Prédire une relation sociale - Introduction aux modèles Exponential Random graph (ERGM)

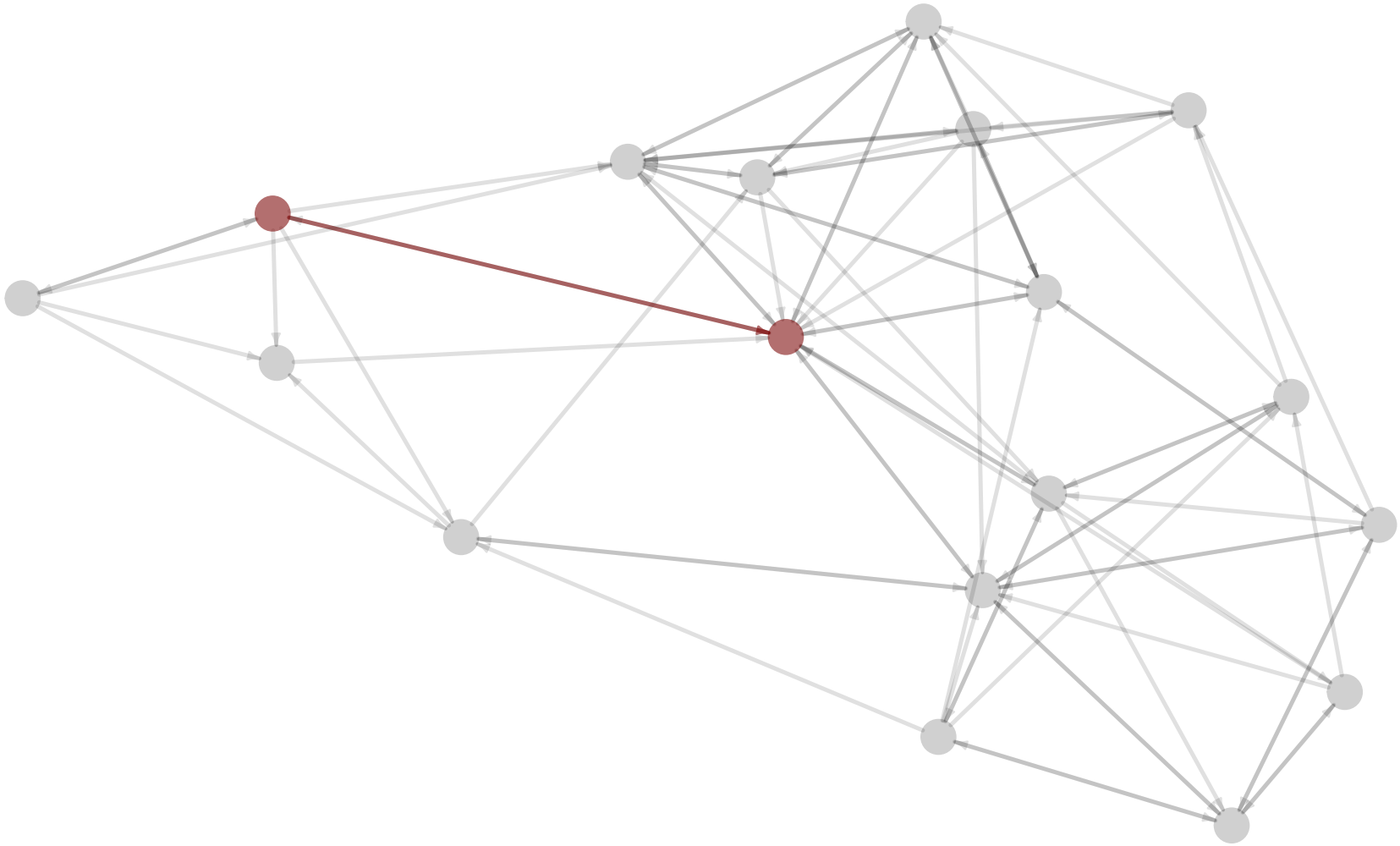
Victorin Luisier
Avec: Olivier Renaud & Eric Widmer

Université de Genève et Université à Distance, Suisse

Rencontres "Méthodes et Recherche"
Organisé par FORS/IMA/MISC
31 mai 2011

Thématique

But : Comprendre ce qui pousse 2 individus à entretenir une certaine relation sociale, sachant qu'elle s'inscrit dans un réseau de relations plus large.



Plan

De quel réseau de relations parle-t-on ?

Nous nous focalisons dans cette présentation sur un type de réseau de relations bien particulier :

- Réseau de relations est **délimité** (on s'intéresse aux relations à l'intérieur d'une entité).
- Réseau n'est **mesuré qu'une fois** (pas de mesures répétées).
- La mesure de la **relation est dichotomique** (1=relation, 0=pas de relation).
- La relation peut-être **dirigée** (*ex : Jules cite Ben comme ami*) ou **non-dirigée** (*ex : Ben et Jules sont amis.*)

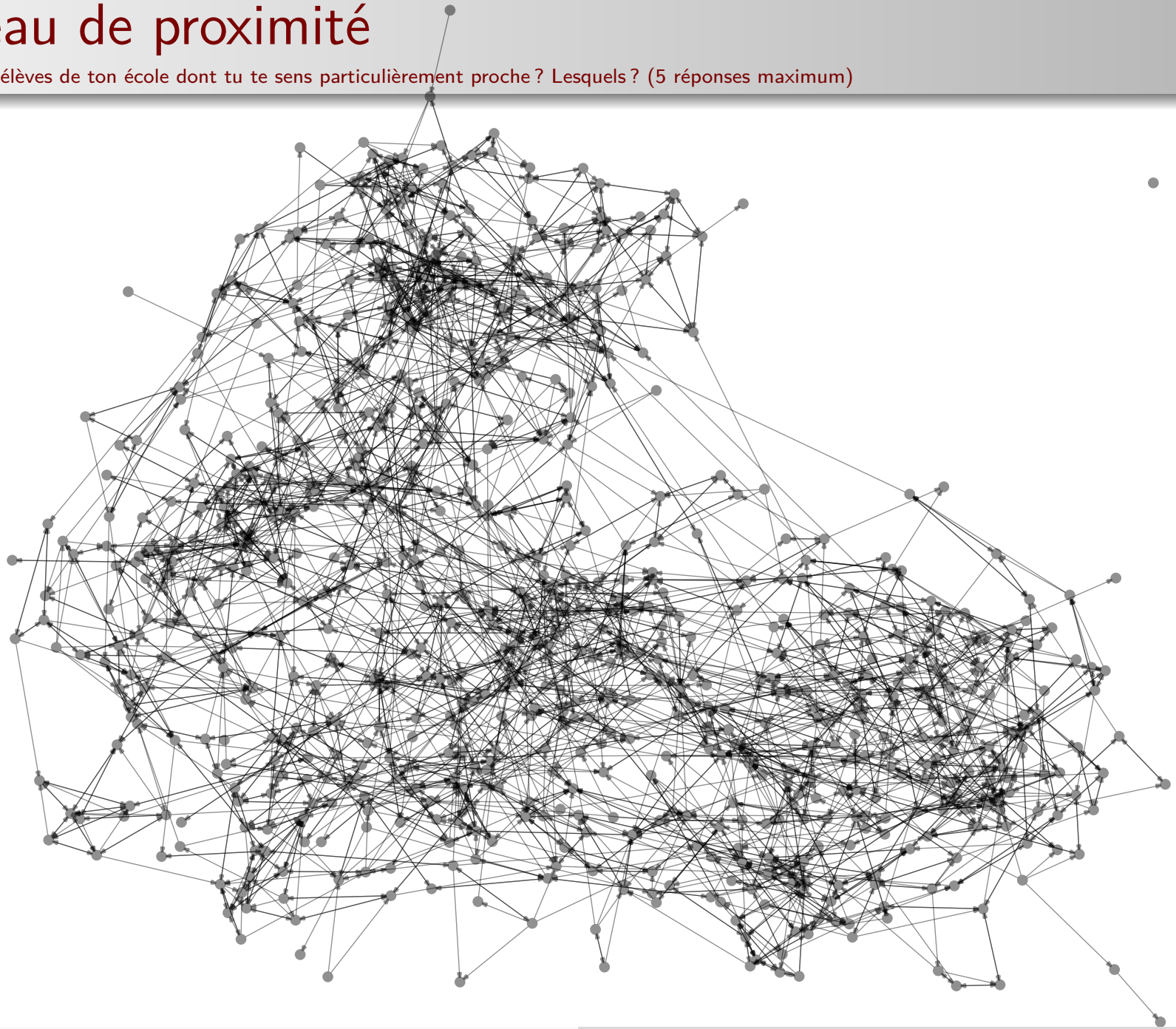
Les sujets (noeuds) ne sont pas forcément des individus : groupes d'individus, concepts, etc.

Exemple de réseau

- Enquête sur les inégalités de genre au collège (Luisier & Widmer, 2009).
- Presque tous les élèves d'un collège genevois (gymnase) : 410 femmes, 340 hommes, entre 15 et 19 ans.
- Relations : Chaque élève a nommé entre 0 et 5 autres élèves du collège avec qui il se sentait particulièrement proche, avec qui il était en conflit, avec qui il se comparait du point de vue des notes, etc.
- Autres variables : sexe, année de formation, classe, féminité et masculinité (BSRI), etc.
- BSRI : 1 facteur "féminité" (tendresse et sensibilité à autrui), 1 facteur "masculinité" (confiance en soi, athlétique et leadership). On utilise les scores latents.

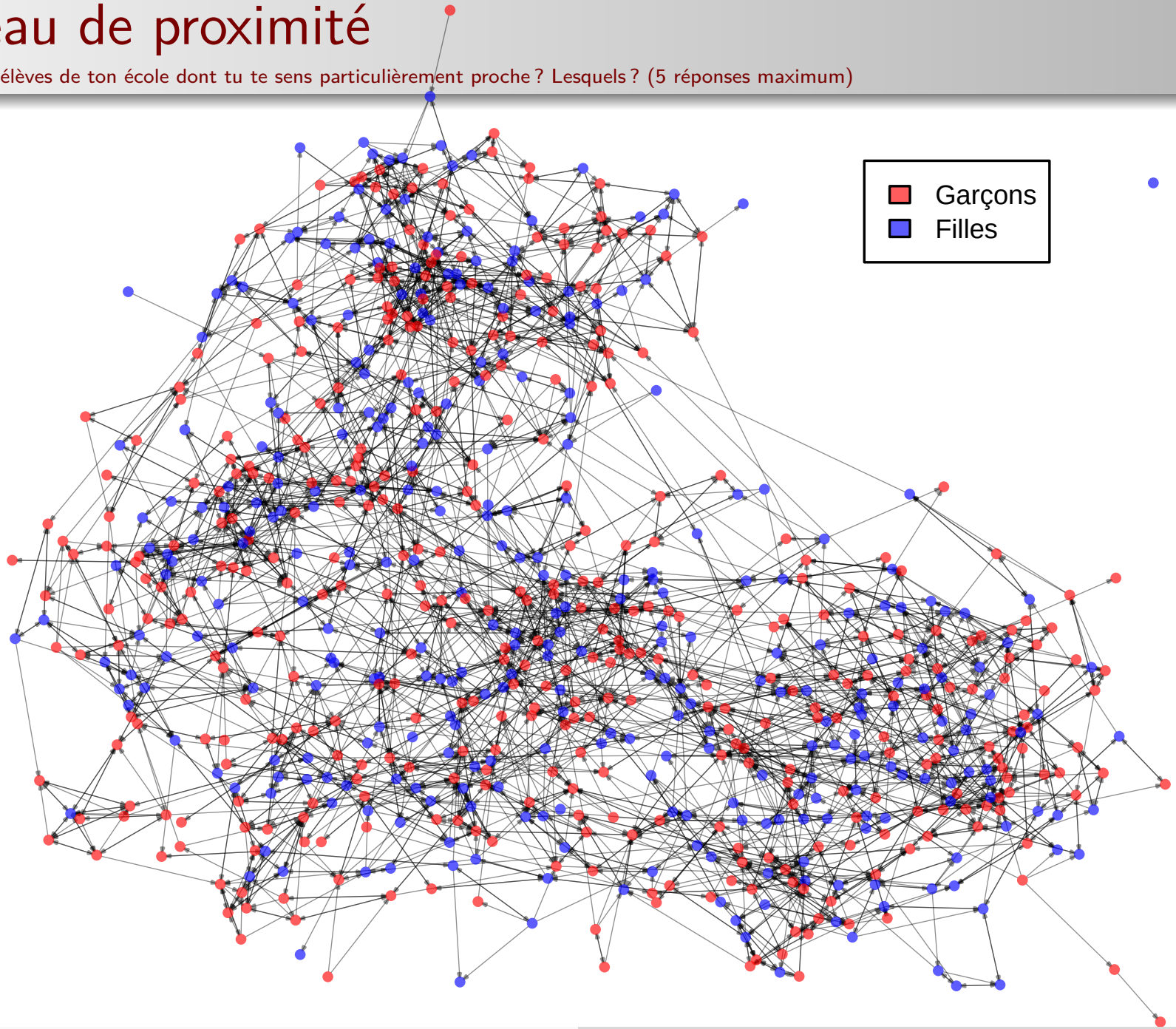
Réseau de proximité

Y a-t-il des élèves de ton école dont tu te sens particulièrement proche ? Lesquels ? (5 réponses maximum)



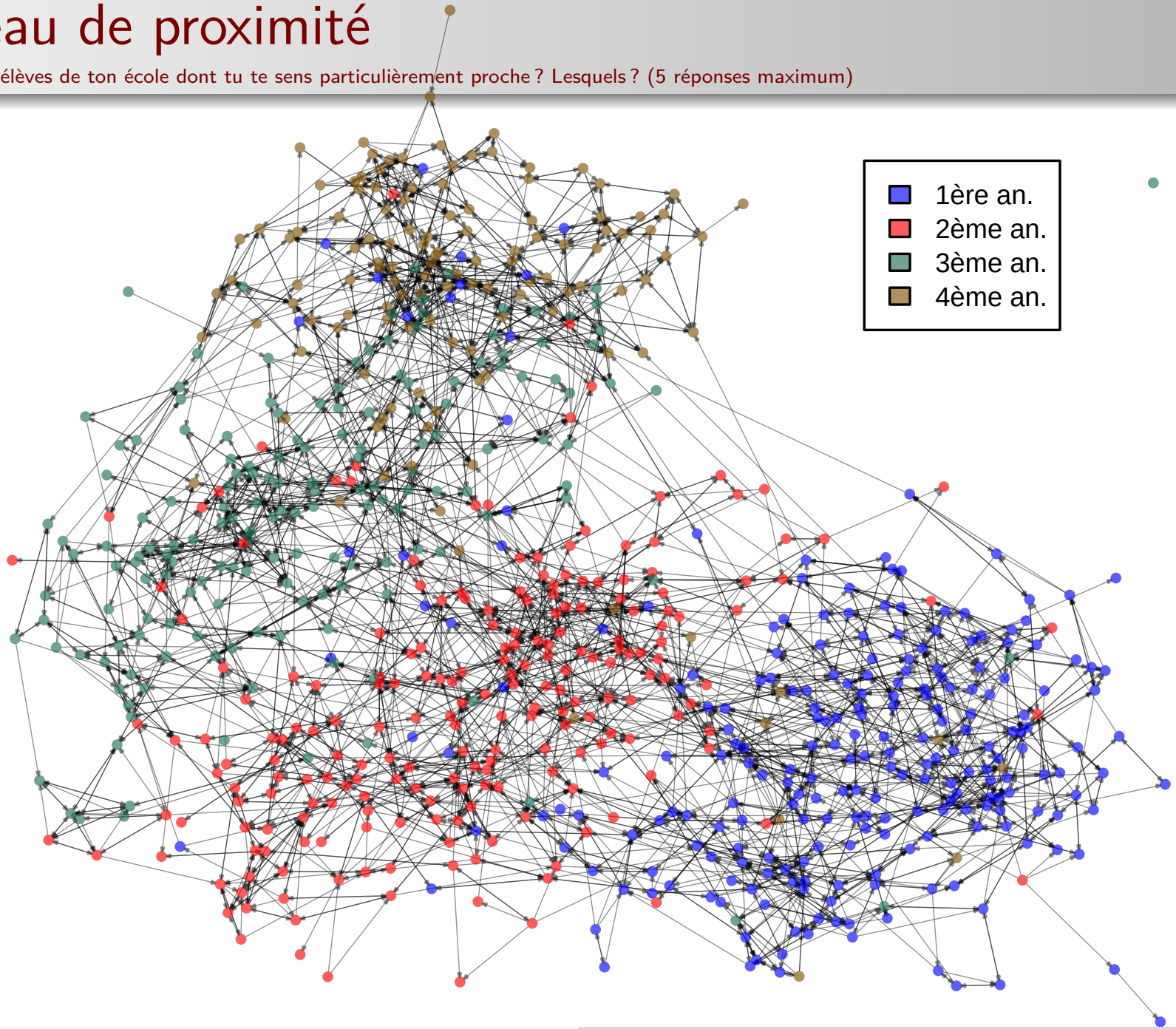
Réseau de proximité

Y a-t-il des élèves de ton école dont tu te sens particulièrement proche ? Lesquels ? (5 réponses maximum)

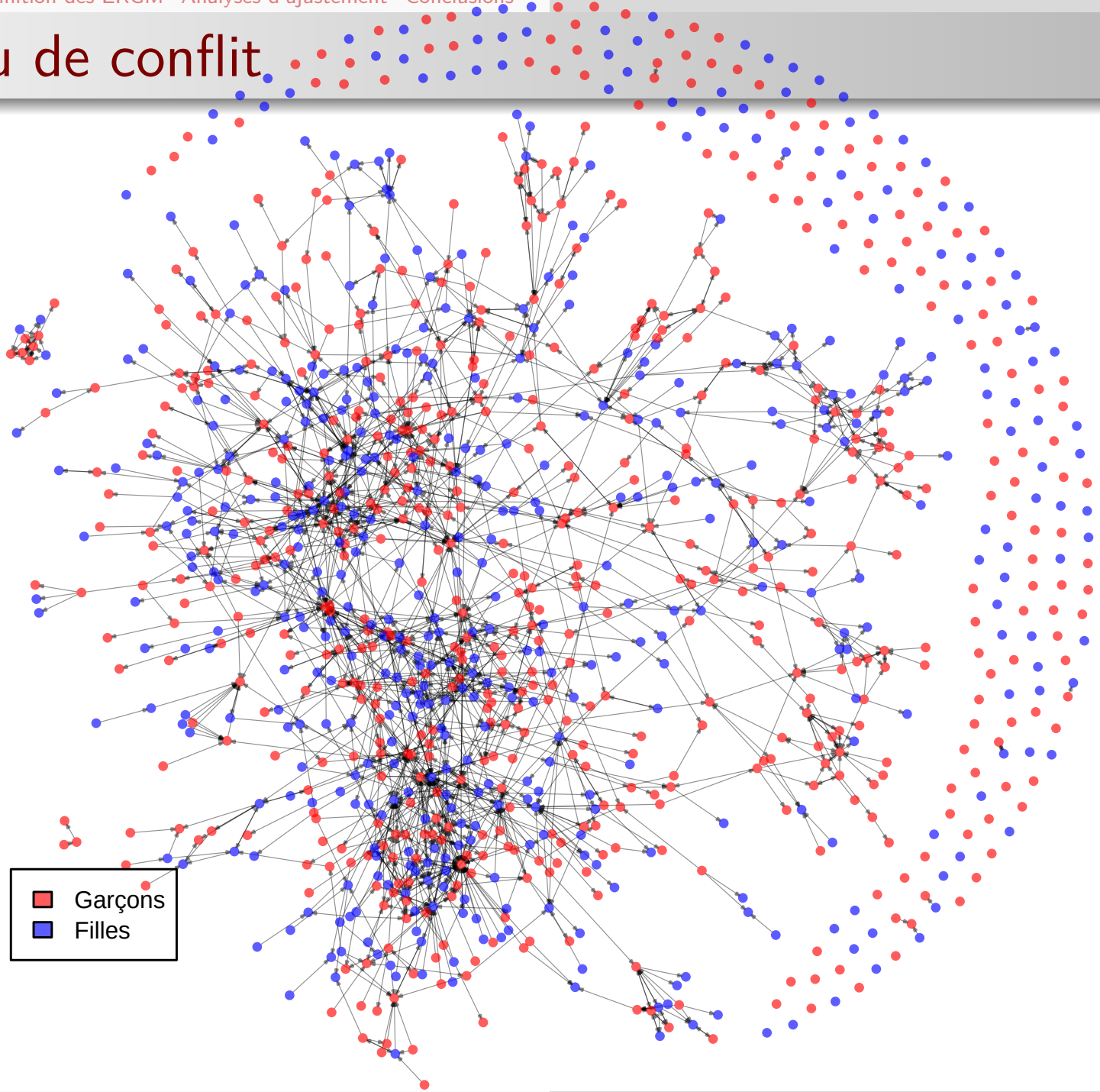


Réseau de proximité

Y a-t-il des élèves de ton école dont tu te sens particulièrement proche ? Lesquels ? (5 réponses maximum)



Réseau de conflit



Réseau de proximité

Y a-t-il des élèves de ton école dont tu te sens particulièrement proche ? Lesquels ? (5 réponses maximum)

	Filles	Garçons
Filles →	82.57	17.43
Garçons →	25.94	74.06

TABLE: Homophilie de sexe (% de ligne)

	0 ami	1 ami	2 amis	3 amis	4 amis	5 amis
Filles →	0.98	1.46	6.34	12.20	20.49	58.54
Garçons →	4.41	3.24	4.71	12.65	20.88	54.12

TABLE: Sociabilité en fonction du sexe (% de ligne)

	Cité 0x	1x	2x	3x	4x	5x	etc.
Filles →	2.93	6.83	12.93	17.07	17.07	12.93	...
Garçons →	6.18	9.41	14.41	16.76	18.82	9.71	...

TABLE: Popularité en fonction du sexe (% de ligne)

Réseau de proximité

Y a-t-il des élèves de ton école dont tu te sens particulièrement proche ? Lesquels ? (5 réponses maximum)

	$1^{ère}$	$2^{ème}$	$3^{ème}$	$4^{ème}$
$1^{ère} \rightarrow$	88.46	9.40	1.43	0.72
$2^{ème} \rightarrow$	8.91	81.40	8.57	1.13
$3^{ème} \rightarrow$	1.40	12.85	75.56	10.20
$4^{ème} \rightarrow$	0.56	1.67	9.67	88.10

TABLE: Homophilie d'année (% de ligne)

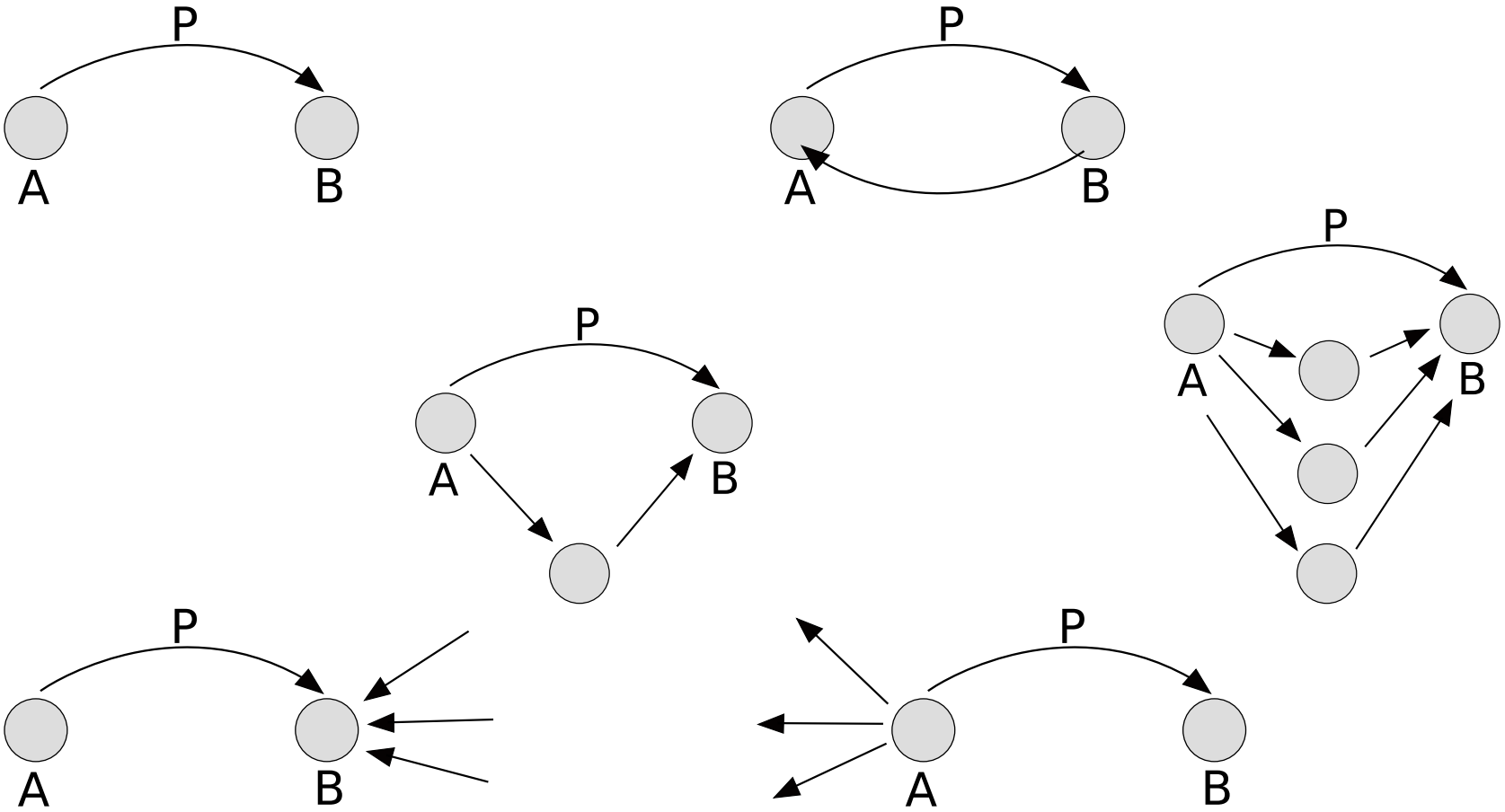
Limites des approches traditionnelles

Ces graphiques et statistiques descriptives ne permettent pas de tester une théorie → Besoin d'un modèle statistique.

Vu que la VD est dichotomique (0 = pas de relation vs 1 = relation) → Régression logistique ?

Problème : L'existence d'une relation entre deux personnes dépend aussi d'autres relations du réseau → structure de dépendance dyadique.

Des relations interdépendantes



→ Régression logistique pas possible (postulats).
→ Vive ERGM !

Les modèles ERGM (Exponential Random Graph Models)

ERGM : modéliser la macrostructure d'un réseau de relations à l'aide de microprocessus d'intérêt.

Exemple : Modéliser le réseau d'amitié du collège à l'aide de :

- Homophilie de sexe, d'année et de classe.
- Tendance que la féminité et masculinité des élèves influencent positivement leur popularité et sociabilité.
- Tendance que les garçons soient plus sociables mais moins populaires.
- Tendance à la réciprocité, etc.

Dans le cadre des ERGM, les **microprocessus d'intérêts** :

- Forces d'attraction locales.
- peuvent aussi introduire une dépendance dyadique (liens interdépendants).
- Virtuellement illimités (*ex : les filles sont plus homophiles d'année que les garçons*)

Définition Formelle

- Y est la matrice nxn des relations entre n individus (y, la matrice observée).
- La dyade Y_{ij} vaut soit 0 (pas de relation), soit 1 (relation). $Y_{ii} \doteq$ n.d.
- Y peut-être dirigée (noeud i cite un lien avec j) ou non-dirigée (i et j sont en lien $\rightarrow Y_{ij} \doteq Y_{ji}$).

Ex. y dirigé :

	s1	s2	s3	s4	...
s1	-	0	1	1	...
s2	0	-	1	0	...
s3	0	0	-	1	...
s4	0	1	1	-	...
...	...	...	...	...	...

Ex. y non-dirigé :

	s1	s2	s3	s4	...
s1	-	0	1	1	...
s2	0	-	1	0	...
s3	1	1	-	1	...
s4	1	0	1	-	...
...	...	...	...	...	...

Définition Formelle

$$P(Y = y) = \frac{\exp\{\eta_1 Z_1(y,x) + \eta_2 Z_2(y,x) + ... + \eta_p Z_p(y,x)\}}{\psi(\boldsymbol{\eta})} \quad (1)$$

- Les $\boldsymbol{\eta}$ sont les paramètres à estimer.
- Les $\boldsymbol{Z}(y,x)$ sont les prédicteurs et dépendent des microprocessus modélisés. [[Techniquement, un prédicteur c'est la somme des situations où le microprocessus se retrouve dans le réseau)]]
- $\psi(\boldsymbol{\eta})$ est une constante de normalisation permettant de borner la distribution du modèle à 1.

$$\psi(\boldsymbol{\eta}) = \sum_{g \in \gamma} \exp\{\eta_1 Z_1(g,x) + \eta_2 Z_2(g,x) + ...\eta_p Z_p(g,x)\}$$

- $\psi(\boldsymbol{\eta})$ est généralement impossible à calculer $\rightarrow$ Les paramètres $\boldsymbol{\eta}$ sont estimés par une méthode **MCMC** (Monte Carlo Markov Chain).

[[Estimer les paramètres par MCMC]]

- La plupart du temps $\psi(\boldsymbol{\eta})$ est incalculable (somme sur γ, i.e. l'ensemble des réseaux possibles de taille $n * n$. Si γ n'est pas contraint, il contient $2^{n*(n-1)}$ éléments) $\rightarrow$ logvraisemblance pas maximisable.
- les $\boldsymbol{\eta}$ estimés par le biais d'une méthode MCMC (Monte Carlo Markov Chain) :
  1. Choix de paramètres de départ $\boldsymbol{\eta}_0$ (pseudovraisemblance)
  2. Simule une gd quantité d'échantillons "pseudoindépendants" sous $\boldsymbol{\eta}_0$ (Metropolis-Hastings)
  3. maximise une approximation du ratio de logvrais : $\ell(\boldsymbol{\eta}) - \ell(\boldsymbol{\eta}_0)$
  4. une estimation $\boldsymbol{\eta}$ est trouvée et devient le nouveau $\boldsymbol{\eta}_0$
  5. On refait points 2-4 plusieurs fois pr améliorer l'estimation de $\boldsymbol{\eta}$

Interprétation des paramètres

Pour être sûr que les paramètres ont été estimés correctement, il faudrait vérifier les diagnostics MCMC avant d'interpréter.

Si l'estimation MCMC s'est bien déroulée :

- un paramètre positif et significatif indique que le microprocessus d'intérêt a une influence positive sur la probabilité d'une relation.
- un paramètre négatif et significatif indique que le microprocessus d'intérêt a une influence négative sur la probabilité d'une relation.

Interprétation des paramètres

Pour notre exemple sur le réseau d'amitié du collège :

	η estimé	Err. standard	p-value
Homo. classe	0.88445	0.03927	***
Homo. année	3.06393	0.13639	***
Homo. sexe	0.84572	0.03779	***
Popularité feminité	0.17719	0.01738	***
Popularité masculinité	0.09454	0.01151	***
Popularité sexe (Garçon)	-0.14355	0.01479	***
Sociabilité feminité	0.10689	0.02251	***
Sociabilité masculinité	-0.06608	0.01582	***
Sociabilité sexe (Garcon)	0.15226	0.01632	***
Réciprocité	5.34121	0.02138	***
Edges (densité)	-11.77477	0.13929	***
Homo. gp de passation	0.29825	0.03926	***
Homo. année (+/-1 an)	1.84928	0.14155	***
Cite 1 ami	0.94486	0.10831	***
Cite 2 amis	4.32663	0.06720	***
Cite 3 amis	7.80147	0.04644	***
Cite 4 amis	11.25402	0.03421	***
Cite 5 amis	15.37701	0.02569	***

[[Interpréter plus finement les paramètres]]

On peut interpréter plus finement les paramètres en réécrivant le modèle pour une dyade Y_{ij} :

$$\log\left\{\frac{P(Y_{ij}=1|Y_{ij}^c=y_{ij}^c)}{P(Y_{ij}=0|Y_{ij}^c=y_{ij}^c)}\right\} = \eta_1\Delta\{Z_1(x,y)\}_{ij}+\ldots+\eta_p\Delta\{Z_p(x,y)\}_{ij} \quad (2)$$

Où :

- Y_{ij}^c renvoie à toutes les dyades sauf Y_{ij}
- $\Delta\{\boldsymbol{Z}(y,x)\}_{ij}$ sont les **statistiques de changement**, i.e. de combien changent $\boldsymbol{Z}(y,x)$ lorsque Y_{ij} et commuté de 0 à 1.

Analyses d'ajustement (Goodness of fit)

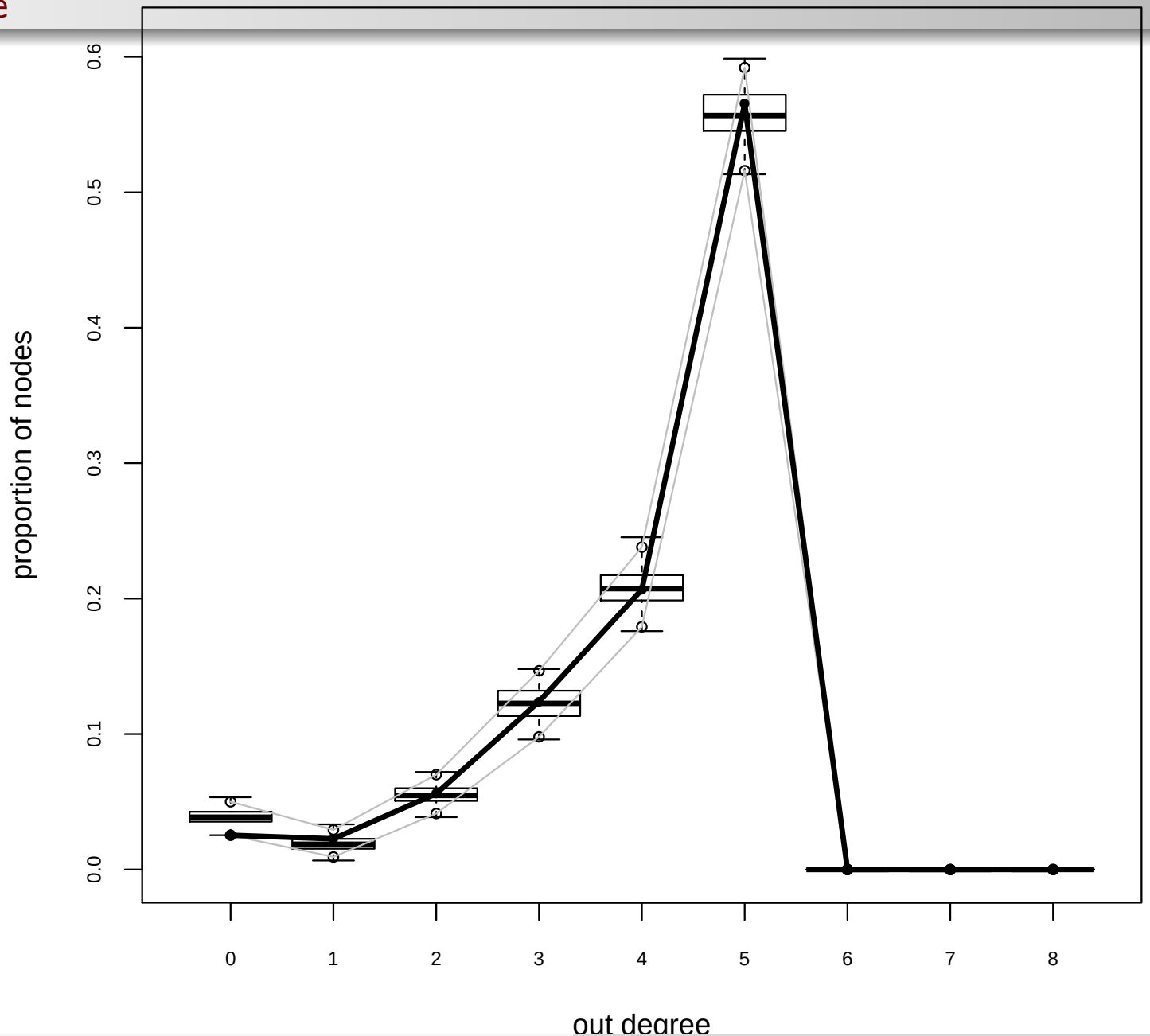
Question : Le modèle estimé explique-t-il convenablement la macrostructure du réseau observé ?

Déroulement :

1. A partir du modèle estimé, on simule (par MCMC) des réseaux de même taille que notre réseau observé.
2. On compare ces réseaux simulés au réseau observé sur certaines caractéristiques d'intérêt $\rightarrow$ Plus les réseaux simulés sont proches de notre réseau observé, mieux le modèle s'ajuste notre réseau de relations observé.

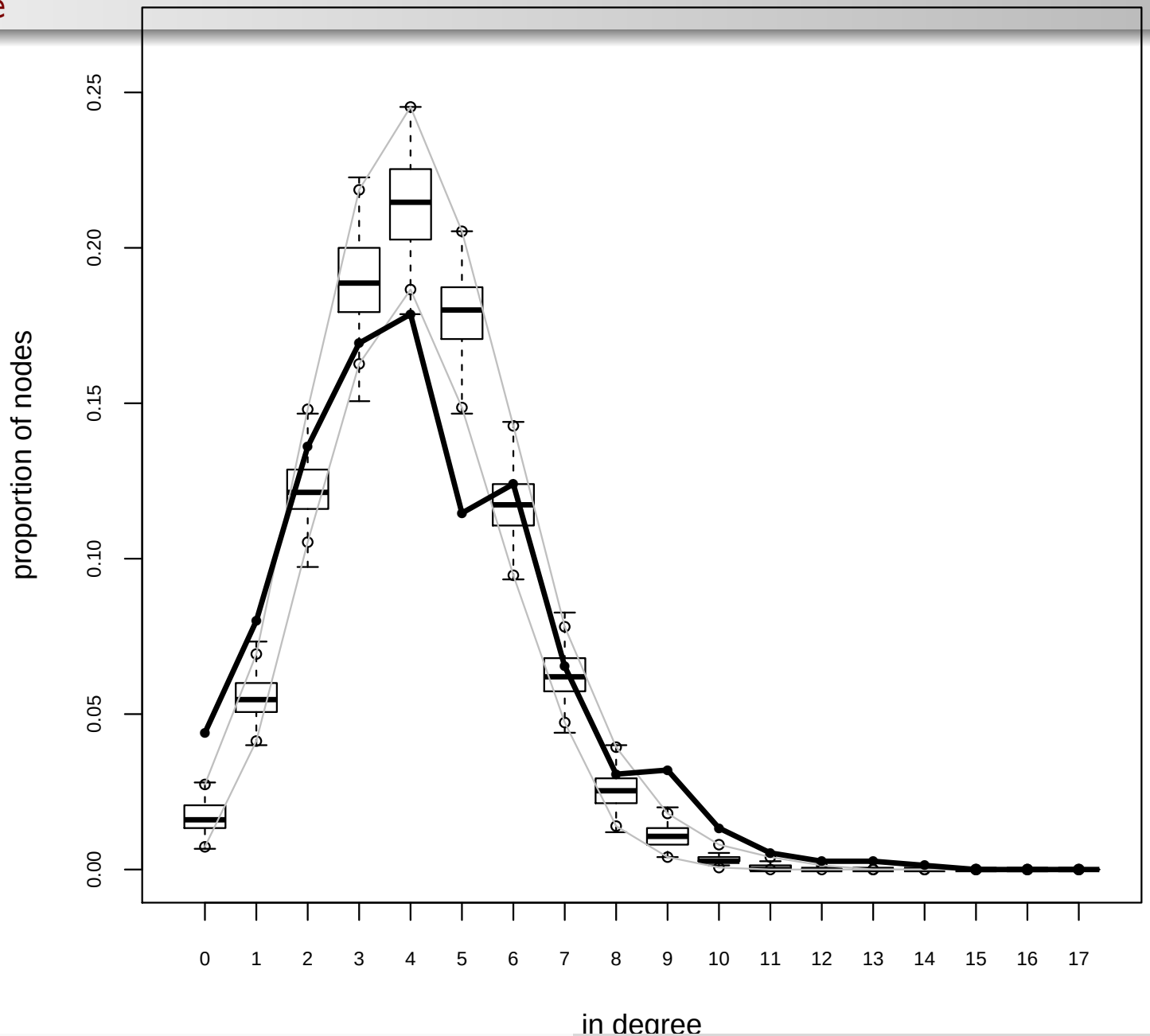
Analyse d'ajustement - exemple

Sociabilité



Analyse d'ajustement - exemple

Popularité



ERGM, applicable en sciences humaines ?

Les bons points :

- Permet de tester l'effet de n'importe quel microprocessus d'intérêt sur la relation.
- Les microp. peuvent aussi introduire une interdépendance des liens.
- Avec MCMC, les analyses GOF sont une "partie de plaisir".

Les points moins évidents :

1. MCMC pour ERGM $\rightarrow$ complexe à comprendre, Tps de calcul.
2. Microprocessus $\rightarrow$ Seuls les principaux existent. Les autres doivent être codés par le chercheur.
3. Pas d'analyse multiniveau (multilevel P2) ou de mesures répétées (SIENA).

Mais... Véritable mine d'or pour de nouvelles questions de recherches.

Références

Le projet STATNET regroupe un ensemble de bibliothèques R. STATNET c'est LE programme pour effectuer des analyses ERGM :

- Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., & Morris, M. (2008). *A statnet Tutorial*. Journal of statistical software, 24(9), 1-27.
- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). *statnet : Software tools for the representation, visualization, analysis and simulation of network data*. Journal of statistical software, 24(1).
- Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). *ergm : A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks*. Journal of Statistical Software, 24(3), 1-29.