1}}(\mathbf{Y},\mathbf{Z})}{p_{\theta^h}(\mathbf{Y},\mathbf{Z})} p_{\theta^h}(\mathbf{Z}|\mathbf{Y})\,\mathrm{d}\mathbf{Z} &= \log \int \frac{p_{\theta^{h+1}}(\mathbf{Y},\mathbf{Z})}{p_{\theta^h}(\mathbf{Y},\mathbf{Z})} \frac{p_{\theta^h}(\mathbf{Y},\mathbf{Z})}{p_{\theta^h}(\mathbf{Y})}\,\mathrm{d}\mathbf{Z} && (8)\\
&= \log \left[\frac{1}{p_{\theta^h}(\mathbf{Y})} \int p_{\theta^{h+1}}(\mathbf{Y},\mathbf{Z})\,\mathrm{d}\mathbf{Z}\right] && (9)\\
&= \log \left[\frac{p_{\theta^{h+1}}(\mathbf{Y})}{p_{\theta^h}(\mathbf{Y})}\right] && (10)
\end{aligned}$$

Finally :

$$\log \left[\frac{p_{\theta^{h+1}}(\mathbf{Y})}{p_{\theta^h}(\mathbf{Y})}\right] \geq 0$$

Convergence

There is no general guaranty about the convergence of the EM algorithm towards the MLE $\widehat{\theta}$. The main property is that the observed likelihood increases at each iteration step.

Although, in practice : very sensible to the initialisation point.

Illustration of the problems of convergence (I)

Mixture model

$$Y_i \sim_{i.i.d.} \frac{1}{2}\mathcal{N}(\mu_1, 1) + \frac{1}{2}\mathcal{N}(\mu_2, 2)$$

or equivalently

$$P(Z_i = 1) = P(Z_i = 2) = \frac{1}{2}$$

$$Y_i | Z_i = k \sim \mathcal{N}(\mu_k, 1)$$

Illustration of the problems of convergence (I)

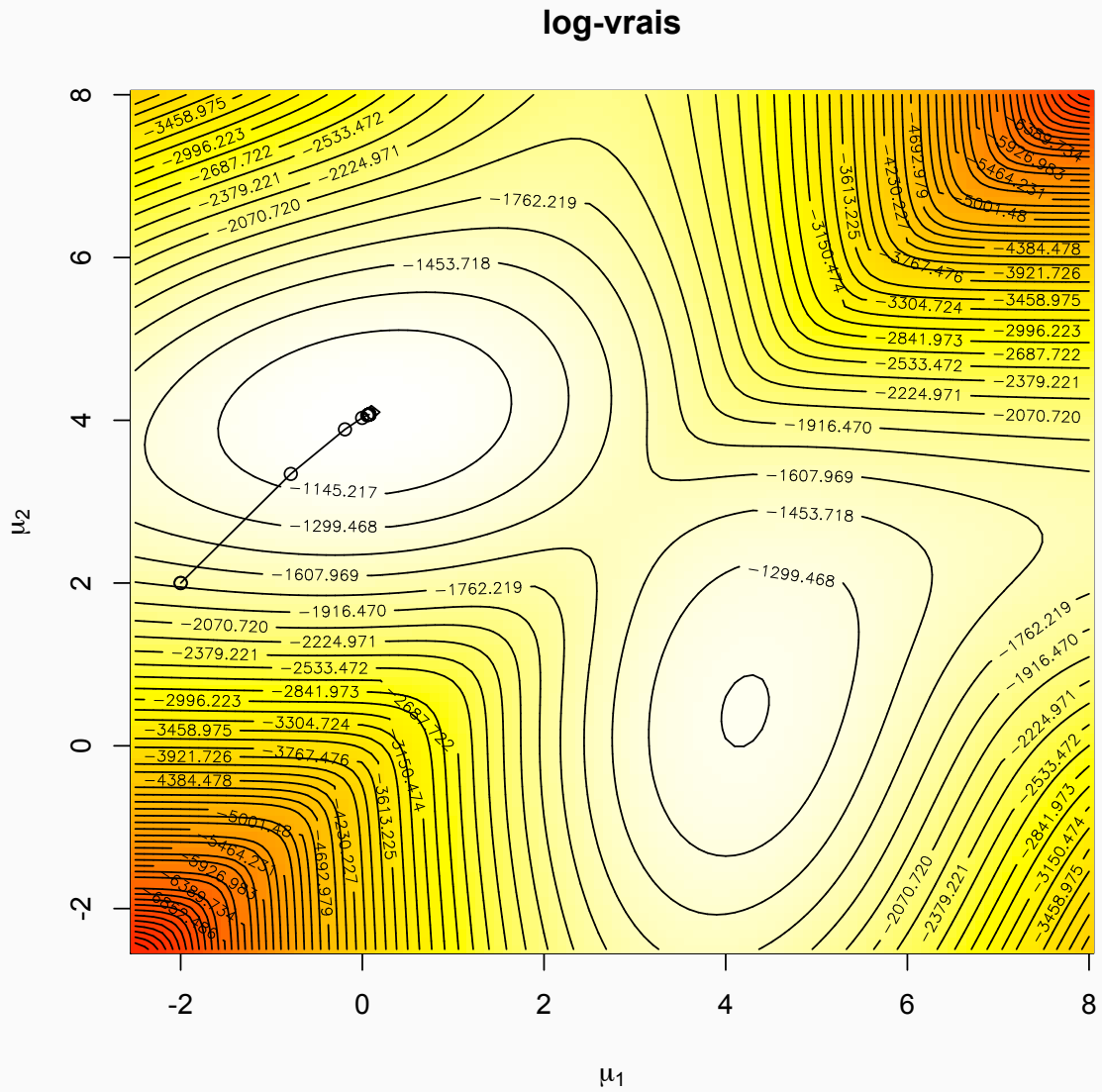


Illustration of the problems of convergence (II)

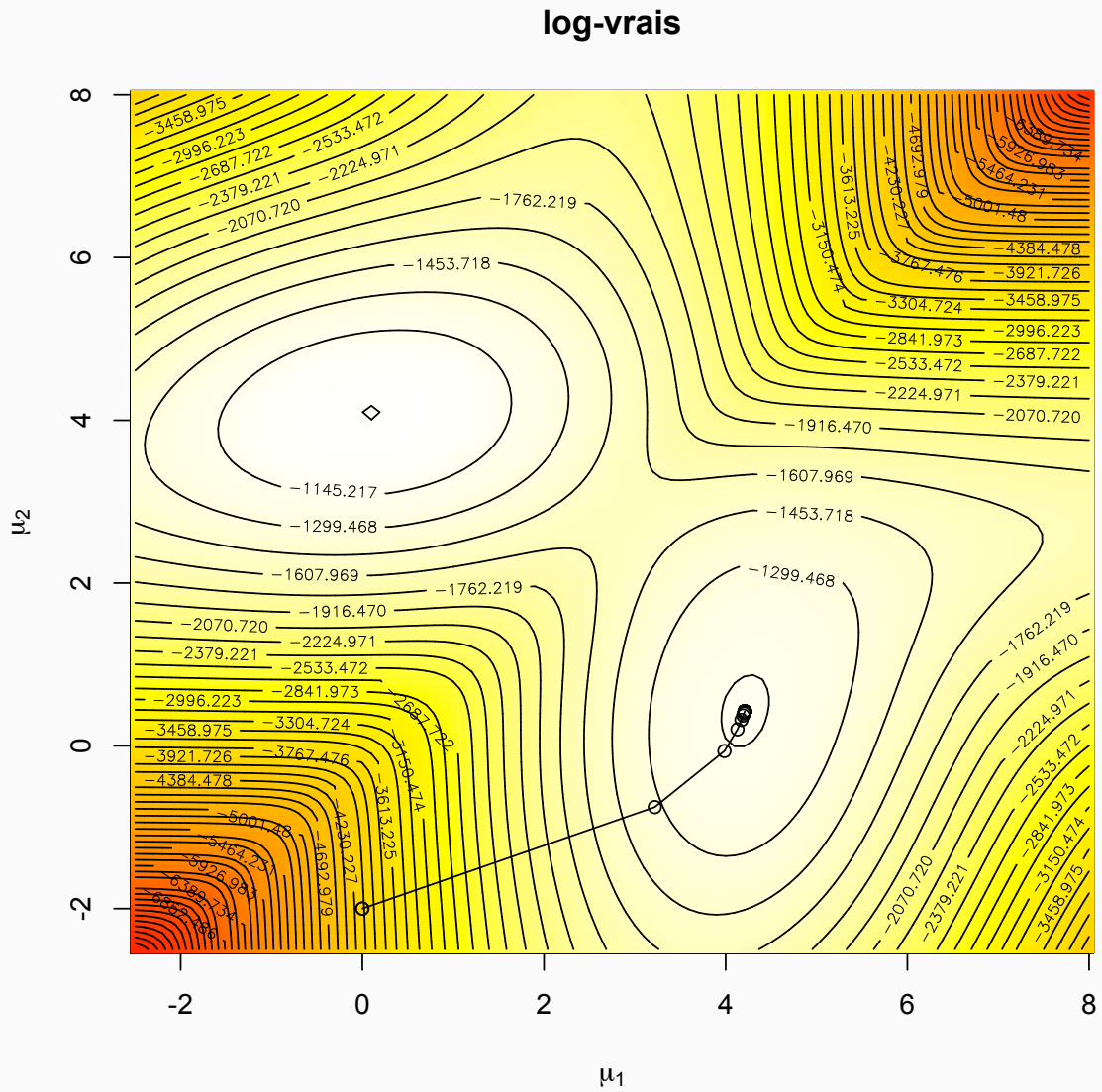
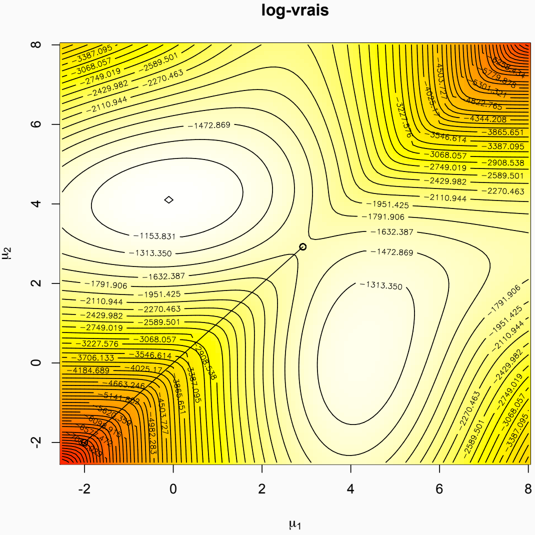


Illustration of the problems of convergence (III)



▸▸ Back to the talk

Proof of the ICL expression

By definition, the marginal complete likelihood is:

$$\begin{aligned}
\log m(\mathbf{Y}, \mathbf{Z}|K) &= \log \int p(\mathbf{Y}, \mathbf{Z}; \theta_K) p(d\theta_K) \\
&= \log \int p(\mathbf{Y}|\mathbf{Z}; \boldsymbol{\alpha}_K) p(d\boldsymbol{\alpha}_K) + \log \int p(\mathbf{Z}; \boldsymbol{\pi}_K) p(d\boldsymbol{\pi}_K) \\
&= \log m(\mathbf{Y}|\mathbf{Z}, K) + \log m(\mathbf{Z}|K)
\end{aligned}$$

Proof: about $\log m(\mathbf{Y}|\mathbf{Z}, K)$ i

$$\begin{aligned}\log \int p(\mathbf{Y}|\mathbf{Z};\boldsymbol{\alpha})p(d\boldsymbol{\alpha}) &= \log \int \prod_{k,\ell=1}^{K} \prod_{i,j|Z_i=k,Z_j=\ell} p(Y_{ij}|Z_i,Z_j;\alpha_{k\ell})p(\alpha_{k\ell})d\alpha_{k\ell} \\ &= \sum_{k,\ell=1}^{K} \log\left[\int e^{\sum_{(i,j)\in\mathcal{S}_{k\ell}} f(Y_{ij},\alpha_{k\ell})} p(\alpha_{k\ell})d\alpha_{k\ell}\right]\end{aligned}$$

with $\mathcal{S}_{k\ell} = \{(i,j)|Z_i = k, Z_j = \ell\}$.

In each term, $(Y_{ij})_{ij}$ are i.i.d: $\forall(k,\ell)$, BIC-like approximation:

$$\begin{aligned}&\log\left[\int e^{\sum_{(i,j)\in\mathcal{S}_{k\ell}} f(X_{ij},\alpha_{k\ell})} p(\alpha_{k\ell})d\alpha_{k\ell}\right] \\ \approx_{n\to\infty} \quad &\max_{\alpha_{k\ell}} \sum_{(i,j)\in\mathcal{S}_{k\ell}} f(X_{ij},\alpha_{k\ell}) - \frac{1}{2}\log\left(\frac{n(n-1)}{2}\right) + O_n(1)\,.\end{aligned}$$

Proof: about $\log m(\mathbf{Y}|\mathbf{Z}, K)$ ii

As a consequence,

$$\log m(\mathbf{Y}|\mathbf{Z}, K) \approx_{n\to\infty} \max_{\alpha} \log p(\mathbf{Y}|\mathbf{Z}; \boldsymbol{\alpha}) - \frac{1}{2}\frac{K(K+1)}{2} \log\left(\frac{n(n-1)}{2}\right) \tag{11}$$

Proof: about $\log m(\mathbf{Z}|K)$ i

- $p(\mathbf{Z};\boldsymbol{\pi}) = \prod_{k=1}^{K} (\pi_k)^{N_k}$ with $N_k = \sum_{i=1}^{n} \mathbf{1}_{Z_i=k}$
- Dirichlet prior distribution is conjugate

$$\log m(\mathbf{Z}|K) = \log \int p(\mathbf{Z};\boldsymbol{\pi}) p(\boldsymbol{\pi}) d\boldsymbol{\pi} = \log \frac{\Gamma(bK)}{\Gamma(b)^K} \frac{\prod_{k=1}^{K} \Gamma(N_k + b)}{\Gamma(n + bK)} \tag{12}$$

where Γ is the Gamma function.

$$\log m(\mathbf{Z}|K) \approx_{n\to\infty} \max_{\boldsymbol{\pi}} \log p(\mathbf{Y}|\mathbf{Z};\boldsymbol{\pi}) - \frac{1}{2}(K-1)\log(n) \tag{13}$$

[Daudin et al., 2008] [Brault, 2014]

⏩ Back to the talk