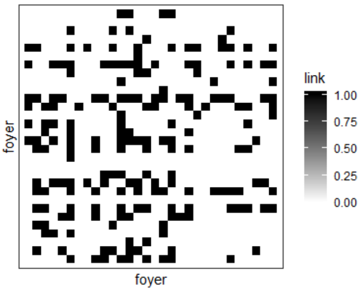
Données de réseaux : modèles probabilistes

Sophie Donnet, François Massol, Nicolas Verzelen
MIA Paris, INRA

Formation Réseaux MIRES / ReSodiv
18-19-06/2019

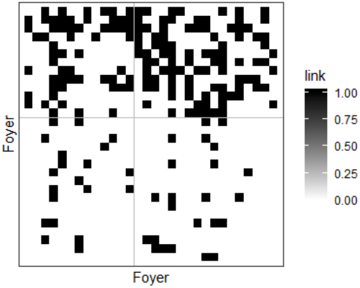
Qu'attendre des SBM ?

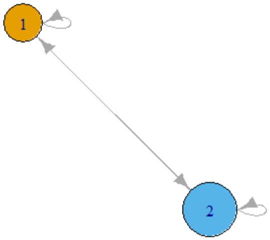
Outils automatiques pour faire des groupes de sommets qui ont le même rôle dans le réseau



Qu'attendre des SBM ?

Outils automatiques pour faire des groupes de sommets qui ont le même rôle dans le réseau





A propos de l'aspect "probabiliste"

- *Hypothèse* : réseau observé Y_{ij} réalisation d'un phénomène aléatoire
- *Modèle* : forme particulière de ce phénomène
- *Aléa* : représente le fait que l'observateur n'est pas capable de prédire Y_{ij}
  - Exemple de l'expérience du Pile ou Face
  - Si je connaissais tous les paramètres physiques de la pièce et du lancer, je serais capable de prédire l'issue du jeu
  - En tant qu'observateur, je ne peux pas prédire : expérience aléatoire

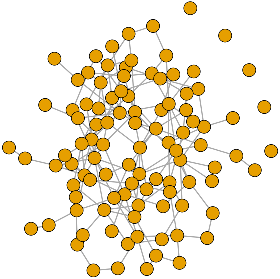
A first random graph model for network : null model

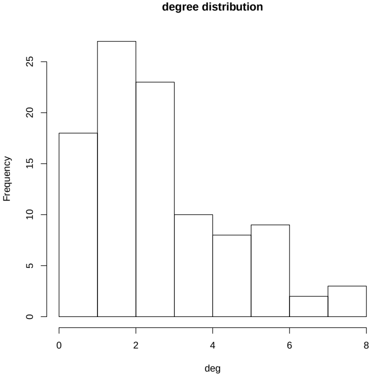
- Erdős-Rényi (1959) Model for n nodes
- Let $(y_{ij})_{i,j=1\dots n}$ be an adjacency matrix (i.e. representing a simple network) with $y_{ij} \in \{0,1\}$
- ER assumes that y_{ij} is the realisation of :

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

where *Bern* is the Bernoulli distribution and $p \in [0,1]$ a probability for a link to exist.

A first random graph model for network : null model





Limitations of an ER graph to describe real networks

- Degree distribution too concentrated, no high degree nodes,
- All nodes are equivalent (no nestedness...),
- No modularity.

Stochastic Block Model (SBM)

Latent block model (LBM)

Stochastic Block Model

Nowicki, & Snijders (2001)
Let (y_{ij}) be an adjacency matrix such that $y_{ij} \in \{0,1\}$. y_{ij} is the realisation of the following processus.

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

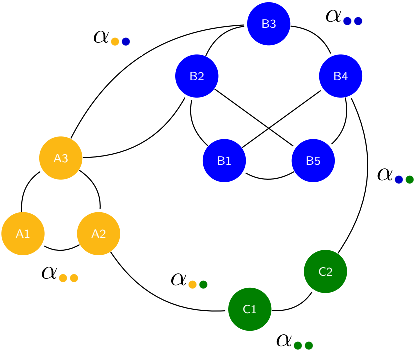
$$\mathbb{P}(Z_i = k) = \pi_k$$

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

(Y_{ij}) independant and

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

Stochastic Block Model : illustration



Parameters

Let n nodes divided into 3 clusters

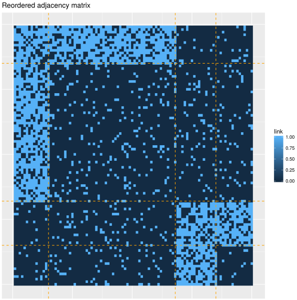
- $\mathcal{K} = \{{\color{orange}\bullet}, {\color{blue}\bullet}, {\color{green}\bullet}\}$ clusters
- $\pi_\bullet = \mathbb{P}(i \in \bullet), \bullet \in \mathcal{K}, i = 1, \dots, n$
- $\alpha_{{\color{orange}\bullet}{\color{blue}\bullet}} = \mathbb{P}(i \leftrightarrow j | i \in {\color{orange}\bullet},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

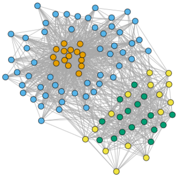
$$
\begin{aligned}
Z_i = \mathbf{1}_{\{i \in \bullet\}} &\sim^{\text{iid}} \mathcal{M}(1, \pi), \quad \forall \bullet \in \mathcal{K}, \\
Y_{ij} \mid \{i \in {\color{orange}\bullet}, j \in {\color{blue}\bullet}\} &\sim^{\text{ind}} \mathcal{B}(\alpha_{{\color{orange}\bullet}{\color{blue}\bullet}})
\end{aligned}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

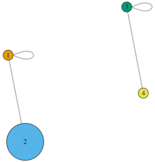
SBM : A great generative model

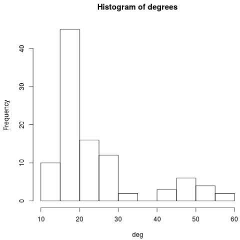
- Generative model : easy to simulate
- No a priori on the type of structure
- Combination of modularity, nestedness, etc...

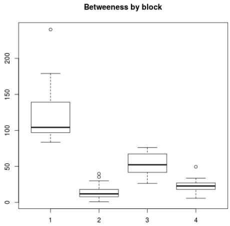
Networks with hubs generated by SBM



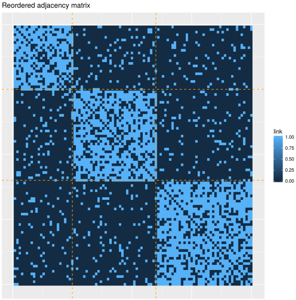


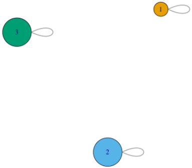


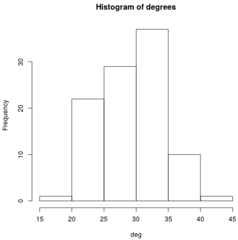




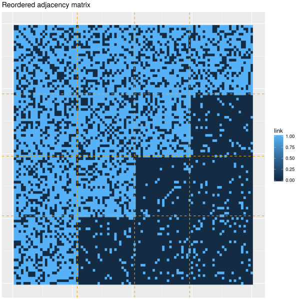
Community network generated by SBM

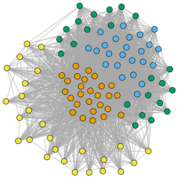


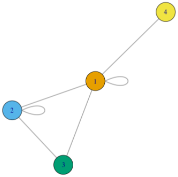


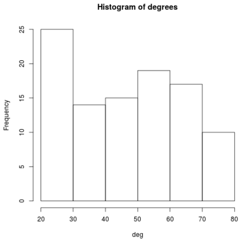


Nestedness generated by SBM

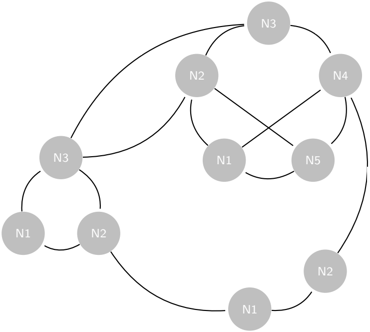








Inférence statistique



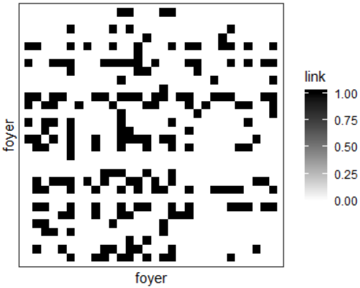
Inférence

- $\mathcal{K} = \{\bullet, \bullet, \bullet\}$, card($\mathcal{K}$) known
  - $\pi_\bullet = ?$,
  - $\alpha_{\bullet\bullet} = ?$
  - $Z_1, \dots, Z_n = ?$
- card($\mathcal{K}$) ?

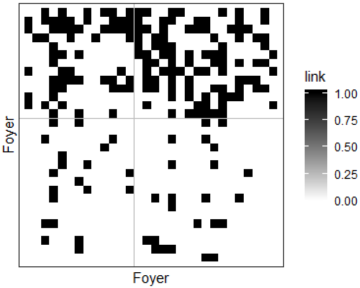
Nowicki, & Snijders (2001), Daudin et al. (2008)

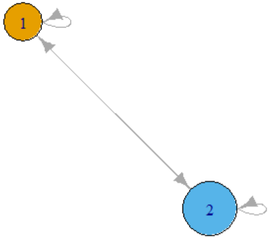
R package : blockmodels.

Exemple du Vanuatu



Exemple du Vanuatu





Stochastic Block Model (SBM)

Latent block model (LBM)

Probabilistic model for binary bipartite networks

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

Latent variables : bi-clustering

- Nodes $i = 1, \dots, n_1$ partitionned into K_1 clusters, nodes $j = 1, \dots, n_2$ partitionned into K_2 clusters
- $$\begin{array}{ll} Z^1_i = k & \text{if node } i \text{ belongs to cluster (block) } k \\ Z^2_j = \ell & \text{if node } j \text{ belongs to cluster (block) } \ell \end{array}$$
- Z^1_i, Z^2_j independent variables

$$\mathbb{P}(Z^1_i = k) = \pi^1_k, \quad \mathbb{P}(Z^2_j = \ell) = \pi^2_\ell$$

Probabilistic model for binary bipartite networks

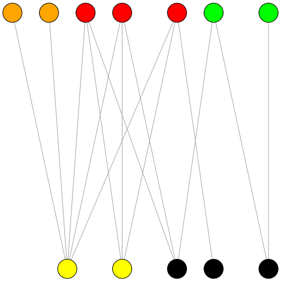
Conditionally to $(Z^1_i)_{i=1,\dots,n_1}, (Z^2_j)_{j=1,\dots,n_2} \cdots$

(Y_{ij}) independent and

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

Govaert & Nadif (2008)

Latent Block Model : illustration



Latent Block Model

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

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

Govaert & Nadif (2008) and R package : blockmodels as well.

Valued-edge networks

Values-edges networks

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

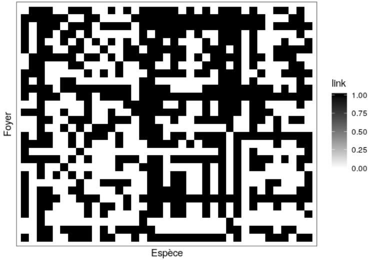
1. a count of the number of observed interactions,
2. a quantity interpreted as the interaction strength,

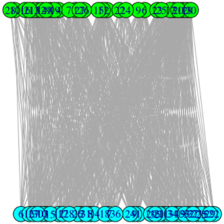
Natural extensions of SBM and LBM

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,

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

Exemple sur Vanuatu Foyers/espèces cultivées





Exemple sur Vanuatu Foyers/espèces cultivées