# Introduction à la fouille de graphes

Fabrice Rossi

Université Paris 1 Panthéon Sorbonne

14 Novembre 2017

Fabrice.Rossi@apiacoa.org
http://apiacoa.org/

# Plan

# Plan

# Graphes

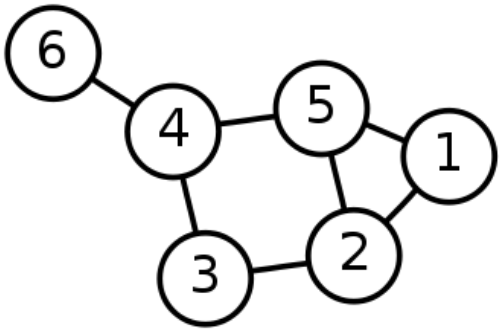
## Un graphe

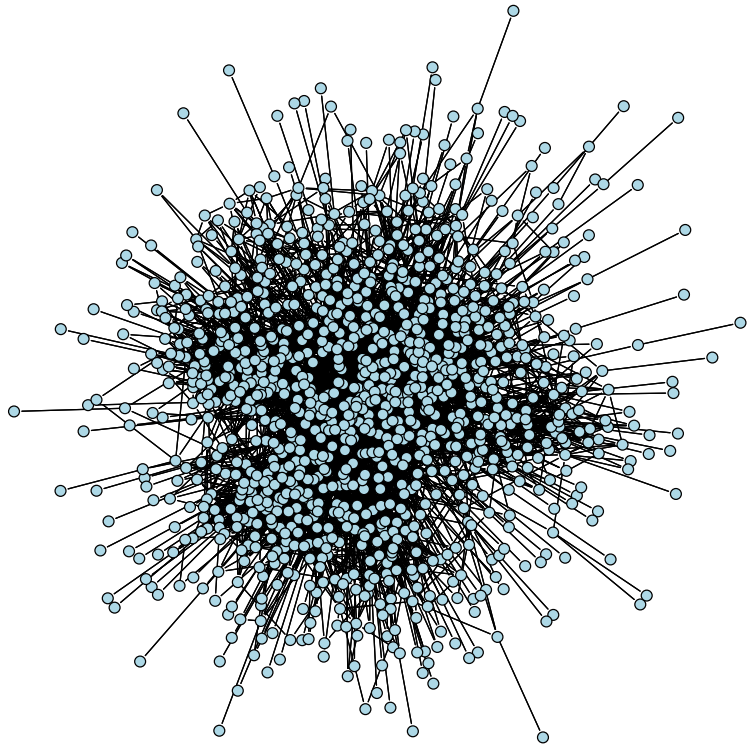
Un ensemble d'objets dont certains sont reliés

# Graphes

## Un graphe

Un ensemble d'objets dont certains sont reliés



élémentaire... ou pas...

Source du graphe de gauche `https://en.wikipedia.org/wiki/File:6n-graf.svg`

# Graphes et réseaux

## Graphe

- modèle très général pour des interactions
- ne pas confondre avec un réseau (objet réel versus objet mathématique)
- ne pas confondre avec un graphique

## Omniprésent concrètement

- réseaux sociaux : téléphonie, email, forums, facebook, twitter, etc.
- internet : réseau physique, web, peer-to-peer, etc.
- infrastructure : réseau routier, téléphonique, électrique, etc.
- biologie : réseau neuronal, interaction entre gènes, etc.
- etc.

# Du réseau au graphe

## Un acte de modélisation

### Réseau « téléphonique »

- infrastructure physique (réseau filaire, RTC) :
  - graphe des points terminaux ?
  - graphe des infrastructures (commutateurs, centraux, NRA) ?
  - graphe « géographique » ?
- réseau social :
  - appels reçus ?
  - appels émis ?
  - durée, fréquence des appels ?

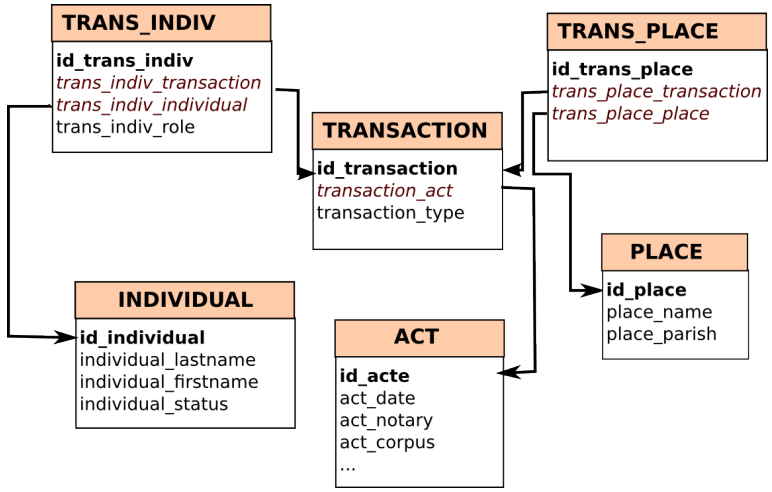
### Choix du modèle

- plusieurs échelles
- agrégation
- inférence de lien

# Les graphes comme modèles

## Omniprésent logiquement

- vision large de l'interaction : modèle relationnel des bases de données
- données relationnelles : données classiques + données d'interaction




## Apports du modèle

- naturel pour les réseaux
- nombreux outils mathématiques (distances, projections, caractérisations locales et globales, etc.)
- complexe, mais pas trop

# Grands graphes

## Graphes « de terrain »

- de quelques centaines à des millions d'objets et des milliards d'interaction
- analyse manuelle impossible

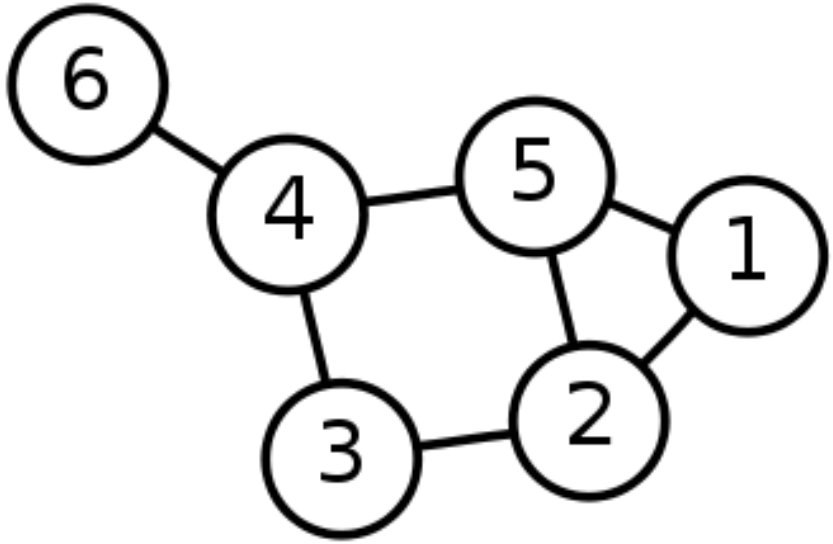
## Croissance constante

- pour les réseaux : de plus en plus de traces numériques
- pour les graphes : « big data » relationnelles

## Fouille de graphes

- visualisation
- caractéristiques locales et globales
- schémas fréquents
- classification

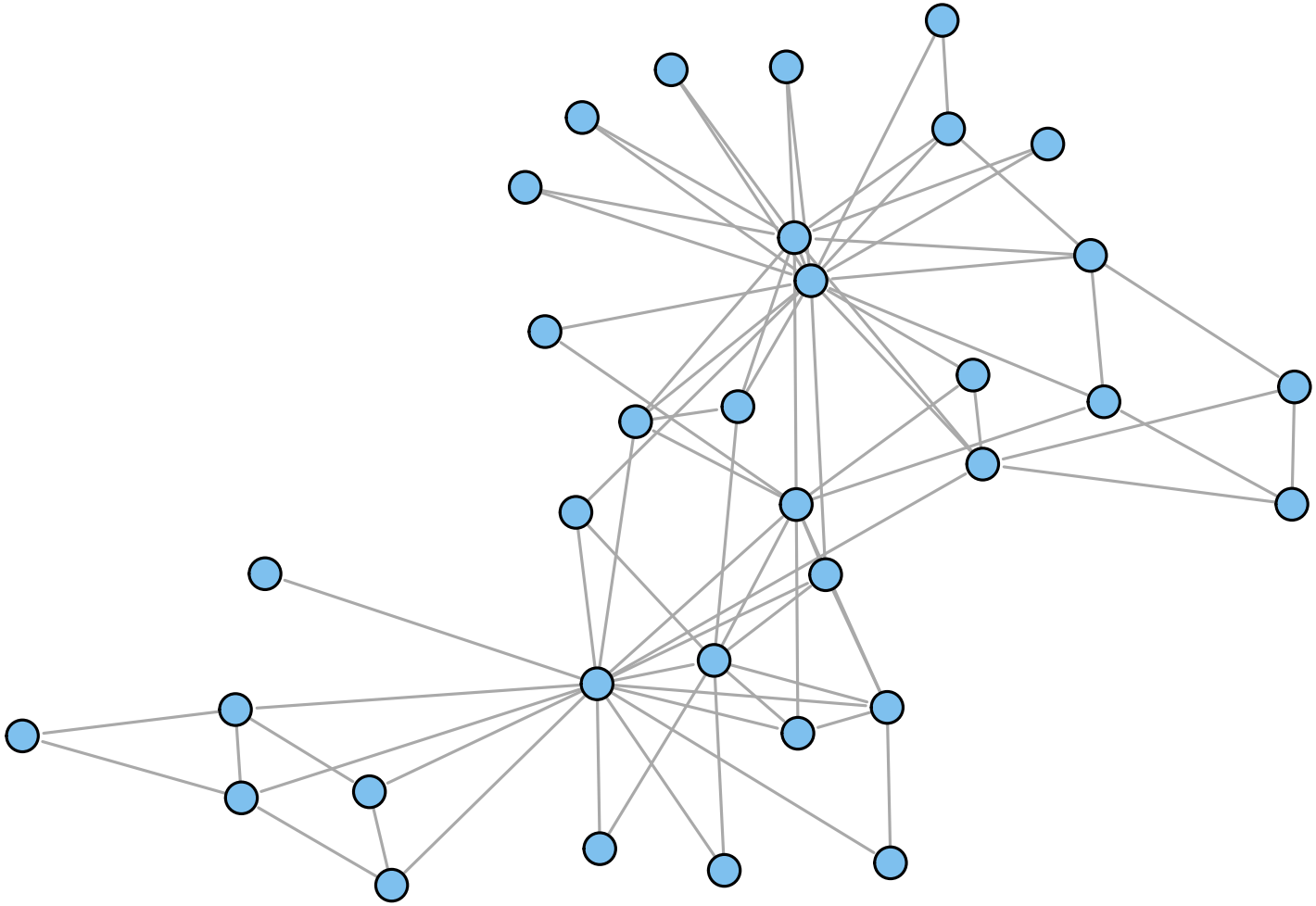
# Visualisation




6 objets

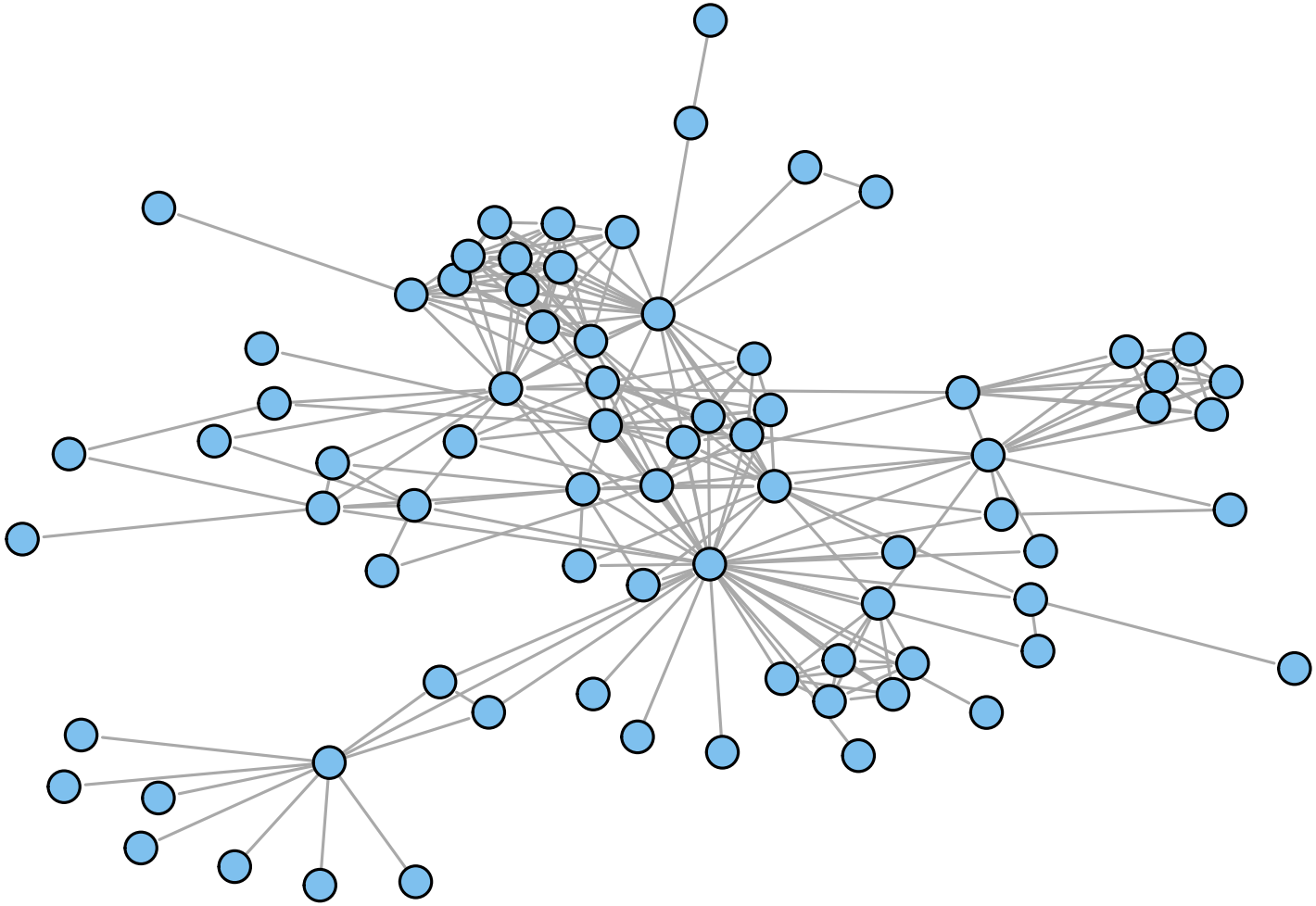
Source `https://en.wikipedia.org/wiki/File:6n-graf.svg`

# Visualisation



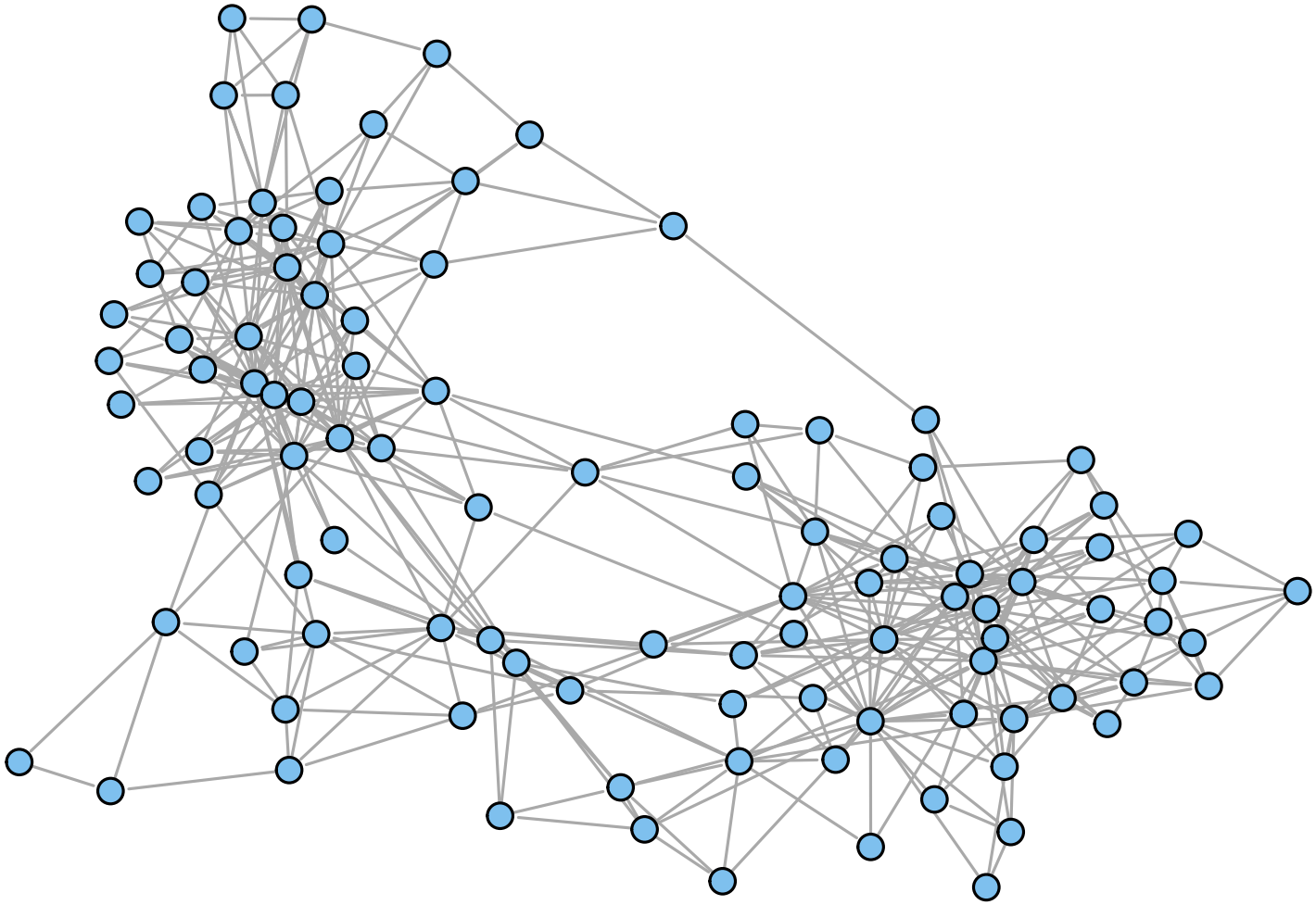
34 objets

# Visualisation



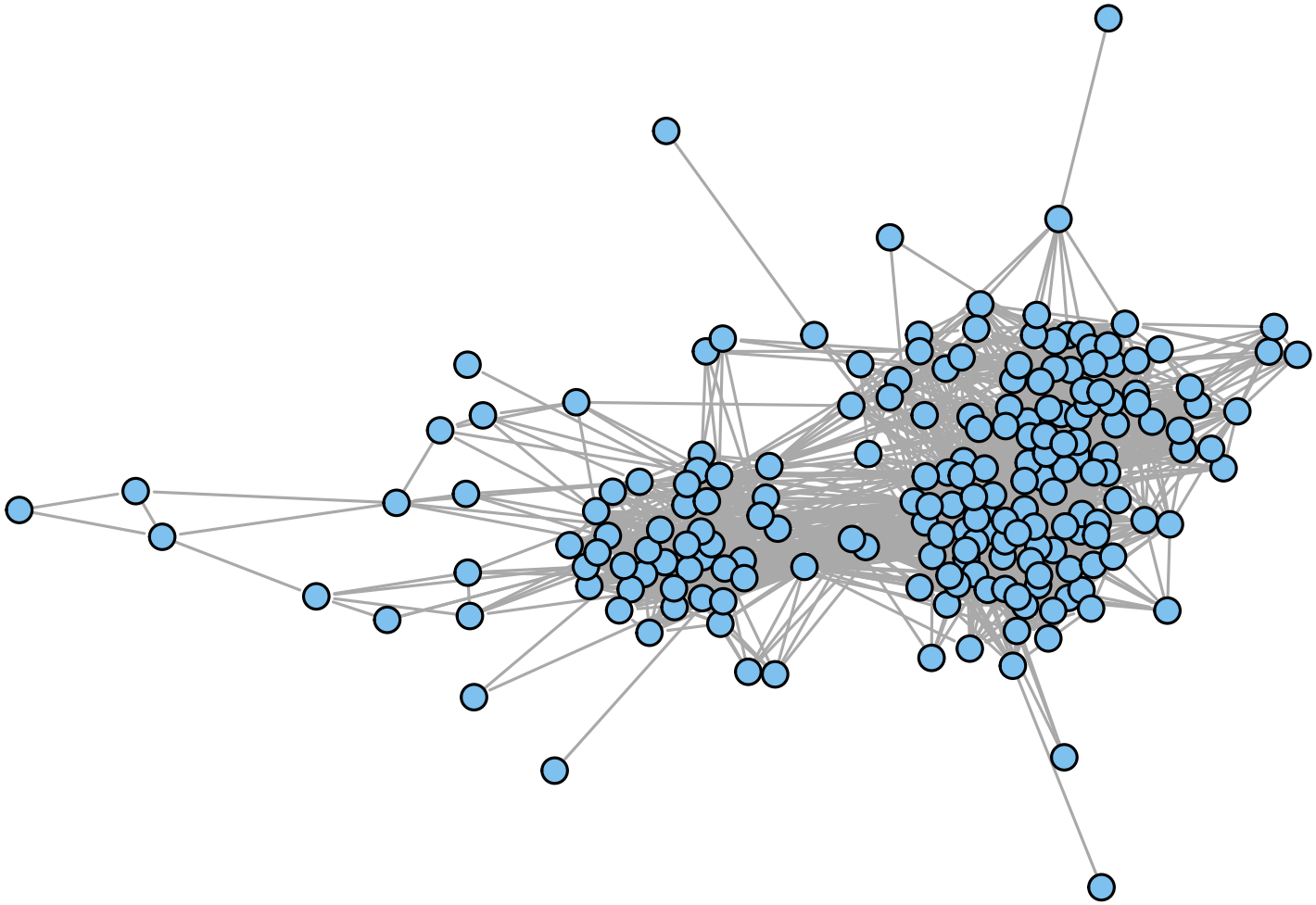
77 objets

# Visualisation



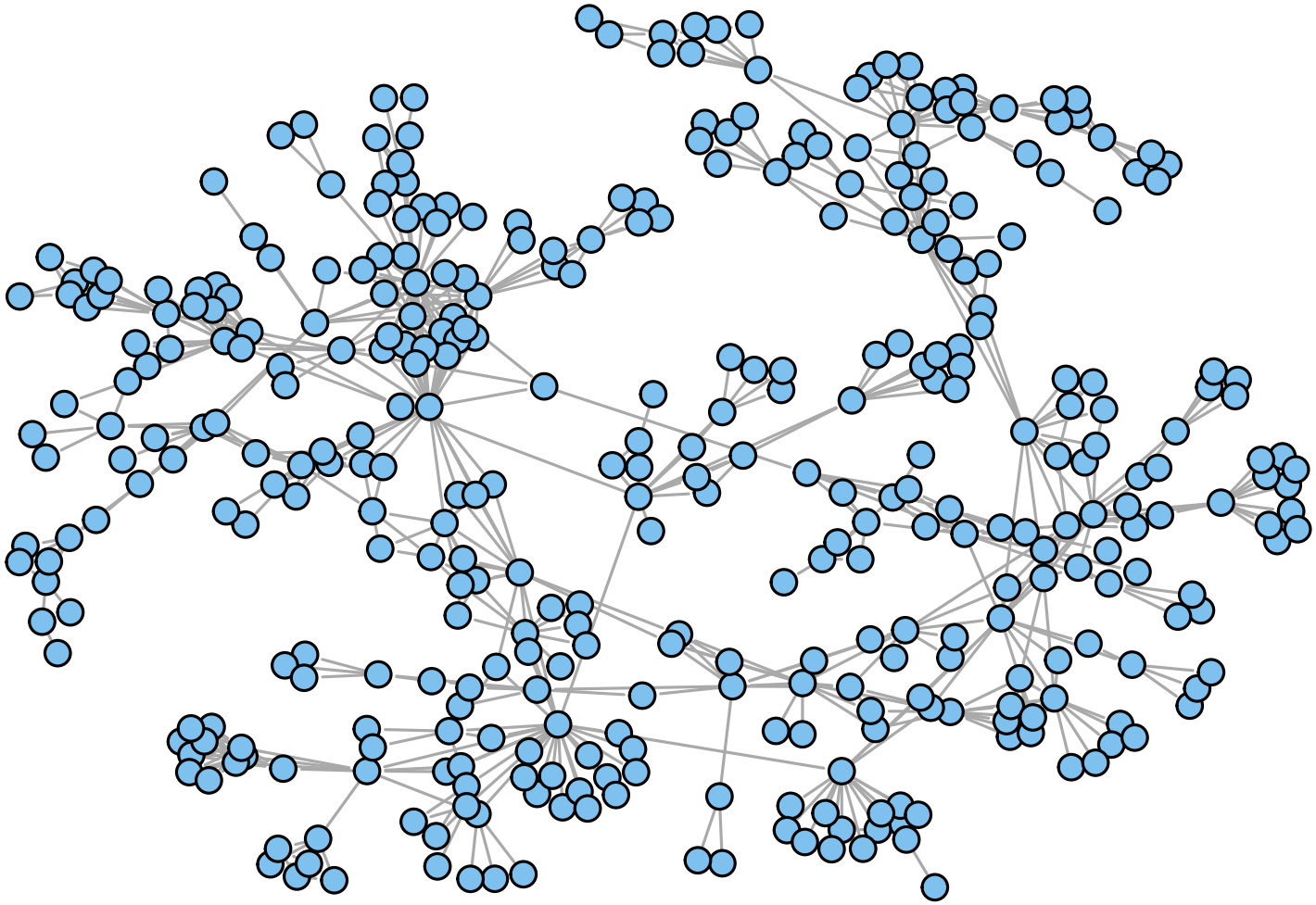
105 objets

# Visualisation



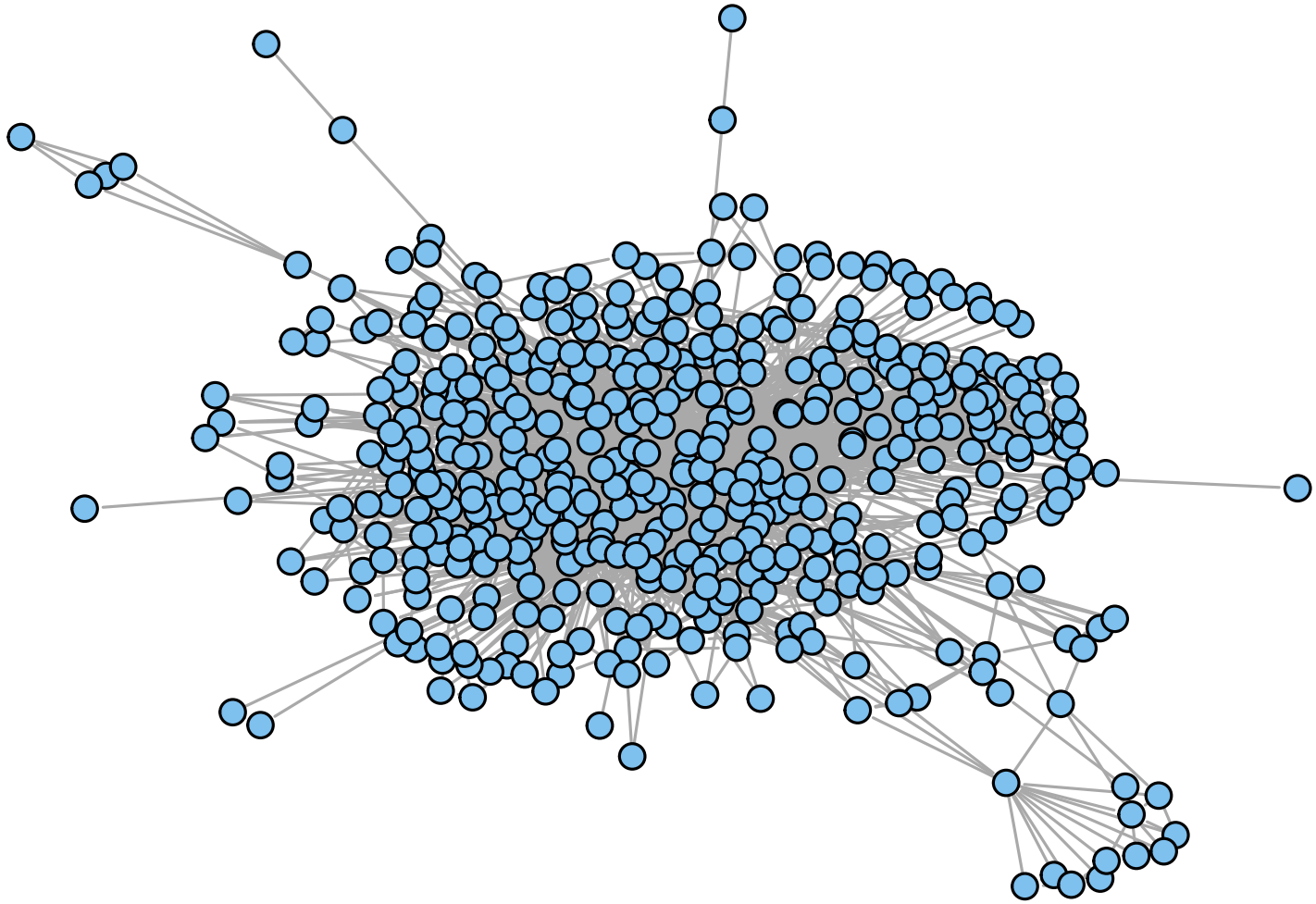
198 objets

# Visualisation



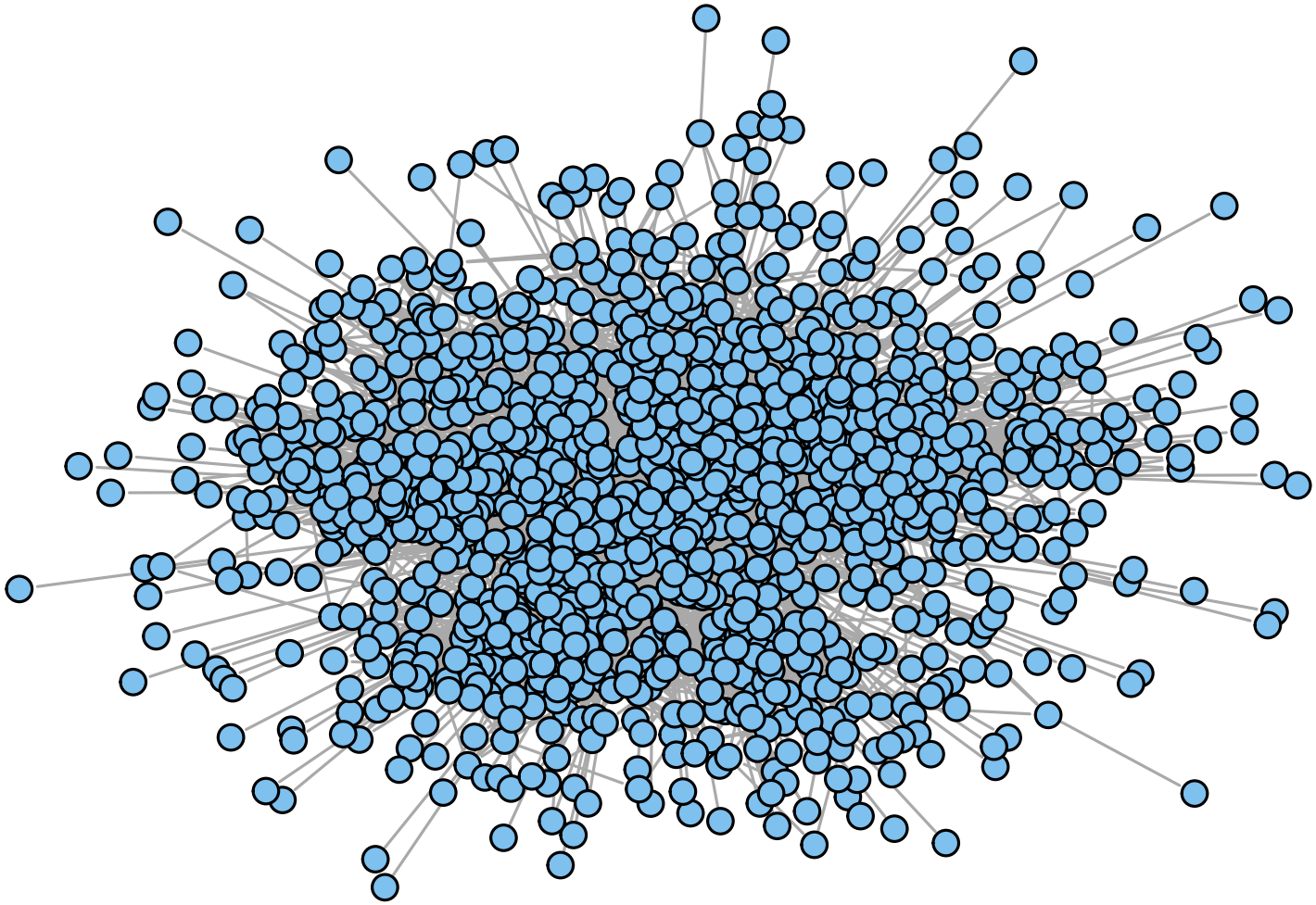
379 objets

# Visualisation



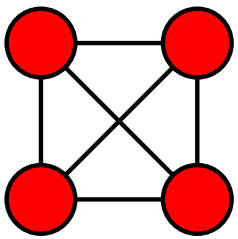
453 objets

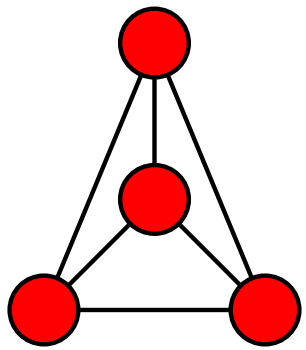
# Visualisation

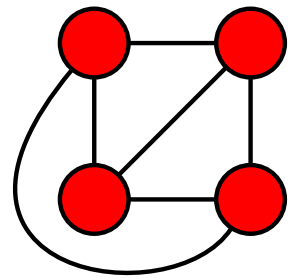


1133 objets

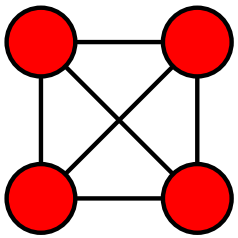
# Danger

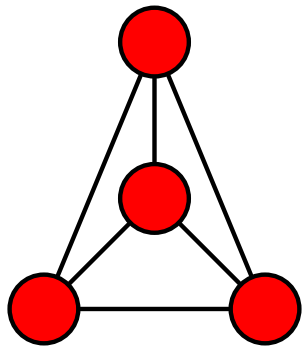


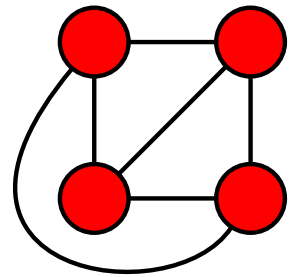




# Danger



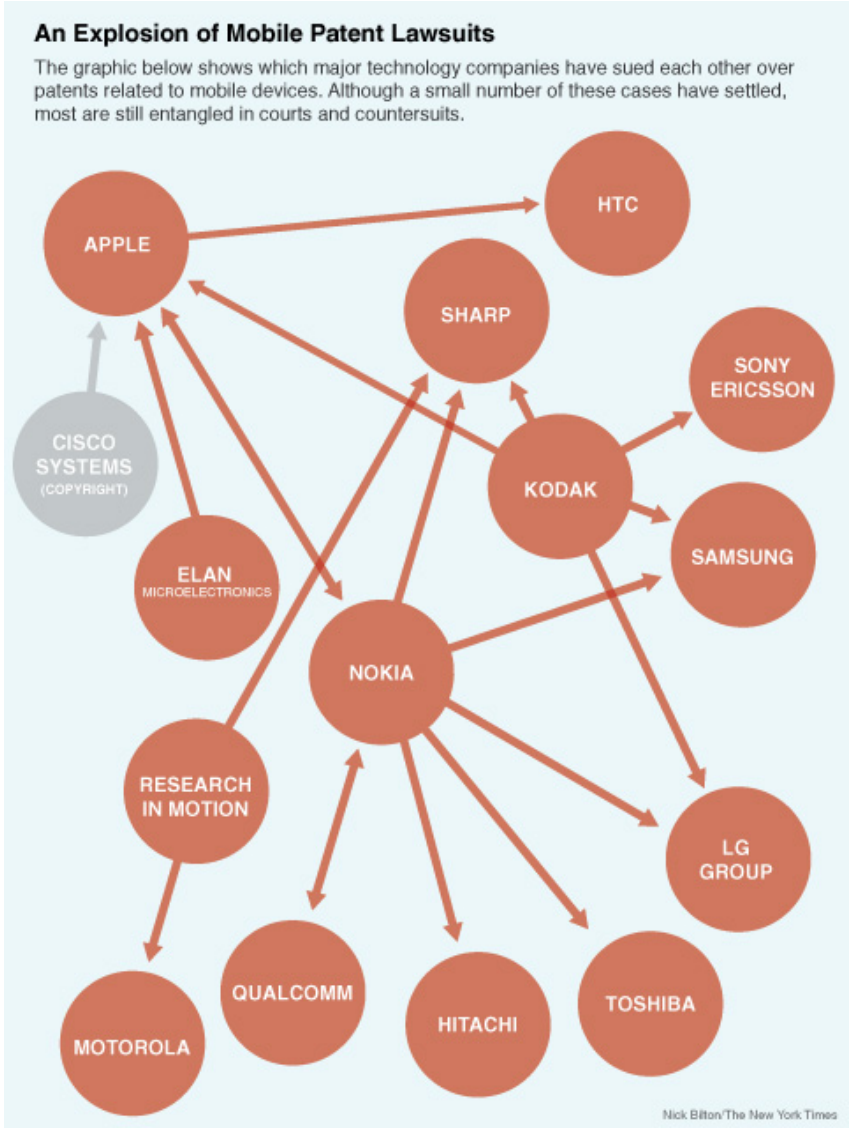




Message de la visualisation ?

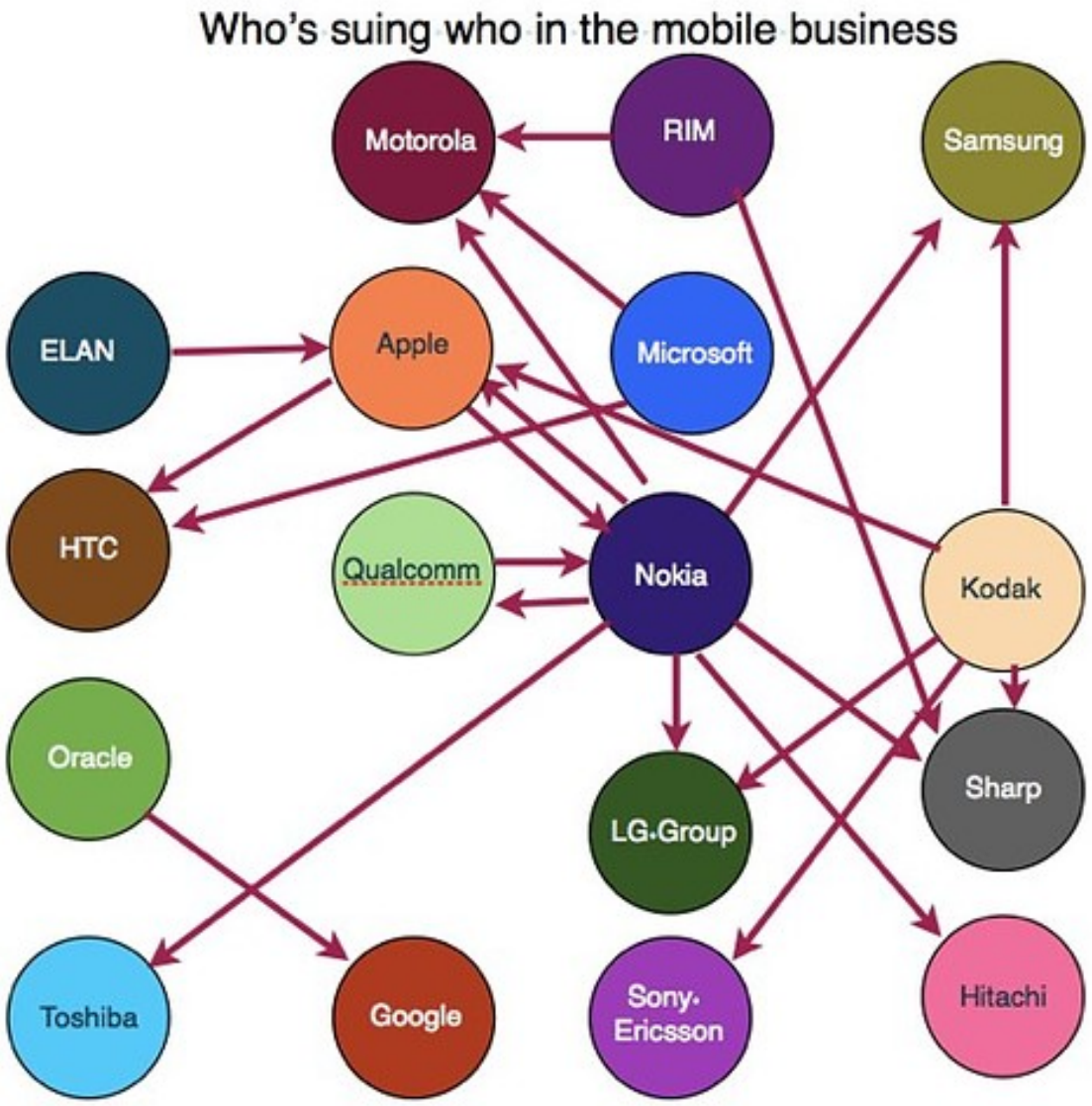
# Exemple concret

## Procès dans le mobile, autour de 2010



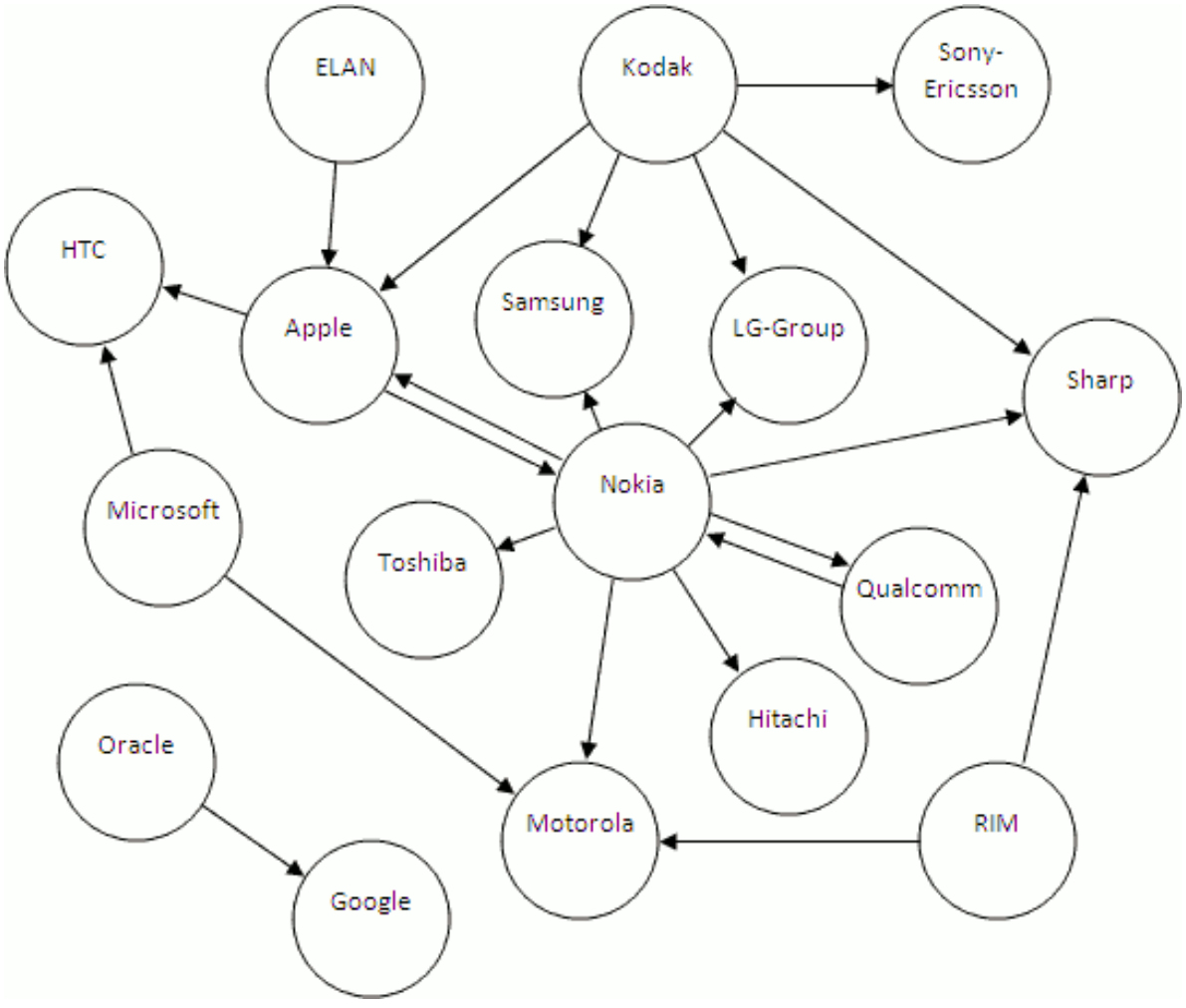
# Exemple concret

## Procès dans le mobile, autour de 2010



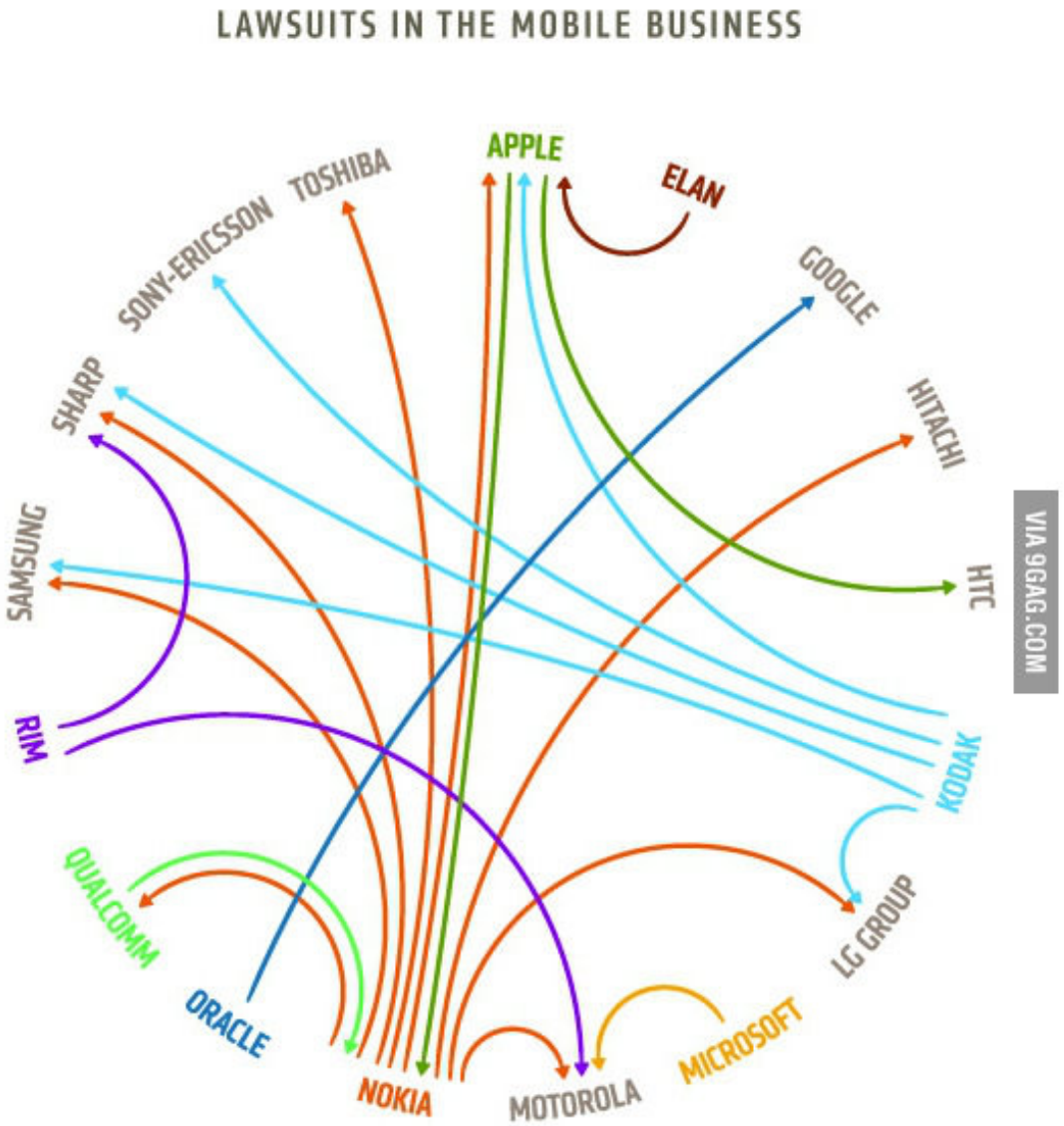
# Exemple concret

## Procès dans le mobile, autour de 2010



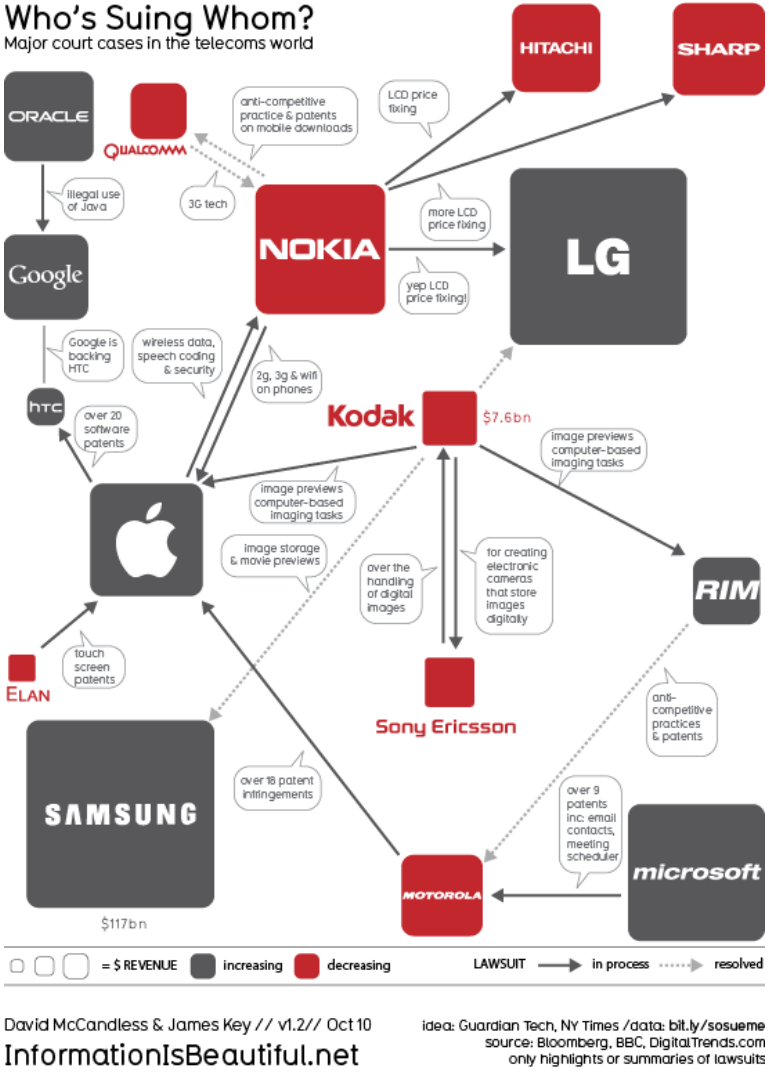
# Exemple concret

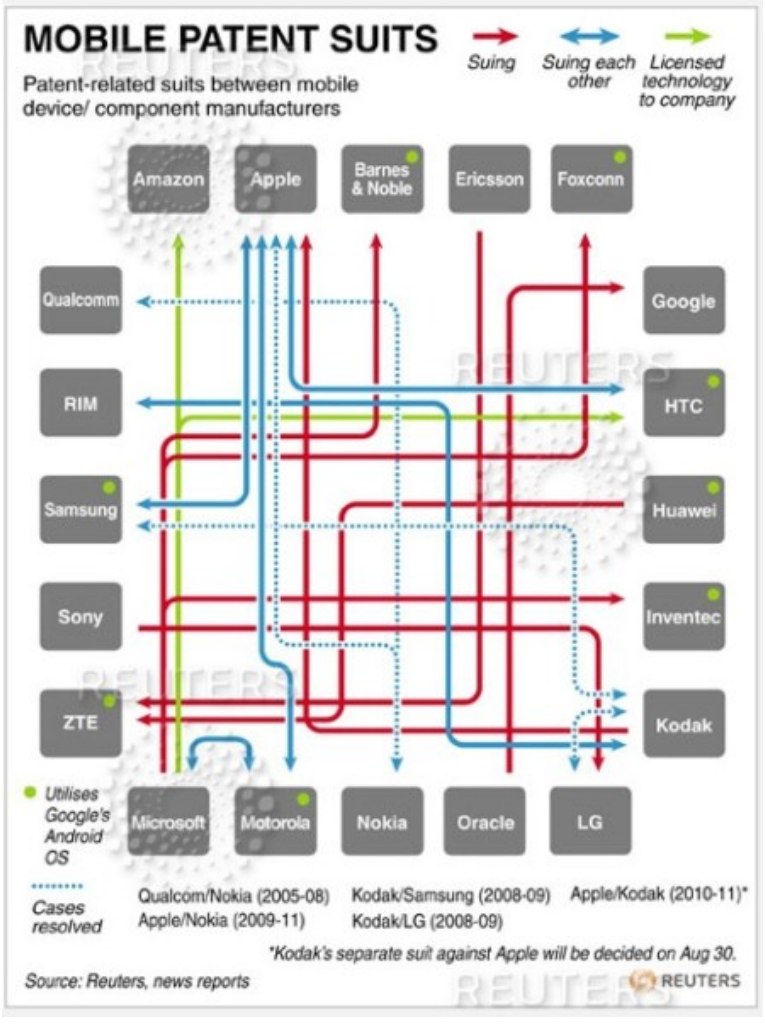
## Procès dans le mobile, autour de 2010



# Exemple concret

## Procès dans le mobile, autour de 2011



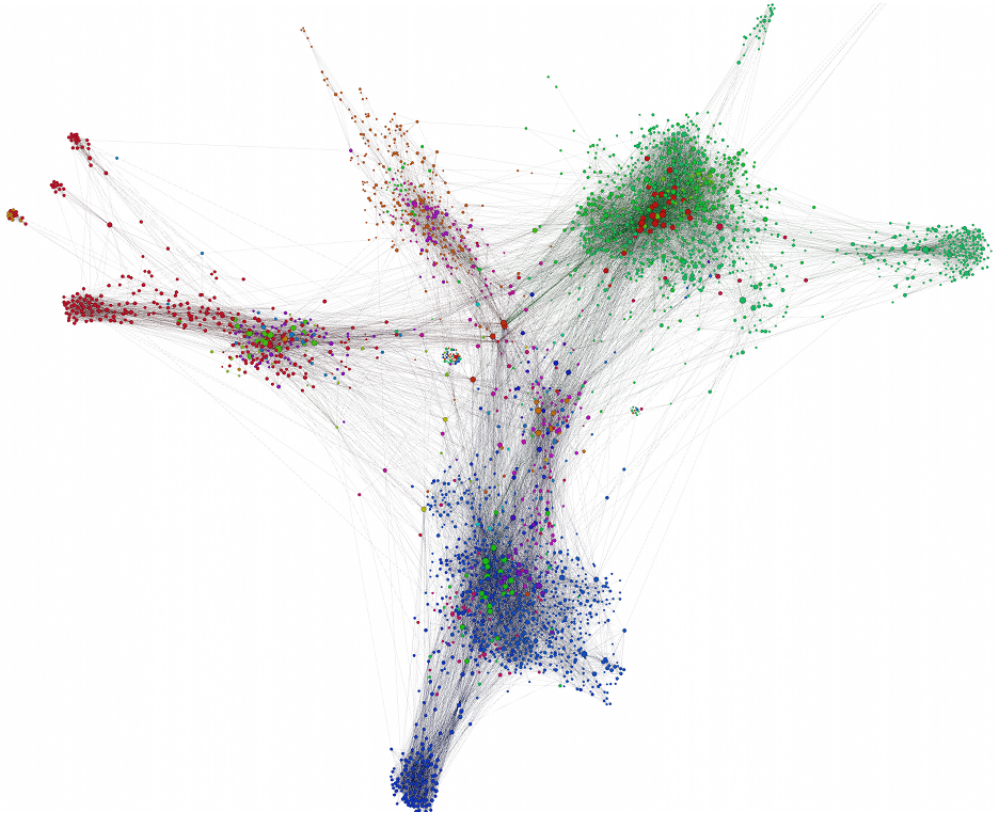
# Visualisation de graphes

## Solution très attractive

- jolis dessins
- vue d'ensemble
- interactivité

## Trompeur

- risque de contresens
- surcharge cognitive
- efficacité contestable



Source : `https://gephi.org/images/screenshots/layout2.png`

# Plan

# Définition

## Graphe

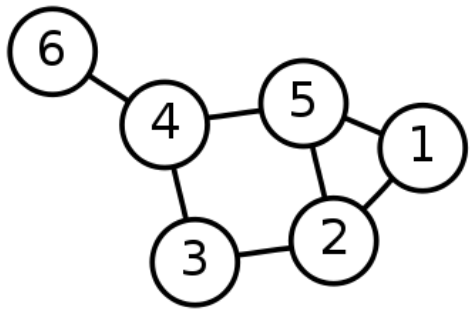
Un **graphe orienté simple** est un couple $(V, E)$, où $V$ est un ensemble de **sommets** ($V$ pour *vertices*) et $E$ une partie de $V \times V$ ($E$ pour *edges*). Un élément $(u, v) \in E$ est un **arc** du graphe.

## Vocabulaire

- si $E$ est un ensemble de couples de $V$, sans distinction d'ordre, le graphe est **non orienté**
- dans ce cas, un élément $\{u, v\} \in E$ est une **arête** du graphe
- un graphe peut être **valué** : $E$ est alors une fonction de $V \times V$ dans $\mathbb{R}^+$ qui à une paire $(u, v)$ associe un poids strictement positif en cas d'arc présent
- un graphe peut être **décoré** (ou étiqueté) : on associe des valeurs aux arêtes et/ou sommets

# Représentation matricielle

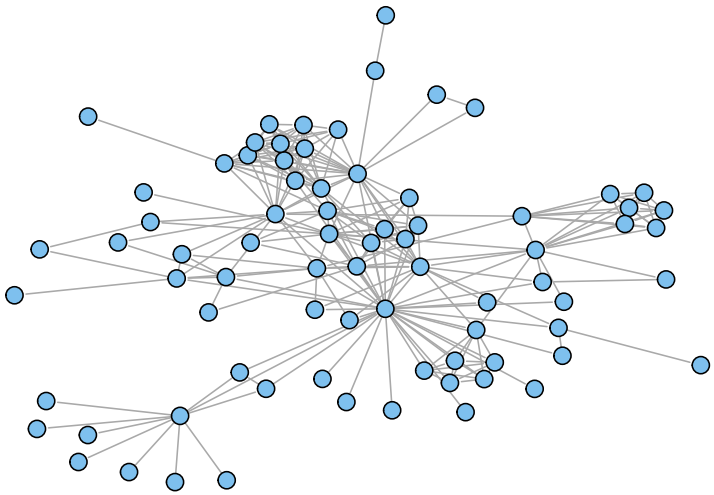
- **matrice d'adjacence** : matrice $A$ de taille $|V| \times |V|$ telle que $A_{ij} = 1$ si et seulement si $(i, j) \in E$.
- exemple




$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

- graphe non orienté $\Leftrightarrow A^T = A$
- graphe valué : $A_{ij}$ poids de l'arc entre $i$ et $j$ (on utilise parfois la matrice de poids $W$ en conservant à $A$ sa sémantique non valuée)

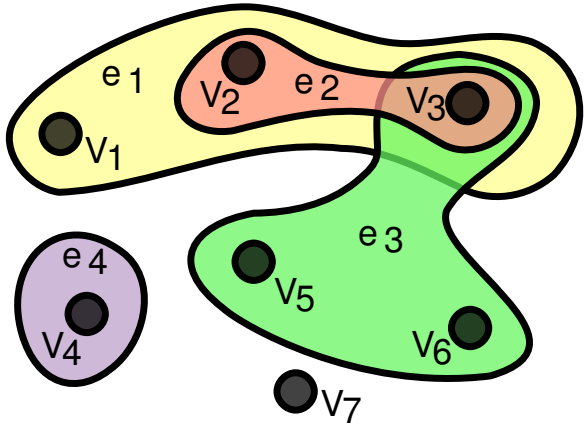
# Visualisation



# Extensions et restrictions

## Extensions

- un **multigraphe** autorise plusieurs arcs/arêtes entre deux sommets
- dans un **hypergraphe**, les sommets sont connectés par des **hyperarêtes** qui peuvent relier plus de 2 sommets




Source : `https://commons.wikimedia.org/wiki/File:Hypergraph.svg`

## Restrictions

- un graphe sans boucle ne contient pas d'arête de la forme $(v, v)$
- un graphe **biparti** est tel que $V = V_1 \cup V_2$ est partitionné en deux classes et aucun arc/arête ne relie des sommets d'une même classe

# Degré

## Définition

- dans un graphe non orienté, le **degré** d'un sommet $v$, $d(v)$, est le nombre d'arêtes qui partent de $v$
- dans un graphe orienté : degré entrant et degré sortant
- dans un graphe valué : somme des poids des arêtes

## Distribution des degrés

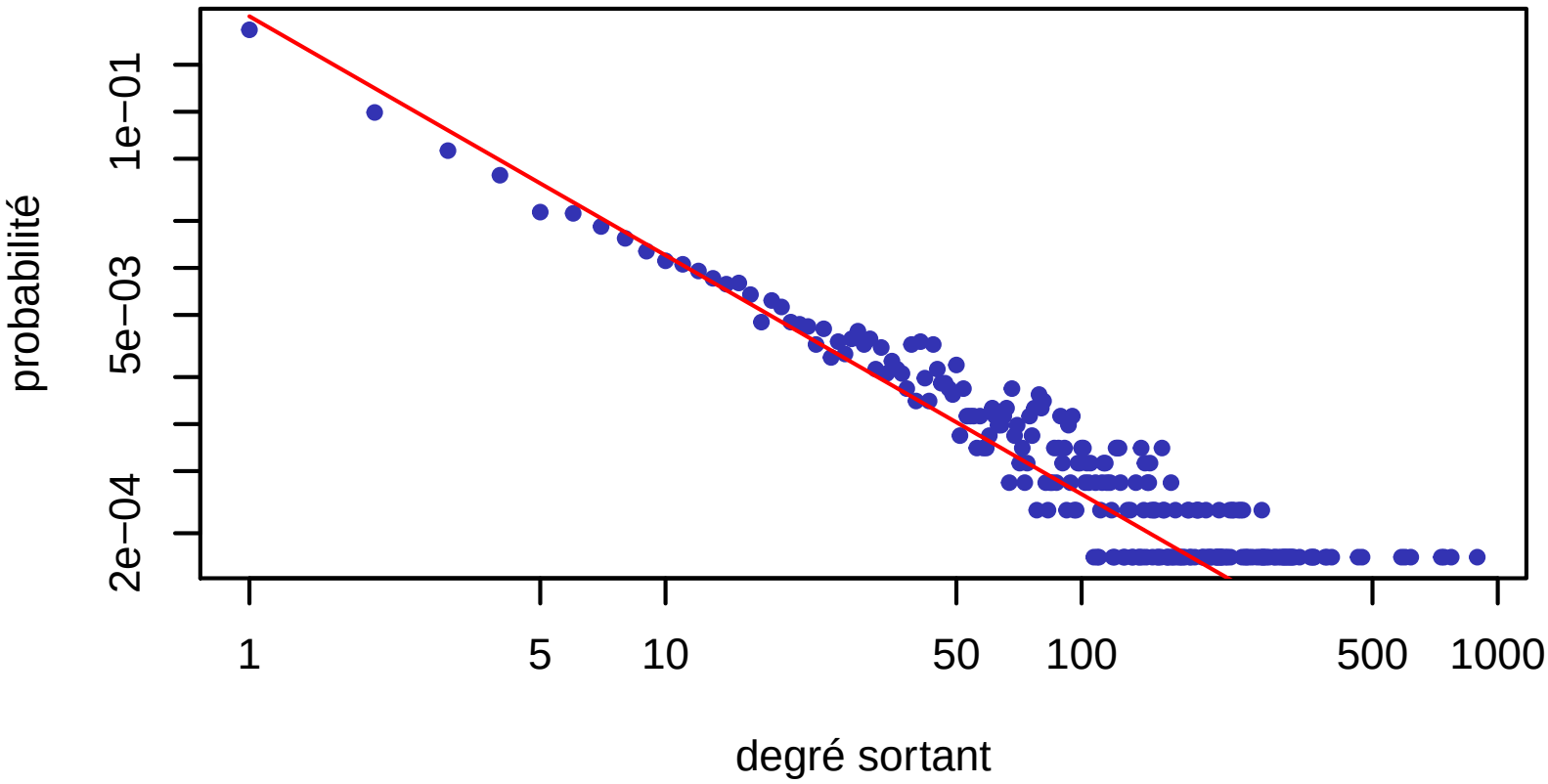
- $P(d = k)$ : « probabilité » d'observer un sommet de degré $k$
- simple comptage en pratique (distribution empirique)
- dans beaucoup de réseaux : loi puissance

$$P(d = k) \sim k^{-\alpha}$$

phénomène de *longue traîne* (distribution *sans échelle*)

# Exemple

- votes sur la Wikipedia avant janvier 2008
- environ 3000 élections
- graphe : a vote pour b
- 7115 sommets et 103689 arcs

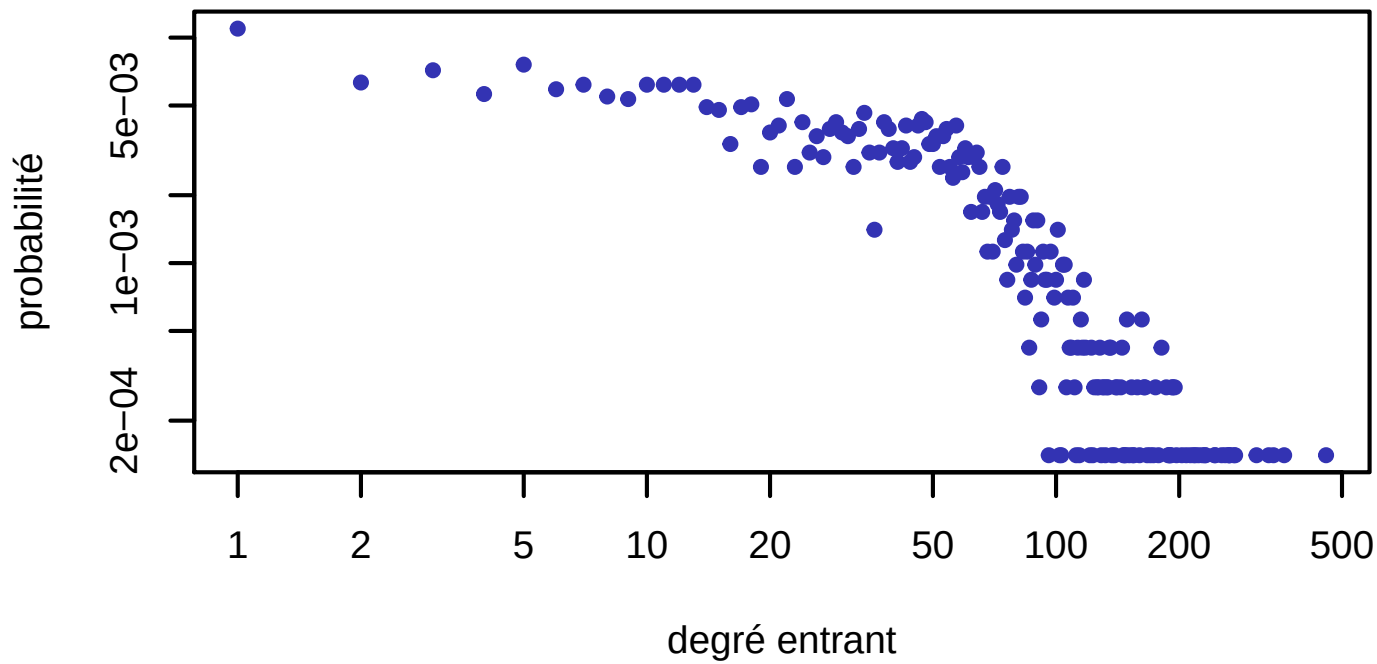



- ici $\alpha \simeq$1,52995

# Remarque

## La loi puissance n'est pas universelle

- phénomène classique dans les réseaux de terrain, mais pas dans les graphes en général
- nombreuses variations comme $P(d = k) \sim k^{-\alpha} e^{-\lambda k}$
- cf *A. Clauset, C.R. Shalizi, and M.E.J. Newman, "Power-law distributions in empirical data" SIAM Review 51(4), 661-703 (2009).*



# Intérêts

## Degré

- une forme de centralité : importance d'un acteur
- applications : détection d'anomalies, recherche d'influenceur, dimensionnement de ressource

## Distribution des degrés

- caractérisation globale : dimensionnement
- caractérise en partie la connectivité globale

## Limitations

- vision très locale
- souvent naïf

# Transitivité

## Principe sociologique

- « mes amis sont amis »
- si $\{u, v\} \in E$ et $\{u, w\} \in E$, on s'attend à avoir $\{v, w\} \in E$
- quantifiable par des mesures locales et globales

## Intérêts

- avant tout descriptif
- caractéristique de beaucoup de réseaux de terrain
- peut représenter un test / une contrainte de réalisme pour certains modèles

# Formalisation

## Dénombrement des triangles

- un triplet : trois sommets formant un sous-graphe connexe
- un triangle : un triplet complet, $\{u, v\}, \{u, w\}, \{v, w\} \in E$
- coefficient de *clustering* :

$$\frac{3 \times \text{nombre de triangles}}{\text{nombre de triplets}}$$

- version locale moyennée :

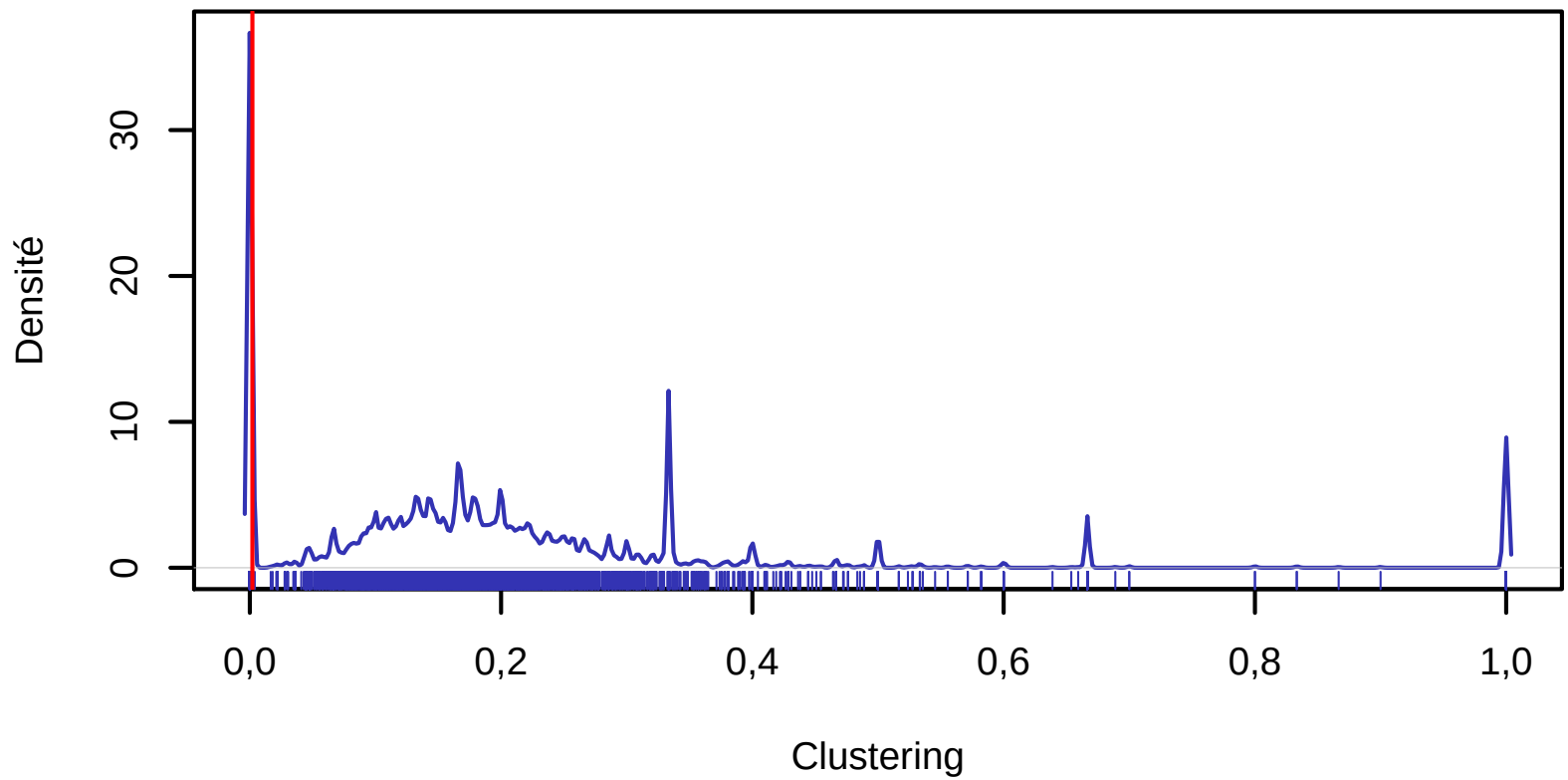
$$\frac{1}{N} \sum_{i=1}^{N} \frac{\text{nombre de triangles contenant i}}{\text{nombre de triplets contenant i}}$$

- diverses extensions (cas pondéré, cas biparti) et généralisation (autres schémas simples comme les carrés)

# Exemple

## Votes Wikipedia

- coefficient global : 0,12548
- référence aléatoire : 0,00205
- moyenne du coefficient local : 0,20552



# Mesures locales

## Bilan

### Mesures

- degré : complètement local
- transitivité : on traverse un lien (test de connexion des voisins)

### Intérêts

- comportement spécifique de certains réseaux (réels)
- peuvent justifier l'étude de certains sommets spécifiques

### Exploitation minimale de la structure graphique

- comptage totalement local ou avec un lien traversé
- agrégation basique (moyenne ou loi)

# Connectivité

- **chemin** :
  - suite d'arc/arêtes entre deux sommets
  - **chemin géodésique** : chemin de longueur minimale entre deux sommets (la longueur tient compte des valuations)
- **composantes connexes** :
  - ensemble de sommets tel qu'il existe au moins un chemin entre chaque paire de sommets
  - connexité **forte** : chemins orientés
- **diamètre** : longueur de la plus longue géodésique

# Petit monde

## Définition (informelle)

Un réseau « petit monde » a un diamètre de l'ordre du logarithme de son nombre de sommets.

## Expérience de Milgram

- routage manuel de lettres
- six étapes : « six degrees of separation » (attention, valeur moyenne pas diamètre)

## Causes et conséquences

- connexions « longue distance » et/ou sommets de grands degrés
- distance moyenne très faible en général
- perte de pertinence de la notion de distance géodésique

# Exemples

## Votes Wikipedia

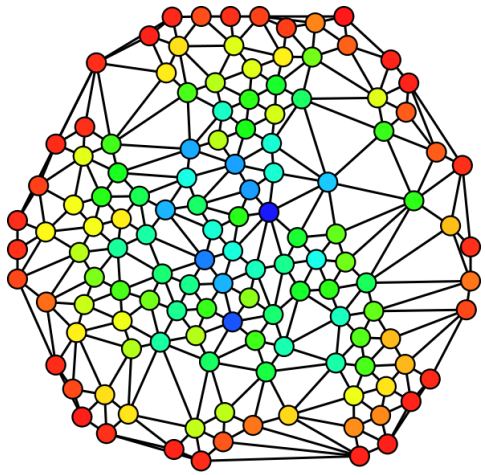
- diamètre dirigé : 10
- diamètre non dirigé : 7
- référence : log(7115) $=$8,86996

## Collaborations en cosmologie

- données :
  - graphe de co-écriture d'articles (source arxiv, de 1993 à 2003)
  - 5242 sommets, 28980 arêtes
  - 355 composantes connexes, la plus grosse avec 4158 sommets
- diamètre (non dirigé) : 17
- référence : log(5242) $=$8,56446

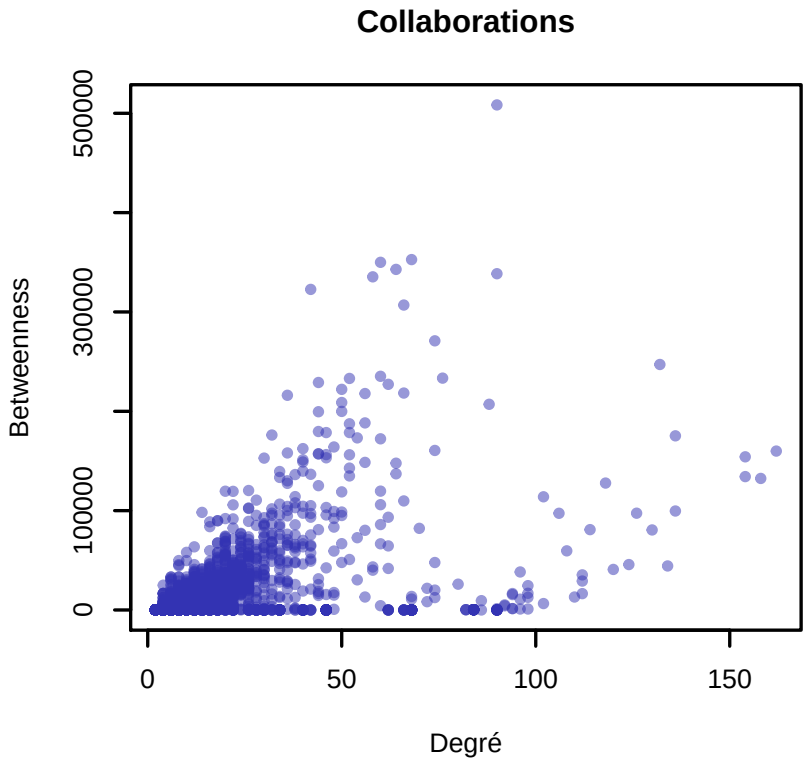
# Centralité

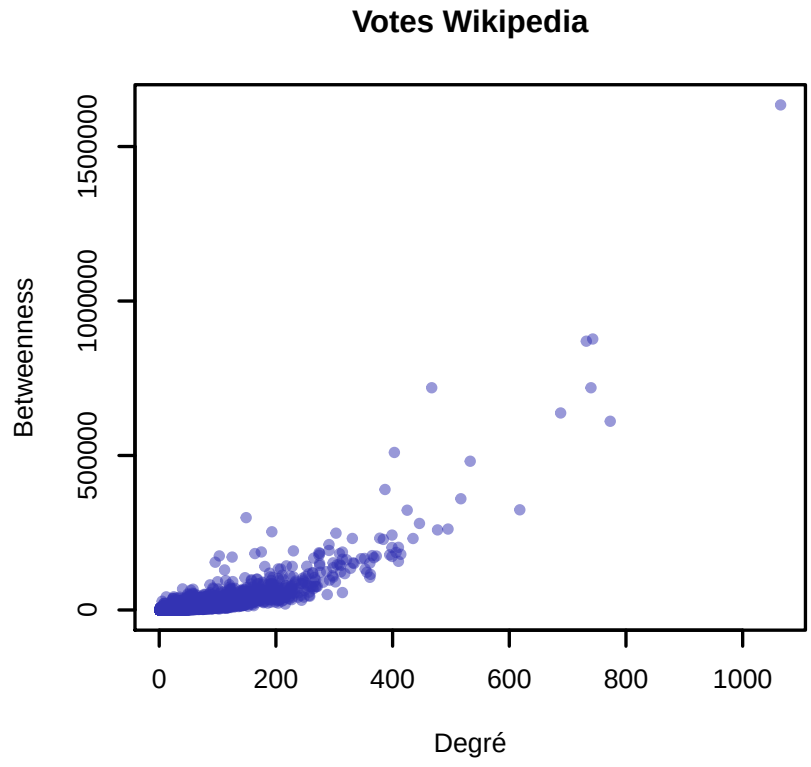
- centralité : importance d'un sommet dans un graphe
- mesure naïve : le degré
- intermédiarité (*betweenness*) :
  - mesure de l'importance d'un sommet en tant qu'agent de lien
  - « nombre » de géodésiques passant par un sommet : somme sur $u$ et $v$ des pourcentages des géodésiques entre $u$ et $v$ passant par le sommet considéré
  - algorithme en $|E||V|$
- d'autres mesures existent...



Source : https://commons.wikimedia.org/wiki/File:Graph_betweenness.svg

# Exemples




## En général

- corrélation importante entre le degré et l'intermédiarité (dépend du réseau)
- **mais** on obtient souvent des sommets extrêmes différents

# Statistiques

## Bilan

### Mesures

- locales : degré, transitivité
- globales : distances géodésiques, diamètre, centralité

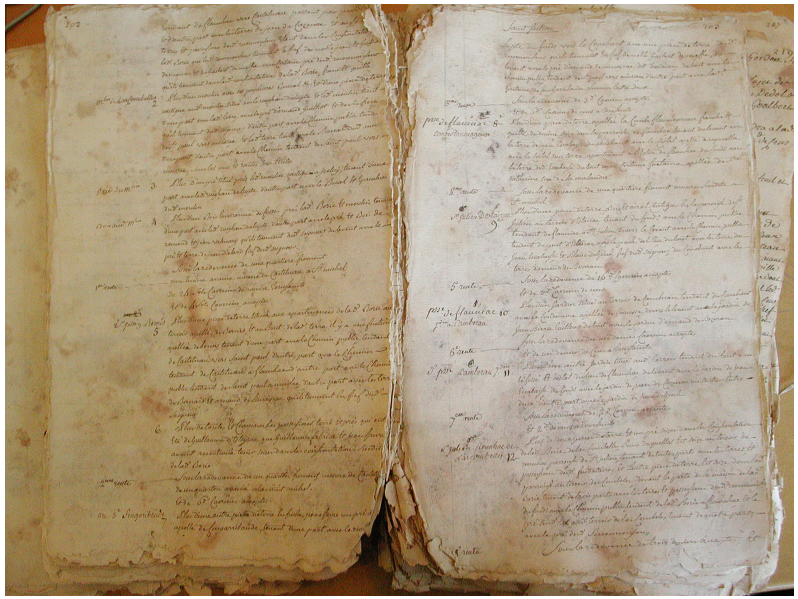
### Questions importantes

- que peut-on espérer trouver avec ces mesures ?
- sont-elles suffisantes pour caractériser certains phénomènes associés au graphe dont elles sont extraites ?

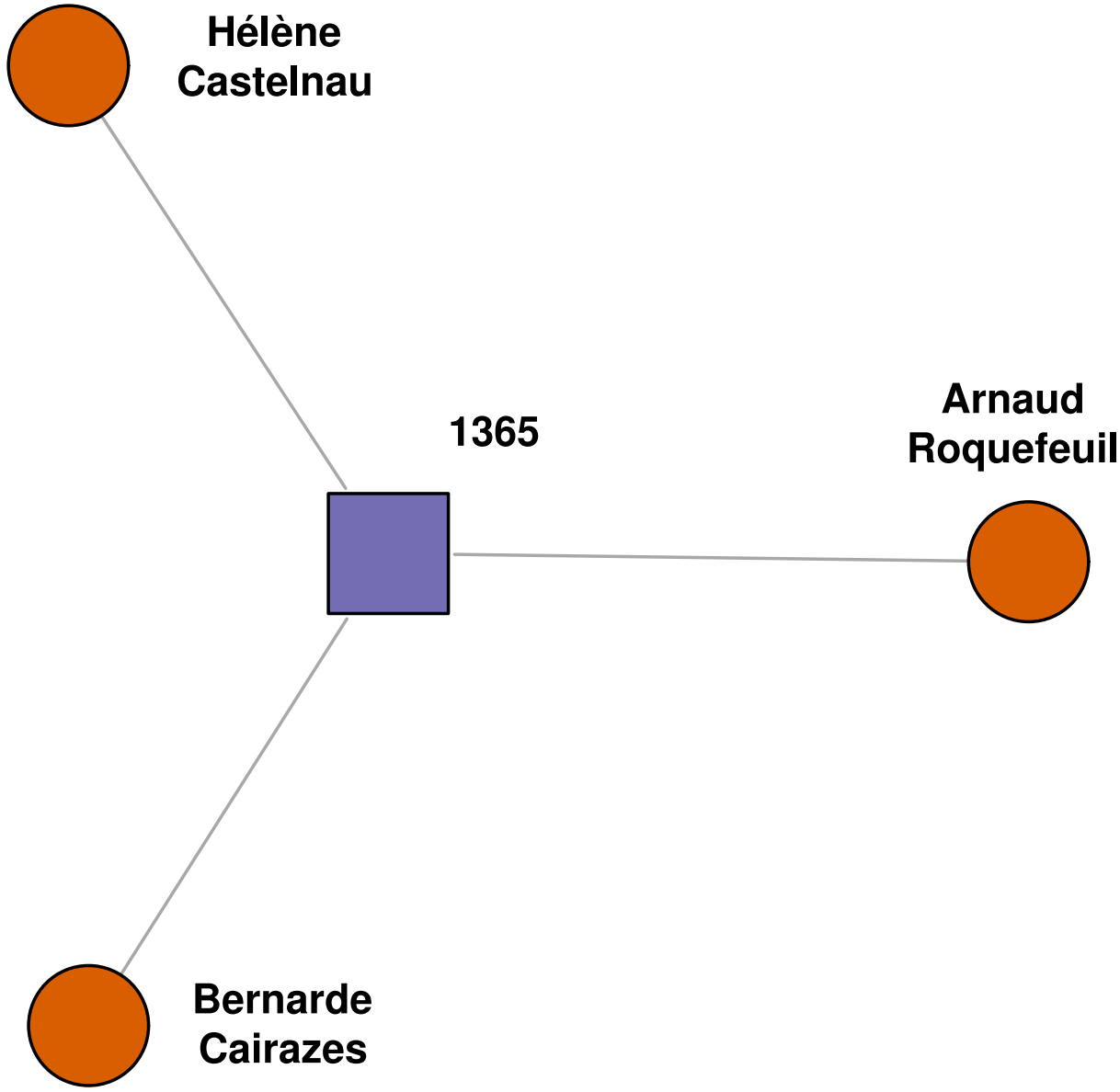
# Une application

## Données

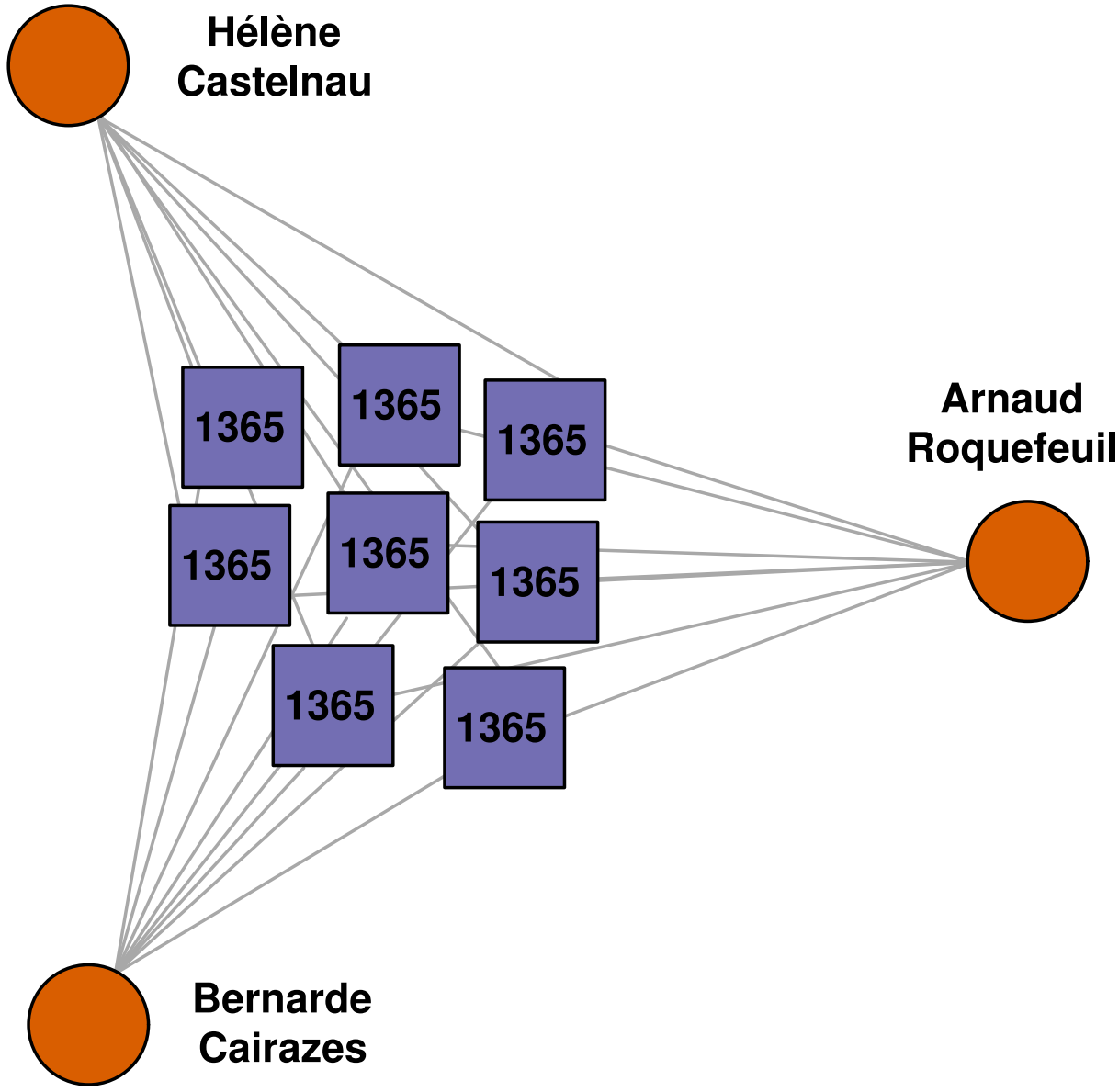
- actes notariés médiévaux (entre 1238 et 1765) :
  - transactions entre seigneurs et paysans
  - droit féodal : fermage, redevance, etc.
- châtellenie de Castelnau Montratier (Bas-Quercy)

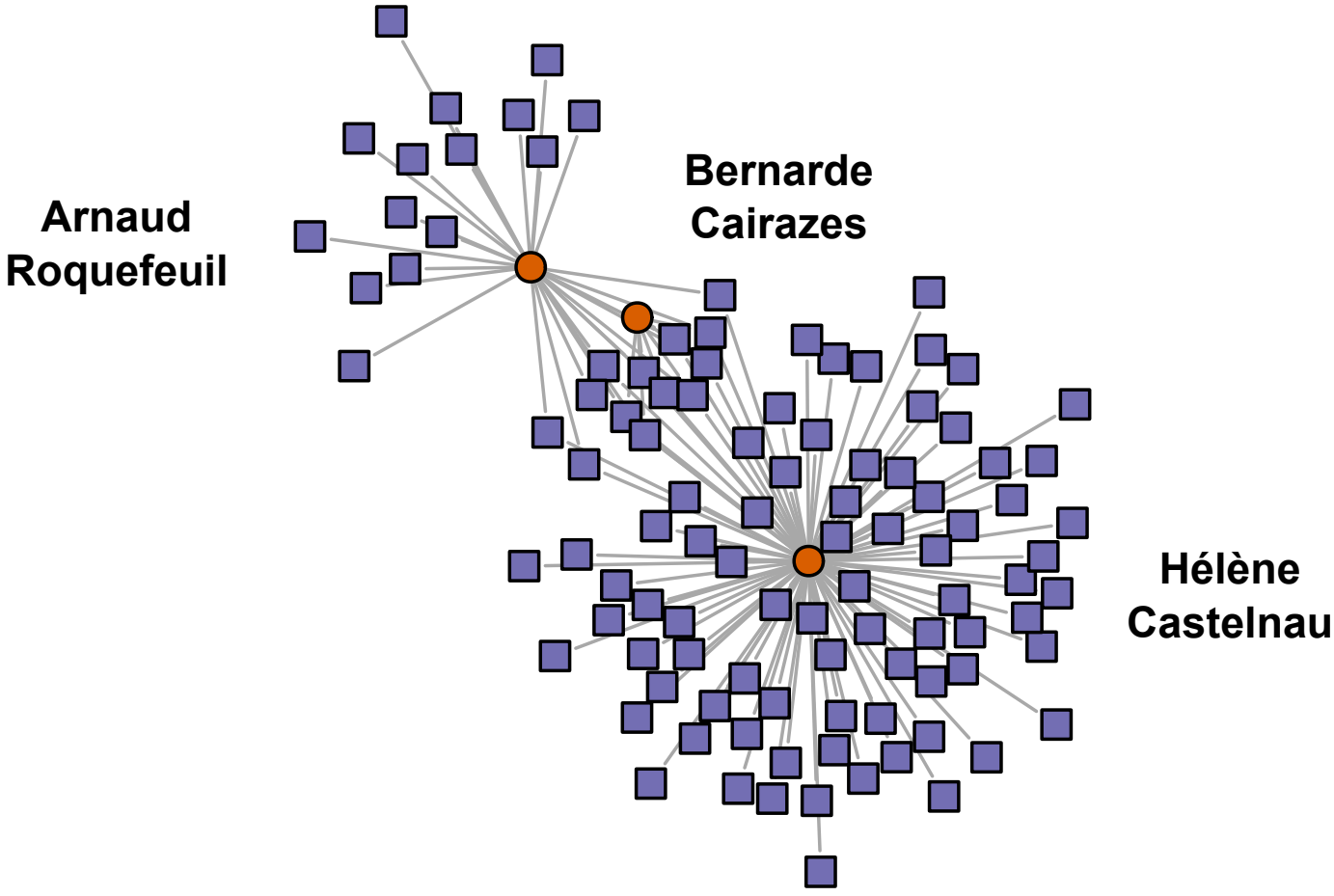
# Réseau biparti complexe



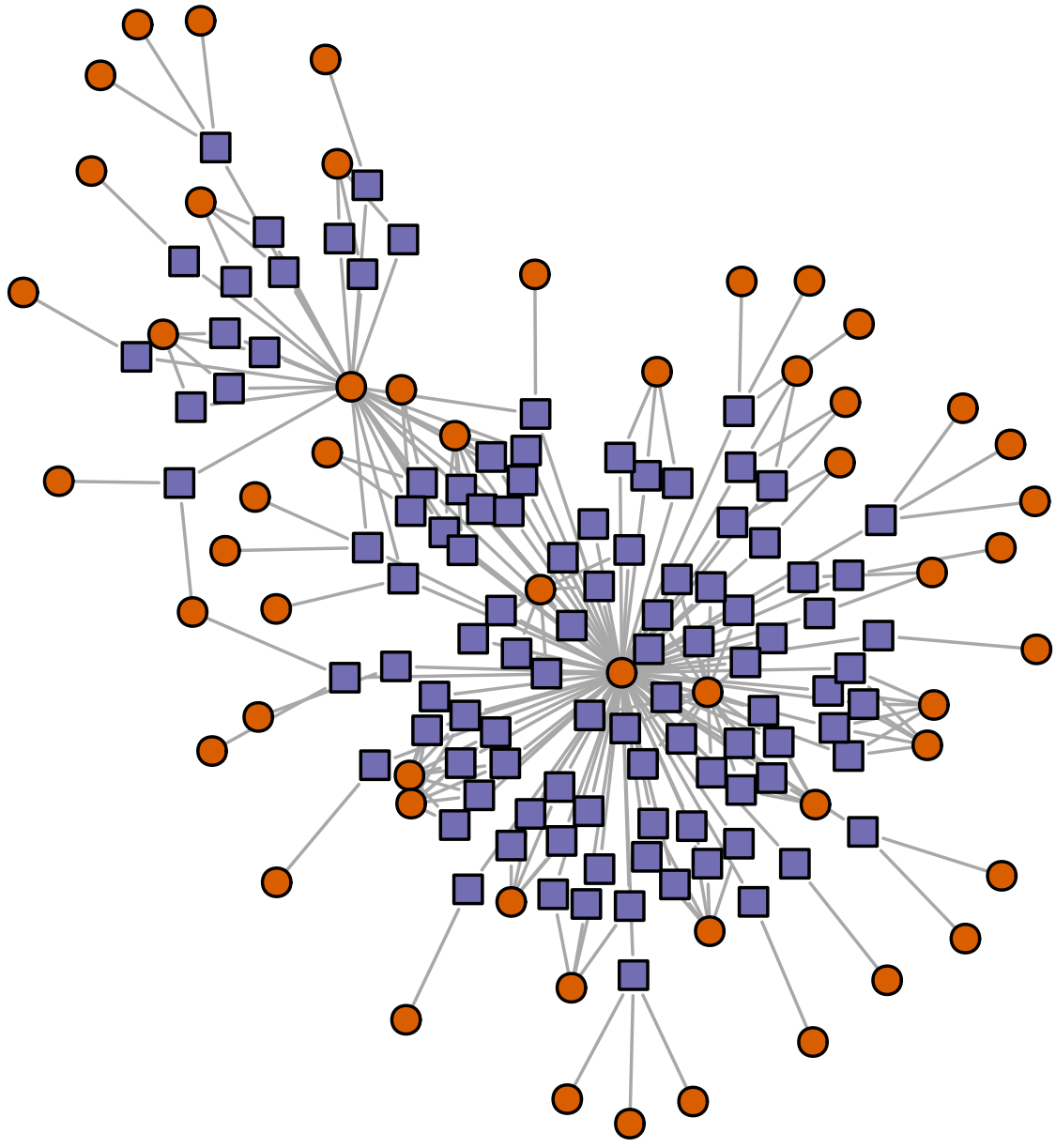
# Réseau biparti complexe



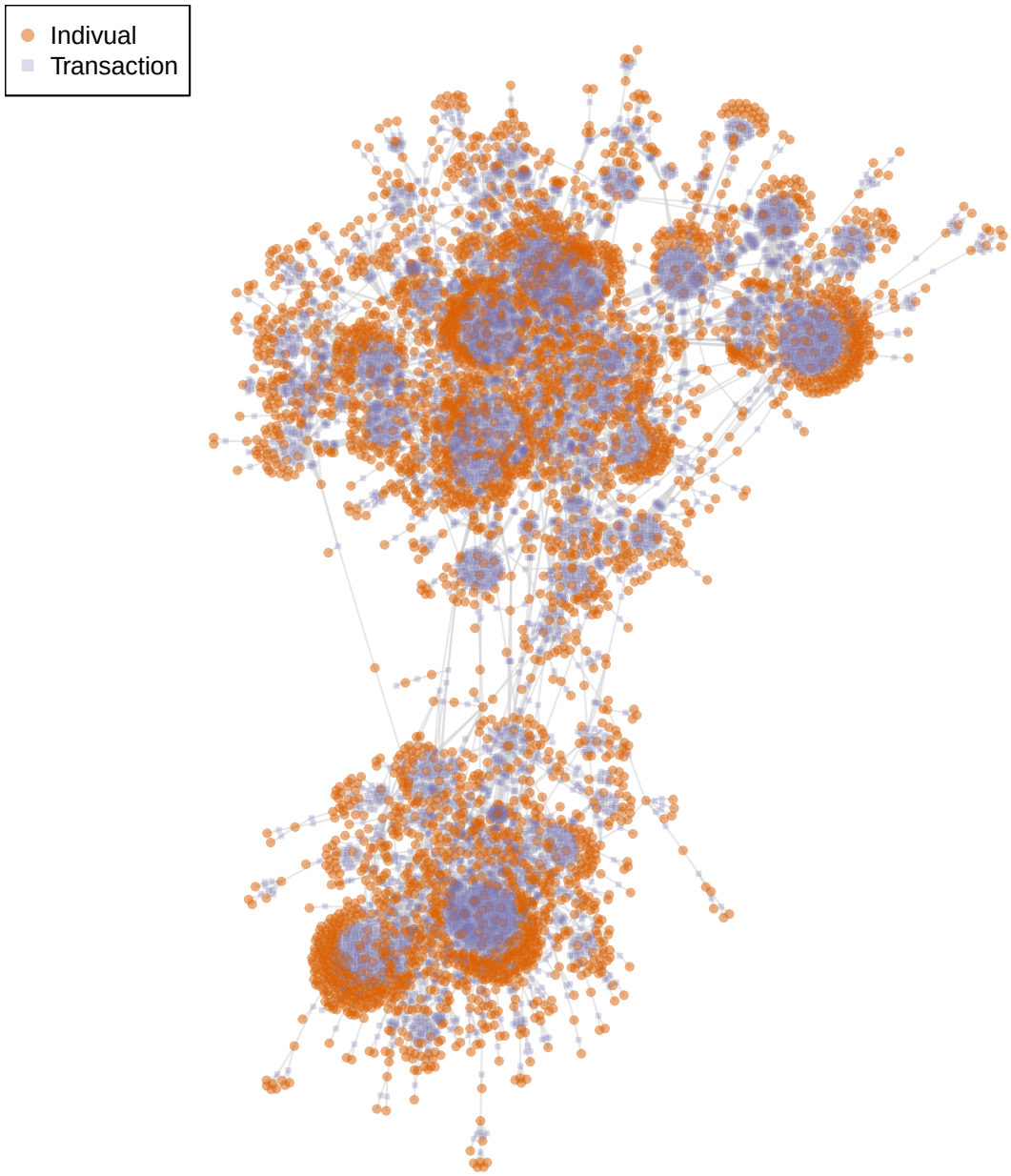
# Réseau biparti complexe



# Réseau biparti complexe



# Réseau biparti complexe



# Intermédiarité et degré

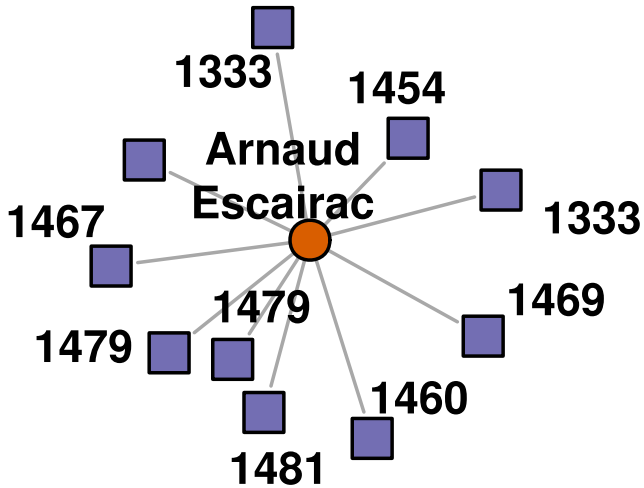
## Stratégie d'analyse

- on cherche les individus centraux au sens de l'intermédiarité mais de degré « faible »
- correspond probablement à des erreurs de transcription des actes (homonymie très forte)

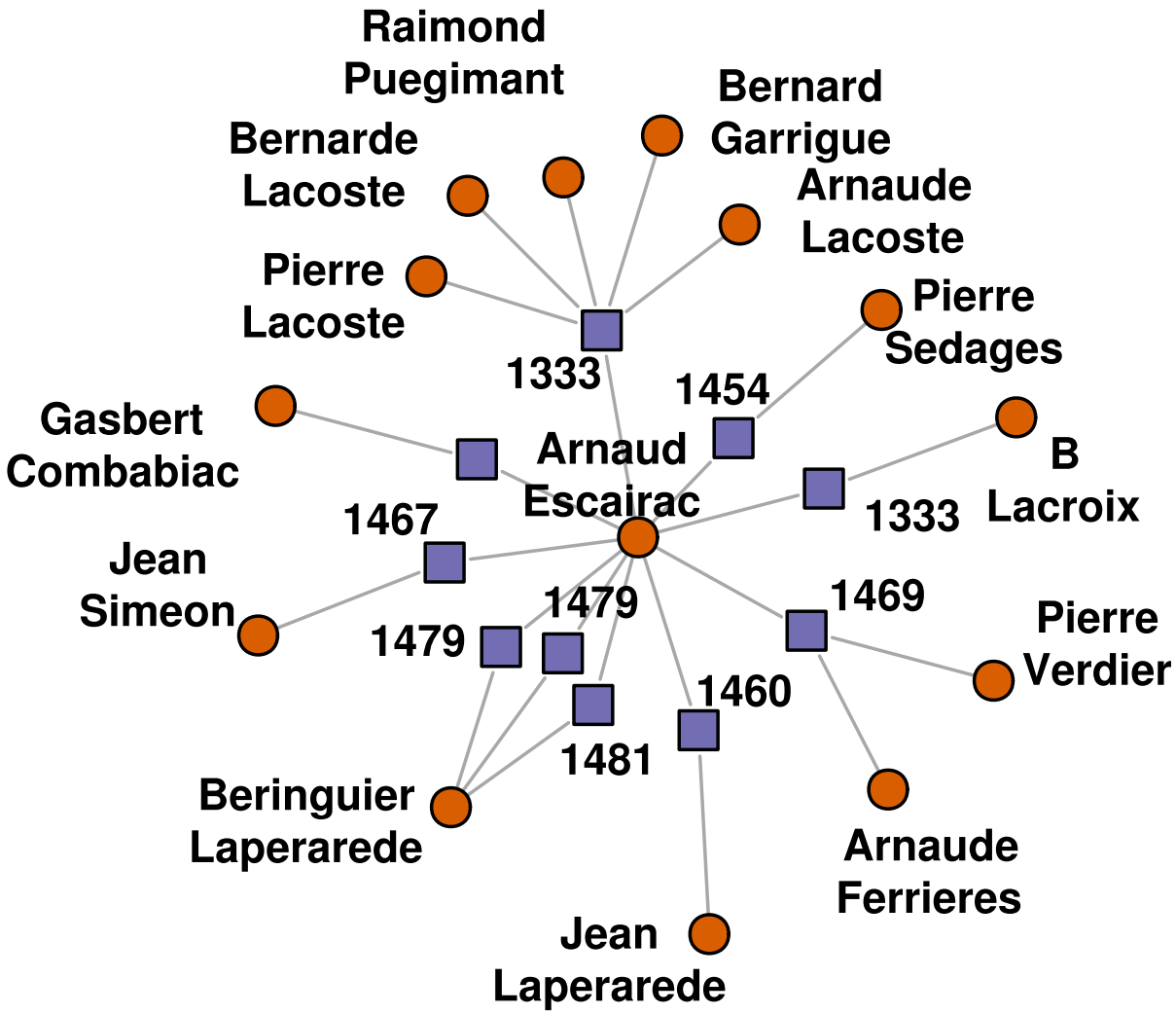
## Résultats

- pour des raisons techniques, on étudie 34 individus « leaders »
- 10 leaders au sens de l'intermédiarité ne sont pas des leaders en degré :
  - quelques « personnes morales » (le Chapitre de Cahors et l'Église de Flaugnac)
  - des erreurs manifestes
  - des erreurs plus subtiles

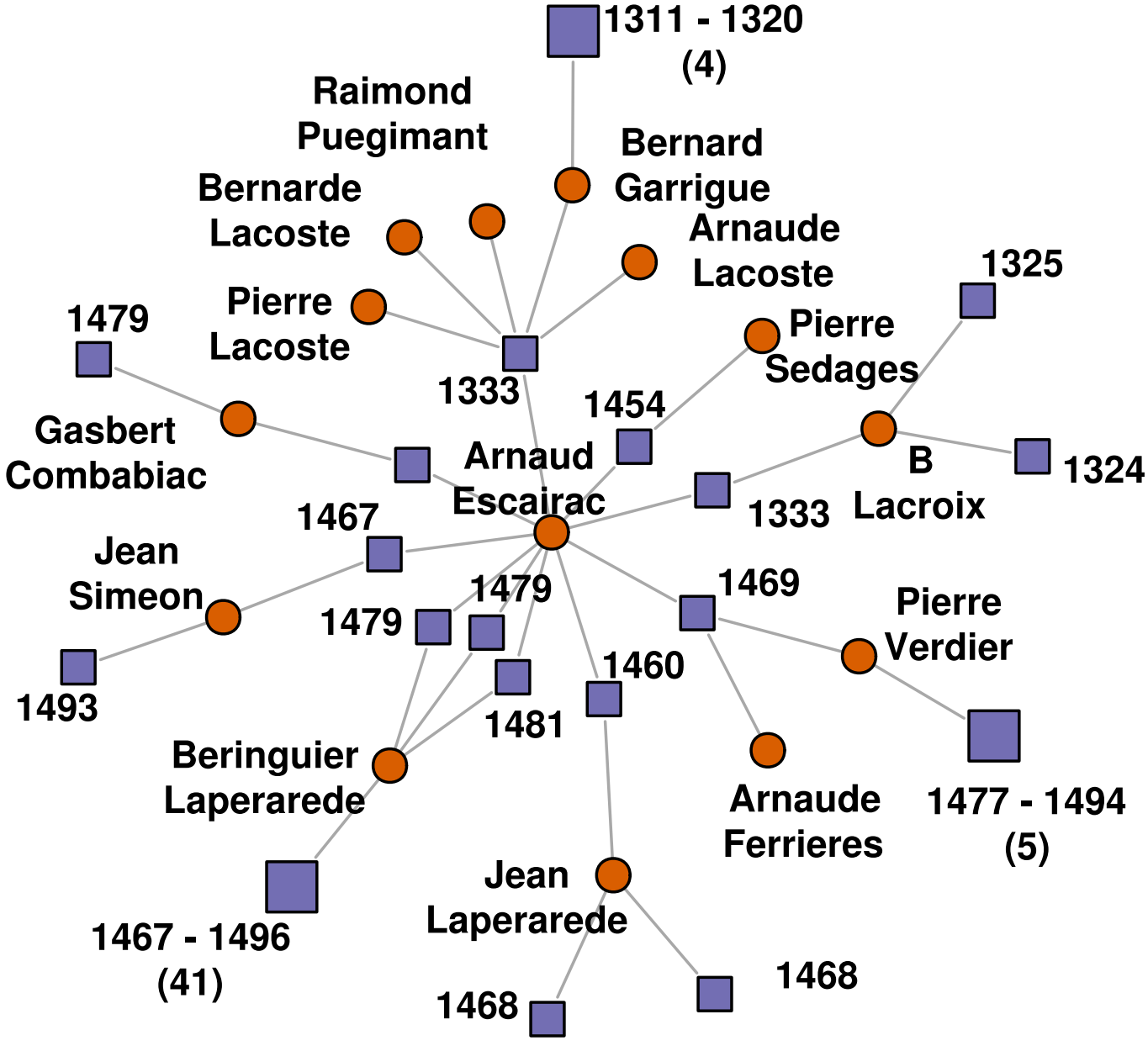
# Erreur manifeste



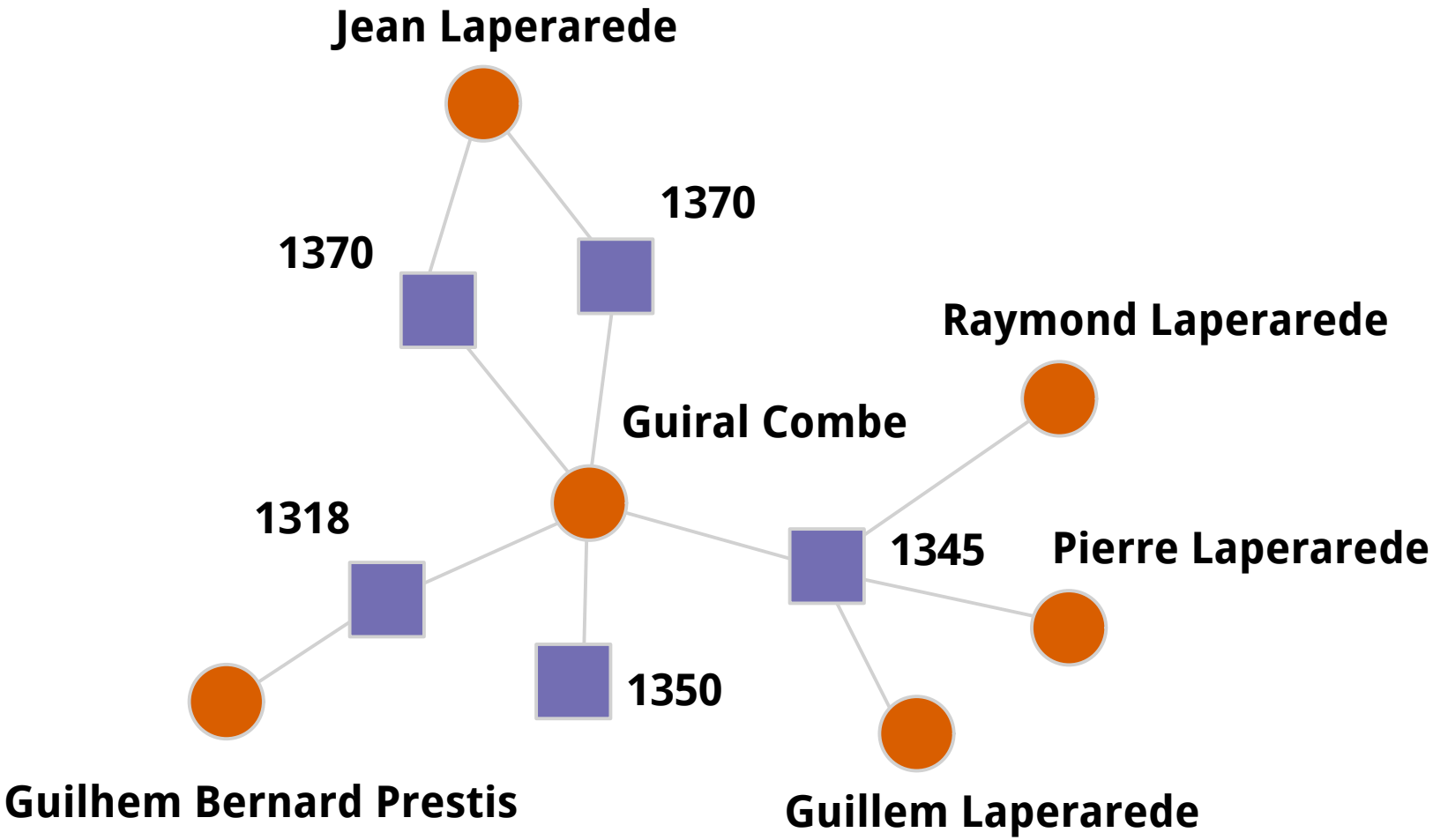
# Erreur manifeste



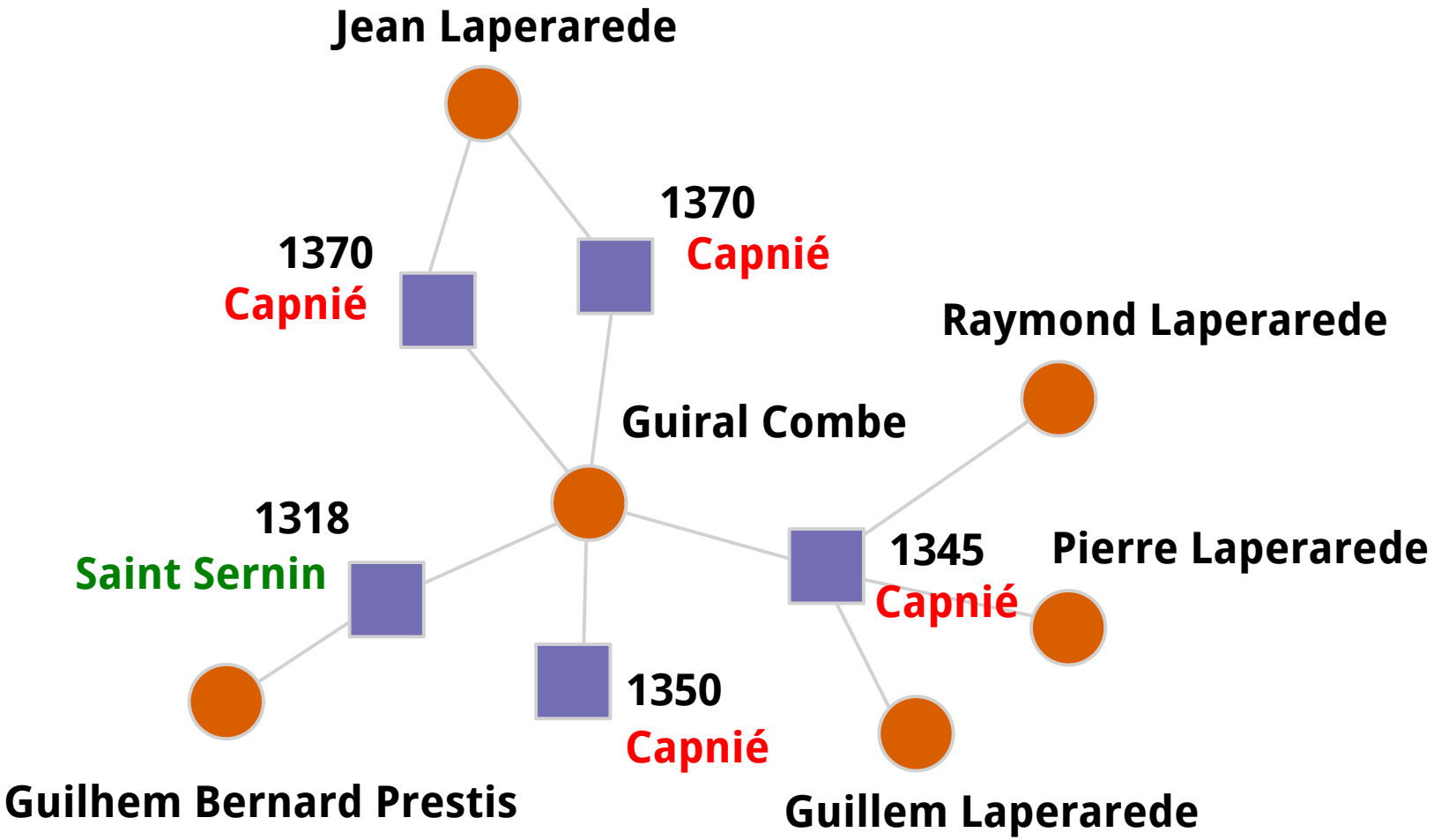
# Erreur manifeste



# Erreur plus subtile



# Erreur plus subtile



# Plan

# Caractérisations macroscopiques des graphes

## Question en suspens

Peut-on analyser certains phénomènes associés à un graphe à partir de caractéristiques macroscopiques ?

## Angle d'attaque

- outil : modèles probabilistes alias « graphes aléatoires »
- démarche :
  1. construction d'un modèle génératif avec quelques paramètres
  2. caractérisation théorique de phénomènes intéressants pour les graphes engendrés par le modèle
  3. évaluation des paramètres adaptés sur un graphe donné
  4. comparaison aux valeurs critiques des paramètres

# Graphes aléatoires

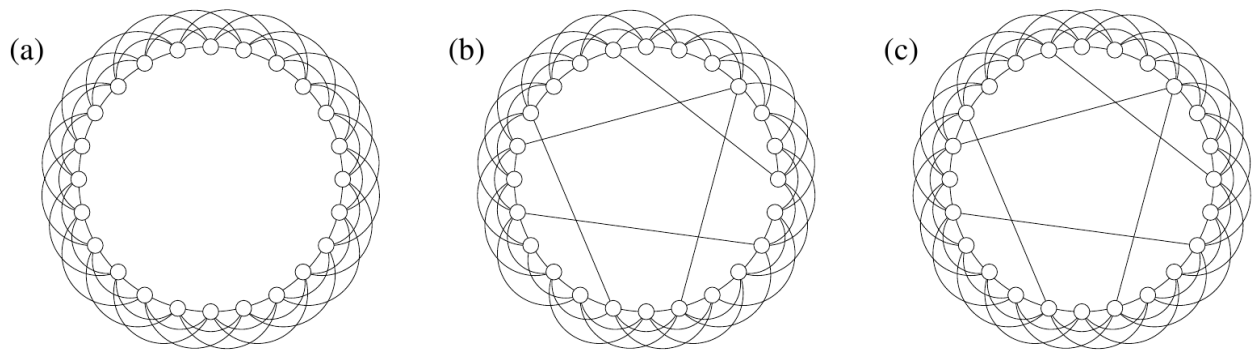
- comment engendrer aléatoirement un graphe intéressant ?
- premier modèle :
  - années 1950
  - Erdös et Rényi, ainsi que Solomonoff and Rapoport
  - graphe aléatoire homogène : probabilité de connexion $p$
  - construction :
    - choix de $n$ sommets
    - pour chaque couple, connexion avec une probabilité $p$ (pas de boucle)
  - petit monde : distance moyenne en $\log n$ (sous degré moyen $z = p(n-1)$ constant)
  - pas de structure : pas de triangle, par exemple
  - distribution des degrés binomiale, asymptotiquement poisson $\Rightarrow$ adéquation limitée avec la réalité

# Modèle de configuration

- modèle plus réaliste : distribution quelconque des degrés
- construction :
  - tirage des degrés des $n$ sommets selon une distribution choisie : chaque sommet $i$ possède $p_i$ pattes
  - choix aléatoire uniforme d'une paire de pattes libres et connexion de celles ci
  - quelques subtilités en pratique (solution de type MCMC)
- pas de structure
  - sommets indépendants
  - pas de triangle, par exemple

# Modèle petit monde

- travaux de Watts et Strogatz (1998)
- idée simple :
  - on part d'un graphe régulier construit sur une structure géométrique simple, par exemple un anneau
  - on change aléatoirement l'une des extrémités d'une connexion (avec une probabilité *p*)
  - variante : on ajoute des arêtes aléatoires




- bonnes propriétés de *clustering* et de diamètre, mais distribution des degrés peu réaliste

Source : The Structure and Function of Complex Networks, M. E. J. Newman, SIAM Review 2003 45 :2, 167-256

# Composante géante

## Connexité approximative

- approximative connexe : une composante géante recouvre (presque) tout le graphe
- taille moyenne de la plus grosse composante connexe pour les graphes aléatoires ?

## Modèle Erdös-Rényi

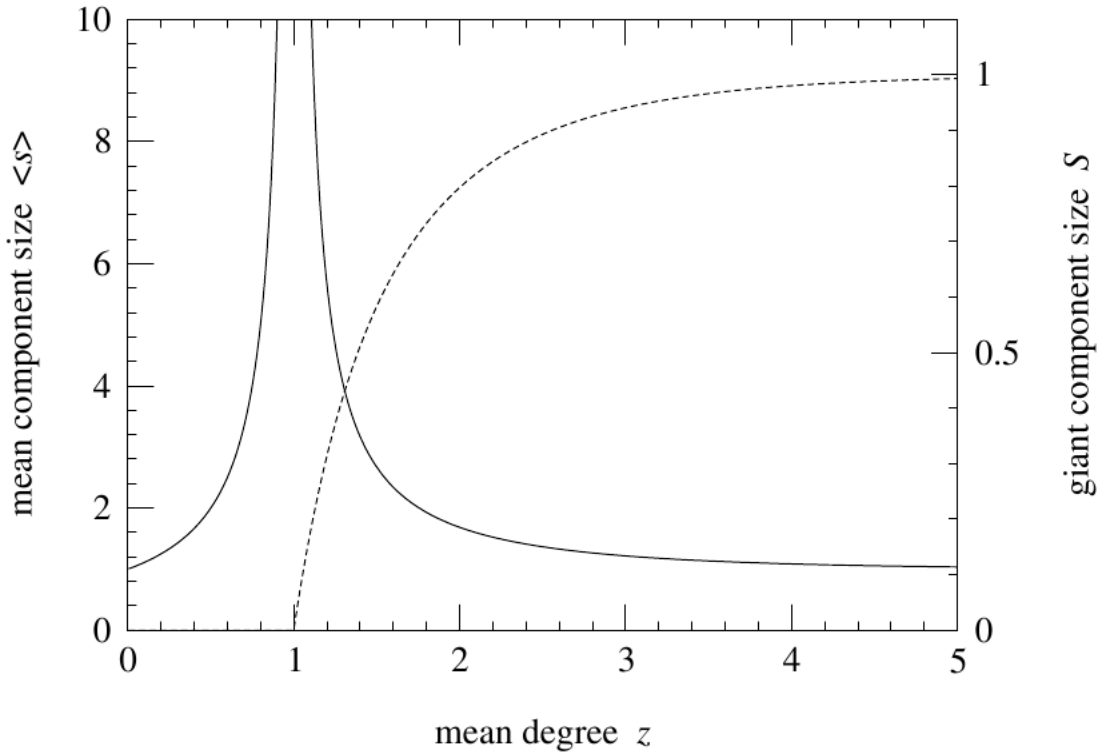
- degré moyen $z = p(n-1)$, $S$ fraction du graphe dans la plus grande composante (en moyenne)
- on a (pour un grand graphe) $S = 1 - e^{-zS}$
- on a aussi comme taille moyenne des composantes en dehors de la composante géante

$$\langle s \rangle = \frac{1}{1 - z + zS}$$

# Transition de phase

## Modèle Erdös-Rényi

- pour $z < 1$, $S = 0$
- pour $z > 1$ $S > 0$
- pour $z = 1$, $\langle s \rangle$ explose :




Source : The Structure and Function of Complex Networks, M. E. J. Newman, SIAM Review 2003 45 :2, 167-256

# Exemples

## Cosmologie

- $z = 11{,}05685$
- en théorie $S = 1$
- en pratique $S = 0{,}79321$

## Votes Wikipedia

- vu comme un graphe non dirigé
- $z = 29{,}14659$
- en théorie $S = 1$
- en pratique $S = 0{,}99311$

# Modèle de configuration

- analyse similaire mais plus complexe ; 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

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

- fraction du graphe dans la plus grande composante, $S$, solution de

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

# Exemples

## Cosmologie

- $\sum_k k(k-2)p_k = 350{,}91339$
- en théorie $S = 1$
- en pratique $S = 0{,}79321$

## Votes Wikipedia

- vu comme un graphe non dirigé
- $\sum_k k(k-2)p_k = 4437{,}66999$
- en théorie $S = 1$
- en pratique $S = 0{,}99311$

# Loi puissance

## Modèle

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

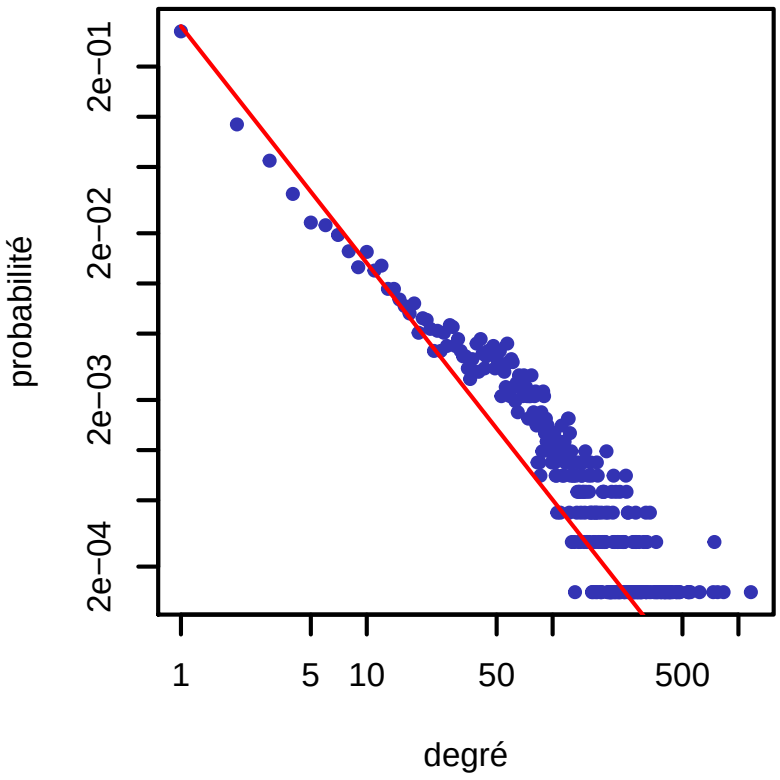
$$\zeta(\alpha - 2) = \alpha\zeta(\alpha - 1),$$

    avec $\zeta(\alpha) = \sum_k k^{-\alpha}$.
  - $\alpha <= 2$, $S = 1$ : une seule composante dans le graphe
  - transition continue entre les deux

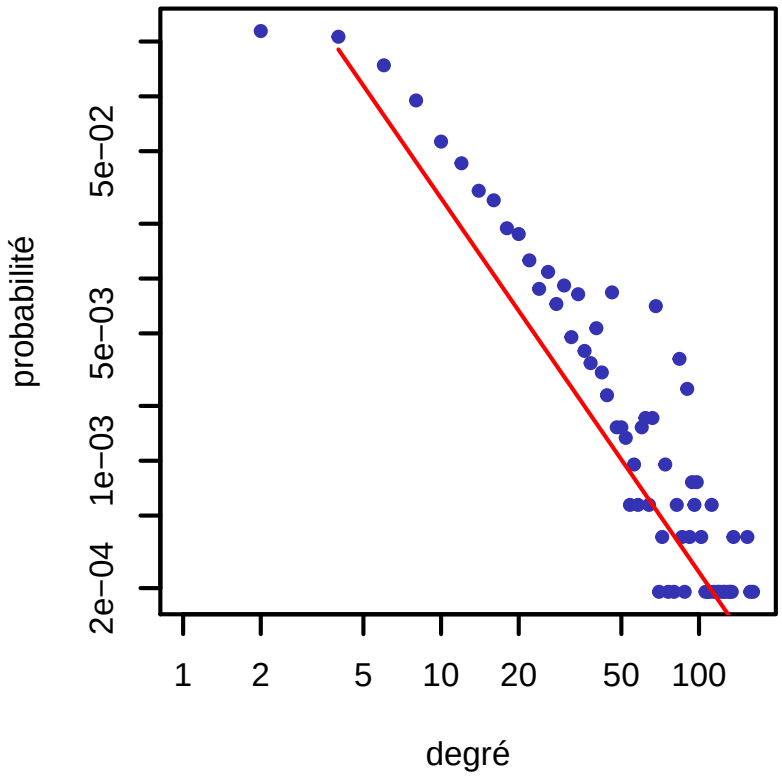
## Examples

- Cosmologie : $\alpha = 2{,}0522$
- Votes Wikipedia : $\alpha = 1{,}42066$

# Examples




Votes Wikipedia




Cosmologie

# Bilan

## Composante géante

- phénomène global non trivial
- caractérisé assez simplement pour certains graphes aléatoires
- dans la limite des grands graphes

## Quel intérêt pratique ?

- si on a le graphe, on connaît la taille de la composante géante...

# Bilan

## Composante géante

- phénomène global non trivial
- caractérisé assez simplement pour certains graphes aléatoires
- dans la limite des grands graphes

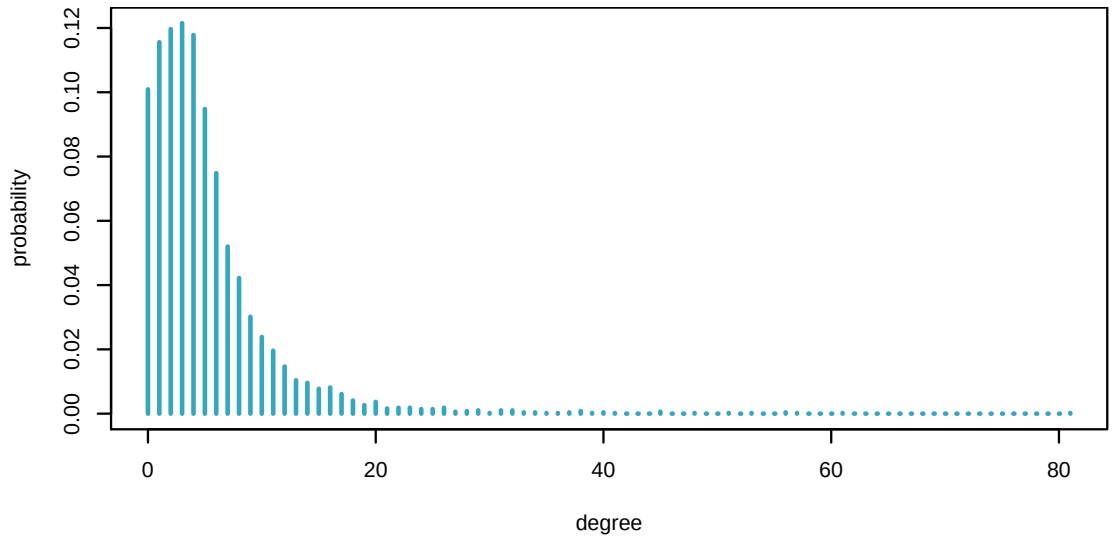
## Quel intérêt pratique ?

- si on a le graphe, on connaît la taille de la composante géante...
- on peut avoir seulement la distribution des degrés (déclaration)
- ou un échantillon du graphe (sondage)

# Application

## épidémie du VIH à Cuba

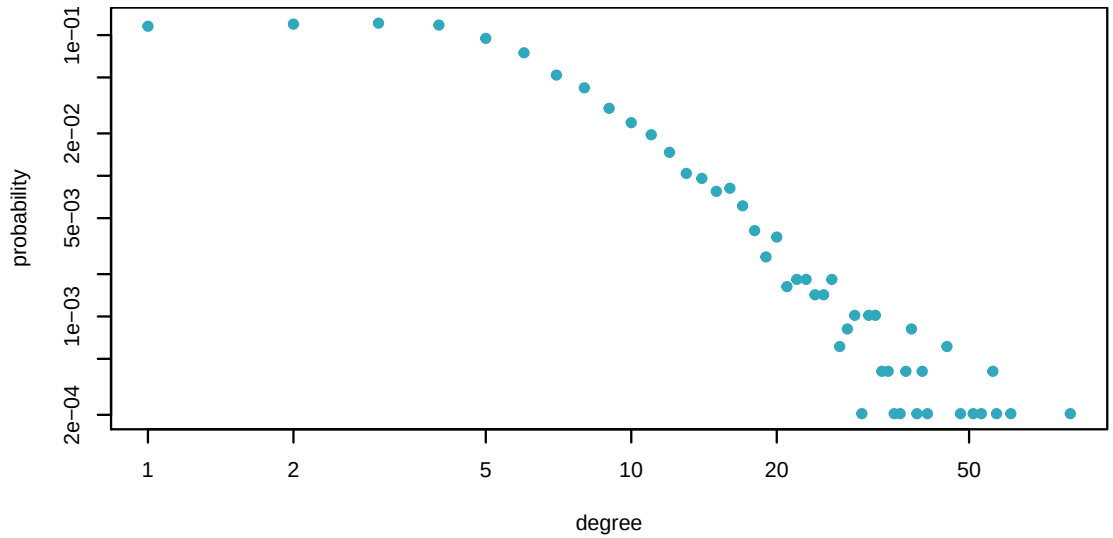
- 5389 individus séropositifs
- déclaration sur le nombre de partenaires sexuels sur les deux années précédant la détection



# Application

## épidémie du VIH à Cuba

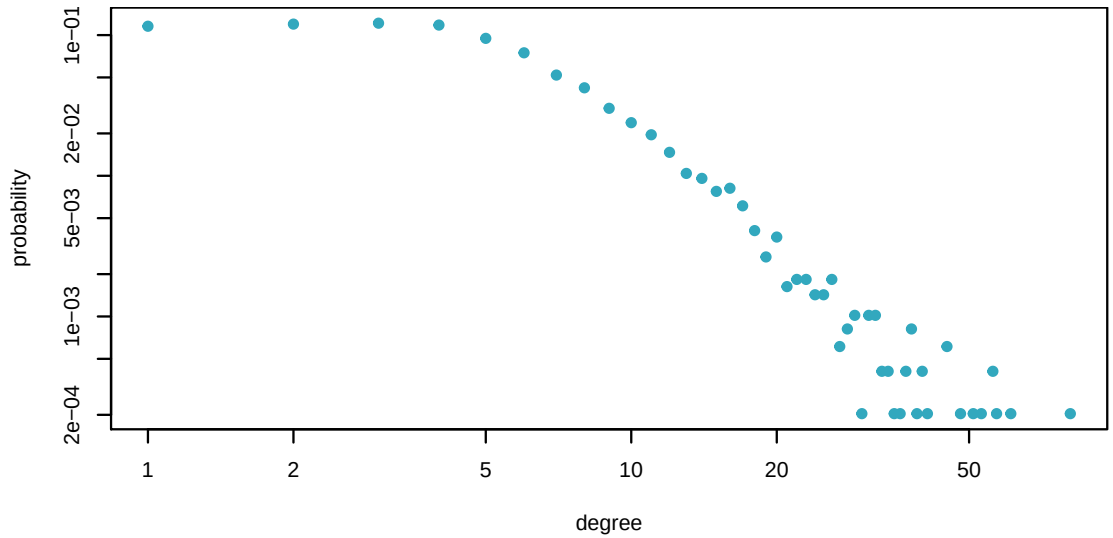
- 5389 individus séropositifs
- déclaration sur le nombre de partenaires sexuels sur les deux années précédant la détection



# Application

## épidémie du VIH à Cuba

- 5389 individus séropositifs
- déclaration sur le nombre de partenaires sexuels sur les deux années précédant la détection

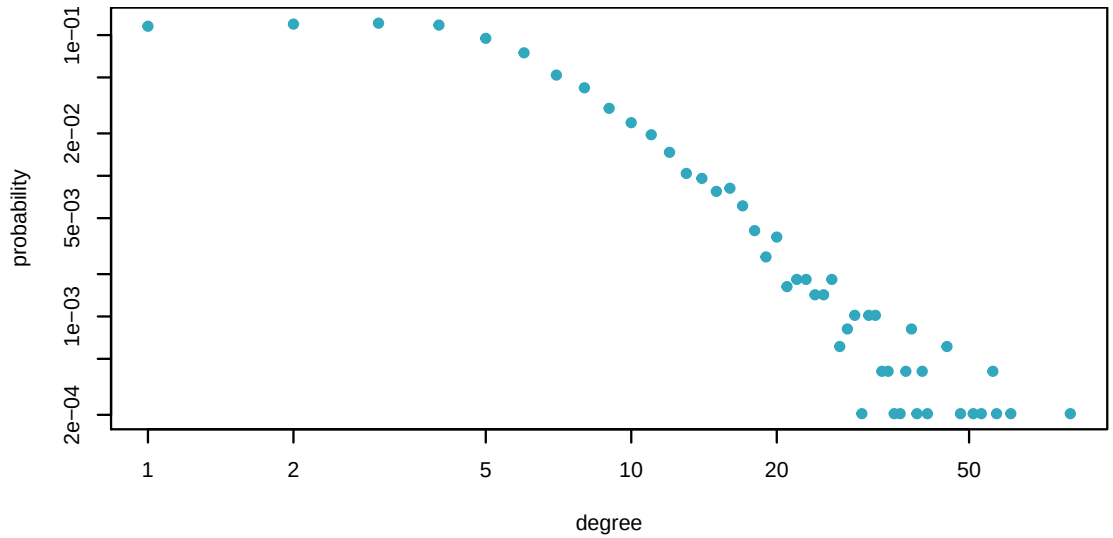



- composante géante : 90.8% !

# Application

## épidémie du VIH à Cuba

- 5389 individus séropositifs
- déclaration sur le nombre de partenaires sexuels sur les deux années précédant la détection




- composante géante : 90.8% !
- réseaux de contacts sexuels :
  - Suède : $\alpha \simeq 2.4 \Rightarrow$ composante géante
  - mais Cuba : $\alpha \simeq 3.5 \Rightarrow$ pas de composante géante

# Phénomènes se déroulant sur un graphe

## Propagation sur un graphe

- d'une épidémie
+ d'une information

## Résilience

- le graphe résiste-t-il à la destruction des sommets/arêtes ?
- influence des pannes/vaccinations/quarantaine sur la composante géante

# Phénomènes se déroulant sur un graphe

## Propagation sur un graphe

- d'une épidémie
+ d'une information

## Résilience

- le graphe résiste-t-il à la destruction des sommets/arêtes ?
- influence des pannes/vaccinations/quarantaine sur la composante géante

## Percolation

- de site : suppression/occupation de sommets
- de lien : suppression/occupation d'arêtes

# Percolation de site

## Modèle simple

- graphe aléatoire $\xrightarrow{\text{choix de sommets}}$ graphe aléatoire
- caractérisation classique par le degré : $q_k$ est la probabilité d'occupation d'un sommet de degré $k$

## Modèle de configuration

- deux sources d'aléa
- nouveau modèle de configuration, par exemple si $q_k = q$

$$p'_{k'} = \sum_{k \geq k'} p_k \binom{k}{k'} q^{k'} (1-q)^{k-k'}.$$

## Question principale

Condition sur $q$ ou $q_k$ pour conserver une composante géante ?

# Analyse

- mêmes outils que pour la composante géante :
  - $F_0(x) = \sum_k p_k q_k x^k$
  - $F_1(x) = \frac{\sum_k k p_k q_k x^{k-1}}{z}$
- condition de transition de phase $F_1'(1) = 1$
- composante géante

$$S = F_0(1) - F_0(u), \quad u = 1 - F_1(1) + F_1(u)$$

- loi puissance et $q_k = q$
  - $q$ critique

$$q_c = \frac{\zeta(\alpha - 1)}{\zeta(\alpha - 2) - \zeta(\alpha - 1)}$$

  - $\alpha < 3 \Rightarrow$ percolation pour tout $q > 0$

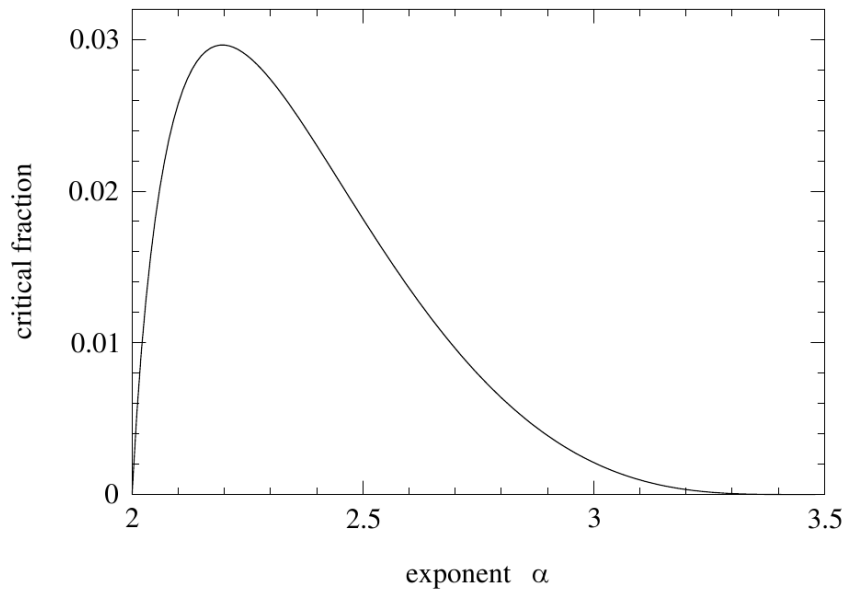
# Conséquences

## Régime épidémique permanent

- loi puissance $\alpha < 3$ : préservation de la composante géante
- \+ forte résistance aux pannes
- \- propagation des épidémies (mais modèle moyennement adapté)

## Attaque ciblée

- attaque coordonnée : tuer les routeurs, c.-à-d. $q_k$ croit avec $k$
- la percolation est assurée majoritairement par les sommets de degrés élevés




Source : The Structure and Function of Complex Networks, M. E. J. Newman, SIAM Review 2003 45 :2, 167-256

# Percolation de lien

## Principe

- $T$ : probabilité d'occupation d'un lien
- même situation : on construit un nouveau graphe aléatoire

## Modèle de configuration

- fonctions génératrices :

$$G_0(x,T) = \sum_k p_k (1 - T + xT)^k, \; G_1(x,T) = \sum_k q_k (1 - T + xT)^k$$

- transition de phase :
  - probabilité critique $T_c = \dfrac{z}{\sum_k k(k-1)p_k}$
  - couverture de la composante géante

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

# Bilan

## Percolation

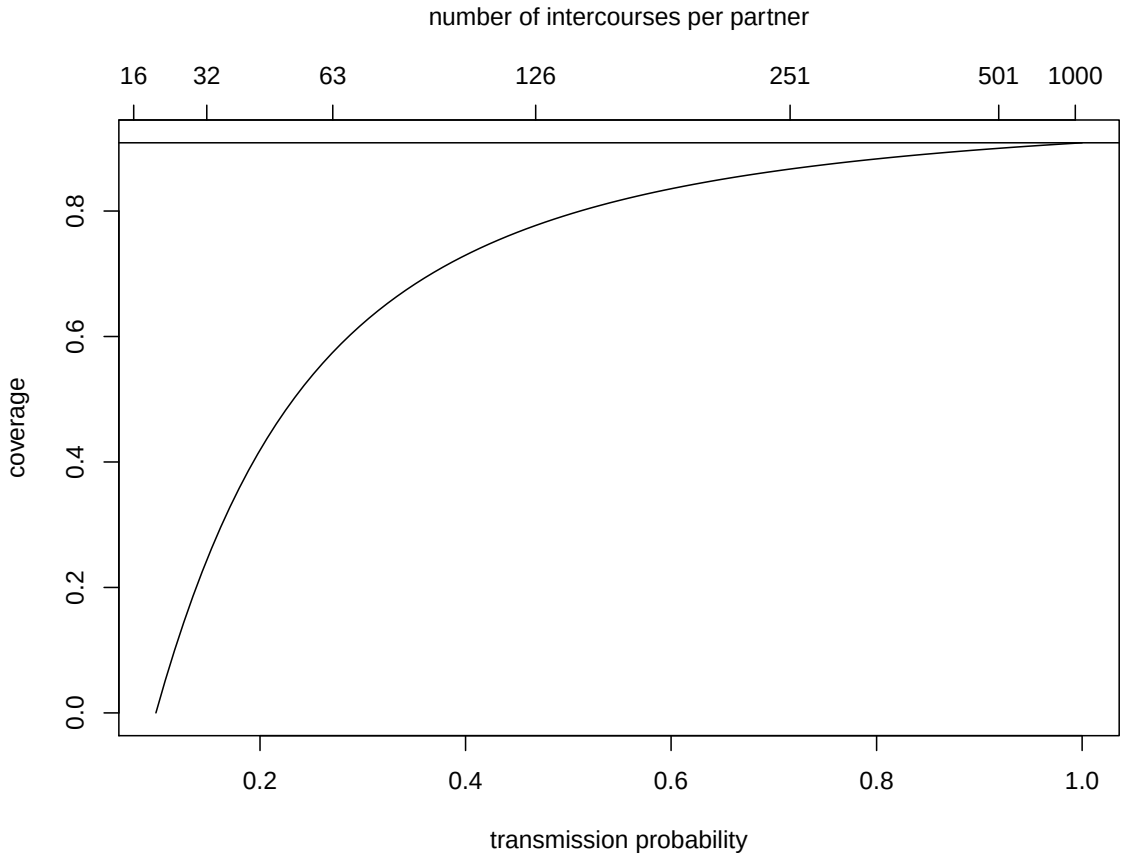
- phénomène global non trivial
- caractérisé assez simplement pour certains graphes aléatoires
- dans la limite des grands graphes

## Quel intérêt pratique ?

- si on a le graphe, une alternative est de faire des simulations :
  - plus précis mais coûteux
  - caractérisation des paramètres ?
- incontournable dans les situations de connaissance incomplète

# Application

- VIH à Cuba
- probabilité critique $T_c \simeq 0.099$ : valeur peu réaliste



# Résumé



## Intérêts

- analyse théorique possible
- résultats utiles : percolation, résilience, etc.
- très nombreux autres résultats (cas orienté, cas biparti, phénomènes plus complexes, etc)

## Limitations

- résultats valables asymptotiquement et en moyenne
- modèles pas assez réalistes :
  - distribution des degrés bien représentée
  - petit monde
  - mais *clustering* mal maîtrisé
  - et surtout pas de corrélation entre les sommets
- fouille de données ?

# Plan

# Schémas fréquents

## Sous-ensembles fréquents

- on se donne une série de sous-ensembles d'un très grand ensemble (par ex. des paniers d'achats)
- on recherche les sous-ensembles fréquents (qui apparaissent au moins x % du temps)
- principe *a priori* :
  - un sous-ensemble d'un sous-ensemble fréquent est fréquent (comptages décroissants)
  - recherche gloutonne et efficace
- très utilisé en fouille de données

## Schémas fréquents

Extension du principe à d'autres structures :

- sous-séquences (prise en compte de l'ordre)
- sous-graphes

# Sous-graphes

## Isomorphisme de graphes

Deux graphes $(V, E)$ et $(W, F)$ sont isomorphes s'il existe une bijection $\phi$ de $V$ dans $W$ telle que $(a, b) \in E \Leftrightarrow (\phi(a), \phi(b)) \in W$.

## Isomorphisme de sous-graphes

Un isomorphisme de sous-graphe de $(V, E)$ dans $(W, F)$ est une fonction injective $\psi$ de $V$ dans $W$ telle que
$(a, b) \in E \Rightarrow (\psi(a), \psi(b)) \in W$.

## Complexité

- vérifier que 2 graphes sont isomorphes est NP (mais on ne sait pas si c'est NP complet)
- vérifier qu'un graphe a un isomorphisme de sous-graphe dans un autre est NP-complet

# Recherche de sous-graphes fréquents

## Cadre classique

- on se donne un ensemble de graphes
- on cherche des sous-graphes fréquents dans l'ensemble des graphes
- on cherche une seule fois un sous-graphe dans chaque graphe
- les graphes sont étiquetés (les isomorphimes ne changent pas les étiquettes)

## Algorithmes classiques

- principe *a priori* (en largeur) : on cherche tous les sous-graphes fréquents de taille 1, puis de taille 2, etc.
- principe de croissance de schéma (en profondeur) : on fait grossir peu à peu les sous-graphes fréquents

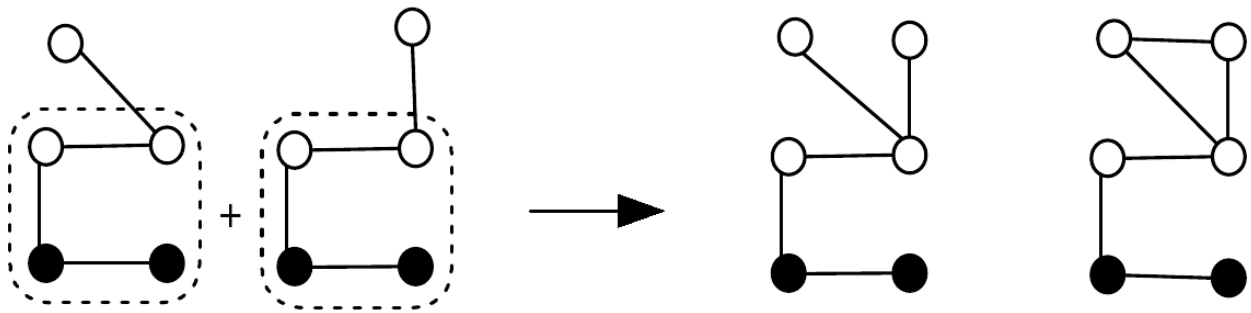
# Génération de candidats

## Cas des sous-ensembles

- *a priori* fonctionne par unions
- si $\{a\}$, $\{b\}$ et $\{c\}$ sont fréquents, de bons candidats sont $\{a, b\}$, $\{a, c\}$ et $\{b, c\}$

## Cas des sous-graphes

- situation beaucoup plus complexe
- nombreuses solutions, par exemple la méthode AGM



Source : figure 9.4 de *Data Mining : Concepts and Techniques* de Jiawei Han et Micheline Kamber

# En pratique

## Un domaine très actif en recherche

- beaucoup d'algorithmes concurrents
- nombreuses extensions du principe de base (contraintes, approximations, graphes sans étiquette, etc.)

## Domaines applicatifs

- analyse de l'usage des sites web (contrainte d'arbres)
- analyse de documents XML (idem)
- chemoinformatique et bioinformatique (molécules)

## Coût algorithmique

- théoriquement énorme
- fonctionnement étonnement rapide sur les graphes creux et bien étiquetés

# Cas d'un seul graphe

## Situation plus complexe

- problème NP-complet associé
- problème de recouvrement des instances d'un sous-graphe
- problème de comptage des instances :
  - il faut une mesure décroissante (principe *a priori*)
  - beaucoup de mesures conduisent à des problèmes NP-complet

## Minimum Image

- un comptage « efficace »
- principe :
  - on considère tous les isomorphismes de sous-graphe de $g$ dans $G$
  - pour chaque sommet de $g$, on compte les sommets distincts de $G$ obtenus grâce aux isomorphismes
  - on dit que $g$ apparaît autant de fois dans $G$ que le plus petit nombre de sommets distincts obtenus

# En pratique

## Synthèse

La situation est très similaire à celle des graphes multiples :

- domaine très actif :
  - nombreux algorithmes
  - variantes : recherche approximative, contraintes, connexions indirectes, etc.
- impossible en théorie (NP-complet) mais solutions approchées efficaces (quelques heures de calcul pour une dizaine de millions de sommets)
- domaines applicatifs similaires mais avec une modélisation différente

## Problèmes (communs aux deux situations)

- méta paramètre : seuil du cas « fréquent »
- on obtient souvent beaucoup de sous-graphes

# Plan

# Classification

## Cas général

- trouver des groupes d'objets similaires et différents des objets d'autres groupes
- applications :
  - résumé/compression
  - structuration
  - découverte de schémas et d'objets atypiques
  - etc.

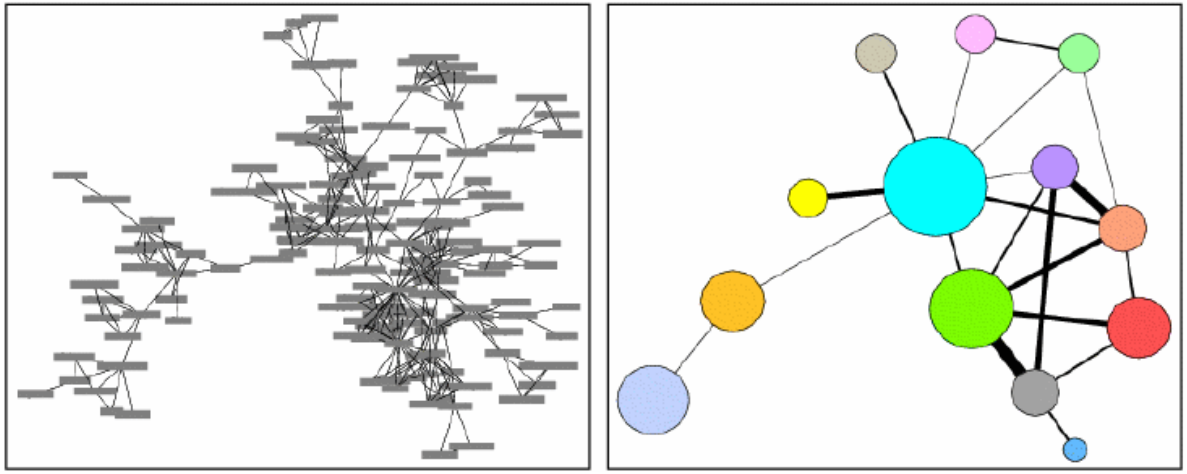
## Cas des graphes

- regrouper les sommets d'un graphe
- en s'appuyant sur les liens

# Classification de sommets

## Applications

- faire apparaître des « communautés »
- simplifier le graphe (par ex. pour la visualisation)



Source :*Finding and evaluating community structure in networks, M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113, 2004*

## Difficultés

- structure non numérique
- interprétation des liens
- cas mixte (liens + valeurs locales)

# Dissimilarités

## Solution « générique »

- définir une dissimilarité entre les sommets d'un graphe
- appliquer un algorithme de classification adapté (hiérarchique, K médiane, etc.)

## Choix de la dissimilarité

- fausse bonne idée : distance géodésique
- solutions plus pertinentes (pas toujours !) :
  - principe des marcheurs aléatoires : le graphe comme une chaîne de Markov
  - par ex. la distance « aller retour » : temps moyen pour aller d'un sommet à un autre et revenir
  - ou encore une propagation inspirée de l'équation de la chaleur

# Marches aléatoires

## Principe de base

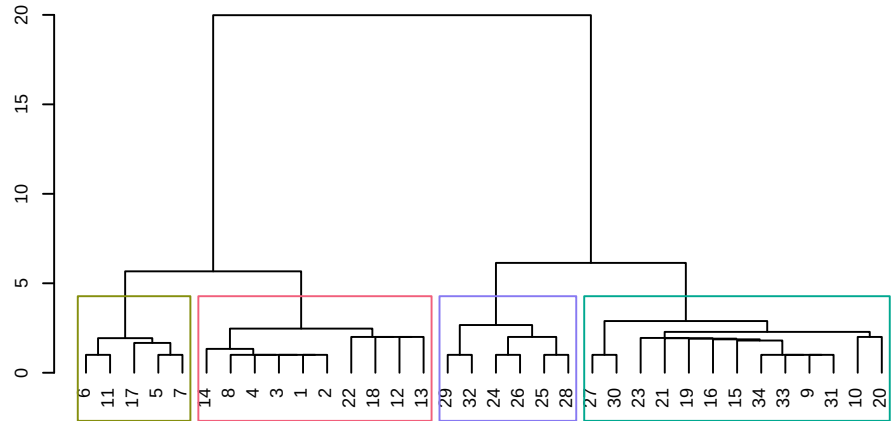
- chaîne de Markov dont les états sont les sommets du graphe $G = (V, E)$
- probabilité de transition de $a$ vers $b$ :
  - 0 si $(a, b) \notin E$
  - $\frac{1}{d(a)}$ si $(a, b) \in E$
- extension évidente au cas pondéré
- autre extension : marcheur paresseux

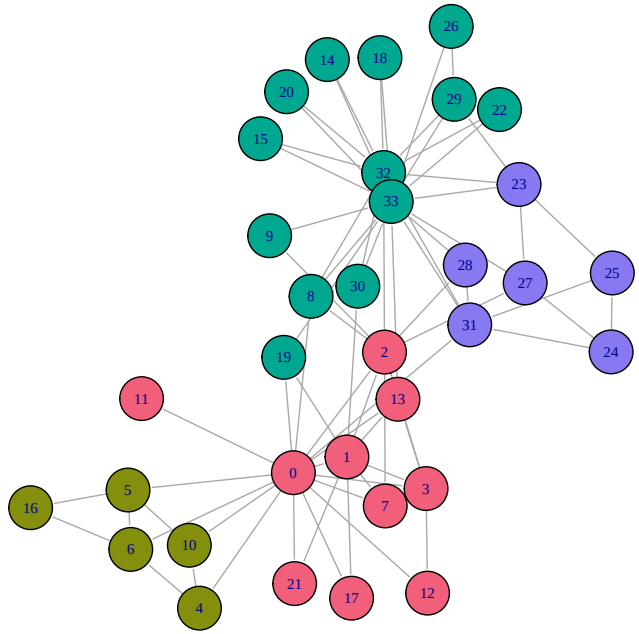
## Concepts importants

- temps d'accès $H(a, b)$ : nombre moyen de saut pour passer de $a$ à $b$
- temps « *commute* » (aller-retour) : $C(a, b) = H(a, b) + H(b, a)$
- distance naturelle : $C(a, b)$

# Exemple

## Distance géodésique

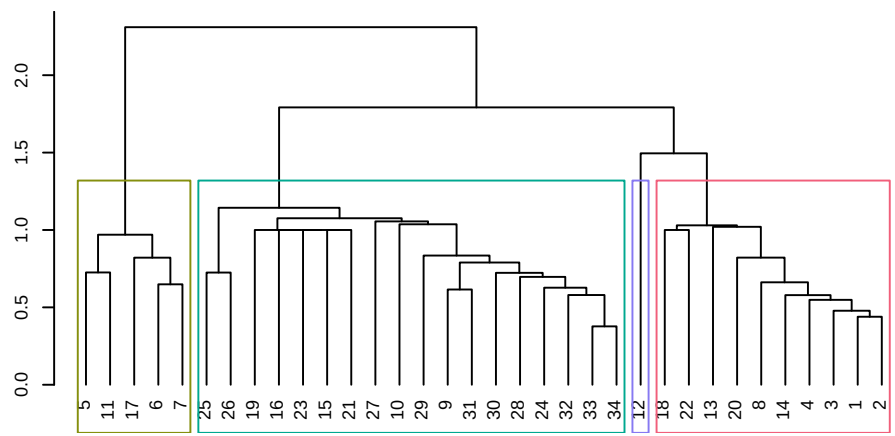


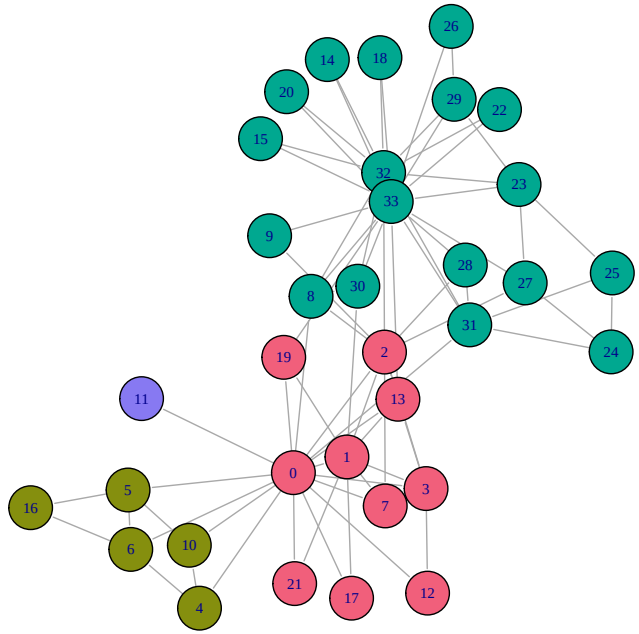




Ward

# Exemple

## Distance aller retour

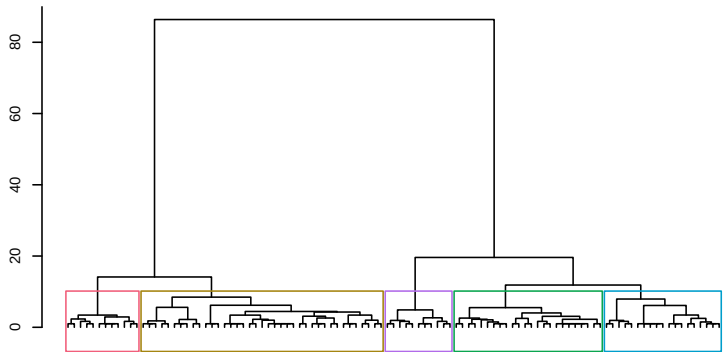



Ward

# Exemple

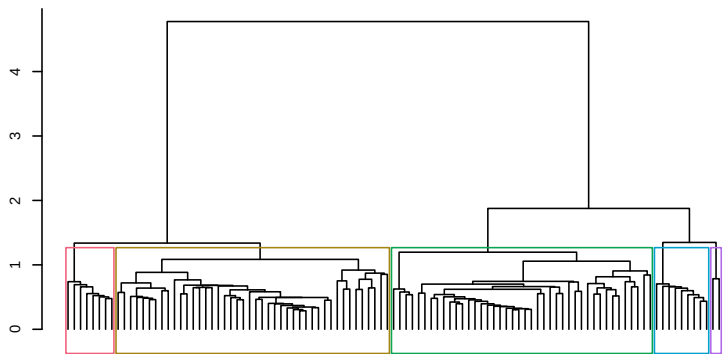
## Distance géodésique

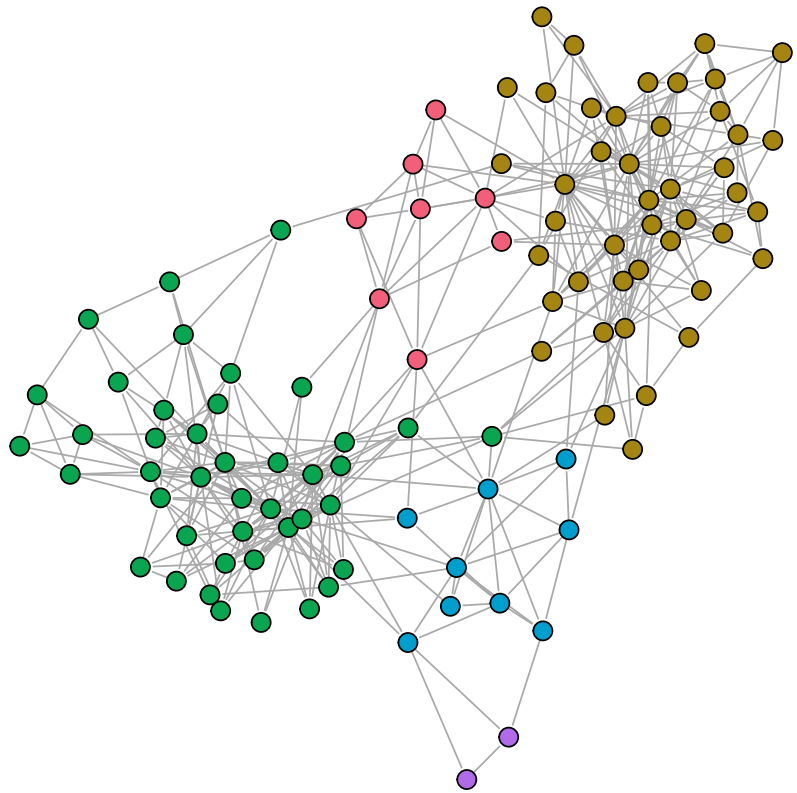



Ward

# Exemple

## Distance aller retour



Ward

# Bilan

## « Métrisation »

- dissimilarité $\Rightarrow$ classification (rien de spécifique aux graphes)
- distance $\Rightarrow$ représentation vectorielle (implicite)

## Souple et riche

- nombreux choix de mesures (combinaison possible avec des informations locales)
- nombreux choix d'algorithmes de classification

## Limitations

- coût algorithmique : calcul du temps aller retour en $O(|V|^3)$
- limites théoriques complexes (pour certains graphes)
- est-ce vraiment graphique ?

# Qualité d'une partition

## Autre approche générique

- définir un critère de qualité pour une partition (classification)
- optimiser le critère

## Modularité

- mesure proposée par Girvan et Newman en 2004
- très populaire depuis
- principe : trouver des classes contenant de nombreuses arêtes internes et peu d'arêtes externes
- originalité : prise en compte du degré des sommets pour mesurer la significativité d'un lien

# Modularité

## Notations

- graphe pondéré : matrice symétrique $W$
- degré valué : $k_i = \sum_j W_{ij}$
- poids total : $m = \frac{1}{2} \sum_{i,j} W_{ij}$
- modèle nul : $P_{ij} = \frac{k_i k_j}{2m}$

## Définition

La modularité de la partition $C = (C_k)_{1 \leq k \leq C}$ du graphe est donnée par

$$Q(C) = \frac{1}{2m} \sum_{k=1}^{C} \sum_{i \in C_k, j \in C_k} (W_{ij} - P_{ij})$$

On cherche à maximiser $Q$.

# Propriétés

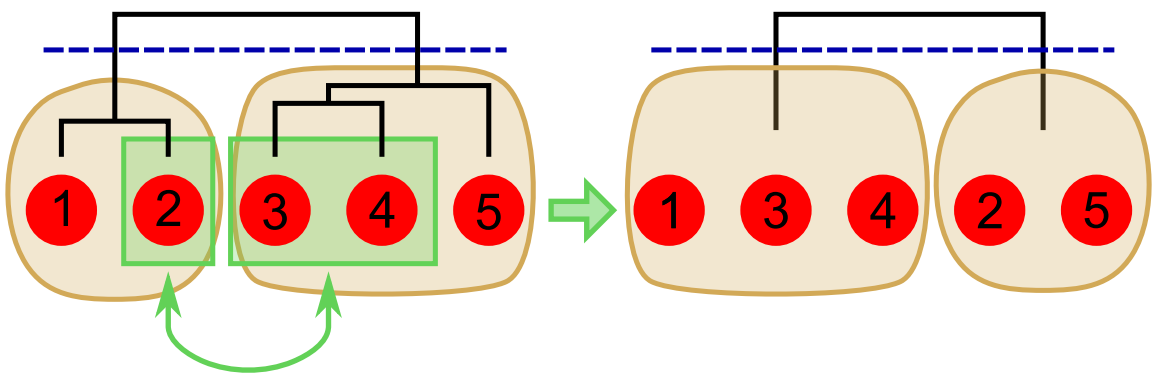
- on note $f(i)$ l'indice de la classe de $i$
- on a
$$Q(C) = \frac{1}{2m} \sum_{i,j} \delta_{f(i)=f(j)} \left(W_{ij} - P_{ij}\right)$$
- et donc si
$$\bar{Q}(C) = \frac{1}{2m} \sum_{i,j} \delta_{f(i)\neq f(j)} \left(W_{ij} - P_{ij}\right)$$
- on a $Q(C) + \bar{Q}(C) = 0$ : situation similaire à celle de l'inertie euclidienne

# Interprétation

- $P_{ij} = \frac{k_i k_j}{2m}$ : modèle de graphe aléatoire de configuration
- significativité d'une arête : $W_{ij} - P_{ij} > 0$
- $Q$ augmente quand les éléments d'une paire $\{i, j\}$ reliés par une arête significative sont placés dans une même classe
- $Q$ augmente quand les éléments d'une paire $\{i, j\}$ reliés par une arête non significative sont placés dans des classes différentes (car $\bar{Q}(C)$ diminue)
- en pratique
  - permet de séparer des sommets de degrés élevés
  - efficace sur les graphes à loi puissance

# Maximisation

- problème combinatoire difficile (NP-complet)
- pas d'astuce simple de type K means
- nombreux algorithmes heuristiques
- par exemple :
  1. placer chaque sommet dans une classe
  2. pour chaque sommet, étudier si son passage dans la classe d'un de ses voisins améliore la modularité et mettre en œuvre la meilleure modification
  3. retourner en 2 jusqu'à stabilisation
  4. construire le graphe induit par la classification et retourner en 1
- raffinement multi-niveau



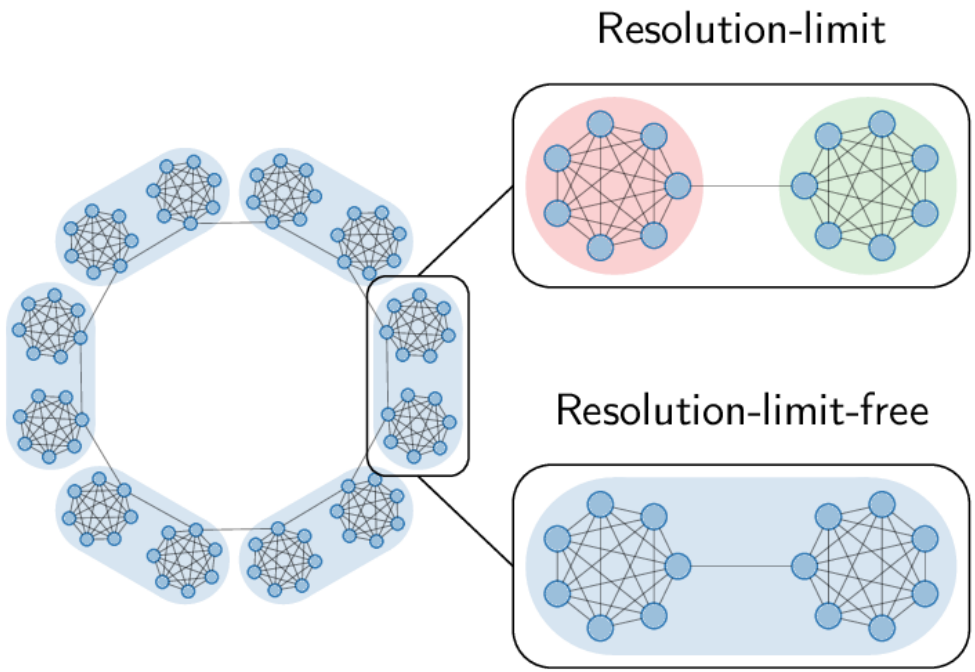
# Choix du nombre de classes

## Problème général

- la qualité d'une classification augmente souvent avec le nombre de classes
- nécessite un critère externe pour le choix du nombre de classes
- nombreuses heuristiques (coude, *gap*, etc.)

## Cas de la Modularité

- choix automatique du nombre de classes
- mais limite de résolution




Source : *Narrow scope for resolution-limit-free community detection, V. A. Traag, P. Van Dooren, and Y. Nesterov, Phys. Rev. E 84, 016114, 2011*

# Significativité de la partition

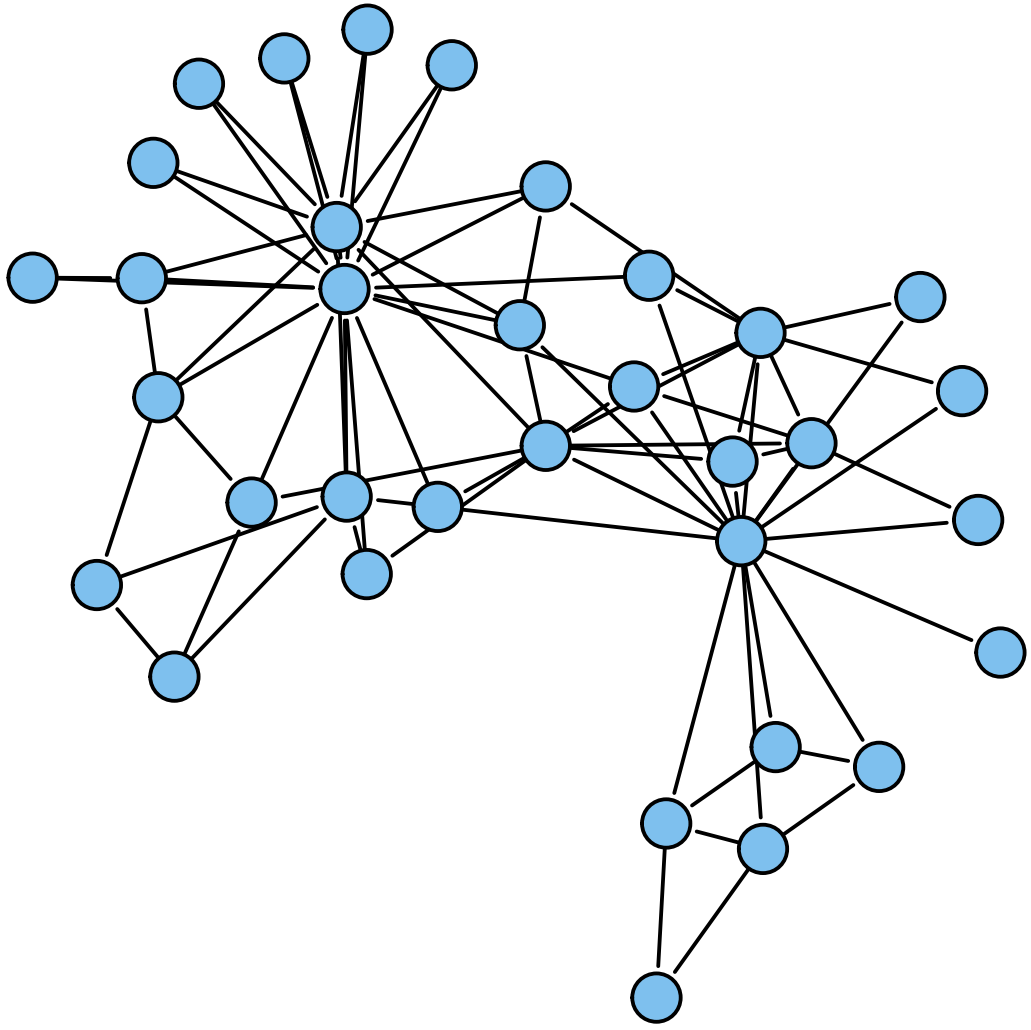
## Problème général

- en général, on trouve toujours $K$ classes si on en cherche $K$
- elles n'ont pas toujours de sens
- cela s'applique même à la modularité

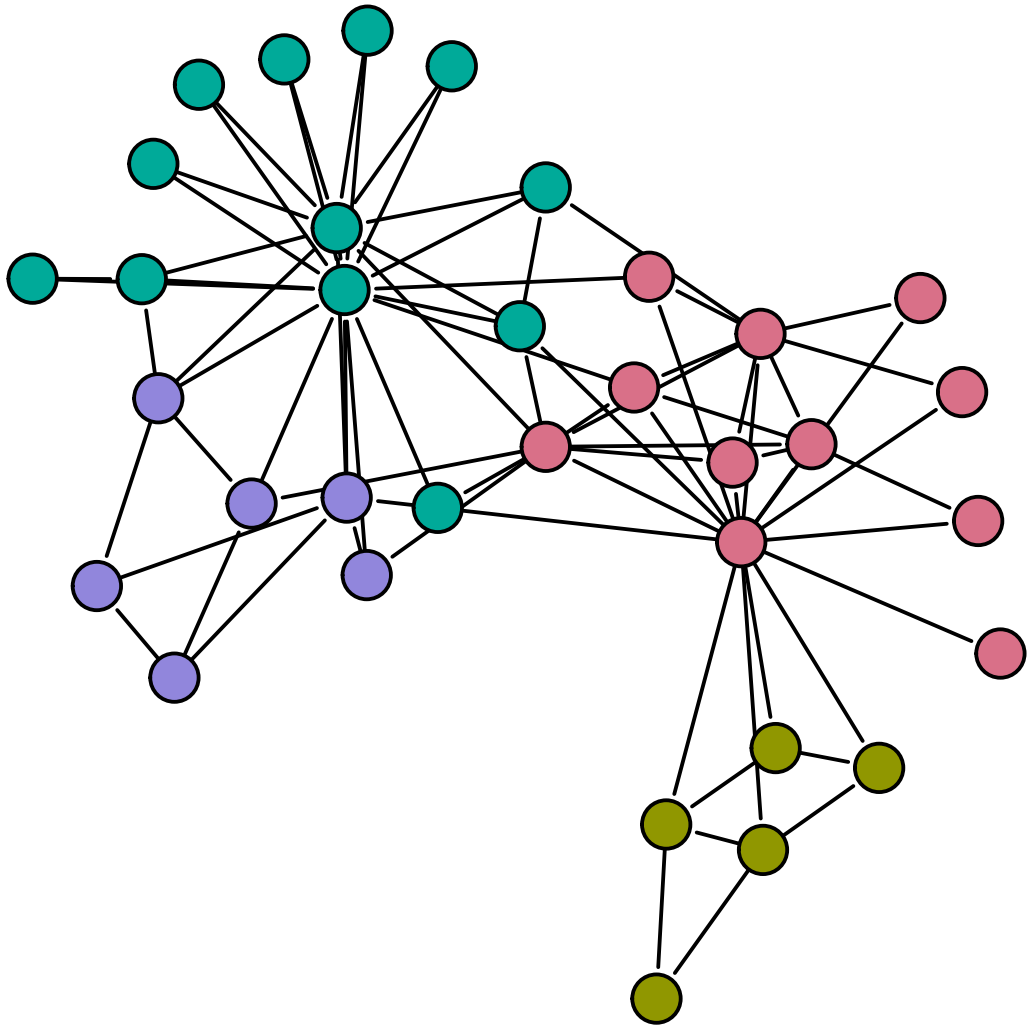
## Une solution par simulation

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

# Exemple

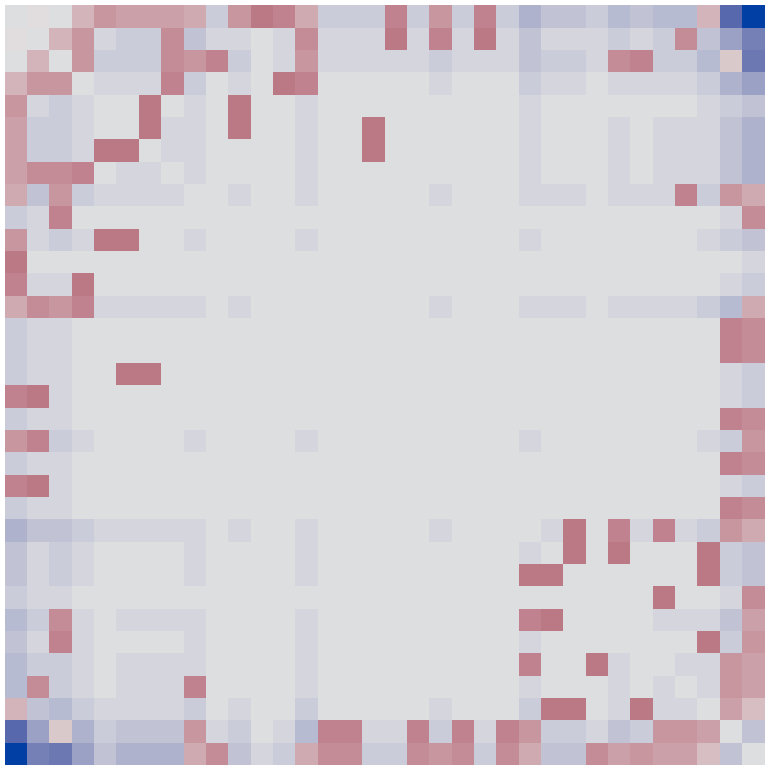


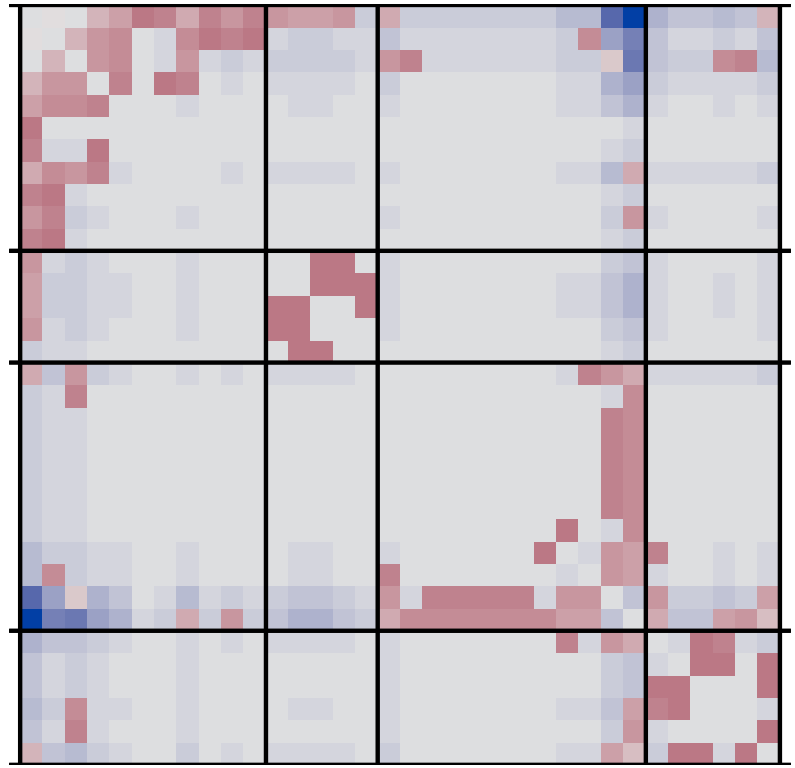
# Exemple



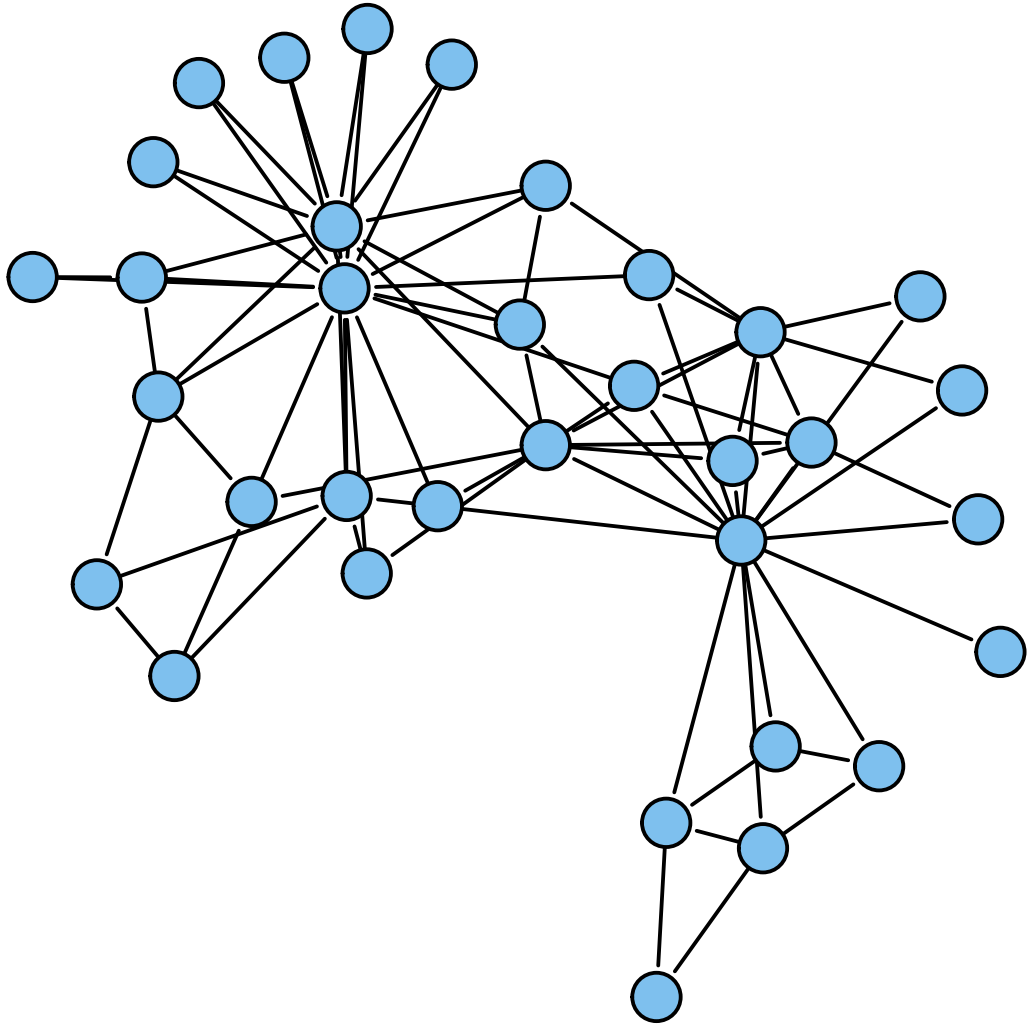
4 classes, modularité $\simeq 0.42$

# Exemple

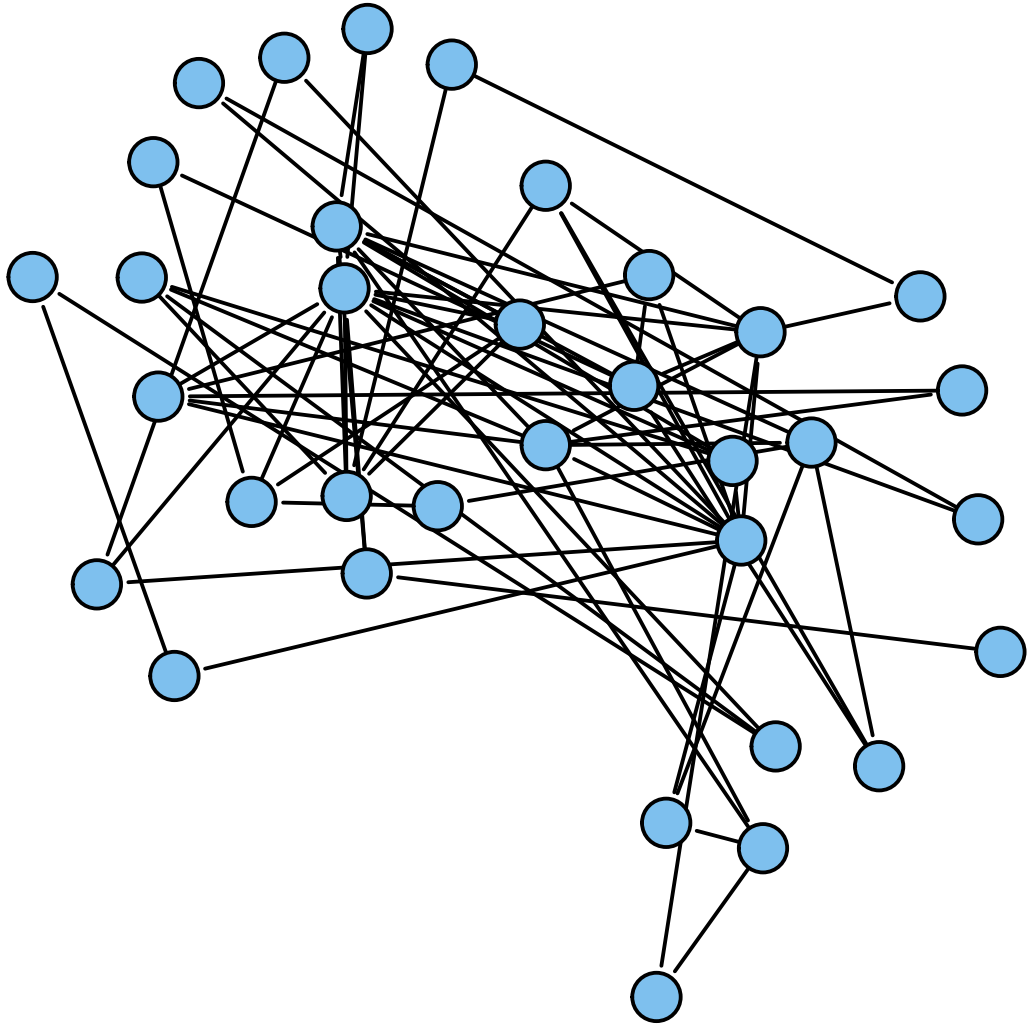




# Simulation

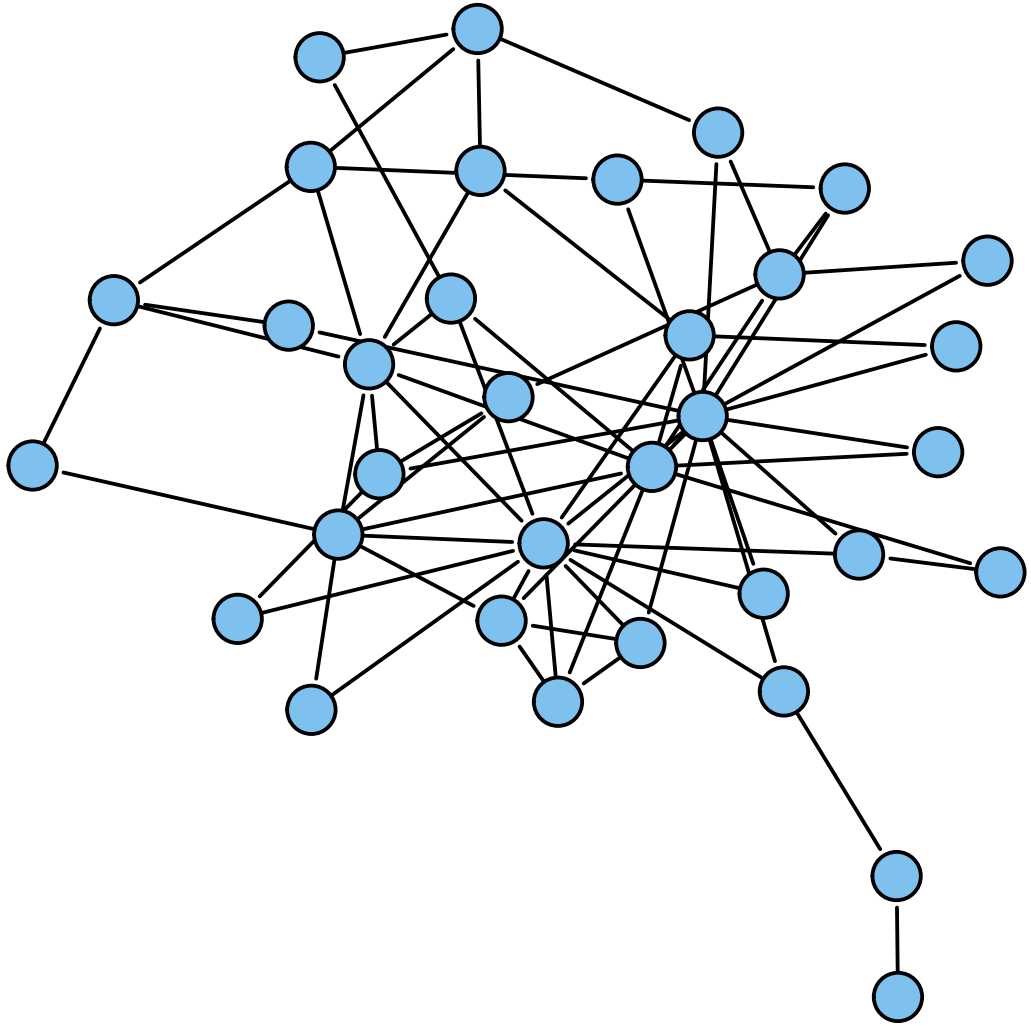


# Simulation



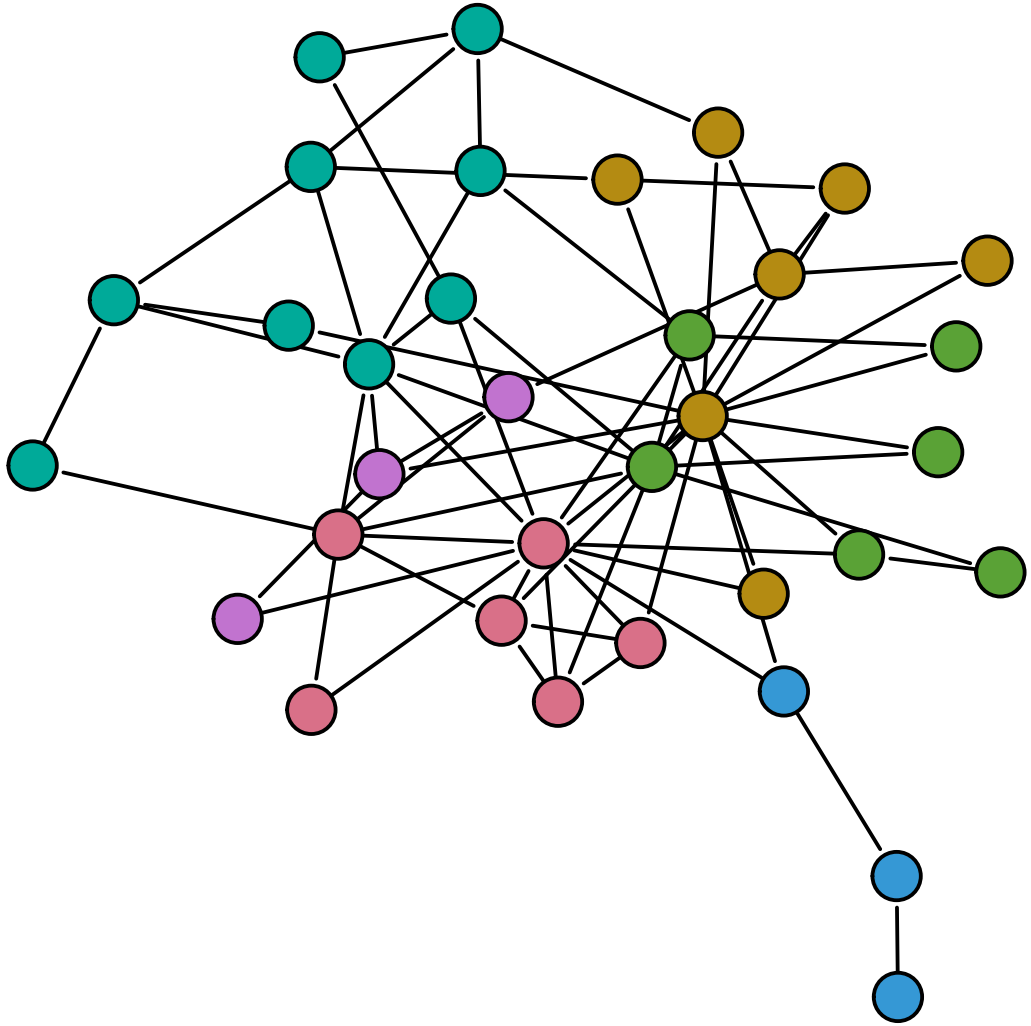
Modèle de configuration : mêmes degrés

# Simulation



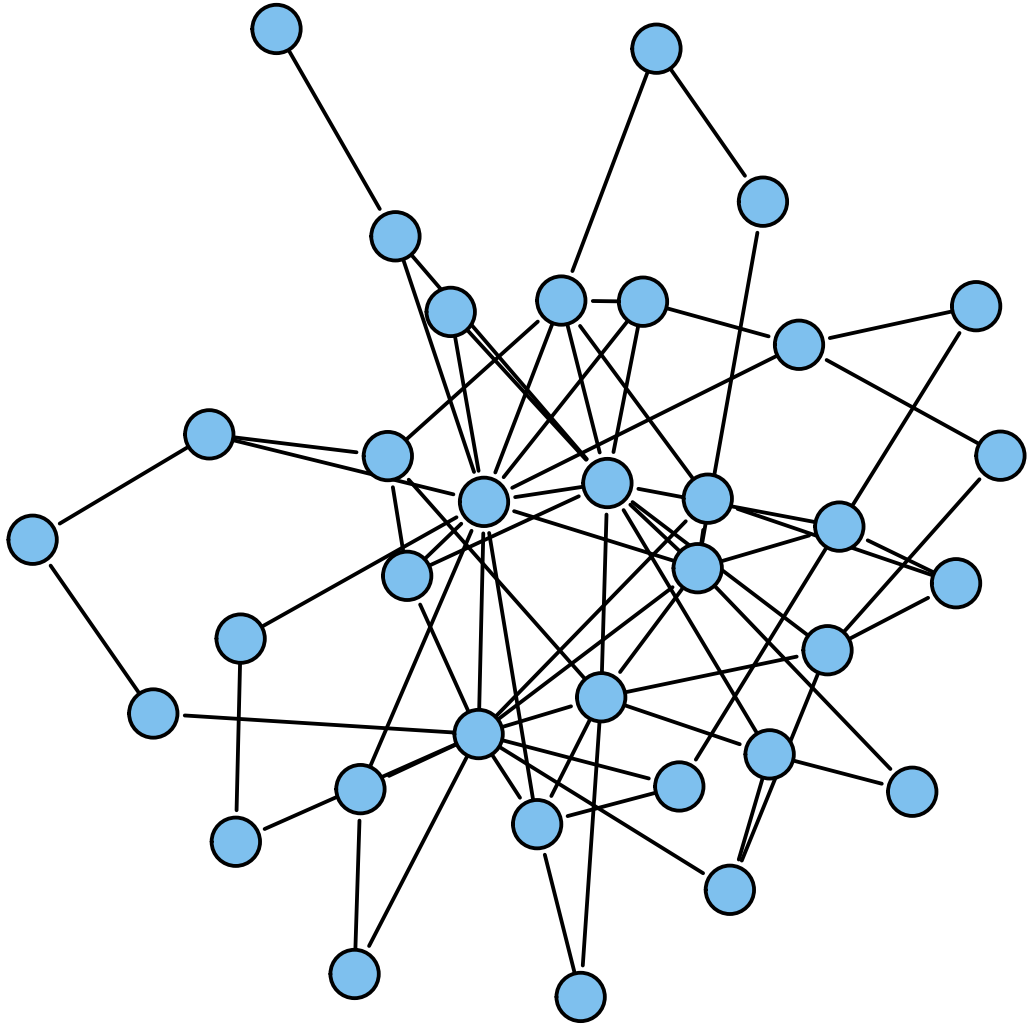
Repositionné

# Simulation



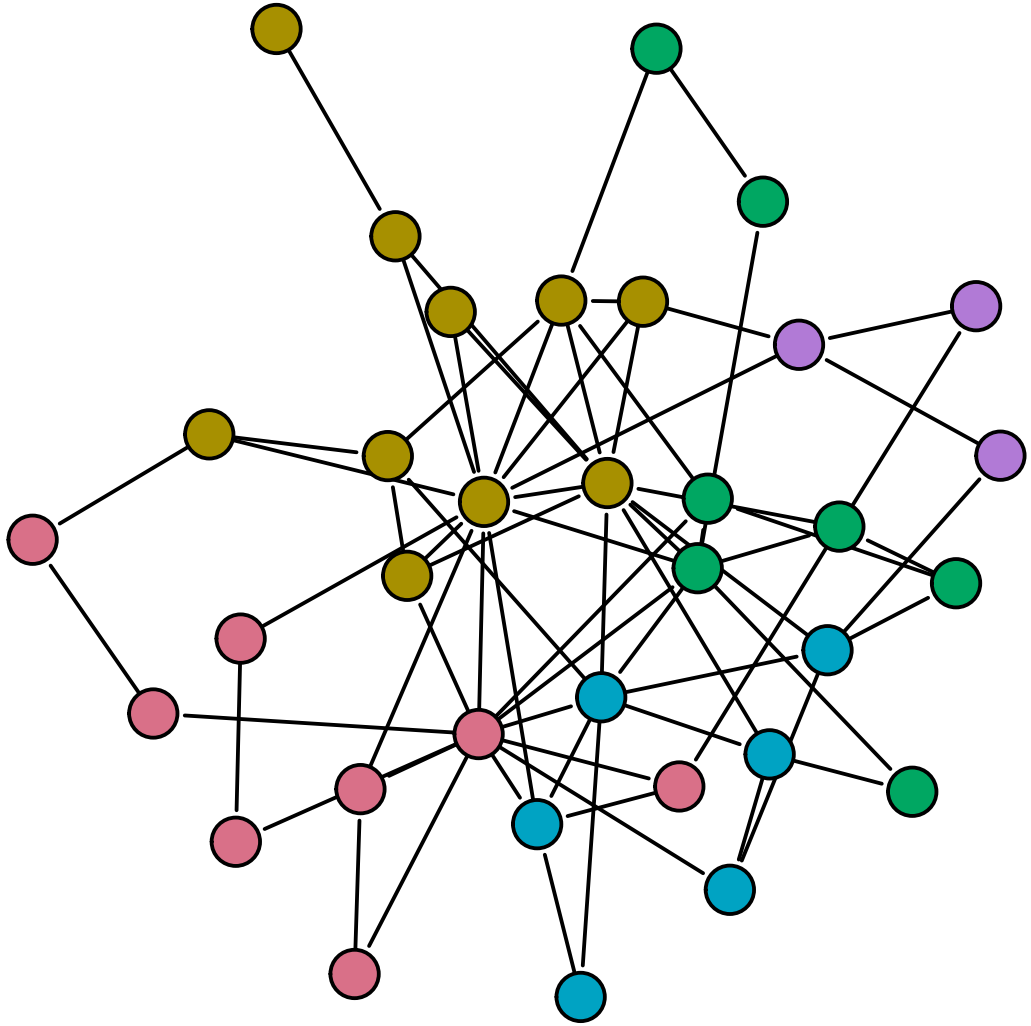
6 classes, modularité $\simeq 0.35$

# Simulation



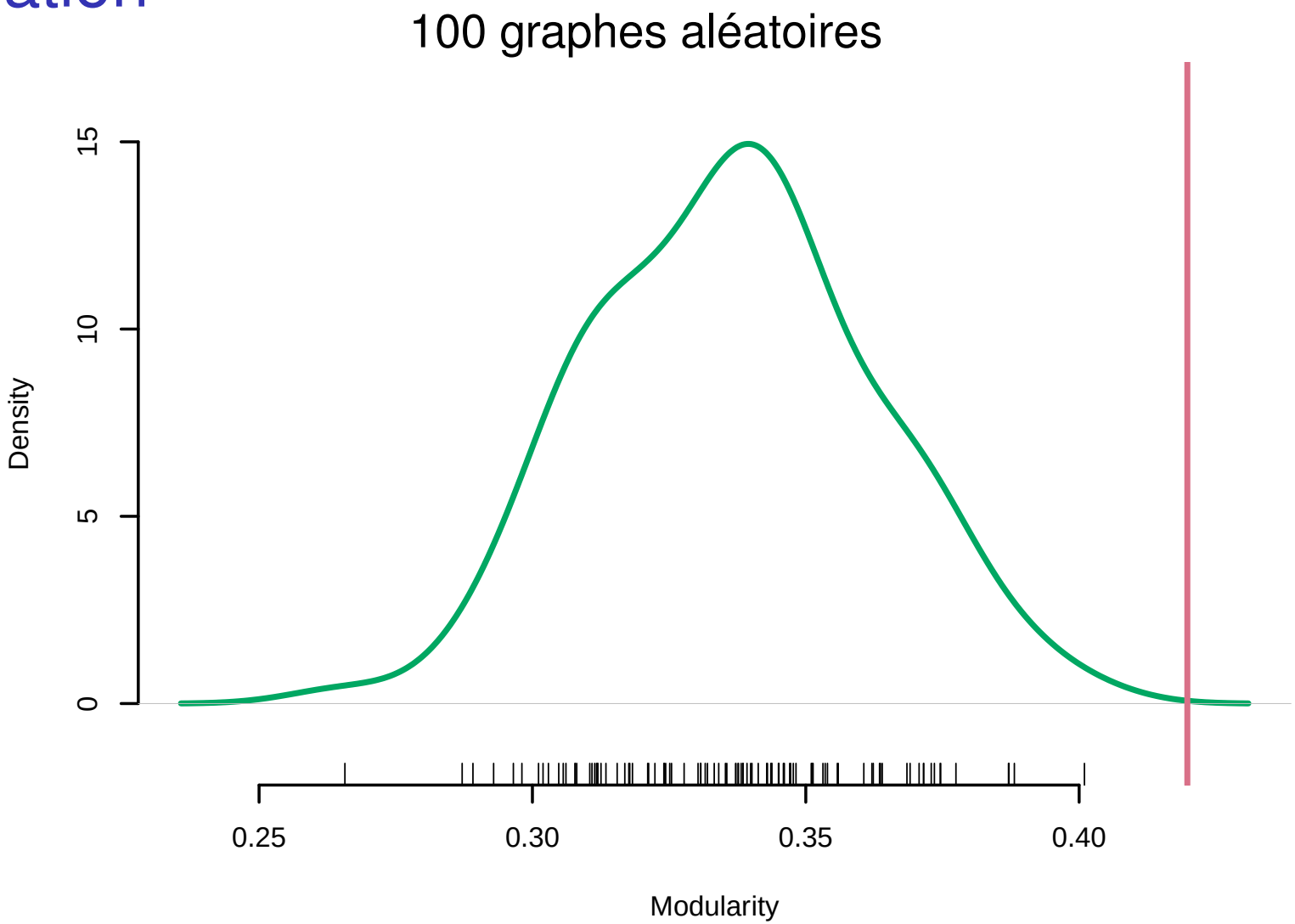
Nouveau tirage

# Simulation



5 classes, modularité $\simeq 0.34$

# Simulation




la classification sur le graphe d'origine a un sens

# Bilan

## Une solution complète

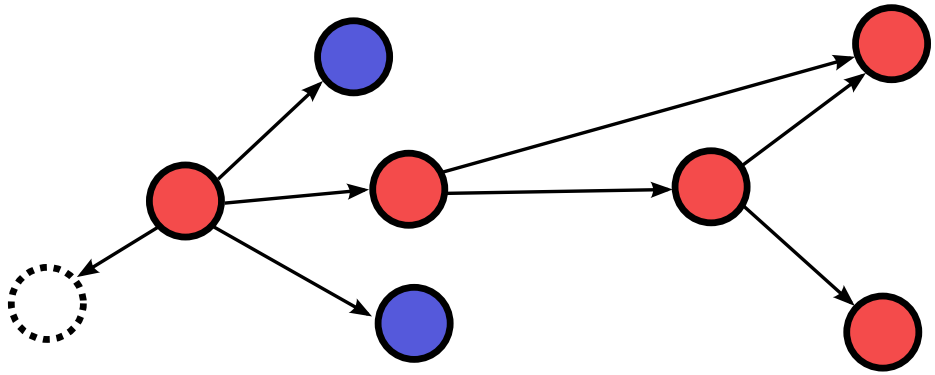
- Modularité : critère spécifique aux graphes (n'induit pas une représentation vectorielle)
- algorithmes d'optimisation rapide : millions de sommets
- choix automatique du nombre de classes (mais limite de résolution)
- test de significativité

## En pratique

- donne de bons résultats, notamment en visualisation
- ne trouve que des classes « modulaires »
- recherches encore très actives (nouveaux algorithmes, extensions, etc.)

# Application

- 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

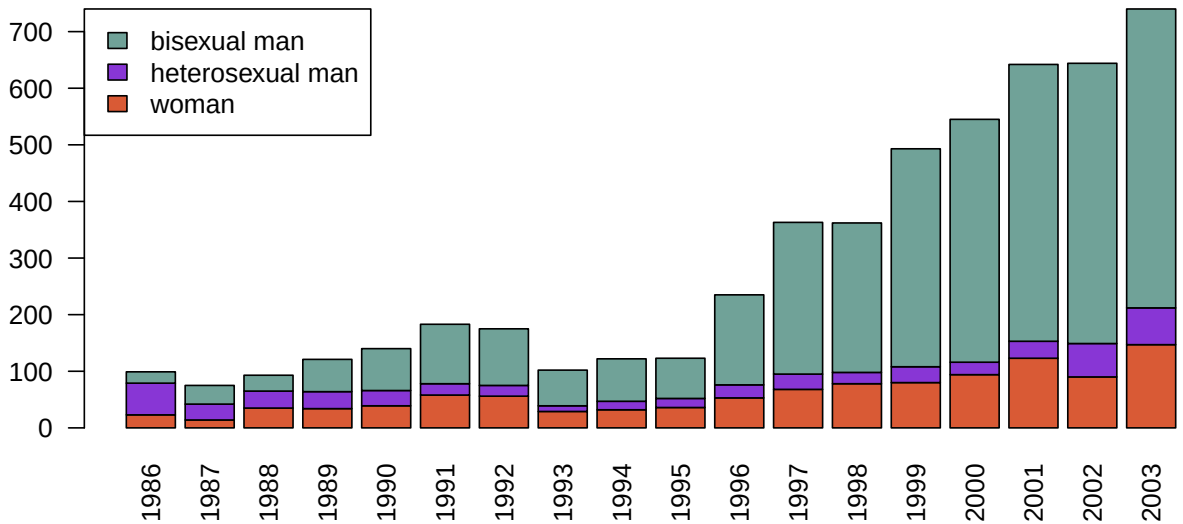
- base de données volumineuse :
  - 5 389 patients décrits par une quinzaine de variables
  - 4 073 relations (graphe assez peu dense)
  - 2 386 patients dans une même composante connexe du graphe (3 168 relations dans cette composante)
- bases « comparables » :
  - Rothenberg et al., 1995
    - étude *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 |

- « détections »

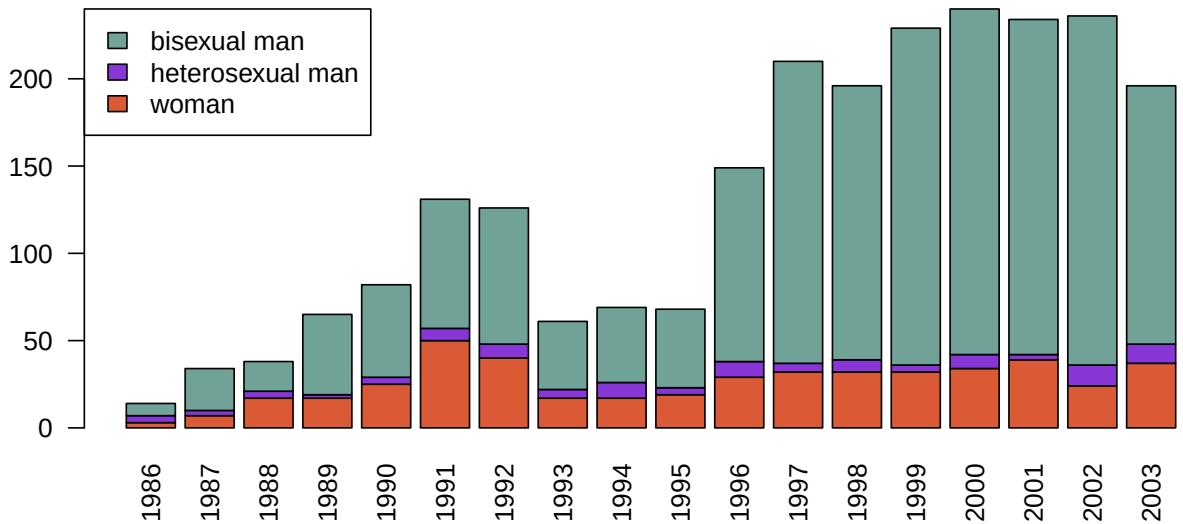


# Aspects macroscopiques

- orientation sexuelle

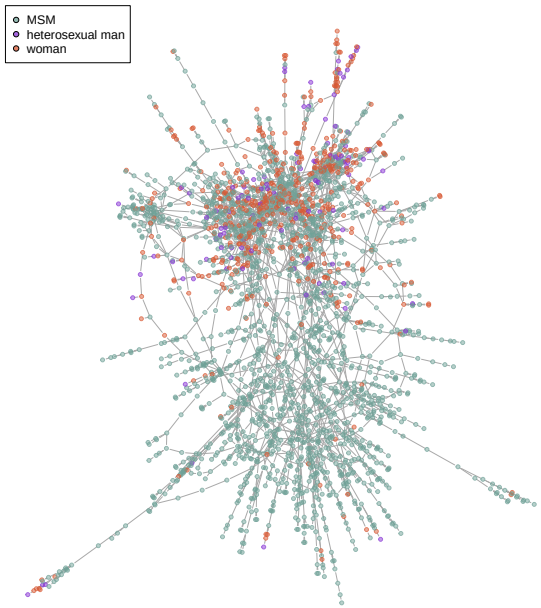
| | population | GCC |
|---|---|---|
| femmes | 0.21 | 0.20 |
| hommes hétéros | 0.11 | 0.05 |
| hommes bisexuels | 0.69 | 0.76 |

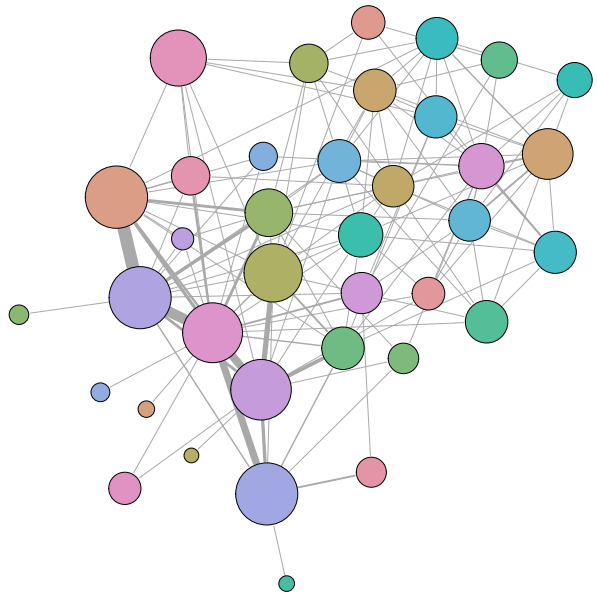
- « détections »



# Visualisation

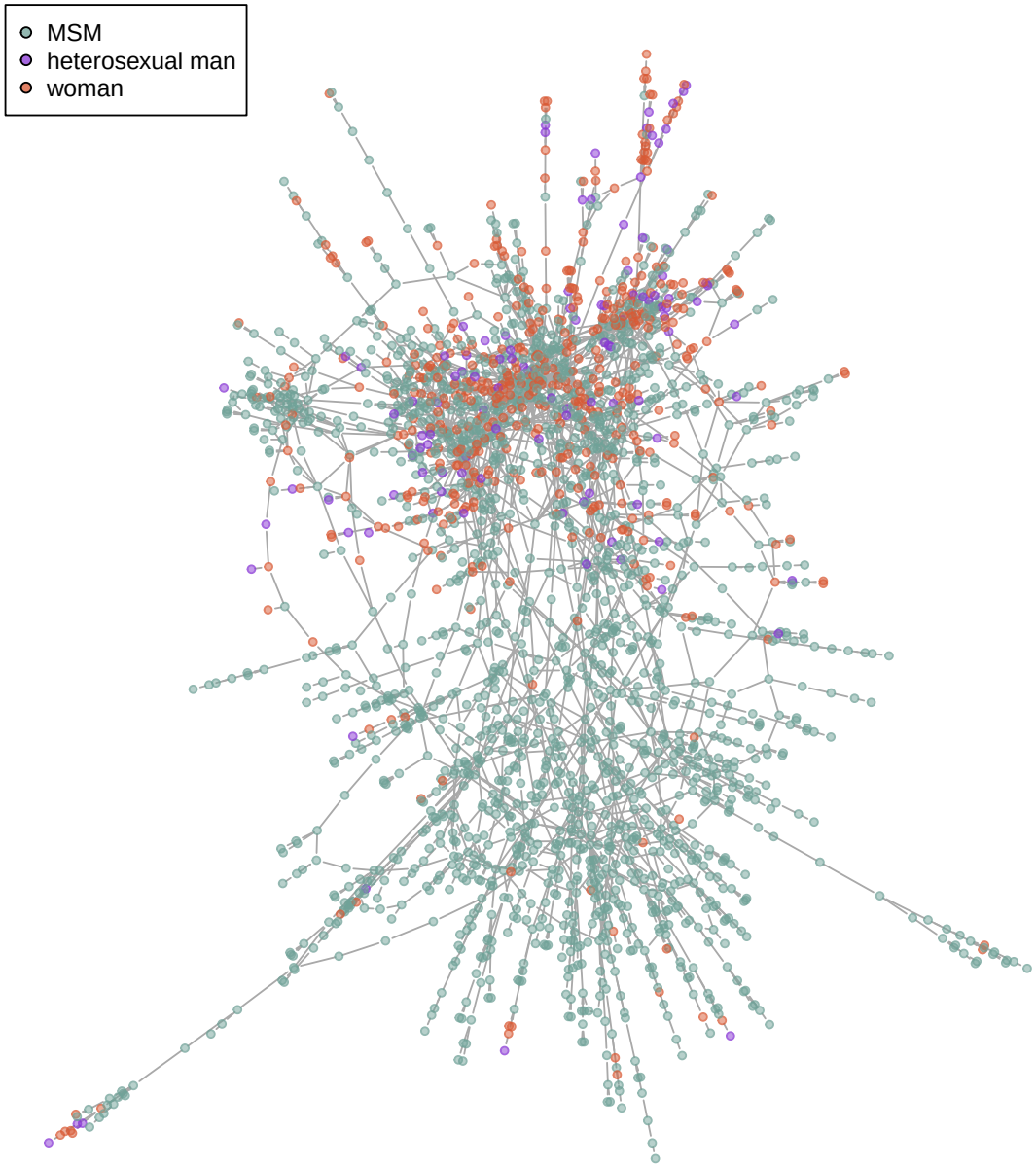
- réduction de complexité :
  - classification des sommets du graphe (modularité)
  - visualisation du graphe des classes



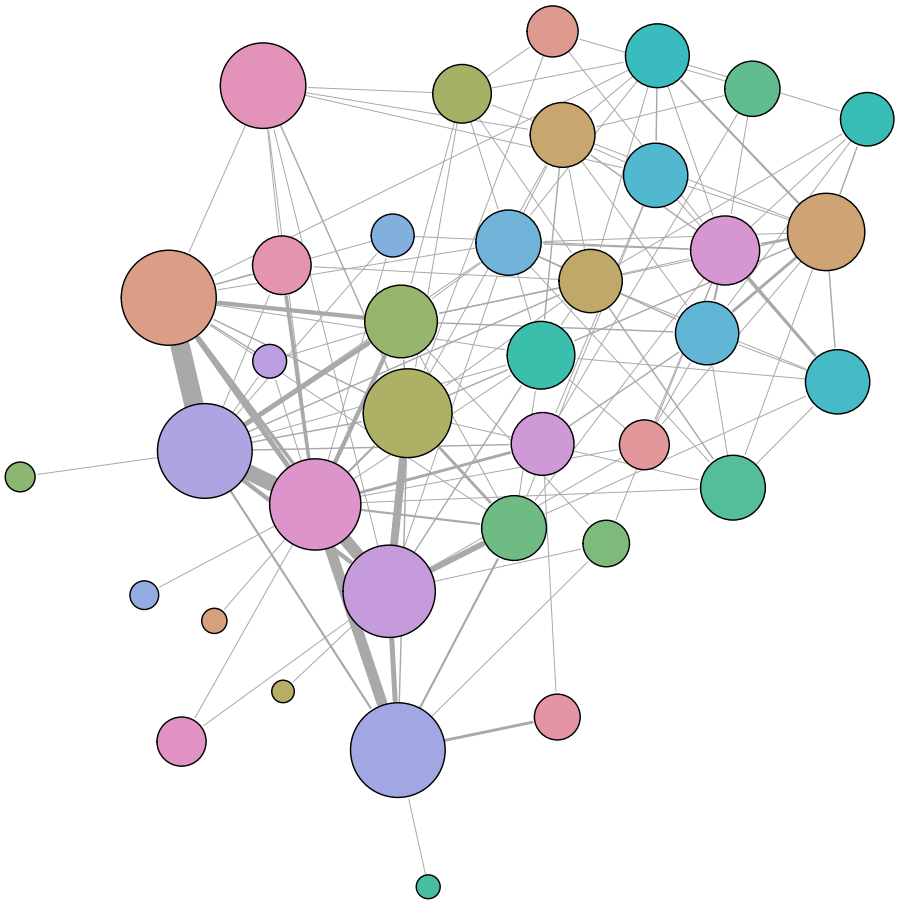



- pertinence ?
  - qualité de la classification
  - lisibilité
  - inférence

# Composante connexe principale

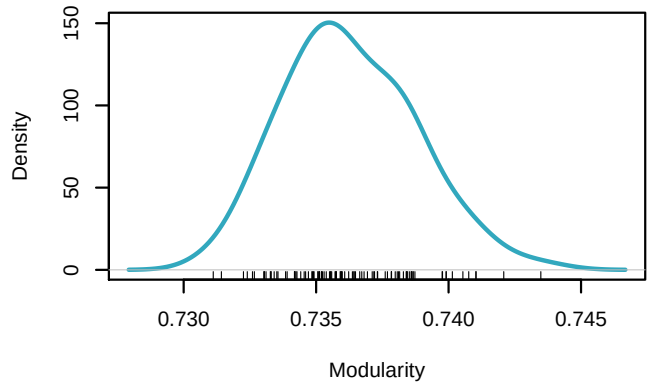


# Classification



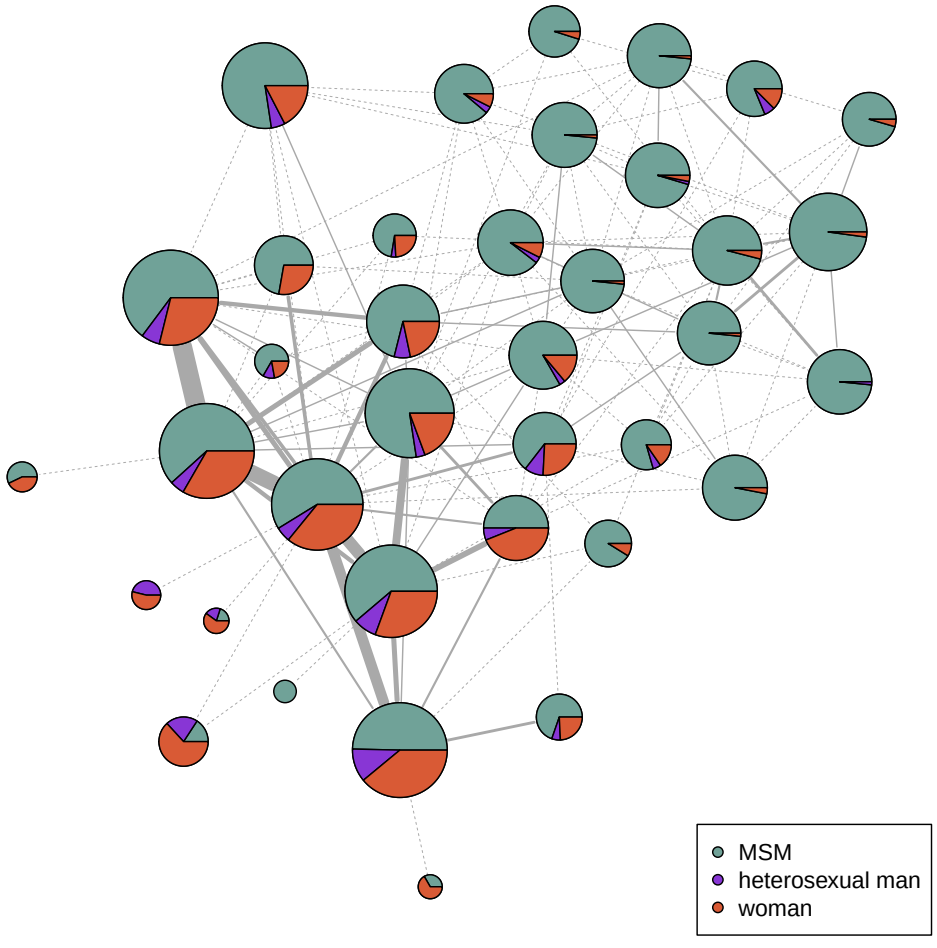
$\Rightarrow$ 37 classes (89.6% des liens internes aux classes)

$\Rightarrow$ modularité $\simeq 0.8522$



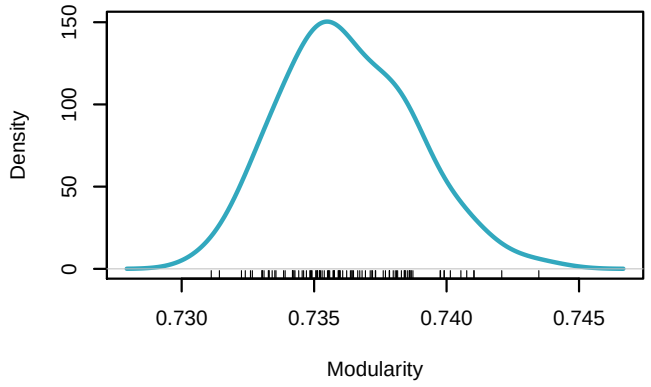

$\Rightarrow$ modularité « aléatoire » $\leq 0.7435$

# Classification




⇒ 37 classes (89.6% des liens internes aux classes)

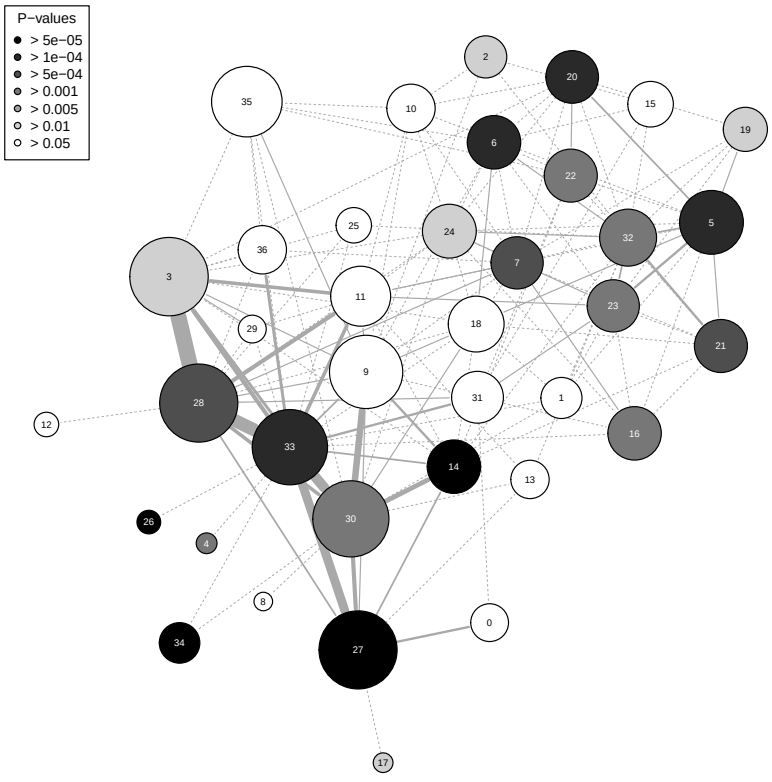
⇒ modularité $\simeq 0.8522$



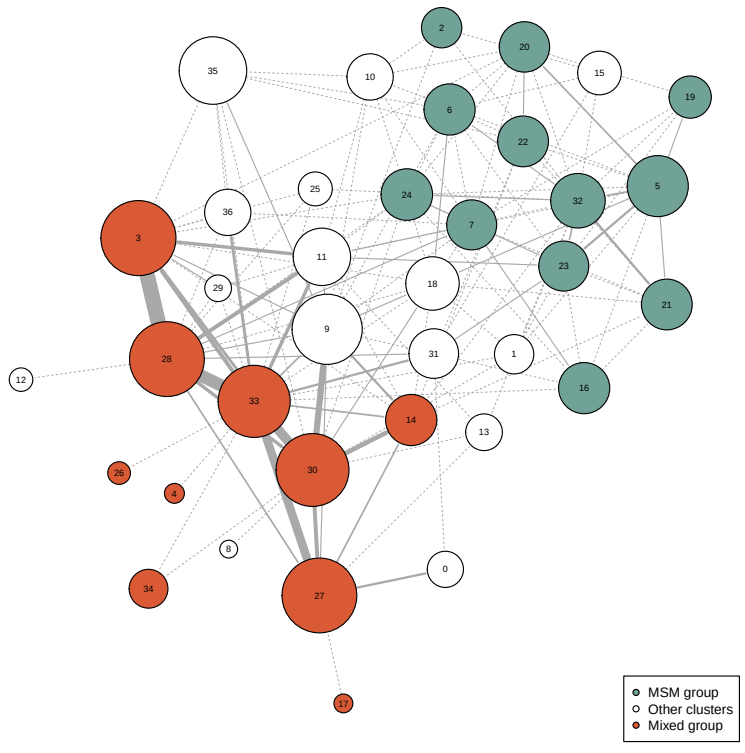

⇒ modularité « aléatoire » $\leq 0.7435$

⇒ visualisation de l'orientation sexuelle

# Statistique graphique

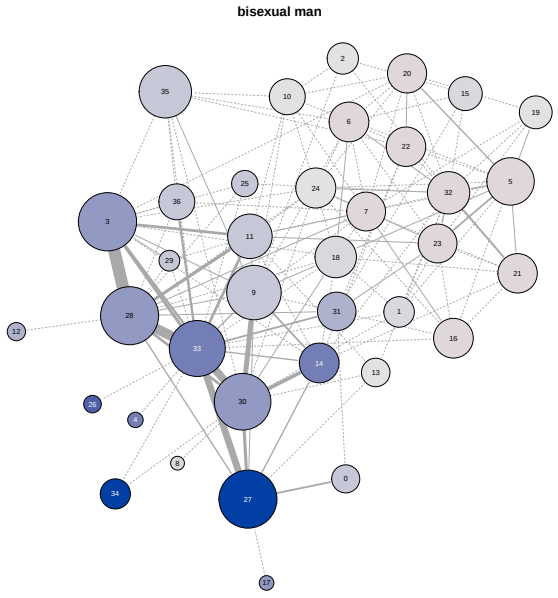


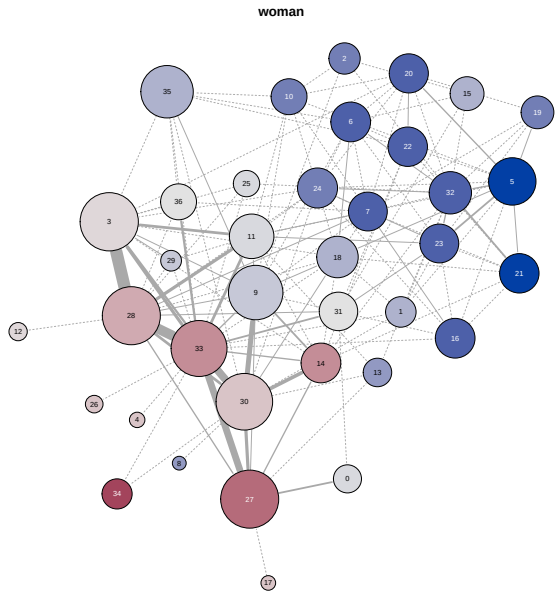

*p-value* d'un test du $\chi^2$ sur la distribution de l'orientation sexuelle

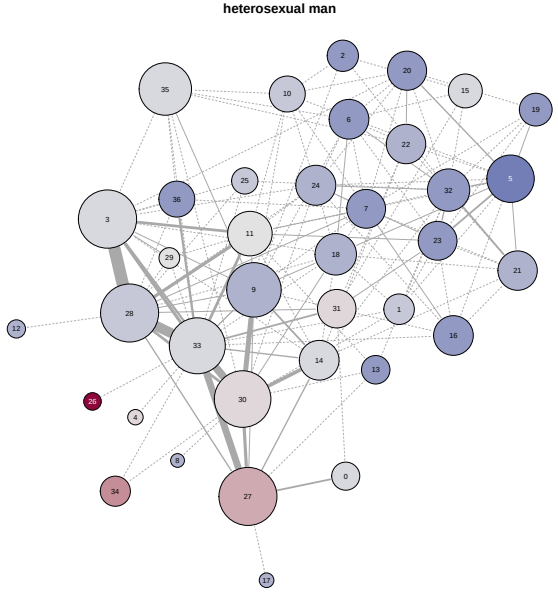



orientation sexuelle atypique

# Statistique graphique

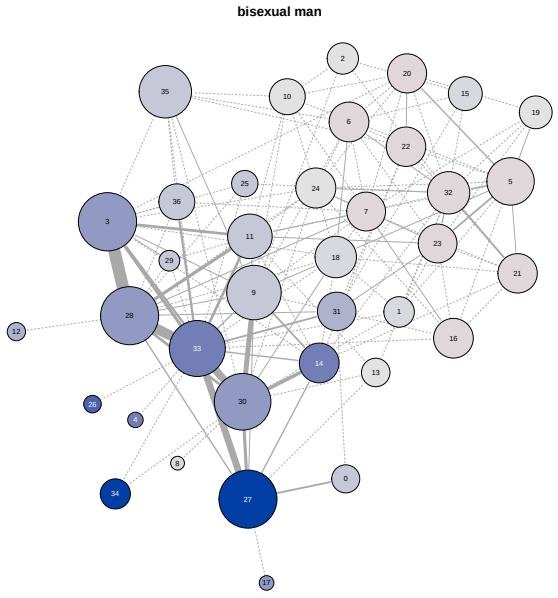


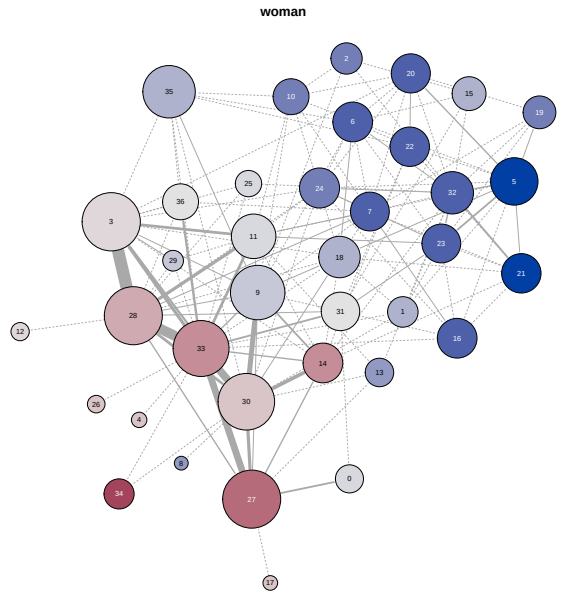



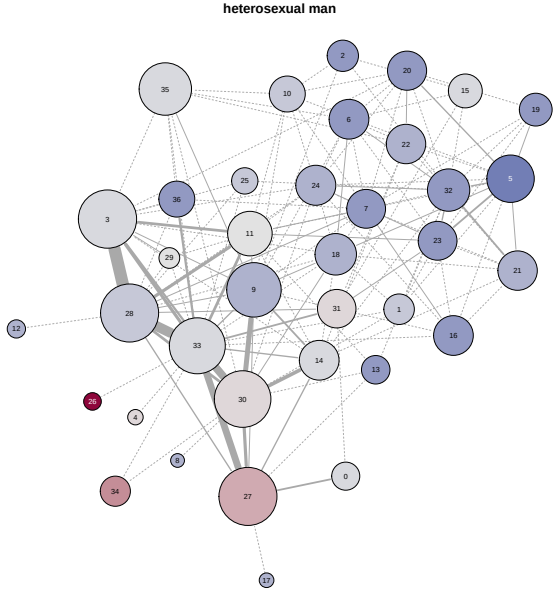



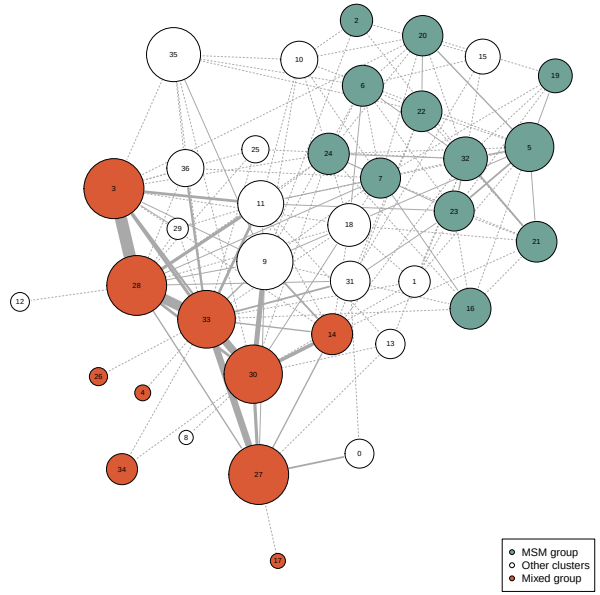

3 macro groupes
clairement identifiés

# Statistique graphique

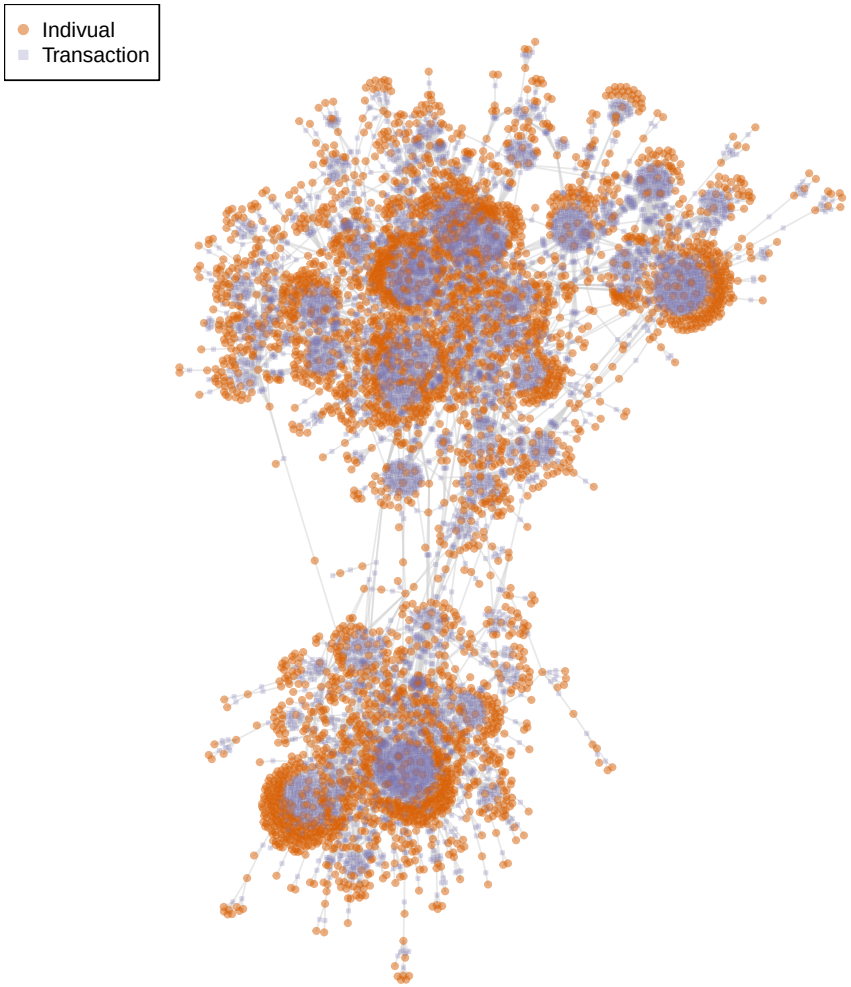


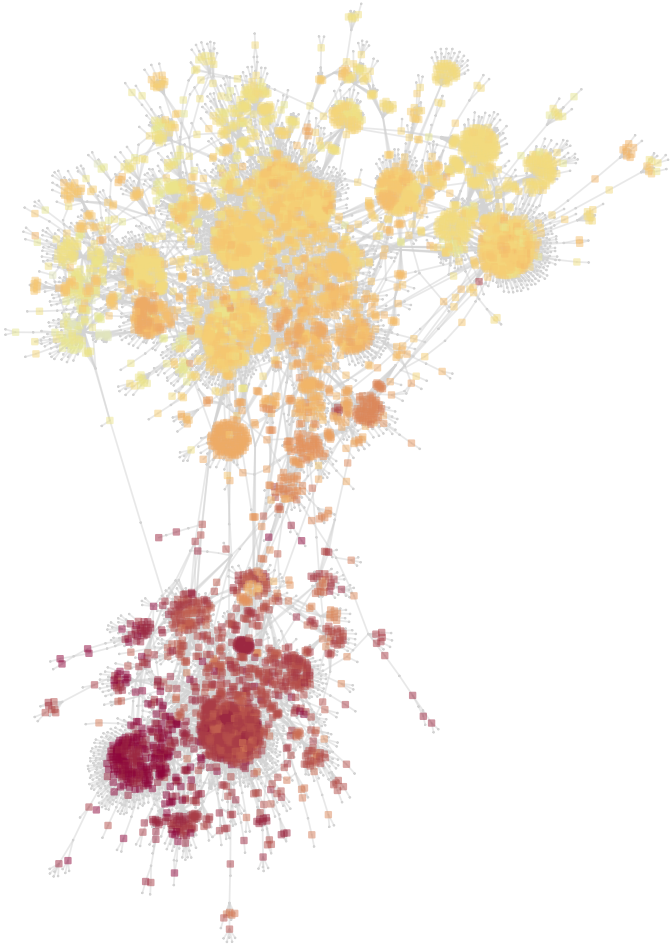



# Données médiévales



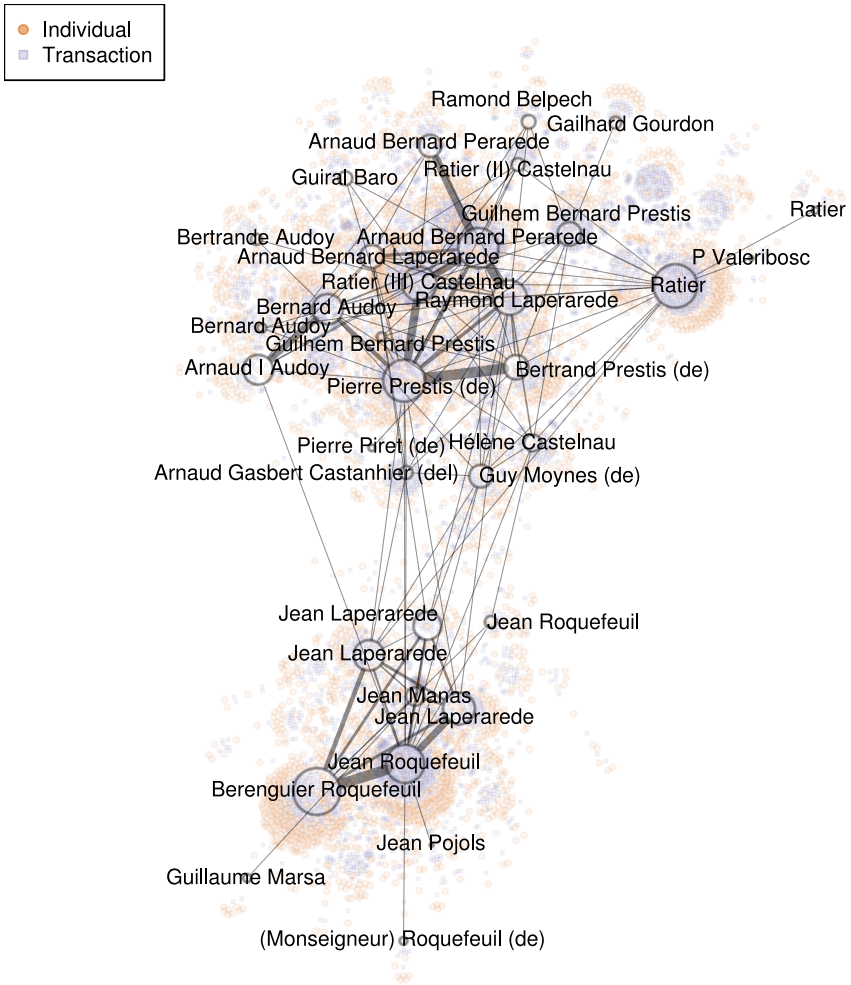

- représentation graphique complexe

# Données médiévales



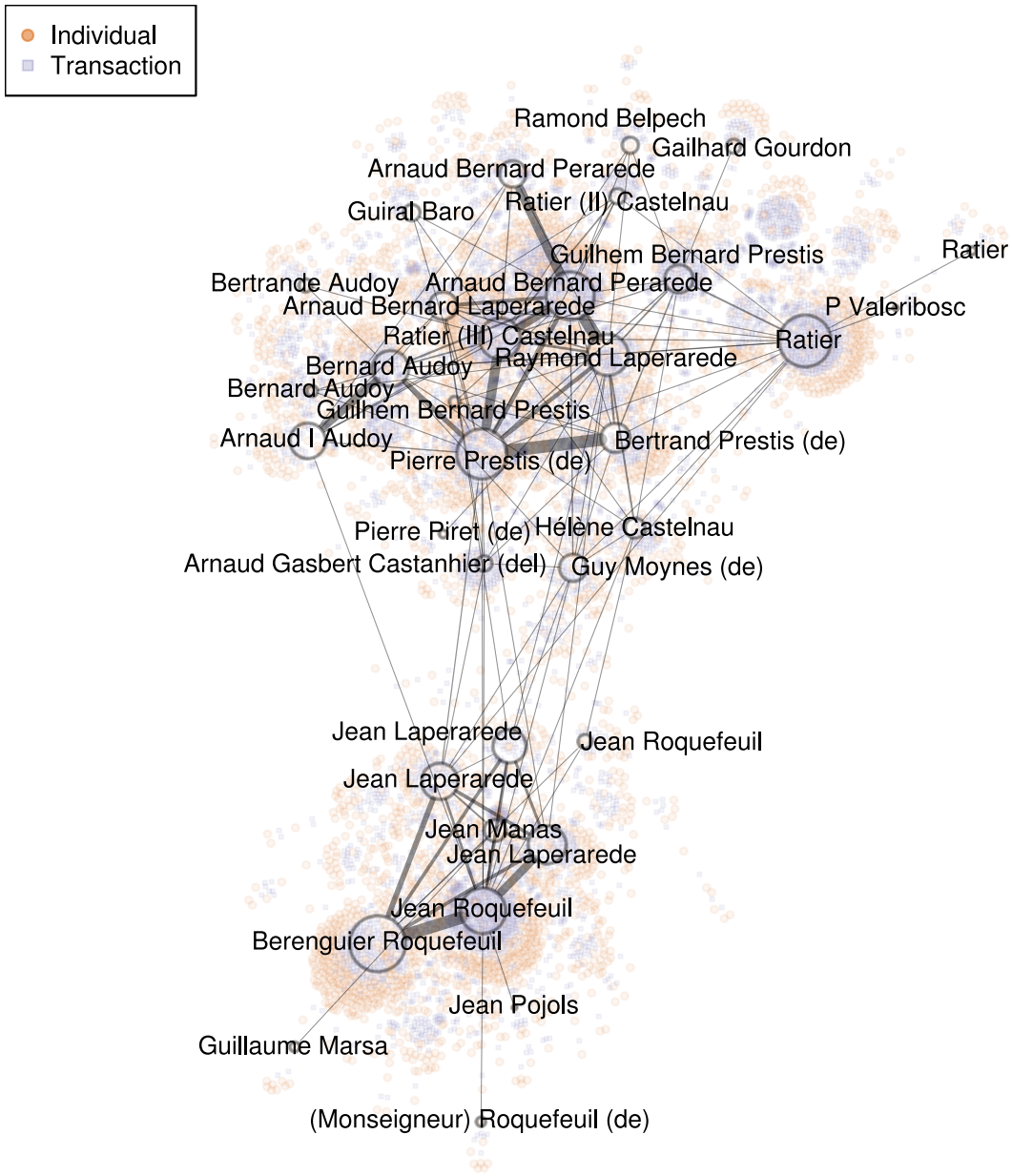
- représentation graphique complexe
- effet historique dominant

# Données médiévales



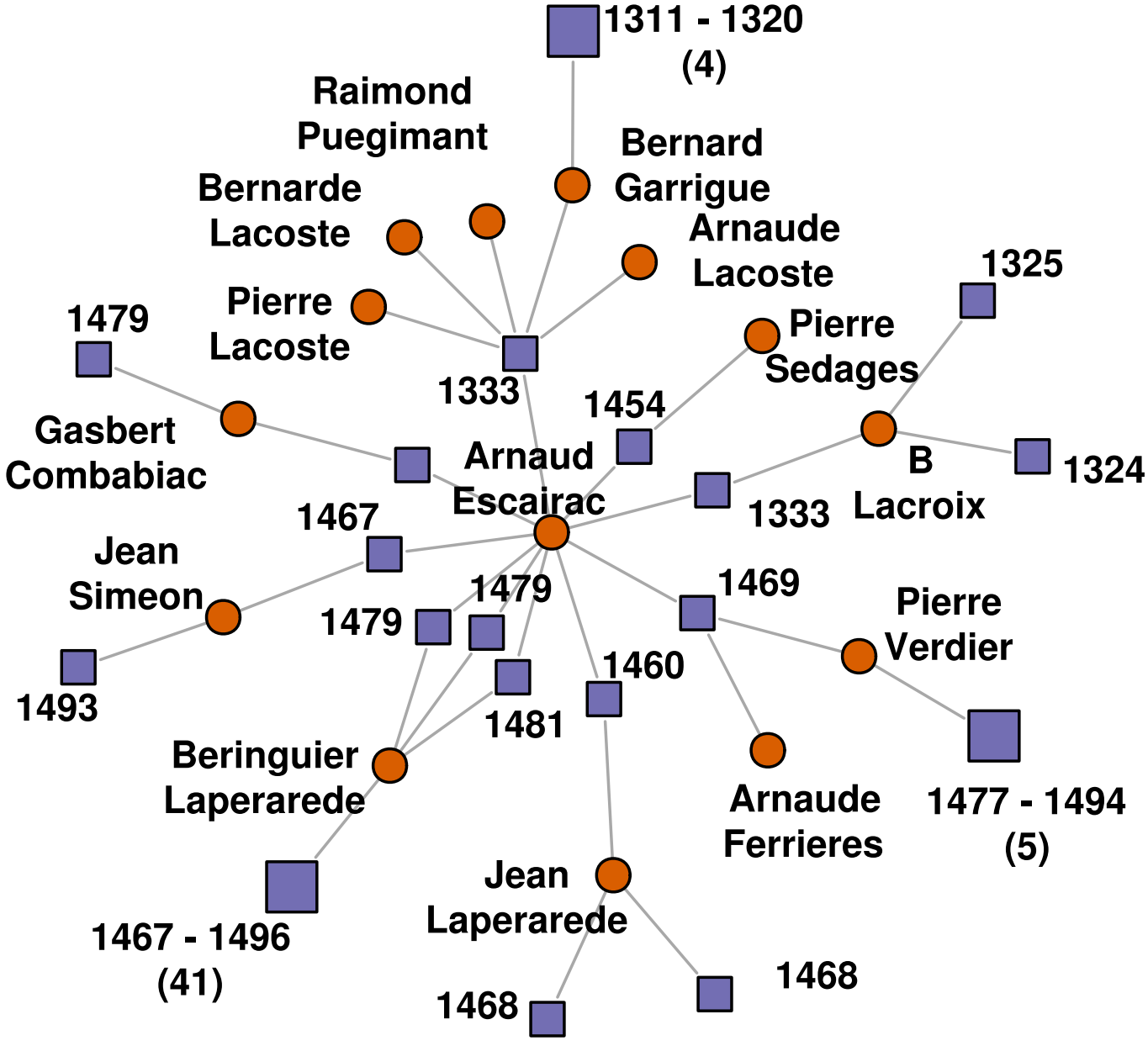

- représentation graphique complexe
- effet historique dominant
- classes + seigneur dominant

# Données médiévales



# Données médiévales

# Données médiévales



# Plan

# Graphe aléatoire et corrélation

## Limites

- graphes aléatoires simples : composante géante, percolation, etc.
- mais : transitivité non maîtrisée et corrélation impossible à prendre en compte

## Classification de sommets

- rupture au modèle d'indépendance
- découverte de classes effectives

## Deux questions cruciales

- Comment réconcilier modèles de graphes aléatoires et corrélation/transitivité ?
- Comment réconcilier modèles de graphes aléatoires et classification ?

# Modèle génératif

## Modèle mathématique

- matrice d'adjacence aléatoire $A$ ($N$ sommets)
- restriction aux graphes simples non dirigés (mais extensions possibles !)
- $\frac{N(N-1)}{2}$ variables aléatoires $A_{12}, A_{13}, \ldots, A_{(N-1)N}$ de Bernoulli
- $\mathbb{P}(A)$ ?

## Modèle de Erdös Rényi

- $\frac{N(N-1)}{2}$ variables indépendantes de même loi $\mathcal{B}(p)$

$$\mathbb{P}(A|p) = \prod_{i<j} p^{A_{ij}}(1-p)^{1-A_{ij}}$$

# Modèle génératif

## Deux ingrédients

1. paramètre pour la loi de $A_{ij}$
2. hypothèses d'indépendance : non corrélation $\Rightarrow$ indépendance entre les $A_{ij}$

## Graphe markovien

- Frank et Strauss, 1986
- $A_{ij}$ indépendant de $A_{kl}$ (conditionnellement au reste du réseau) si $i, j, k$ et $l$ sont distincts
- pour tout graphe markovien, $\mathbb{P}(A)$ est caractérisé par des décomptes de structures (arêtes, triangles et étoiles)

# Exponential Random Graph Model

## ERGM

- en pratique les graphes markovien ne sont pas très utiles (estimation par pseudo-vraisemblance, comportement peu adapté)
- modèle plus contraint (et plus général !), les ERGM (Wasserman and Pattison, 1996)

$$\mathbb{P}(A = a) = \exp\{\theta^T u(a) - \phi(\theta)\},$$

où $u(a)$ désigne un ensemble de statistiques sur $a$ (nombre de triangles, degrés, etc.)

## *Triad model*

1. $u_1(a) = \sum_{i<j} a_{ij}$ : nombre d'arêtes
2. $u_2(a) = \sum_{i<j} \sum_{k \neq i,j} a_{ik} a_{jk}$ : nombre de voisins communs
3. $u_3(a) = \sum_{i<j<k} a_{ij} a_{ik} a_{jk}$ : nombre de triangles

# Estimation

Snijders 2002

## MCMC

- Monte Carlo Markov Chain
- méthode fondamentale pour simuler des tirages de lois complexes
- ici, on change aléatoirement l'état d'une arête conditionnellement aux paramètres et au reste du réseau

## Algorithme de Robbins-Monro

- optimisation stochastique
- utilisé ici pour résoudre $E_\theta\{u(A)\} = u(a)$ (méthode des moments, maximum de vraisemblance)
- s'appuie sur une approche MCMC pour estimer les grandeurs utiles

# Interprétation

## Approche « économétrique »

- proche du principe de la régression logistique/linéaire
- les paramètres $\theta$ sont intéressants (significativité, etc.)
- prise en compte simple de grandeurs non graphiques

## Limites

- approche par agrégation seule
- problèmes de dégénérescence (graphes vides ou complets seulement)
- problèmes de multimodalité

# Stochastic Block Model

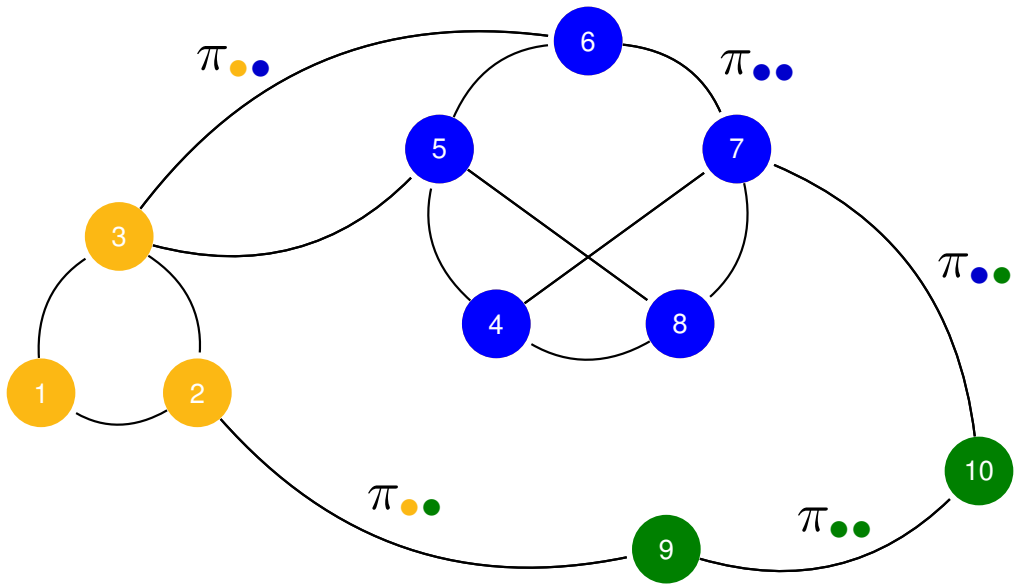
## Principe

- approche complètement différente des ERGM
- idée de base :
  - chaque sommet a un rôle caché
  - la connectivité s'explique uniquement par le rôle
- proche des principes sociologique d'équivalence

## Modèle de base

- $K$ classes, $N$ variables cachées $Z_i$ à valeurs dans $\{1, \ldots, K\}$
- indépendance conditionnelle $\mathbb{P}(A|Z) = \prod_{i<j} \mathbb{P}(A_{ij}|Z_i, Z_j)$
- influence des rôles seuls $\mathbb{P}(A_{ij} = 1|Z_i = k, Z_j = l) = \pi_{kl}$
- pas d'hypothèse sur les valeurs $\pi_{kl}$ : communautés, hub, etc.

# Illustration



# Estimation

## EM

- candidat à priori pour l'algorithme EM (variables cachées)
- mais dépendance croisée et calcul de $\mathbb{P}(A)$ impossible explicitement
- diverses solutions :
  - approximation variationnelle
  - approche Bayésienne variationnelle
  - approche Bayésienne avec MCMC
  - etc.

## Produits de l'estimation

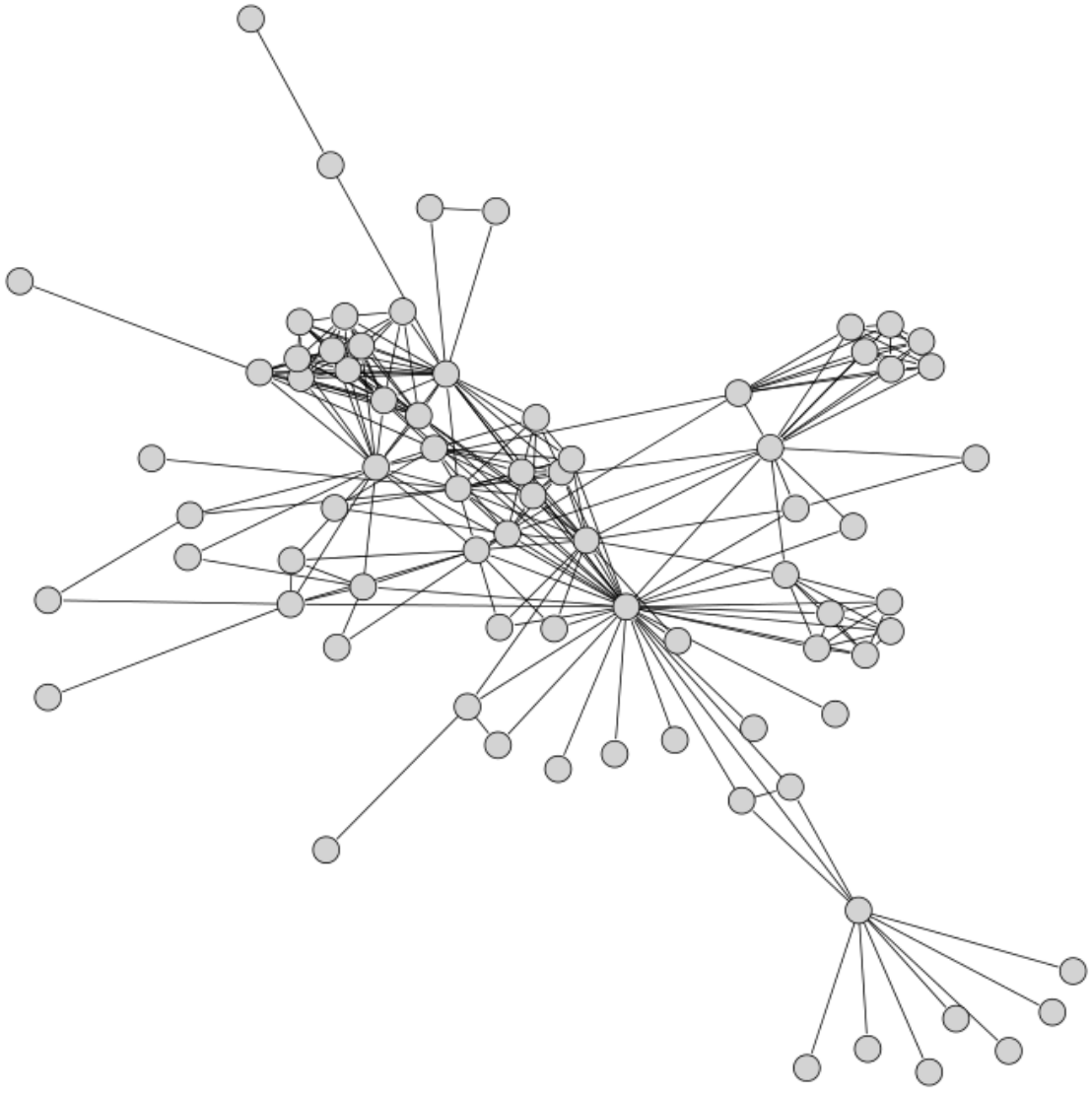
- approximation de $\mathbb{P}(Z_i|A)$ : classification (« floue ») des sommets
- $\pi_{kl}$ : connections entre classes

# SBM et Modularité

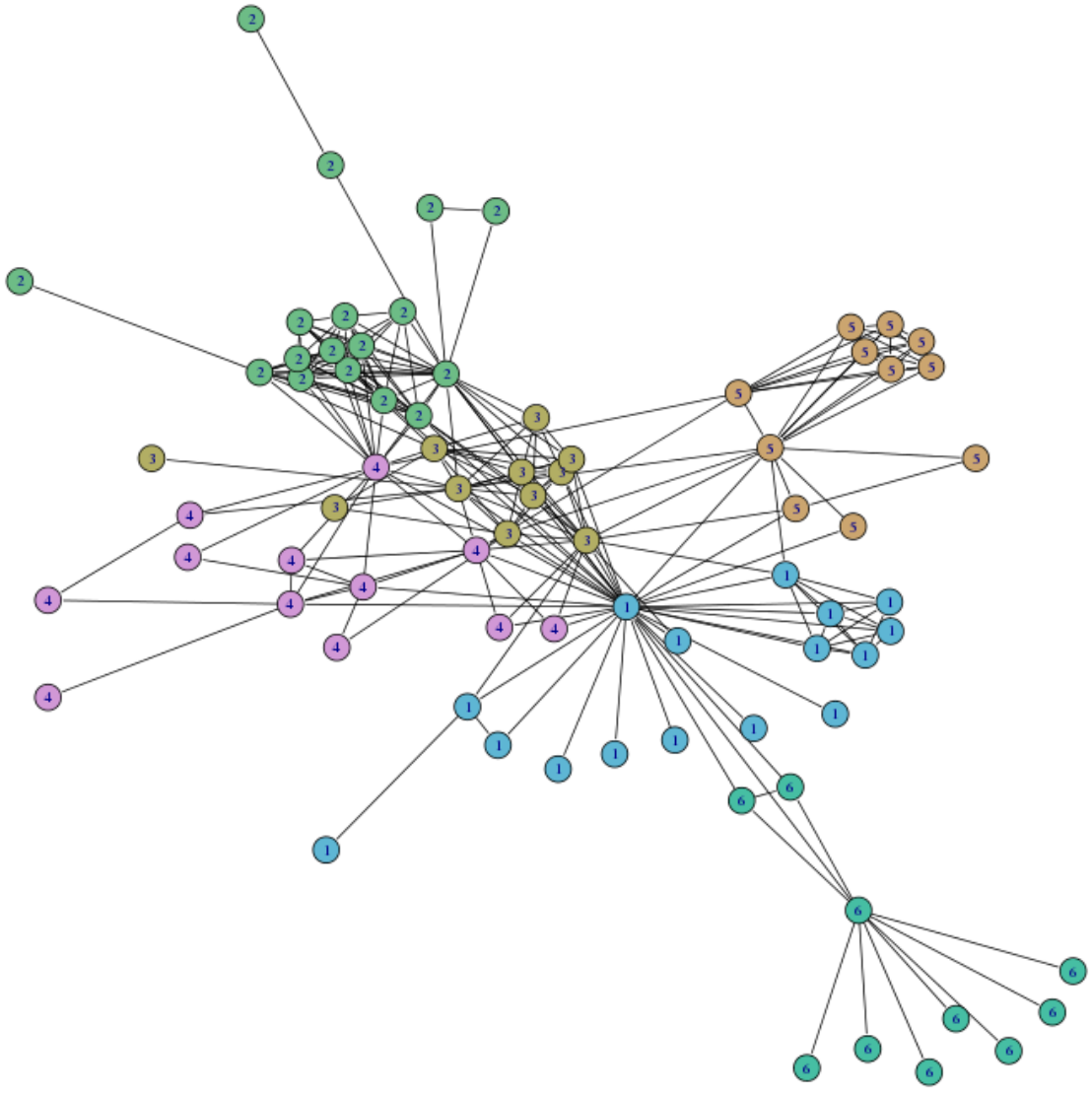
## Comparaison

- Modularité :
  - \+ algorithmes rapides
  - \+ mélange de degrés possible dans une classe
  - \- s'intéresse seulement à la diagonale
  - \- homogénéité moyenne, même sur la diagonale
- SBM :
  - \+ modèle très riche
  - \+ nombreuses extensions naturelles (valuation, orientation, aspects temporels)
  - +/- classes homogènes au sens des degrés
  - \- algorithmes lents

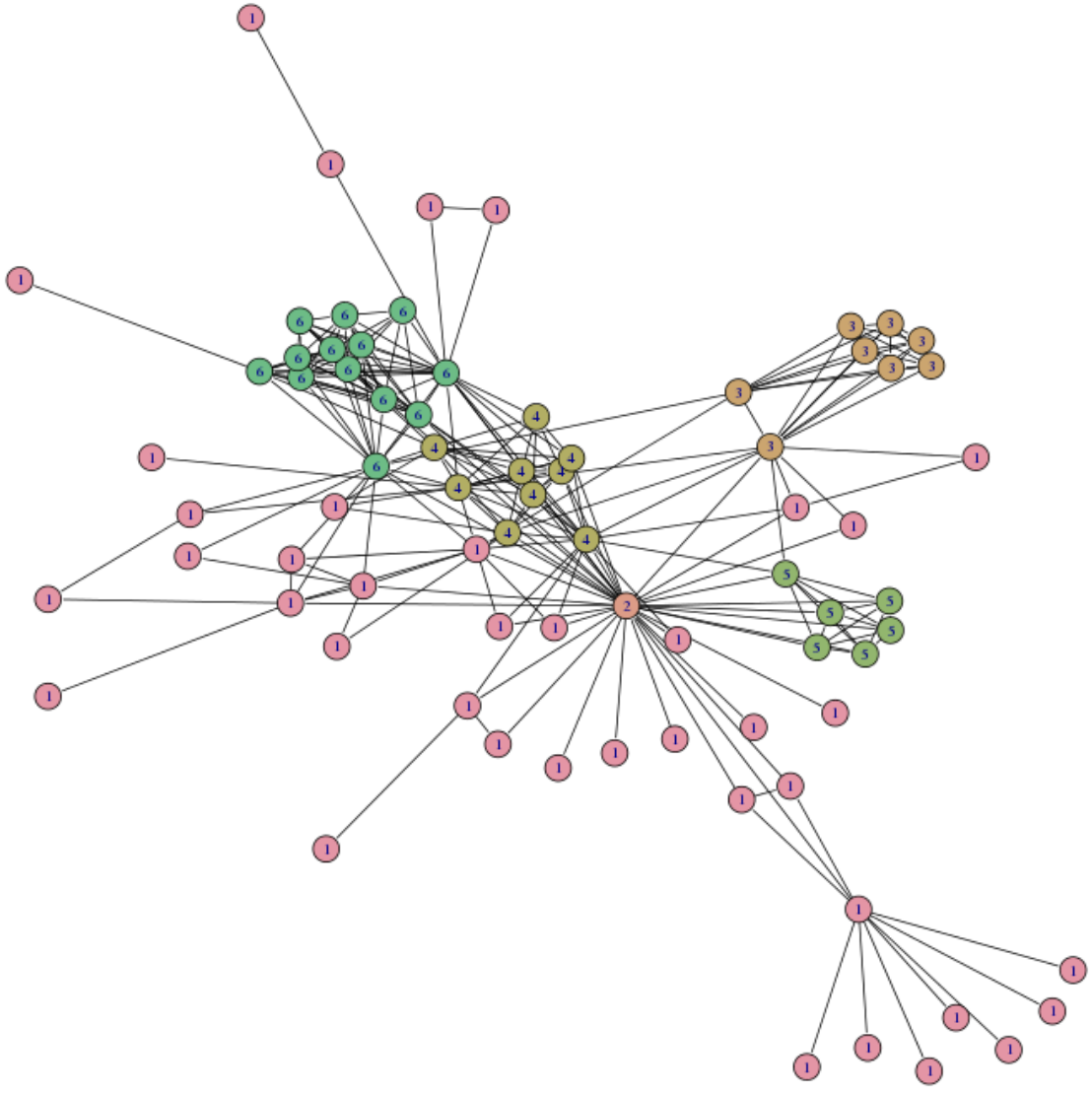
# Exemple de traitements



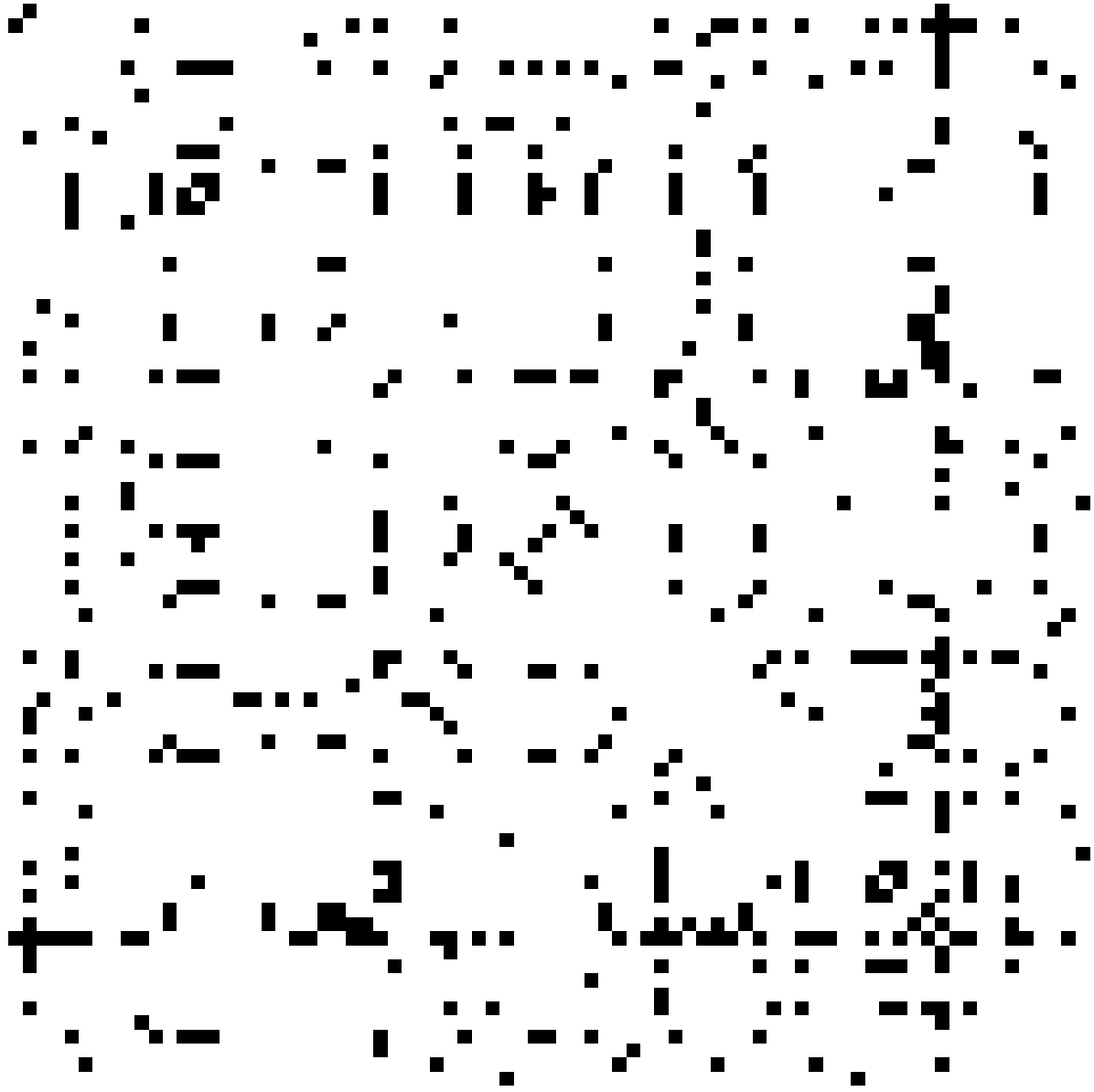
# Maximum de modularité



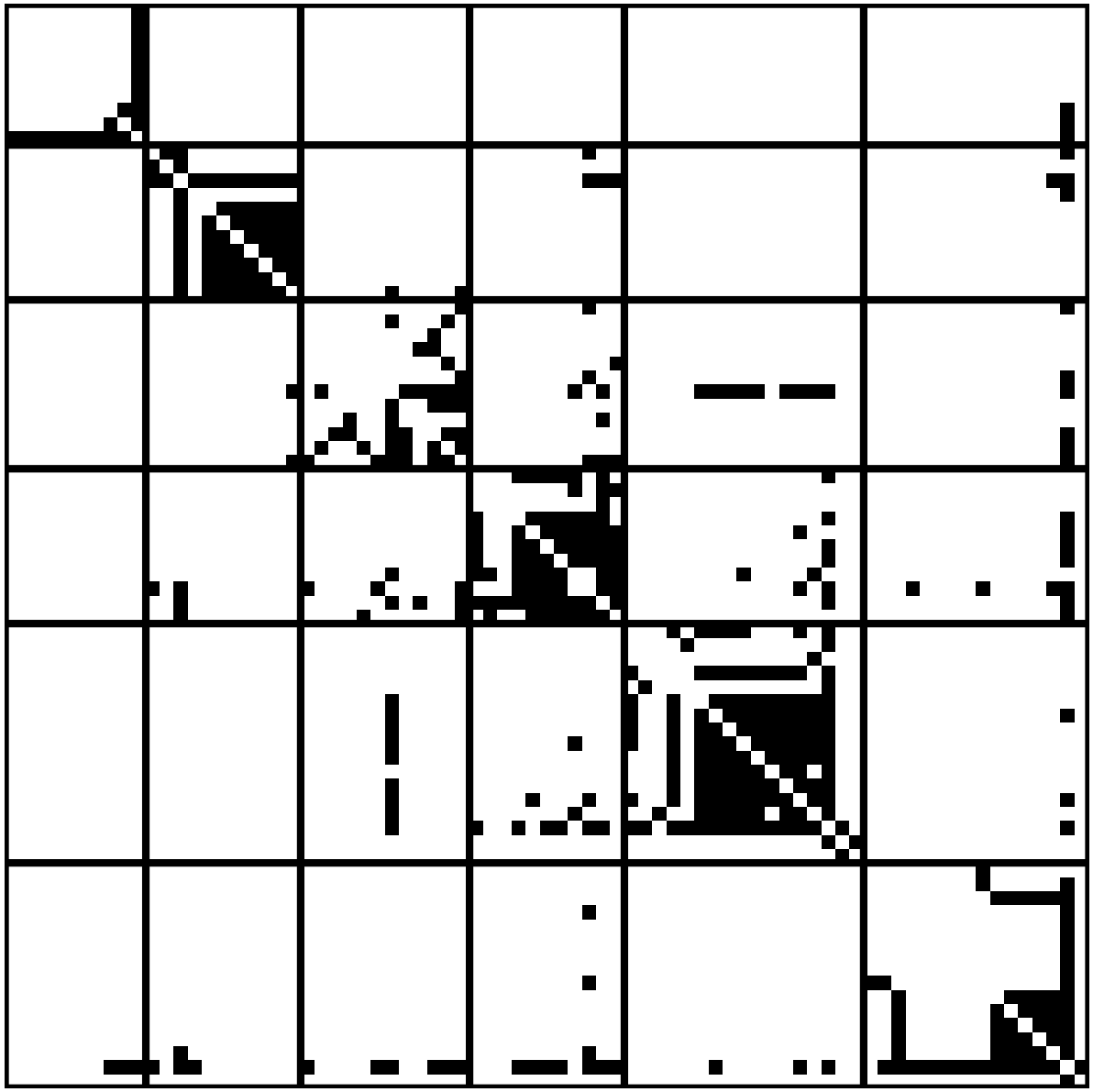
# Stochastic block model



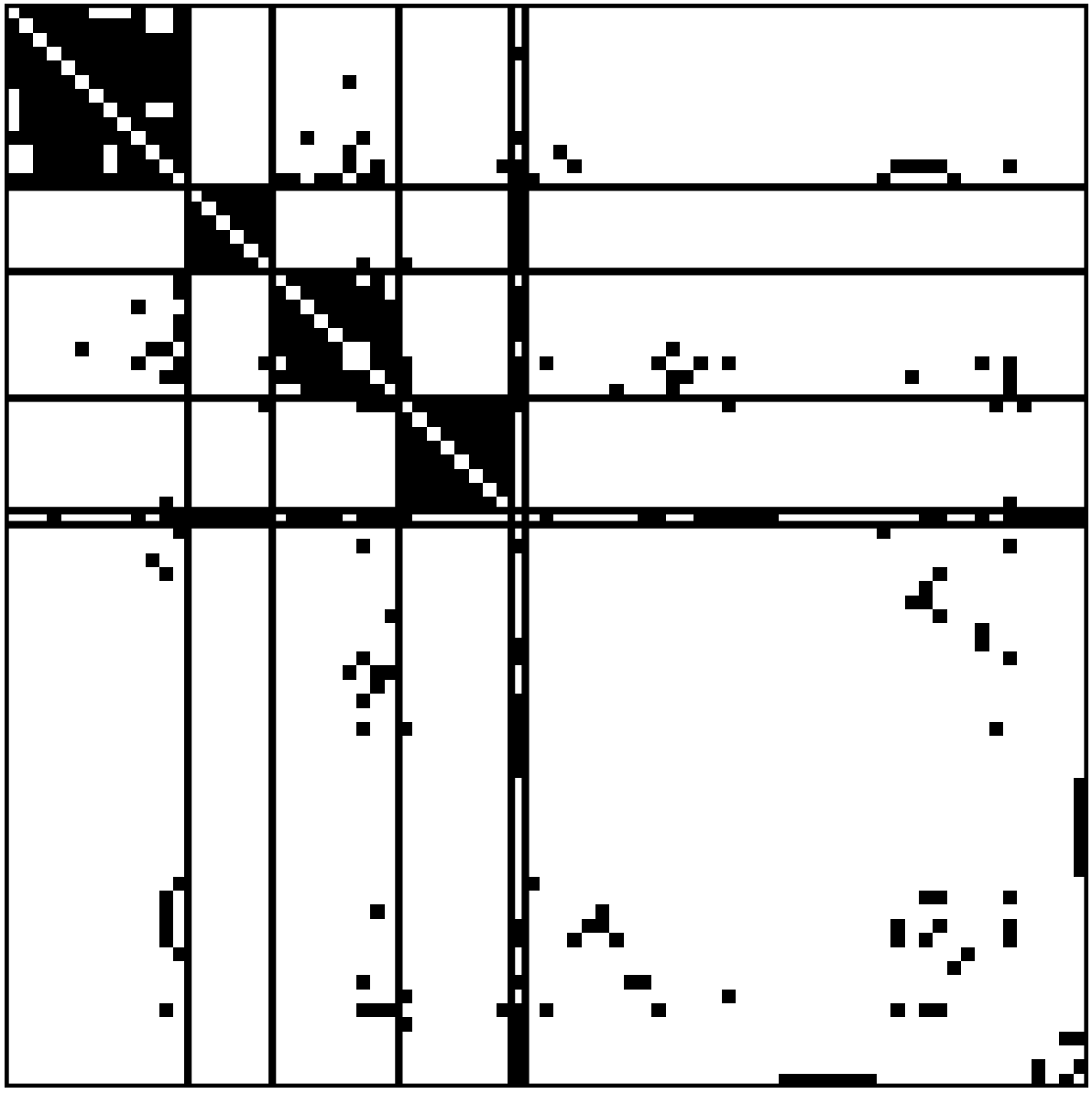
# Matrice d'adjacence



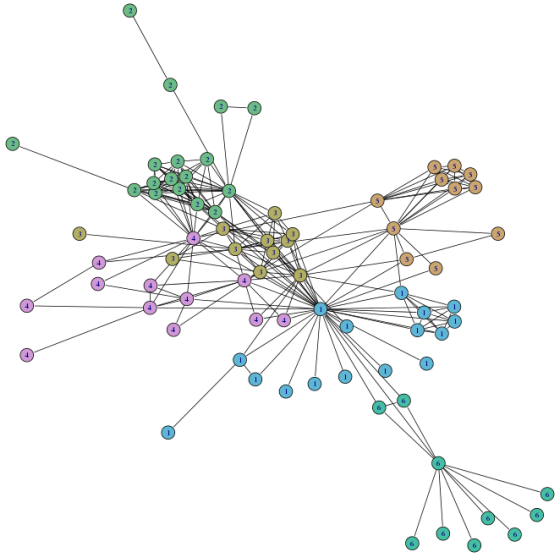
# Maximum de modularité

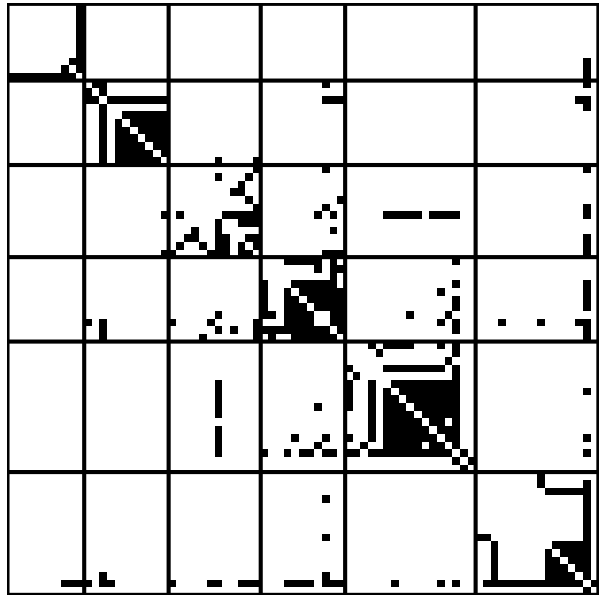


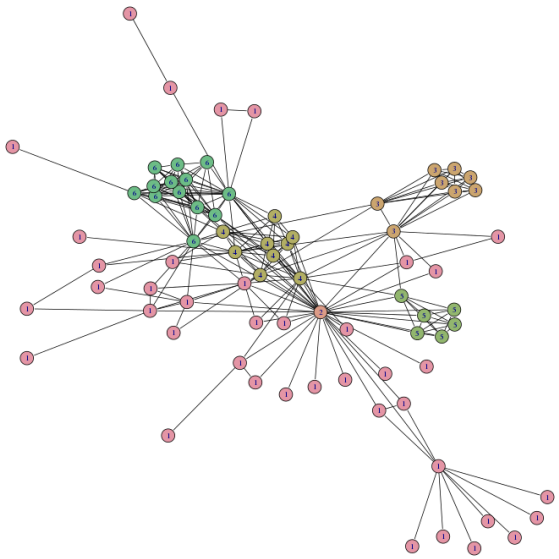
# Stochastic block-model

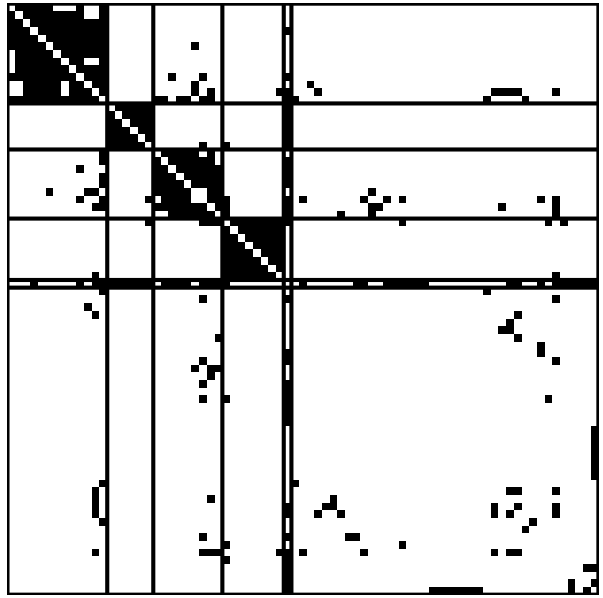


# Comparaison



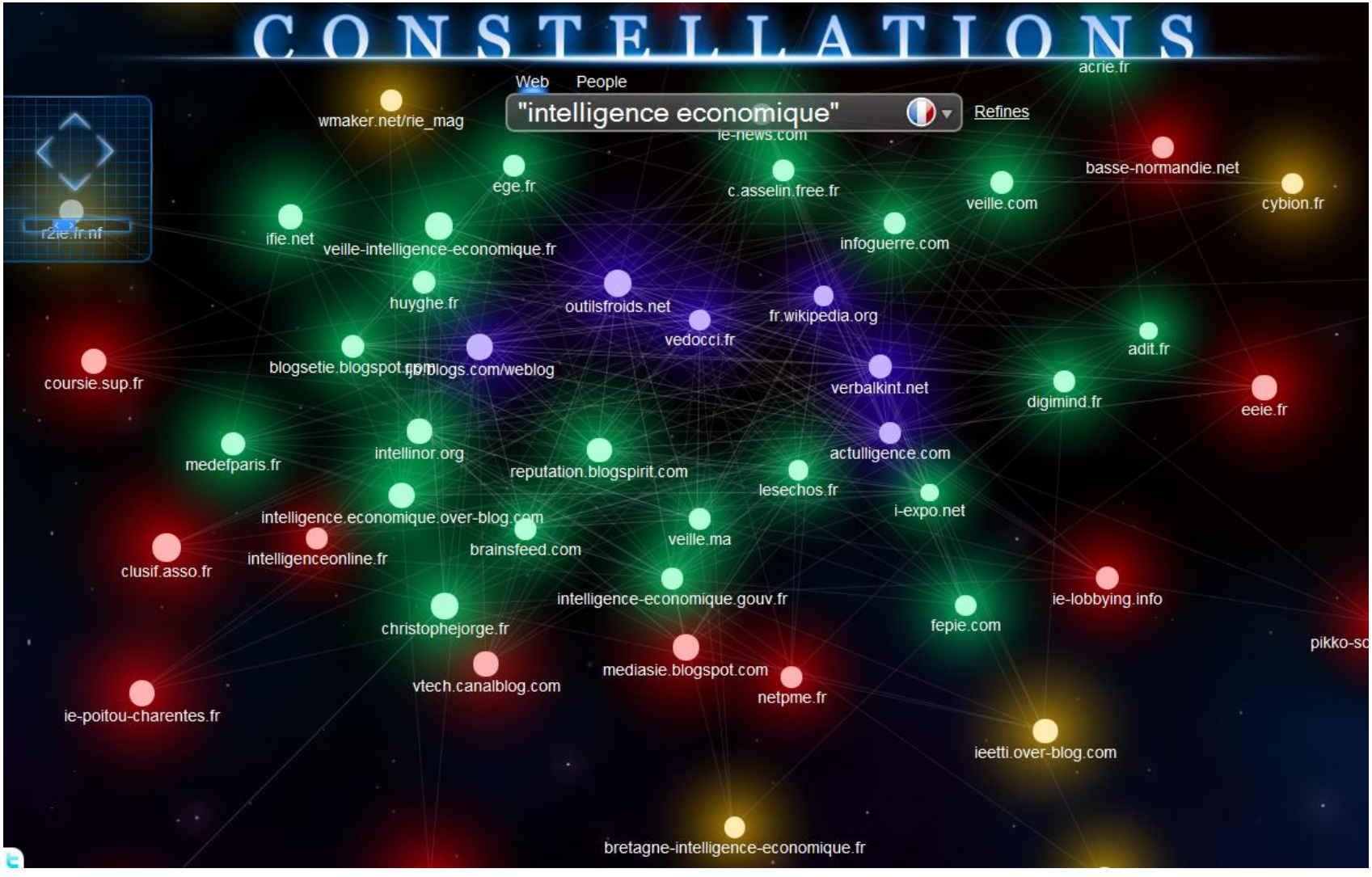






# Exemple d'application

## Constellations




Source : `http://labs.exalead.com/project/constellations`

# Bilan

## Modèles à rôles cachés

- cas simple du SBM
- extensions :
  - rôles continus (position)
  - rôles multiples
  - rôles stochastiques
  - etc.

## En pratique

- résultats utiles en pratique
- cas d'utilisation assez opposés à ceux de la modularité
- algorithmes plutôt lents (adaptés aux graphes denses et petits)
- solutions approximatives si besoin

# Plan

# En résumé

## Graphes

- objets riches et complexes
- modèles pour de très nombreuses situations concrètes

## Difficultés

- quel graphe pour des données ?
- visualisation peu informative (et parfois trompeuse)
- dépendances entre les objets
- problèmes associés NP complets

# Outils

## Caractérisations locales/globales

- équivalent des statistiques de base
- caractérisation grossière mais utiles

## Graphes aléatoires simples

- phénomènes sur un graphe
- échantillon ou informations locales

## Sous graphes fréquents

- comportements répétés
- sous-structures

# Outils

## Classification des sommets

- le graphe comme une structure de proximité
- communautés
- simplification et exploration des graphes peu denses

## Modèles avancés de graphes aléatoires

- introduire des dépendances entre sommets, expliquer la connectivité par des schémas et des covariables
- classification au delà de la notion de proximité

# Sujets non traités

Non exhaustif...

## Graphes multiples

- chaque objet d'étude est un graphe entier
- problèmes classiques de la *data science* :
  - classification
  - prédiction
- outils fondamentaux : distances et noyaux entre graphes

## Prévisions sur un graphe

- étiquettes pour les sommets non étiquetés
- arêtes entre sommets
- outils :
  - propagation d'étiquettes
  - matrice de covariance creuse
  - etc.

# Le futur ?

## Temporalité

- graphes évolutifs
- temps discret ou temps continu

## Très grands graphes

- algorithmes approximatifs
- effets du volume

## Multi, hyper, ?

- plusieurs graphes sur un même ensemble de sommets (contradiction, renforcement, etc.)
- données relationnelles
- hypergraphes comme modèle réaliste des interactions

# Licence

# Crédits photos et illustrations I

- Illustrations wikipedia :
  - https://en.wikipedia.org/wiki/File:6n-graf.svg
  - https://commons.wikimedia.org/wiki/File:Hypergraph.svg
  - https://commons.wikimedia.org/wiki/File:Graph_betweenness.svg
- Procès dans le mobile :
  - illustration *An Explosion of Mobile Patent Lawsuits* de Nick Bilton pour le New York Times :

    https://bits.blogs.nytimes.com/2010/03/04/an-explosion-of-mobile-patent-lawsuits/
  - illustration *Who's suing who in the mobile business* de Josh Halliday and Charles Arthur pour le Guardian : https://www.theguardian.com/technology/2010/oct/04/microsoft-motorola-android-patent-lawsuit
  - illustration *Lawsuits in the mobile business* de George Kokkinidis pour Design Language News : http://news.designlanguage.com/post/1252039209
  - illustration *Who's Suing Whom ?* de David McCandless et James Key pour Information is Beautiful :

    http://www.informationisbeautiful.net/2010/whos-suing-whom-in-the-telecoms-trade/
  - illustration *Mobile Patent Suits* de Reuters :

    https://blogs.thomsonreuters.com/answerson/mobile-patent-suits-graphic-day/
- illustration de démonstration de Gephi : https://gephi.org/images/screenshots/layout2.png

# Crédits photos et illustrations II

- illustrations issue de *The Structure and Function of Complex Networks, M. E. J. Newman, SIAM Review 2003 45 :2, 167-256* : https://arxiv.org/abs/cond-mat/0303516
  - construction du modèle petit monde : figure 11
  - transition de phase du modèle de Erdös-Rényi : figure 10
  - attaque ciblée : figure 14
- illustration sur la méthode AGM : figure 9.4 de *Data Mining : Concepts and Techniques* de Jiawei Han et Micheline Kamber
- illustration sur la simplification d'un graphe : figure 10 de *Finding and evaluating community structure in networks, M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113, 2004* : https://arxiv.org/abs/cond-mat/0308217
- limite de la modularité : figure 1 de *Narrow scope for resolution-limit-free community detection, V. A. Traag, P. Van Dooren, and Y. Nesterov, Phys. Rev. E 84, 016114, 2011* : http://arxiv.org/abs/1104.3083
- capture d'écran du logiciel constellations : http://labs.exalead.com/project/constellations