Analyse exploratoire d'un graphe : le cas de la contamination par le VIH à Cuba

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

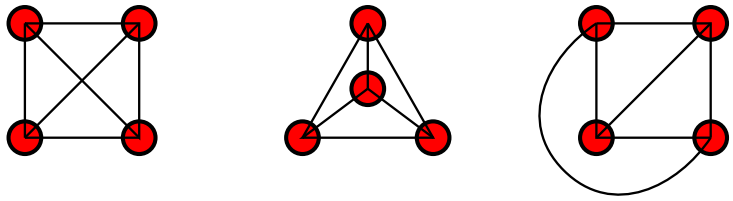
SMAI 2011

Données relationnelles

- données :
  - observations classiques : $(x_i)_{1 \leq i \leq N} \in \mathbb{R}^p$
  - graphe d'interaction : les x_i sont les sommets
- analyse exploratoire :
  - comprendre la structure du graphe
  - lier graphe et valeurs attachées aux sommets
- outil classique : visualisation de graphe

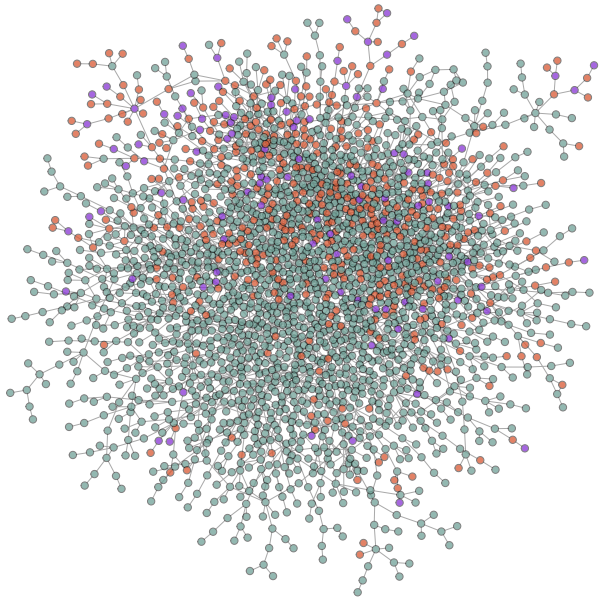
Données relationnelles

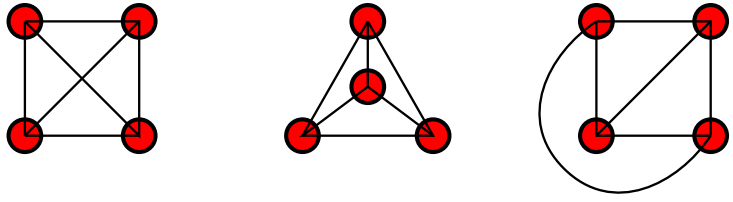
- données :
  - observations classiques : $(x_i)_{1 \leq i \leq N} \in \mathbb{R}^p$
  - graphe d'interaction : les x_i sont les sommets
- analyse exploratoire :
  - comprendre la structure du graphe
  - lier graphe et valeurs attachées aux sommets
- outil classique : visualisation de graphe, mais



Données relationnelles

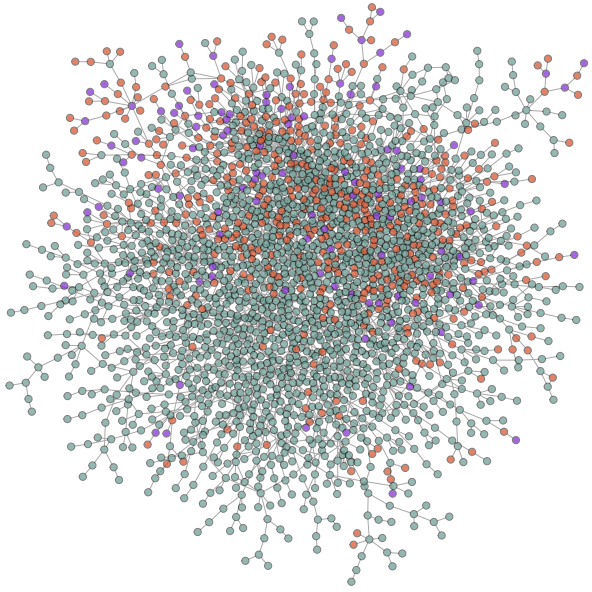
- données :
  - observations classiques : $(x_i)_{1 \leq i \leq N} \in \mathbb{R}^p$
  - graphe d'interaction : les x_i sont les sommets
- analyse exploratoire :
  - comprendre la structure du graphe
  - lier graphe et valeurs attachées aux sommets
- outil classique : visualisation de graphe, mais

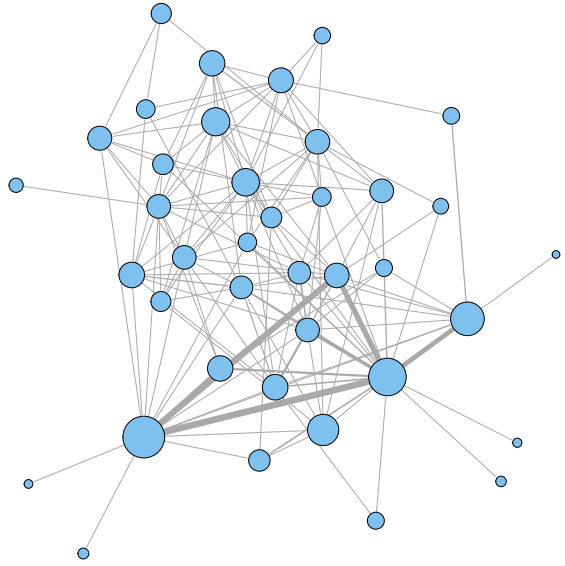




Visualisation hiérarchique

- 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 »

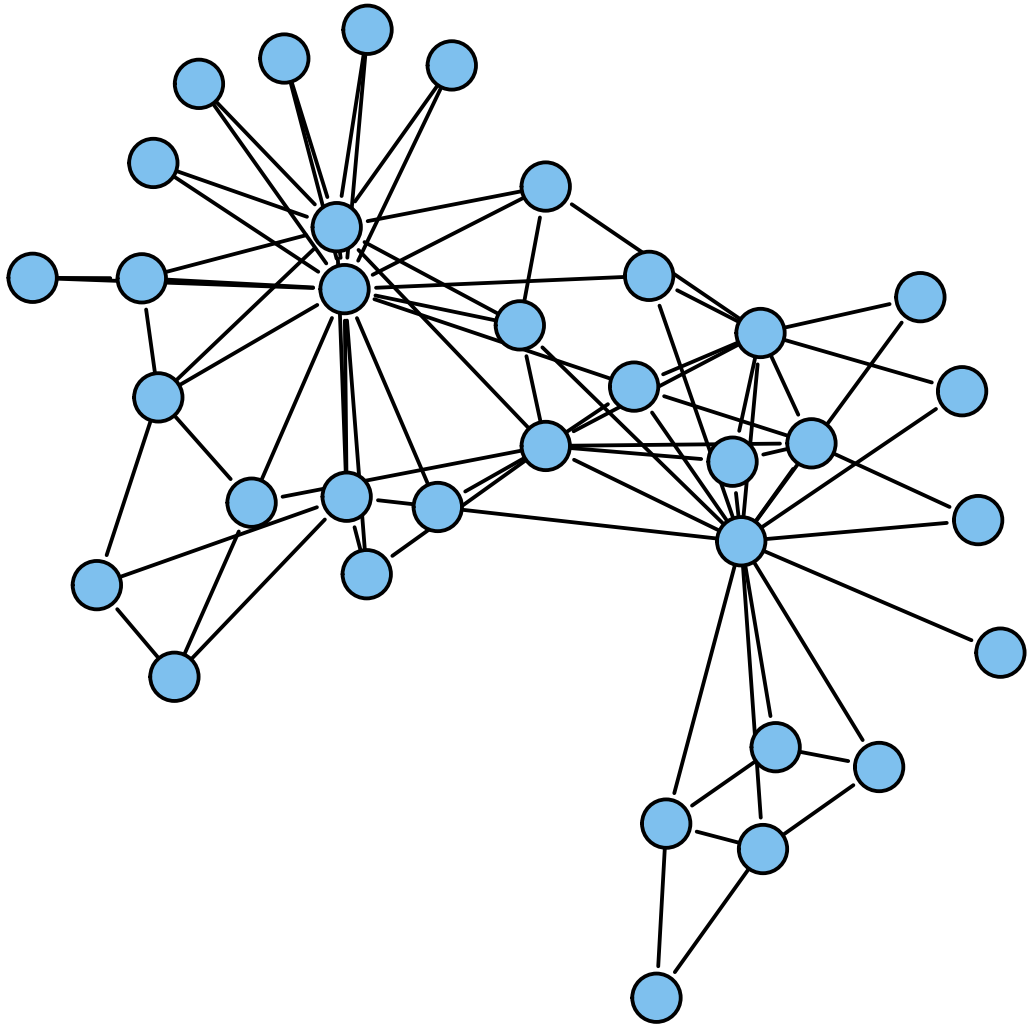
Sensibilité

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

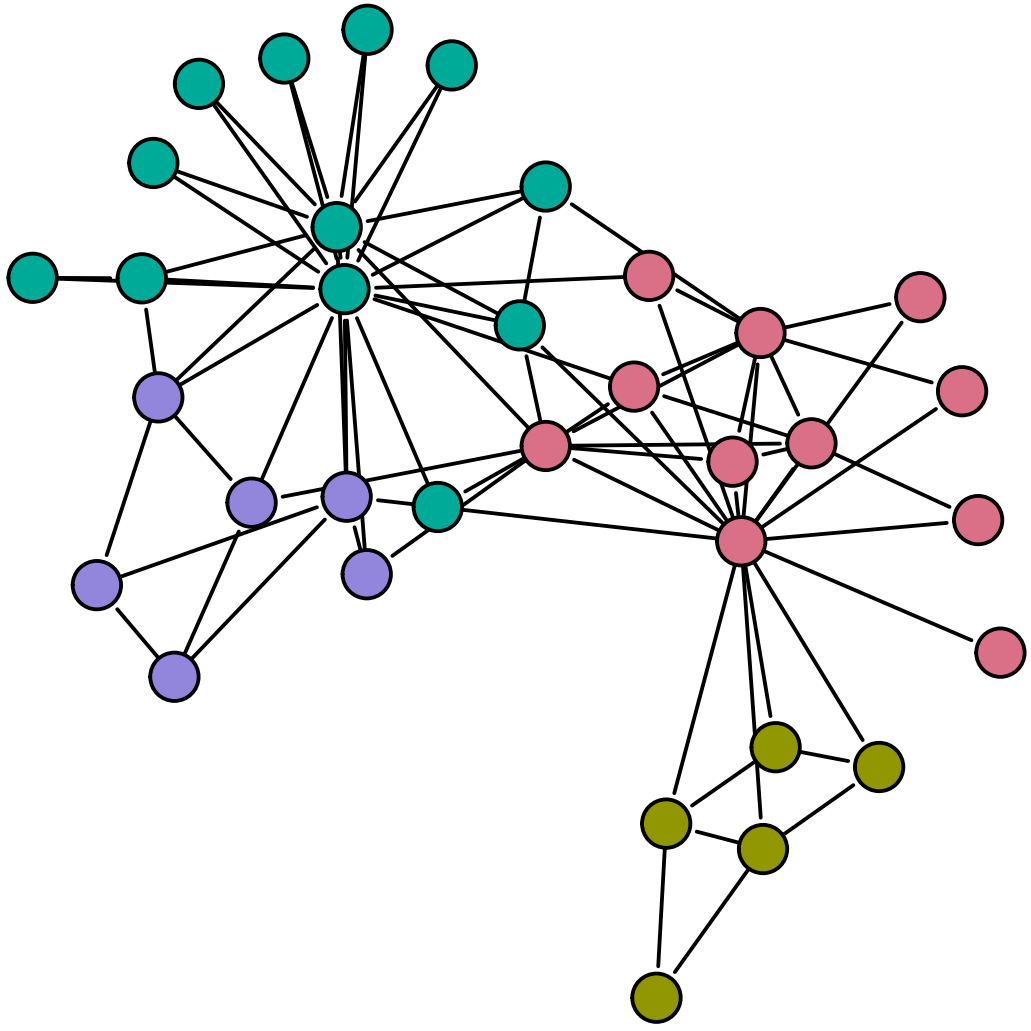
Sensibilité

- 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 (sans structure)
  - classification $\Rightarrow$ modularité
  - niveau « ambiant » de modularité : *p-value* de la modularité sur le graphe étudié

Exemple

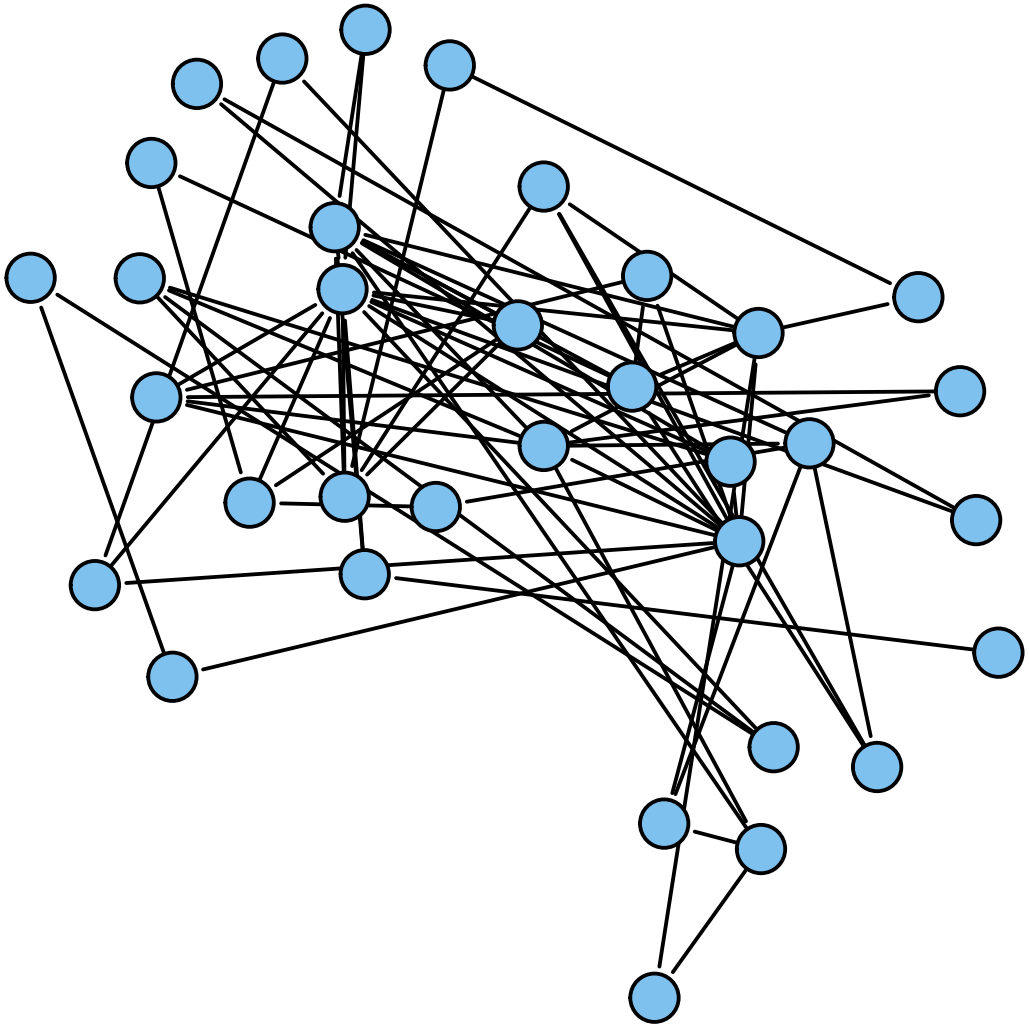


Exemple



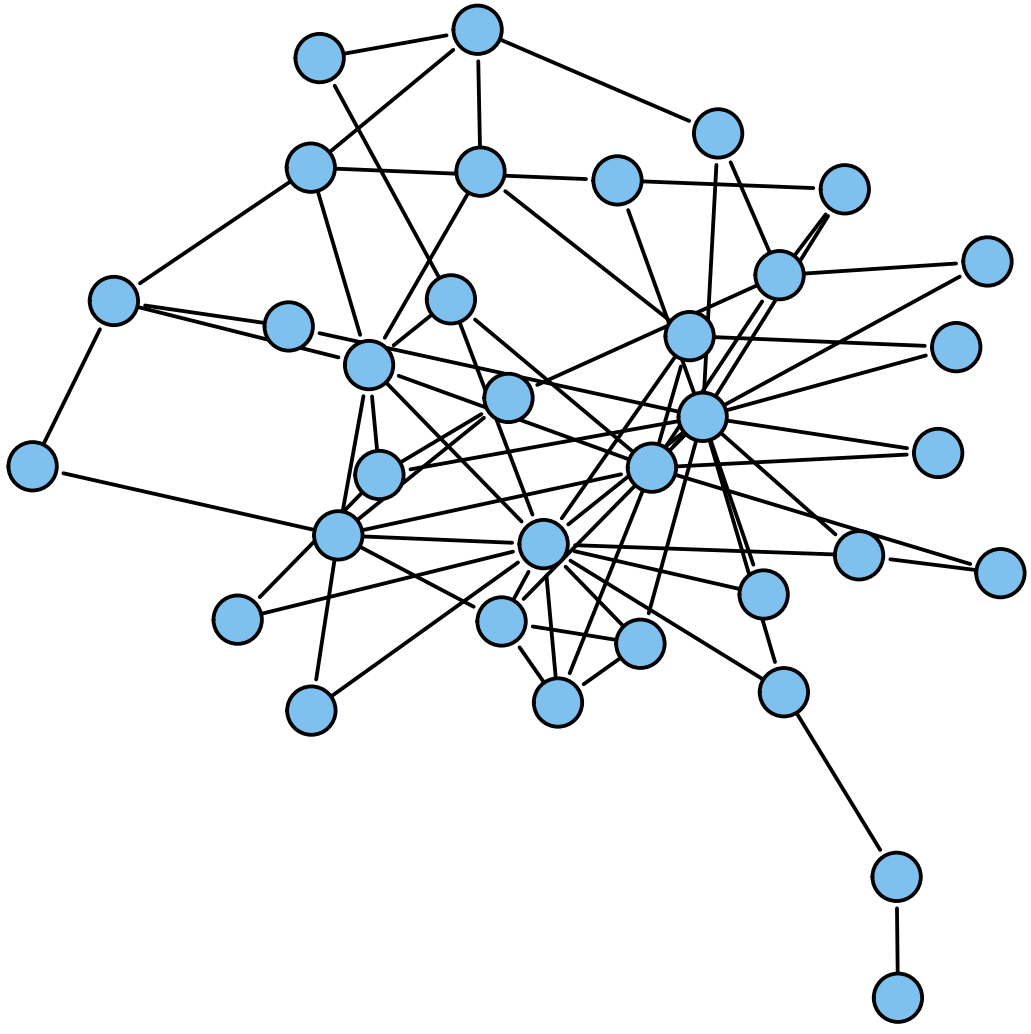
4 classes, modularité $\simeq 0.42$

Exemple



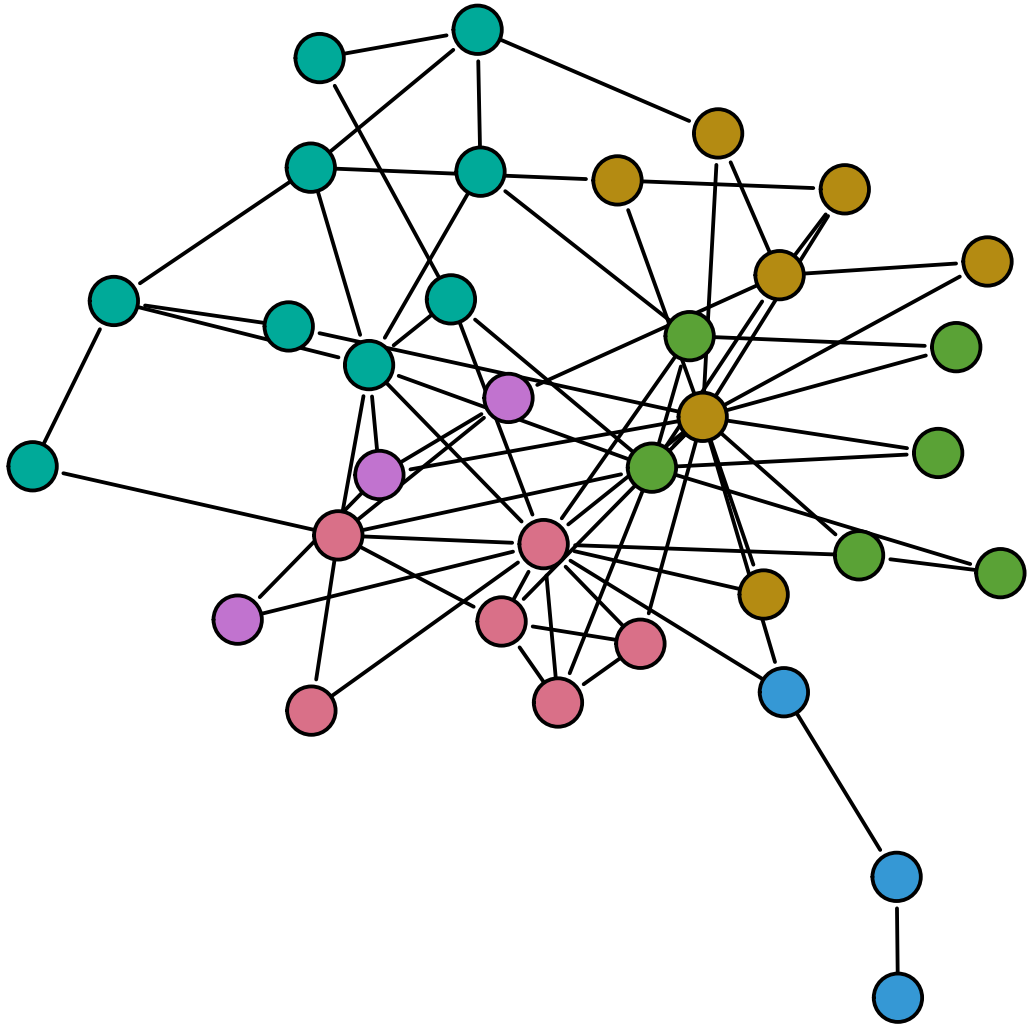
Modèle de configuration : mêmes degrés

Exemple



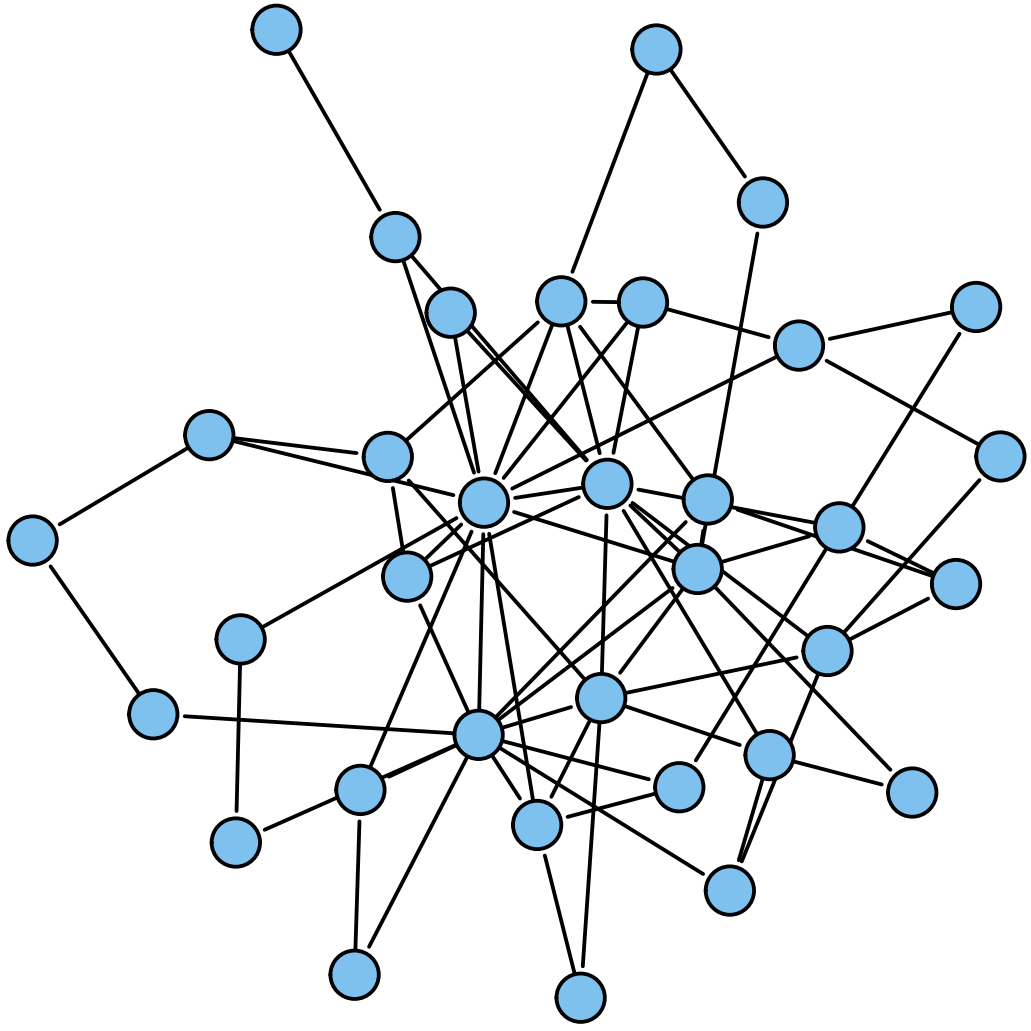
Repositionné

Exemple



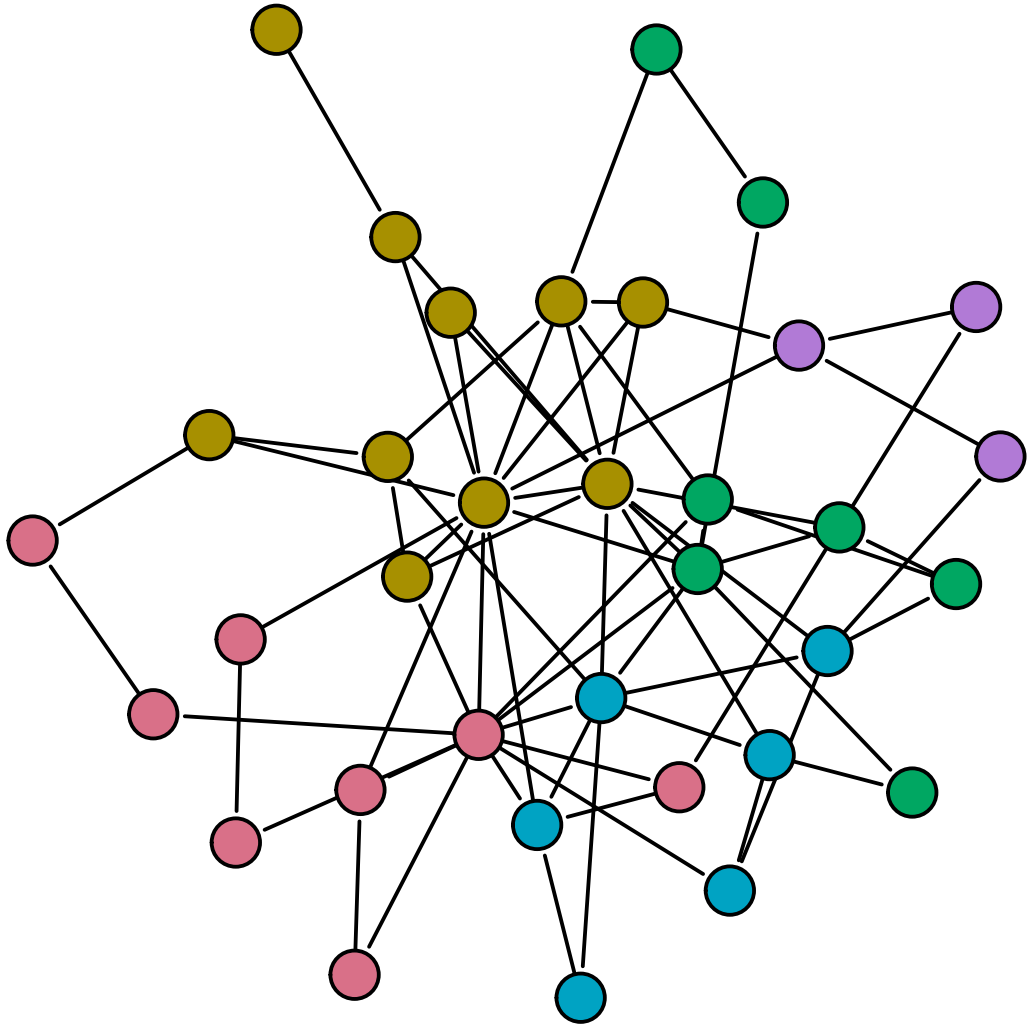
6 classes, modularité $\simeq 0.35$

Exemple



Nouveau tirage

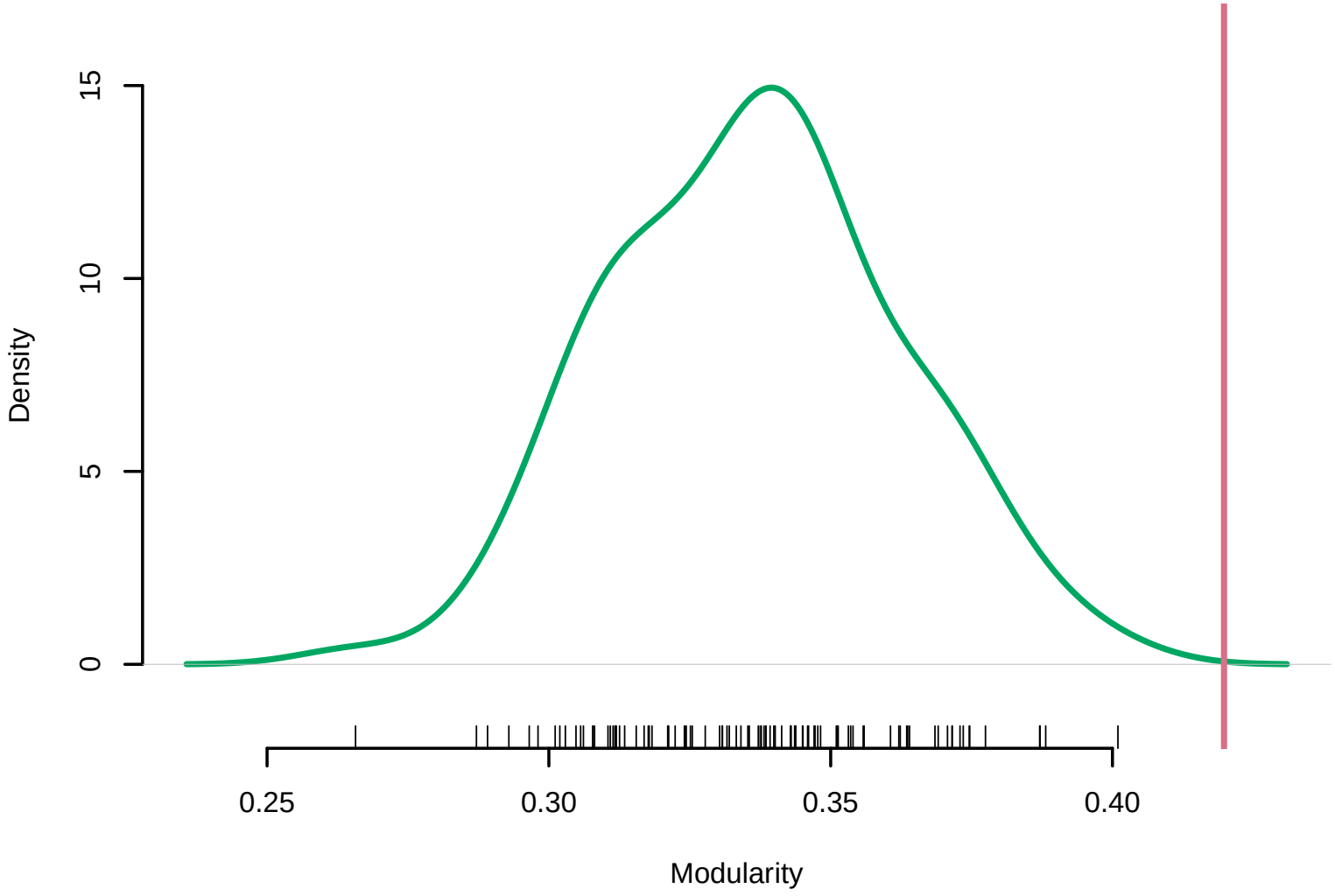
Exemple



5 classes, modularité $\simeq 0.34$

Exemple

100 graphes aléatoires

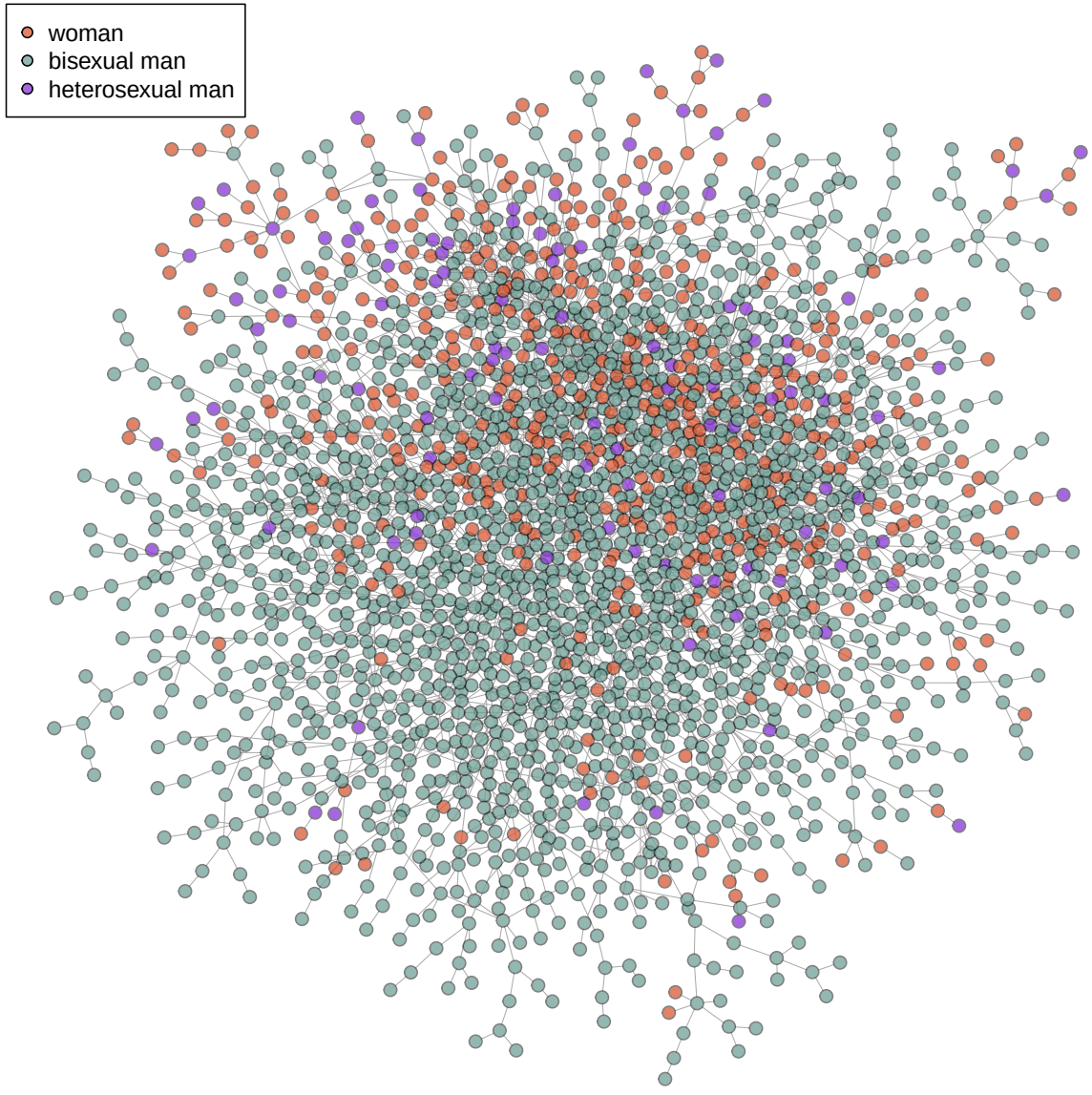


la classification sur le graphe d'origine a un sens

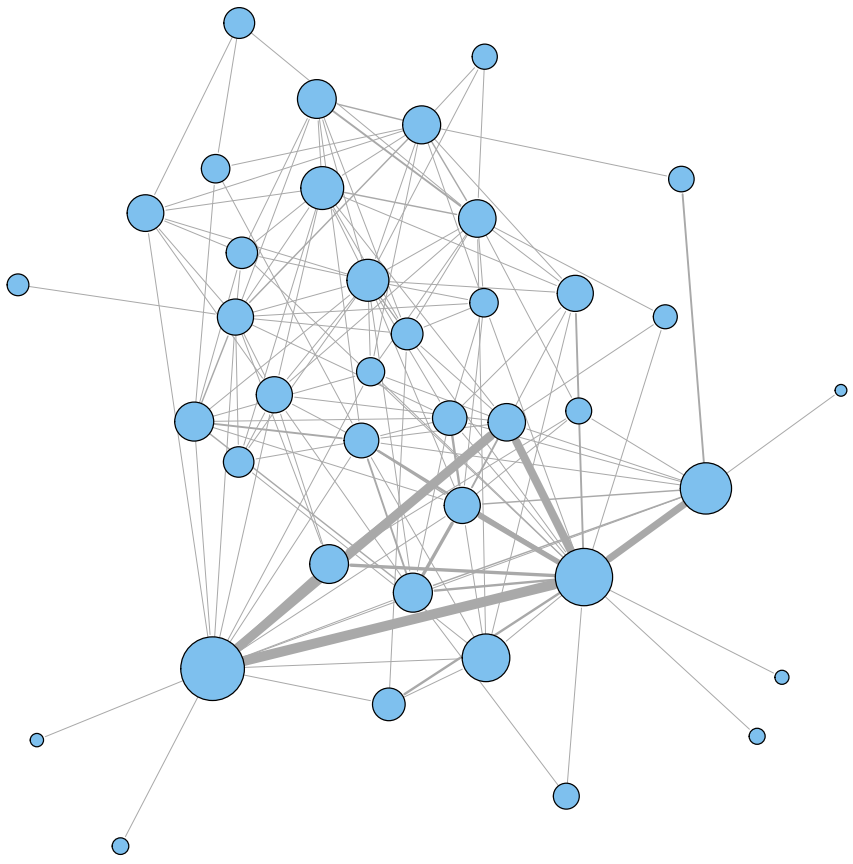
VIH à Cuba

- suivi du VIH/SIDA à Cuba de 1986 à 2004
- suivi d'infection étendu : on demande à chaque nouveau patient d'identifier ses partenaires sexuels durant les deux années avant sa détection
- base de données volumineuse :
  - 5 389 patients décrits pas diverses variables (genre, orientation sexuelle, date de naissance, etc.)
  - 4 073 relations (graphe assez peu dense)
  - 2 386 patients dans une même composante connexe du graphe (3 168 relations dans cette composante)
- objectifs de l'analyse :
  - structure de l'épidémie
  - prévention

Composante connexe principale

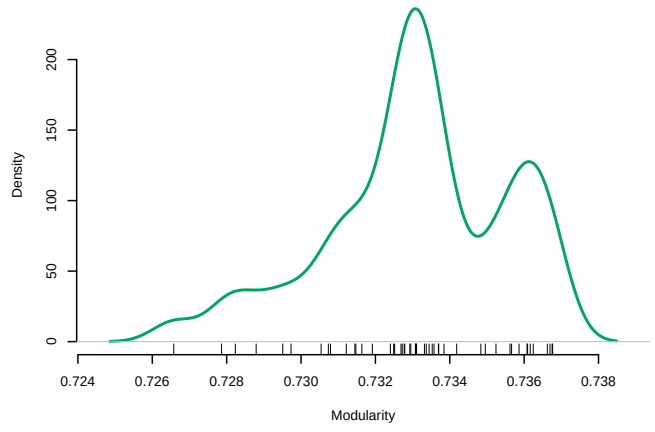


Classification



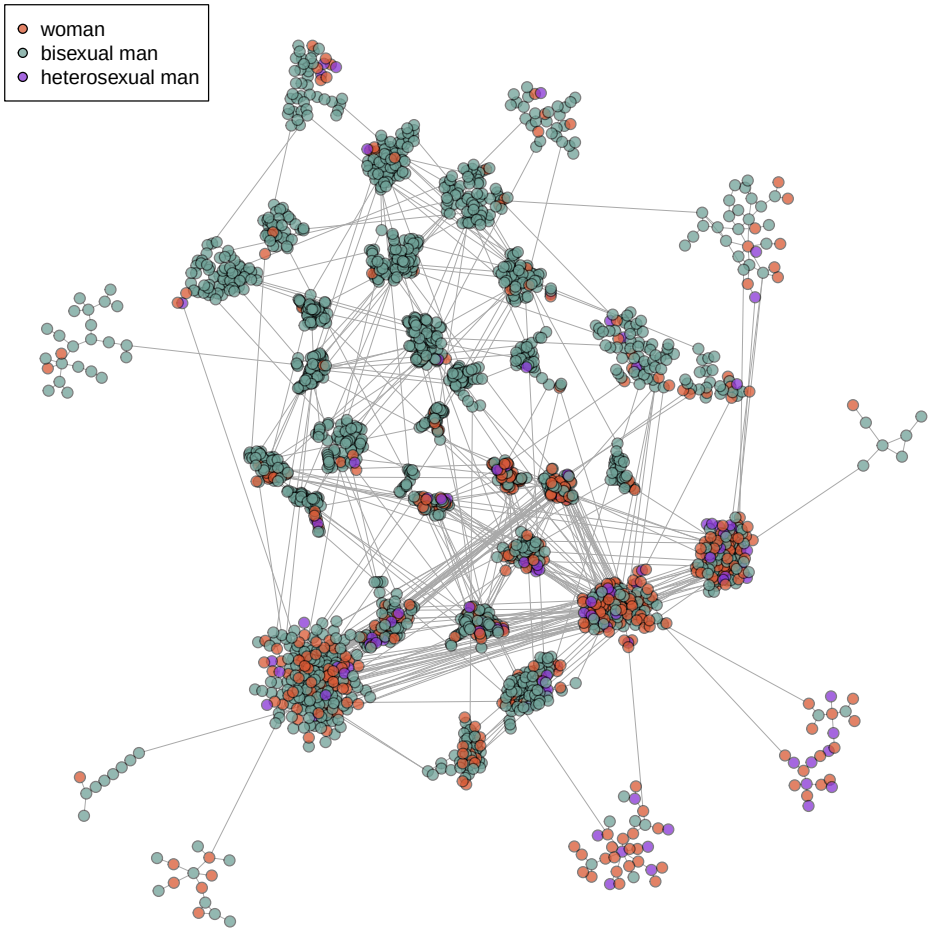
$\Rightarrow$ 39 classes (89.5% des liens internes aux classes)

$\Rightarrow$ modularité $\simeq 0.85$



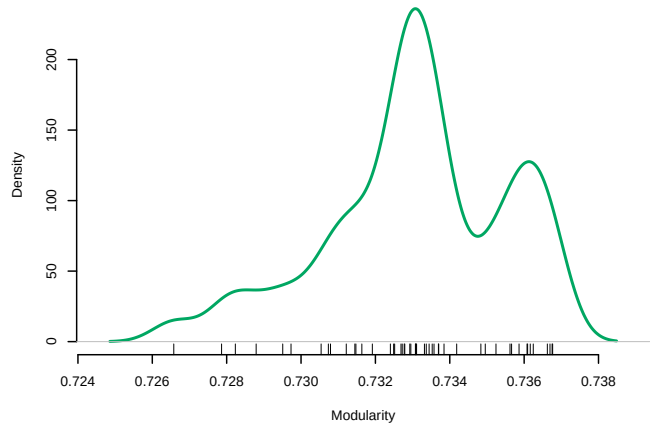
$\Rightarrow$ modularité « aléatoire » ≤ 0.74

Classification



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

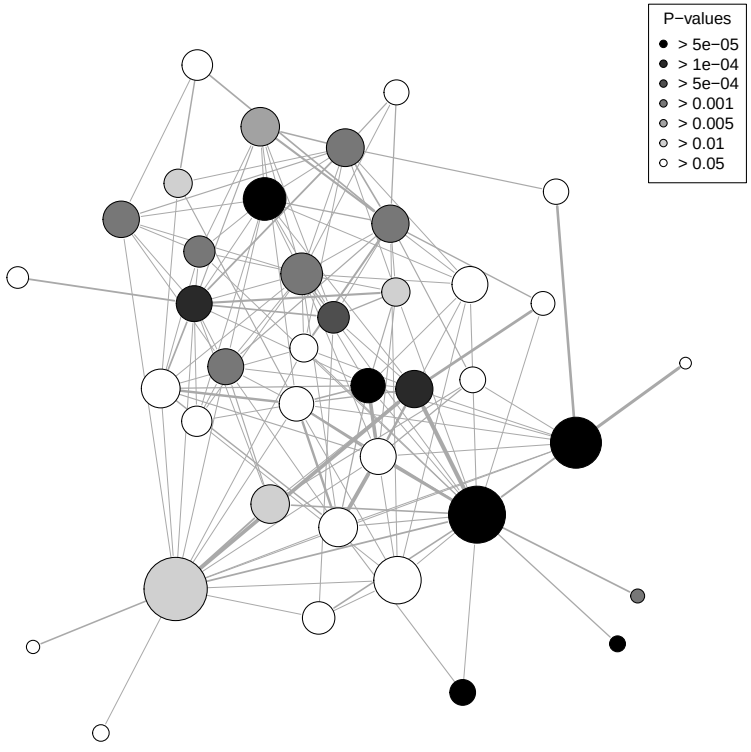
⇒ modularité $\simeq 0.85$



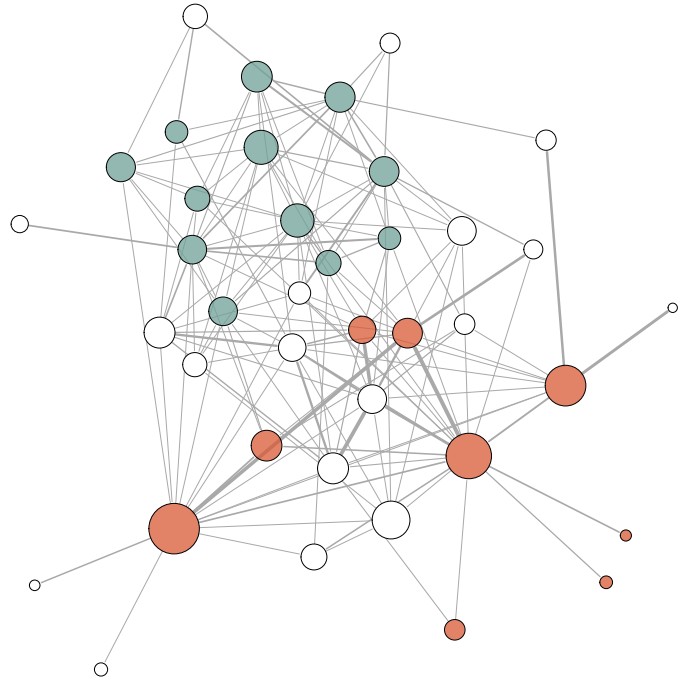
⇒ modularité « aléatoire » ≤ 0.74

⇒ visualisation hiérarchique de l'orientation sexuelle

Statistique graphique

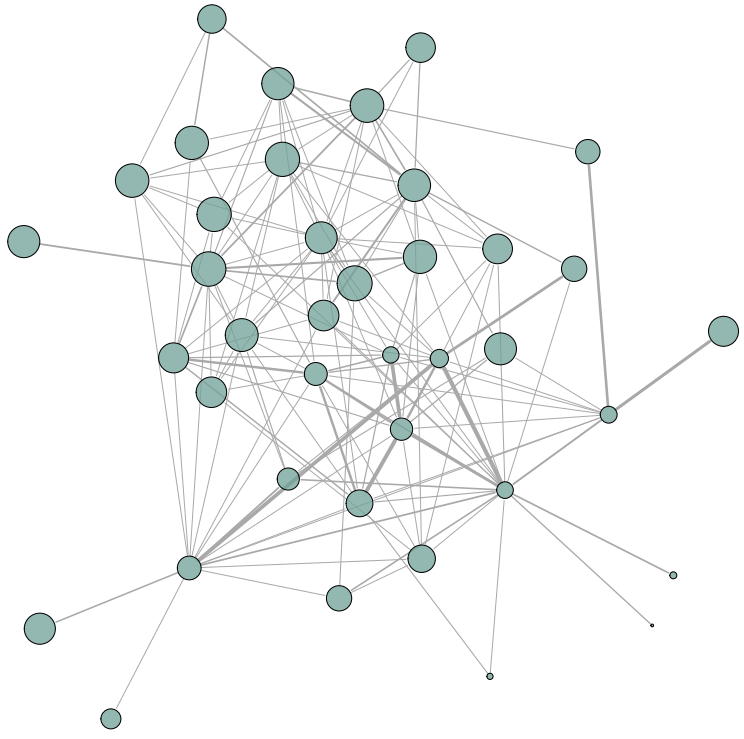


p-value d'un test du χ^2 sur la distribution de l'orientation sexuelle

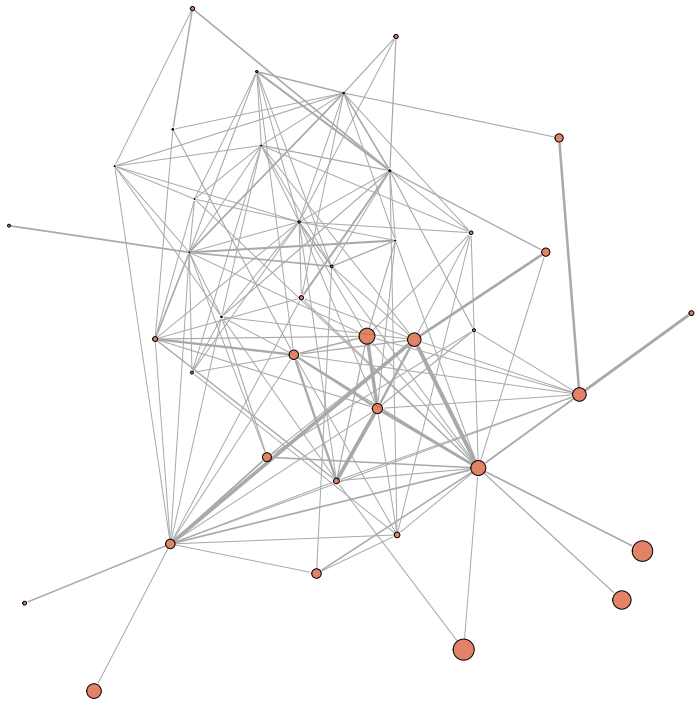


orientation sexuelle atypique

Statistique graphique



Hommes homosexuels



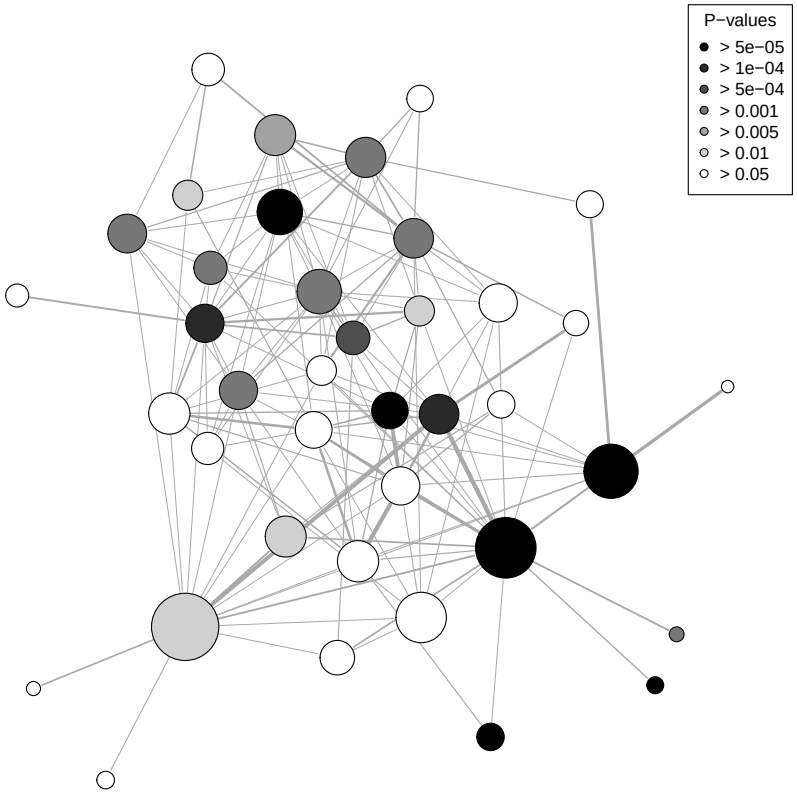
Femmes

Pourcentages

Hiérarchie

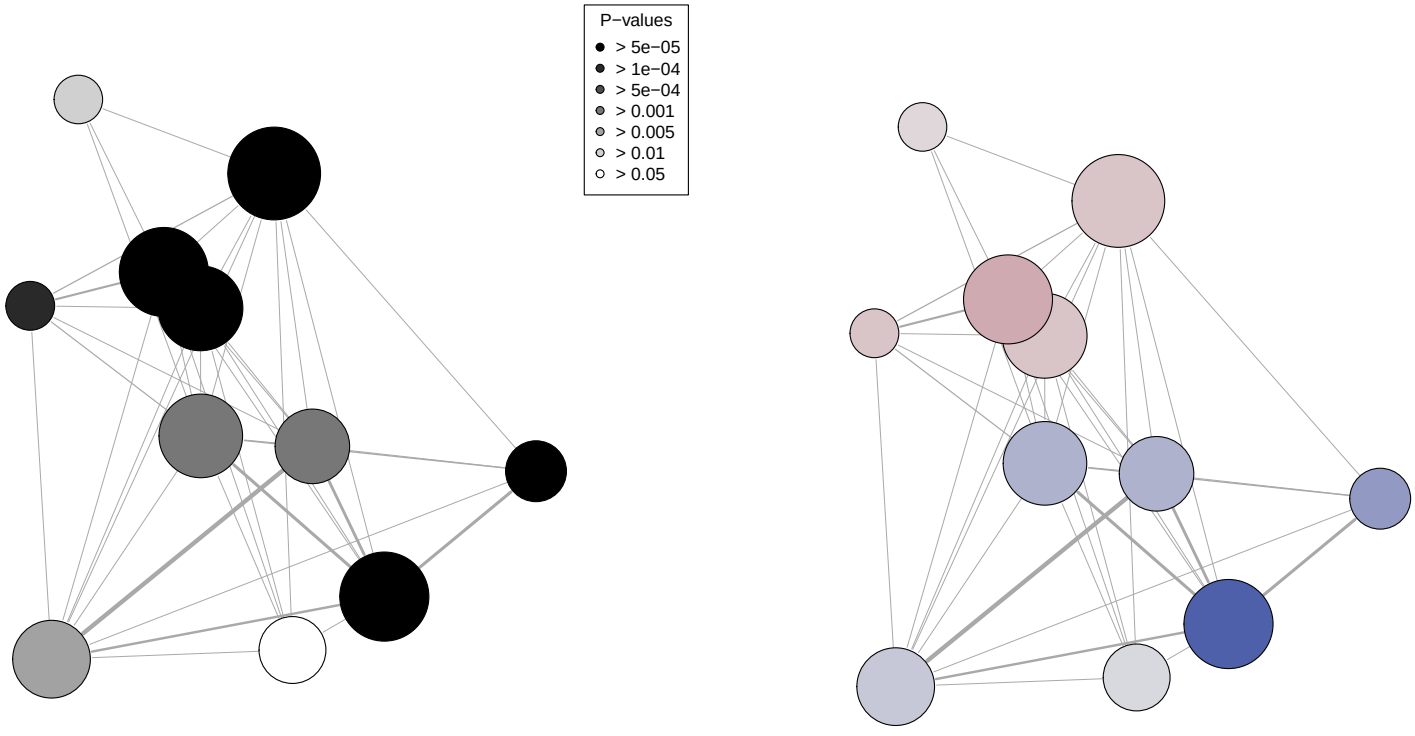
- moins de détails :
  - poursuite de l'algorithme glouton de classification
  - fusion de classes (contrainte hiérarchique)
  - visualisation barycentrique
  - maintient de la modularité au dessus du seuil aléatoire
- plus de détails :
  - classification des classes
  - liens externes supprimés
  -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

Simplification



p-value

Simplification

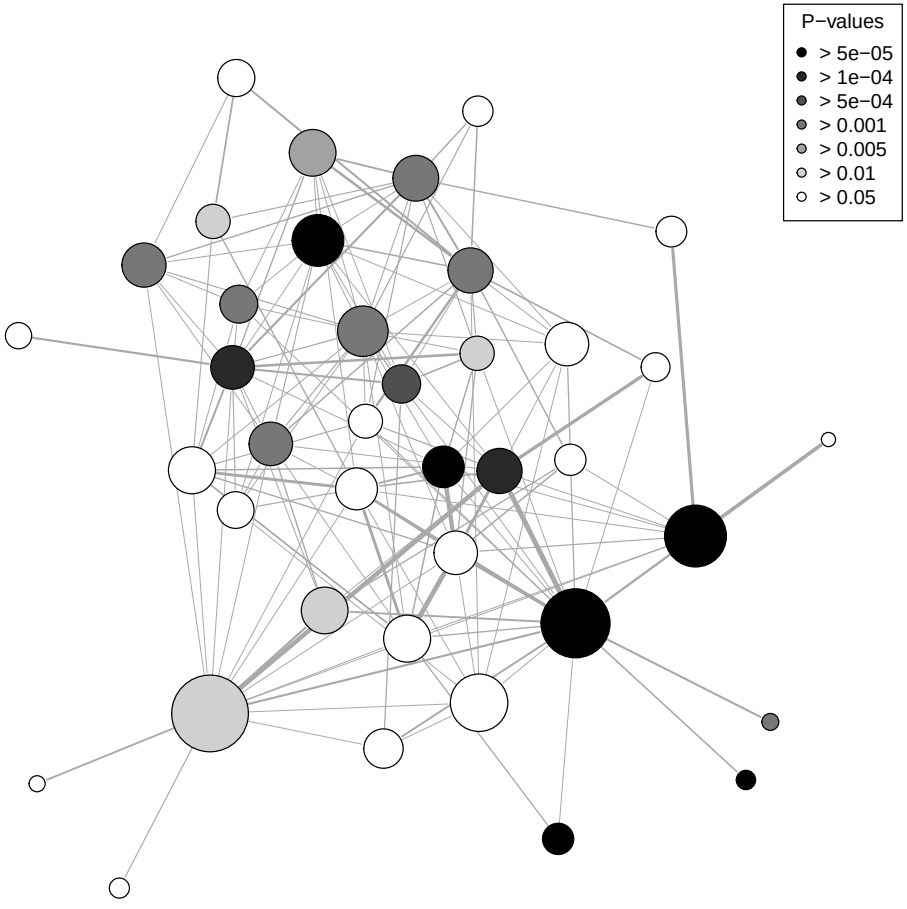


p-value

résidus de *Pearson*

Confirme la structure en deux parties

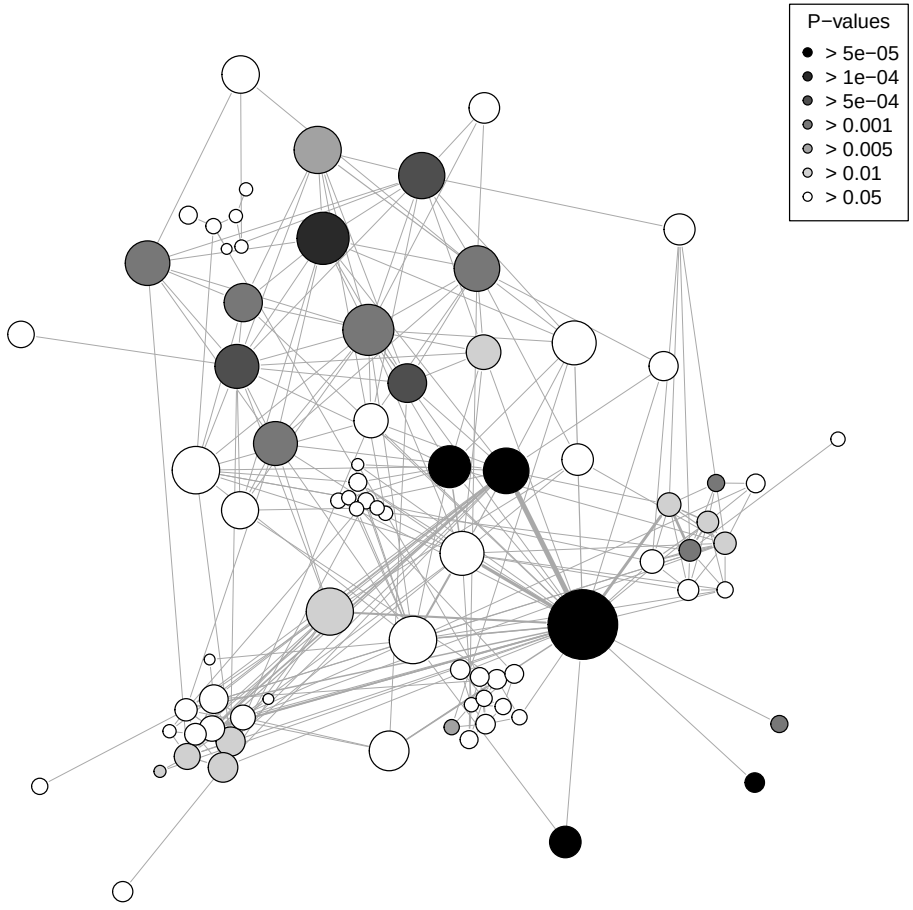
Sous-structures



$\Rightarrow$ 5 classes possèdent une sous structure

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

Sous-structures

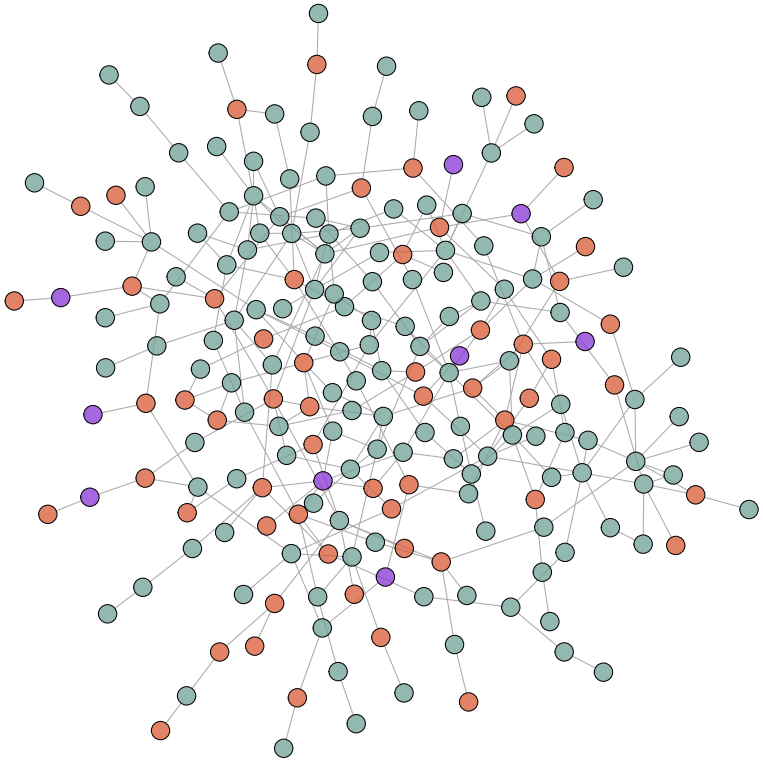


⇒ 5 classes possèdent une sous structure

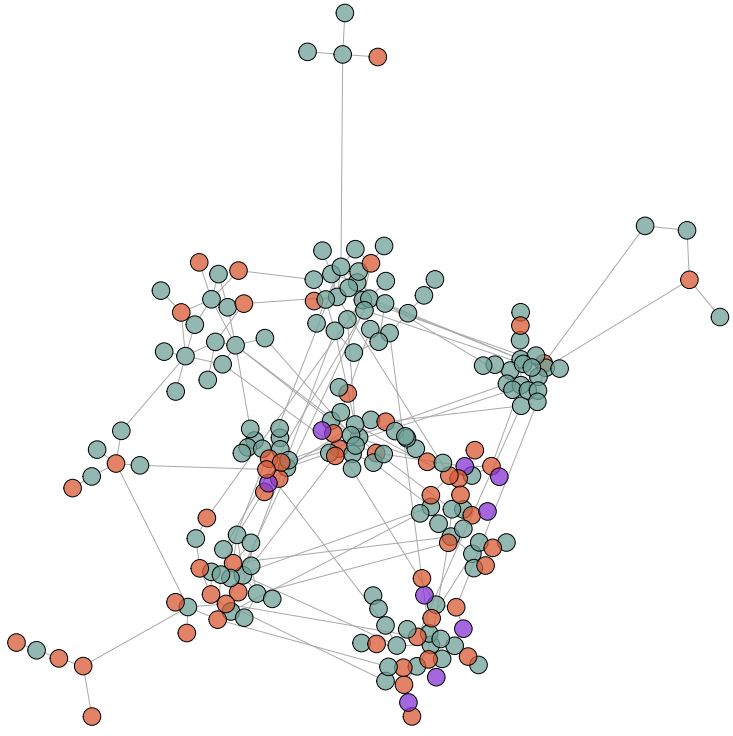
⇒ la modularité se maintient au dessus de 0.81

⇒ sous structures atypiques

Analyse locale



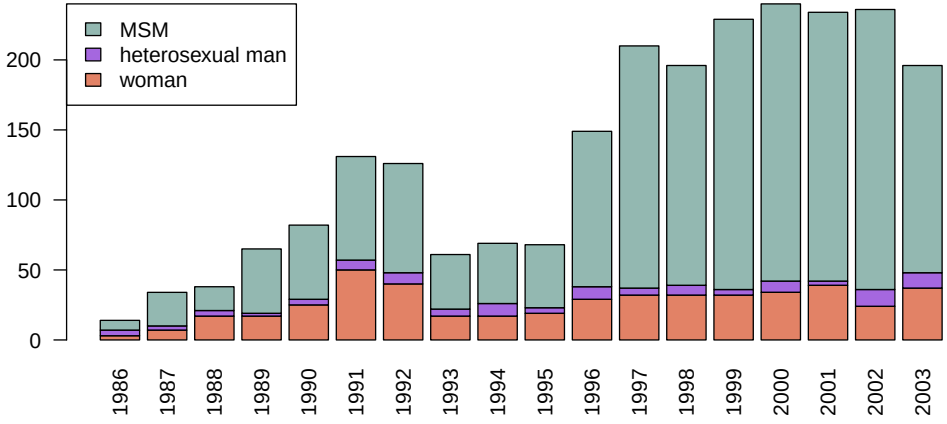
Visualisation classique

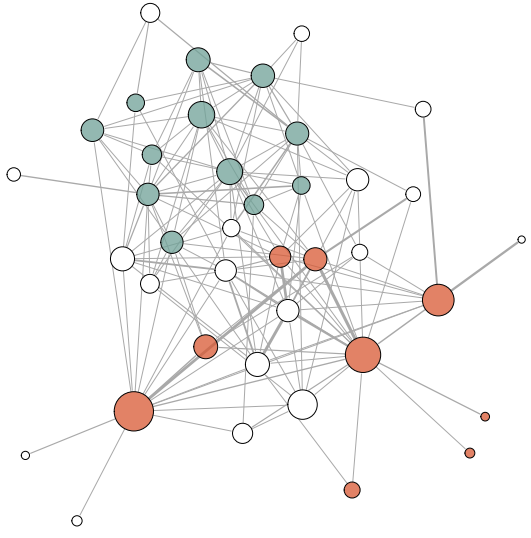


Visualisation hiérarchique

Aspect temporel

Recrutement annuel

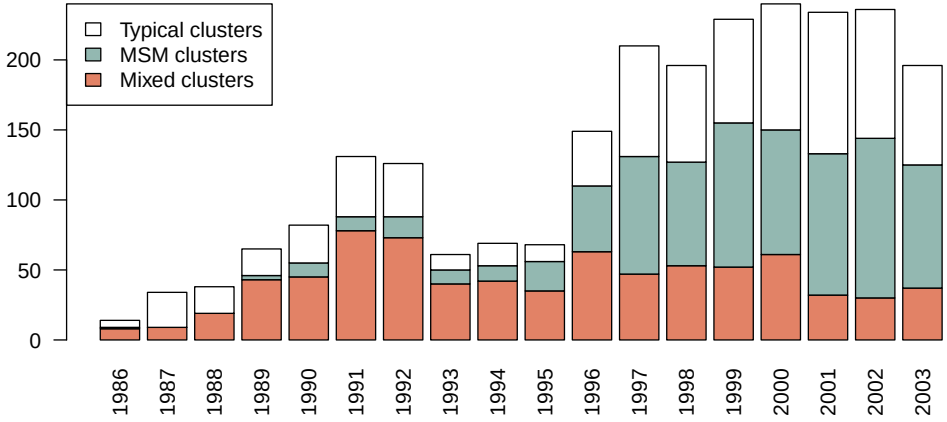


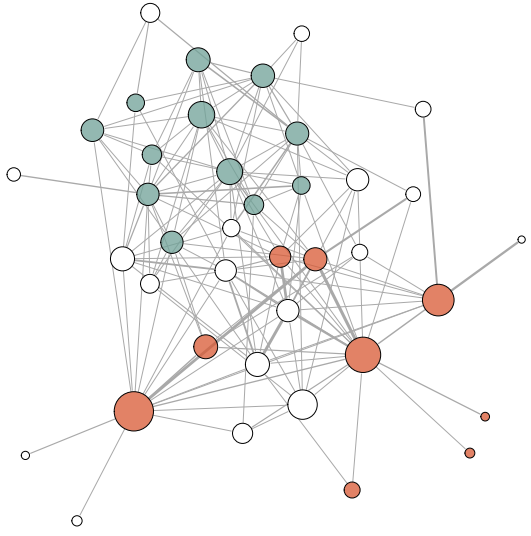


Pas de prise en compte explicite du temps, mais connexions « datées »

Aspect temporel

Recrutement annuel





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

Conclusion

- exploration visuelle de données relationnelles :
  - classification
  - rendu hiérarchique avec simplification ou détails
  - classes significatives
  - statistique graphique
- perspectives :
  - aspect temporel explicite
  - graphes bipartis
  - autres critères
  - ...