Graphes petits mondes

Nicolas Hanusse

Chercheur CNRS au LaBRI

hanusse@labri.fr

Objectifs

- Caractériser les données les plus fréquemment rencontrées
  - En partant de données réelles …
  - En proposant de bons « modèles »
- Naviguer visuellement dans les masses de données:
  - Recherche d’information
  - Fouille de données
  - Exploration

Plan

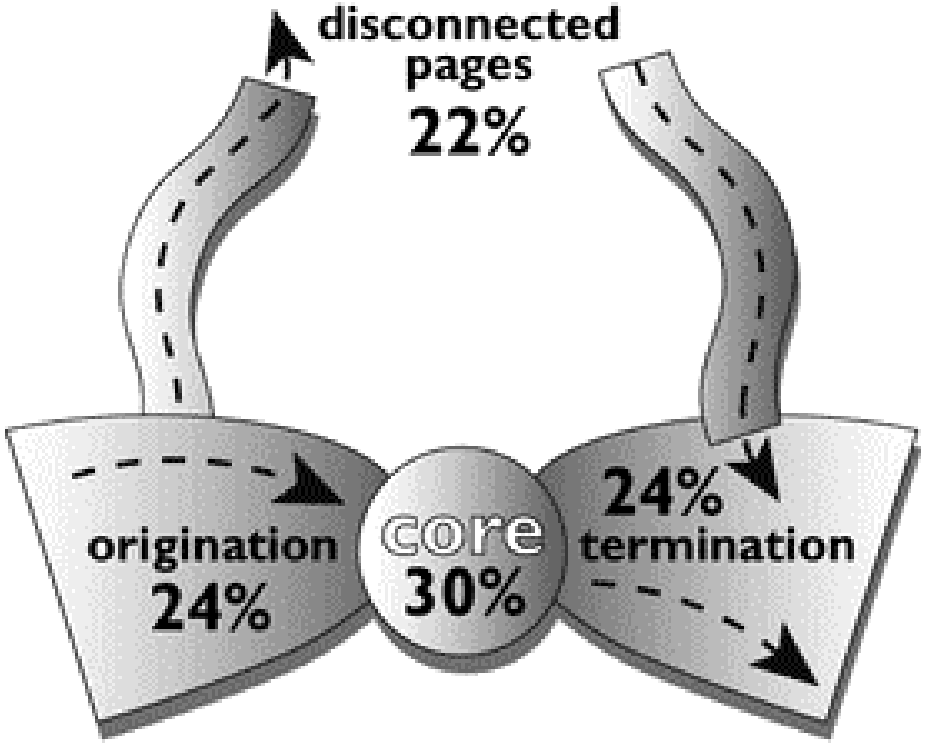
- Exemples de quelques graphes et problèmes issus de la réalité;
- Modèles de graphes « petits-mondes »:
  - Approche statistique;
  - Approche routage.
- Applications;
- Problèmes de recherche.

Quel est le rapport entre …

- Graphe du Web ?
- Réseaux sociaux:
  - collaborations scientifiques – Erdös Number;
  - Hollywood graph.
- Réseaux d’intéractions: protéïne-protéïne;
- Réseaux de transport aériens;
- Ontologies lexicales et sémantiques;

Graphe du Web

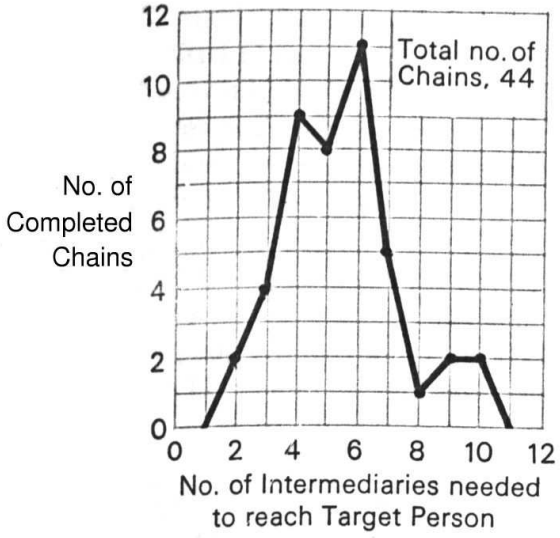
- Graphe du web:
  - Nœuds = page web
  - Arcs = liens hypertexte
- Obtenu par un crawl de 2 millions de nœuds.
- Résultats:
  - Forme en « nœud papillon »;
  - Diamètre estimé à 20.



Graphe des collaborations scientifiques

- Graphe des citations:
  - Noeud = article;
  - Lien = référence d’un article vers un autre.
- Statistiques sur les degrés:
  - Majorité d’article ont très peu de citations:
    - 633,391 sur 783,339 articles ont moins de 10 citations;
    - En moyenne, un article est cité moins de 2 fois.
  - Existence d’articles trés populaires:
    - 64 articles ont plus de 1000 citations;
    - 1 article a 8907 citations.

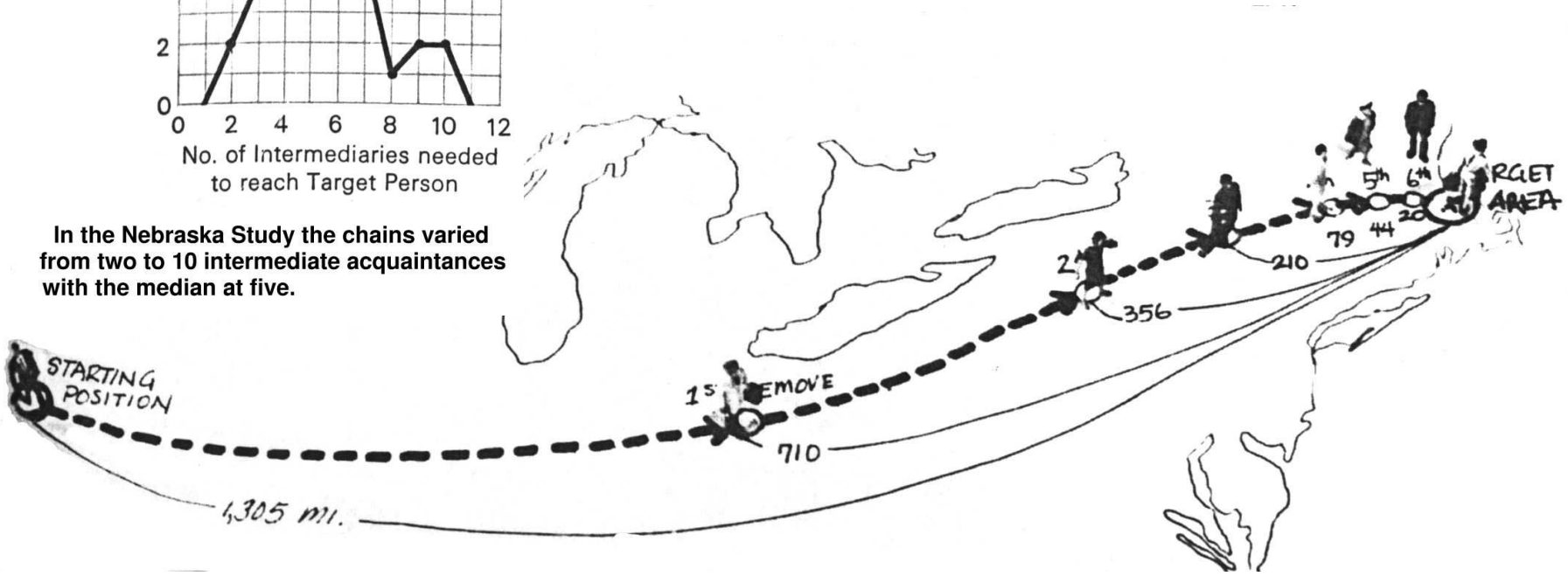
Réseaux sociaux: L'expérience de Milgram - 1967



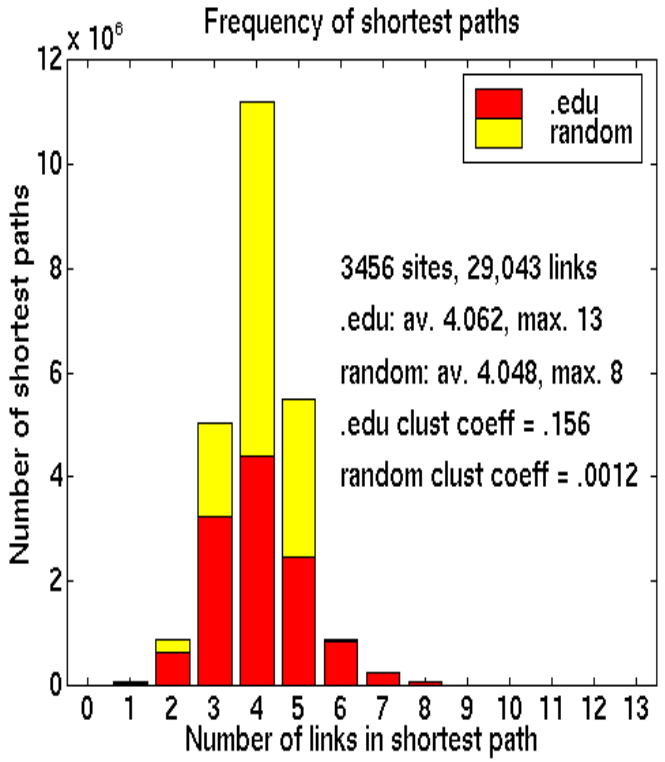
In the Nebraska Study the chains varied from two to 10 intermediate acquaintances with the median at five.

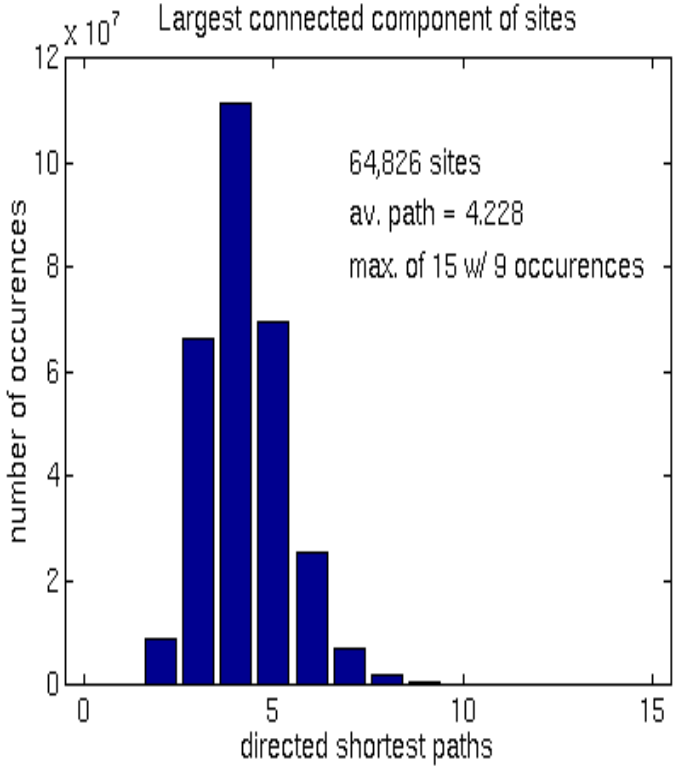
- Du Nebraska à Boston …

Pourquoi le routage glouton est-il efficace ?



Distribution des distances





Lada A. Adamic. The Small World Web. 2000.

Ce sont tous des graphes petits mondes !

Pas de définition précise *mais* existence de nombreux courts chemins.

Caractérisation des petits mondes

- Statistiques sur des paramètres de graphes:
  - Petit diamètre;
  - Densité locale forte;
  - Densité globale faible = degré moyen faible;
  - Distribution des degrés: hétérogène ou homogène;
  - …
- Algorithmique: on peut router facilement et rapidement dans un petit monde.

Le monde réel est-il aléatoire ?

Non !

- *Pouvons-nous identifier des structures communes dans les données issues du monde réel ?*

Réalité versus Aléatoire

- Les graphes réels ont tendance à être creux
  - Graphe aléatoire creux contiennent peu de motifs ou structures.
- Graphes réels comportent de nombreux courts chemins:
  - Comme les graphes aléatoires
- Graphes réels sont localement denses
  - Contrairement aux graphes aléatoires creux.

- ***d***: degré moyen faible

- *L*: longueur moyenne faible
- *D*: diamêtre petit

- *C*: coefficient de regroupement

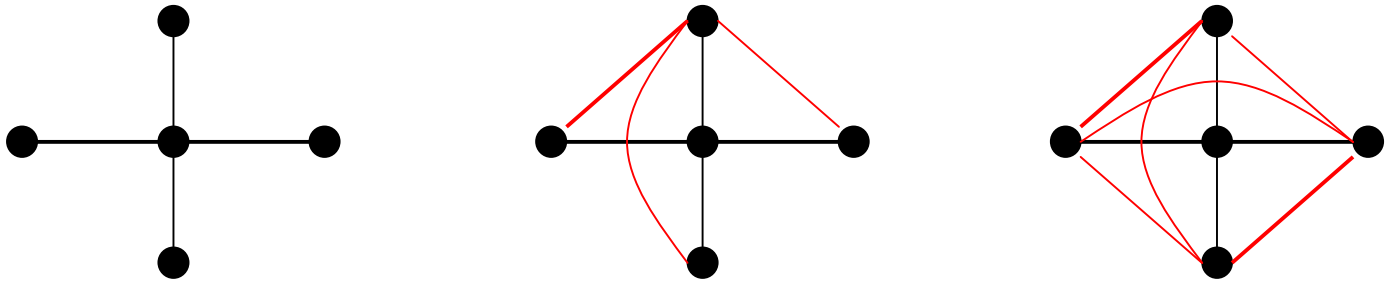
(fraction des voisins connectés entre eux)

Le modèle de graphe aléatoire $G_{n,p}$

- Graphe à *n* sommets;
- Pour toute paire de sommets *(i,j):*
  - On met une arête avec probabilité p.
- Comportement « moyen »:
  - Environ *m* = *np* arêtes;
  - Si *p/n = o(log n),* graphe non connexe;

Densité locale – Coefficient de regroupement C

- d_u = degré sommet u
- $C(u) = 2 \dfrac{\#\ \text{liens entre voisins de } u}{d_u(d_u - 1)}$



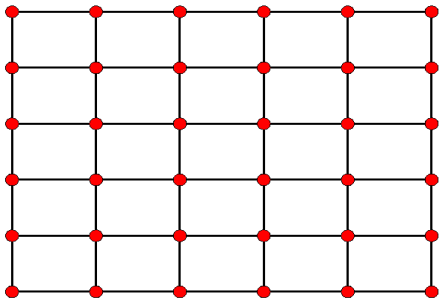
C=0 C=1/2 C=1

Grosso-modo, *C* estime une proportion de triangles (on peut proposer des variantes).

Comportement asymptotique

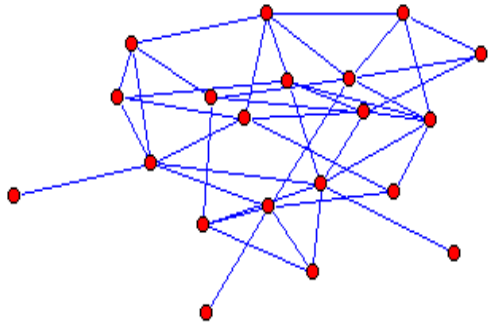
Grille ou variante

$$L(N) = N^{1/2}$$
$$C(N) \approx const\ .$$



Graphe aléatoire

$$L(N) = \log N$$
$$C(N) \approx N^{-1}$$

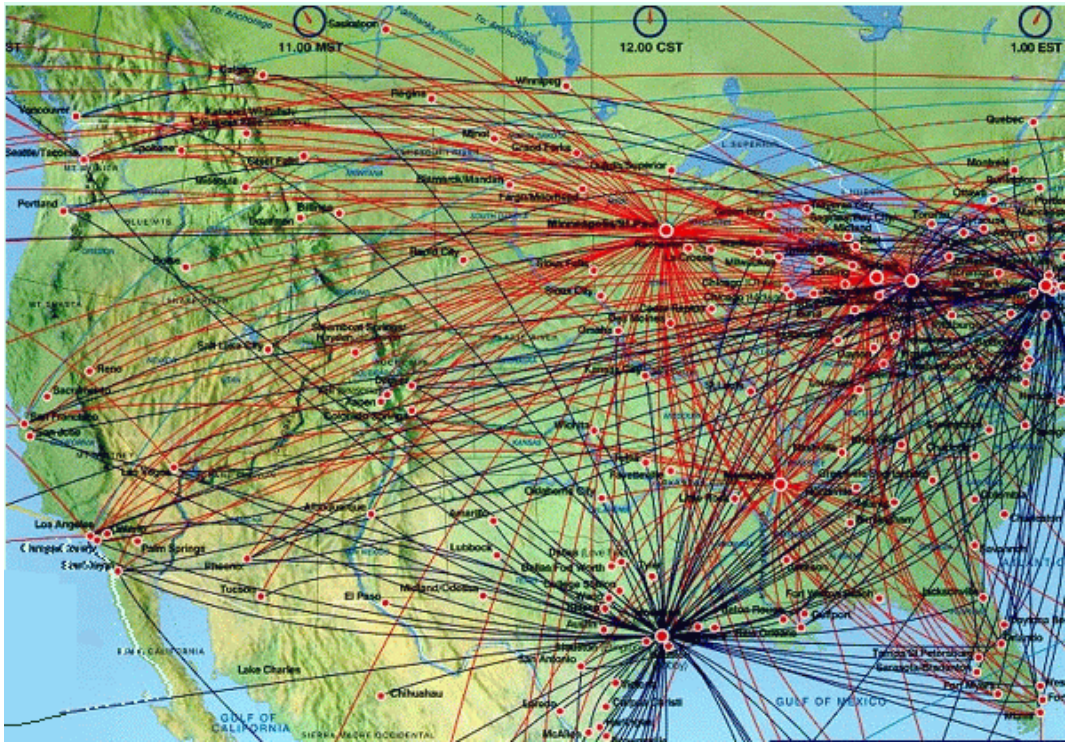


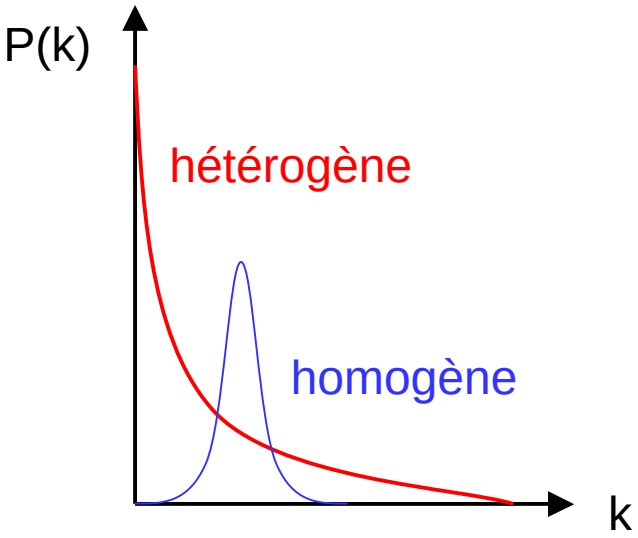
Comparaison entre mesures et modèle $G_{n,p}$

Réseaux	*L mesurée*	*L - $G_{n,p}$*	*C mes.*	*C - $G_{n,p}$*	*n*
WWW	3.1	3.35	0.11	0.00023	153127
Hollywood	3.65	2.99	0.79	0.00027	225226
Electrique	18.7	12.4	0.08	0.005	4014
C. Elegans	2.65	2.25	0.28	0.05	282

$G_{n,p}$ n'est pas un bon de modèle de graphe petit-monde (coef *C* trop faible)

Distribution des degrés

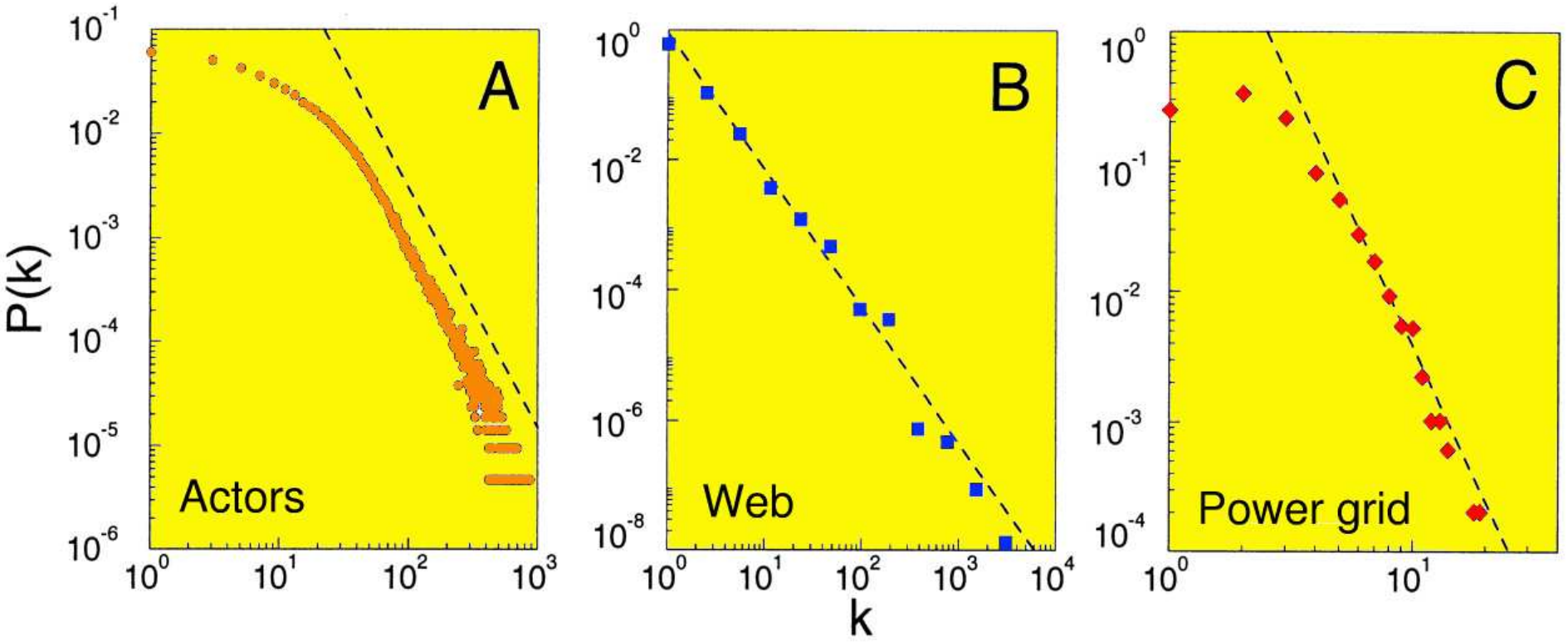




homogène: comme dans $G_{n,p,}$ - loi de poisson (concentré autour de la moyenne)

Hétérogène: loi en puissance – $Prob(d_u = k) \approx 1/k^{\alpha}$ avec $\alpha \in [2,3]$

Réseaux hétérogènes (sans échelle)



Noeuds:	Acteurs	Pages Web	Stations elec.
Liens:	Films	Hyper-liens	Lignes HT

Barabasi & Albert, *Science* **286**, 509 (1999)

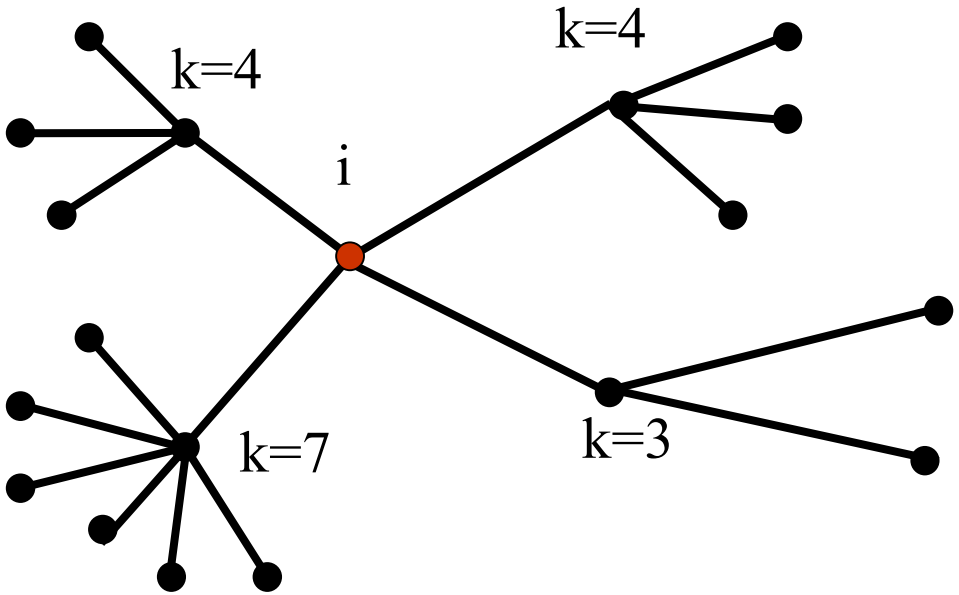
Corrélation degré et dégrés du voisinage: assortativité

$$\mathbf{k_{nn,i}} = \frac{\mathbf{1}}{\mathbf{k_i}} \mathbf{\Sigma_j}\, \mathbf{a_{ij} k_j}$$

Degré moyen de mes voisins – arêtes pondérés

$$\mathbf{k_{nn}(k)} = \frac{\mathbf{\Sigma_i}\, \delta(\mathbf{k_i} - \mathbf{k}) \mathbf{k_{nn,i}}}{\mathbf{\Sigma_i}\, \delta(\mathbf{k_i} - \mathbf{k})}$$

Assortativité du graphe



$k_i=4$

$k_{nn,i}=(3+4+4+7)/4=4.5$

Assortativité

- Comportement assortatif: $k_{nn}(k)$ croissant
  - Exemple : réseaux sociaux (popularité)
  - Gros sites reliés aux gros sites.
- Comportement non-assortatif: $k_{nn}(k)$ décroissant
  - Exemple: Internet
  - Gros sites reliés à des petits sites – structure hiérarchique

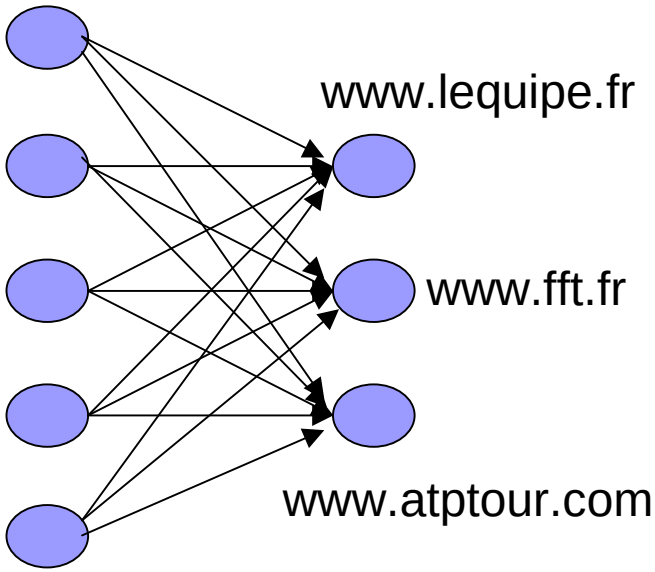
Autres paramètres à étudier

- Excentricité du sommet: distance du sommet le plus éloigné
- Centralité d'un sommet: distance moyenne aux autres sommets.
- « Betweenness »: nombre de plus courts chemins passant par une arête.

Paramètres importants pour des heuristiques de partitionnement !

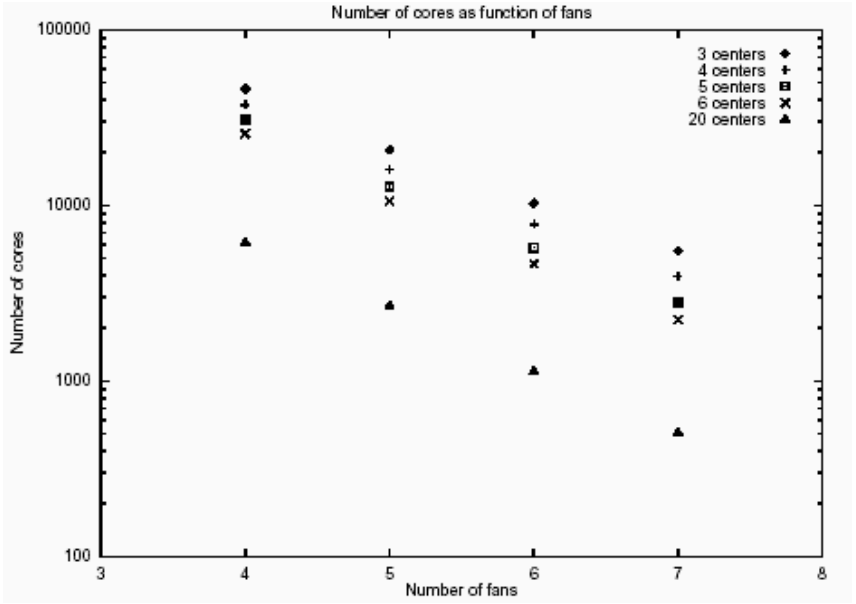
Communités dans les petits mondes

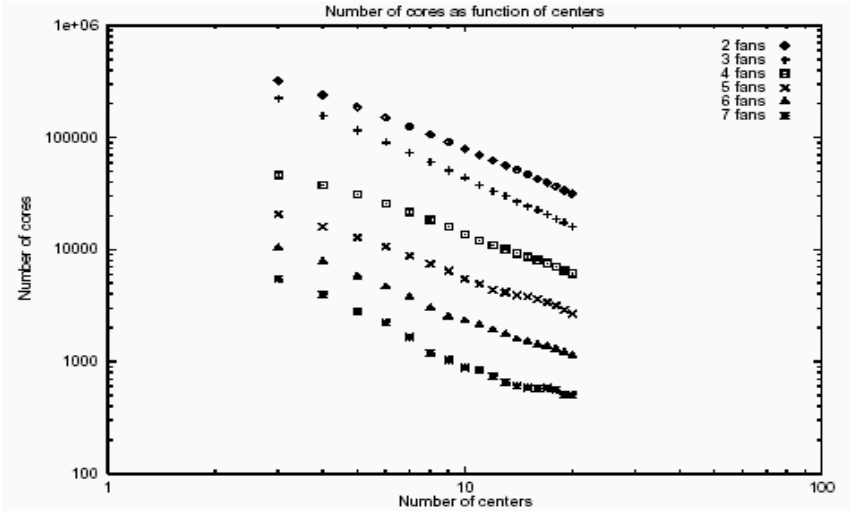
- Partionnement +/- naturel
- Communauté idéale: clique (C=1)
- En pratique, une communauté $C_{i,j}$ nommée « core » contient des petits graphes bipartis complets $K_{i,j}$:
  - i sortants « fans »;
  - j entrants « centers »;



$K_{5,3}$

Petits “cores” bipartis





Nombre de core C_{ij} comme fonction de $i.j$

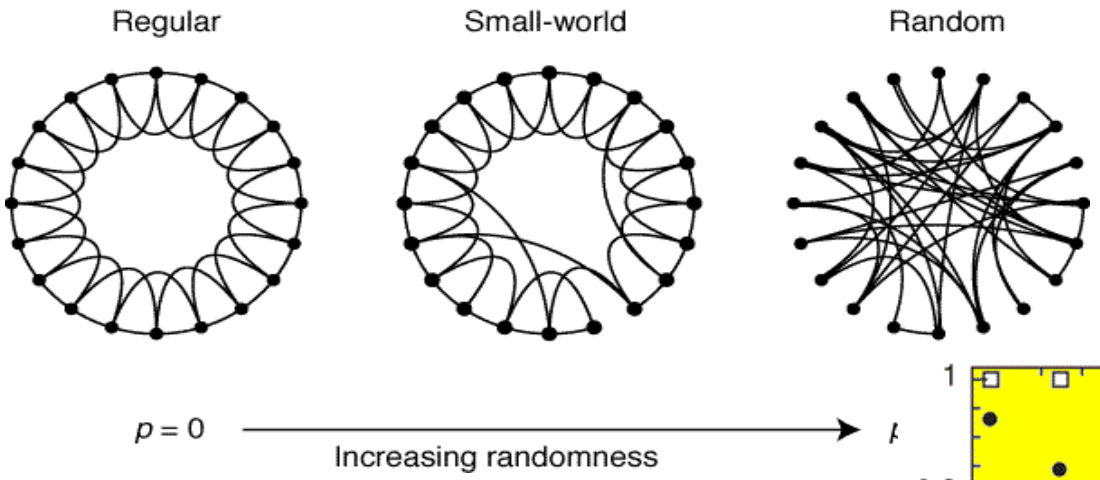
Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew Tomkins. Extracting large-scale knowledge bases from the web.2000.

Quelques modèles de graphes petits mondes

Différents modèles de graphes aléatoires:

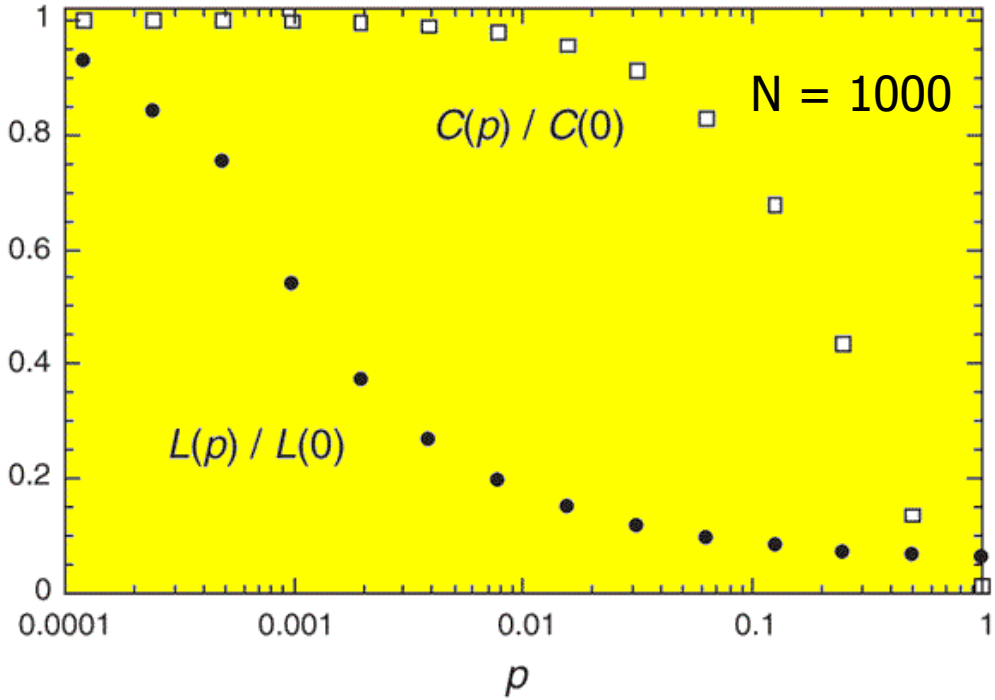
- Redirection d'arêtes;
- Graphes augmentés;
- Graphes évolutifs;
- Ad-hoc:
  - distribution de degrés fixés (par exemple, selon loi en puissance), arbres de cliques, …

Redirection d'arêtes



N noeuds forment un graphe régulier. Avec probabilité p, chaque arête est redirigée aléatoirement.

=>Création de raccourci

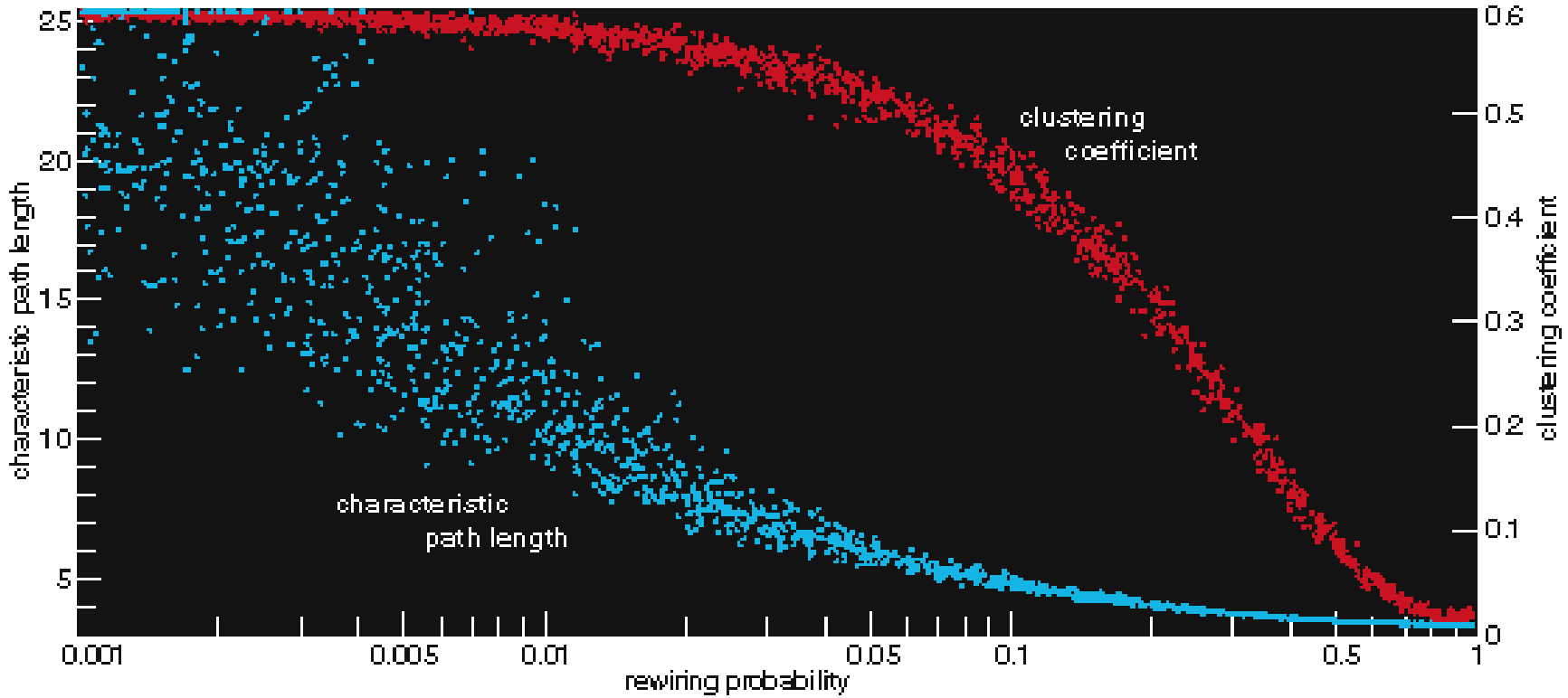


- Grand coef. Reg. C
- petite Long. chemins L
- mais graphe homogène

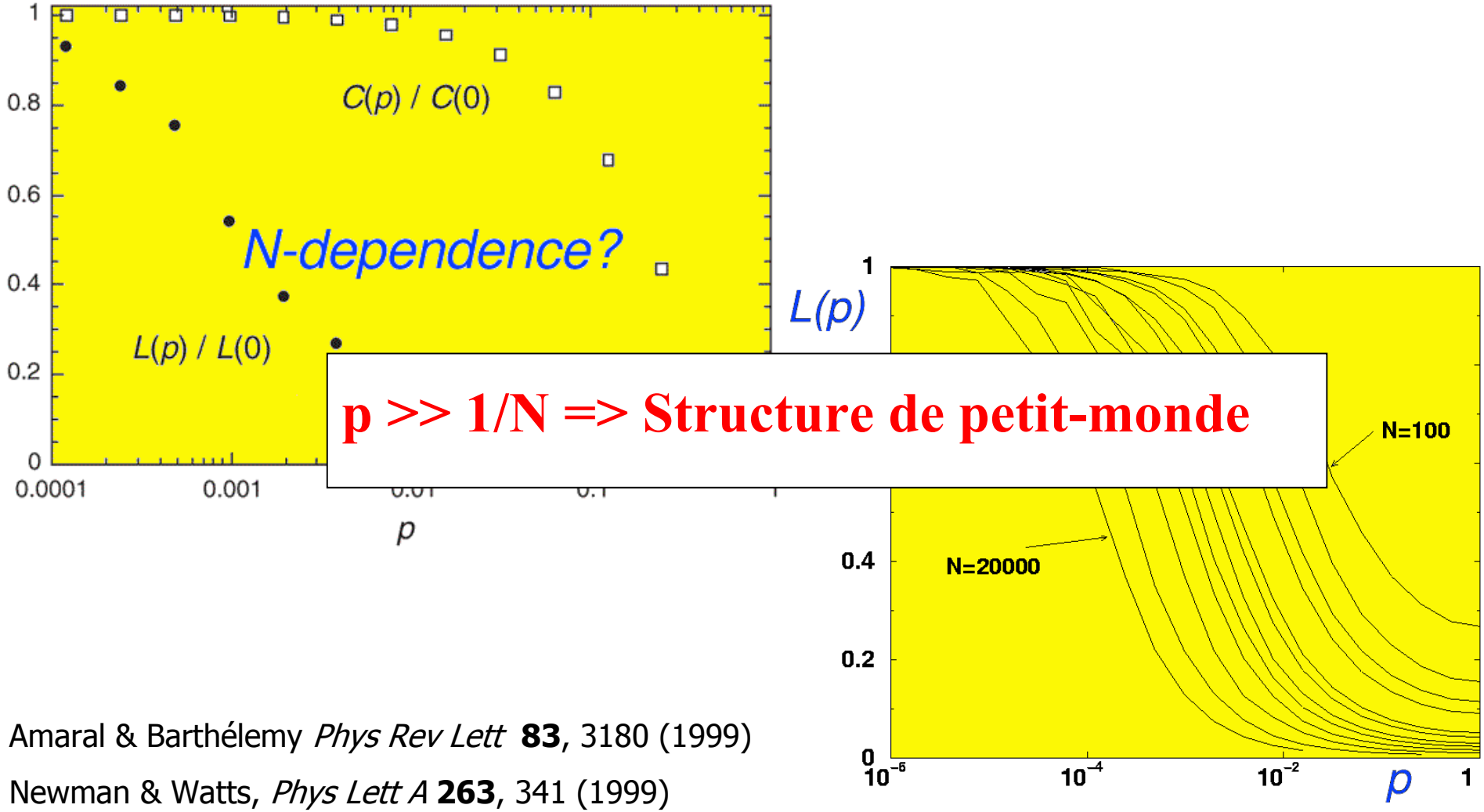
Watts & Strogatz,

Nature **393**, 440 (1998)

Relation longueur moyenne L et coefficient de regroupement C



Dépendance par rapport à la taille



p >> 1/N => Structure de petit-monde

Amaral & Barthélemy *Phys Rev Lett* **83**, 3180 (1999)

Newman & Watts, *Phys Lett A* **263**, 341 (1999)

Barrat & Weigt, *Eur Phys J B* **13**, 547 (2000)

Observations importantes

(1) Le nombre de noeuds (N) n'est pas fixé.

Les réseaux grandissent continuellement avec l'ajout de nouveaux noeuds.

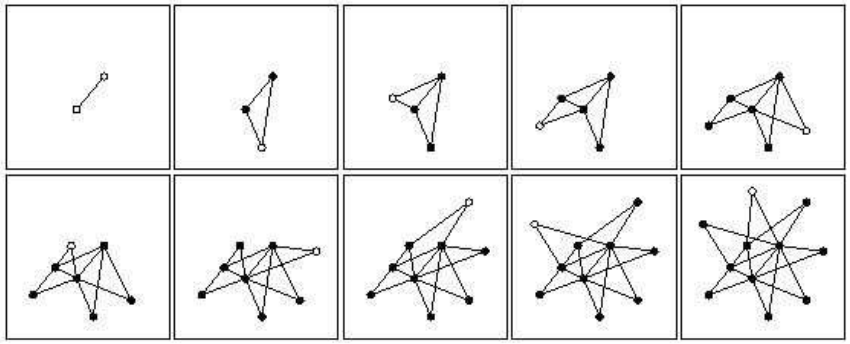
Exemples:
WWW :addition de nouvelles pages
Citation : publication d'articles

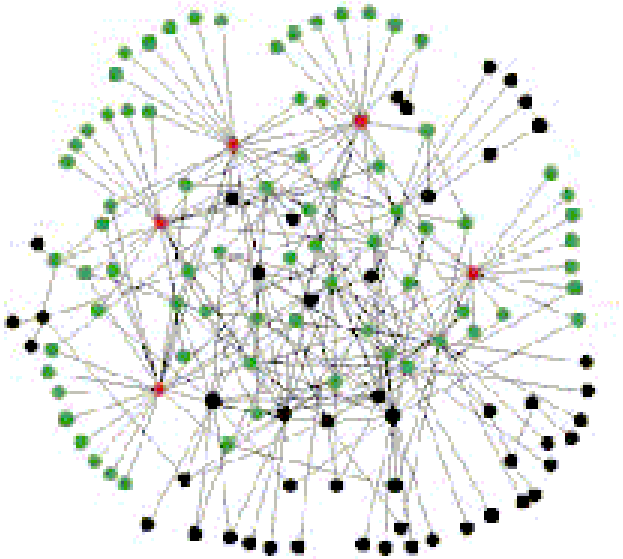
(2) L'attachement n'est pas uniforme.

Un noeud se connecte avec grande probabilité avec des noeuds ayant déja un grand nombre de liens.

Exemples :
WWW : nouvelles pages se connectent à des sites bien connus (CNN, YAHOO, NewYork Times, etc)
Citation : articles populaires ont tendance à être à nouveau cité.

Modèle évolutif probabiliste

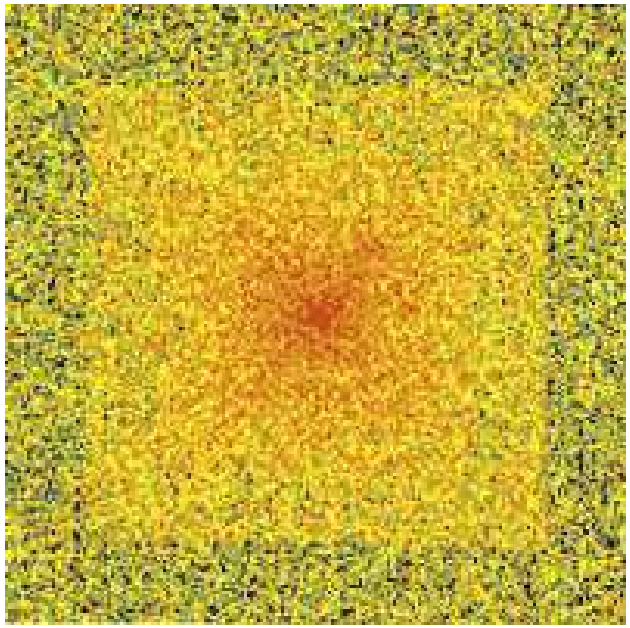


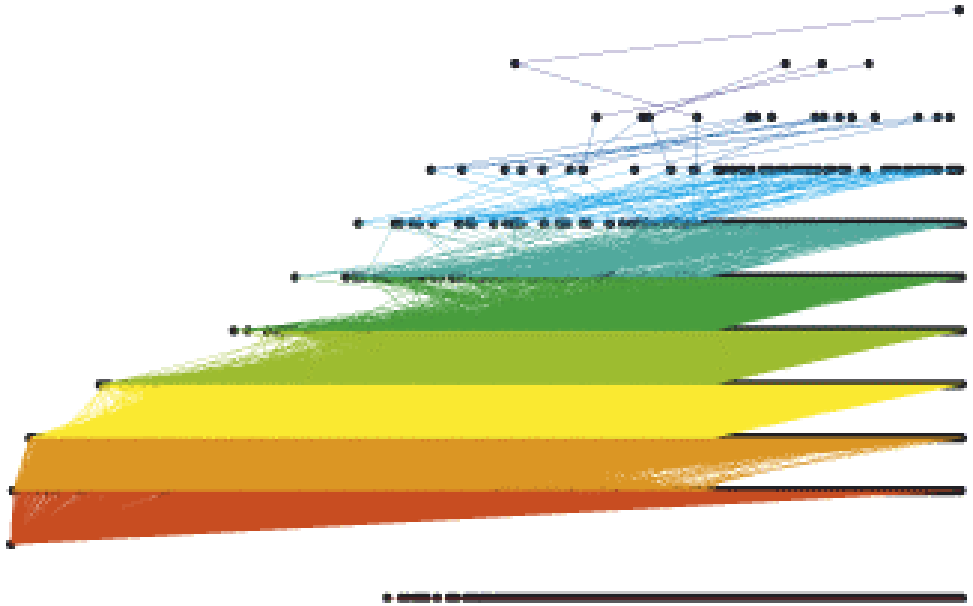


- Graphe germe
- Attachement préférentiel des nouveaux noeuds aux anciens selon les degrés.
  - Prob(attach au noeud j) proportionel au degré du noeud j
  - Prob(degree = k) = a/k^3

En partant d’une distribution des degrés ayant une loi en puissance

- en plaçant en spirale en commençant par les nœuds de plus fort degrés.
- Les Couleurs désignent la distance au sommet de plus fort degrés

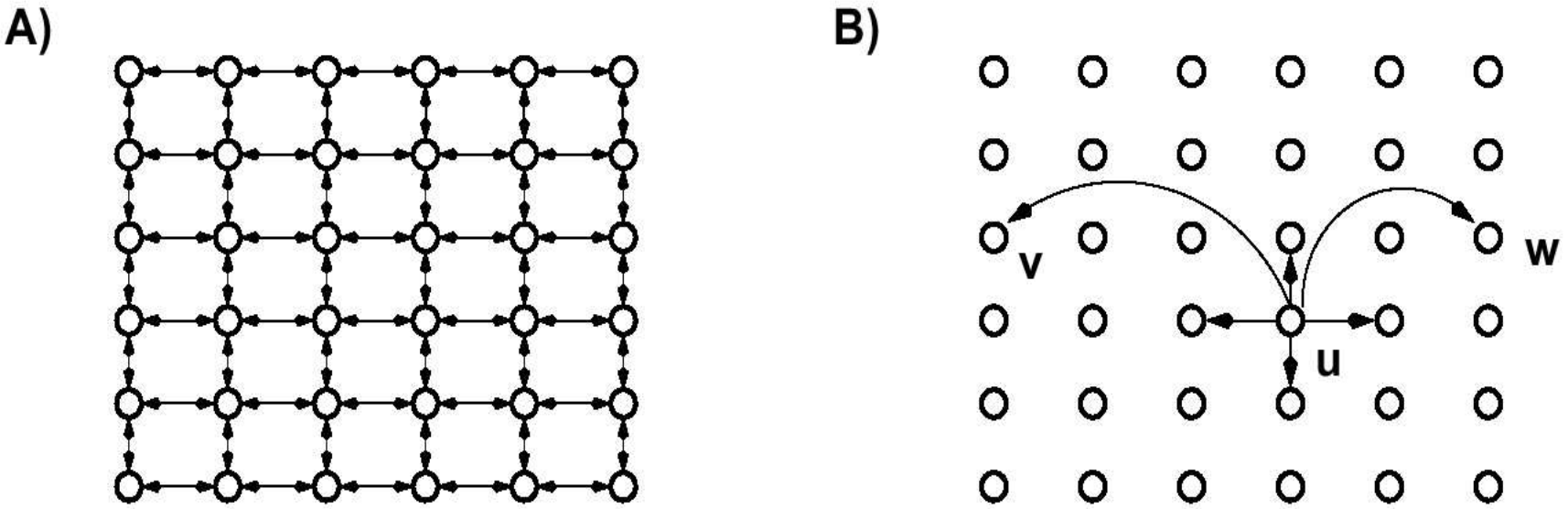




Obtention de petit diamètre

Graphe augmenté (G,L)

- graphe déterministe G + liens aléatoires L



- Kleinberg (D,s) : grille dim. D + Prob(L_u=v) ≈ 1/dist(u,v)^s
- routage glouton basé sur le voisinage (cf. Milgram)

Conséquences algorithmiques

Le graphe est petit-monde …
et alors ?

Conséquences algorithmiques

- Applications:
  - Routage, recherche information;
  - Partitionnement;
  - Tolérance aux pannes;
  - Diffusion d’information, propagation d’épidémie et rumeurs;
  - Visualisation de petits-mondes.
- Contraintes:
  - Algorithmique SIMPLE et UNIVERSEL dédié aux grandes masses de données;
  - Temps: sub-quadratique requis;
  - Espace mémoire: sub-quadratique requis.

Routage et Diamètre du graphe de Kleinberg

- Pour un graphe à *n* sommets et 1 lien longue distance par sommet.

Mesure	$0 \leq s < D$	$s = D$	$D<s<2D$	$s = 2D$	$s > 2D$
Routage	n^{α}	$log^2 n$	n^{α}	n^{α}	n^{α}
Diamètre	*? log n*	*log n*	$log^{\beta} n$	*?*	n^{β}

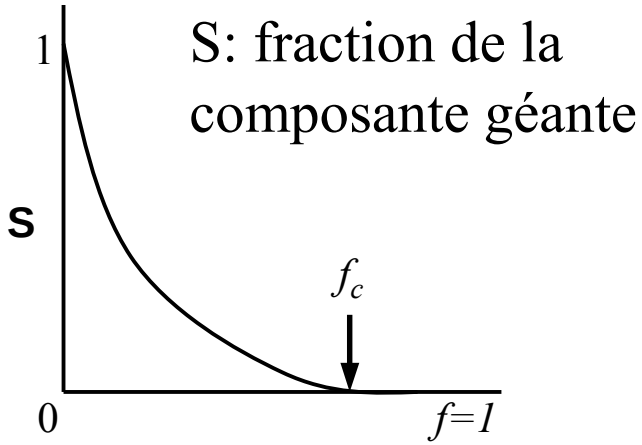
Routage (Kleinberg) et Diamètre (Martel-NGuyen pour s=D et Duchon-Hanusse-Simonklein)

•Routage glouton efficace pour s=D;

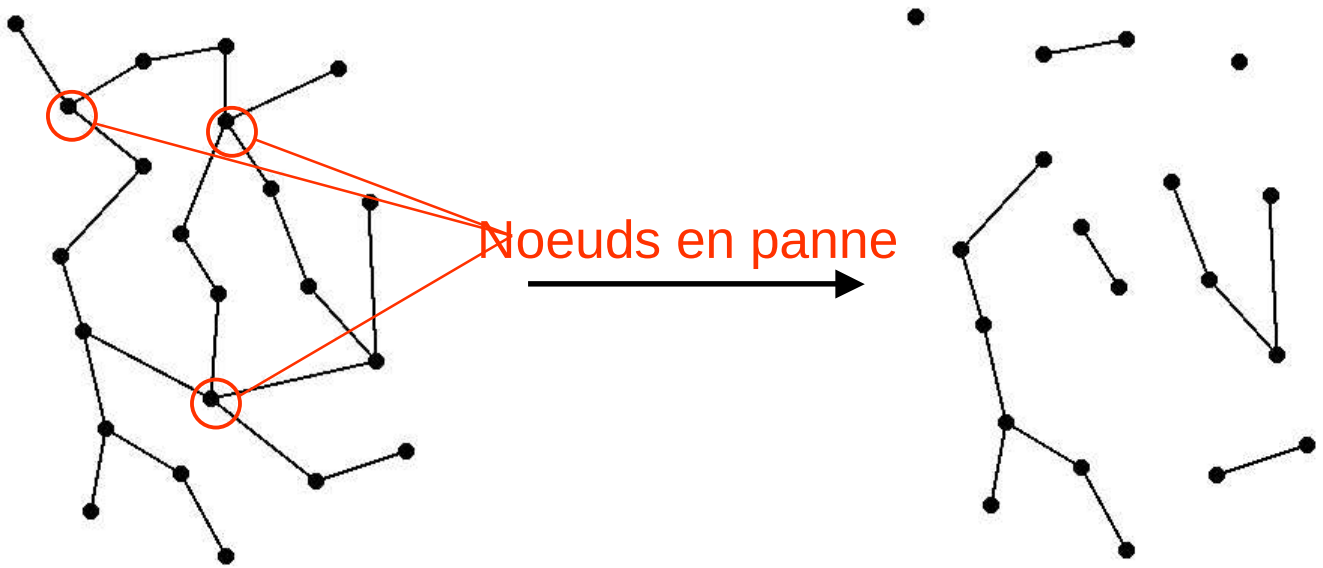
•Rajouter des liens longue distance ne changent pas vraiment les résultats.

Robustesse

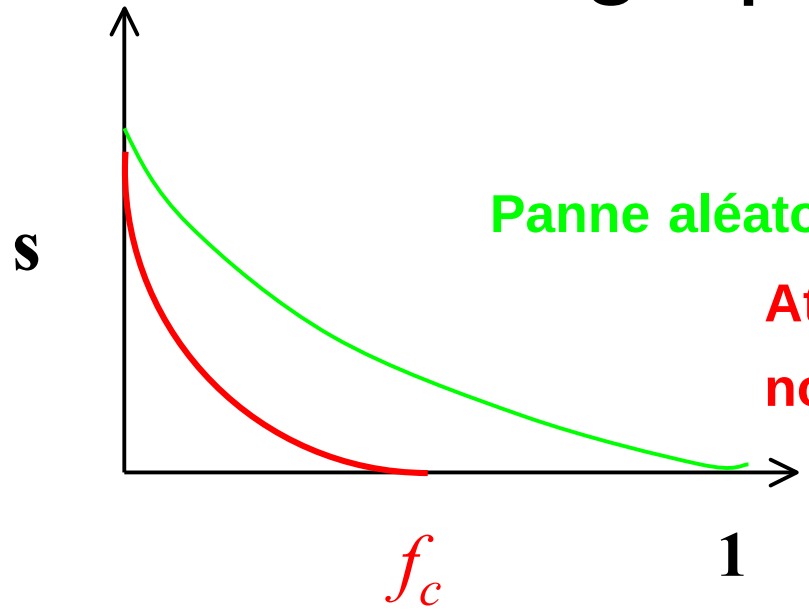
- Les graphes petits-mondes préservent leur structure, propriétés même en présence d'erreurs ou de pannes:
  - cellules $\rightarrow$ mutations;
  - Internet $\rightarrow$ pannes de routers;



Fraction des noeuds supprimés, f



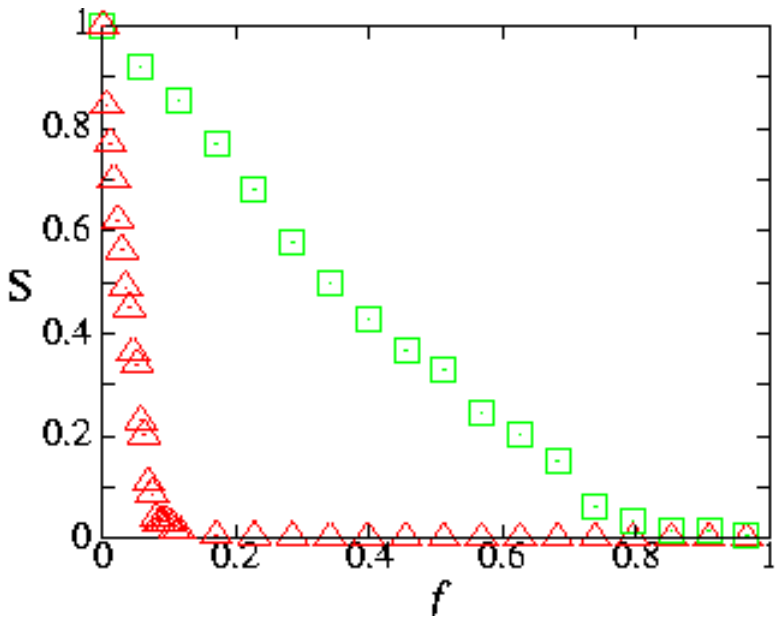
Les cas des graphes sans-échelle



Panne aléatoire $\rightarrow f_c = 1$ $\quad (2 < \gamma \leq 3)$

Attaque =panne progressive des noeuds de plus haut degrés $\rightarrow f_c < 1$

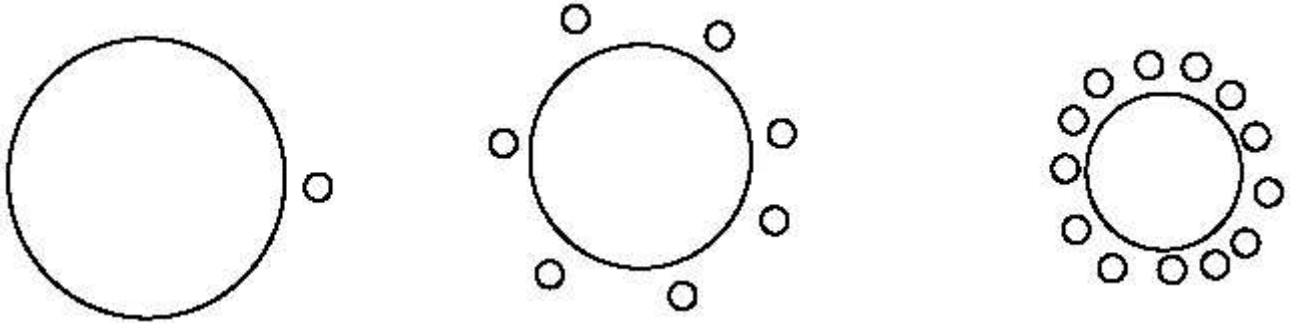
Internet

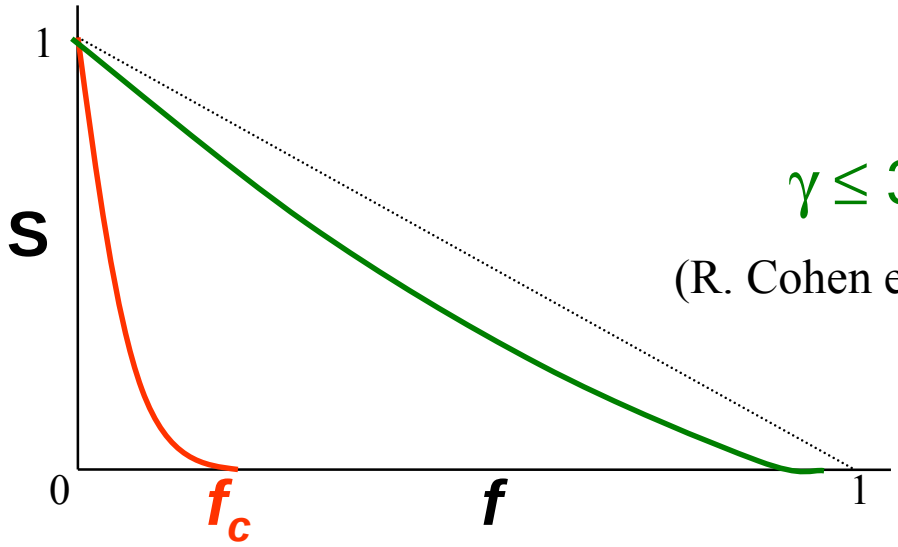


R. Albert, H. Jeong, A.L. Barabasi, Nature **406** 378 (2000)

Comparaison Pannes et Attaques

Pannes

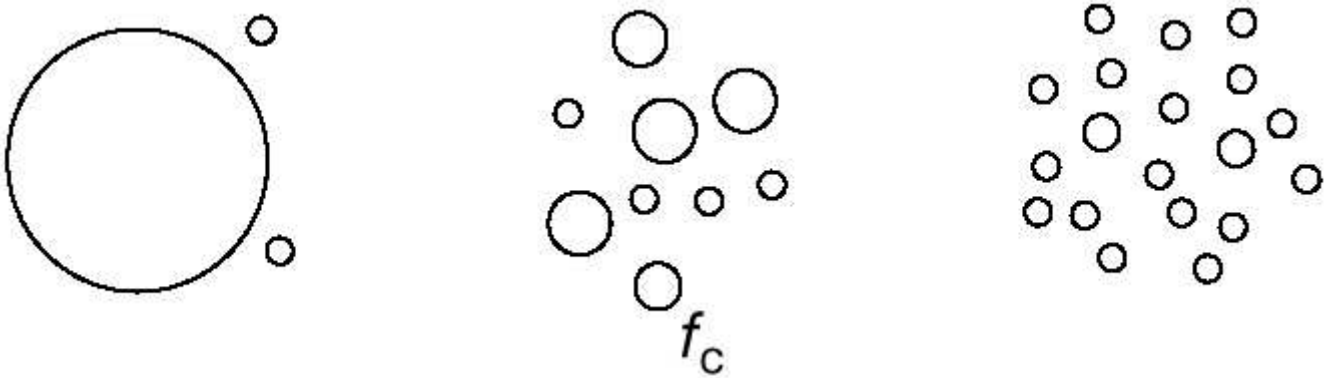




$\gamma \leq 3 : f_c=1$

(R. Cohen et al PRL, 2000)

Attaques



Autres stratégies d’attaques

- Elimination de noeuds ou arêtes avec la plus grande centralité
- Noeuds ou arêtes avec la plus grande “betweenness” Cascades
- Cascades: élimination de sous-graphes ...

Complexité pour le calcul de la centralité d'un sommet

- centralité(i) = $\sum_j \frac{dist\,(i,\,j)}{n-1}$

- Calcul des plus courts chemins pour toute paire se fait en temps: $O(n^3)$ ou $O(nm+n^2 \log n)$.
- Pour des graphes à plusieurs millions d'éléments, même si m = O(1), trop long
- Et si on veut simplement approximer la centralité ?

Approximer la centralité à un facteur 1+ε en temps O(m log n)

Algorithme RAND:

- Soit *k* un nombre d'itérations paramétrant l'erreur de mesure;
- A l'itération *i*, choisir au hasard un sommet v_i et résoudre le problème du plus court chemin avec v_i comme source;
- Cent(*u*) = $\dfrac{kn \sum_{i=1}^{k} dist\,(v_i, u)}{n-1}$

Eppstein et Wang, « Approximating Centrality » 2004

Et bien d’autres pistes

- Recherche applicatives:
  - Algorithmes de visualisation dédié aux petits-mondes
  - Recherche d’information:
    - Image, vidéo: descripteurs = dimension (cf. modèle Kleinberg)
    - Réseaux pairs à pairs
  - Diffusion et/ou limitatation de virus, information
- Plus formellement:
  - Petit-monde: beaucoup d’intéractions locales + qq intéractions globale
  - Peut-on transformer tout graphe en petit-monde avec peu de modifications