Détection de structures à l’aide de modèles probabilistes sur les graphes

Modèles

Pierre Barbillon

21 juin 2024

My collaborators

On the R packages



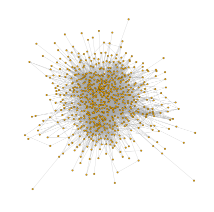

J. Chiquet	S. Donnet	J.B. Léger	Saint-Clair Chabert-Liddell
(INRAE)	(INRAE)	(Univ. Tech. Compiègne)	(INRAE)
sbm	sbm, GREMLINS	blockmodels	colSBM

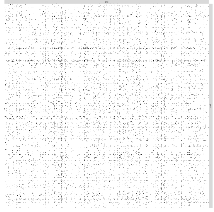
Other collaborators

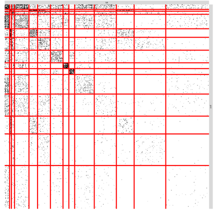
T. Tabouy (ex PhD student), E. Lazega (Sciences Po), L. Lacoste (new PhD student), E. Anakok (PhD student) + E. Thébault (iEES) + C. Fontaine (MNHN) + T. Vanrenterghem (INRAE), ANR Econet + ANR Pastodiv + GDR Resodiv

Recall

from the observation of a network determine structure







Outline

Stochastic and Latent Block Models

Other latent space models and other methods

Extensions of SBM

Outline

A first random graph model for network

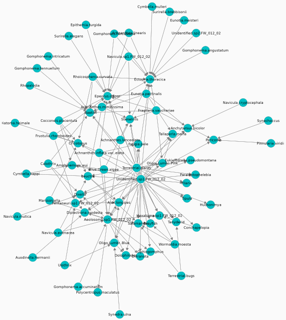
[Erdős and Rényi, 1959] Model for n nodes

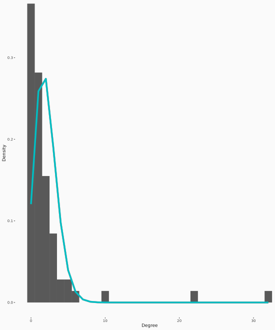
$$\forall 1 \leq i, j \leq n, \quad Y_{ij} \overset{i.i.d.}{\sim} \mathcal{B}\text{ern}(p),$$

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

Consequence

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





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 modularity, no hubs

Stochastic Block Model and Latent Block Model

Model on a simple network with n nodes:

SBM: [Nowicki and Snijders, 2001]

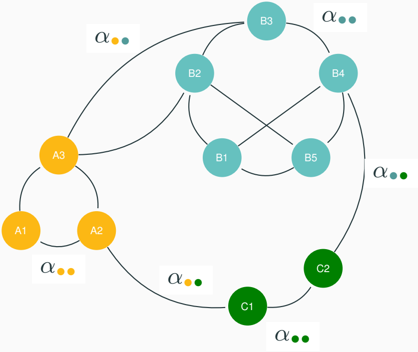
- Q blocks of nodes sharing similar connection structure,
- $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ independent latent variables s.t. $\mathbb{P}(Z_i = k) = \pi_k$ for $k \in \{1, \ldots, Q\}$ and $i \in \{1, \ldots, n\}$,
- $Y_{ij} | Z_i, Z_j \overset{ind}{\sim} \mathcal{F}(\alpha_{Z_i, Z_j})$ for all dyads (i, j)

Model on a bipartite network with n_1 and n_2 nodes:

LBM: [Govaert and Nadif, 2010]

- Q_1 and Q_2 blocks of nodes sharing similar connection structure,
- $\boldsymbol{Z}^1 = (Z_1^1, \ldots, Z_n^1)$ and $\boldsymbol{Z}^2 = (Z_1^2, \ldots, Z_m^2)$ independent latent variables s.t. $\mathbb{P}(Z_i^1 = k) = \pi_k^1$ for all $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, Q_1\}$ and $\mathbb{P}(Z_j^2 = l) = \pi_l^2$ for all $j \in \{1, \ldots, m\}$, $l \in \{1, \ldots, Q_2\}$
- $Y_{ij} | Z_i^1, Z_j^2 \overset{ind}{\sim} \mathcal{F}(\alpha_{Z_i^1, Z_j^2})$ for all dyads (i, j).

Stochastic Block Model : illustration



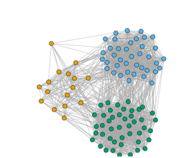
Parameters
Let n nodes divided into 3 clusters

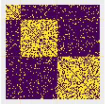
- $\{\bullet, \bullet, \bullet\}$ clusters
- $\pi_{\bullet} = \mathbb{P}(i \in \bullet),\, i = 1, \ldots, n$
- $\alpha_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

$$\mathbf{Y} \sim \text{SBM}_n(Q, \boldsymbol{\pi}, \boldsymbol{\alpha})$$

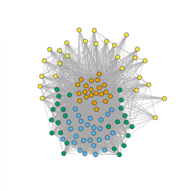
Simulations under the SBM

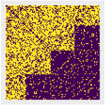
$$\boldsymbol{\alpha} = \begin{pmatrix} 0.70 & 0.09 & 0.09 \\ 0.09 & 0.70 & 0.09 \\ 0.09 & 0.09 & 0.70 \end{pmatrix}$$



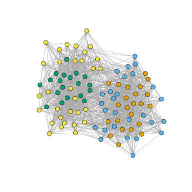


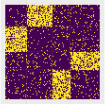
$$\boldsymbol{\alpha} = \begin{pmatrix} 0.70 & 0.70 & 0.70 & 0.70 \\ 0.70 & 0.70 & 0.70 & 0.09 \\ 0.70 & 0.70 & 0.09 & 0.09 \\ 0.70 & 0.09 & 0.09 & 0.09 \end{pmatrix}$$





$$\boldsymbol{\alpha} = \begin{pmatrix} 0.09 & 0.70 & 0.09 & 0.09 \\ 0.70 & 0.09 & 0.09 & 0.09 \\ 0.09 & 0.09 & 0.09 & 0.70 \\ 0.09 & 0.09 & 0.70 & 0.09 \end{pmatrix}$$





Likelihood for SBM

Complete likelihood $(\mathbf{Y})$ **et** (Z)

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

Marginal likelihood $(\mathbf{Y})$

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

Remark
$\mathcal{Z} = \{1, \ldots, Q\}^n \Rightarrow$ when Q and n increase, impossible to compute.

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

Complete likelihood $(\mathbf{Y})$ et (Z)

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

Marginal likelihood $(\mathbf{Y})$

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

Remark
$\boldsymbol{\mathcal{Z}} = \{1,\dots,Q\}^n \Rightarrow$ when Q and n increase, impossible to compute.

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

Complete likelihood $(\mathbf{Y})$ et $(\boldsymbol{Z})$

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

Marginal likelihood $(\mathbf{Y})$

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

Remark

$\boldsymbol{\mathcal{Z}} = \{1,\dots,Q\}^n \Rightarrow$ when Q 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

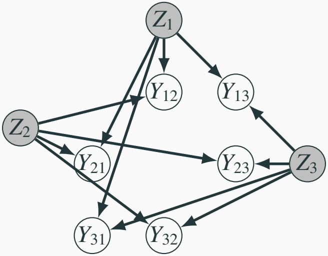
$$\mathcal{Q}(\theta|\theta^{(t-1)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}\left[\log \ell_c(\mathbf{Y},\mathbf{Z};\theta)\right]$$

- **Step M**:

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

Limitation of EM

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



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

Variational EM : maximization of a lower bound

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

Let $\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}}$ be any distribution on $\boldsymbol{Z}$

Central identity

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

Note that:

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

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

Let $\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}}$ be any distribution on $\boldsymbol{Z}$

Central identity

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

Note that:

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

Variational EM : maximization of a lower bound

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

Let $\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}}$ be any distribution on $\boldsymbol{Z}$

Central identity

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

Note that:

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

Variational EM

- Maximization of $\log \ell(\mathbf{Y}; \theta)$ w.r.t. θ replaced by maximization of the lower bound $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}})$ w.r.t. τ and θ.
- **Benefit** : we choose $\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}}$ such that the maximization calculus can be done explicitly
  - In our case: mean field approximation : neglect dependencies between the (Z_i)

$$P_{\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}}}(Z_i = q) = \tau_{iq}$$

Variational EM

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

- **Step 1** Maximization w.r.t. τ

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

Variational EM

Algorithm

- **Step 2** Maximization w.r.t. θ

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

In practice

- Really fast
- Strongly depends on the initial values

A penalized likelihood criterion

- Selection of the number of clusters Q
- Integrated Classification Likelihood (ICL) [Biernacki et al., 2000]

$$ICL(\mathcal{M}_Q) = \log \ell_c(\mathbf{Y}, \hat{\mathbf{Z}}; \hat{\theta}_Q) - \text{Pen}(\mathcal{M}_Q)$$

  where

$$\hat{Z}_i = \underset{q \in \{1,\dots,Q\}}{\arg\max}\ \hat{\tau}_{iq}.$$

-

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

  where

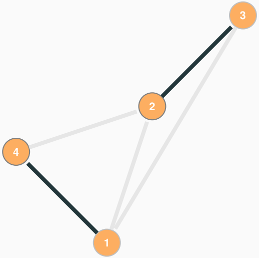
$$\text{Pen}(\mathcal{M}_Q) = \frac{1}{2}\left\{\underbrace{(Q-1)\log(n)}_{\text{Clust.}} + \underbrace{Q^2\log(n^2-n)}_{\text{Conn.}}\right\}$$

Penalty for bipartite networks

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

Recall on missing value

Data: a graph G with missing data.



Adjacency matrix:

$$A = \begin{pmatrix} 0 & \mathtt{NA} & \mathtt{NA} & 1 \\ \mathtt{NA} & 0 & 1 & \mathtt{NA} \\ \mathtt{NA} & 1 & 0 & 0 \\ 1 & \mathtt{NA} & 0 & 0 \end{pmatrix}$$

Goal: Cluster nodes in spite of missing data and predict NA to $\{0, 1\}$ or predict most likely existing links.

Inferring the SBM from an observed network (Missing data)

[Timothée Tabouy and Chiquet, 2020].

Observation of a network: $n \times n$ binary matrix $\mathbf{R}$ such that $R_{ij} = 1$ if Y_{ij} is observed, $R_{ij} = 0$ otherwise ($Y_{ij} = \texttt{NA}$).

Observation process: [Rub76] MCAR, MAR or NMAR?

(MCAR)

(MAR or NMAR)

(NMAR)

(NMAR)

Inference under M(C)AR scheme: Likelihood on the observed data.

Inference under NMAR scheme

Need for accounting for the complete likelihood where we have missing data ($\mathbf{Y}^{\mathrm{m}}$) and latent variables $\boldsymbol{Z}$

Variational distribution on $(\mathbf{Y}^{\mathrm{m}}, \boldsymbol{Z})$ in the VEM algorithm:

$$\mathcal{R}_{(\mathbf{Y}^{\mathrm{m}},\mathbf{Z})} = \mathcal{R}_{(\mathbf{Y}^{\mathrm{m}})} \cdot \mathcal{R}_{(\mathbf{Z})} = \prod_{(i,j),Y_{ij}=NA} \nu_{ij}^{Y_{ij}} (1-\nu_{ij})^{1-Y_{ij}} \cdot \prod_{i=1}^{n} \prod_{k=1}^{Q} (\tau_{ik})^{\mathbb{I}_{Z_i=k}},$$

where

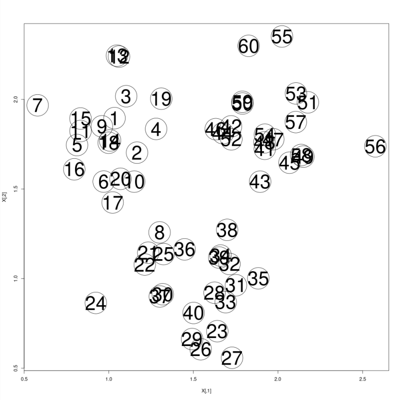
- ν_{ij}s and τ_{ik}s parameters to be optimized in the VE step,
- τ_{ik} is almost generic,
- ν_{ij} is specific to the sampling design.

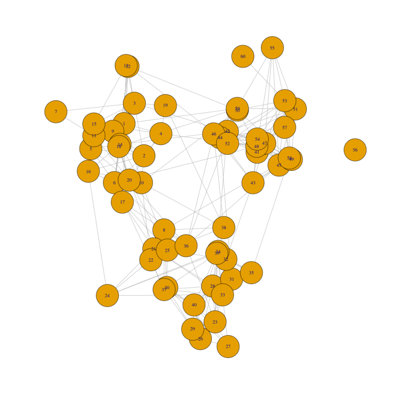
Contributions:

- Derived variational steps for some NMAR sampling schemes,
- Importance of accounting for sampling illustrated on synthetic and real data,
- Implementation in an `R` package `missSBM` [Tabouy et al., 2019].

Outline

- $\forall i \in \{1, \ldots, N\}$, $Z_i \overset{ind}{\sim}$ Mixture$\mathcal{N}((\mu_k)_k, (\Sigma_k)_k)$,
- $\forall (i,j)$,$Y_{ij}|Z_i, Z_j \overset{ind}{\sim} b(\exp(-\|Z_i - Z_j\|/\sigma^2))$.





(Generalised) random dot product graph

Alternative to the distance between latent positions, the dot product can be used:

$$\forall (i,j),\ Y_{ij}|Z_i, Z_j \overset{ind}{\sim} b(Z_i \cdot Z_j = Z_i^\top Z_j).$$

[Rubin-Delanchy et al., 2022] proposed a generalisation:

$$\forall (i,j),\ Y_{ij}|Z_i, Z_j \overset{ind}{\sim} b(Z_i I_{p,q} Z_j)$$

with

$$I_{p,q} \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}.$$

Modularity

In [Newman, 2006], definition of modularity, for a given clustering:

$$Mod = \frac{1}{C} \sum_{i,j} \left[A_{ij} - \frac{d_i d_j}{C} \right] \delta_{ij}$$

where

- $C = \sum_i \sum_j A_{ij}$,
- d_i is the degree of species i (i.e. $d_i = \sum_j A_{ij}$),
- δ_{ij} are dummy variables indicating whether species/nodes i and j are assumed to belong to the same module/community/cluster.

Goal: look for the partitioning of nodes into communities/modules that maximizes the associated modularity score.

Algos: edge-betweenness algorithm (EB), the leading-eigenvector algorithm (LE) and the Louvain algorithm (ML).

Laplacian of a graph

For $G = (V, E)$ an undirected graph s.t. $V = \{1, \ldots, N\}$ and A the corresponding adjacency matrix.

- Degree of a vertex/node: $d_i = \sum_j A_{ij}$,
- Unnormalized Laplacian: $L = D - A$ with $D = \mathrm{diag}(d_1, \ldots, d_N)$,

Properties:

- for $x \in \mathbb{R}^n,\ x^\top L x = \frac{1}{2} \sum_j^N A_{ij}(x_i - x_j)^2$,
- L is symmetric and positive definite,
- the smallest eigenvalue is 0 and associated with the vector $\mathbb{1}$,
- the order of multiplicity of 0 is the number of connected components.

[Von Luxburg, 2007]

Normalized Laplacians:

$$
\begin{aligned}
L_{sym} &= D^{-1/2} L D^{-1/2} = I_N - D^{-1/2} A D^{-1/2} \\
L_{rw} &= D^{-1} L = I_N - D^{-1} A
\end{aligned}
$$

Properties:

- for $x \in \mathbb{R}^n,\ x^\top L_{sym} x = \frac{1}{2} \sum_j^N A_{ij} (x_i/\sqrt{d_i} - x_j/\sqrt{d_j})^2$,
- L_{sym} and L_{rw} are symmetric and positive definite,
- the smallest eigenvalue is 0,
- the order of multiplicity of 0 is the number of connected components.

Spectral Clustering:

Input: Adjacency Matrix $A \in \mathbb{R}^{N \times N}$, number k of clusters to construct.

- Compute the unnormalized Laplacian L.
- Compute the first k eigenvectors $u_1, \dots, u_k$ of L.
- Let $U \in \mathbb{R}^{N \times k}$ be the matrix containing the vectors $u_1, \dots, u_k$ as columns.
- For $i = 1, \dots, N$, let $z_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U.
- Cluster the points $(z_i)_{i=1,\dots,N}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \dots, C_k$.

Outline

Stochastic and Latent Block Models

Other latent space models and other methods

Extensions of SBM

Degree corrected SBM

[Karrer and Newman, 2011]

$$\mathbb{P}(Y_{i,j} = 1 | Z_i = k, Z_j = l) = 1/\big(1 + \exp(-\alpha_{kl} - \nu_i - \nu_j)\big)$$

with

- $(\nu_i)_{1 \leq i \leq n}$ parameters to be estimated,
- find clusters besides the degree/popularity.

Mixed membership SBM

[Airoldi et al., 2008]

For each node $p \in N$:

- Draw a K-dimensional mixed membership vector $\vec{\pi}_p \sim \text{Dirichlet}(\vec{\alpha})$.

For each pair of nodes $(p, q) \in N \times N$:

- Draw membership indicator for the initiator, $\vec{z}_{p \to q} \sim \text{Multinomial}(\vec{\pi}_p)$.
- Draw membership indicator for the receiver, $\vec{z}_{q \to p} \sim \text{Multinomial}(\vec{\pi}_q)$.
- Sample the value of their interaction, $Y(p, q) \sim \text{Bernoulli}(\vec{z}^T_{p \to q} A \vec{z}_{q \to p})$.

Other extensions exist...

Multilayer networks

- Large variety of multilayer networks,
- SBM as a probabilistic generative model easy to extend to numerous cases.
- e.g. dynamic or spatial SBM ([Matias and Miele, 2017, Longepierre and Matias, 2019]) or Topic SBM [Bouveyron et al., 2018].

Our contributions

- Multiplex network [Barbillon et al., 2017, Lazega et al., 2016],
- Multilevel network [Chabert-Liddell et al., 2019],
- Multipartite network [Bar-Hen et al., 2018].

Adaptation of the VEM algorithm and ICL criterion to select the numbers of blocks.

Multiplex network

In collaboration with A. Bar-Hen, S. Donnet and E. Lazega [Barbillon et al., 2017, Lazega et al., 2016].

Multiple relations between individuals:

$$\mathbf{Y}_{ij}|Z_i, Z_j \overset{ind}{\sim} \text{Bern}^M((\alpha^w_{Z_i,Z_j})_w) \quad \text{with} \sum_w \alpha^w_{Z_i,Z_j} = 1.$$

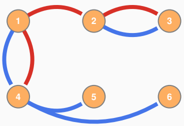


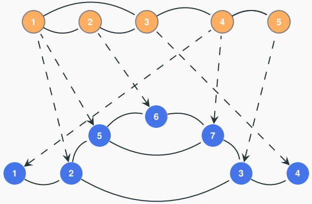
Figure 1: Illustration of a multiplex network. For each dyad, two kinds of link may exist. They are respectively displayed by red and blue edges.

Model inference implemented in the `R` package: `sbm`.

Application to a network of French researchers in cancerology (advice relation and indirect relation through the labs of the researchers).

Multilevel network

In collaboration with S. Donnet and S.-C. Chabert-Liddell (Ph.D. Thesis) and E. Lazega [Chabert-Liddell et al., 2019].



- Organizational level: SBM for $(\mathbf{Y}^O, \mathbf{Z}^O)$,
- Individual level: SBM for $(\mathbf{Y}^I, \mathbf{Z}^I)$,
- Interlevel dependence $i \in \{1, \ldots, n_I\}$, $k \in \{1, \ldots, K_I\}$,
  $\mathbb{P}(Z_i^I = k | Z_j^O, A_{ij} = 1) \overset{ind}{=} \gamma_{kZ_j^O}$, where A is the affiliation matrix.

Implemented in S.-C. Chabert-Liddell's `R` package: `MLVSBM`.

Application to a dataset of a program trade fair (Organizations = audiovisual firms, individuals = sales representatives).

Generalized Multipartite Networks

In collaboration with A. Bar-Hen and S. Donnet [Bar-Hen et al., 2018].

- Pre-specified functional groups (colors of nodes),
- Looking for blocks within functional groups,
- Each network between 2 functional groups is either an SBM or an LBM.

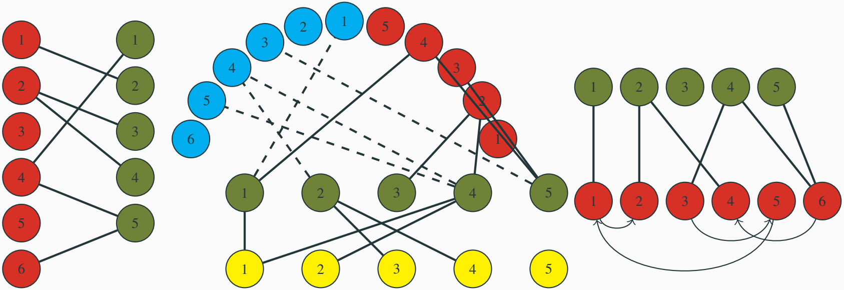
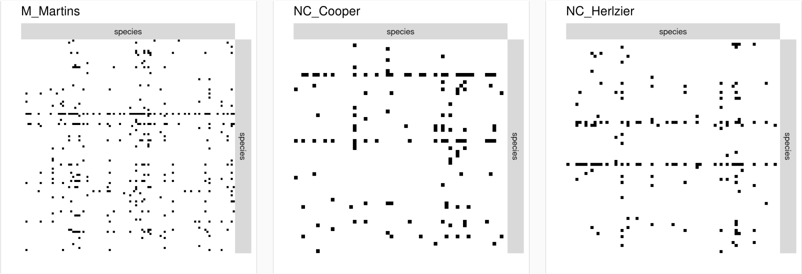


Figure 2: Illustrations of bipartite (left), multipartite (center) and generalized multipartite networks (right). The colors stand for the different functional groups.

Implemented in an `R` package: `GREMLINS`.
Application for ecological interaction (see practical session).

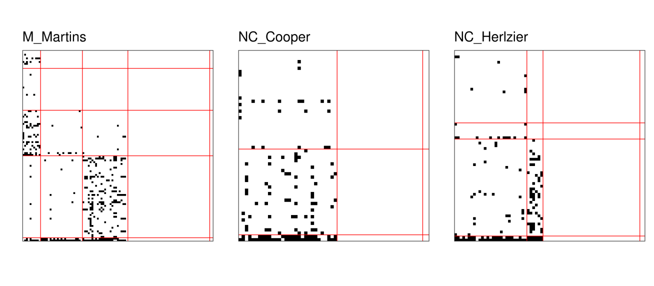
Towards collection of networks: Three foodwebs

- Pine-firest stream food webs issued from Maine, North-Caroline and New-Zealand [Thompson and Townsend, 2003]
- Involve respectively 105, 58 and 71 species.
- $Y_{ij} = 1$ if i is eaten by j. Directed relation



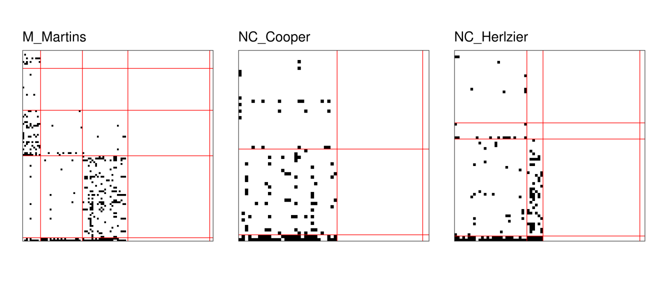
- Look for similarities and differences between network structures.

Separate SBMs



- Fitted SBM on each separately
- Reordered the matrices following the blocks
- Label the blocks following the average out-degrees order

Towards collection of networks: Separate SBMs



- Two bottom groups in each matrix are basal species : eaten by many species and not eating anybody.
-
    - **Martins**: has a separation into 5 blocks, the third one is a medium trophic level, which preys on basal species and is highly preyed by species of the 1st block.
    - **Cooper**. Higher trophic levels grouped together in the same block (lack of statistical power).
    - **Herlzier**: higher trophic level is separated into 2 blocks determined on how much they prey on the less preyed basal block.

Collection of networks: Joint modeling of the networks

- Need to model jointly the networks
- Identify the groups playing the same role through out the networks, with an unsupervised strategy.
- Let $(\mathbf{Y}^m)_{m=1,\dots,M}$ denote the collection of networks each involving n_m nodes.
- $(\mathbf{Y}^m)$ independent.
- $$\mathbf{Y}^m \sim \text{SBM}_{n_m}(Q^m, \boldsymbol{\pi}^m, \boldsymbol{\alpha}^m)$$
- Conditions on the parameters $(\boldsymbol{\pi}^m)_{m=1,\dots,M}$ and $(\boldsymbol{\alpha}^m)_{m=1,\dots,M}$

First naive model

iid-colSBM

$$\mathbf{Y}^m \sim \mathsf{SBM}_{n_m}(Q, \boldsymbol{\pi}, \boldsymbol{\alpha})$$

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

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

Same structure of connection α, specific proportions of blocks in each network

π-colSBM

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

On the block proportions

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

π-colSBM: different proportions

$M = 2$ networks

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

- Same connection structure between blocks
- Different block proportions
- $2 \times (3-1) + 3^2 = 15$ parameters.

π-colSBM: nested structures

$\pi_q^m \geq 0$

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{12} & \alpha_{22} & \alpha_{23} \\ \alpha_{13} & \alpha_{23} & \alpha_{33} \end{pmatrix} \qquad \begin{array}{c} \pi^1 = [.25, .25, .50] \\ \pi^2 = [.40,\ 0\ , .60] \end{array}.$$

- Blocks 1 and 3 are represented in the two networks while block 2 only exists in network 1.
- $3 - 1 + 3 - 2 + 3^2 = 14$ parameters

π-colSBM: partially nested structures

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \cdot \\ \alpha_{31} & \cdot & \alpha_{33} \end{pmatrix} \qquad \begin{matrix} \pi^1 = [.25, .75, \ 0 \ ] \\ \pi^2 = [.40, \ 0 \ , .60] \end{matrix}.$$

- The two networks share block 1 (for instance super predators or basal species)
- The remaining nodes of each network not equivalent in terms of connectivity.
- Blocks 2 and 3 never interact because their elements do not belong to the same network and so α_{23} and α_{32} are not required to define the model.
- $(2-1)+(2-1)+7 = 11$ parameters.

Summary

M independent networks.

$$\mathbf{Y}^m \sim \text{SBM}(Q^m, \boldsymbol{\pi}^m, \boldsymbol{\alpha}^m)$$

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


colSBM
πcolSBM
δcolSBM
δπcolSBM

Inference and model selection

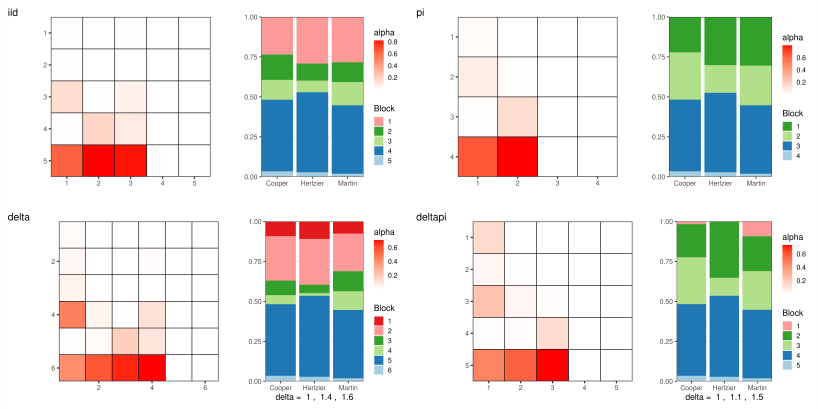
VEM algorithm

- Direct extension of VEM previously described for iid-colSBM and π-colSBM
- Less obvious with $\delta_m\alpha$: M step not explicit.
- Sensitive to initializations: need to match blocks among networks.

Model Selection

- ICL can be directly extended for iid-colSBM and the δ-colSBM
- for $\pi(\delta)$-colSBM, taking into account empty blocks...

Our 4 consensus models



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

- separated SBMs gives an ICL of -2080.
- iid-colSBM : prefered model. Make 5 blocks
- π-colSBM: block proportion quite similar. Make no use of its flexibility

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008).
Mixed membership stochastic blockmodels.
Journal of machine learning research, 9(Sep):1981–2014.

Bar-Hen, A., Barbillon, P., and Donnet, S. (2018).
Block models for multipartite networks. applications in ecology and ethnobiology.

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.

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.

Bouveyron, C., Latouche, P., and Zreik, R. (2018).

The stochastic topic block model for the clustering of vertices in networks with textual edges.
Statistics and Computing, 28(1):11–31.

Chabert-Liddell, S.-C., Barbillon, P., Donnet, S., and Lazega, E. (2019).
A stochastic block model for multilevel networks: Application to the sociology of organisations.
arXiv preprint arXiv:1910.10512.

Erdős, P. and Rényi, A. (1959).
On random graphs, I.
Publicationes Mathematicae Debrecen, 6:290–297.

Govaert, G. and Nadif, M. (2010).
Latent block model for contingency table.
Communications in Statistics—Theory and Methods, 39(3):416–425.

Karrer, B. and Newman, M. E. (2011).
Stochastic blockmodels and community structure in networks.
Physical review E, 83(1):016107.

Lazega, E., Bar-Hen, A., Barbillon, P., and Donnet, S. (2016).
Effects of competition on collective learning in advice networks.

Social Networks, 47:1–14.

Longepierre, L. and Matias, C. (2019).
Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model.
Electronic Journal of Statistics, 13(2):4157–4223.

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.

Newman, M. E. (2006).
Modularity and community structure in networks.
Proceedings of the national academy of sciences, 103(23):8577–8582.

Nowicki, K. and Snijders, T. A. B. (2001).
Estimation and prediction for stochastic blockstructures.
Journal of the American Statistical Association, 96(455):1077–1087.

Rubin-Delanchy, P., Cape, J., Tang, M., and Priebe, C. E. (2022).

A statistical interpretation of spectral embedding: the generalised random dot product graph.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(4):1446–1473.

Tabouy, T., Barbillon, P., and Chiquet, J. (2019).
misssbm: An r package for handling missing values in the stochastic block model.

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.

Timothée Tabouy, P. B. and Chiquet, J. (2020).
Variational inference for stochastic block models from sampled data.
Journal of the American Statistical Association, 115(529):455–466.

Von Luxburg, U. (2007).
A tutorial on spectral clustering.
Statistics and computing, 17:395–416.