

II) Analyse de données d'échange

b) Les modèles à blocs stochastiques

UMR MIA-Paris, AgroParisTech, INRA

Formation Analyse de Réseaux

11-12 Juin 2018



Objectifs

- ▶ Quelques modèles de graphes aléatoires. Miment-ils les propriétés de réseaux observés ?
- ▶ Focus sur le modèle à blocs stochastique [SBM] qui suppose que les liens entre individus découlent de leur appartenance à un groupe. Comment l'exploiter ?
- ▶ Focus sur SBM lorsqu'on dispose également d'informations sur les individus. Les liens découlent alors de l'appartenance au groupe et aussi de ces informations.
- ▶ Quelques références et packages R sur les extensions du SBM.

Sommaire

Exemples de modèles de graphes aléatoires

- Modèle d'Erdős-Rényi

- Modèle d'attachement préférentiel

- Modèle d'ERGM

Modèle à blocs stochastiques

Graphes aléatoires

Réseau d'interaction = Graphe aléatoire $\mathcal{G} = (\mathcal{V} = \{1, \dots, n\}, \mathcal{E})$

Données : \mathbf{Y} la matrice d'adjacence de \mathcal{G}

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

$Y_{ii} = 0$ pour tous les i et $Y_{ij} = Y_{ji}, \forall i \neq j$ si liens non dirigés.

Y_{ij} sont des variables aléatoires, i.e. les relations s'établissent aléatoirement

Modèle d'Erdős-Rényi

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

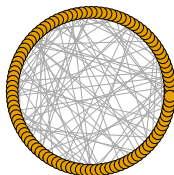
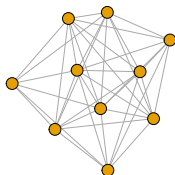
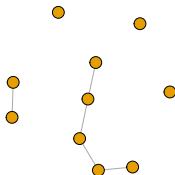
$$Y_{ij} \sim^{iid} \mathcal{B}(p)$$

Tous les noeuds ont même probabilité de connexion

Erdős-Rényi – Exemple (1)

```
G1 <- igraph::sample_gnp(10, 0.1)
G2 <- igraph::sample_gnp(10, 0.9)
G3 <- igraph::sample_gnp(100, .02)

par(mfrow=c(1,3))
plot(G1, vertex.label=NA)
plot(G2, vertex.label=NA)
plot(G3, vertex.label=NA, layout=layout.circle)
```



Erdős-Rény – Caractéristiques

```
> average.path.length(G3)
[1] 4.938823
> diameter(G3)
[1] 11
```

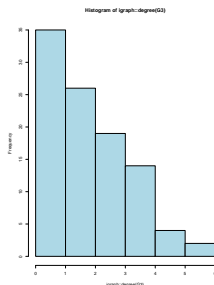
Les chemins le plus court et le plus long se constituent de relativement peu de nœuds

```
> transitivity(G3)
[1] 0.02955665
```

Le coefficient de clustering est assez faible

```
hist(degree(G3), col="lightblue")
```

La distribution des degrés est assez homogène



Modèle d'attachement préférentiel

Modèle d'attachement préférentiel (Barabási 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 $D_i^\alpha + \text{constante}$,
où $D_i = \sum_{j \neq i} Y_{ij}$ est le degré du nœud i

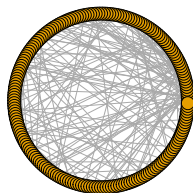
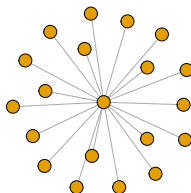
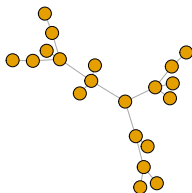
Les nœuds qui ont un fort degré ont de grandes chances d'être connectés : les riches s'enrichissent.

Modèle d'attachement préférentiel – Exemple

```
G1 <- igraph::sample_pa(20, 1, directed=FALSE)
```

```
G2 <- igraph::sample_pa(20, 5, directed=FALSE)
```

```
G3 <- igraph::sample_pa(200, directed=FALSE)
```



Modèle d'attachement préférentiel – Caractéristiques

```
> average.path.length(G3)
[1] 6.05397
> diameter(G3)
[1] 13
```

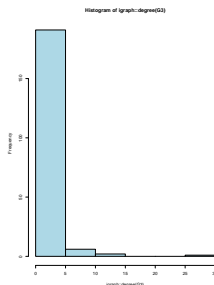
Les chemins le plus court et le plus long se constituent de relativement peu de nœuds

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

Le coefficient de clustering est nul

```
hist(degree(G3), col="lightblue")
```

La distribution des degrés est hétérogène et caractéristique d'une loi de puissance



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

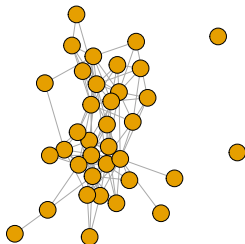
- ▶ \mathbf{y} une réalisation de \mathbf{Y}
- ▶ H une configuration, e.g. arête, triangle, étoile, etc.
- ▶ $g_H(y)$ le nombre de fois où cette configuration apparaît dans \mathbf{y}
- ▶ θ_H le coefficient de dépendance
- ▶ κ la constant de normalisation

La distribution des arêtes est due à la présence de différents motifs dans le réseau observé

ERGM – Exemple

```
library(sand); data(lazega); A <- get.adjacency(lazega)
lazega.s <- network::as.network(as.matrix(A), directed=FALSE)
my.ergm <- formula(lazega.s ~ edges + kstar(2)
                    + kstar(3) + triangle)
```

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



Limites

► Modèle d'Erdős-Rényi

- modélisation d'une structure homogène donc inadaptée aux réseaux réels

► Modèle d'attachement préférentiel

- modélisation d'une structure où la distribution des degrés est une loi de puissance, i.e. existence d'un petit groupe de nœuds fortement connectés
- défini par un algorithme – cadre statistique non propice à l'estimation des paramètres

► Modèle ERGM

- modélisation de structures très particulières et de petites tailles
- justifications théoriques analogues à celle du glm non établies

Modèle à blocs stochastiques

- modélisation d'une structure de groupes courante de réseaux réels
- cadre statistique propice à l'estimation des paramètres

Sommaire

Exemples de modèles de graphes aléatoires

Modèle à blocs stochastiques

- SBM

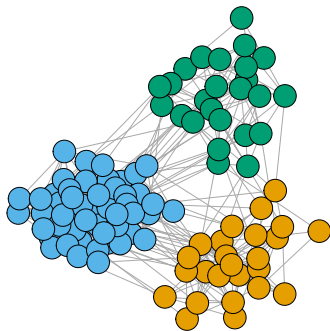
- SBM et covariables

- Autour du SBM – packages R

SBM – Exemple de topologie (1)

Réseau de communauté

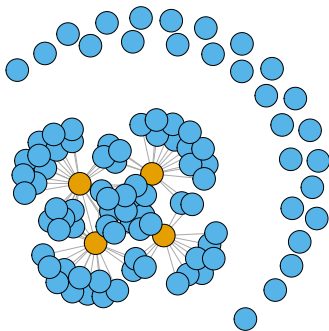
```
pi <- matrix(c(0.3,0.02,0.02,0.02,0.3,0.02,0.02,0.02,0.3),3,3)
communities <- igraph::sample_sbm(100, pi, c(25, 50, 25))
plot(communities, vertex.color = rep(1:3,c(25, 50, 25)))
```



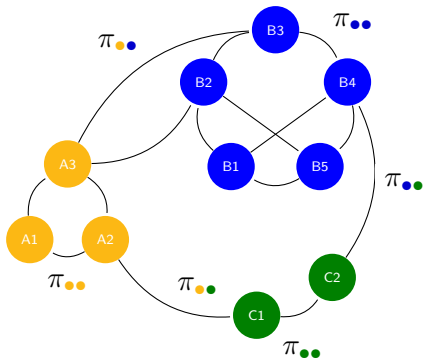
SBM – Exemple de topologie (2)

Réseau en étoile

```
pi <- matrix(c(0.05,0.3,0.3,0),2,2)
star <- igraph::sample_sbm(100, pi, c(4, 96))
plot(star, vertex.label=NA, vertex.color = rep(1:2,c(4,96)))
```



Modèle à blocs stochastiques [SBM] (1)



SBM (Nowicki et Snijders, 2001)

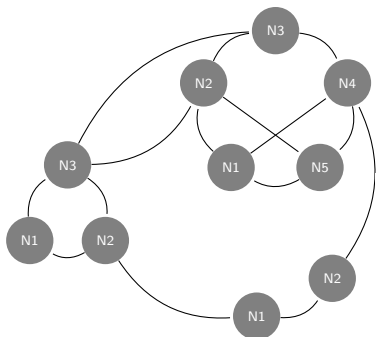
Soient n nœuds répartis ainsi :

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

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

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 :

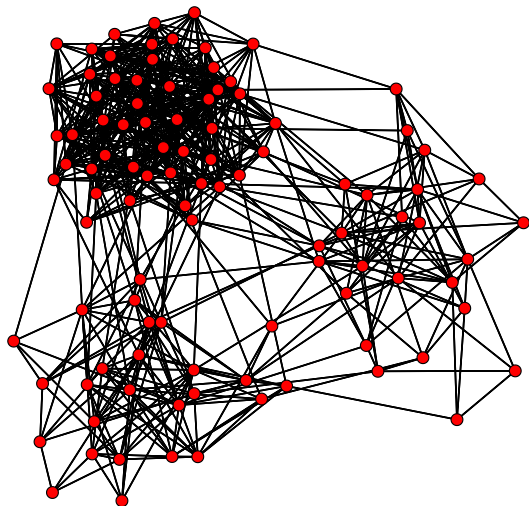
- ▶ $\mathcal{Q} = \{\bullet, \bullet, \bullet\}$, $\text{card}(\mathcal{Q})$ connu
- ▶ $\alpha_{\bullet} = ?$,
- ▶ $\pi_{\bullet\bullet} = ?$

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

SBM – Estimation – Sélection de modèle

- ▶ Estimation de α le vecteur des probabilités d'appartenance aux Q groupes
via un algorithme EM variationnel
- ▶ Estimation de π la matrice des probabilités de connexion au sein des groupes et entre les groupes
via ce même vEM
- ▶ 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)

```
library(blockmodels)

# matrice d'adjacence du graphe de communautés
communities=as.matrix(get.adjacency(communities,type="both"))

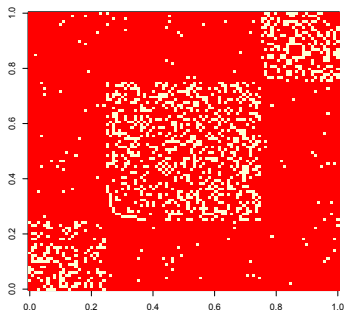
my_model <- BM_bernoulli("SBM_sym",communities,verbosity=0,plotting='')
m=my_model$estimate()

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

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

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

```
# probabilités a posteriori d'appartenance de chacun des noeuds aux groupes  
head(my_model$memberships[["Q"]$Z)  
      [,1]      [,2]      [,3]  
[1,] 0.000999001 0.998002 0.000999001  
[2,] 0.000999001 0.998002 0.000999001
```

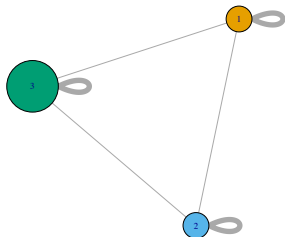


SBM – Communautés – Package **blockmodels** (3)

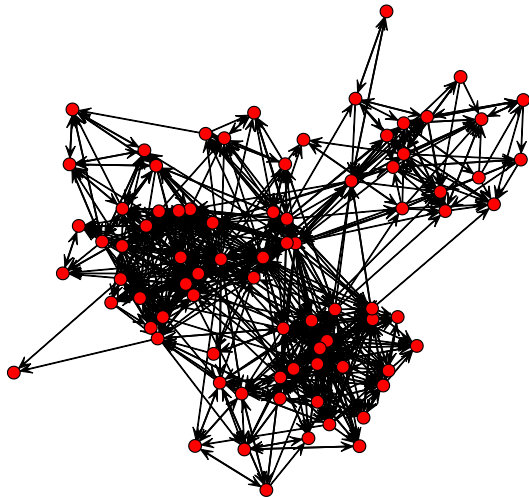
```
my_model$model_parameters[[Q]]$pi
      [,1]      [,2]      [,3]
[1,] 0.33136563 0.01974406 0.01676524
[2,] 0.01974406 0.28168363 0.01833345
[3,] 0.01676524 0.01833345 0.32752832

> colSums(my_model$memberships[[Q]]$Z)/100
[1] 0.2502498 0.2502498 0.4995005
```

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



SBM – UKfaculty $n = 181, \rho = 0.13$

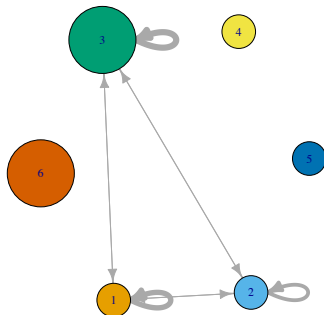


SBM – Réseau UKfaculty – Package **blockmodels**

```
# matrice d'adjacence
Net=as.matrix(UKfaculty_adj_cov$Net)

sbm.faculty <- BM_bernoulli("SBM", Net); sbm.faculty$estimate()

# nombre de groupes sélectionné
> which.max(sbm.faculty$ICL)
[1] 6
```



SBM – UKfaculty – Package mixer

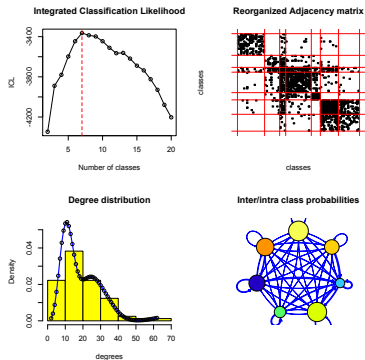
```
library(mixer)
```

```
> mix.sbm.faculty<- mixer(x=Net,qmin=2,qmax=20)
```

Mixer: the adjacency matrix has been transformed in a directed edge list

```
plot(mix.sbm.faculty); faculty.output <- getModel(mix.sbm.faculty)
```

```
> faculty.output$q  
[1] 7
```



SBM et covariables

SBM avec covariables (Mariadassou et al., 2010)

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \alpha), \quad \forall \bullet \in \mathcal{Q}$$
$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{B} \left(g(\pi_{\bullet\bullet} + x_{ij}^T \beta) \right)$$

avec g la fonction logistique et x_{ij} le vecteur de covariables sur la dyade $(i ; j)$.

Pour p covariables, $\mathbb{P}(i \sim j)$ est une fonction croissante de $\pi_{\bullet\bullet} + x_{ij}^1 \beta_1 + \dots + x_{ij}^p \beta_p$.

La présence d'une arête dépend de l'appartenance de chaque nœud à un groupe ainsi que des covariables qui portent sur la dyade correspondante. La structure de groupe dépend alors d'informations autres à ces covariables.

Construction des covariables sur les dyades – UKfaculty

Covariable **école d'affiliation** (qualitative-4 niveaux)

↪ version quantitative : valeur absolue de la différence

↪ version binaire pour chaque niveau ℓ du facteur

$$x_{ij}^{(\ell)} = \begin{cases} 1 & \text{si } i \text{ et } j \text{ de même niveau } \ell \\ 0 & \text{sinon} \end{cases}$$

Pour chaque école, la covariable vaut 1 si 2 individus sont affiliés à la même école et 0 sinon

↪ version ternaire: pour chaque niveau ℓ du facteur

$$x_{ij1}^{(\ell)} = \begin{cases} 1 & \text{si } i \text{ et } j \text{ de même niveau } \ell \\ 0 & \text{sinon} \end{cases}$$

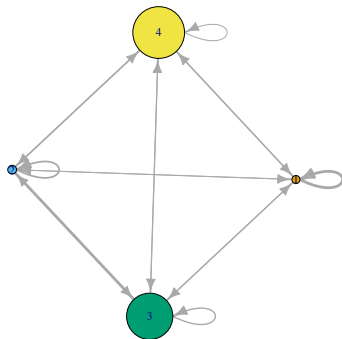
et

$$x_{ij2}^{(\ell)} = \begin{cases} 1 & \text{si } i \text{ ou } j \text{ est de niveau } \ell \\ 0 & \text{sinon} \end{cases}$$

SBM et covariables (version binaire) – UKfaculty

```
listVar=list(EdgeCovar[, ,1], ..., EdgeCovar[, ,4])  
sbm.cov.faculty <- BM_bernoulli_covariates_fast("SBM", Net, listVar)  
sbm.cov.faculty$estimate() ; Q=which.max(sbm.cov.faculty$ICL)  
> Q  
[1] 4
```






4 groupes au lieu de 6 : prise en compte des covariables mais il reste d'autres facteurs non identifiés qui engendrent une structure



Autour du SBM – packages R

- ▶ SBM valué avec covariables : lois gaussienne et de Poisson
package **blockmodels**
- ▶ SBM tenant compte des données manquantes
package **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
package **codes R**
- ▶ Test pour savoir si les covariables collectées sont suffisantes pour expliquer le réseau
package **gofnetwork**

Références ER, PA, ERGM, SBM, SBM covariables

-  Barabási, A-L et Albert, R., (1999). Emergence of Scaling in Random Networks, *American Association for the Advancement of Science*, **286**, 509–512.
-  Erdős, P. et Rényi, A, (1959). On random graphs, / *Publicationes Mathematicae (Debrecen)*, **6**, 290–297.
-  Mariadassou, M., Robin, S. et Vacher, C, (2010). Uncovering latent structure in valued graphs: a variational approach, *Ann. Appl. Stat.*, **4**, 715–42.
-  Nowicki, K. et Snijders, T.A.B., (2001). Estimation and prediction for stochastic block-structures, *JASA*, **96**, 1077–87.
-  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.