Analyse de données d'échange

Les modèles à blocs stochastiques

Vanesse Labeyrie & Sarah Ouadah

Formation Analyse de Réseaux

11-15 Juin 2019

Objectifs

- Présentation de quelques modèles de graphes aléatoires.
  On se demandera s'ils miment les propriétés de réseaux observés.
- Focus sur le modèle à blocs stochastiques [SBM] qui suppose que les liens entre individus découlent de leur appartenance à un groupe.
  Comment le mettre en oeuvre et l'interpréter ?
- Quelques références et packages R sur les extensions du SBM.
- Focus sur le modèle à blocs latents [LBM] pour les graphes bipartites.

Sommaire

Graphe aléatoire

Un graphe aléatoire $\mathcal{G} = (\mathcal{V} = \{1, \dots n\}, \mathcal{E})$ est la représentation mathématique d'un réseau d'interaction.

Variable d'intérêt - Données : $\mathbf{Y}$ la matrice d'adjacence de $\mathcal{G}$

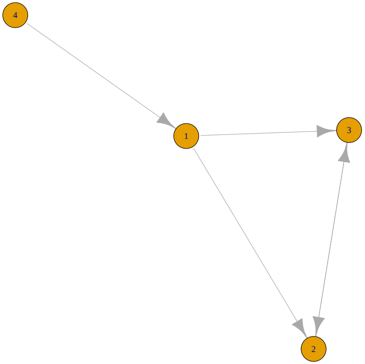
$$Y_{ij} = \begin{cases} 1 \text{ ou autre valeur} & \text{si } (i,j) \in \mathcal{E} \text{ (arête)} \\ 0 & \text{sinon} \end{cases}$$

et $Y_{ii} = 0, \forall i$. *Lorsque le graphe est non dirigé* $Y_{ij} = Y_{ji}, \forall i \neq j$.

Les Y_{ij} sont des variables aléatoires. Leur réalisations, i.e. les valeurs que l'on observe, se réalisent donc avec une certaine probabilité et proviennent d'un échantillon de la population.

Matrice d'adjacence

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



Modèle d'Erdös-Rényi

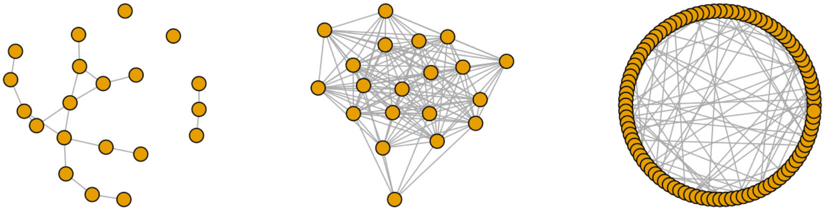
Modèle d'Erdös-Rényi (Erdös et Rényi, 1959)

$$Y_{ij} \text{ i.i.d.} \sim \mathcal{B}(p)$$

Tous les nœuds ont même probabilité de connexion

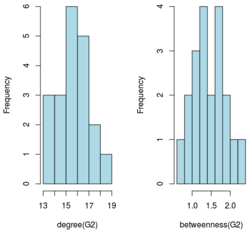
Erdös-Rényi – Exemple (1)

```
G1 <- sample_gnp(20, 0.1)
G2 <- sample_gnp(20, 0.8)
G3 <- sample_gnp(100, .02)
```



Erdös-Rény – Caractéristiques

```
> hist(degree(G2)); hist(betweenness(G2))
```



Les distributions des degrés et de la betweenness sont assez homogènes

```
> average.path.length(G2); diameter(G2)
[1] 1.152632
[1] 2
```

La moyenne et le maximum de la longueur du plus court chemin se constituent de très peu de nœuds. La modularité est quasi nulle :

```
> modularity(G2.clustering)
[1] 0.03841326
```

Extensions du modèle d'Erdös-Rényi

Modèle d'Erdös-Rényi hérérogène

$$Y_{ij} \text{ ind.} \sim \mathcal{B}(p_{ij})$$

Chaque paire de nœuds a sa propre probabilité de connexion

Modèle linéaire généralisé

$$\begin{cases} Y_{ij} \text{ ind.} \sim \mathcal{B}(p_{ij}) \\ logit(p_{ij}) = x_{ij}^{\intercal}\beta + \alpha \end{cases}$$

où x_{ij} est le vecteur de covariables sur l'arête (i, j).

Chaque paire de nœuds a sa propre probabilité de connexion qui dépend de covariables, e.g. différence d'âge

Modèle d'attachement préférentiel

Modèle d'attachement préférentiel (Barabàsi et Albert, 1999)

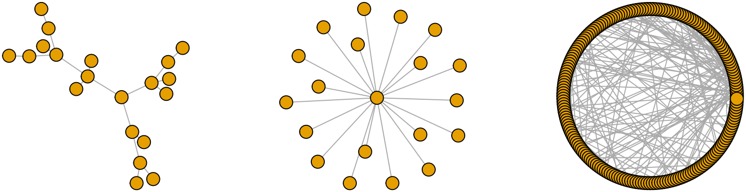
Le graphe se construit ainsi à partir d'un graphe initial $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$:

1. au temps t, on ajoute un nouveau nœud V_t
2. V_t est connecté à $i \in V_{t-1}$ avec probabilité $D_i^\alpha + \mathrm{constante}$, où $D_i = \sum_{j \neq i} Y_{ij}$ est le degré du nœud i

Les nœuds qui ont un fort degré ont de grandes chances d'être connectés : les riches s'enrichissent.

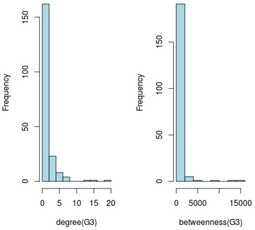
Modèle d'attachement préférentiel – Exemple

```
G1 <- sample_pa(20, 1)
G2 <- sample_pa(20, 5)
G3 <- sample_pa(200)
```



Modèle d'attachement préférentiel – Caractéristiques

```
> hist(degree(G3)); hist(betweenness(G3))
```



Les distributions des degrés et de la betweenness sont hétérogènes et caractéristiques d'une loi de puissance

```
> average.path.length(G3); diameter(G3)
[1] 6.704372
[1] 17
```

La moyenne et le maximum de la longueur du plus court chemin se constituent de relativement peu de nœuds ($n = 200$). Aucun triangle ne se forme :

```
> transitivity(G3)
[1] 0
```

Modèle exponentiel de graphe [ERGM]

Modèle exponentiel de graphe [ERGM] (review de Wasserman et Pattison, 1996)

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y}) = \left(\frac{1}{\kappa}\right) exp\left(\sum_H \theta_H g_H(\mathbf{y})\right)$$

avec

- $\mathbf{y}$ une réalisation de $\mathbf{Y}$
- H une configuration/motif, e.g. arête, triangle, étoile, etc.
- $g_H(y)$ le nombre de fois où cette configuration apparaît dans $\mathbf{y}$
- θ_H le coefficient de dépendance
- κ la constante de normalisation

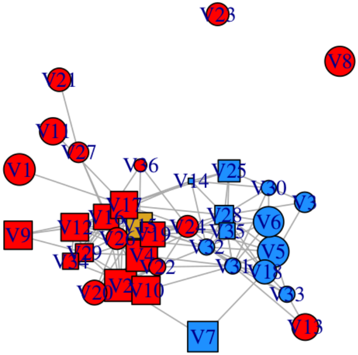
La distribution des arêtes est due à la présence de différents motifs dans le réseau observé. On peut également ajouter dans le modèle des attributs sur les nœuds et les arêtes.

ERGM – Exemple

Lazega est un réseau de collaboration entre 36 avocats appartenant à différents cabinets et compte 115 liens non dirigés

```
my.ergm <- formula(lazega ~ edges + kstar(2) + kstar(3) + triangle)
```

```
> ergm::summary.statistics(my.ergm)
   edges    kstar2    kstar3 triangle
     115       926      2681      120
```



Limites

- Modèle d'Erdös-Rényi
  - modélisation d'une structure homogène, pas de degré fort, ni de modularité
  - peu adapté aux réseaux observés
- Modèle d'attachement préférentiel
  - modélisation d'une structure où la distribution des mesures de centralité est une loi de puissance, i.e. existence d'un petit groupe de nœuds centraux, une transitivité nulle
  - non propice à un cadre d'inférence statistique (mécanistique)
- Modèle ERGM
  - modélisation de structures très particulières et de petites tailles
  - justifications théoriques (pendant du glm) non établies

Modèle à blocs stochastiques

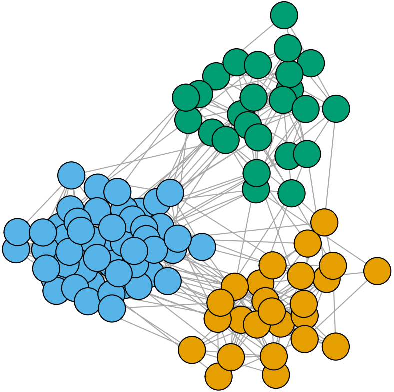
- modélisation de réseaux structurés en groupes : situation courante des réseaux réels
- propice à l'inférence statistique : estimation des interactions et de la composition des groupes

Sommaire

SBM – Exemple de topologie (1)

Réseau de communauté

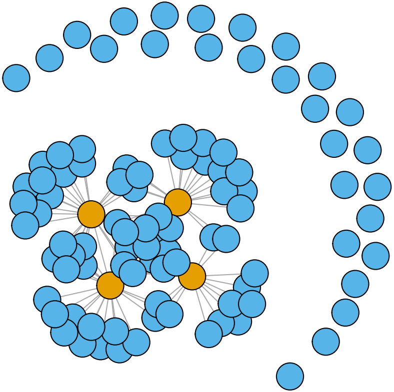
```
pi <- matrix(c(0.3,0.02,0.02,0.02,0.3,0.02,0.02,0.02,0.3),3,3)
communities <- sample_sbm(100, pi, c(25, 50, 25))
```



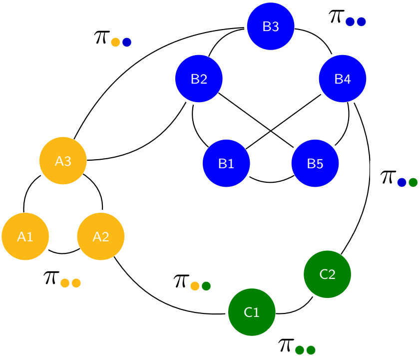
SBM – Exemple de topologie (2)

Réseau en étoiles (hubs)

```
pi <- matrix(c(0.05,0.3,0.3,0),2,2)
star <- sample_sbm(100, pi, c(4, 96))
```



Modèle à blocs stochastiques [SBM] (1)



SBM (Nowicki et Snijders, 2001)

Soient n nœuds répartis ainsi :

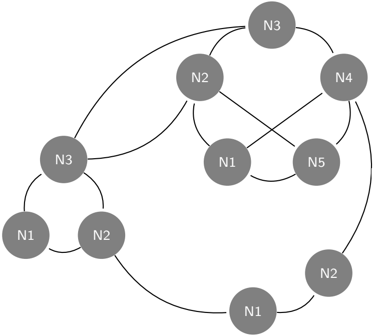
- $\mathcal{Q} = \{\bullet, \bullet, \bullet\}$ groupes
- $\pi_\bullet = \mathbb{P}(i \in \bullet)$ $\bullet \in \mathcal{Q}, i = 1, \dots, n$
- $\alpha_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \text{ i.i.d.} \sim \mathcal{M}(1, \pi), \quad \forall \bullet \in \mathcal{Q}$$

$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \text{i.i.d} \sim \mathcal{B}(\alpha_{\bullet\bullet})$$

Toute paire de nœuds a une probabilité de connexion induite par un caractère spécifique à chacun des nœuds : le groupe d'appartenance

SBM (2)



SBM

Soient n nœuds répartis ainsi :

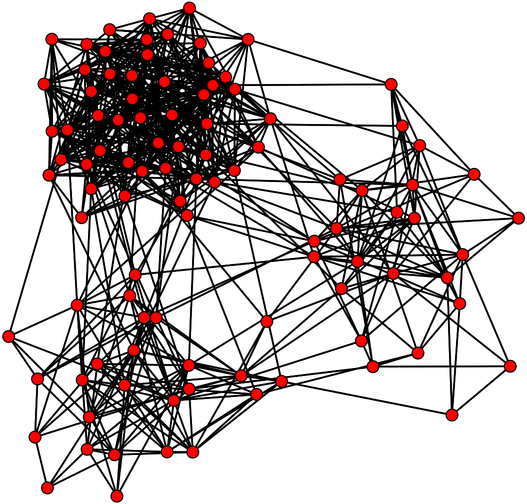
- $\mathcal{Q} = \{\textcolor{orange}{\bullet}, \textcolor{blue}{\bullet}, \textcolor{green}{\bullet}\}$, card$(\mathcal{Q})$ connu
- $\pi_{\bullet} = ?$,
- $\alpha_{\textcolor{orange}{\bullet}\textcolor{blue}{\bullet}} = ?$

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \text{ i.i.d.} \sim \mathcal{M}(1, \pi), \quad \forall \bullet \in \mathcal{Q},$$
$$Y_{ij} \mid \{i \in \textcolor{orange}{\bullet}, j \in \textcolor{blue}{\bullet}\} \text{ i.i.d.} \sim \mathcal{B}(\alpha_{\textcolor{orange}{\bullet}\textcolor{blue}{\bullet}})$$

SBM – Estimation – Sélection de modèle

- Constitution des groupes : estimation de π le vecteur des probabilités d'appartenance aux Q groupes
  via un algorithme EM variationnel
- Interactions : estimation de α la matrice des probabilités de connexion au sein des groupes et entre les groupes
  via ce même vEM
- Nombre de groupes : estimation de Q le nombre de groupes
  via la maximisation du critère vICL

SBM – Réseau de communautés $n = 100$, $\rho = 0.12$



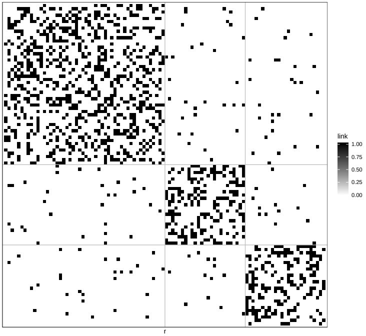
SBM – Communautés – Package **blockmodels** (1)

```
# appel du package
> library(blockmodels)

# définition de l'objet
> communities.sbm <- BM_bernoulli("SBM_sym",communities_adjacency)

# méthode d'inférence
> communities.sbm$estimate()

# nombre de groupes sélectionné avec vICL
>  which.max(communities.sbm$ICL)
[1] 3
```

Le critère de sélection de modèle vICL retrouve le nombre de groupes égal à 3

SBM – Communautés – Package **blockmodels** (2)

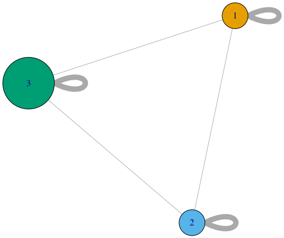
```
# extraction des paramètres estimés
> paramEstimSBM <- extractParamBM(communities.sbm,Q)

# appartenance des noeuds aux groupes
> paramEstimSBM$Z
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 ...
 [56] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 ...
```



SBM – Communautés – Package **blockmodels** (3)

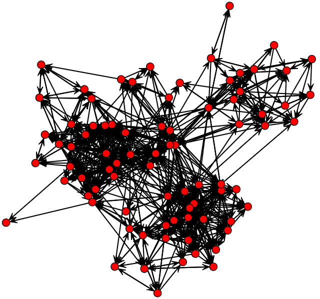
```
> paramEstimSBM$pi
[1] 0.4995005 0.2502353 0.2502642

> paramEstimSBM$alpha
           [,1]       [,2]       [,3]
[1,] 0.30473086 0.02469296 0.02309172
[2,] 0.02469296 0.32480004 0.02138467
[3,] 0.02309172 0.02138467 0.31149209
```

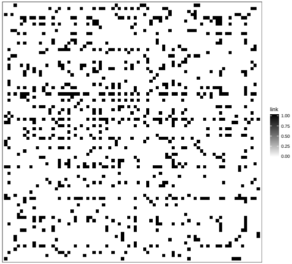
*On retrouve bien les probabilités d'appartenance des nœuds aux groupes (*0.5, 0.25 *et* 0.25*), ainsi que les probabilités de connexion intra et inter groupes (*0.3 *et* 0.02*).*



SBM – UKfaculty $n = 81$, $\rho = 0.13$

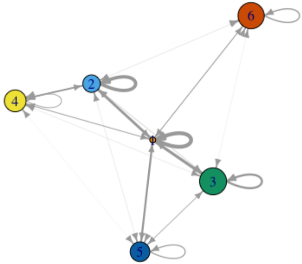
UKfaculty est un réseau d'amitiés entre 81 individus appartenant à différentes "écoles" et compte 817 liens dirigés

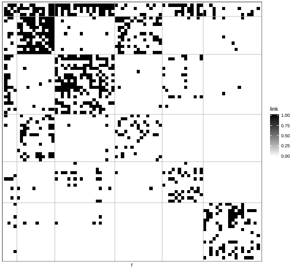




SBM – Réseau UKfaculty – Package **blockmodels**

```
> UK.sbm <- BM_bernoulli("SBM",UK_adjacency)
```





Autour du SBM – packages R

- SBM valué et/ou covariables : lois gaussienne et de Poisson
  package blockmodels
- SBM tenant compte des données manquantes
  missSBM
- Overlapping SBM : possibilité d'appartenir à plusieurs groupes
  package OSBM
- Modèles à blocs latents [LBM] : SBM pour graphes bipartites
  package blockmodels
- SBM multiplex
  package blockmodels (binaire) et codes R
- Tests d'ajustement à ER, HER, W-graphe, SBM, EDD
  codes R
- Test pour savoir si les covariables collectées sont suffisantes pour expliquer le réseau
  package gofnetwork

Modèle à blocs latents [LBM]

LBM (Govaert and Nadif, 2003)

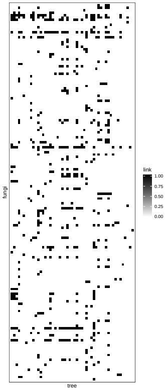
$$
\begin{array}{ll}
(Z_i^R) \text{ i.i.d.} & Z_i^R \sim \mathcal{M}(1, \pi^R) \\
(Z_j^C) \text{ i.i.d.} & Z_j^C \sim \mathcal{M}(1, \pi^C)
\end{array}
$$

$$(Y_{ij}) \text{ indep.} \mid (Z_i^R, Z_j^C) \quad (Y_{ij} \mid Z_i^R = k, Z_j^C = \ell) \sim \mathcal{B}(\alpha_{k\ell})$$

Toute paire de nœuds (constituée d'un nœud du "haut" et d'un nœud du "bas") a une probabilité de connexion induite par un caractère spécifique à chacun de ces nœuds : leur groupe d'appartenance. Les groupes se constituent de nœuds de même nature.

LBM - hôte-parasite

154 espèces de champignons et 51 espèces d'arbres interagissent lorsqu'un champignon parasite un arbre et de manière équivalente lorsqu'un arbre est hôte d'un champignon.



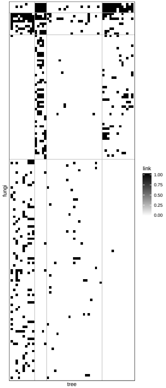
LBM – hôtes-parasites – Package **blockmodels** (1)

```
# définition de l'objet
> fungi_tree.lbm <- BM_bernoulli("LBM",as.matrix(fungi_tree))

# méthode d'inférence
> fungi_tree.lbm$estimate()

# nombre de groupes sélectionné avec vICL
> paramEstimLBM <- extractParamBM(fungi_tree.lbm,Q)
> paramEstimLBM$Q
QRow QCol
   4    4
```

Le critère de sélection de modèle trouve 4 groupes d'arbres et 4 groupes de champignons

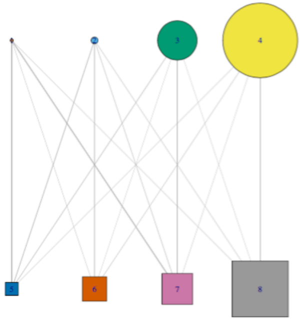
LBM – hôtes-parasites – Package **blockmodels** (2)

```
# extraction des paramètres estimés
> paramEstimLBM <- extractParamBM(fungi_tree.lbm,Q)

# appartenance des noeuds aux groupes
> paramEstimLBM$ZRow; paramEstimLBM$ZCol
```



LBM – hôtes-parasites – Package **blockmodels** (3)

```
> paramEstimLBM$piRow; paramEstimLBM$piCol
[1] 0.02655516 0.05570334 0.31666508 0.60107642
[1] 0.1050494 0.1963968 0.2477304 0.4508234

> paramEstimLBM$alpha
           [,1]        [,2]        [,3]        [,4]
[1,] 0.96813478 0.077538579 0.840370657 0.067563355
[2,] 0.52055882 0.584398216 0.230893917 0.107930384
[3,] 0.32450427 0.003624764 0.098526840 0.005780612
[4,] 0.01834547 0.154334411 0.001330278 0.019219920
```



Références ER, PA, ERGM, SBM, LBM

- Erdös, P. et Rényi, A, (1959). On random graphs, *I Publicationes Mathematicae (Debrecen)*, **6**, 290–297.
- Barabási, A-L et Albert, R., (1999). Emergence of Scaling in Random Networks,*American Association for the Advancement of Science*, **286**, 509–512.
- Wasserman, S. et Pattison, P. (1996)., Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp",*Psychometrika*, **61**, 401–425.
- Nowicki, K. et Snijders, T.A.B., (2001). Estimation and prediction for stochastic block-structures, *JASA*, **96**, 1077–87.
- Govaert, G. and Nadif, M (2003). Clustering with block mixture models. Pattern Recognition, 36(2): 463?473.