Stochastic Block Model pour les réseaux multiplex. Application à un réseau de chercheurs en cancérologie.

P. Barbillon[1], A. Bar-Hen[2], S. Donnet[1], E. Lazega[3]

[1]AgroParisTech / INRA MIA UMR 518, [2]MAP 5, Université Paris-Descartes,

[3]Institut d'Études Politiques de Paris (Sciences Po)

NetBio, 29-30 septembre 2015, Paris

Outline

Contexte et données

Données

Catching up with big fish in the big pond ? Multi-level network analysis through linked design

Emmanuel Lazega, Marie-Thérèse Jourda, Lise Mounier, Rafaël Stofer. Social Networks, **30 :2**, 159-176 (2008)

- "Elite", des chercheurs français en cancérologie à la fin des années 1990.
- Parmi les 168 chercheurs identifiés, 128 ont accepté l'entretien (76%)
  - Info sur les chercheurs : âge, spécialité, laboratoire, 2 indicateurs de performance de publication, directeur du laboratoire ou non
  - Connexions entre chercheurs
- Concernant les laboratoires, 76 directeurs ont acceptés l'entretien.
  - Informations sur les laboratoires : localisation, taille (# chercheurs)
  - Connexions entre laboratoires

Relations étudiées

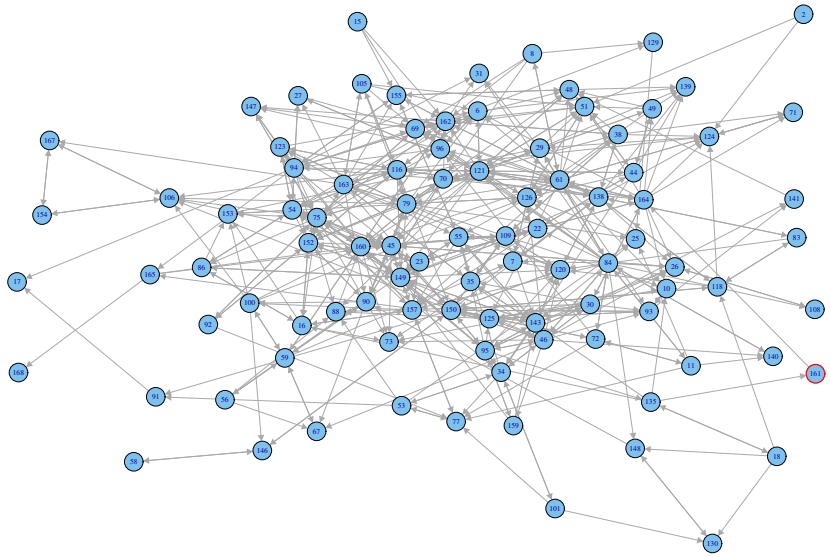
- *Relations entre chercheurs*. 5 types de relations de conseils : conseils concernant la direction des projets, conseils pour trouver des financements institutionnels, conseils pour gérer les ressources financières, conseils pour les recrutements, conseils à propos des articles avant soumission aux journaux
- *Relations entre laboratoires* : échanges de ressources. Recrutement de chercheurs ou post-docs, développement de programmes commun de recherche, réponses conjointes à des programmes de recherche, partage d'équipement techniques, partage de matériel expérimental, mobilité de personnels administratifs, invitation à des conférences et séminaires

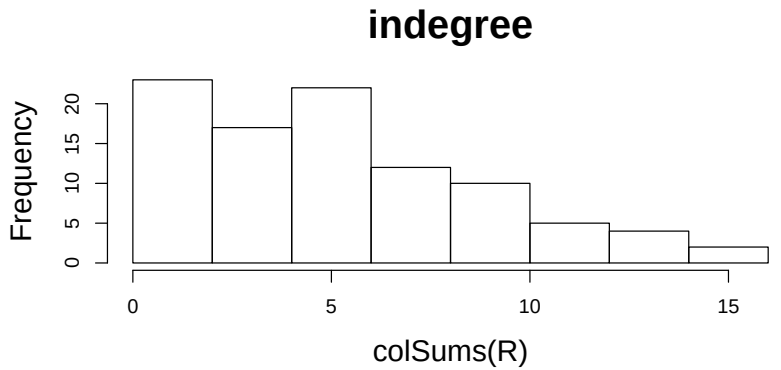
Agrégation des réseaux

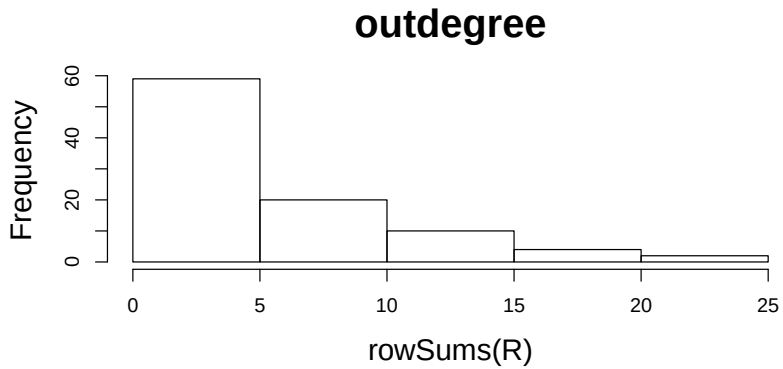
- Pour les niveaux individuel et organisationnel, les différents types de relations ont été agrégés.

  $\Rightarrow$ Deux réseaux :
  - Réseau de conseil entre 95 chercheurs,
  - Réseau d'échange de ressources entre 76 laboratoires.
- **Remarques**
  - Hiérarchique (mais peu d'individus par laboratoire)
  - Réseaux dirigés

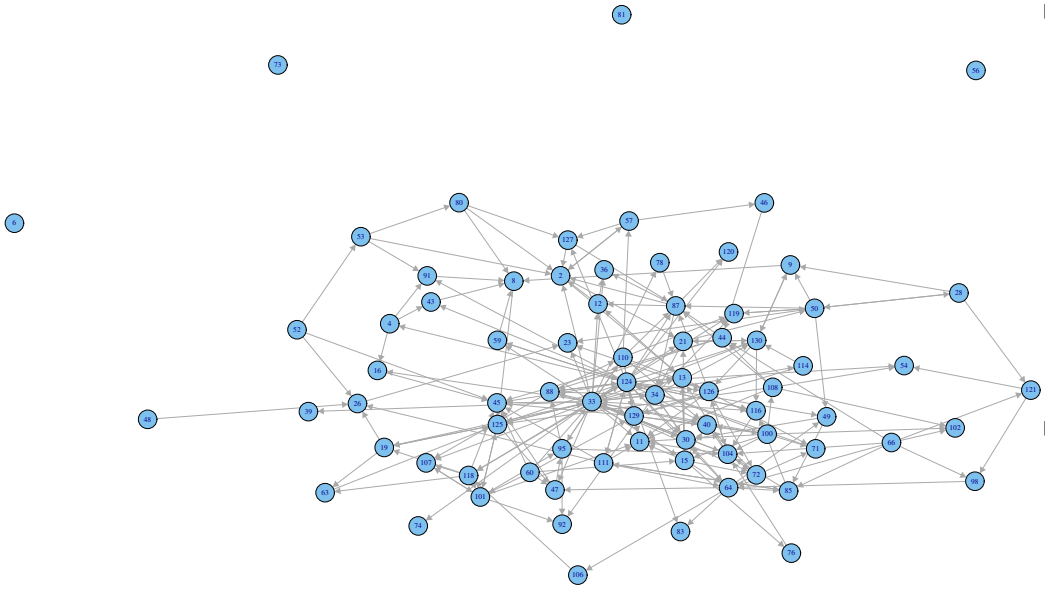
Réseau des chercheurs

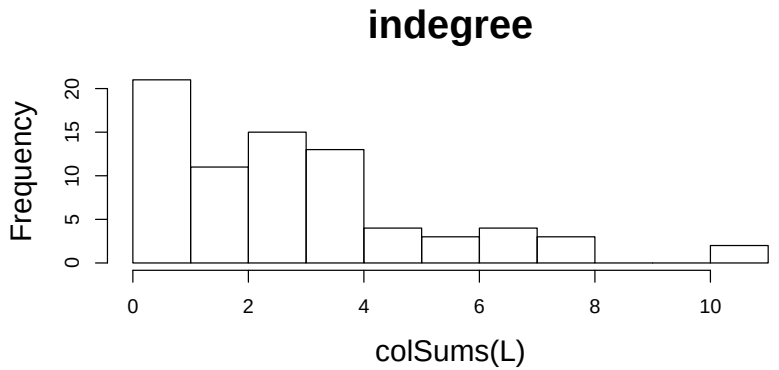


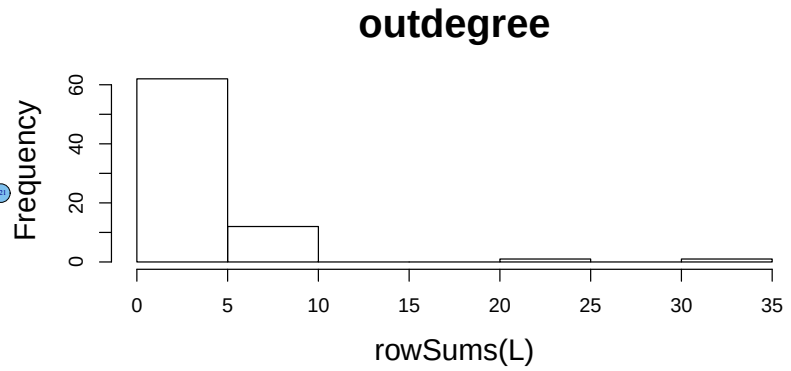




Réseau des laboratoires







Problématiques

- Décrire et comprendre les interactions entre les 2 niveaux (individuels et organisationnel), les individus appartenant à des institutions
- Le fait que deux laboratoires sont liés influence-t-il la relation entre ses membres ? De façon positive ? Négative ?
- Tous les chercheurs profitent-il de la même façon des relations de leurs laboratoires ?
- Objectif : faire des groupes de chercheurs partageant les mêmes propriétés de connections

De multi-niveaux à multiplex

- Soient i et j deux chercheurs $\in \{1, \ldots, n\}$.
- On définit 2 types de relations entre i et j.

$$R_{ij} = X_{ij}^1 = 1 \qquad \text{si } i \text{ demande conseil directement à } j$$
$$L_{ij} = X_{ij}^2 = 1 \qquad \text{si le lab. de } i \text{ fournit des ressources au lab. de } j$$

- $$\forall (i,j), \quad X_{ij}^{1:2} = (X_{ij}^1, X_{ij}^2) \in \{0,1\}^2$$
- Relations directes et indirectes $\Rightarrow$ Multiplex
- Démarche qui semble raisonnable puisqu'on a peu de chercheurs par laboratoire

Erdós-Rényi multiplex

- Modèle le plus simple :

$$\forall (i,j) \in \{1,\dots,n\}^2, i \neq j, \forall w \in \{0,1\}^2,$$

$$\mathbb{P}(X_{ij}^{1:2} = w) = \pi^{(w)} \text{ where } \sum_{w \in \{0,1\}^2} \pi^{(w)} = 1\,,$$

- $(X_{ij}^{1:2})_{i,j}$ sont mutuellement indépendants
- Mais types de relations (directe / indirecte) non indépendants
- Probabilités de connections identiques pour tous les individus

SBM for multiplex

- Idée : introduire des motifs de connexion non purement réguliers
- [Snijders and Nowicki, 1997]
- Supposons que les individus $i \in \{1, ..., n\}$ appartiennent à un des Q groupes (appartenance non-observée)
- $$\boxed{Z_i = q \quad \text{si } i \text{ appartient au groupe } q}$$
- $\forall (i,j) \in \{1, \dots, n\}^2, i \neq j, \forall w \in \{0,1\}^2, \forall (q,l) \in \{1, \dots, Q\}^2,$

$$\boxed{\begin{array}{rcl} \mathbb{P}(X_{ij}^{1:2} = w | Z_i = q, Z_j = l) & = & \pi_{ql}^{(w)} \\ P(Z_i = q) & = & \alpha_q \, . \end{array}} \quad (1)$$

- Conditionnellement aux affectations $(Z_i)_{i=1\dots n}$, les connexions sont indépendantes
- Hétérogénéité dans les connexions introduites par les Z_i

Estimation de paramètres

- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$, $\quad \boldsymbol{\pi} = (\pi_{ql}^{(w)})_{w \in \{0,1\}^2, (q,l) \in \{1,\ldots,Q\}^2}$, $\quad \theta = (\boldsymbol{\alpha}, \boldsymbol{\pi})$
- $(2^2 - 1)Q^2 + (Q - 1)$ paramètres
- Vraisemblance :

$$\begin{aligned} \ell(\mathbf{X}^{1:2}; \theta) &= \int_{\mathbf{z} \in \{1,\ldots Q\}^n} p(\mathbf{X}^{1:2} | \mathbf{Z}; \boldsymbol{\pi}) p(\mathbf{Z}; \boldsymbol{\alpha}) d\mathbf{Z}\,, \\ &= \sum_{\mathbf{Z} \in \{1,\ldots Q\}^n} \prod_{i,j,i \neq j} \pi_{Z_i Z_j}^{(X_{ij}^{1:2})} \prod_{i=1}^{n} \alpha_{Z_i}\,, \end{aligned} \tag{2}$$

- Dès que n ou Q grands, la vraisemblance observée (2) ne peut être calculée : maximisation difficile.

EM variationnel

- Adaptation du variationnel EM [Daudin & al. (2008)]
- En quelques mots
  - EM standard nécessite de pouvoir calculer des quantités du type $\mathbb{E}[\log \ell(\mathbf{X}^{1:2}, \mathbf{Z}; \theta) | \mathbf{X}^{1:2}, \theta']$
  - Impossible dans le cas du SBM : $p(\cdot|\mathbf{X}^{1:2}; \theta)$ non explicite
  - Variational EM : optimisation d'une borne inférieure de la log-vraisemblance

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X}}^{1:2}) = \log \ell(\mathbf{X}^{1:2}|\theta) - \mathbf{KL}[\mathcal{R}_{\mathbf{X}^{1:2}}, p(\cdot|\mathbf{X}^{1:2}; \theta)]$$

  - avec
    - **KL** is the Kullback-Leibler divergence
    - $\mathcal{R}_{\mathbf{X}^{1:2}}$ approximation de la loi conditionnelle $p(\cdot|\mathbf{X}^{1:2}; \theta)$
  - Remarque : $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X}}^{1:2}) = \log \ell(\mathbf{X}^{1:2}|\theta) \Leftrightarrow \mathcal{R}_{\mathbf{X}^{1:2}} = p(\cdot|\mathbf{X}^{1:2}; \theta)]$.
  - EM variationnel alterne entre maximisation de $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{X}}^{1:2})$ par rapport à$\mathcal{R}_{\mathbf{X}}^{1:2}$ par rapport à θ.

Propriétés théoriques

- [Bickel & al. (2013)] : consistance des estimateurs du variationnel EM dans le cas des SBM "standards"
- Adaptation au cas multiplex en cours.

Critère ICL pour choisir Q

- Critère de vraisemblance pénalisé BIC

$$\text{BIC} = \log \ell(\mathbf{X}^{1:2}; \widehat{\theta}, \mathcal{M}_Q) - \text{Pen}(\text{BIC}), \tag{3}$$

- Avec
  - Issu d'une approximation de Laplace de $\int \ell(\mathbf{X}^{1:2}; \theta, \mathcal{M}_Q) p(\theta) d\theta$ (sous conditions de régularités)
- MAIS :
  - Approximation non valable dans le cas des SBM
  - $\log \ell(\mathbf{X}^{1:2}; \widehat{\theta}, \mathcal{M}_Q)$ sans expression explicite

Critère ICL pour choisir Q

- ICL : alternative à BIC
- Repose sur une approximation de $\log \int_\theta \ell(\mathbf{X}^{1:2}, \mathbf{Z}; \theta, \mathcal{M}_Q) p(\theta) d\theta$ [**?**] :

$$\log \int_\theta \ell(\mathbf{X}^{1:2}, \mathbf{Z}; \theta, \mathcal{M}_Q) p(\theta) d\theta = \log \ell(\mathbf{X}^{1:2}, \mathbf{Z}; \widehat{\theta}, \mathcal{M}_Q) - \frac{1}{2} \left\{ Q^2(2^2 - 1) \log(2n(n-1)) + (Q-1) \log n \right\}$$

- $\text{ICL} = E_{\mathbf{Z}|\mathbf{X}^{1:2}, \widehat{\theta}, \mathcal{M}_Q} \left[\log \ell(\mathbf{X}^{1:2}, \mathbf{Z}; \widehat{\theta}, \mathcal{M}_Q) \right] - \text{Pen(ICL)}$
- Remarques
  - Critère de vraisemblance complète conditionnelle pénalisée :

$$E_{\mathbf{Z}|\mathbf{X}^{1:2}, \widehat{\theta}, \mathcal{M}_Q} \left[\log \ell(\mathbf{X}^{1:2}|\mathbf{Z}, \widehat{\theta}, \mathcal{M}_Q) \right] = \log \ell(\mathbf{X}^{1:2}|\widehat{\theta}, \mathcal{M}_Q) - H(\mathbf{Z}|\mathbf{X}^{1:2}, \widehat{\theta}, \mathcal{M}_Q) \quad (4)$$

  - avec $H(\mathbf{Z}|\mathbf{X}^{1:2}, \widehat{\theta}, \mathcal{M}_Q)$ entropie de $p(\mathbf{Z}|\mathbf{X}^{1:2}, \widehat{\theta}, \mathcal{M}_Q)$,
  - Pen(ICL) pénalise la complexité du modèle, et entropie encourage les classifications avec des groupes bien séparés.

Calcul de ICL

- $$E_{\mathbf{Z}|\mathbf{X}^{1:2},\widehat{\theta},\mathcal{M}_Q}\left[\log \ell(\mathbf{X}^{1:2},\mathbf{Z},\widehat{\theta},\mathcal{M}_Q)\right] = \sum_{i,j=1,i\neq j}^{n}\sum_{q,l=1}^{Q} E_{\mathbf{Z}|\mathbf{X}^{1:2},\widehat{\theta},\mathcal{M}_Q}\left[\mathbb{I}_{Z_i=q,Z_j=l}\right]\log \widehat{\pi}_{ql}^{(X_{ij}^{1:2})} + \sum_{i=1}^{n}\sum_{q=1}^{Q} E_{\mathbf{Z}|\mathbf{X}^{1:2},\widehat{\theta},\mathcal{M}_Q}\left[\mathbb{I}_{Z_i=q}\right]\log \widehat{\alpha}_q$$

- Utilisation de l'approximation variationnelle :

$$E_{\mathbf{Z}|\mathbf{X}^{1:2},\widehat{\theta},\mathcal{M}_Q}\left[\mathbb{I}_{Z_i=q}\right] = \widehat{\tau}_{iq}, \qquad E_{\mathbf{Z}|\mathbf{X}^{1:2},\widehat{\theta},\mathcal{M}_Q}\left[\mathbb{I}_{Z_i=q,Z_j=l}\right] = \widehat{\tau}_{iq}\widehat{\tau}_{jl}$$

Résultats : probabilités marginales et conditionnelles

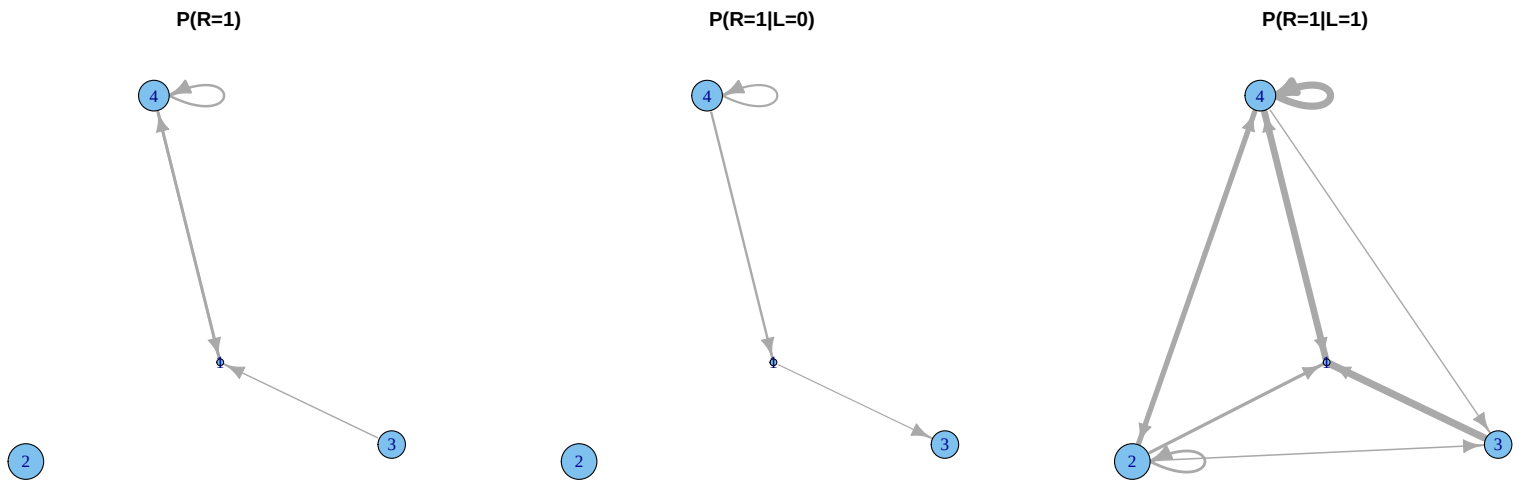


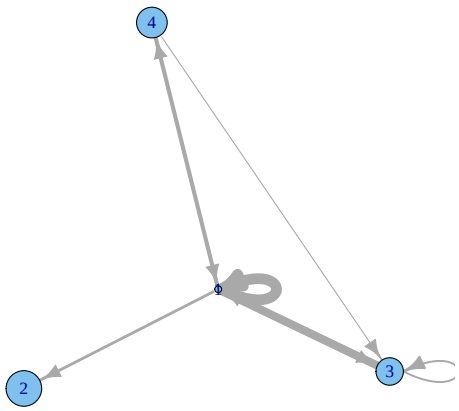
Figure: Marginal probabilities of Researcher connections between and within blocks (top) and probabilities of Researcher connections between and within blocks conditionally on absence (bottom left-hand-side) or presence (bottom right-hand-side) of Lab connection. Vertex size is proportional to the block size. Edge width is proportional to the probabilities of connection ; if this probability is smaller than 0.1, edges are not displayed.

Résultats : commentaires

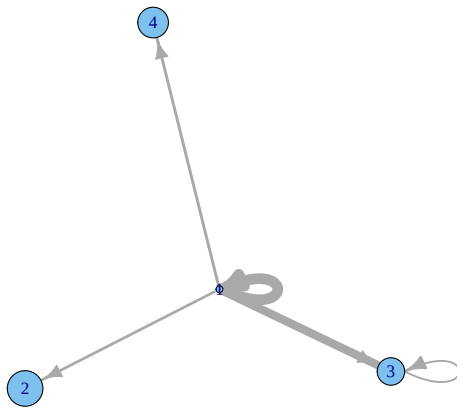
- L'existence d'un échange de ressources entre laboratoires augmente clairement la probabilité de connexion entre chercheurs
- Particulièrement vrai pour groupe 2 (avec groupe 4).
- Idem au sein du groupe 4
- Le groupe 3 est le moins affecté par l'existence de relations entre laboratoires
- Les chercheurs ne profitent pas de la même façon de l'insertion de leur laboratoire dans le tissu institutionnel
- Vrai aussi au niveau des laboratoires

Résultats : probabilités marginales et conditionnelles

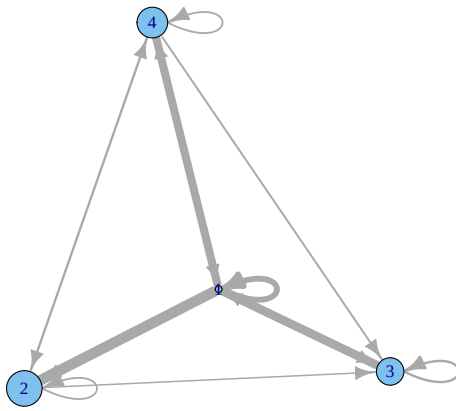
P(L=1)



P(L=1|R=0)



P(L=1|R=1)



Résultats : description des 4 blocs

(a)

	block
1	2
2	48
3	19
4	26

(b)

	not idf	idf
1	1	1
2	26	22
3	10	9
4	11	15

(c)

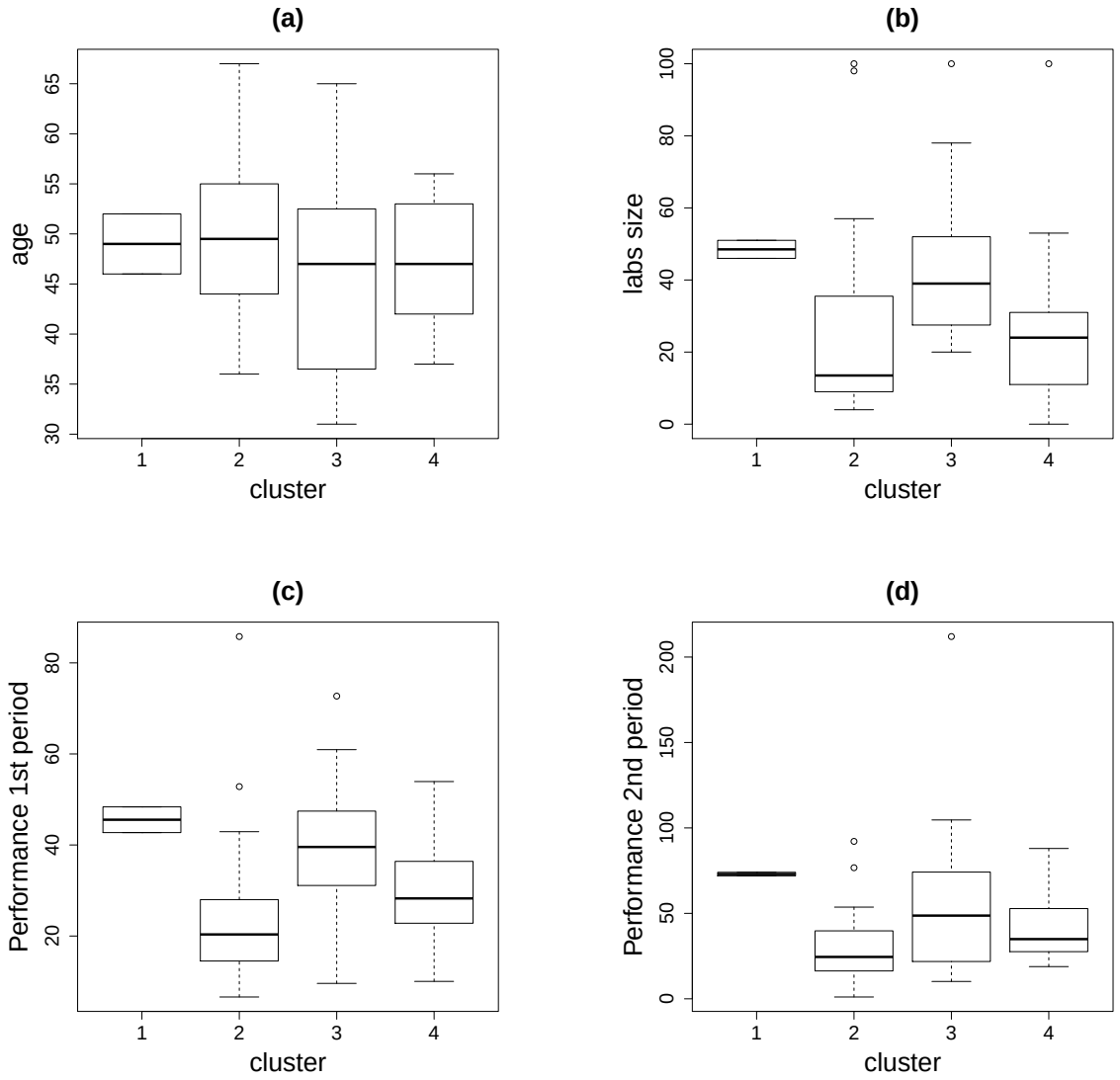
	not director	director
1	1	1
2	21	27
3	12	7
4	12	14

(d)

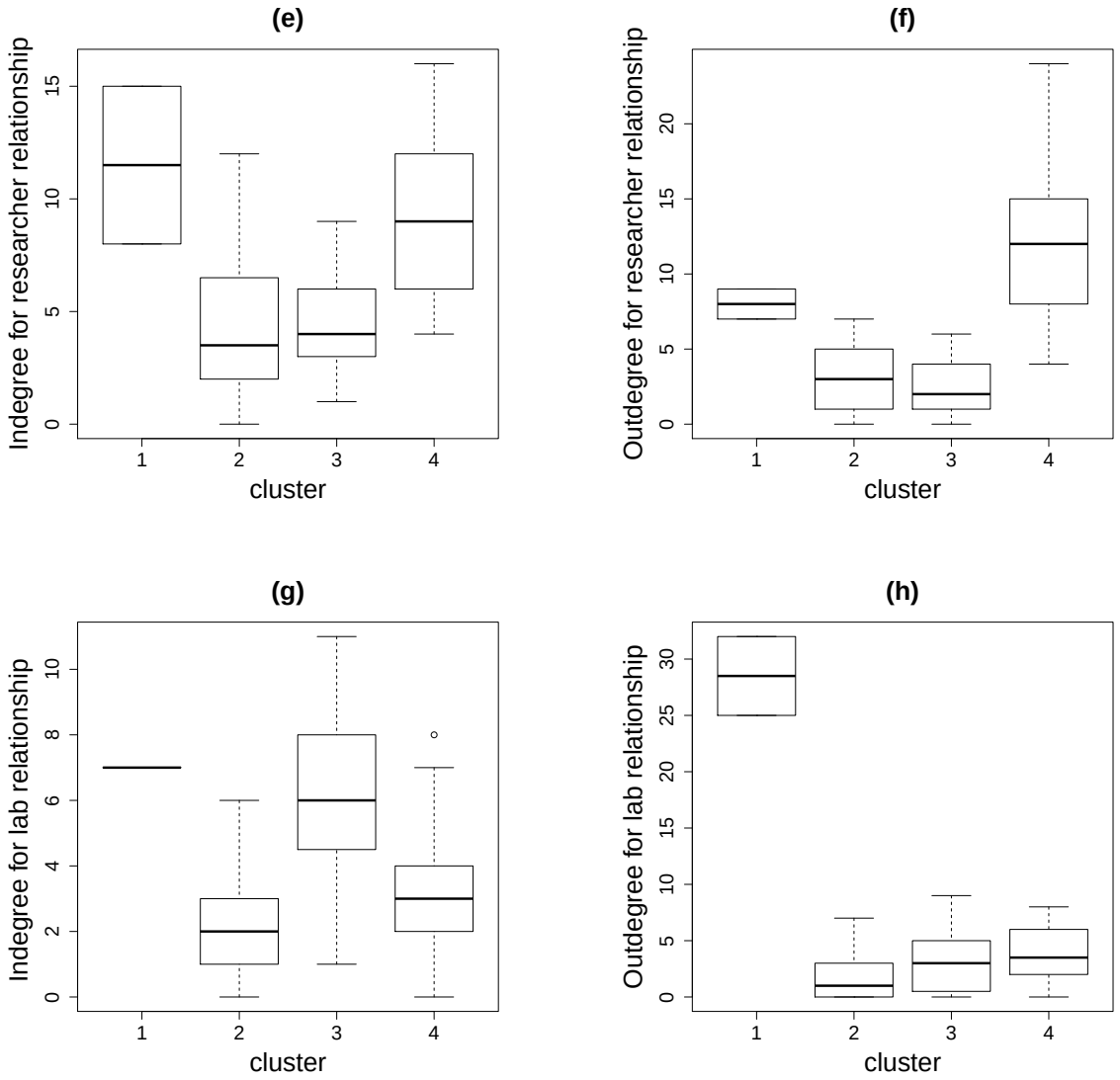
	PH	Su	He	ST	FPh	FMo	FMoG
1	0	0	0	0	0	1	1
2	12	5	6	10	7	7	1
3	0	1	3	1	1	10	3
4	6	1	7	4	2	2	4

PH : Public Health ;Su : Surgery ; He : Hematology ; ST : Solid Tumours ; FPh : Fundamental pharmacology ; FMo : Fundamental molecular research ; FMoG : Fundamental molecular genetic research

Résultats : descriptions des 4 blocs



Résultats : descriptions des 4 blocs



Commentaires

- **Bloc 1** : petit groupe, dans de gros laboratoires, gros in et out degrees (chercheurs et laboratoires). Profils relationnels similaires car fournisseurs de souris.
- **Bloc 2** : petits in et out degree (chercheurs et laboratoires), chercheurs un peu plus âgés, dans de plus petits laboratoires. Hétérogènes en terme de spécialités, plus faibles performances. Individus les plus affectés par les connexions des laboratoires
- **Bloc 3** : jeunes chercheurs en recherche fondamentale. Faibles in et out degree dans des laboratoires avec des forts indegrees et out-degrees moyens. 70% sont dans les plus gros publiants. Peu affectés par les connexions des laboratoires
- **Bloc 4** : beaucoup d'hématologistes. In et out degree moyens dans des laboratoires avec de faibles in et out degrees.

Conclusions

- Une façon de gérer le multiniveau dans les graphes
- Mis en évidence des patterns de connexions intéressants sur ce jeu de données
- Autre application : étude conjointe des relations de conseils et de concurrence sur ce même jeu de données.

References

Barbillon, P., Donnet, S., Lazega, E., Bar-Hen, A.

Stochastic Block Models for Multiplex networks : an application to networks of researchers.

http ://arxiv.org/abs/1501.06444

Daudin, J. -J., Picard, F. and Robin, S. (2008)

A mixture model for random graphs

Statistics and Computing, 18(2) :173–183.

Bickel, Peter and Choi, David and Chang, Xiangyu and Zhang, Hai (2013)

Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels

The Annals of Statistics, 4 :1922–1943.

Lazega, E., Jourda, M.-T., Mounier, L., and Stofer, R. (2008).

Catching up with big fish in the big pond ? multi-level network analysis through linked design.

Social Networks, 30(2) :159 – 176.

Leger, J.-B. (2014).

Wmixnet : Software for clustering the nodes of binary and valued graphs using the stochastic block model.

ArXiv e-prints.

Snijders, T. A. B. and Nowicki, K. (1997).

Estimation and prediction for stochastic blockmodels for graphs with latent block structure.

J. Classification, 14(1) :75–100.