Analyse exploratoire de graphes d'infection

Fabrice Rossi

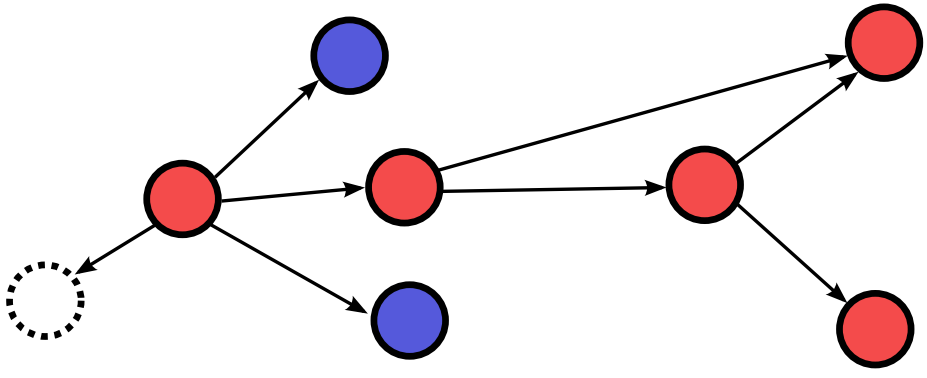
avec Stéphan Clémençon, Hector De Arazoza, et Viet-Chi Tran

ANR Viroscopy (ANR-08-SYSC-016-03)

Télécom ParisTech, Universidad de la Habana, and Université Lille 1

Graphe d'infection

- suivi du VIH/SIDA à Cuba de 1986 à 2004
- suivi d'infection étendu : partenaires sexuels durant les deux années avant une détection



- nombreuses caractéristiques pour chaque patient : genre, orientation sexuelle, date de naissance, etc.
- objectifs d'étude :
  - effets des caractéristiques sur la propagation
  - efficacité du suivi d'infection
  - etc.

Données volumineuses

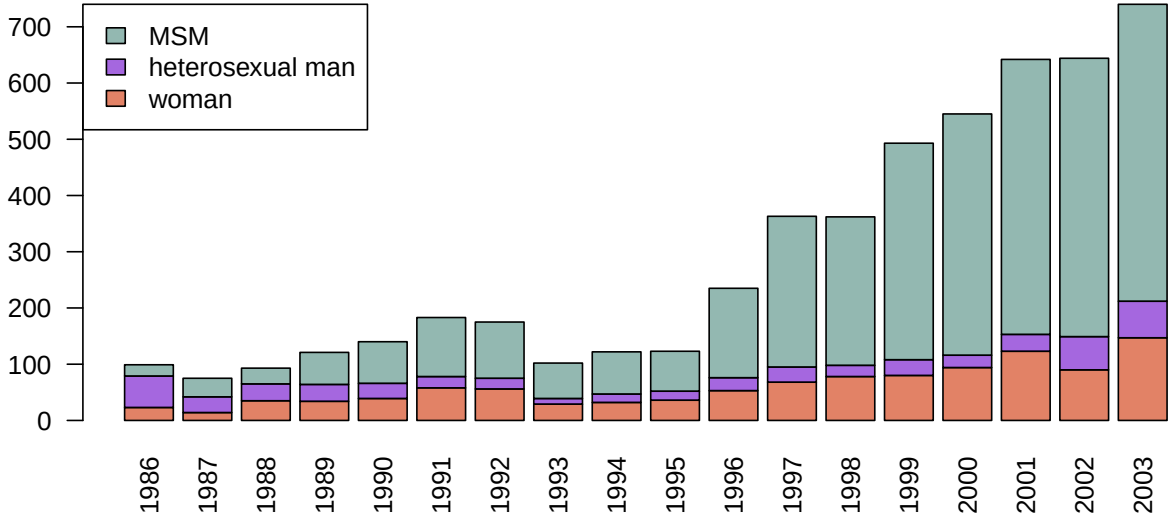
- base de données volumineuse :
  - 5 389 patients décrits par une quinzaine de variables
  - 4 073 relations (graphe assez peu dense)
  - 2 386 patients dans une même composante connexe du graphe (3 168 relations dans cette composante)
- bases « comparables » :
  - Rothenberg et al., 1995
    - étude *Colorado Springs*, suivi de contact
    - 2 200 personnes (quelques VIH+), 965 dans la plus grande composante connexe
  - Wylie et Jolly, 2001
    - étude *Manitoba*, suivi d'infection
    - 4 544 personnes (MST), 82 dans la plus grande CC
  - Bearman, Moody et Stovel, 2004
    - sexualité des adolescents américains (pas de MST), suivi de contact
    - 573 personnes, 288 dans la plus grande CC

Aspects macroscopiques

- orientation sexuelle

	population	GCC
femmes	0.21	0.20
hommes hétéros	0.11	0.05
hommes bisexuels	0.69	0.76

- « recrutement »

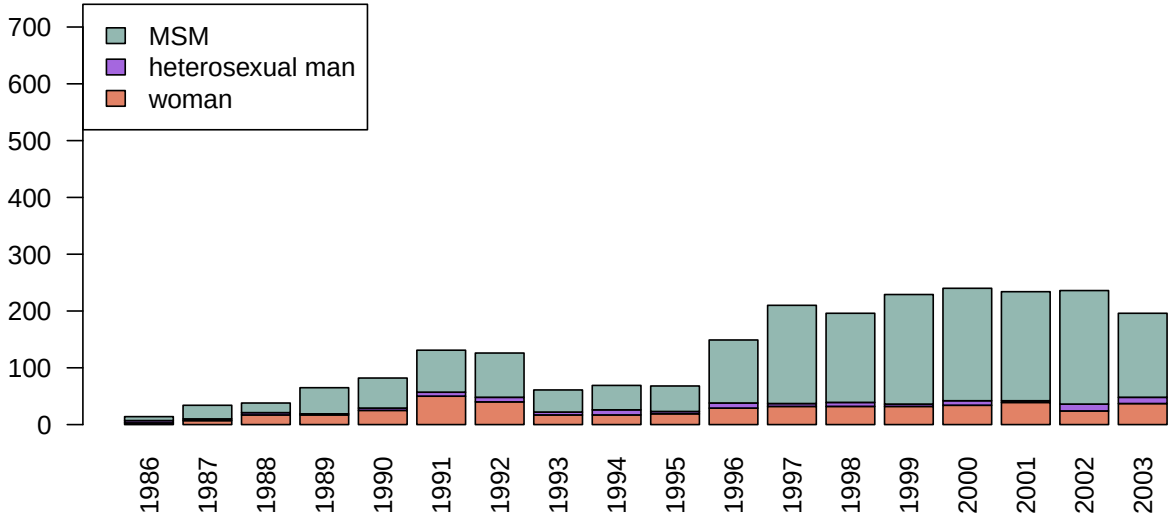


Aspects macroscopiques

- orientation sexuelle

	population	GCC
femmes	0.21	0.20
hommes hétéros	0.11	0.05
hommes bisexuels	0.69	0.76

- « recrutement »

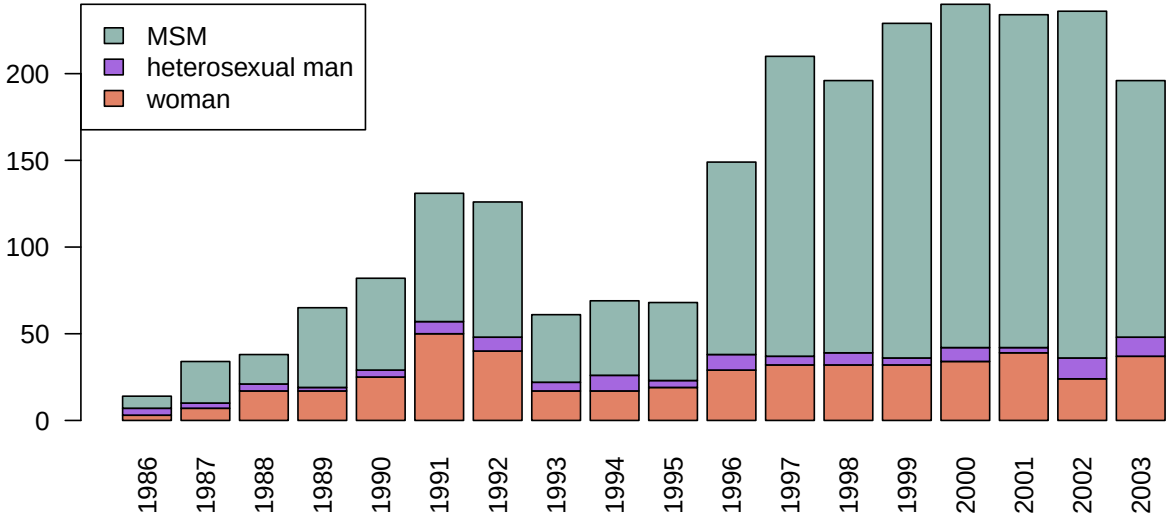


Aspects macroscopiques

- orientation sexuelle

	population	GCC
femmes	0.21	0.20
hommes hétéros	0.11	0.05
hommes bisexuels	0.69	0.76

- « recrutement »



Graphes aléatoires

- modélisation macroscopique du réseau
- modèle de configuration :
  - degrés des noeuds fixés
  - distribution uniforme sur les graphes (simples) avec ces degrés
  - principe d'appariement de pattes
- résultats (asymptotiques) connus sur :
  - composante connexe « géante »
  - percolation
  - etc.

Transition de phase

- ingrédients :
  - p_k distribution des degrés, $z = \sum_k k p_k$, degré moyen
  - q_k degré « en excès » $q_k = \frac{(k+1)p_{k+1}}{z}$
  - fonctions génératrices associées

$$G_0(x) = \sum_k p_k x^k, \quad G_1(x) = \sum_k q_k x^k$$

- transition de phase autour de

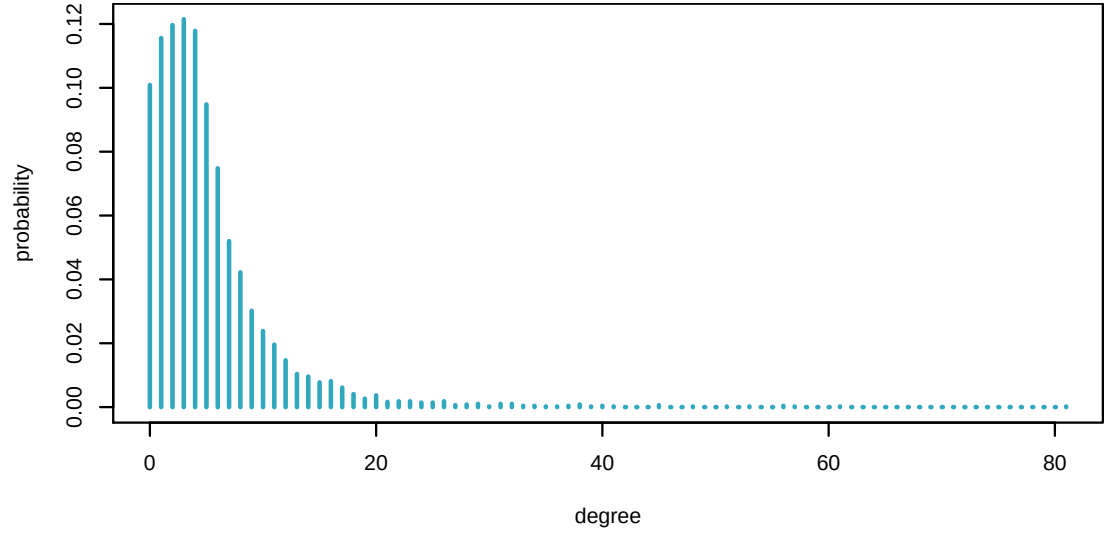
$$\sum_k k(k-2)p_k = 0$$

- fraction du graphe dans la plus grande composante, $S = \lim_n |C_{\max}|/n$, solution de

$$S = 1 - G_0(u), \quad u = G_1(u)$$

Degré déclaratif

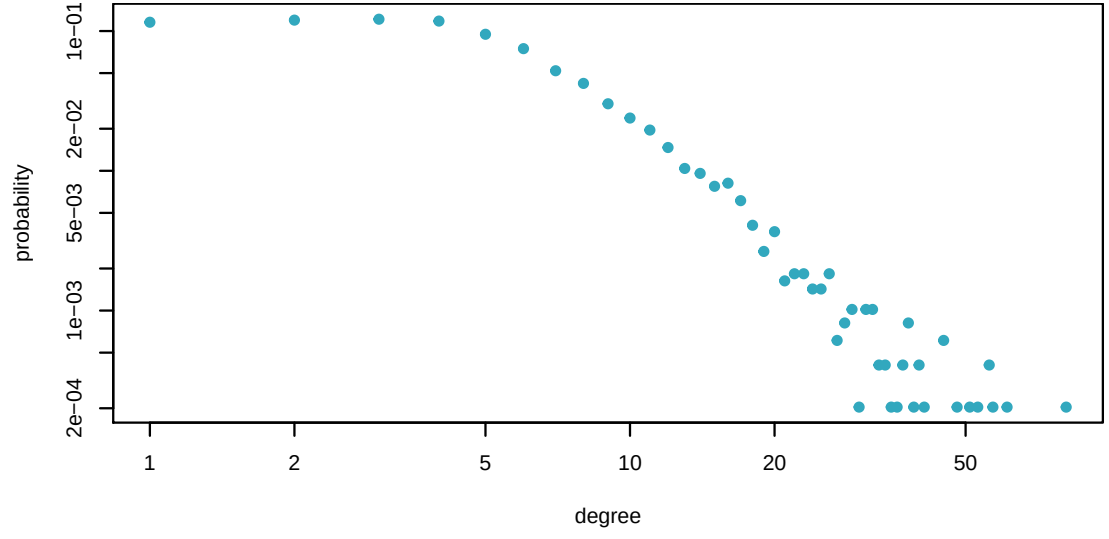
- retour à Cuba : nombre de partenaires sexuels sur les deux années précédant la détection



- sous le modèle de configuration, la composante géante occupe 90.8% du réseau

Degré déclaratif

- retour à Cuba : nombre de partenaires sexuels sur les deux années précédant la détection



- sous le modèle de configuration, la composante géante occupe 90.8% du réseau

Loi puissance

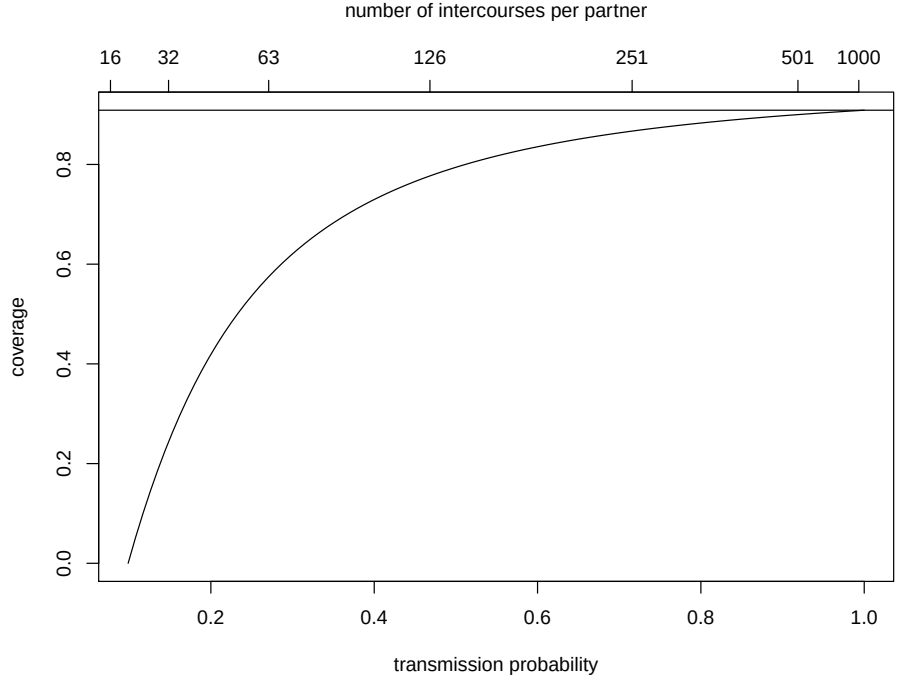
- cas particulier du modèle de configuration
- $p_k \sim k^{-\alpha}$
- transition de phase :
  - $\alpha > 3.4788$: pas de composante géante ($S = 0$)

$$\zeta(\alpha - 2) = \alpha\zeta(\alpha - 1),$$

    avec $\zeta(\alpha) = \sum_k k^{-\alpha}$.
  - $\alpha <= 2$, $S = 1$: une seule composante dans le graphe
  - transition continue entre les deux
- exemples, réseaux de contacts sexuels :
  - Suède (Liljeros et al, 2001) : $\alpha \simeq 2.4$
  - mais Cuba : $\alpha \simeq 3.5$

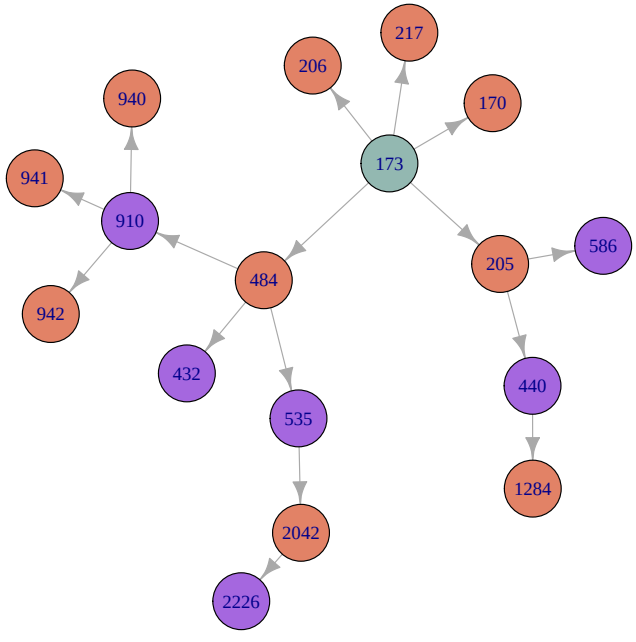
Percolation de lien

- modèle naïf de contamination :
  - probabilité T d'occupation d'un lien
  - graphe aléatoire support $\Rightarrow$ graphe d'infection potentiel
- analyse asymptotique sur des modèles simples :
  - composante géante du graphe d'infection
  - transition de phase sur T
- à Cuba, $T_c \simeq 0.099$



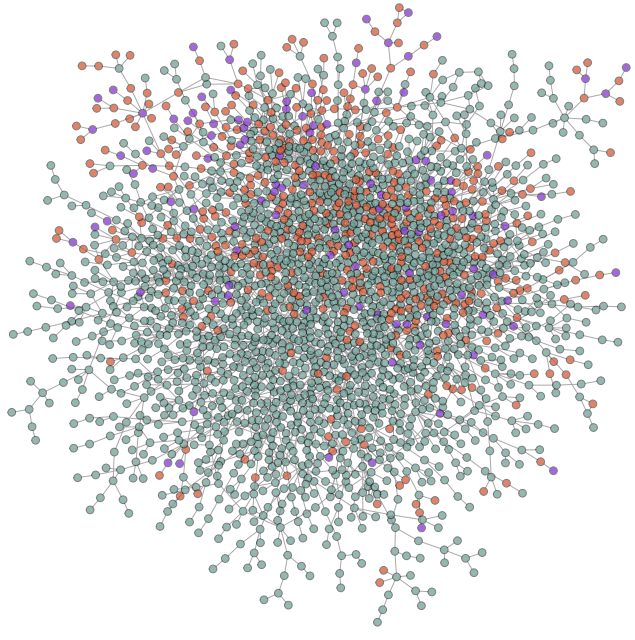
Analyser plus finement

- modèles très éloignés de la réalité :
  - pas de prise en compte de l'orientation sexuelle
  - plus généralement : noeuds anonymes
  - résultats asymptotiques
- analyse exploratoire :
  - visualisation
  - vérification à posteriori
  - ré-échantillonnage



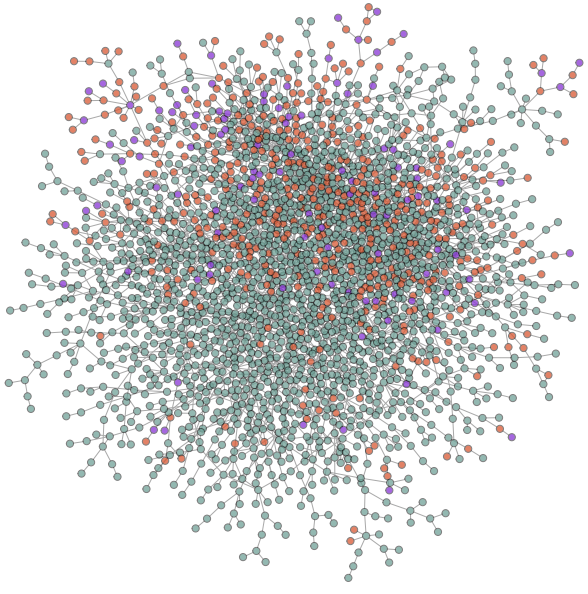
Analyser plus finement

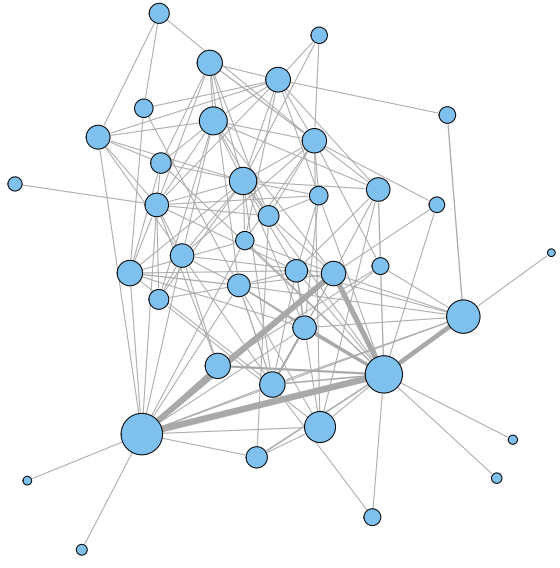
- modèles très éloignés de la réalité :
  - pas de prise en compte de l'orientation sexuelle
  - plus généralement : noeuds anonymes
  - résultats asymptotiques
- analyse exploratoire :
  - visualisation
  - vérification à posteriori
  - ré-échantillonnage



Visualisation hiérarchique

- réduction de complexité :
  - classification (hiérarchique) des sommets du graphe
  - visualisation du graphe des classes





- pertinence ?
  - qualité de la classification
  - lisibilité
  - inférence

Classification

- classification des sommets d'un graphe :
  - domaine très étudié (détection de communautés)
  - dizaines de techniques
  - objectif ici : résumer la structure du graphe
- mesure de qualité
  - Modularité (Girvan et Newman, 2004) :

$$Q = \frac{1}{2m} \sum_{l=1}^{L} \sum_{i,j \in C_l} \left(w_{ij} - \frac{k_i k_j}{2m} \right)$$

  + favorise les classes denses
  + gère correctement les sommets de haut degré
  + nombre de classes « optimal »
  + adapté à la visualisation (Noack, 2009)
  - optimisation NP difficile
  - résolution limitée
  - sensible au « bruit »

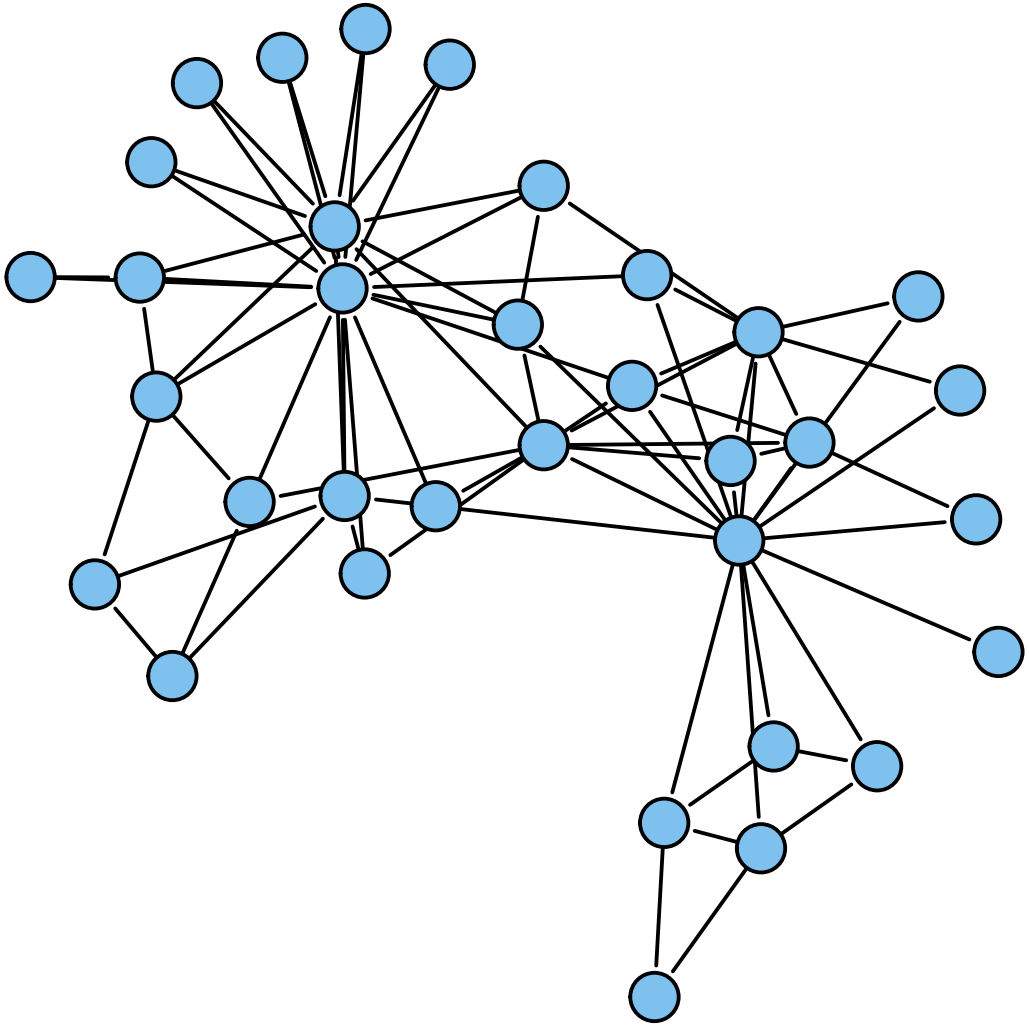
Sensibilité

- algorithmes de maximisation
  - méthodes gloutonnes
  - fusion de classes et raffinement (échange de sommets)
  - trouvent toujours une classification...
  - valeur de la modularité peu informative

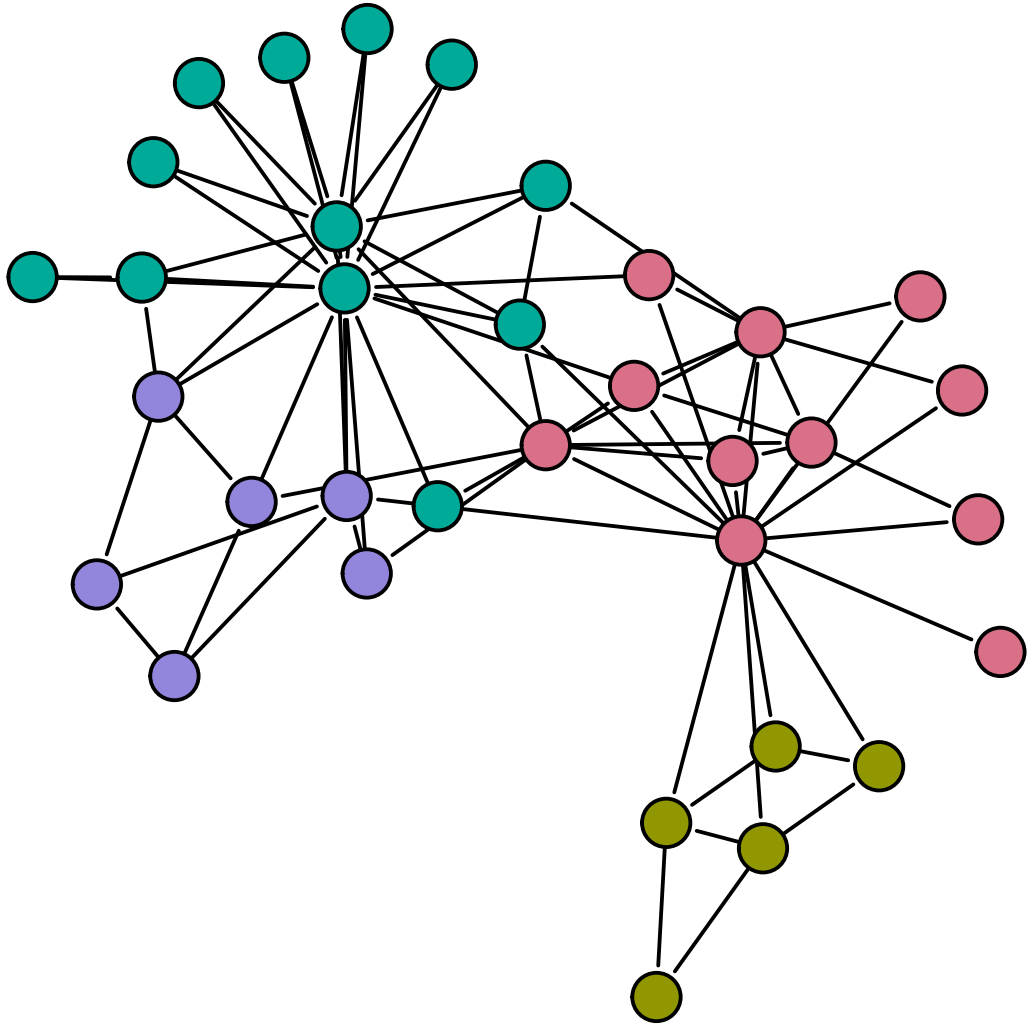
Sensibilité

- algorithmes de maximisation
  - méthodes gloutonnes
  - fusion de classes et raffinement (échange de sommets)
  - trouvent toujours une classification...
  - valeur de la modularité peu informative
- test sur la modularité :
  - graphe aléatoire (modèle de configuration)
  - classification $\Rightarrow$ modularité
  - niveau « ambiant » de modularité : *p-value* de la modularité sur le graphe étudié

Exemple

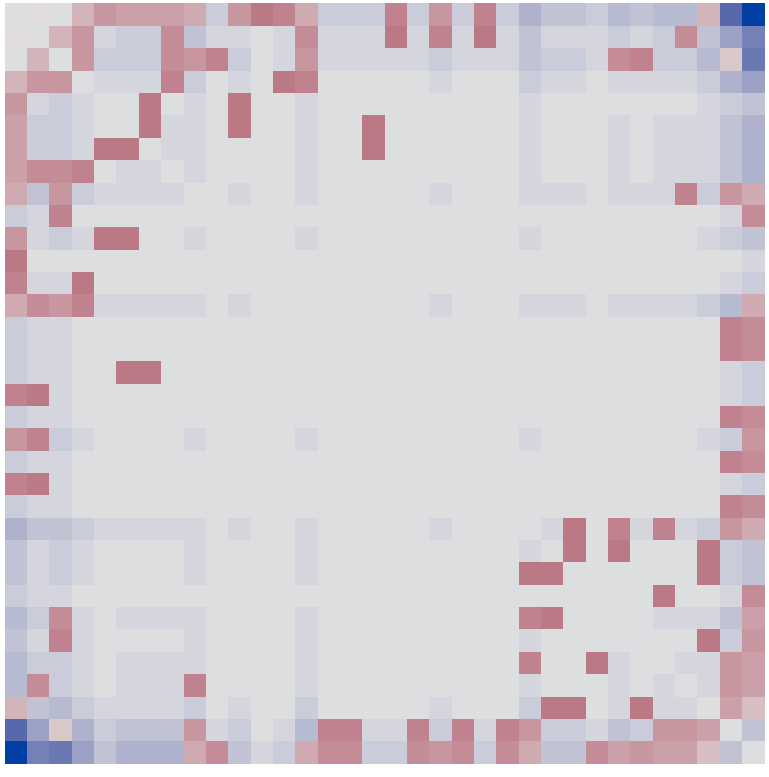


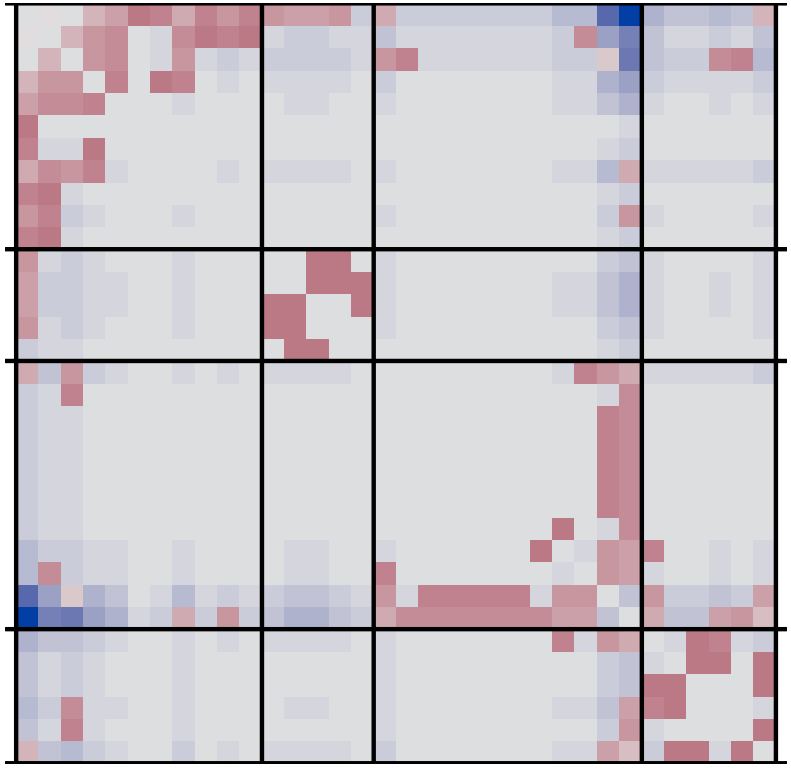
Exemple



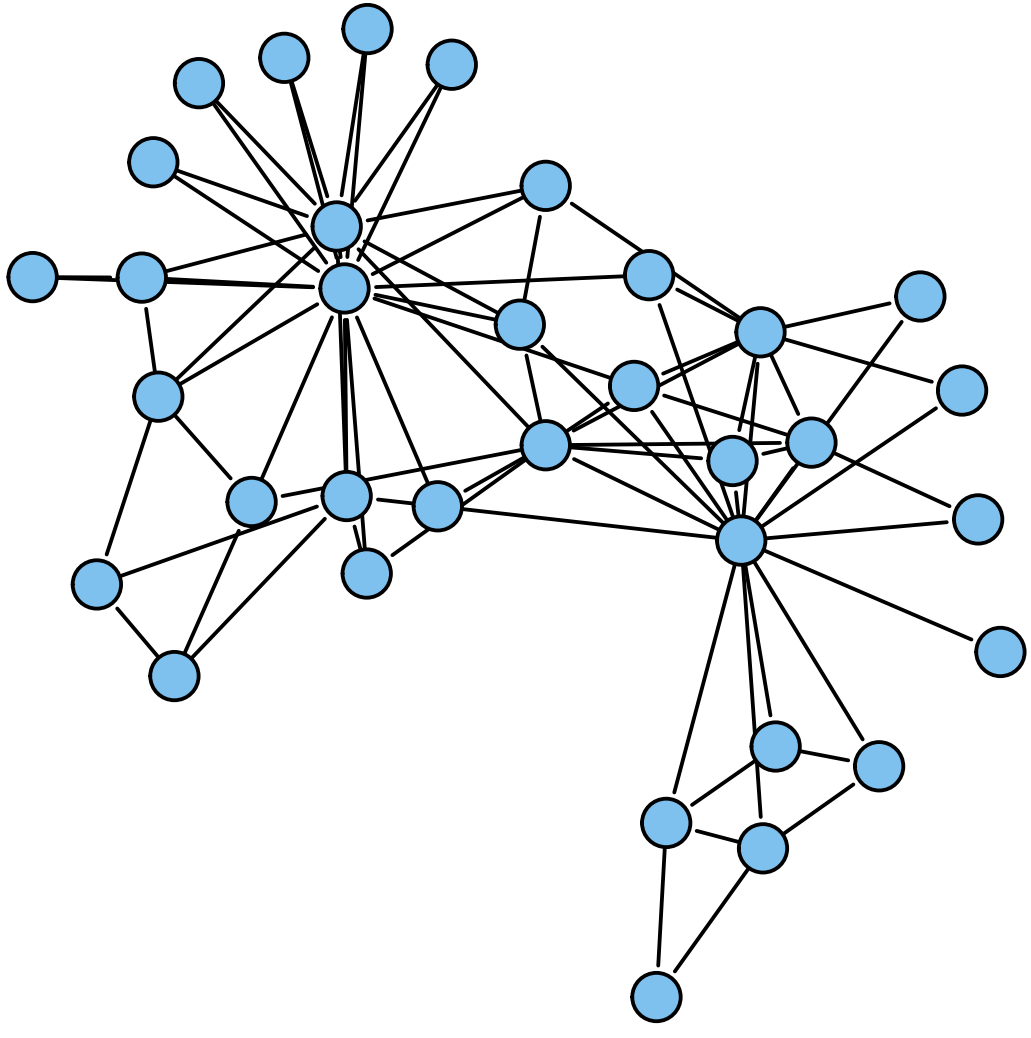
4 classes, modularité $\simeq 0.42$

Exemple

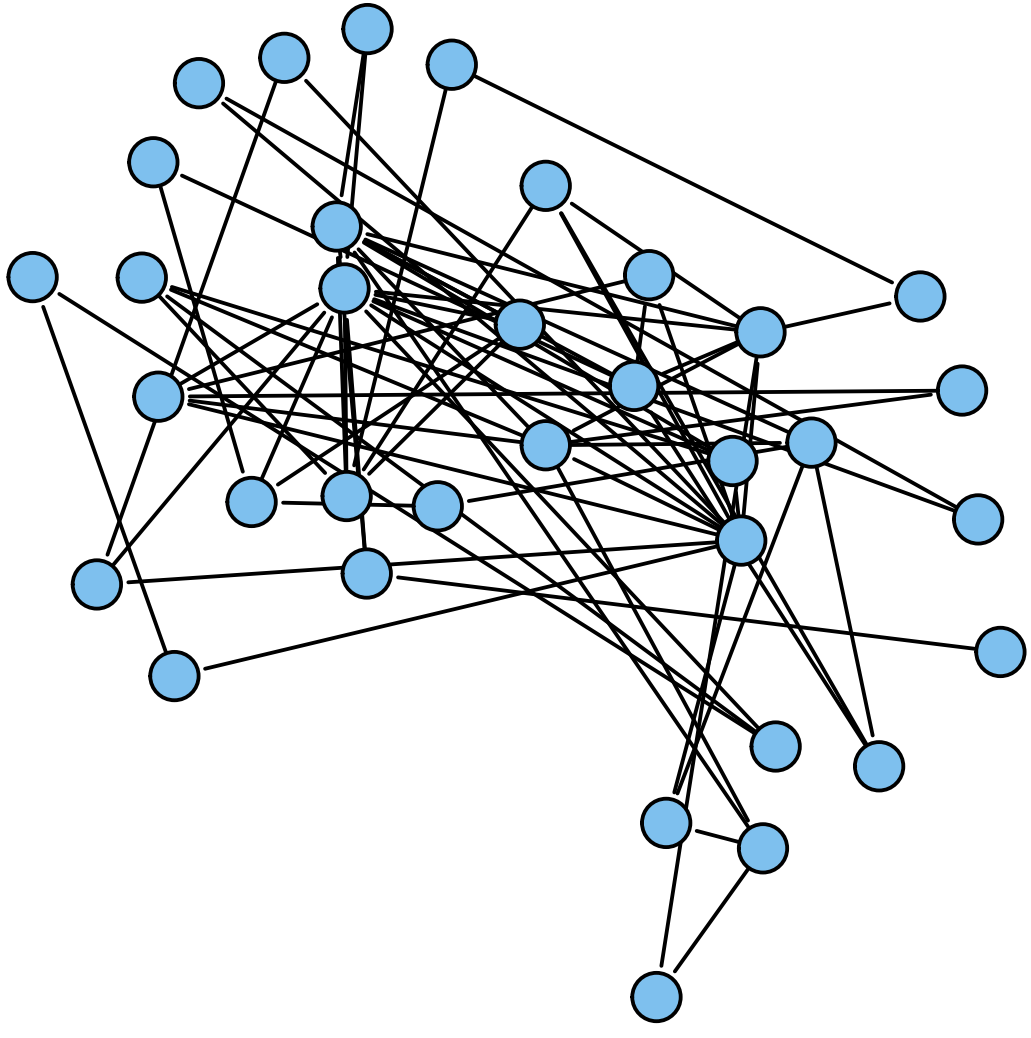




Simulation

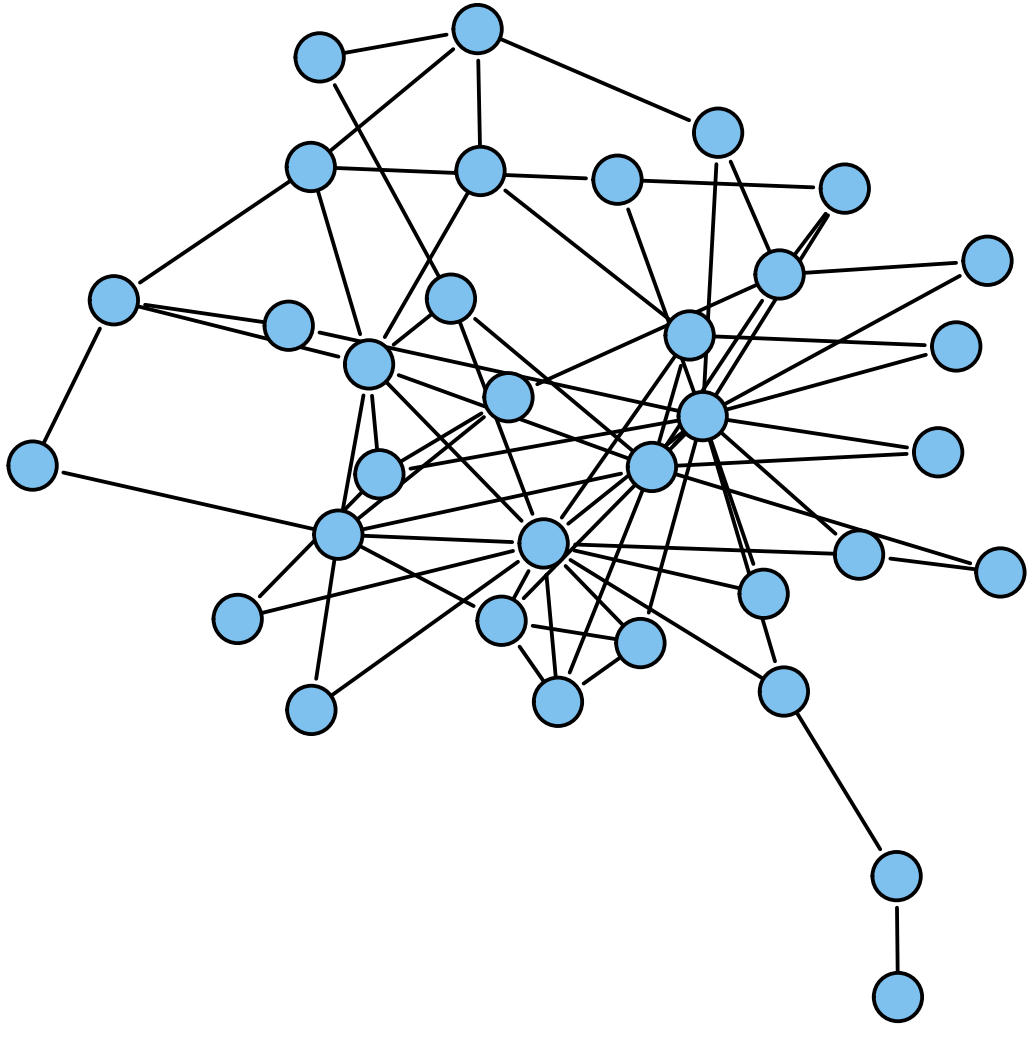


Simulation



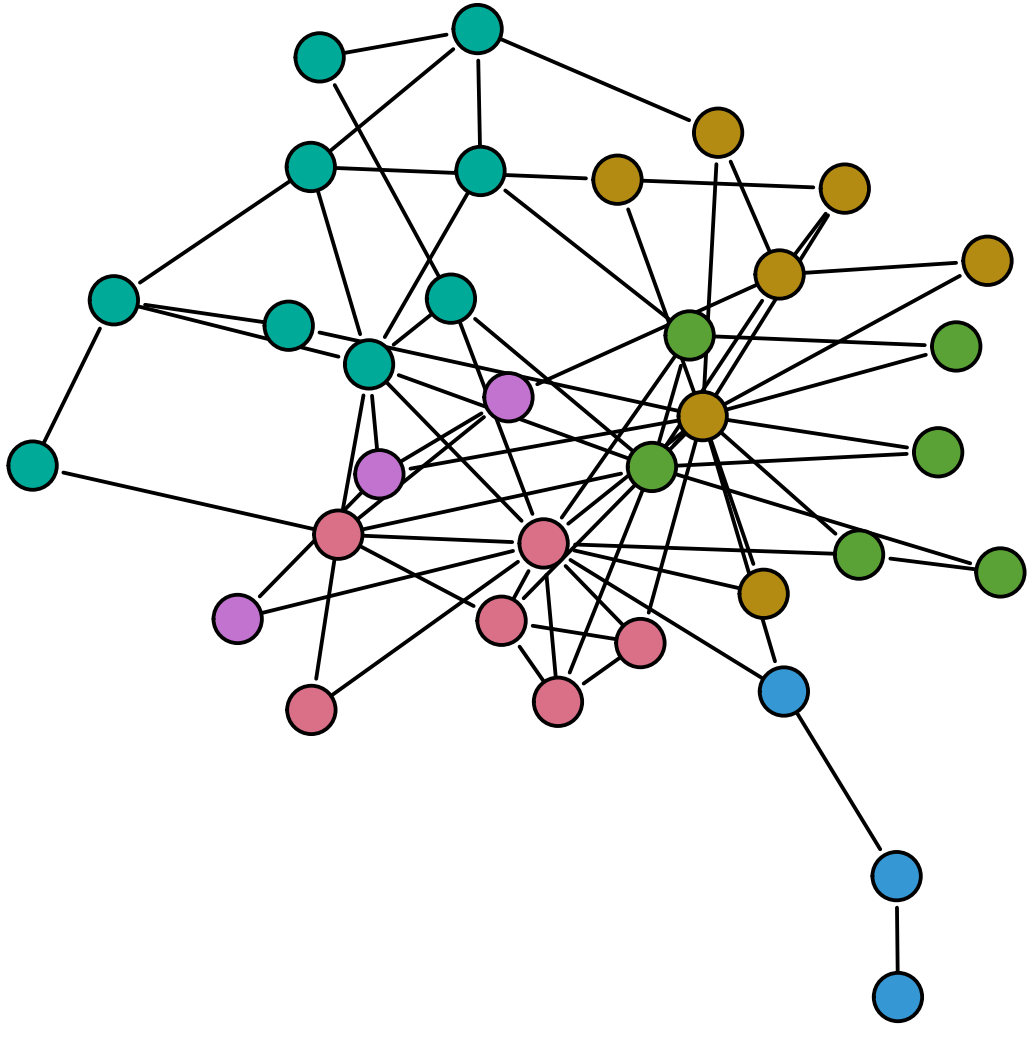
Modèle de configuration : mêmes degrés

Simulation



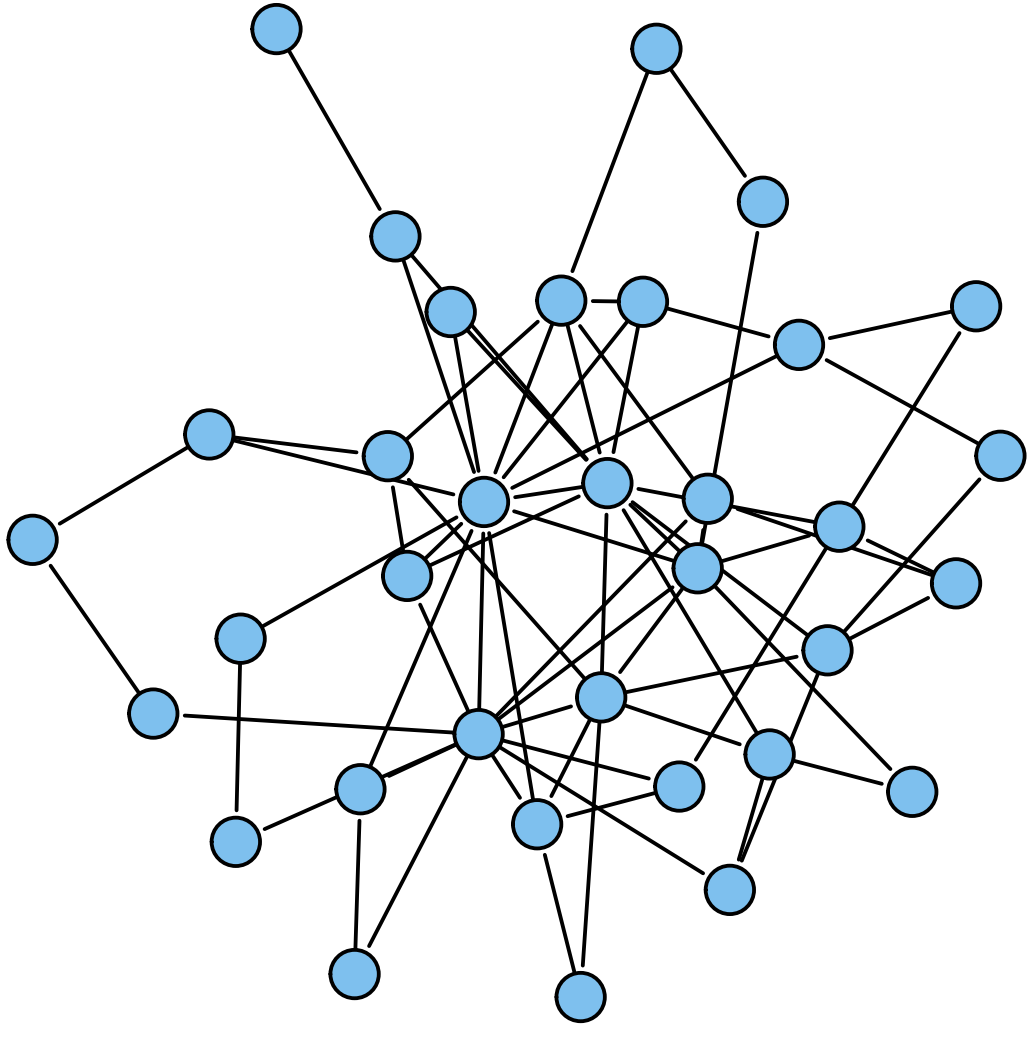
Repositionné

Simulation



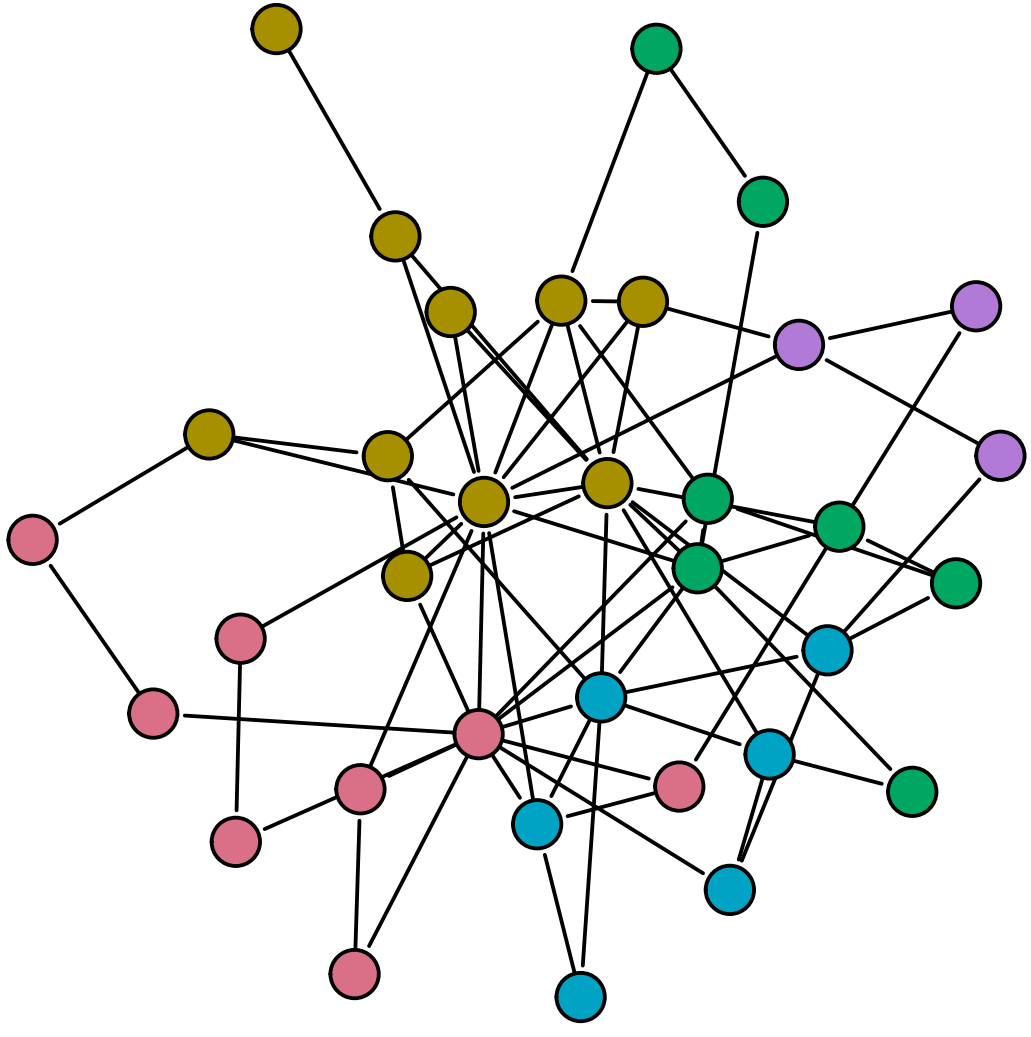
6 classes, modularité $\simeq 0.35$

Simulation



Nouveau tirage

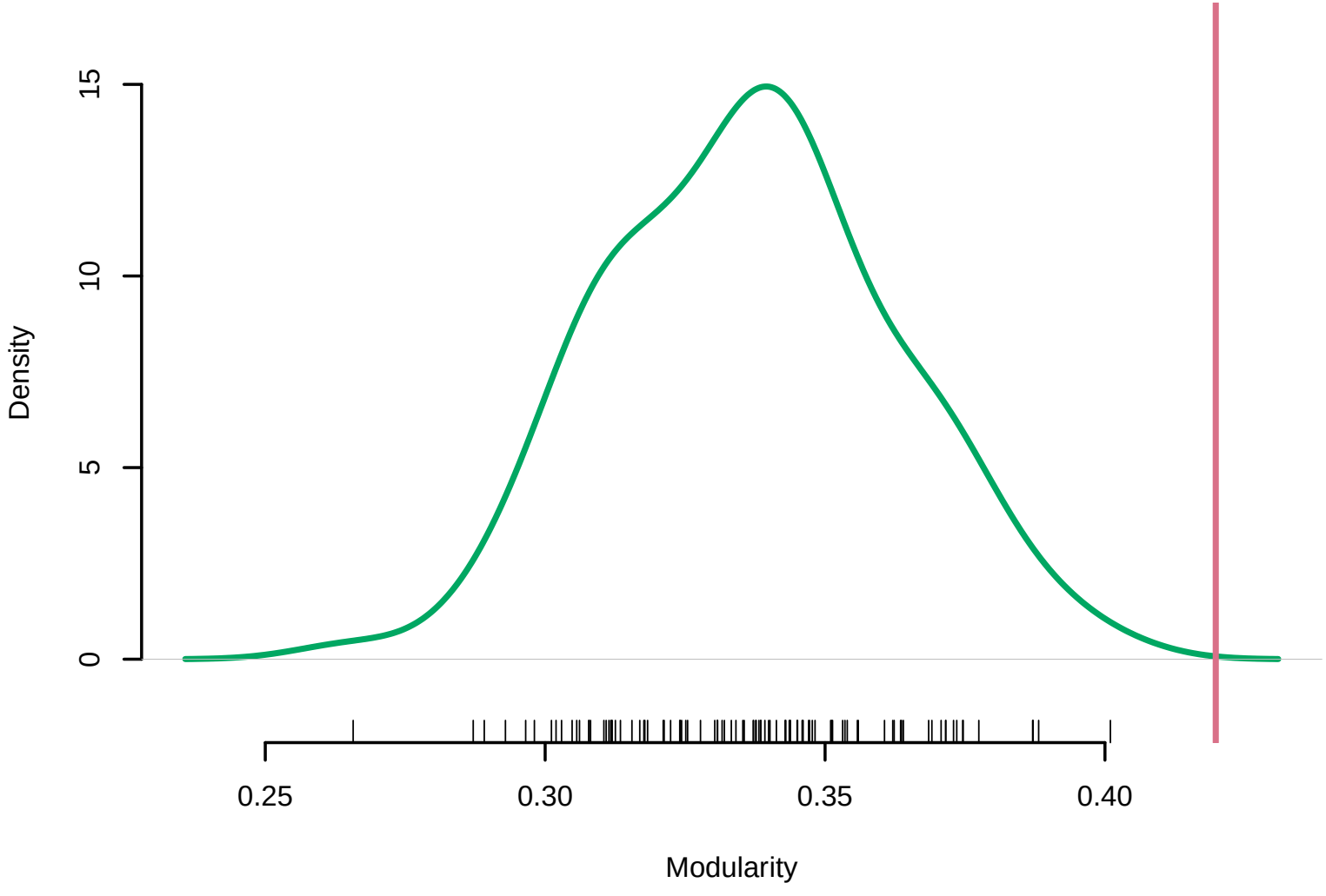
Simulation



5 classes, modularité $\simeq 0.34$

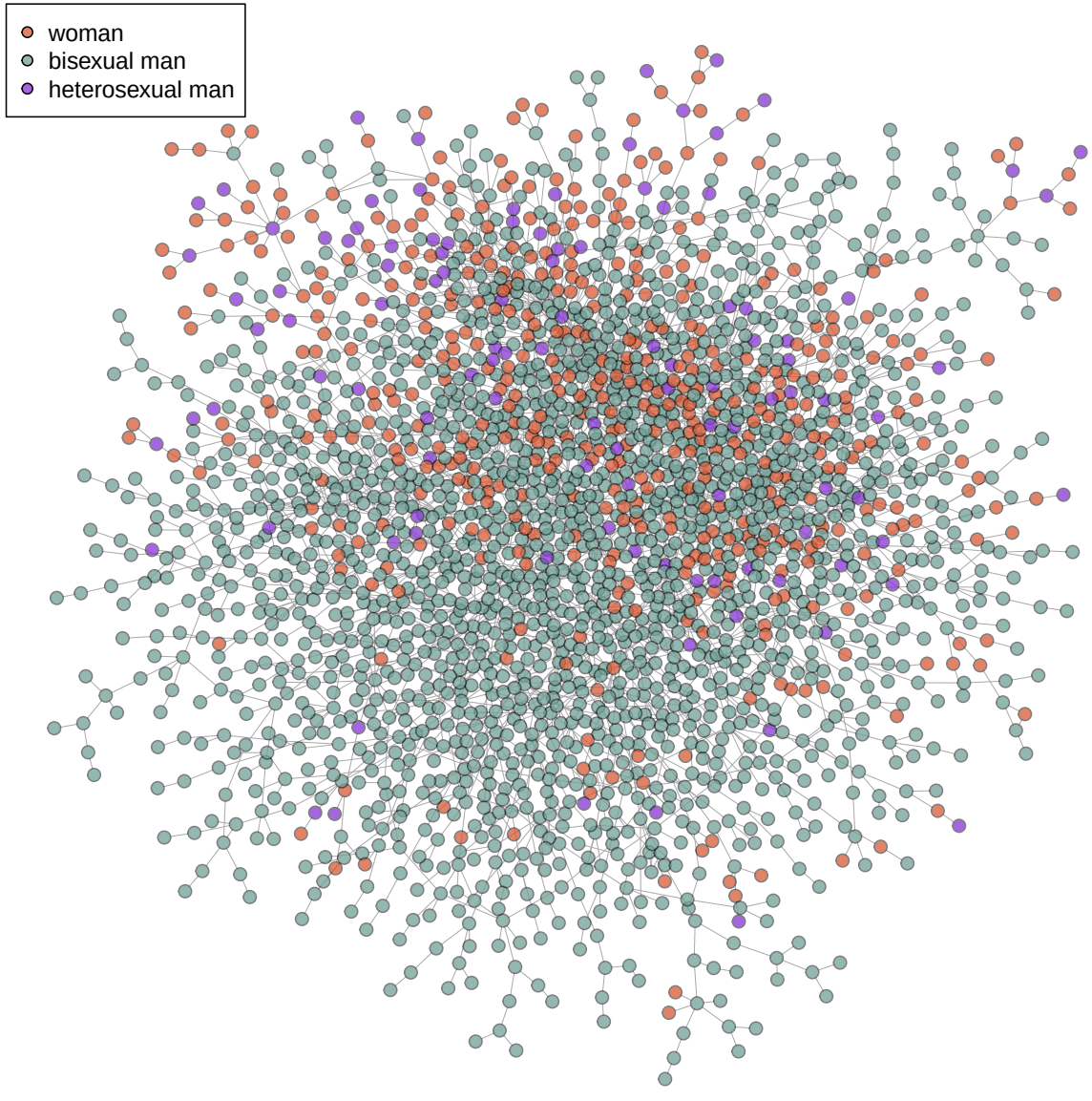
Simulation

100 graphes aléatoires

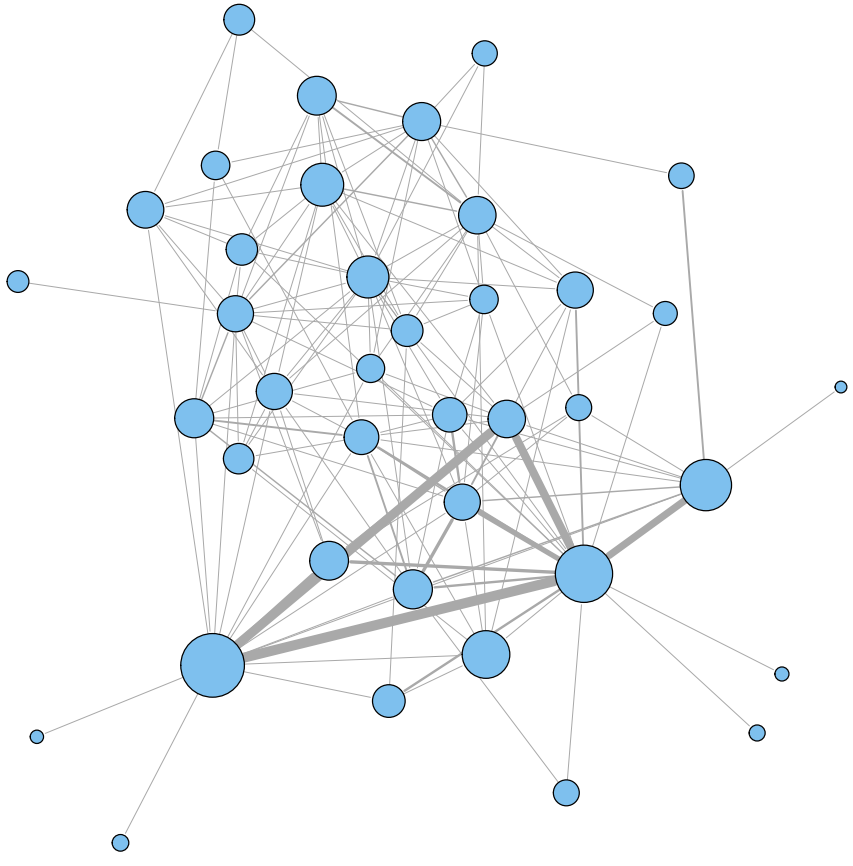


la classification sur le graphe d'origine a un sens

Composante connexe principale

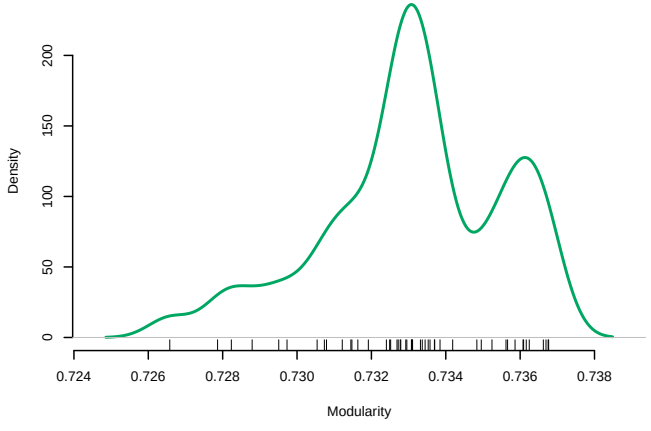


Classification



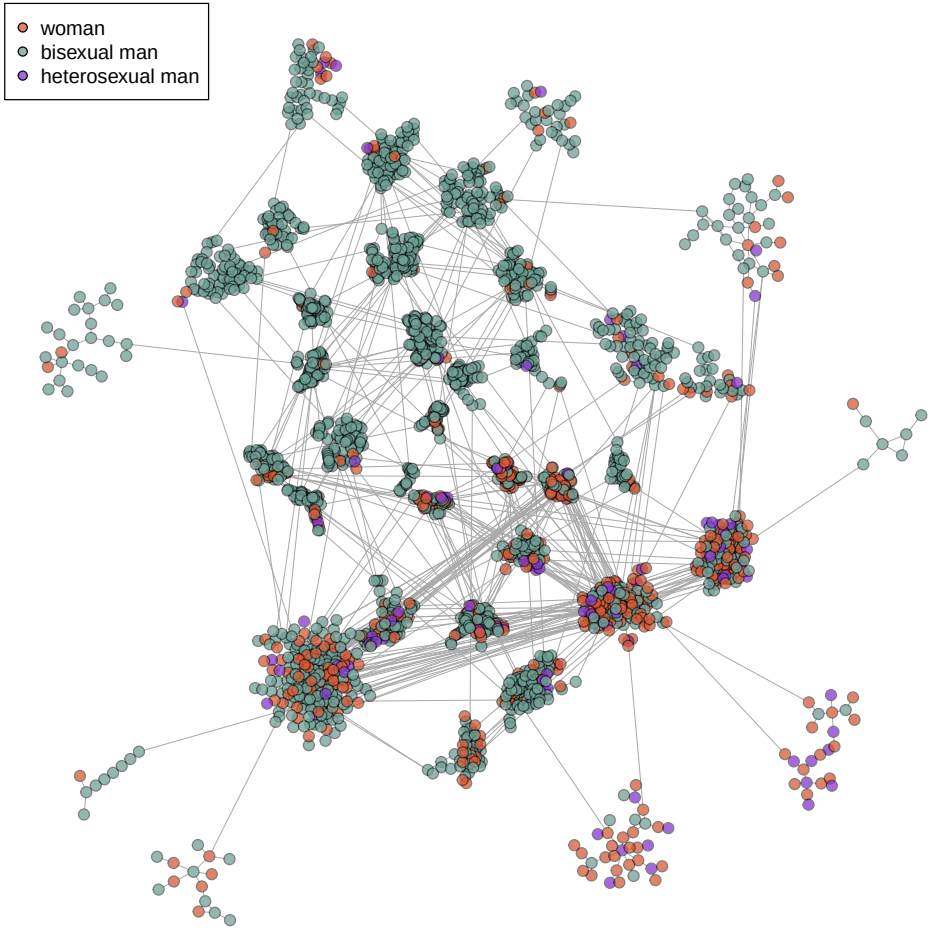
$\Rightarrow$ 39 classes (89.5% des liens internes aux classes)

$\Rightarrow$ modularité $\simeq 0.85$



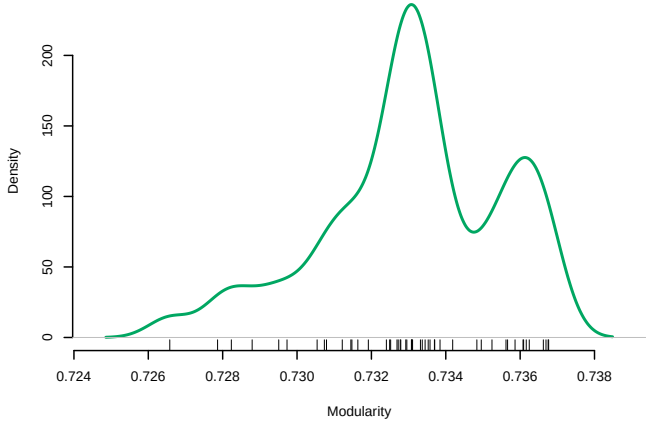
$\Rightarrow$ modularité « aléatoire » ≤ 0.74

Classification



⇒ 39 classes (89.5% des liens internes aux classes)

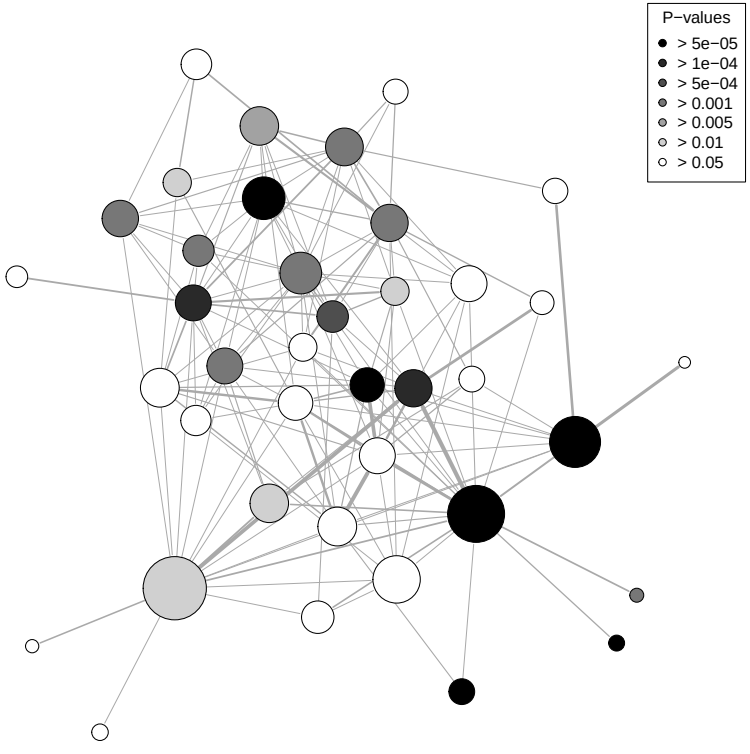
⇒ modularité $\simeq 0.85$



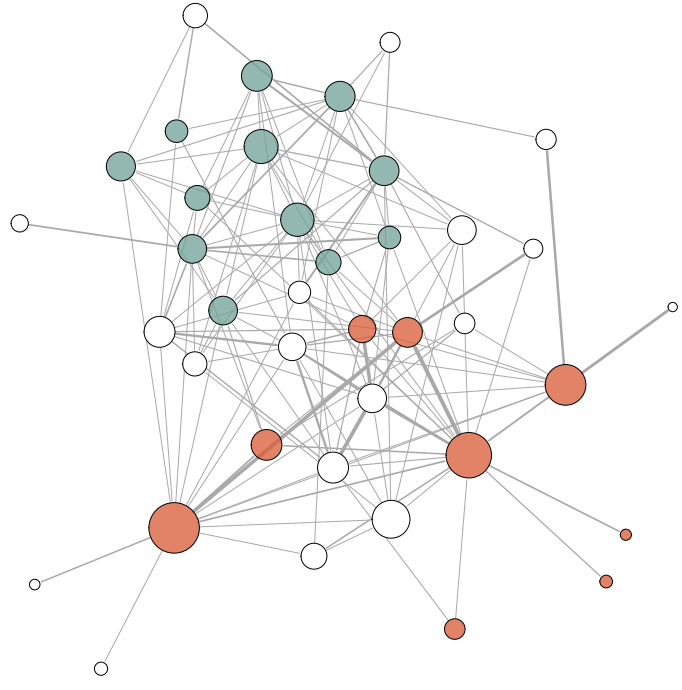
⇒ modularité « aléatoire » ≤ 0.74

⇒ visualisation hiérarchique de l'orientation sexuelle

Statistique graphique

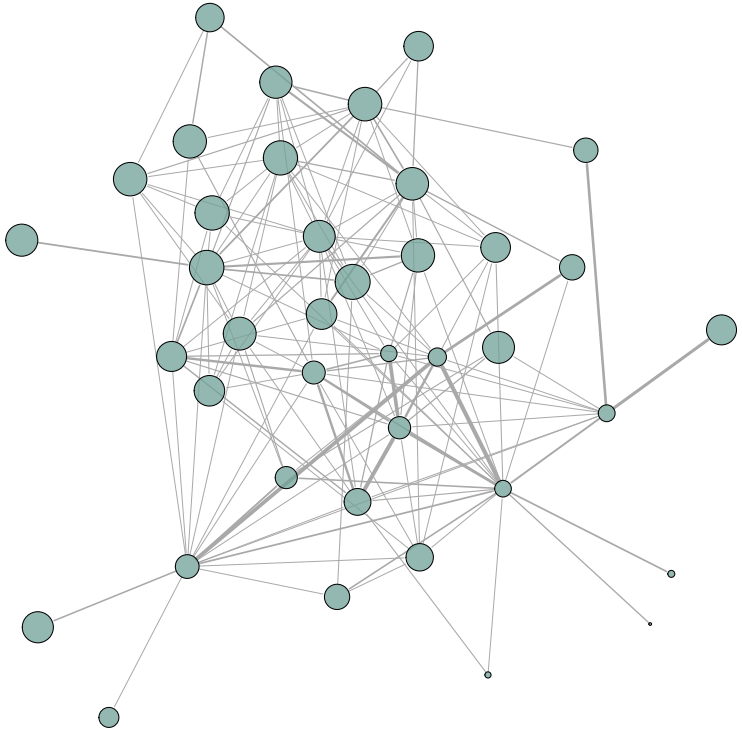


p-value d'un test du χ^2 sur la distribution de l'orientation sexuelle

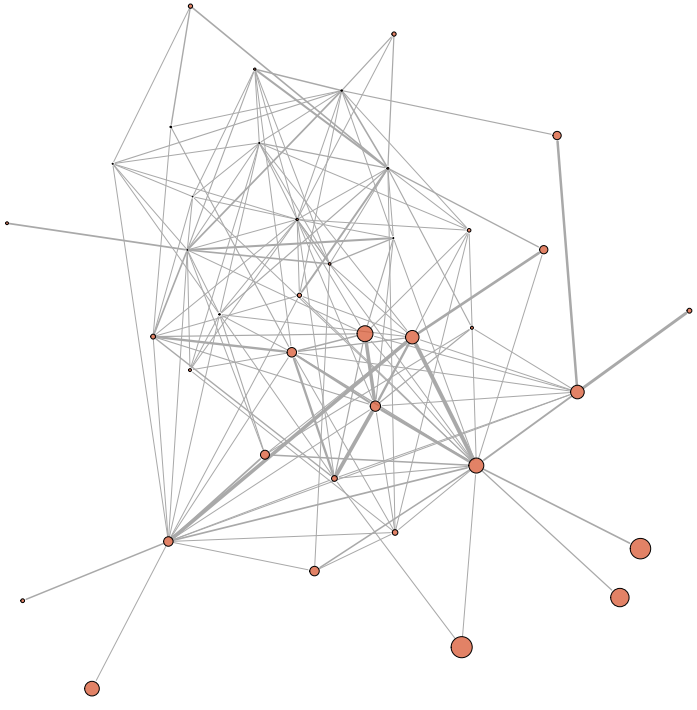


orientation sexuelle atypique

Statistique graphique



Hommes homosexuels



Femmes

Pourcentages

Confirmation

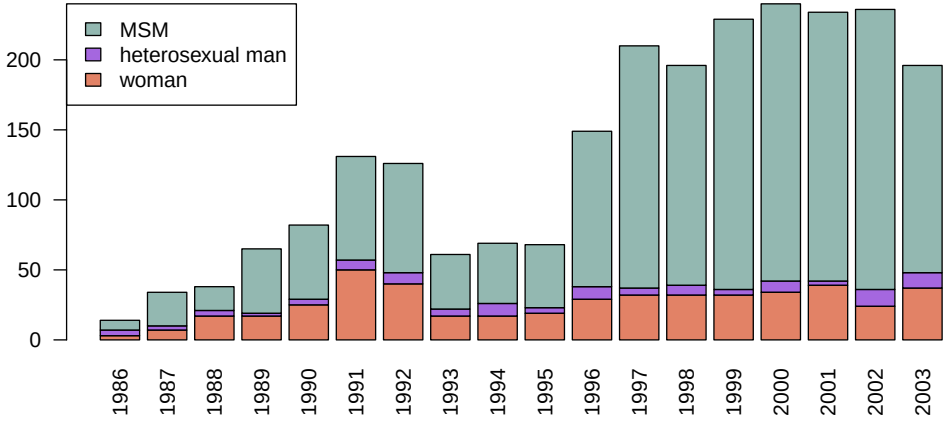
- distances géodésiques :

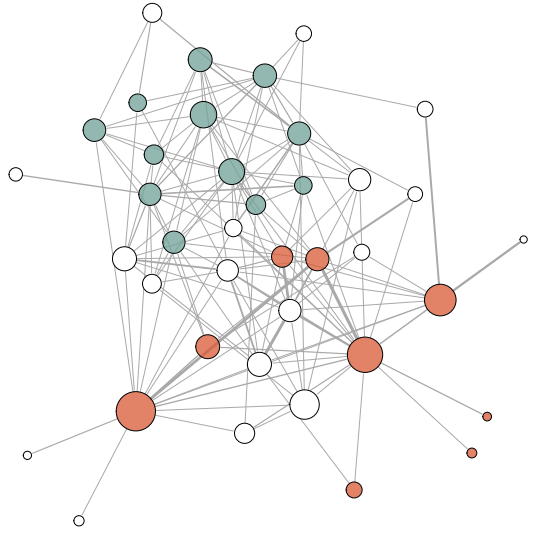
	Bisexual	Mixed	Typical
Bisexual	9.79	12.28	11.93
Mixed	12.28	7.56	9.24
Typical	11.93	9.24	12.04

- connections directes :
  - 333 connexions entre groupes (sur 3168)
  - seulement 16 connexions entre les méta-groupes
  - 1 % de chance d'obtenir si peu de connexions entre méta-groupes

Aspect temporel

Recrutement annuel

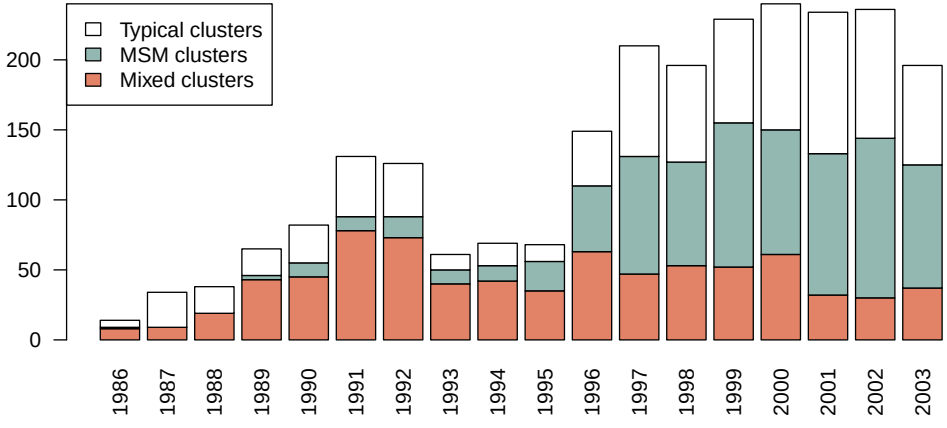


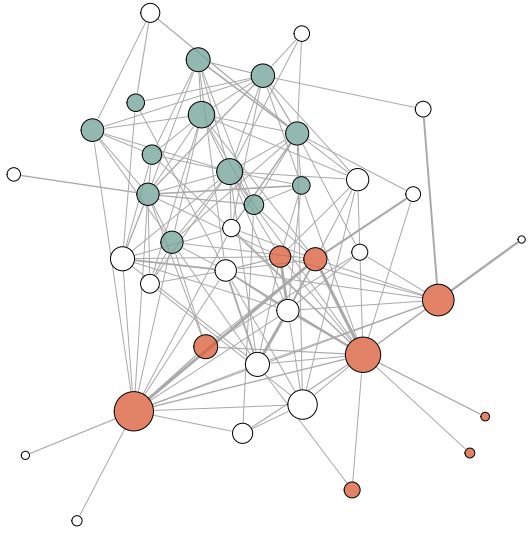


Pas de prise en compte explicite du temps, mais connexions « datées »

Aspect temporel

Recrutement annuel





Pas de prise en compte explicite du temps, mais connexions « datées »

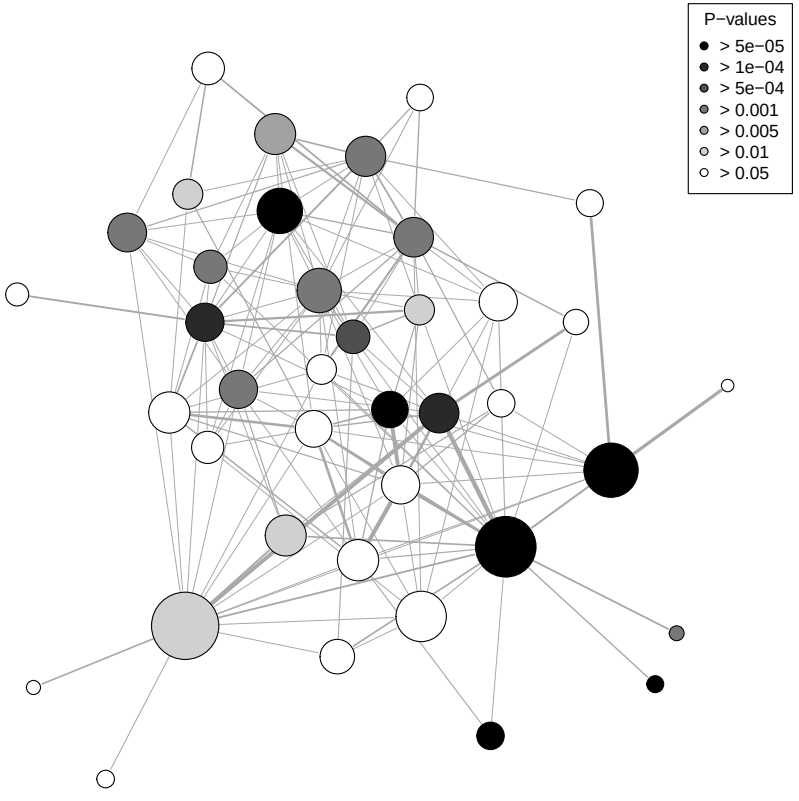
Conclusion

- exploration visuelle de graphes :
  - ne montrer que ce qui existe
  - statistique graphique
  - rendu hiérarchique : simplification ou détails
- validation par simulation :
  - coûteux
  - reste très naïf : ne dispense que de l'asymptotique
- perspectives :
  - aspect temporel explicite
  - graphes multipartis
  - etc.

Hiérarchie

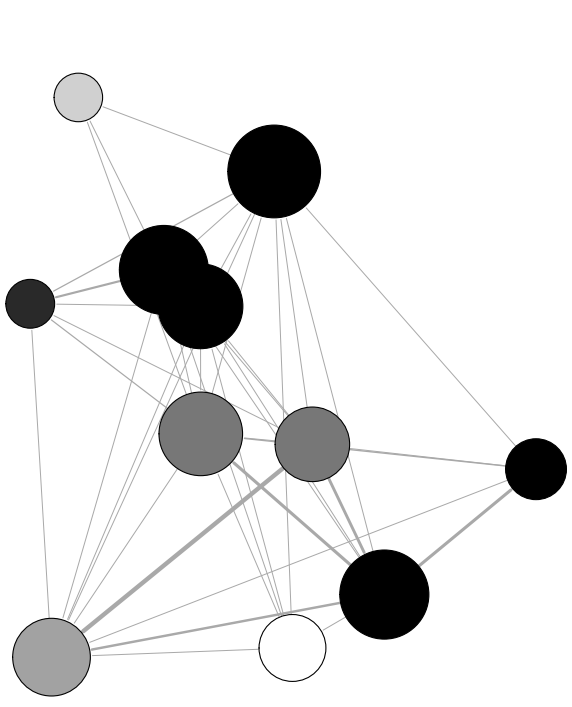
- moins de détails :
  - poursuite de l'algorithme glouton de classification
  - fusion de classes (contrainte hiérarchique)
  - visualisation barycentrique
  - maintient de la modularité au dessus du seuil aléatoire
- plus de détails :
  - classification des classes
  - liens externes supprimés
  - pas de sous-classes non significatives
  - maintient de la modularité globale au dessus du seuil aléatoire

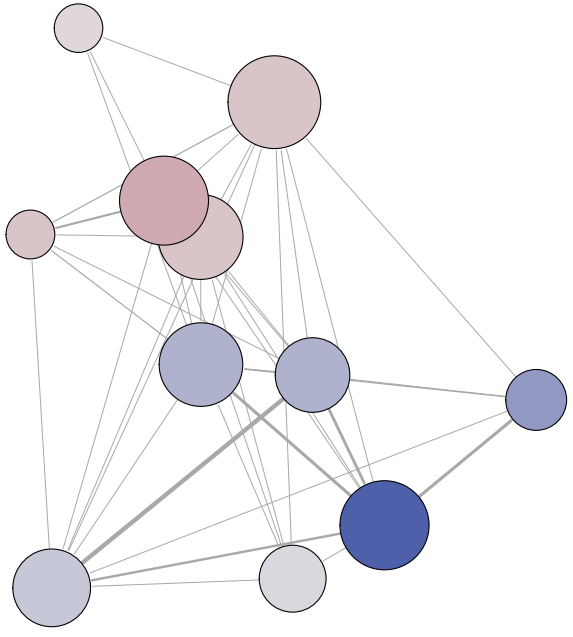
Simplification



p-value

Simplification



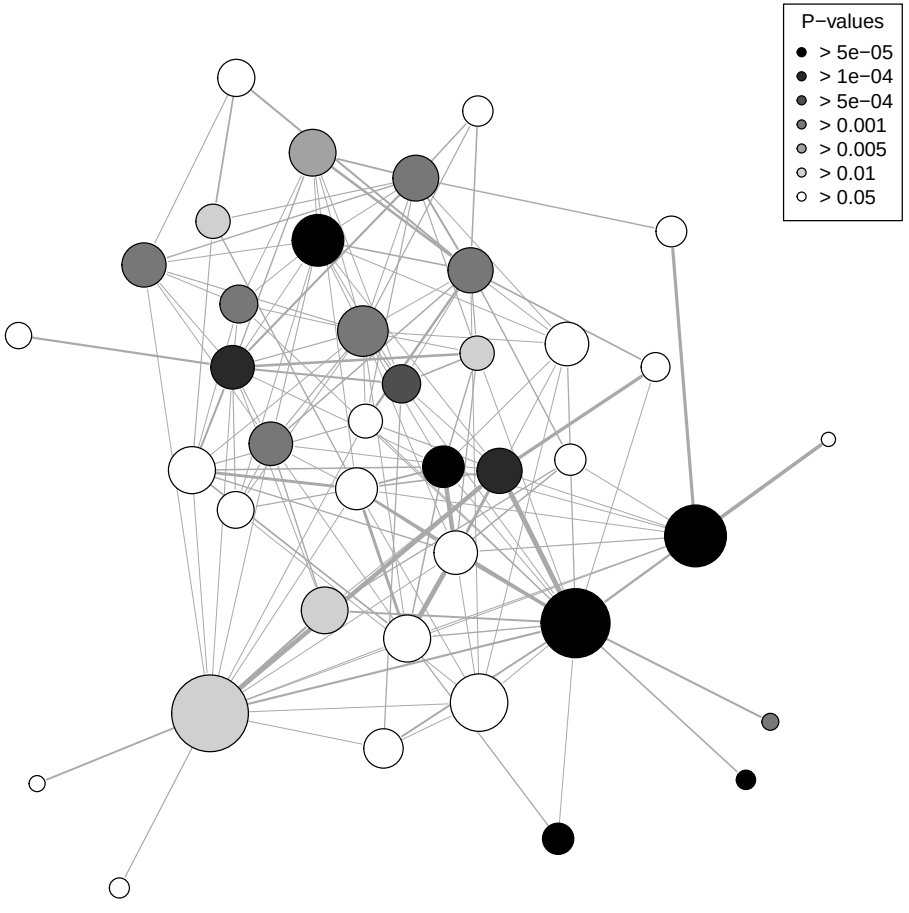


p-value

résidus de *Pearson*

Confirme la structure en deux parties

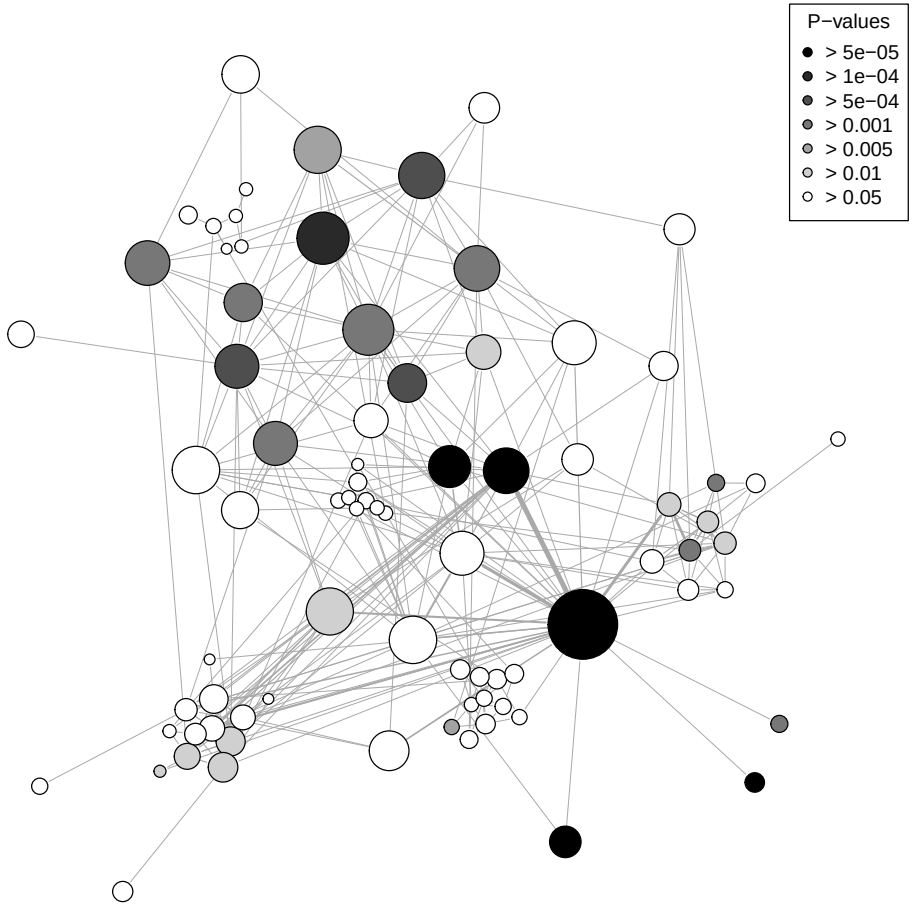
Sous-structures



$\Rightarrow$ 5 classes possèdent une sous structure

$\Rightarrow$ la modularité se maintient au dessus de 0.81

Sous-structures

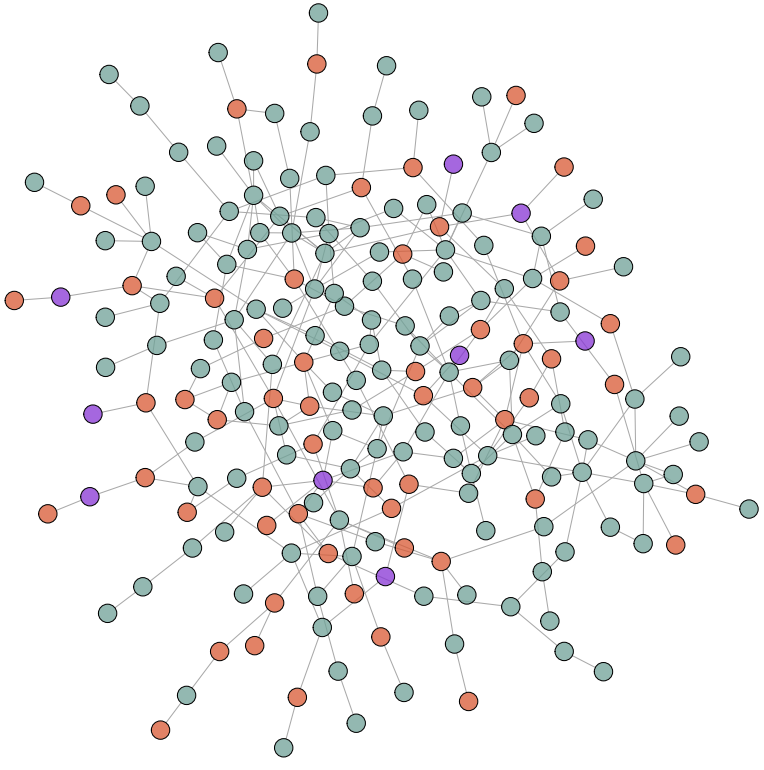


⇒ 5 classes possèdent une sous structure

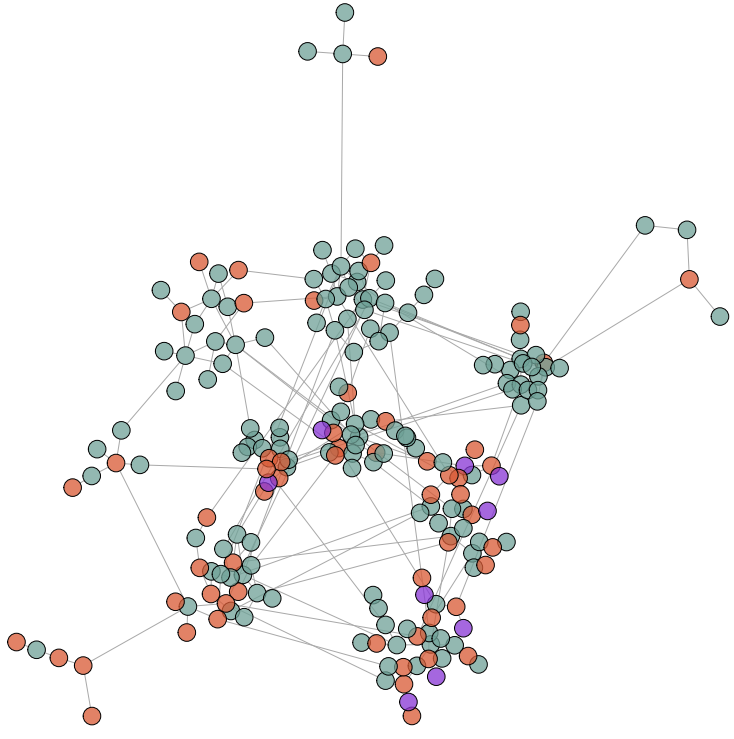
⇒ la modularité se maintient au dessus de 0.81

⇒ sous structures atypiques

Analyse locale



Visualisation classique



Visualisation hiérarchique