Clustering de nœuds dans les réseaux à l'aide de mixer

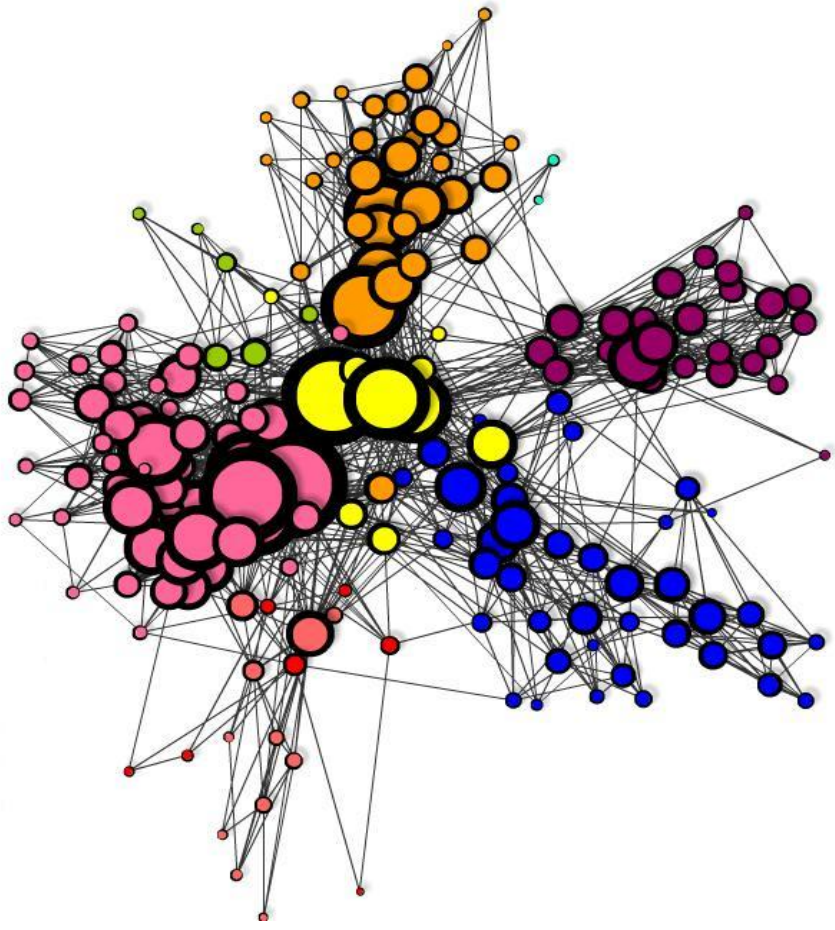
Pierre Latouche

Université Paris 1 Panthéon-Sorbonne
Laboratoire SAMM

Semin-R, 29/03/2012

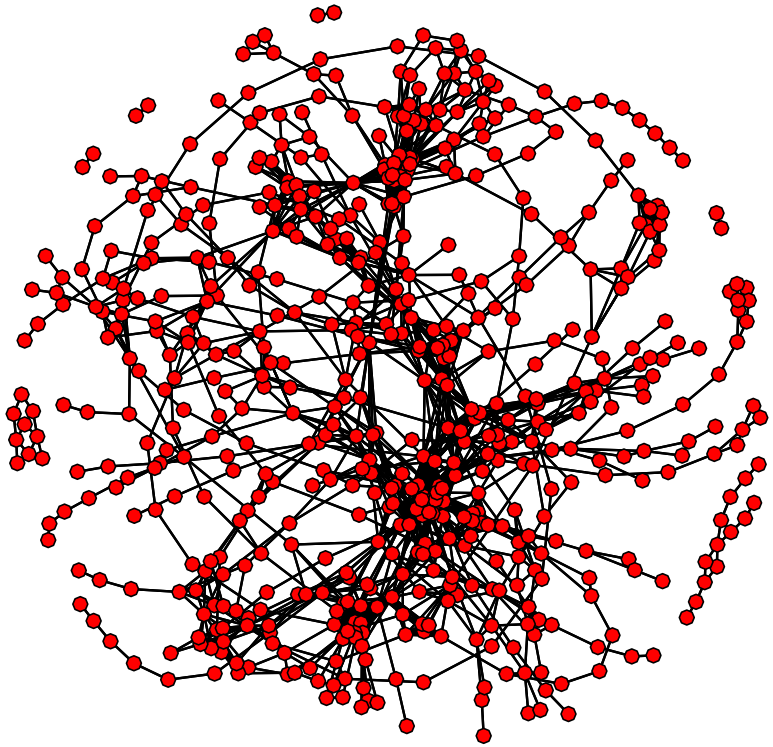
Réseaux "réels"

- **Utilisés dans de nombreux domaines scientifiques** :
  - WWW
  - Biologie, sociologie, physique, ...
- **Nature des données**:
  - Interactions entre N objets
  - $\mathcal{O}(N^2)$ interactions possibles
- **Topologie d'un réseau** :
  - Décrit la manière dont les nœuds interagissent, liens entre structures et fonctions



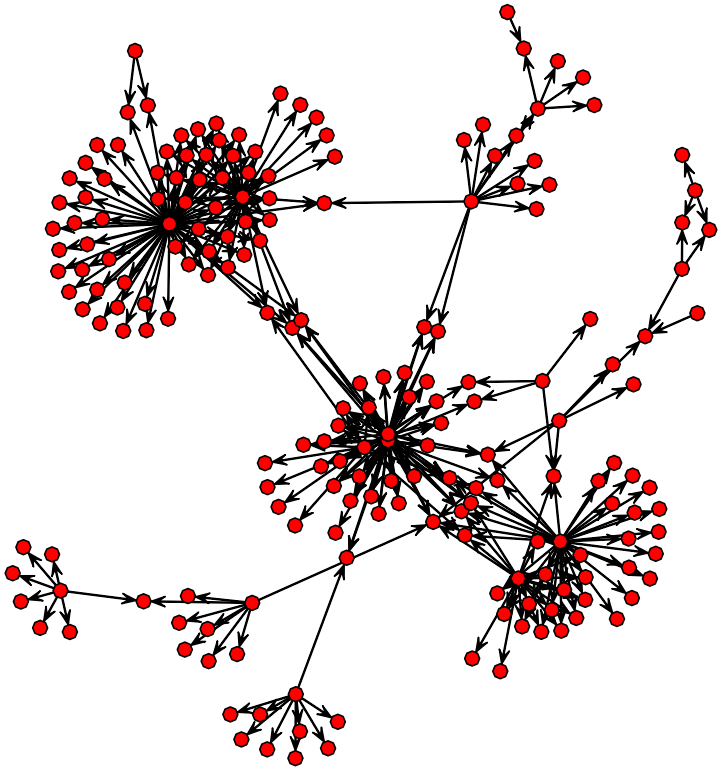
Sous-réseau issu du réseau de la Blogosphère politique française.

En Biologie



Réseau métabolique d'*Escherichia coli* (Lacroix et al., 2006).

En Biologie



Sous-réseau issu du réseau de régulation de la levure (Milo et al., 2002).

Réseaux “réels”

- **Propriétés** :
  - Creux
  - Composante géante
  - Hétérogénéité
  - Attachement préférentiel
  - Petit monde

$\hookrightarrow$ Structures (groupes de nœuds)

Réseaux “réels”

- **Propriétés** :
  - Creux
  - Composante géante
  - Hétérogénéité
  - Attachement préférentiel
  - Petit monde

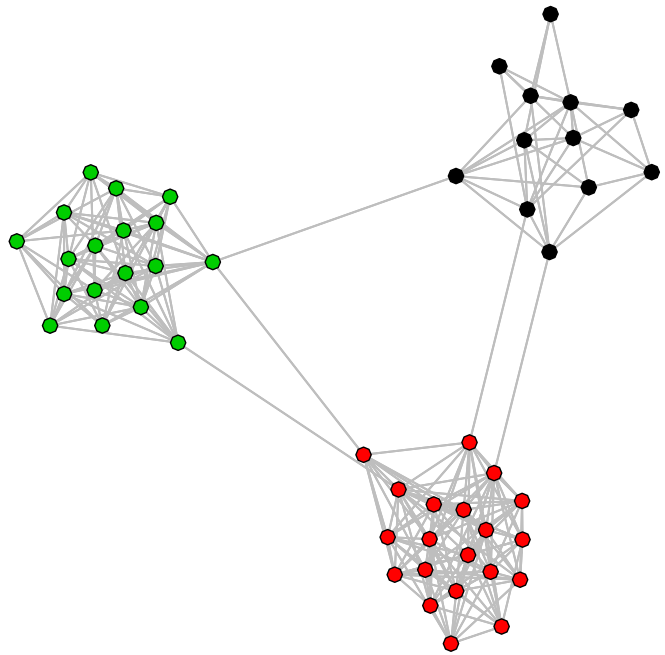
↪ Structures (groupes de nœuds)

Clustering dans les réseaux

- **Les méthodes existantes cherchent** :
  - Communautés
  - Attachement “inverse” (disassortative mixing)
  - Structures hétérogènes

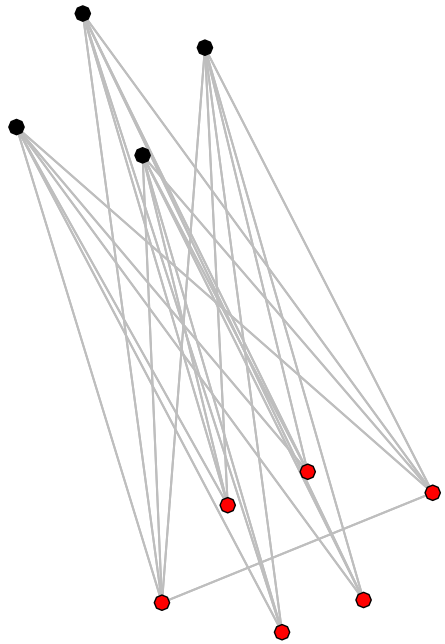
Clustering dans les réseaux

- **Les méthodes existantes cherchent** :
  - Communautés
  - Attachement "inverse" (disassortative mixing)
  - Structures hétérogènes



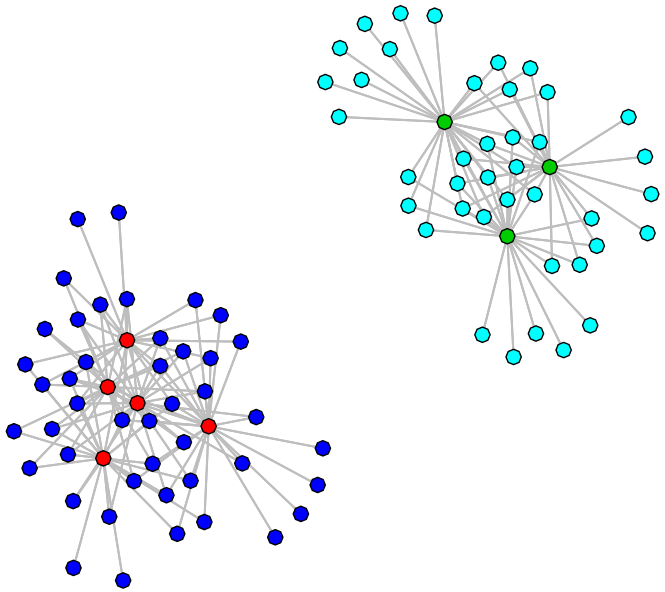
Clustering dans les réseaux

- **Les méthodes existantes cherchent** :
  - Communautés
  - Attachement “inverse” (disassortative mixing)
  - Structures hétérogènes



Clustering dans les réseaux

- **Les méthodes existantes cherchent** :
  - Communautés
  - Attachement “inverse” (disassortative mixing)
  - Structures hétérogènes

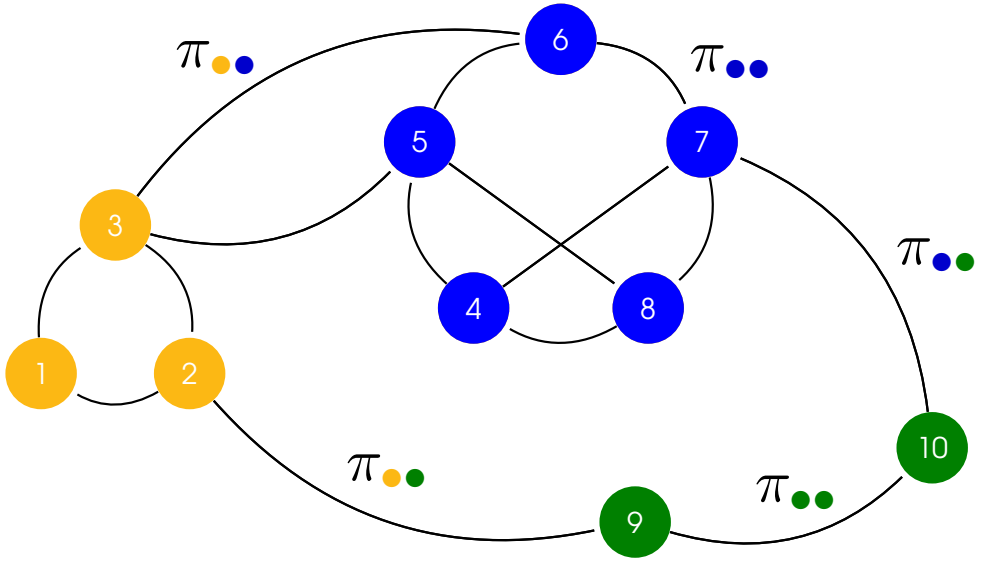


Plan

Modèle à Blocs stochastiques

- Stochastic Block Model (SBM) en anglais
- Nowicki et Snijders (2001)
- $\mathbf{Z}_i$ variables indépendantes, cachées:
  - $\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\alpha})$
  - $Z_{iq} = 1$: nœud i appartient au groupe q
- $\mathbf{X} \,|\, \mathbf{Z}$ arêtes tirées indépendamment :

$$X_{ij} | \{i \in q, j \in l\} \sim \mathcal{B}(\pi_{ql})$$


$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
6
5
7
3
4
8
1
2
10
9

Maximum de vraisemblance

- **Maximum de vraisemblance** :
  - Données observées : $\log p(\mathbf{X} \mid \boldsymbol{\alpha}, \boldsymbol{\Pi})$
    $\hookrightarrow$ Maximisation=Problème
- Solutions classiques (EM) pas applicables !
  - Données complétées : $\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\Pi})$

Sélection de modèles

Critères "classiques"

Comme $\log p(\mathbf{X} \mid \boldsymbol{\alpha}, \boldsymbol{\Pi})$ n'est pas calculable, impossible d'utiliser les critères classiques tels que:

- $AIC = \log p(\mathbf{X} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Pi}}) - K$
- $BIC = \log p(\mathbf{X} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Pi}}) - \frac{K}{2} \log \frac{N(N-1)}{2}$

Critère ICL pour SBM

- Daudin et al. (2008)

Package R mixer, premiers pas

```
>install.packages("mixer")
>library(mixer)
>help(mixer)
```

mixer {mixer} R Documentation

MIXtures of Erdos-Renyi random graphs

Description

Estimate the parameters, the hidden class variables, as well as the number of classes `q` of a MIXture of Erdos Renyi random graphs. The estimation is performed for binary graphs (edges are assumed to be drawn from Bernoulli distributions).

Usage

```
mixer( x, qmin=2, qmax=NULL, method="variational", directed = NULL,
nbiter=10, fpnbiter=5, improve=FALSE, verbose=TRUE)
```

Générer des réseaux

- Réseau avec communautés

```
>g<-graph.affiliation(n=100,c(1/3,1/3,1/3),0.6,0.1)
```

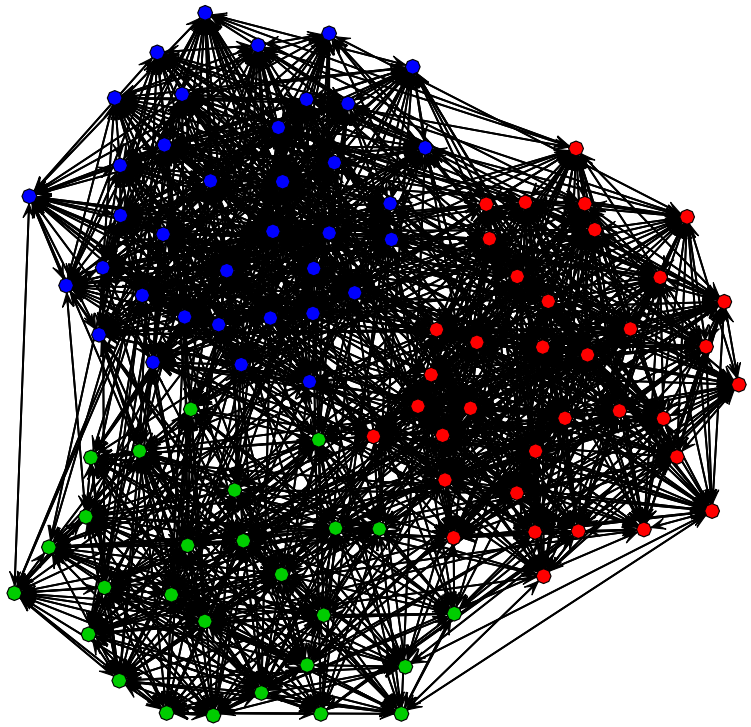
- Attachement “inverse”

```
>g<-graph.affiliation(n=100,c(1/3,1/3,1/3),0.1,0.6)
```

- `str(g)`

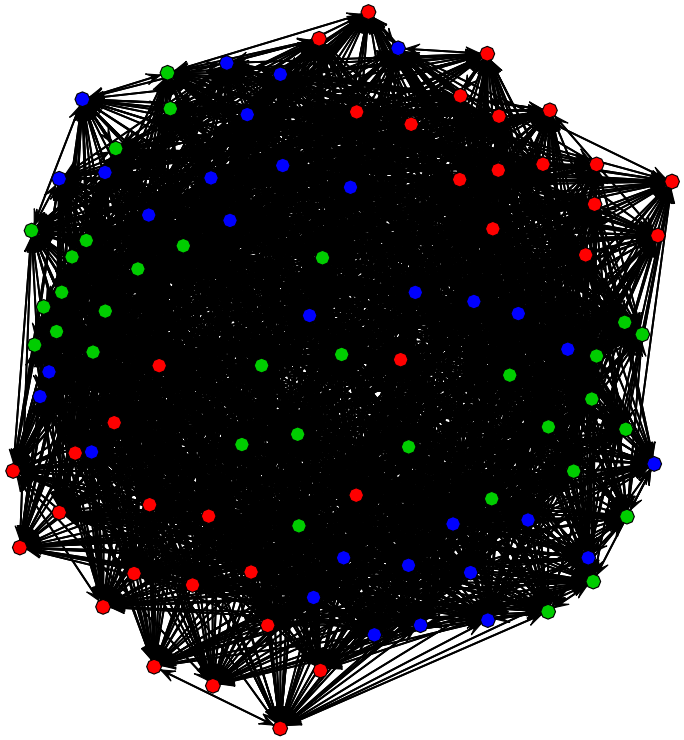
Afficher des réseaux

```
>library(sna)
>gplot(g$x, vertex.col=g$cluster+1)
```



Afficher des réseaux

```
>library(sna)
>gplot(g$x, vertex.col=g$cluster+1)
```

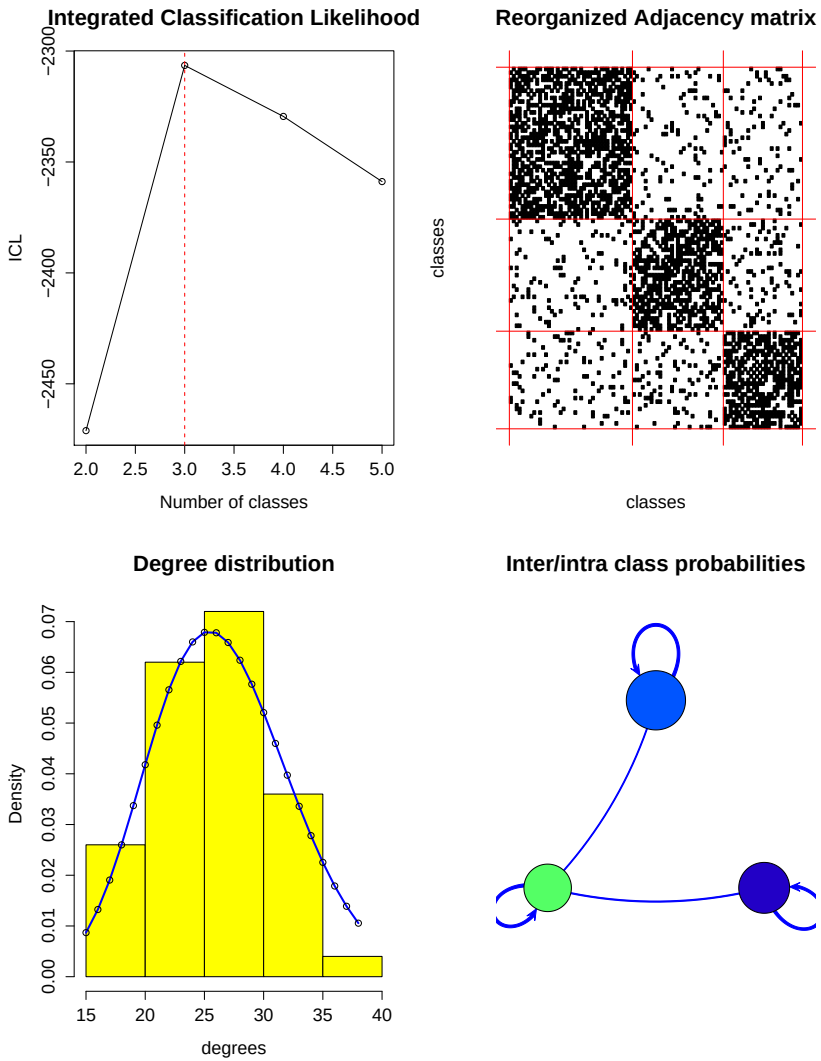


Plusieurs algorithmes d'optimisation

- variationnel EM + ICL (Daudin et al. 2008)
- classification EM + ICL (Zanghi et al. 2008) $\hookrightarrow$ Plus rapide mais biaisé
- variationnel Bayes EM + ILvb (Latouche et al. 2012) $\hookrightarrow$ Plus lent mais meilleurs résultats

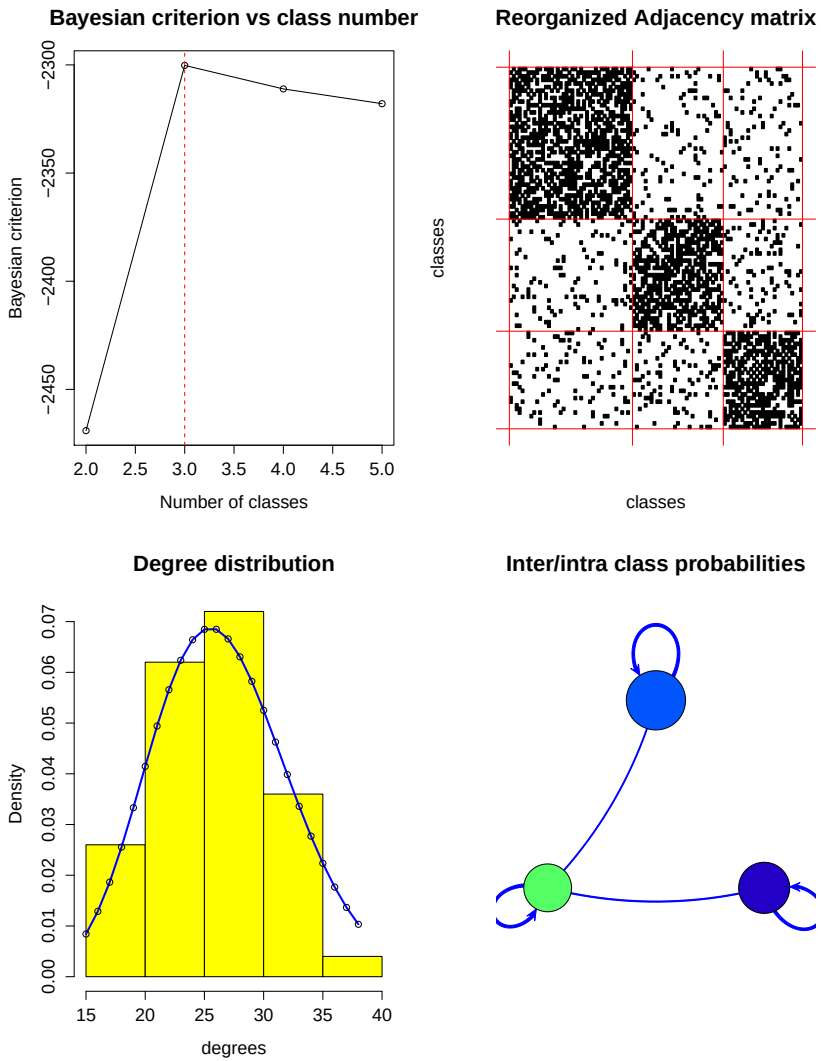
Utilisation de mixer

```
>xout<-mixer(g$x, qmin=2, qmax=5, method="variational")
>plot(xout)
```



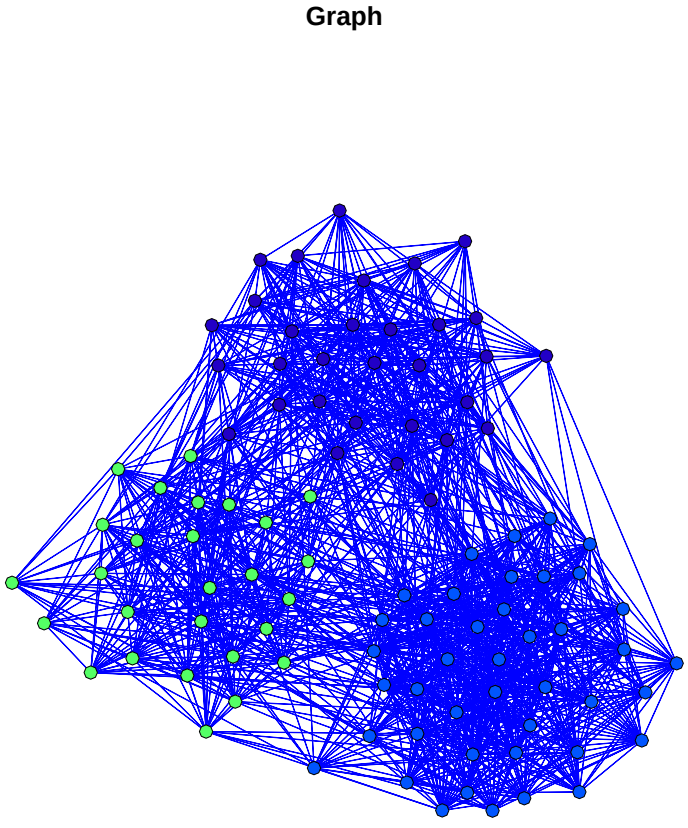
Utilisation de mixer

```
>xout<-mixer(g$x, qmin=2, qmax=5, method="bayesian")
>plot(xout)
```



Utilisation de mixer

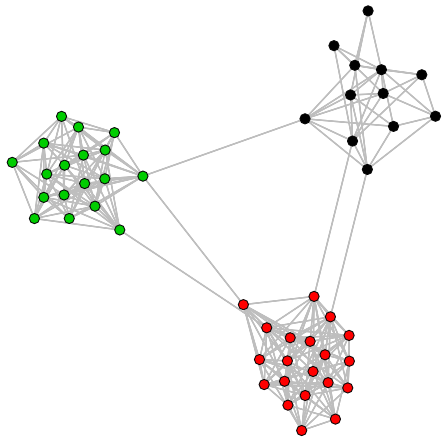
```
>plot(xout, frame=5)
```



Série d'expériences (Latouche et al. 2012)

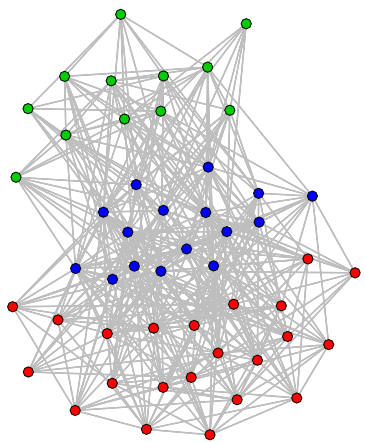
- **Deux types de structures** :
  - Affiliation :

$$\mathbf{\Pi} = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon \\ \epsilon & \lambda & & \vdots \\ \vdots & & \ddots & \epsilon \\ \epsilon & \dots & \epsilon & \lambda \end{pmatrix}$$



  - Affiliation et une classe de hubs :

$$\mathbf{\Pi} = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon & \lambda \\ \epsilon & \lambda & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \lambda & \dots & \dots & \dots & \lambda \end{pmatrix}$$



Réseaux d'affiliation

(a) $Q_{True} \backslash Q_{ICL}$

	2	3	4	5	6	7
3	0	100	0	0	0	0
4	0	0	100	0	0	0
5	0	0	23	**77**	0	0
6	0	1	28	59	**12**	0
7	0	8	49	42	1	**0**

(b) $Q_{True} \backslash Q_{ILvb}$

	2	3	4	5	6	7
3	0	100	0	0	0	0
4	0	0	100	0	0	0
5	0	0	0	**99**	1	0
6	0	0	4	23	**73**	0
7	0	2	14	44	27	**13**

Réseaux d'affiliation et une classe de hubs

(c) $Q_{True} \backslash Q_{ICL}$

	2	3	4	5	6	7
3	0	**100**	0	0	0	0
4	0	0	**100**	0	0	0
5	0	0	12	**88**	0	0
6	0	0	19	59	**22**	0
7	0	3	29	56	12	**0**

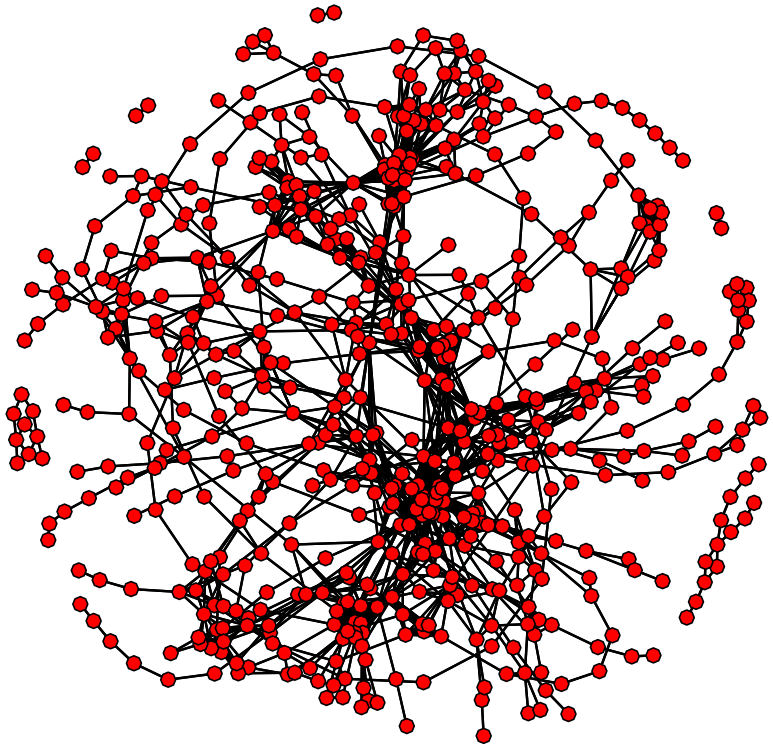
(d) $Q_{True} \backslash Q_{ILvb}$

	2	3	4	5	6	7
3	0	**100**	0	0	0	0
4	0	0	**100**	0	0	0
5	0	0	2	**98**	0	0
6	0	0	1	29	**70**	0
7	0	0	3	34	45	**18**

Réseau métabolique d'*Escherichia coli*

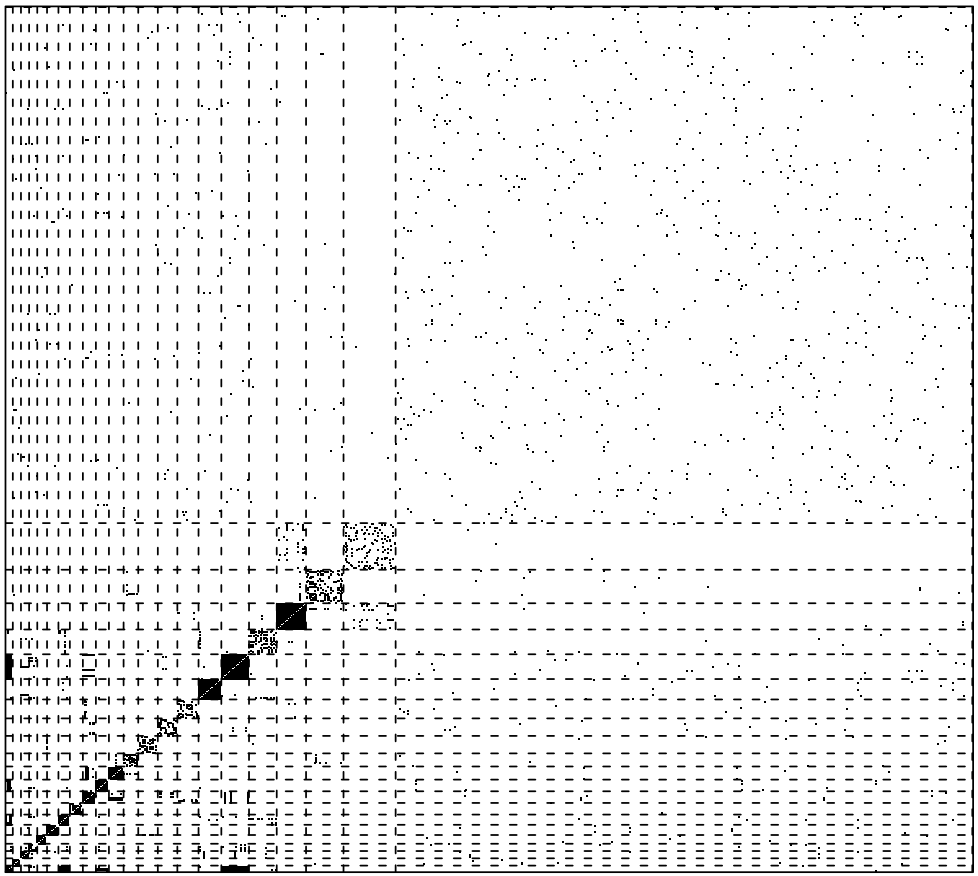
- Lacroix et al. (2006)
- Lab : Biométrie et Biologie Évolutive (Lyon 1)
- 605 nœuds, 1782 arêtes

Réseau métabolique d'*Escherichia coli*



Réseau métabolique d'*Escherichia coli* (Lacroix et al., 2006).

Résultats (1)



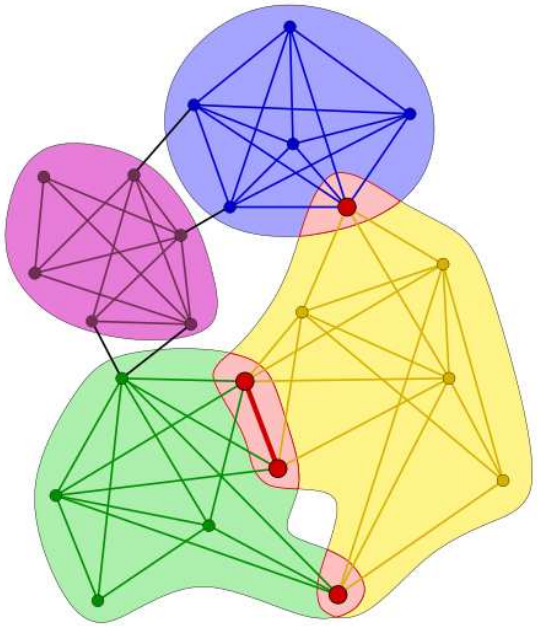
Représentation “dot blot” du réseau métabolique après classification des nœuds en $K = 22$ classes.

Résultats (2)

- Parmi les classes, huit sont des cliques
- Six ont une probabilité intra-classe supérieure à 0.5
- Les cliques et pseudo-cliques partagent des réactions faisant intervenir un même composé
  - Cliques : chorismate, pyruvate, L-aspartate, L-glutamate, D-glyceraldehyde-3-phosphate and ATP
- Les classes 1 and 17 sont associées au pyruvate

Contents

Chevauchement dans les réseaux



Palla et al. (2005)

Problème, limite

SBM et la plupart des méthodes de clustering existantes font l'hypothèse qu'un nœud appartient à un groupe et à un seul

Modèle à blocs stochastiques (SBM)

- Nowicki et Snijders (2001)
- $\mathbf{Z}_i$ indépendantes, cachées :

$$\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\alpha})$$

Modèle à blocs stochastiques chevauchants (OSBM)

- Latouche et al. (2011)
- Z_{iq} indépendantes, cachées :

$$\mathbf{Z}_i \sim \prod_{q=1}^{Q} \mathcal{B}(Z_{iq};\, \alpha_q) = \prod_{q=1}^{Q} \alpha_q^{Z_{iq}} (1-\alpha_q)^{1-Z_{iq}}$$

Modèle à blocs stochastiques chevauchants (OSBM)

- Latouche et al. (2011)
- Z_{iq} indépendantes, cachées :

$$\mathbf{Z}_i \sim \prod_{q=1}^{Q} \mathcal{B}(Z_{iq};\ \alpha_q) = \prod_{q=1}^{Q} \alpha_q^{Z_{iq}} (1-\alpha_q)^{1-Z_{iq}}$$

- $\mathbf{X} \,|\, \mathbf{Z}$ Tirage des arêtes :

$$X_{ij} |\, \mathbf{Z}_i, \mathbf{Z}_j \sim \mathcal{B}\big(X_{ij};\ g(a_{\mathbf{Z}_i,\mathbf{Z}_j})\big)$$

- $g(t) = 1/\left(1+\exp(-t)\right)$ fonction logistique

$$a_{\mathbf{Z}_i,\mathbf{Z}_j} = \mathbf{Z}_i^\intercal\, \mathbf{W}\, \mathbf{Z}_j + \mathbf{Z}_i^\intercal\, \mathbf{U} + \mathbf{V}^\intercal\, \mathbf{Z}_j + W^*$$

Package osbm : très bientôt ..

- Autorise le chevauchement entre les groupes
- Sélection de modèles

	UMP	UDF	liberal	PS	analysts	others
cluster 1	30 + 3	0 + 1	0	0	0 + 1	0
cluster 2	2 + 3	29 + 1	0	0	1 + 3	0
cluster 3	0	0	24	0	1 + 1	0
cluster 4	0	0 + 2	0	40	0 + 4	1
outliers	5	1	1	17	5	30

Analyse de la Blogosphere en $K = 4$ classes.

Références

- Groupe SSB ! ANR NeMo !
- Package R mixer = Christophe Ambroise, Gilles Grasseau, Mark Hoebeke, Vincent Miele, Franck Picard, Pierre Latouche
- J-J. Daudin, F. Picard et S. Robin (2008), A mixture model for random graphs. Statistics and Computing, 18, 2, 151-171.
- H. Zanghi, C. Ambroise et V. Miele (2008), Fast online graph clustering via Erdos-Renyi mixture. Pattern Recognition, 41, 3592-3599
- P. Latouche, E. Birmelé, C. Ambroise (2011), Overlapping stochastic block models with application to the French political blogosphere network. Annals of Applied Statistics, 5, 1, 309-336
- P. Latouche, E. Birmelé, C. Ambroise (2012), Variational Bayesian inference and complexity control for stochastic block models. Statistical Modelling, 12, 1, 93-115