COURS SYRRES RÉSEAUX SOCIAUX

INTRODUCTION

Jean-Loup Guillaume

Le cours

- Enseignant :
  - Jean-Loup Guillaume – équipe Complex Network
- Page du cours :
  - http://jlguillaume.free.fr/www/teaching-syrres.php
- Évaluation :
  - Mini-projet se basant sur les méthodes de similarité.

Expérience de Milgram (1967)

- Objectif :
  - Faire transiter une lettre depuis le Nebraska à un agent de change de Boston.
  - Une personne initie la chaine.
  - Transitions de la main à la main par des personnes que l'on connait.

Expérience de Milgram (1967)

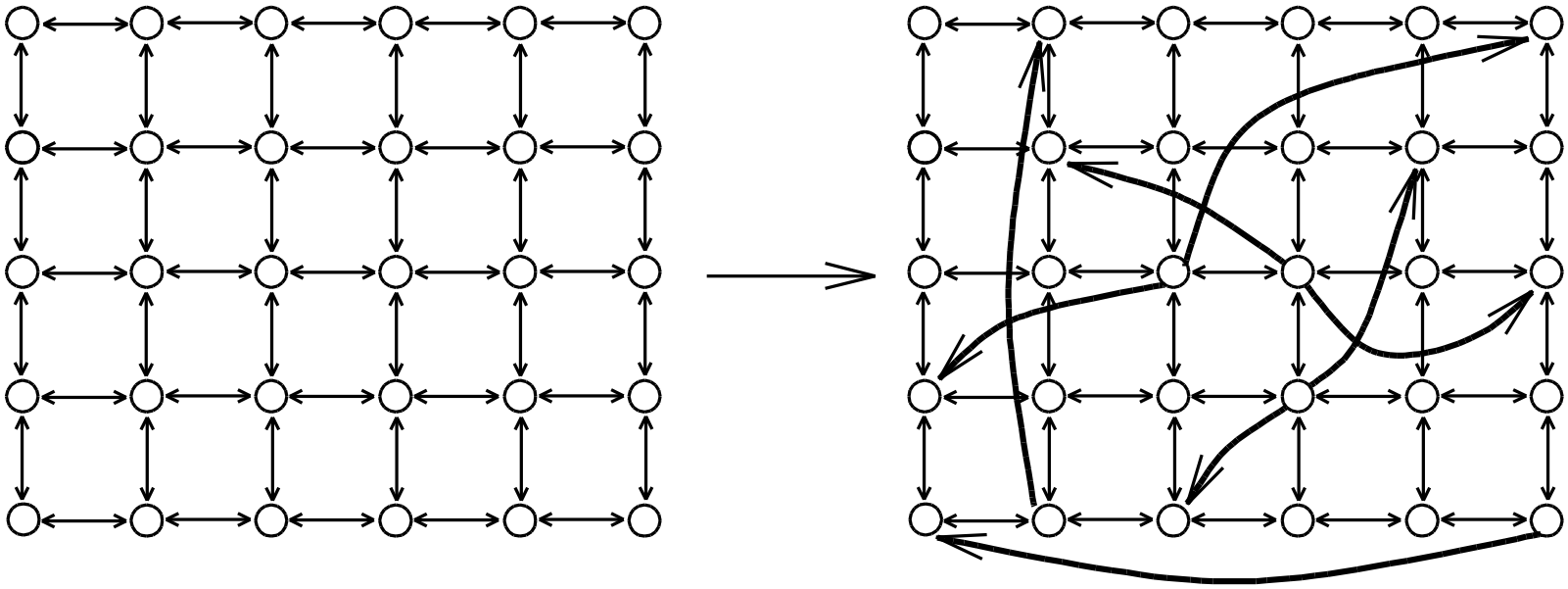
- Résultats :
  - 44 lettres arrivent sur 160.
  - Chemins avec 5 intermédiaires en moyenne.

- Remarques :
  - Chemin interrompu ≠ Il n'existe pas de chemin.
  - Chemin de longueur x ≠ Il n'existe pas de chemin de longueur <x

- Conclusions :
  - Il existe des chemins courts.
  - Les intermédiaires arrivent à les trouver sans connaissance globale du réseau.

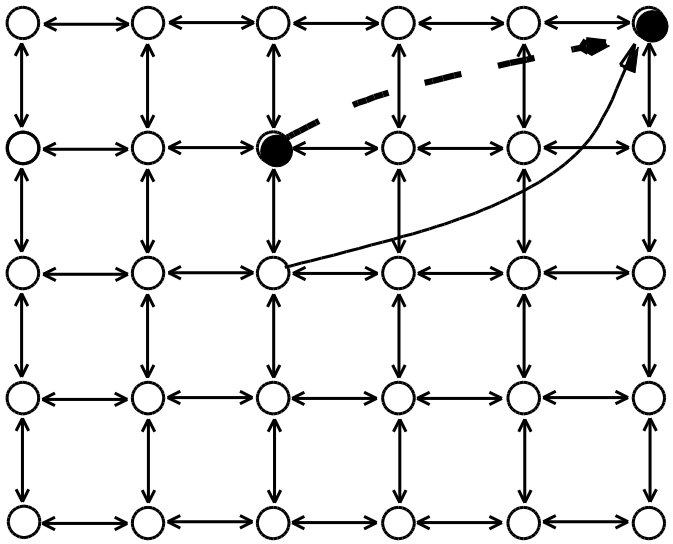
Modélisation

- Objectif : formaliser l'expérience de Milgram
  - Initialement une grille (amis proches).
  - On ajoute q voisins quelconques à chaque sommet (amis lointains).



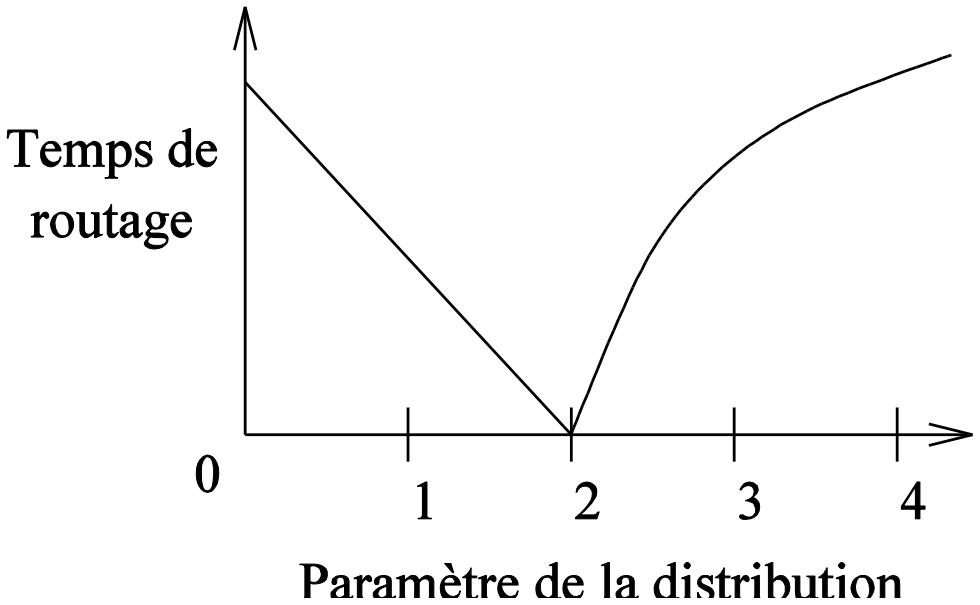
Modélisation

- Un sommet connait :
  - Sa position, celle de ses voisins, celle de la destination.
  - Il envoie le message à son voisin le plus proche de la destination.



Modélisation

- Un seul lien supplémentaire pour chaque sommet u.
- La destination choisie avec une probabilité dépendant de sa distance à u.
- Dans la majorité des cas, pas de chemins courts.



Nombre d'Erdös

- Graphe de collaboration :
  - Deux scientifiques sont connectés s'ils ont co-écrit un article.
  - Chaque scientifique à un nombre d'Erdös :
    - 0 = Erdös
    - 1 = collaborateurs d'Erdös
    - 2 = collaborateurs de collaborateurs d'Erdös
  - http://www.oakland.edu/enp/

Modélisation

- Récupération de la liste des co-auteurs de tous les articles scientifiques
- Ensuite il ne reste qu'à faire des calculs de plus courts chemins d'Erdös vers les autres chercheurs.

Kevin Bacon Game

- Graphe des acteurs :
  - Deux acteurs sont reliés s'ils ont joué dans un même film.
  - Distance entre acteurs ?
    - http://oracleofbacon.org/
    - Distance entre Tom Cruise et Clint Eastwood ?
    - Distance entre Mickey Mouse et Omar Sy ?

Modélisation

- Graphe des acteurs simple à construire :
  - http://www.imdb.com/interfaces

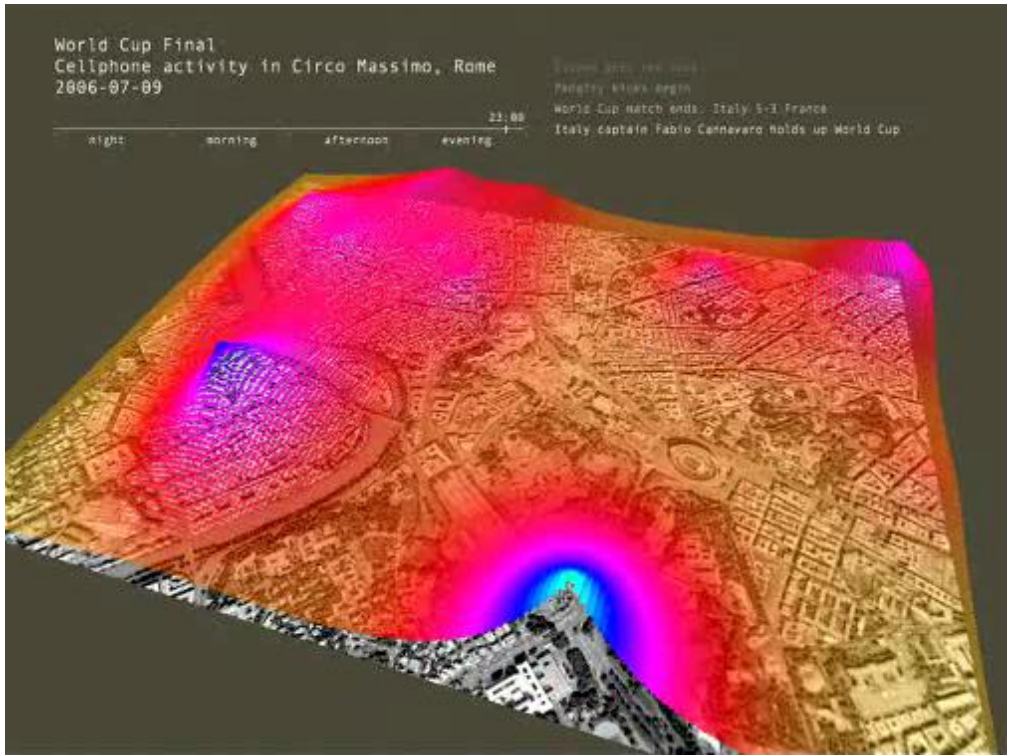
- Ensuite il ne reste qu'à faire des calculs de plus courts chemins.

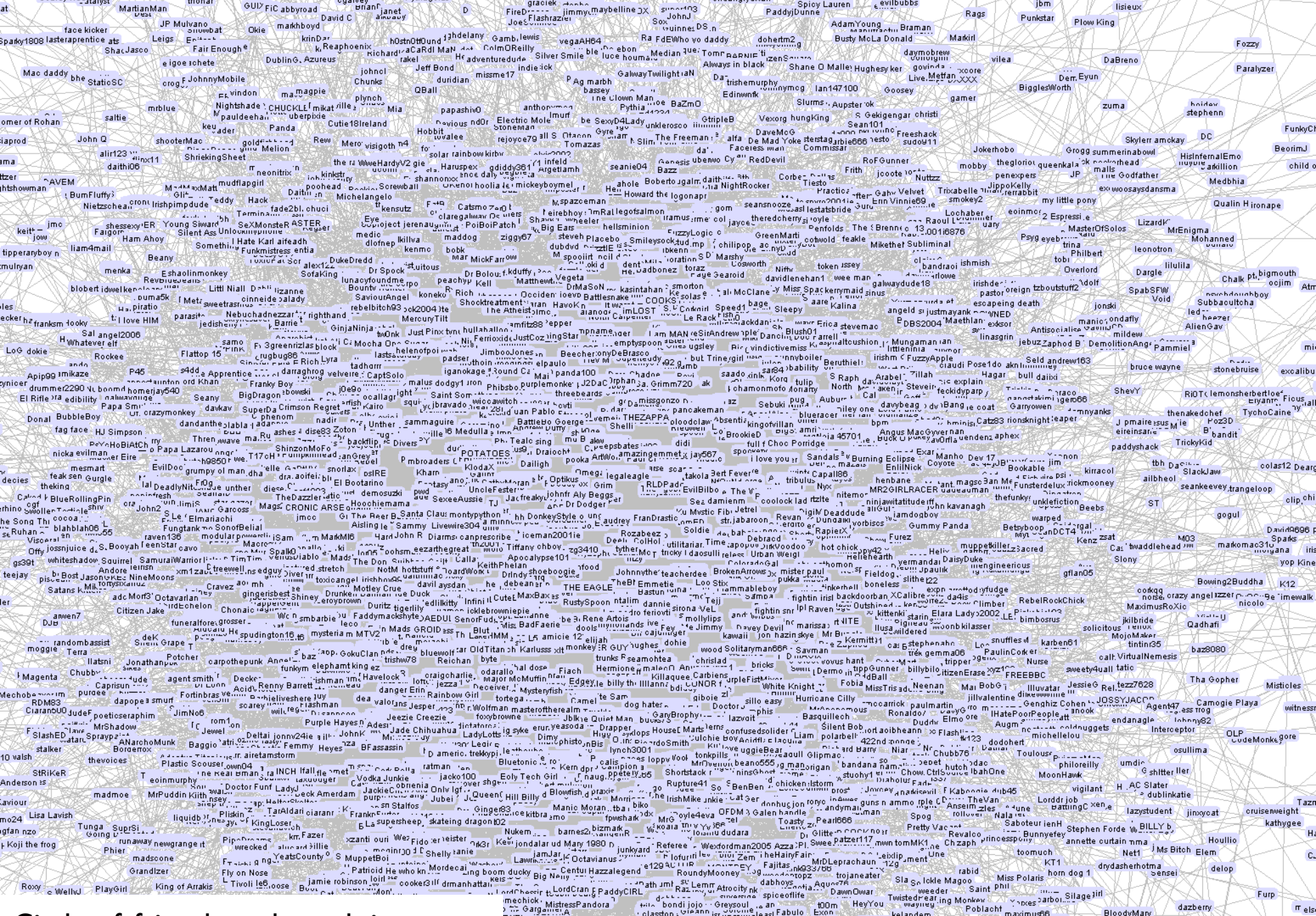
Fichiers P2P

- Propagation d'un fichier d'utilisateurs en utilisateurs
- Mesure ?

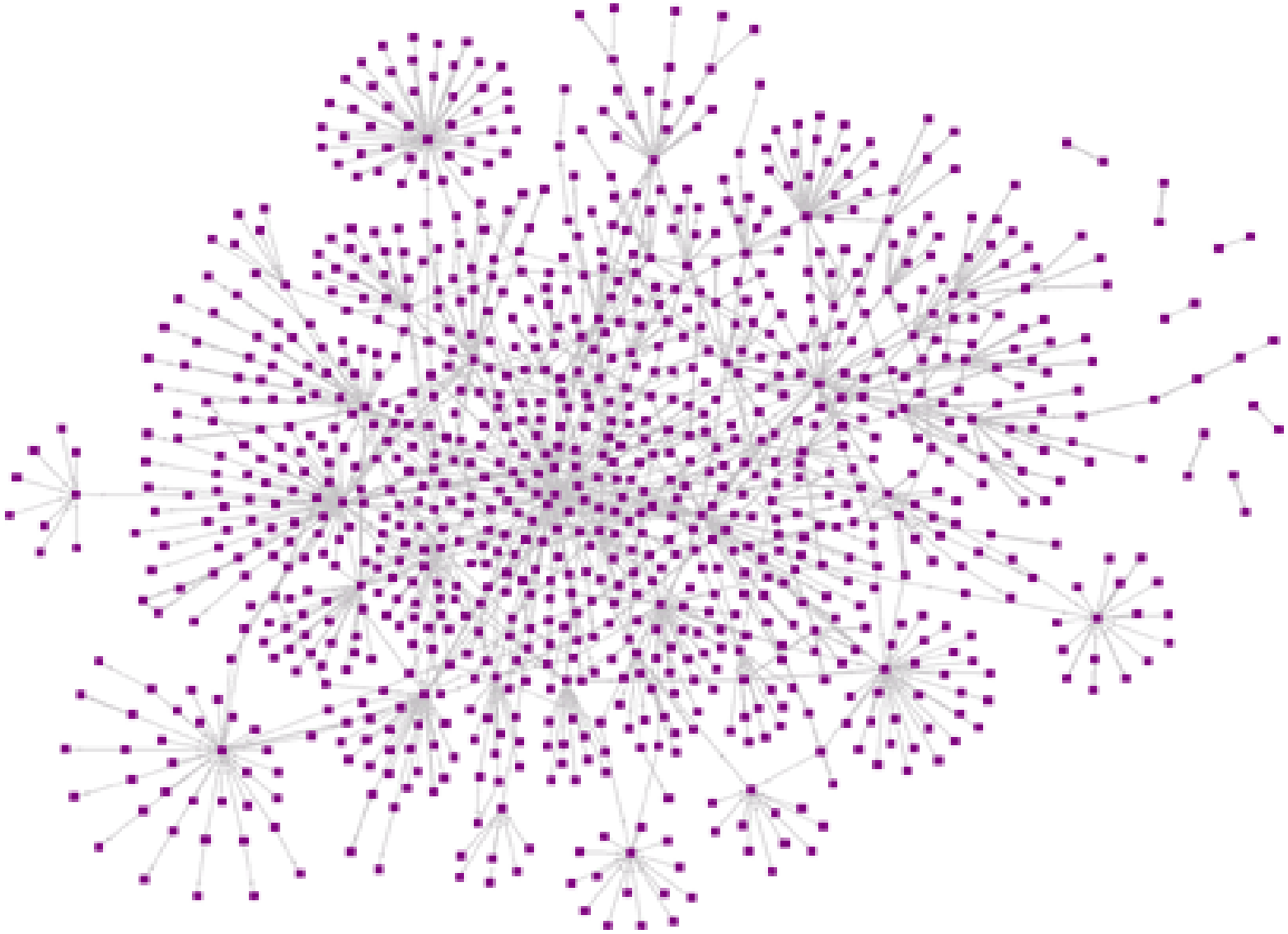
Téléphone

- Suivi de communications :
  - Date, heure, durée, type, correspondant
  - Type d'appelant, mobilité, …
  - http://senseable.mit.edu/





Circle of friends on boards.ie



Relationship between key recovery agencies after Katrina

FRIDAY
MAY 12, 2006

The Philadelphia Inquirer

50¢
75 cents in some locations outside the metropolitan area

City Edition

WWW.PHILLY.COM

177th Year, No. 346

A Wide Surveillance Net

A firestorm follows a report that as part of the terror fight, three firms turned over the phone-call records of millions of Americans.

Data Mining: How It Works

Each colored square on this diagram represents one of the 9/11 hijackers. With **social-networking analysis** software, a researcher used publicly available data to show the connections among each hijacker and his associates (in gray). Heavier lines indicate stronger connections.

The hijackers

Pilots have a thick outline:

- American Airlines Flight 77
- American Airlines Flight 11
- United Airlines Flight 93
- United Airlines Flight 11

SOURCE: Diagram © Valdis Krebs, http://orgnet.com/hijackers.html. For more information on how the diagram was created, go to **http://go.philly.com/hijackers**

Uproar: Issues of privacy, national security weighed.

By Ron Hutcheson and James Kuhnhenn
INQUIRER WASHINGTON BUREAU

WASHINGTON — Revelations that the government collected the phone-call records of millions of Americans drew calls yesterday for a congressional investigation and fueled questions about President Bush's choice for CIA director.

At least three companies — AT&T Corp., Verizon Communications Inc., and BellSouth Corp. — turned over call records for tens of millions of their customers to the National Security Agency as part of the administration's antiterrorism effort, USA Today reported.

Officer Gary Skerski will be laid to rest.

Phalanx of police to attend funeral

By Julie Stoiber
INQUIRER STAFF WRITER

There are few occasions more somber, more dreaded, than the one that will draw hundreds of men and women in uniform to the city today.

They will come from around the country to say goodbye to a fellow police officer who was cut down in the line of duty Monday night. Gary Skerski, a 46-year-old father of two and a popular community-relations officer, will be laid to rest with all the ceremony the Philadelphia Police Department can muster.

Cardinal Justin Rigali will conduct his funeral Mass at the ornate Port Richmond church where Skerski's family worships.

Portions of Interstate 95 will be closed this afternoon as an honor guard of police motorcycles escorts the officer to his final resting place at Resurrection Cemetery in

Beaucoup d'autres réseaux

- informatique : internet, web, pair-à-pair, usages, …
- sciences sociales : collaboration, amitié, contacts sexuels,
- échanges, économie, …
- biologie : cerveau, gènes, protéines, écosystèmes, …
- linguistique : synonymie, co-occurrence, …
- transport : routier, aérien, électrique, …
- Etc.

Contextes différents

Propriétés et problématiques communes

Objectifs

« Comprendre le comportement d'entités qui interagissent par des lois gouvernant le système. »

- On cherche à comprendre :
  - La structure de ces graphes.
  - Leur évolution.
  - Les phénomènes agissant sur ces réseaux.

Quelques applications

- Informatique :
  - Réseaux : routage, protocoles, sécurité
  - P2P : conception de systèmes, déviances
  - Web : indexation, moteurs de recherche
  - Dessin de graphes, etc.
- Sociologie :
  - Diffusion d'innovations, rumeurs
  - Identification de communautés
- Epidémiologie :
  - Diffusion de virus, vaccination

Méthodologie

- Utilisation d'outils formels
  - Théorie des graphes
  - Analyse statistique
  - Modélisation probabiliste
- Études expérimentales
  - Simulation
  - Utilisation de données réelles
- Étudier des applications
  - Comprendre en profondeur certains réseaux
  - Extraction de concepts généraux

Dans ce cours

- Présentation plus en profondeur du domaine et des problématiques :
  - Métrologie
  - Analyse
  - Modélisation
  - Algorithmique
- Détection de communautés
- Réputation, prédiction, innovations et leaders

Projet

- Construire des graphes à partir des données JJ :
  - Construction de graphes de similarité.
  - Tester un algorithme de détection de communauté.
  - Faire de la prédiction de liens.
  - Comparer tout ça, essayer de comprendre des choses et essayer de l'expliquer.

COURS SYRRES
RÉSEAUX SOCIAUX

AXES DE RECHERCHE

Jean-Loup Guillaume

Axes de recherche

- Métrologie :
  - Comment mesurer les réseaux réels ?
- Analyse :
  - A quoi ressemblent-ils ?
- Modélisation :
  - Peut-on créer des réseaux artificiels similaires ?
- Algorithmique :
  - Comment calculer des choses sur ces grands graphes ?

ANALYSE

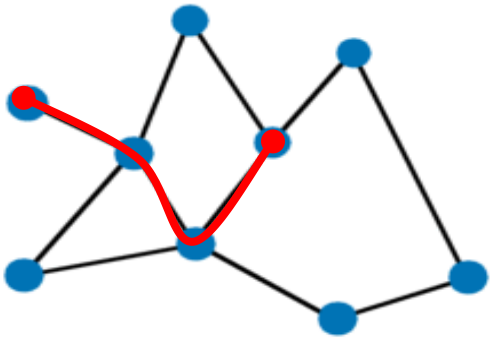
Analyse

- Objectifs de l'analyse (statistique) :
  - Description (statistique)
  - Obtenir de l'information pertinente
  - Interprétation des résultats obtenus

- Comment ?
  - Propriétés connues
  - Définition de propriétés (statistiques) pertinentes
  - Corrélations entre ces propriétés
  - Comparaison avec des graphes aléatoires
  - Observation de la croissance des graphes, …

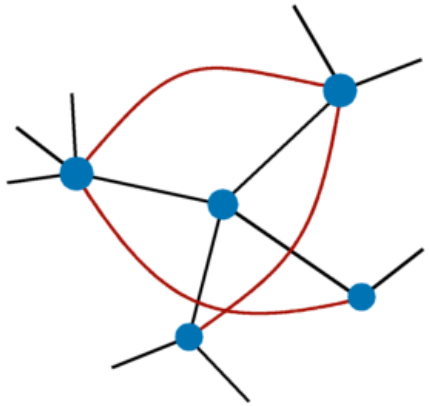
Propriétés classiques

- Distance moyenne :
  - À quelle distance sont les sommets les uns des autres ?



Propriétés classiques

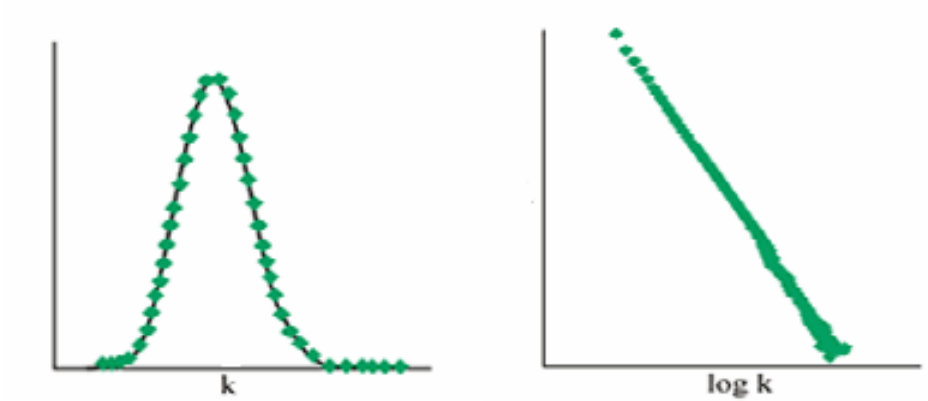
- Clustering :
  - Les amis de mes amis… / densité locale
  - A comparer à la densité globale



$$c(u) = \begin{cases} \frac{|\{(x,y) \in E, x,y \in N(u)\}|}{\binom{d(u)}{2}} & \text{si } d(u) \geq 2 \\ 0 & \text{sinon,} \end{cases}$$

Propriétés classiques

- Distribution des degrés (nombre de voisins) :
  - Taille ou salaire des individus ?



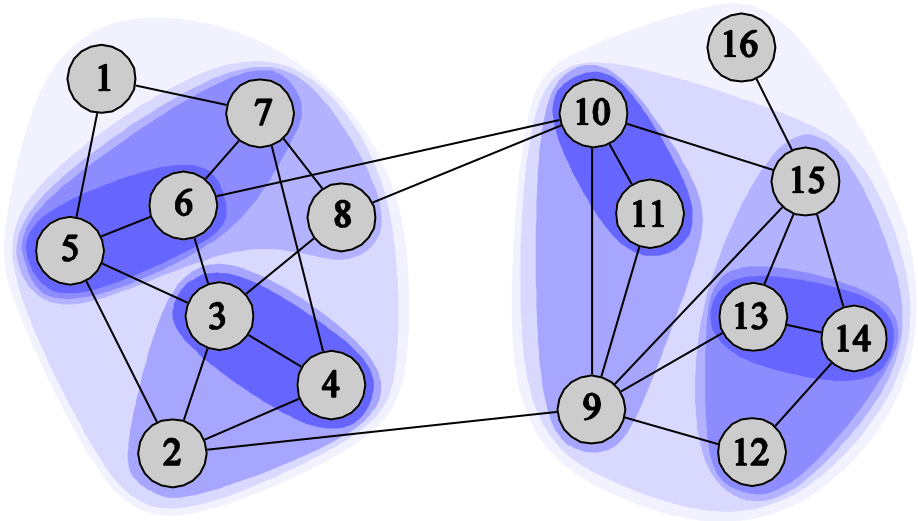
$$P_d \sim e^{-\lambda} \cdot \frac{\lambda^d}{d!} \qquad P_d \sim d^{-\alpha}$$

Propriétés classiques

- Composantes connexes :
  - Ensemble maximal de sommet tel qu'il existe un chemin entre toute paire de sommets de l'ensemble
  - Graphe connexe = une seule composante connexe

Propriétés classiques

- Communautés = sous-groupes :
  - Denses : beaucoup de liens dans les groupes
  - Peu connectés les uns aux autres



Propriétés classiques

- Autres propriétés :
  - Centralité
    - Nombre de plus courts chemins passant par un sommet, etc.
  - Corrélations entre propriétés
    - Degré-degré
    - Degré-clustering
  - Taille des cliques, cliques biparties, etc.

Propriétés communes aux GdT

- Faible densité
- Fort clustering (forte densité locale)
- Faible distance moyenne
- Distribution des degrés très hétérogène
- Une composante géante
- Présence de communautés

Tous les graphes ne partagent pas ces propriétés

Notion de graphe aléatoire

ANALYSE – EXEMPLE

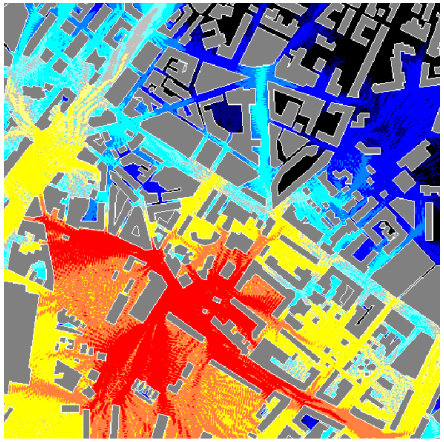
RÉSEAUX DE CONTACTS

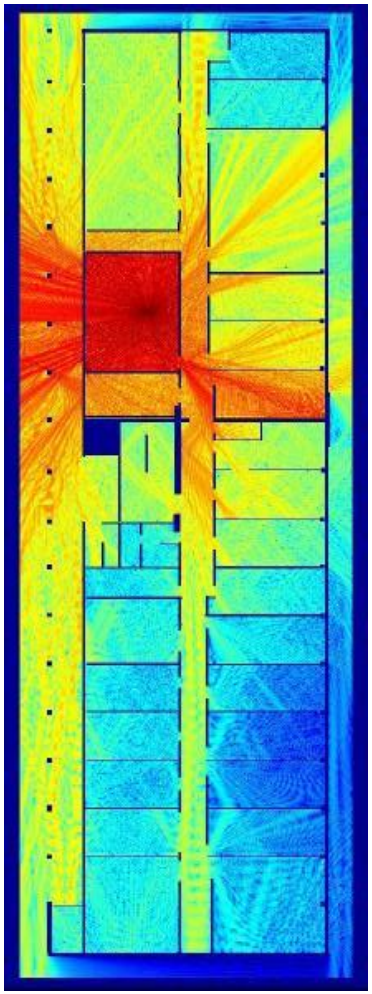
Contexte

- Nombreux équipements avec capacités sans-fil :
  - Ordinateurs, téléphones, PDA, GPS, cartes Navigo…
  - Réseaux sans-fils de plus en plus omniprésents
- Contacts physiques ou virtuels permanents :
  - Rencontres physiques, appels téléphoniques, envoi de mails…
- Objectifs :
  - Tirer parti des contacts naturels des individus
  - Transmission de l'information de proche en proche
  - Réseau dynamique, non connexe : problèmes de routage…

Proximité physique ou radio ?

- Quels contacts entre individus ?
  - Contacts physiques ?
  - Proximité géographique ?
  - Déplacements des individus ?
- Comment mesurer la mobilité ?
  - Suivi des déplacements
  - Géolocalisation (gps)
    - Couteux, difficile à mettre en œuvre
    - Équiper chaque individu



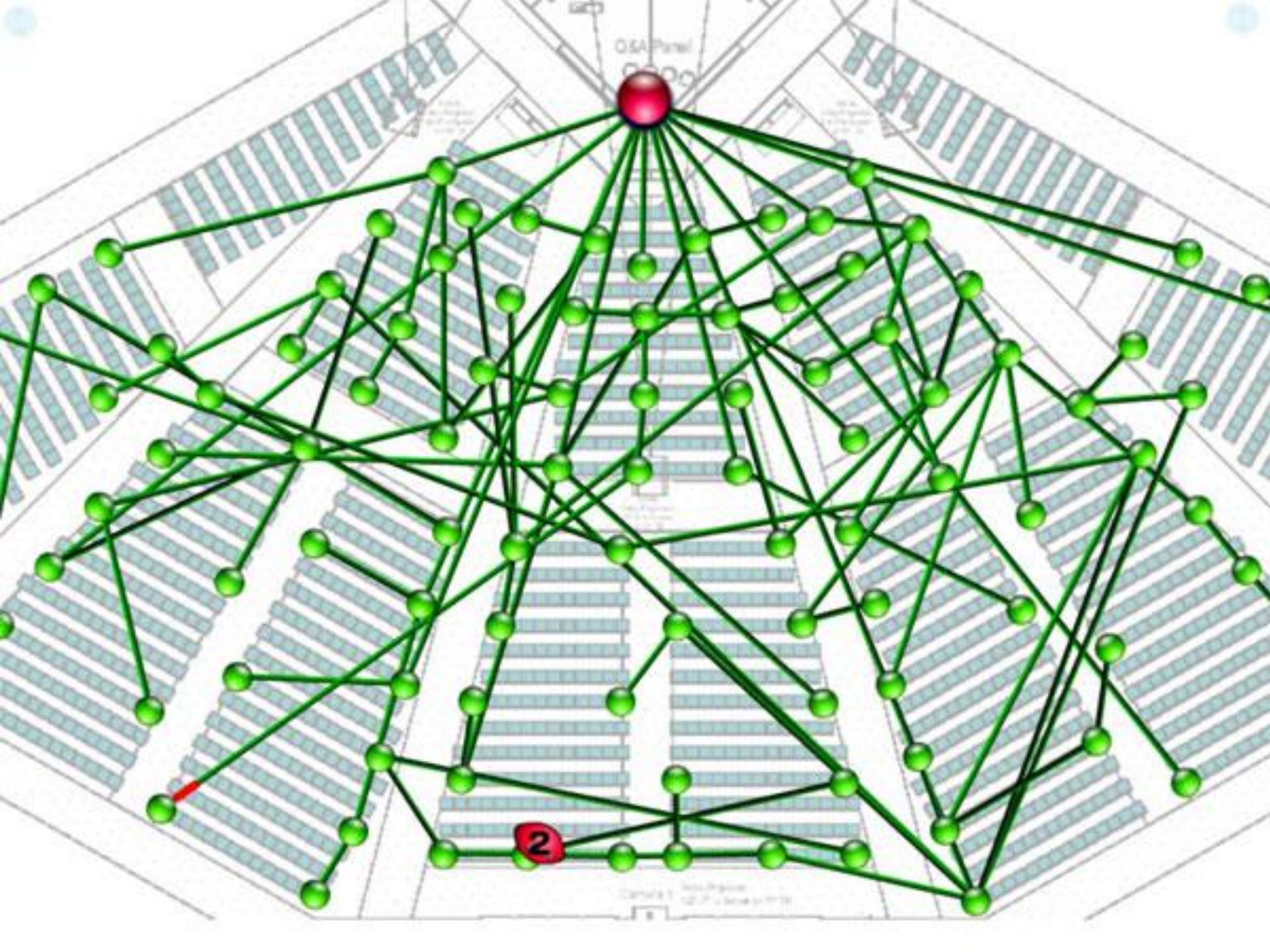


En pratique

- Types de réseaux mesurables :
  - Réseaux de contacts
  - Réseaux et mobilité
- Applications en informatique :
  - Déploiement de réseaux dans des environnements "hostiles"
    - Zones militaires, forêts, …
- Une étude de cas :
  - 41 capteurs durant 3 jours
  - Propriétés dynamiques du réseau

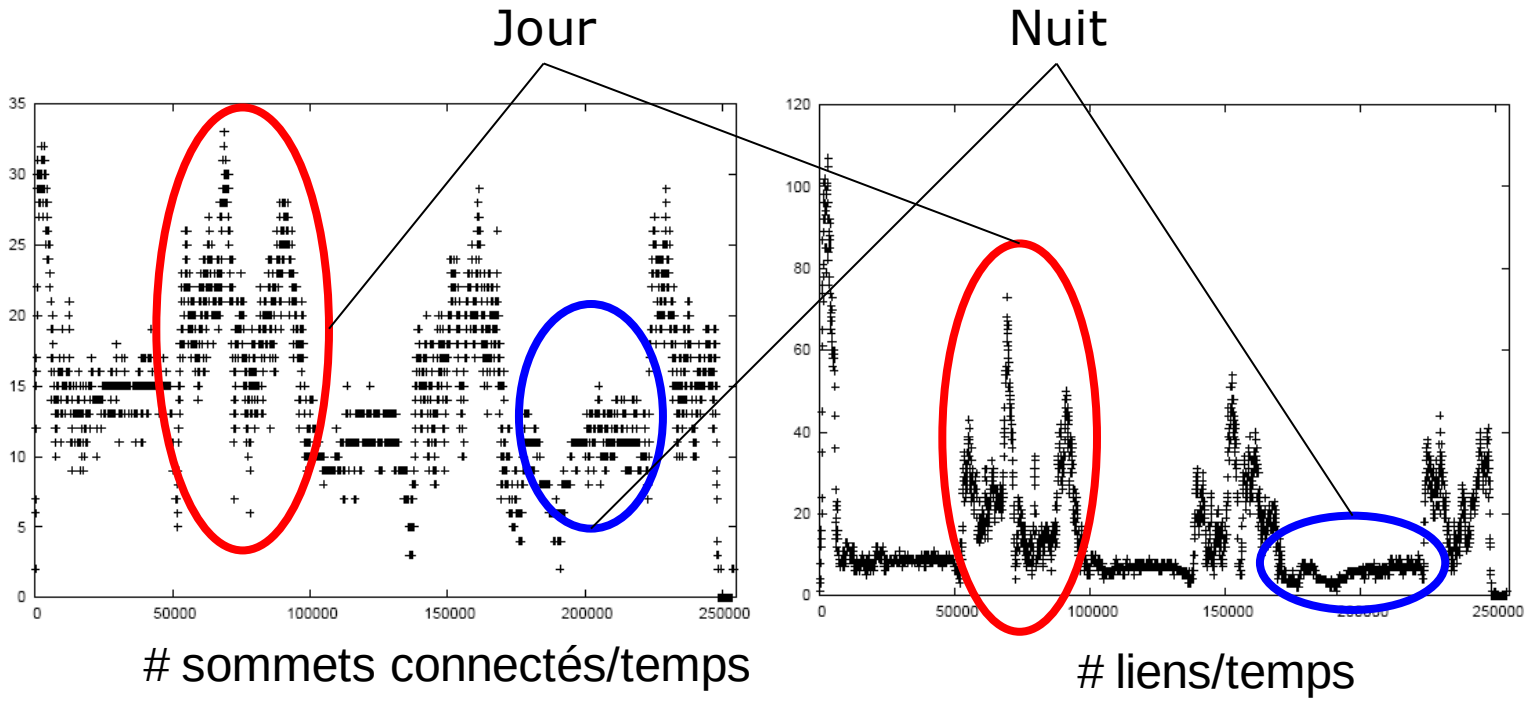
Étude de cas

- Conférence Infocom 2005 :
  - 54 capteurs (11 en panne, 2 perdus)
  - 3 jours (254 151 sec)
- Capteurs bluetooth :
  - Recherche de contacts (5s)
  - Attente (108-132s)
  - Pas de géolocalisation
- Données :
  - Un ensemble de liens à chaque instant
  - Liens non symétriques

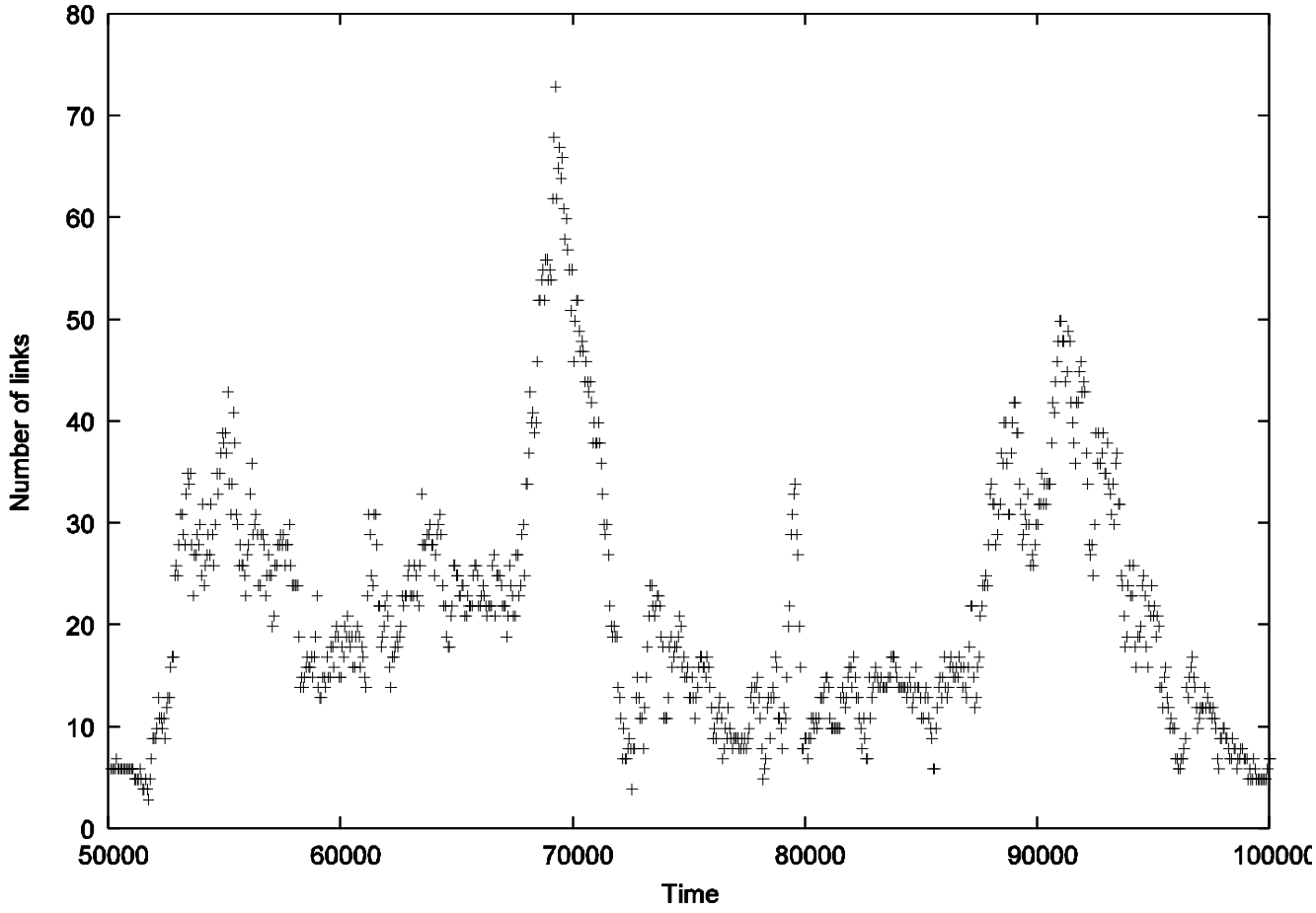

2

Évolution du réseau

- Effets sociologiques :
  - Jours, nuits, repas, pauses…
  - Beaucoup de petites variations. 50% de sommets isolés
  - Au maximum 34 sommets connectés

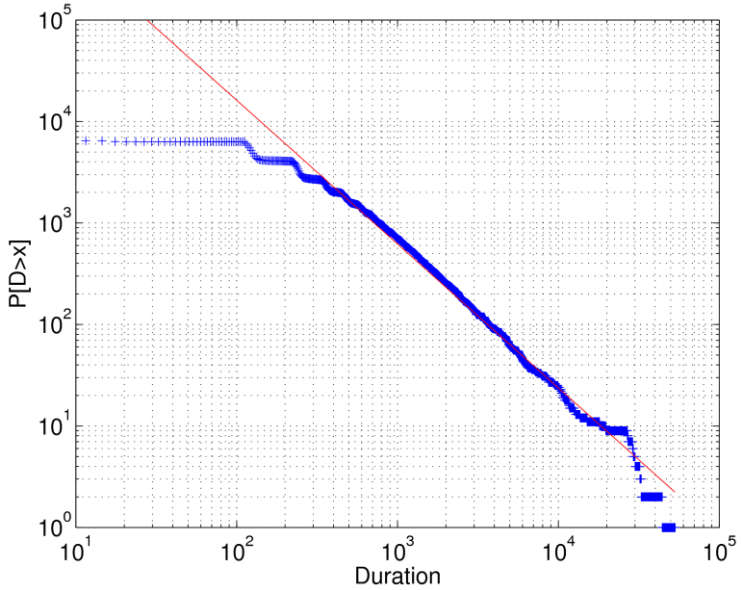


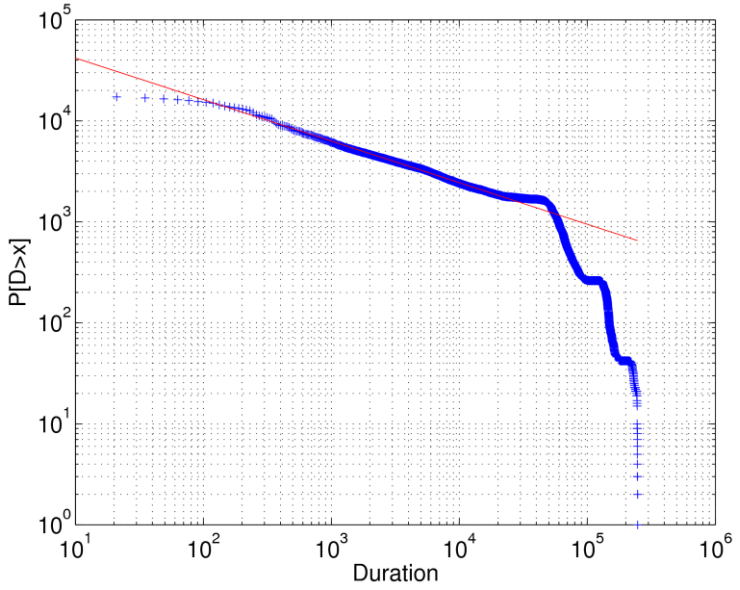
Sur une journée



Durée des contacts

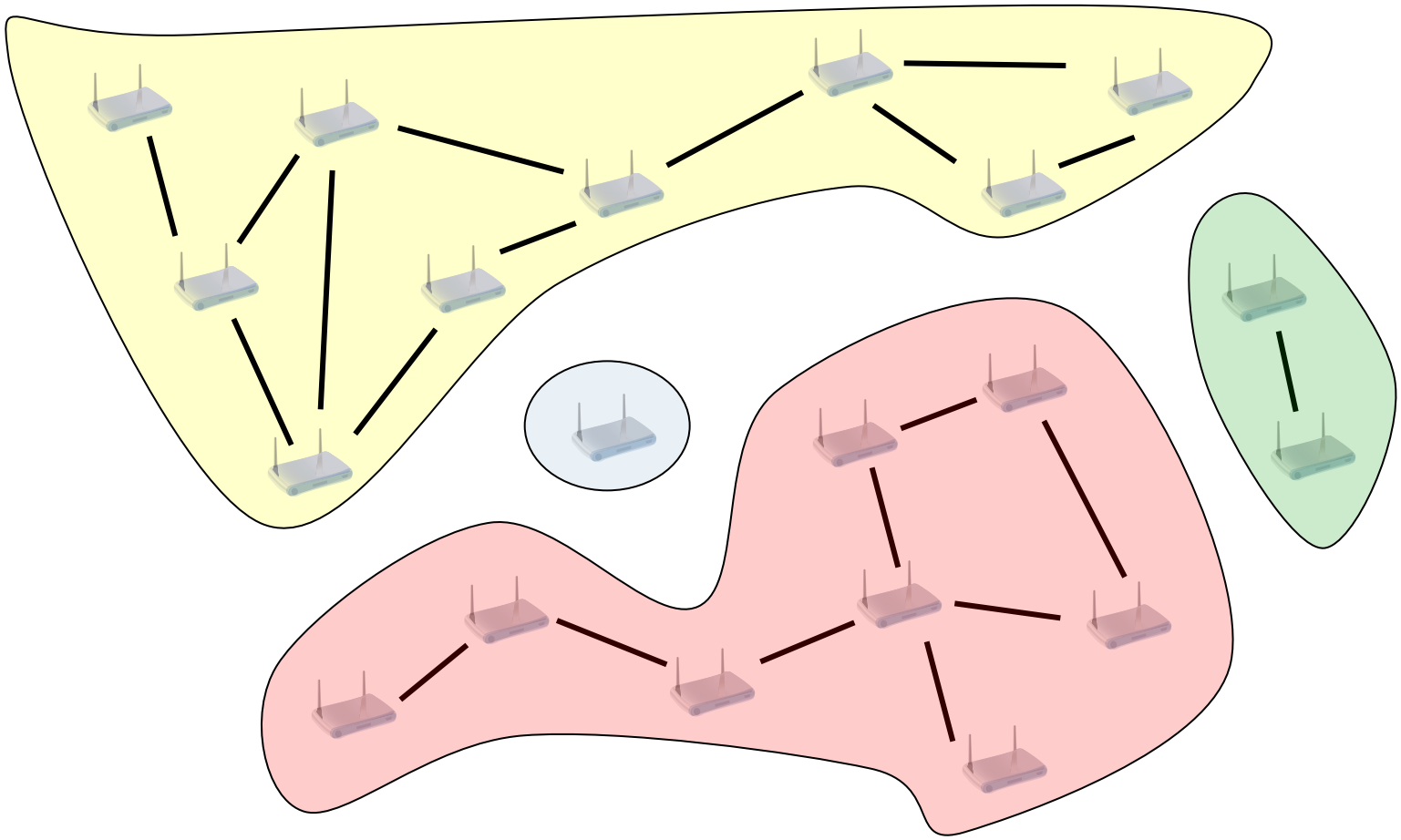
- Distribution en loi puissance (hétérogène)
  - Certains liens sont fréquents, d'autres pas :
    - Liens fréquents utilisables pour router
    - Liens non fréquents pour atteindre des zones particulières





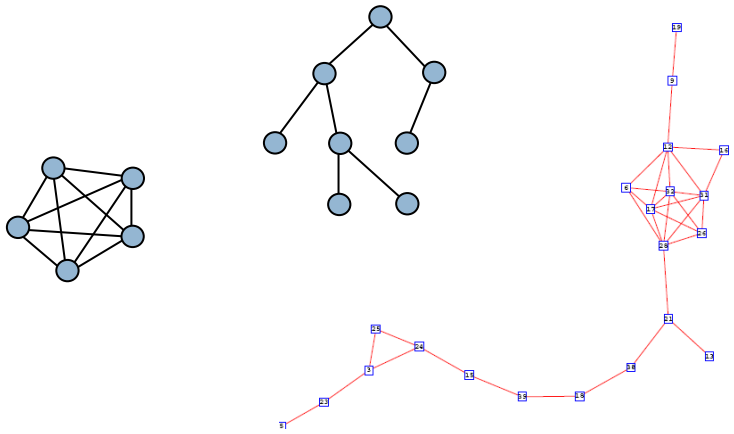
Composantes connexes

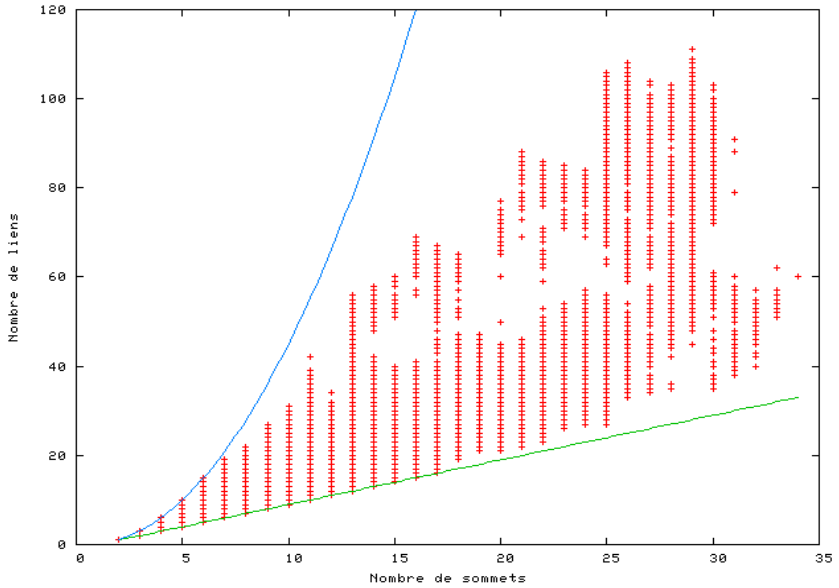
- À chaque instant, composantes connexes :



Composantes connexes

- Ensemble de composantes :
  - Petites composantes : densité variable.
  - Grosse composantes : faible densité :
    - max(nb_liens) ~ 4.5*nb_sommets

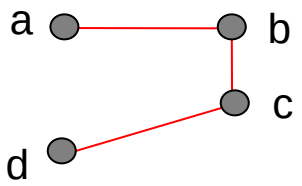


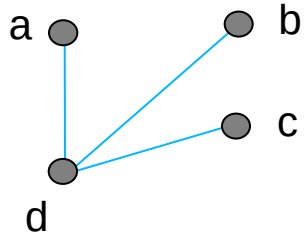


Approche fouille de données

- Graphe dynamique (liens x temps) :
  - Rectangle maximaux de 1
  - Calcul exhaustif ?
    - Graphes fréquent : seuils sur la durée.
    - Graphes significatifs : seuils sur le nombre de liens.

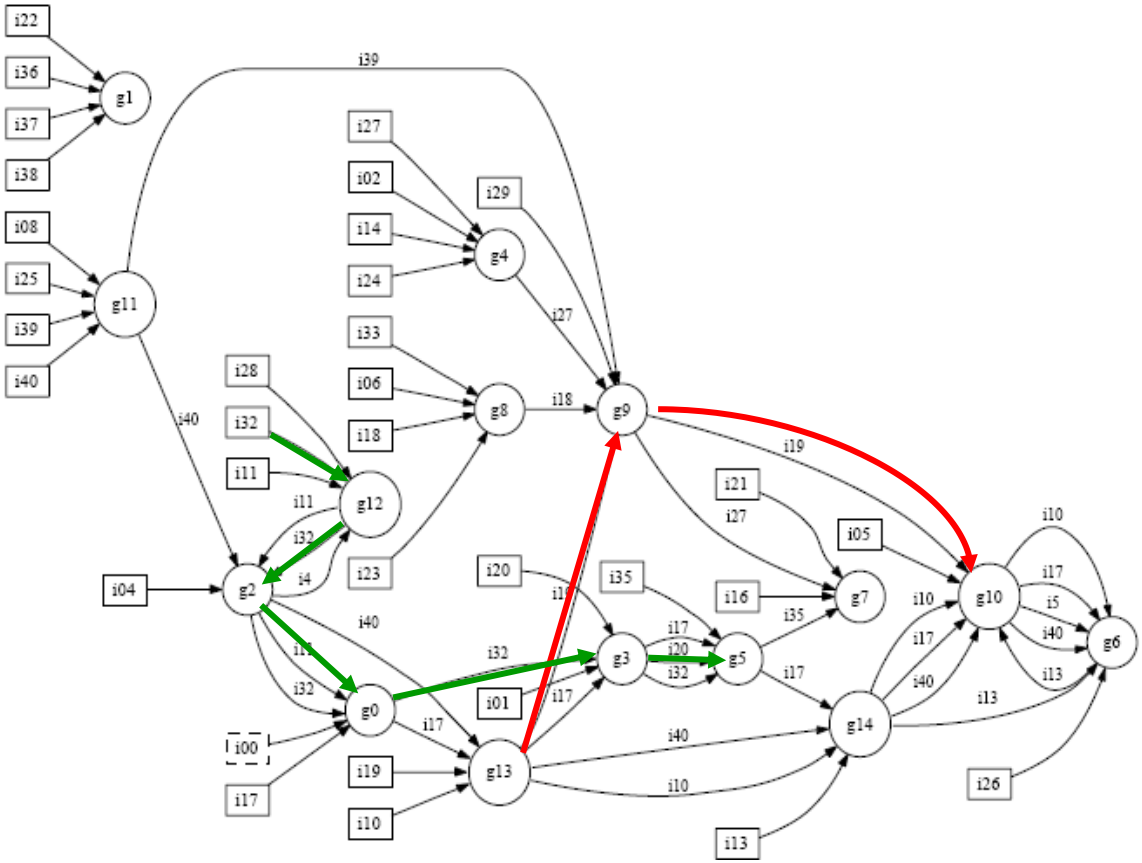
	t1	t2	t3	t4	t5	t6
a-b	1	1	1	1	0	0
a-c	0	0	0	0	0	0
a-d	0	0	0	1	1	1
b-c	1	1	1	1	0	0
b-d	0	0	0	1	1	1
c-d	1	1	1	1	1	1





Identifier les groupes sociaux

- Recherche de groupes fréquents fortement connectés :
  - 19 : entre dans le groupe 13, va dans 9 puis dans 10



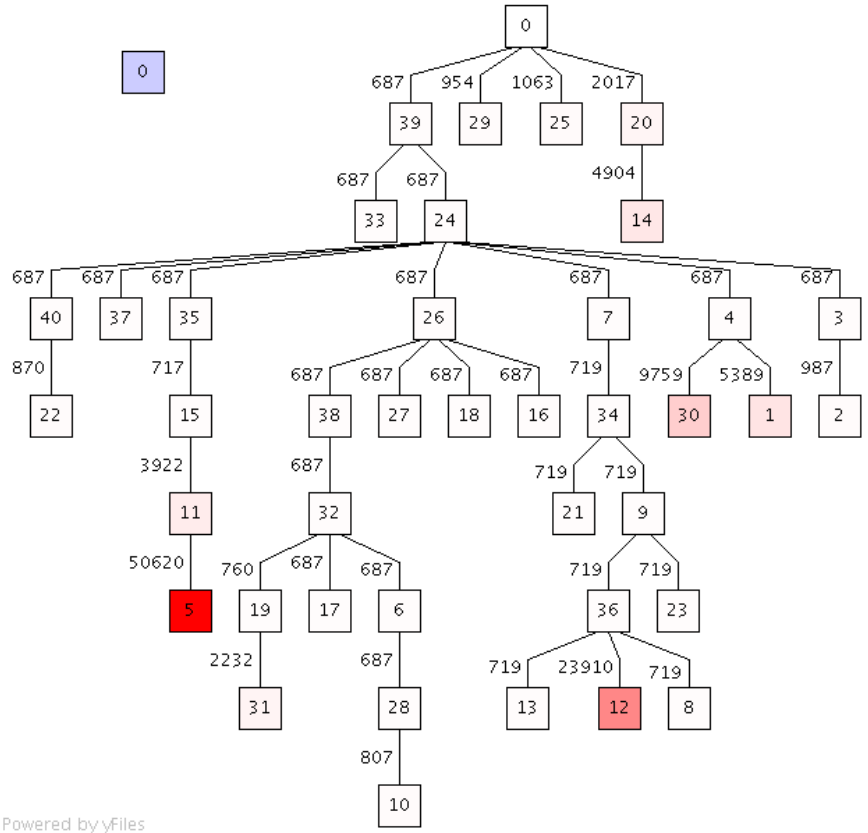
Routage : envoi différé

- Une idée de modélisation :
  - Simulation de propagation d'une rumeur

- Envoi d'un message par u au temps t :
  - v reçoit le message au temps t'
  - v envoi le message immédiatement ou pas ?
  - Paramètres:
    - Qui veut envoyer le message
    - À quel moment
    - Temps de réflexion avant de faire suivre

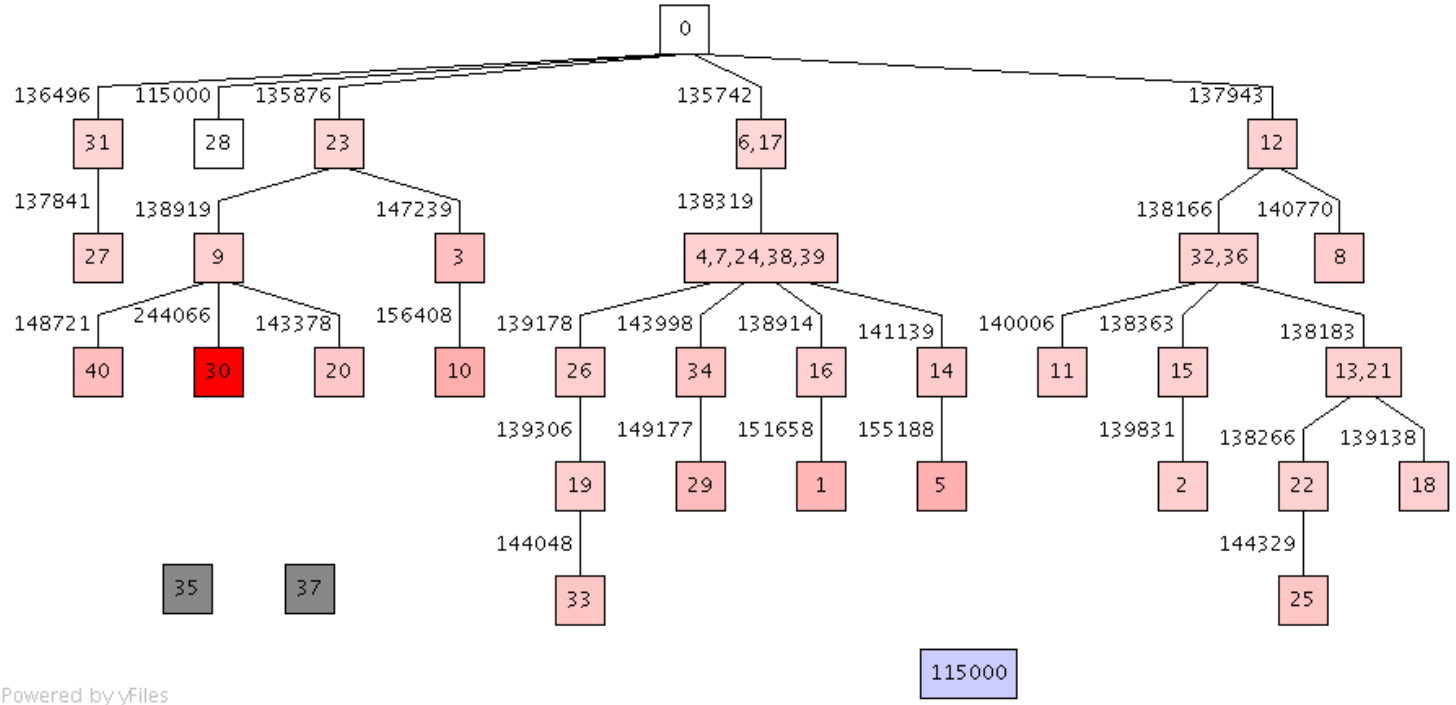
Diffusion rapide

- Diffusion initiée au temps t=0
  - (presque) tout le monde reçoit en quelques minutes



Diffusion lente

- Diffusion initiée durant la nuit (t=115000)
  - Fin de la nuit au temps 138000
  - Nombre maximal de liens au temps 139000



MODÉLISATION

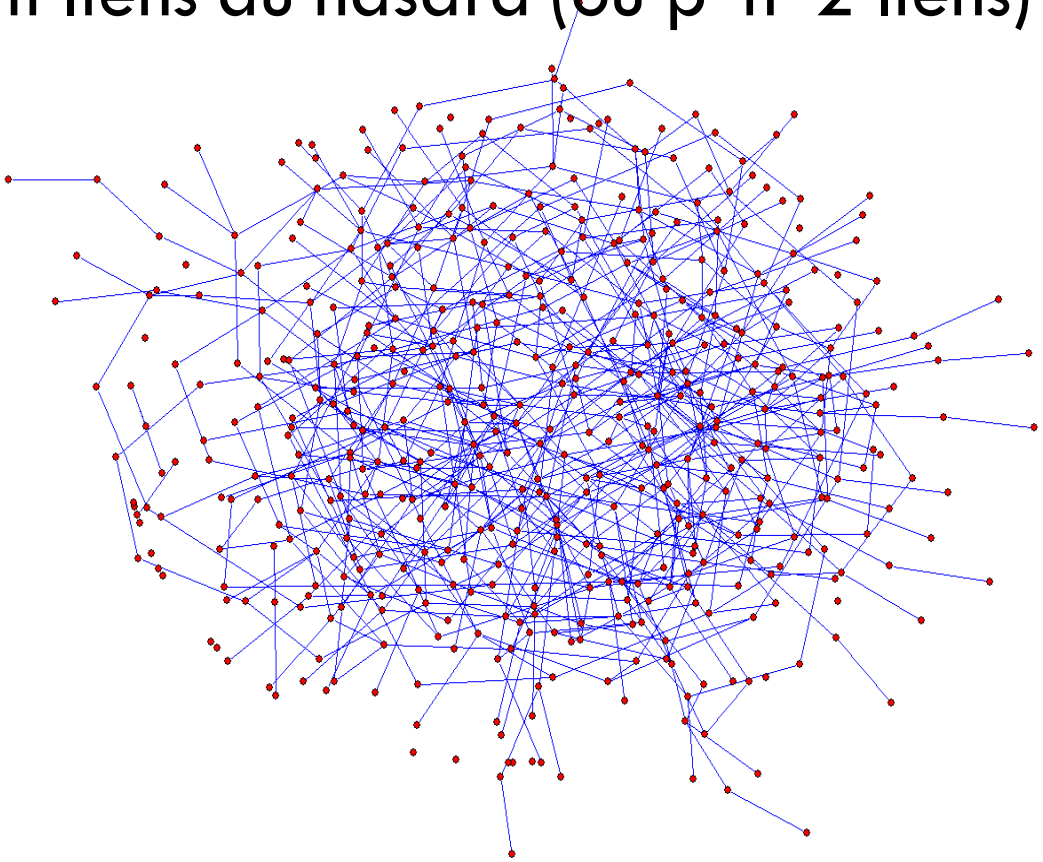
Générer des graphes réalistes

- Propriétés observées normales ?
  - Comparer avec un graphe aléatoire ayant certaines propriétés

- Mais encore…
  - Simuler des phénomènes (attaques, diffusion, …)
  - Évaluer des protocoles, des algorithmes, …
  - Comprendre
  - Prévoir

Tout aléatoire

- Créer n sommets
- Ajouter m liens au hasard (ou p*n^2 liens)



Notion de propriété attendue

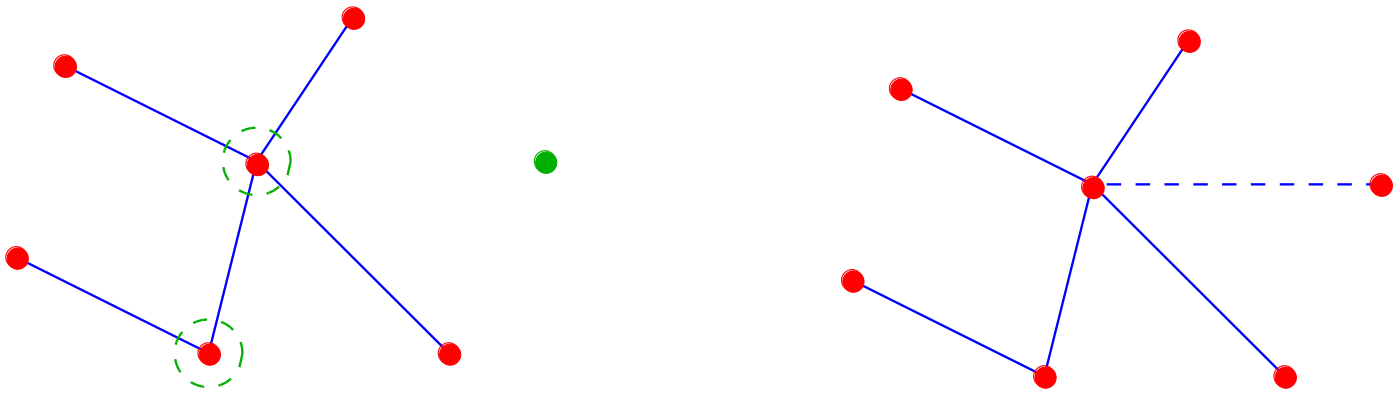
- Exemple : graphe aléatoire, $n = m = 4950$
  - Résultat (réel) : clique de 100 sommets (les autres ont degré 0)
  - Étonnant ?
    - Probabilité d'avoir degré 0 : $q = (1 - 2/n)^n \sim 0.14$.
    - Nombre attendu de sommets de degré 0 : $nq \sim 683$
    - Très peu probable

Propriétés observées

- Densité fixée
- Connexité : composante géante de taille O(n)
  - (pour $m \geq O(n)$)
- Distance moyenne, diamètre ~ log(n)
  - (pour $m \geq O(n)$)
- Distribution des degrés homogène
- Clustering proche de 0
- Pas de structure communautaire

Distribution de degrés

- Attachement préférentiel (rich get richer) :
  - Ajout de sommets un à un.
  - Ajout de lien vers des sommets déjà connectés.



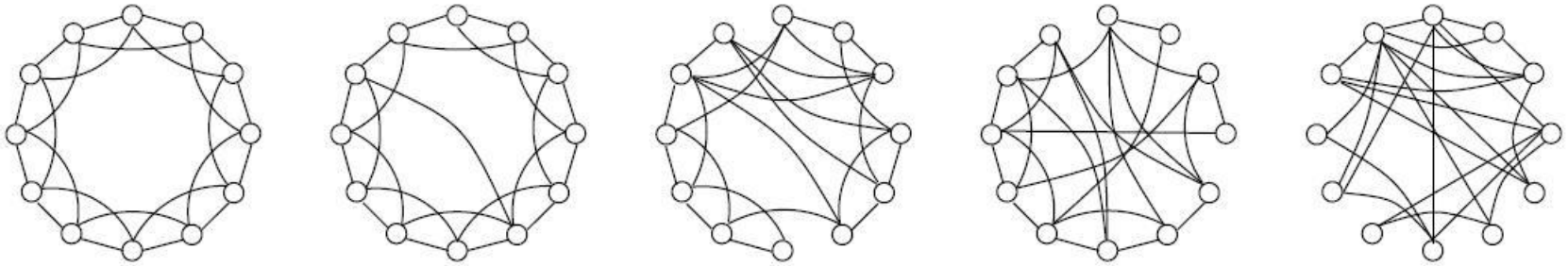
Distribution de degrés

- Modèle configurationnel :
  - On prend n sommets
  - On fixe le degré de chaque sommet
  - On ajoute les liens au hasard en respectant les degrés

- Ces modèles ne génèrent pas de clustering !

Coefficient de clustering

- Mélanger un graphe très rigide :
  - Donne du clustering et une distance moyenne courte
  - Ne donne pas de degrés hétérogènes !



MODÉLISATION – EXEMPLE ROBUSTESSE

Application : robustesse

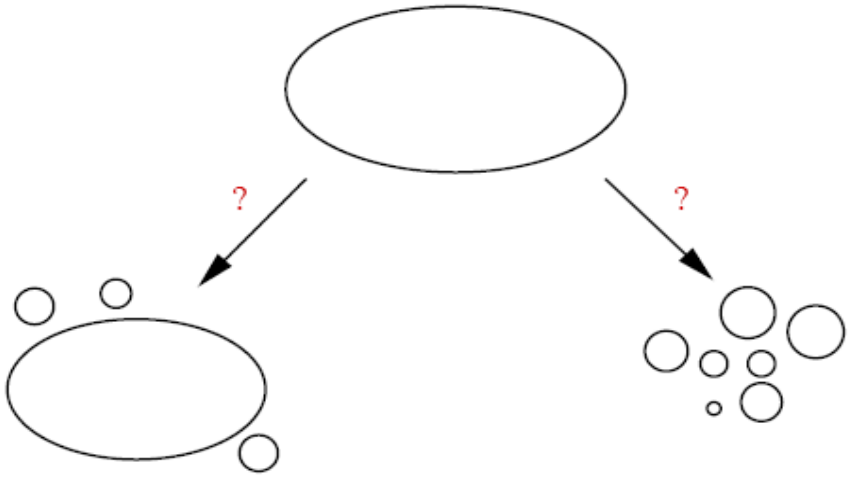
- Étude des phénomènes visant des sommets :
  - Internet : pannes ou attaques sur routeurs.
  - Réseaux sociaux : maladies, rumeurs, …
  - Échanges d'e-mails : virus informatiques.

- Deux types d'atteintes
  - Pannes : aléatoires.
  - Attaques : ciblées.

- But : Comprendre ces phénomènes pour pouvoir :
  - Prédire.
  - Construire des stratégies d'attaque/défense.

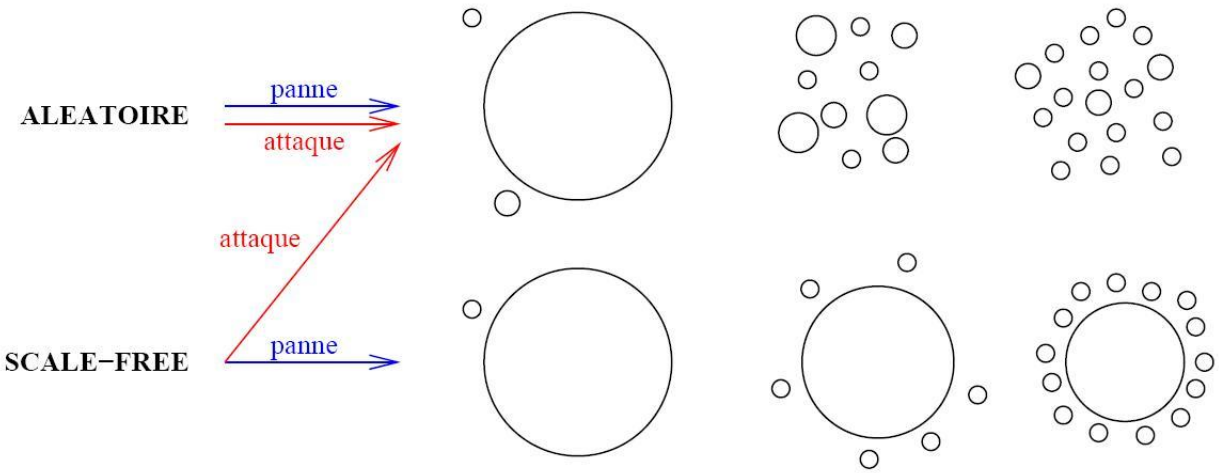
Impact d'une panne/attaque

- Critères :
  - Basés sur la distance.
  - Tailles des composantes connexes.
  - …



Résultats

- Supression :
  - Panne = aléatoire
  - Attaque = ciblée (plus fort degré d'abord)

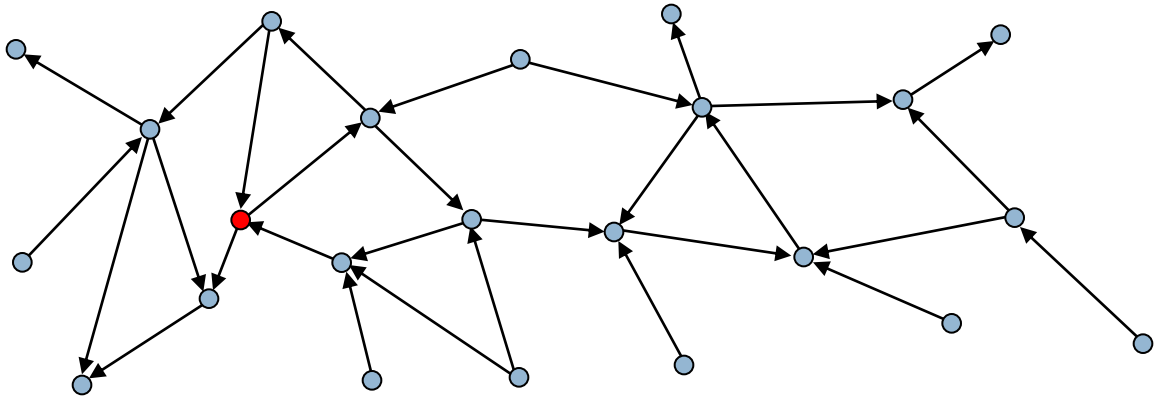


- Question : qui vacciner pour limiter une épidémie ?

MÉTROLOGIE

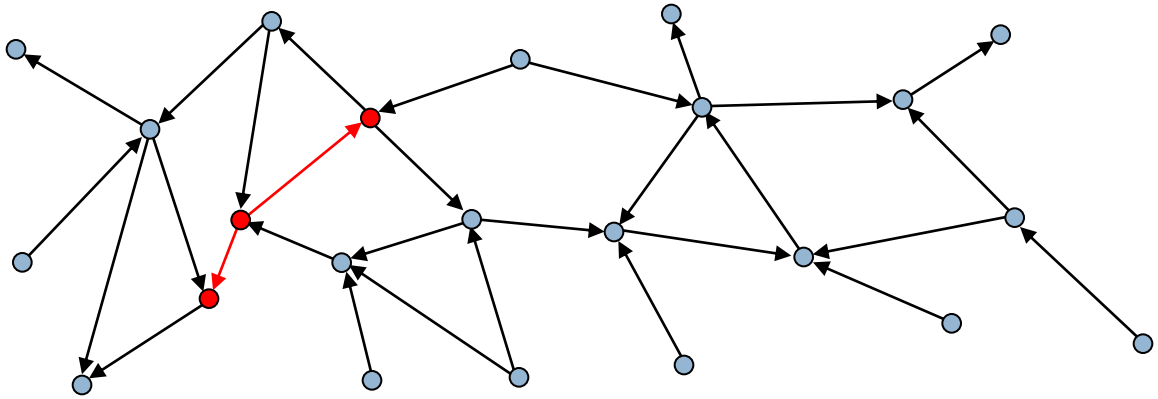
Métrologie du web

- Processus de mesure :
  - Parcours en largeur depuis plusieurs sources
- Réseau : orienté, non connexe, dynamique



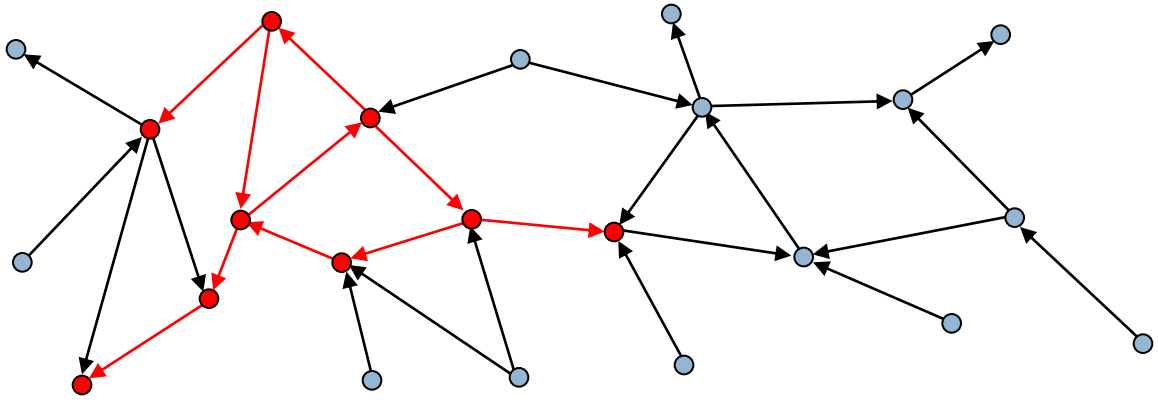
Métrologie du web

- Processus de mesure :
  - Parcours en largeur depuis plusieurs sources
- Réseau : orienté, non connexe, dynamique



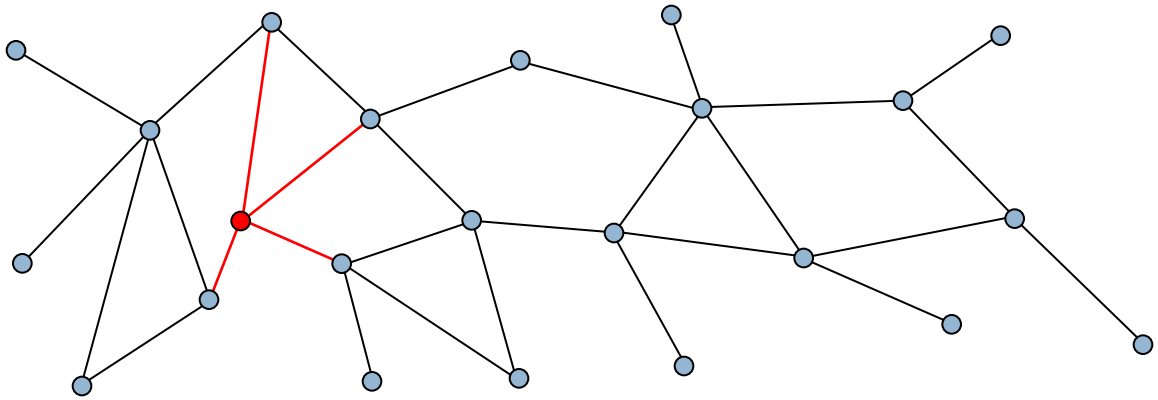
Métrologie du web

- Processus de mesure :
  - Parcours en largeur depuis plusieurs sources
- Réseau : orienté, non connexe, dynamique



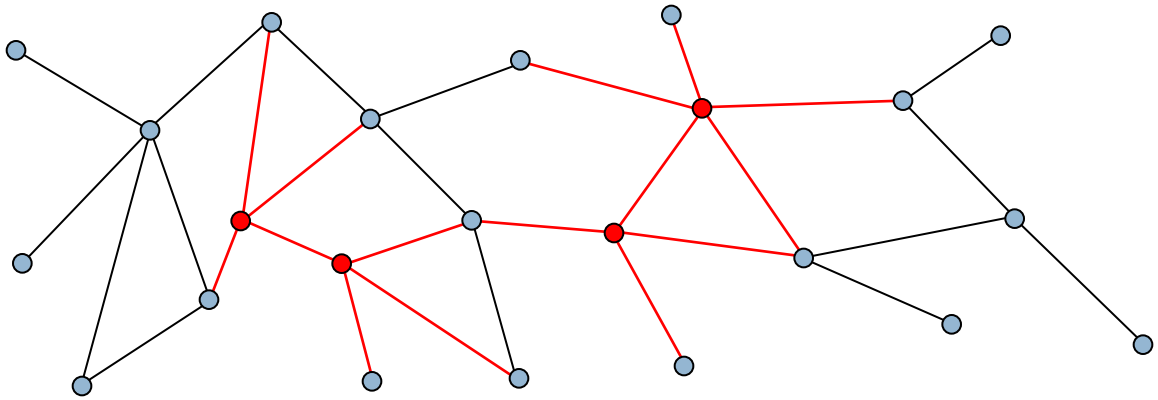
Métrologie des réseaux sociaux

- Processus de mesure :
  - Réseaux égocentrés
  - Listes de diffusion, communautés, …
- Réseau : orienté ou pas



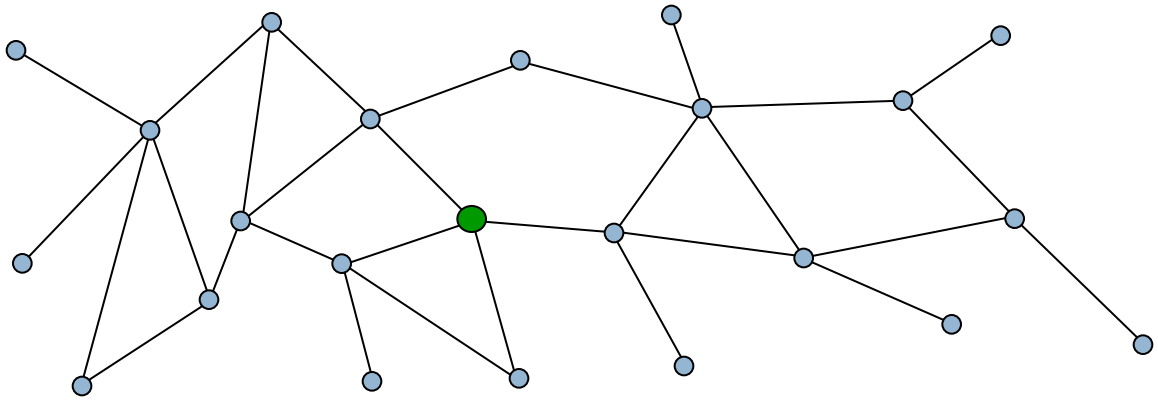
Métrologie des réseaux sociaux

- Processus de mesure :
  - Réseaux égocentrés
  - Listes de diffusion, communautés, …
- Réseau : orienté ou pas



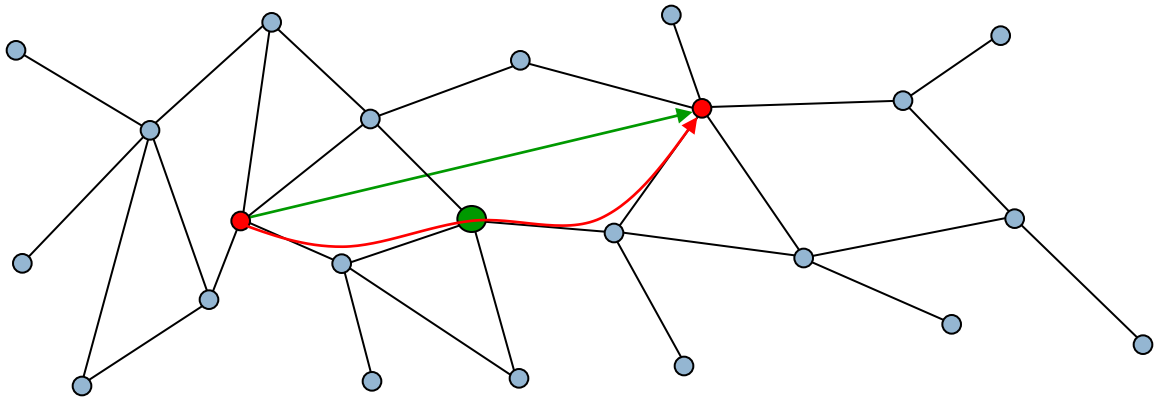
Métrologie des réseaux d'échanges

- Processus de mesure :
  - Trafic passant par un sommet
- Réseau : orienté, pondéré



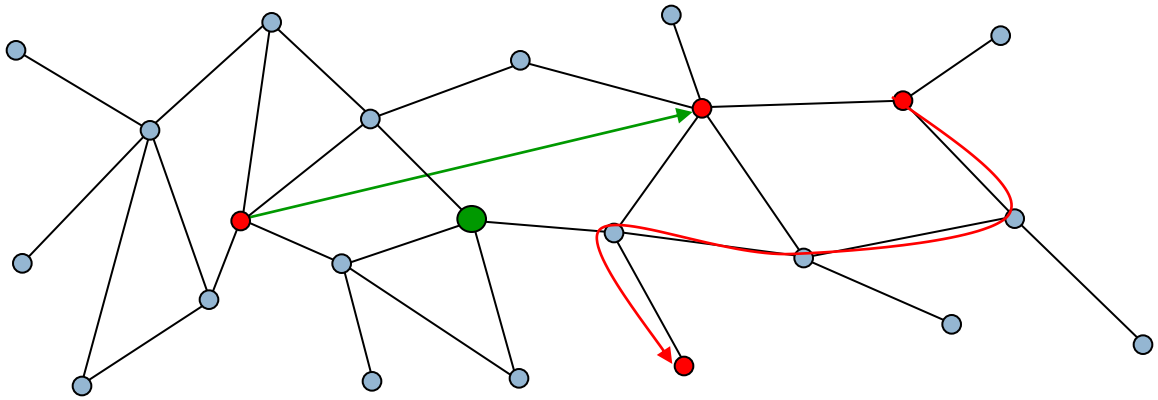
Métrologie des réseaux d'échanges

- Processus de mesure :
  - Trafic passant par un sommet
- Réseau : orienté, pondéré



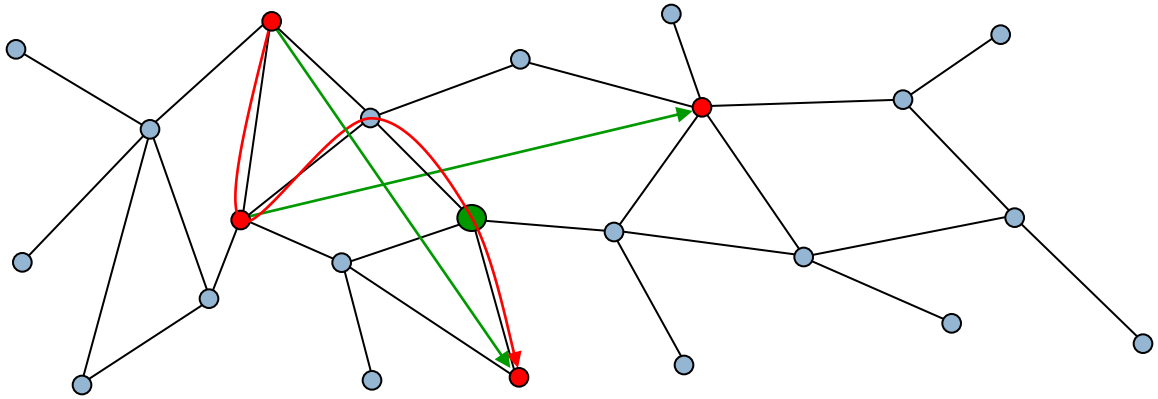
Métrologie des réseaux d'échanges

- Processus de mesure :
  - Trafic passant par un sommet
- Réseau : orienté, pondéré



Métrologie des réseaux d'échanges

- Processus de mesure :
  - Trafic passant par un sommet
- Réseau : orienté, pondéré



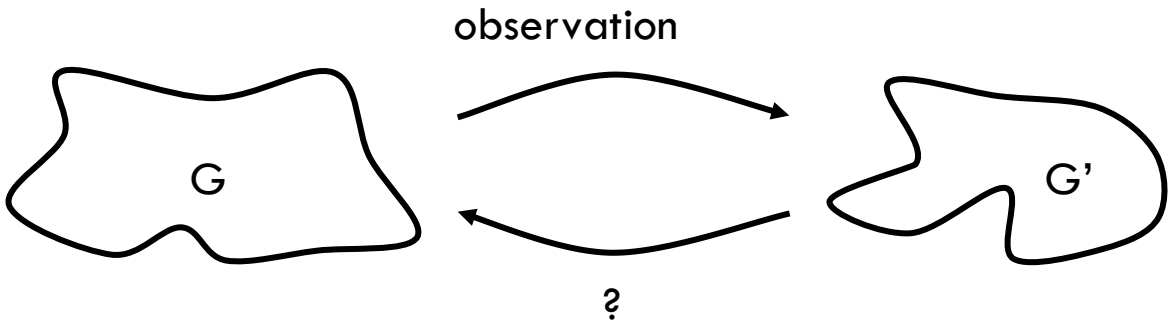
Métrologie

- En général :
  - Impossibilité d'étudier l'objet réel, juste une mesure

- Questions à se poser :
  - Qui a fait la mesure ?
  - Quelle proportion de l'objet a été mesurée ?
  - Combien de temps la mesure a duré ?
  - Y-a-t'il des contraintes spécifiques ?
  - La mesure peut-elle être reproduite ?

Métrologie

- Étude du biais introduit par l'observation
- Que dire de l'objet réel à partir de l'observation ?
- Nouveaux protocoles de mesures, etc.



- Évaluer la représentativité des "cartes"

Une approche

- On simule la mesure sur un graphe aléatoire

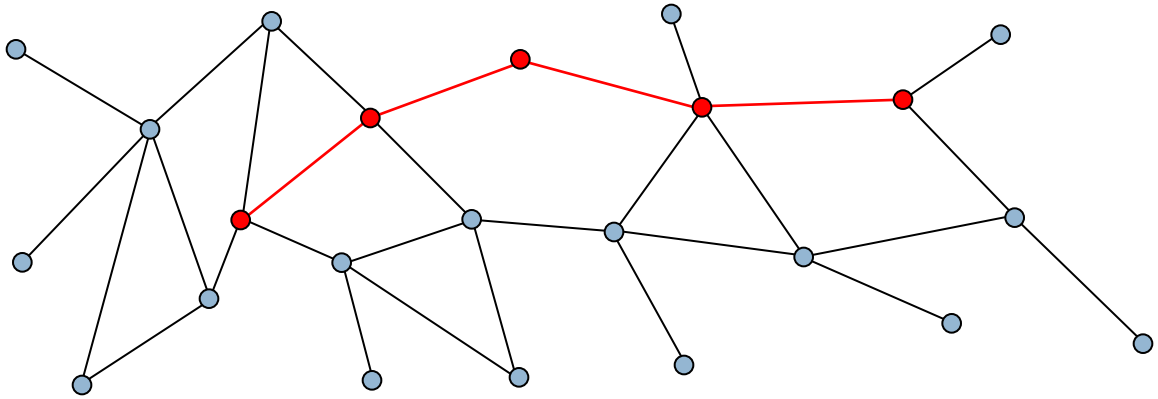
- Modélisation du processus de mesure :
  - Internet : traceroute = chemins courts
  - Web : crawl = parcours en largeur

- Modélisation du réseau :
  - Graphes aléatoires
  - Respect des degrés, du clustering ou autre
  - …

MÉTROLOGIE – EXEMPLE INTERNET

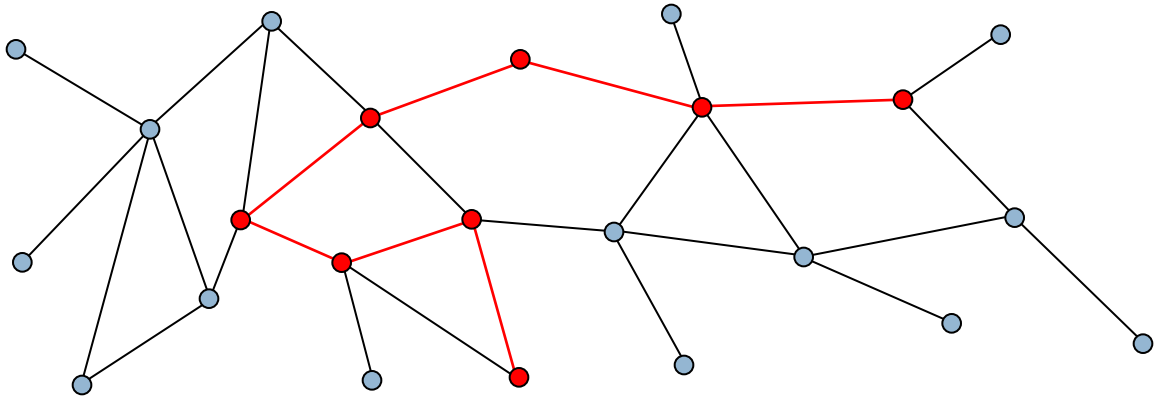
Métrologie de l'internet

- Processus de mesure :
  - Traceroute ~ (plus courts) chemins de plusieurs sources vers plusieurs destinations
- Réseau : (non) orienté, pondéré (RTT, …)



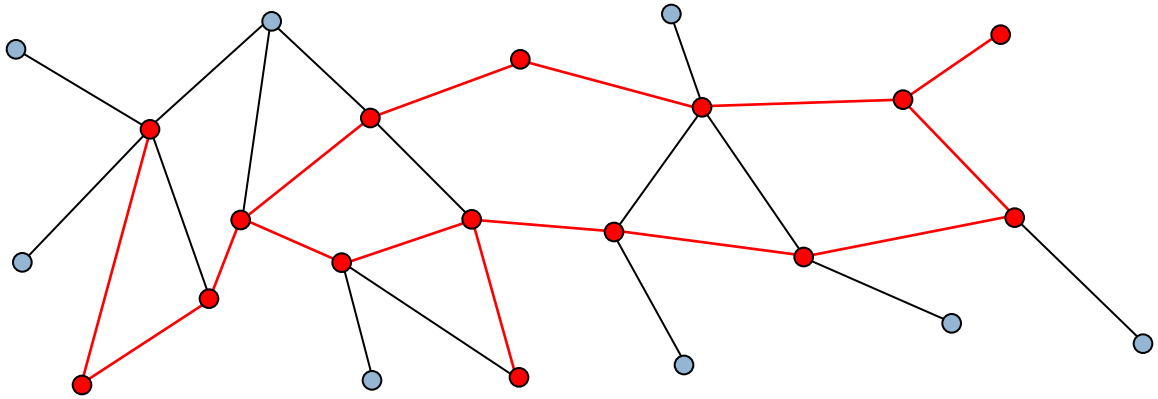
Métrologie de l'internet

- Processus de mesure :
  - Traceroute ~ (plus courts) chemins de plusieurs sources vers plusieurs destinations
- Réseau : (non) orienté, pondéré (RTT, …)



Métrologie de l'internet

- Processus de mesure :
  - Traceroute ~ (plus courts) chemins de plusieurs sources vers plusieurs destinations
- Réseau : (non) orienté, pondéré (RTT, …)



Questions

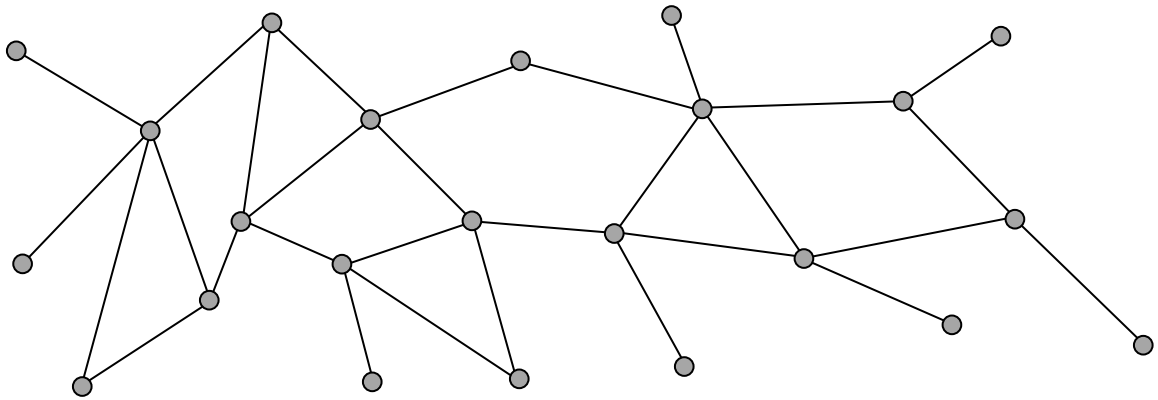
- Influence de :
  - Nombre de sources et destinations
  - Propriétés du réseau
  - Localisation des sources et destinations

  sur le résultat obtenu ?

- Modélisation :
  - Traceroute = plus courts chemins (un ou tous)
  - Graphe = graphe aléatoire (modèle à choisir)

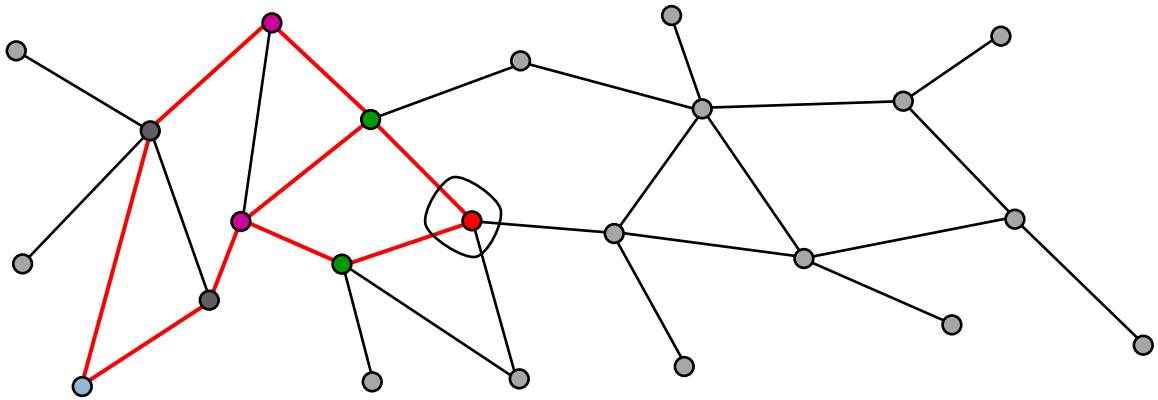
Que voit-on ?

- D'une source vers tout le monde



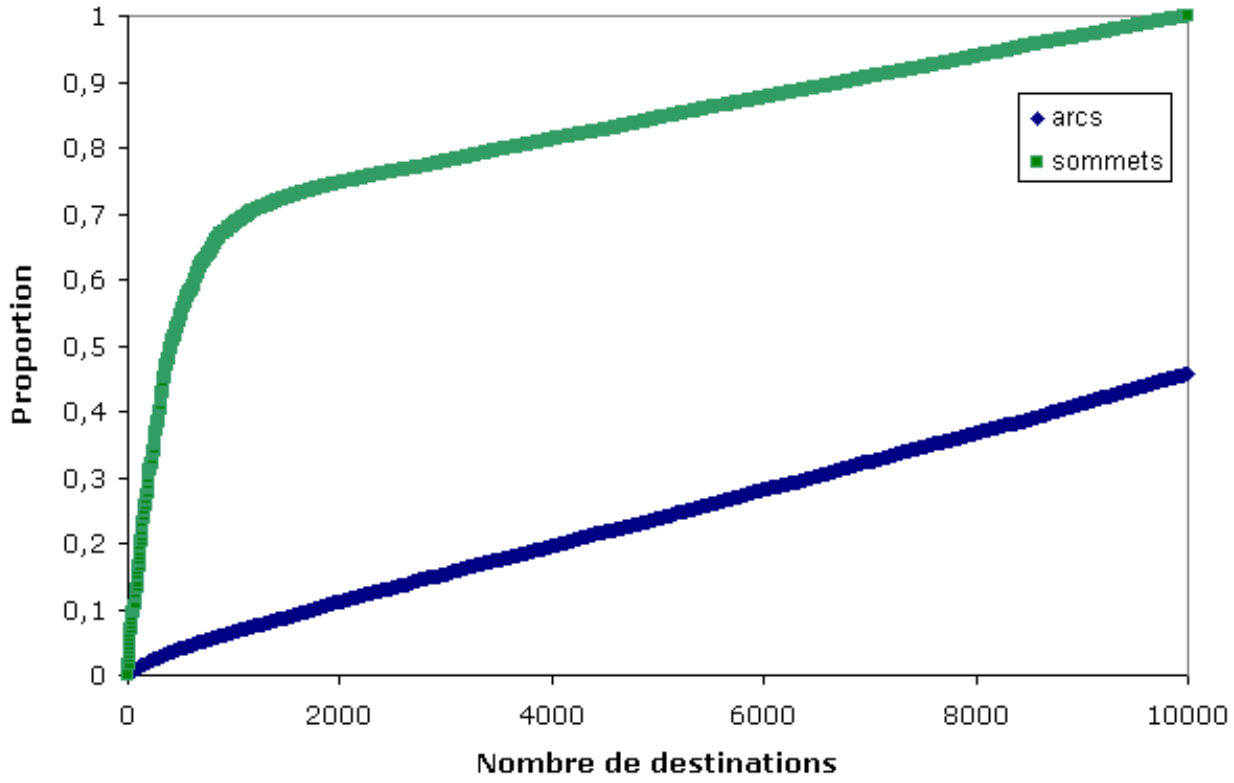
Que voit-on ?

- D'une source vers tout le monde :
  - Liens rouge découverts (sur des plus courts chemins)
  - On répète pour les autres destinations.
- Résultat ?



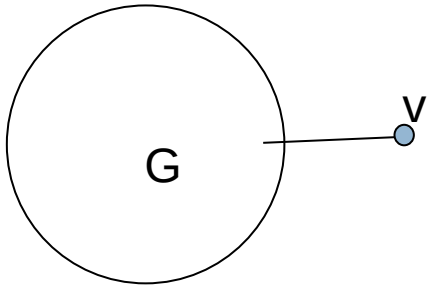
Exemple : une source, k destinations

- Graphe aléatoire, tous les chemins :
  - Modification du nombre de destinations

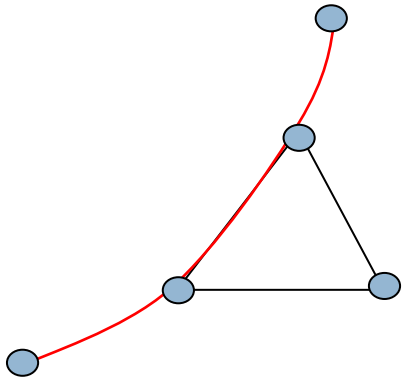


Zones dures à mesurer

- Sommet de degré 1 : uniquement visible si source ou destination

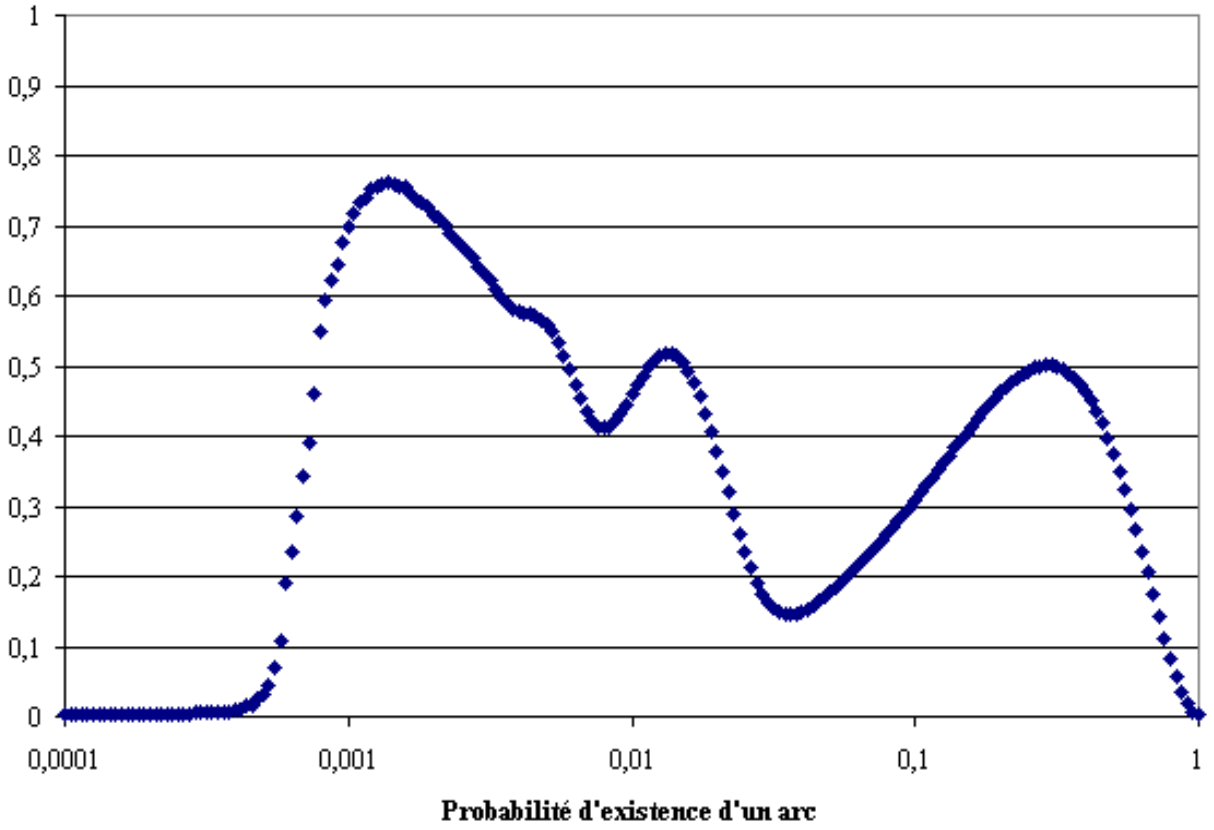


- Graphe complet : visiter tous les liens



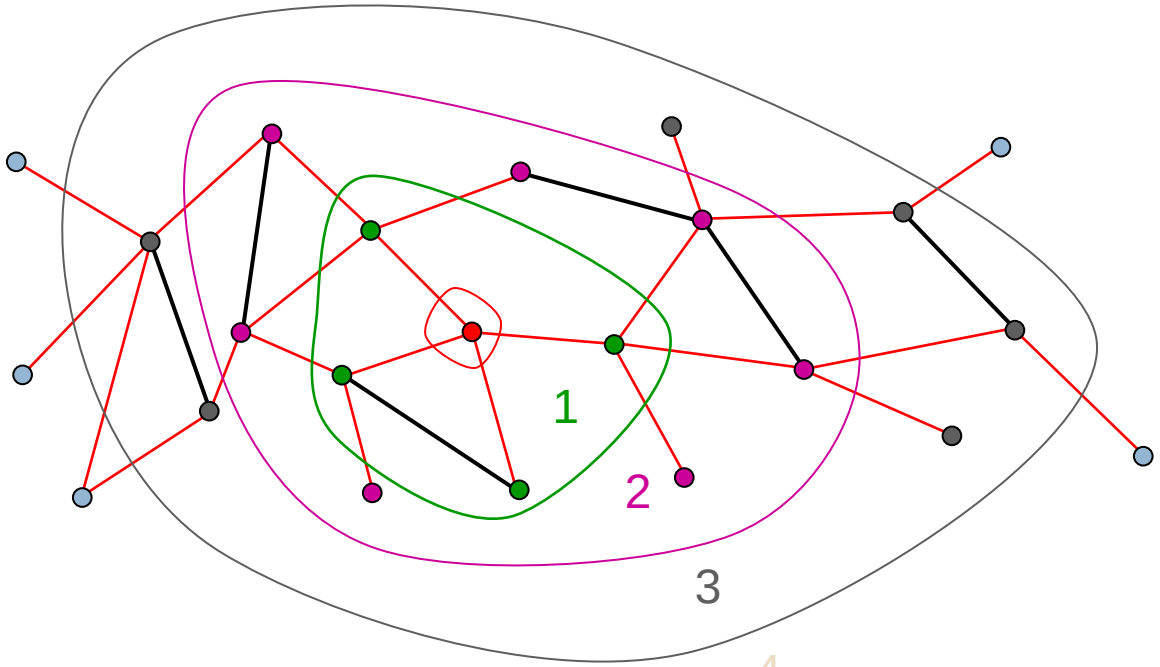
Variation de la densité

- De un vers tous :

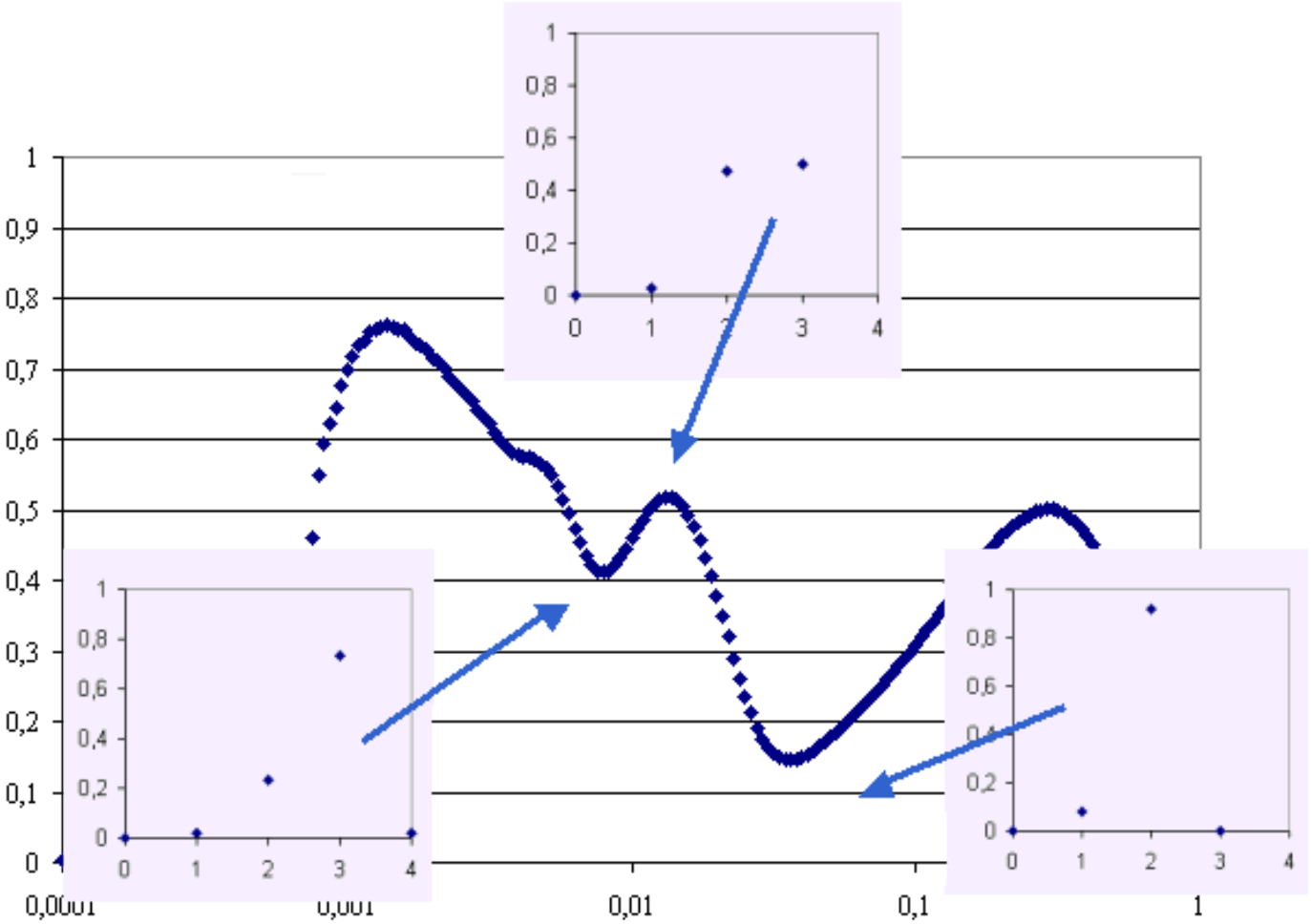


Que voit-on ?

- D'une source vers tout le monde :
  - Liens rouge découverts (sur des plus courts chemins)
  - Liens noirs invisibles

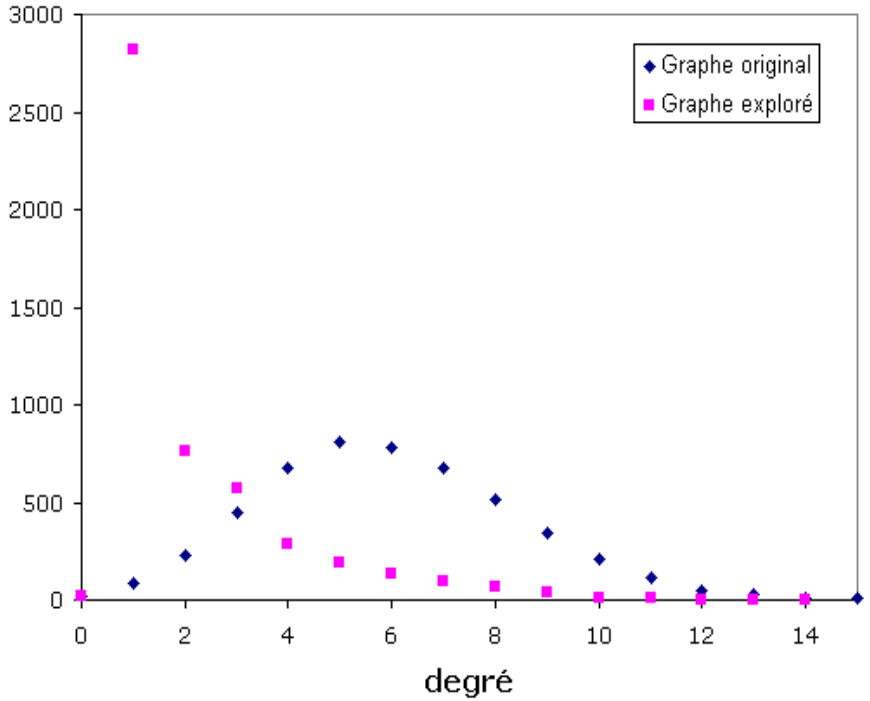


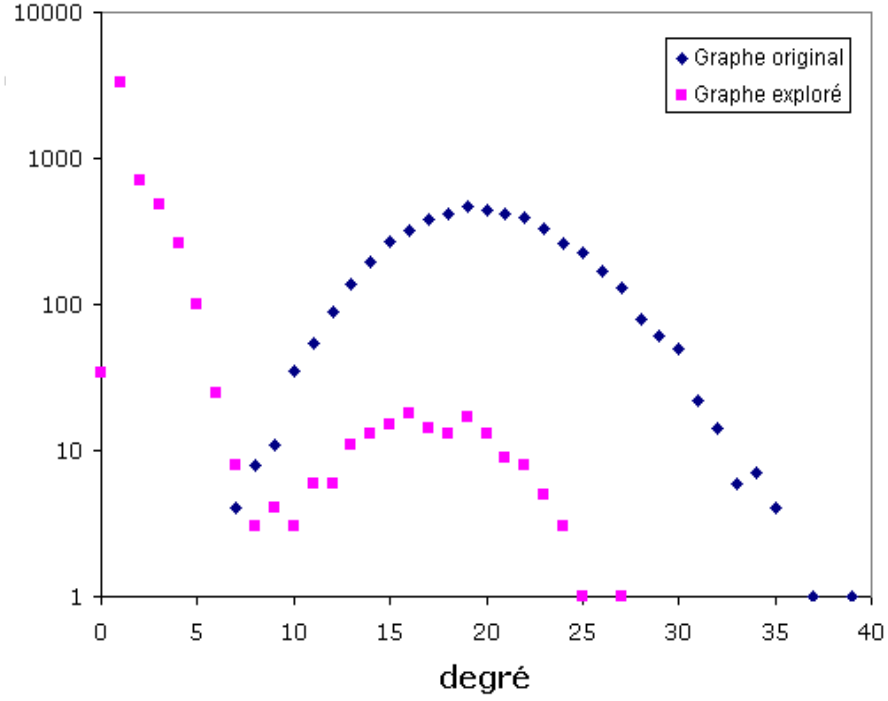
Variation de la densité



Distribution des degrés

- Différences entre l'original et la mesure :
  - Beaucoup de sommets de faible degré
  - Peu de sommets de fort degré
- Mauvaise estimation de la propriété réelle !





ALGORITHMIQUE

Besoin d'algorithmes spécifiques ?

- Gros problème = taille :
  - Internet = Millions de sommets (routeurs)
  - Facebook = 500 millions d'utilisateurs actifs
  - Web = Google connait plus de 1000 milliards d'URL distinctes (pas de pages distinctes)

- Il est non trivial de :
  - Stocker le graphe en mémoire
  - Faire des calculs sur le graphe

Exemples

- Compter les triangles d'un graphe (clustering) :
  - naïvement O(n*n*n)
  - O(m*n^1/a) si distribution des degrés en loi puissance d'exposant a
- Diamètre :
  - Complexité théorique : O(nm)
  - Approximation en O(m)
- Problèmes NP-complets ?

Exemples

- Beaucoup de problèmes spécifiques aux graphes de terrains :
  - Exemple : détection de communautés, NP-complet
  - Approximation (non prouvée) : linéaire

PROJET TME

Sujet ouvert

- Construction d'un réseau social :
  - Similarité entre utilisateurs de Jester Jokes
  - Basé sur les notes et les méthodes de similarité vues dans les 3 premiers cours
- Calculs de communautés :
  - Liens entre qualité et similarité
- Prédiction de liens ?