- Certaines propriétés sont monotones en fonction de p
- Par exemple : A = le graphe est sans triangle
- Plus on ajoute des arêtes plus il y a de chances d'avoir un triangle
- Soit X la variable aléatoire du nombre de triangles.
- $E(X) = \binom{n}{3}p^3$
- En posant $\lambda = pn$ on obtient : $lim_{n\to\infty}\binom{n}{3}p^3 = \frac{\lambda^3}{6}$.
- avec $\lambda = 10^4$ il existera des triangles avec une forte probabilité
- et inversement lorsque $\lambda = 10^{-4}$ il n'en existera pas avec une forte probabilité.
- Le nombre de V est en moyenne $n\binom{n}{2}p^2 = n\frac{\lambda^2}{2}$. Le clustering global est donc de l'ordre de $\frac{\lambda}{n} = p$.

Phénomène de seuil

Définition

$r(n)$ est un **seuil** pour une propriété de graphe A si :

1. $p(n) \ll r(n)$ implique $lim_{n\to\infty} P[\mathcal{G}(n,p) \models A] = 0$
2. $p(n) \gg r(n)$ implique $lim_{n\to\infty} P[\mathcal{G}(n,p) \models A] = 1$

Phénomène de seuil

Définition

$r(n)$ est un **seuil** pour une propriété de graphe A si :

1. $p(n) \ll r(n)$ implique $lim_{n\to\infty} P[\mathcal{G}(n,p) \models A] = 0$
2. $p(n) \gg r(n)$ implique $lim_{n\to\infty} P[\mathcal{G}(n,p) \models A] = 1$

Seuil et triangles

- $r(n) = 1/n$ est une fonction de seuil pour avoir un triangle
- Il n'y a pas d'unicité de la fonction de seuil : $r(n) = 10/n$ marche aussi pour avoir un triangle.
- $\mathcal{G}(n, \frac{1}{n \log n})$ est sans triangle avec une forte probabilité.

Transition de phases

Une vision dynamique de $\mathcal{G}(n,p)$

En faisant croître p, on regarde ce que deviennent les propriétés du graphe. On observe comme en physique des phénomènes de **transition de phase** et de **percolation**

Des recherches de physiciens

Ce que l'on peut étudier rigoureusement...

...même si ce n'est pas facile !

Diamètre

On a vu que...

Pour p fixé et $n \to \infty$ le diamètre est 2

Et si p est fonction de n ?

- Pour $p\frac{n}{\log n} \to \infty$ le diamètre est presque sûrement constant **constant** (avec proba 1).
- Pour $p\frac{n}{\log n} = c > 1$ le diamètre est p. s. **sous-logarithmique** (de l'ordre de $\frac{\log n}{\log np} = \Theta(\frac{\log n}{\log \log n})$).
- Pour $np \geq c > 1$, le diamètre de la composante géante est p. s. **logarithmique** ($\exists f()$ t.q. au plus $\frac{\log n}{\log np} + f(c)\frac{\log n}{np}$).

cf. The Diameter of Sparse Random Graphs [Chung and Lu, 2001.].

Connexité : plusieurs seuils !

- Si $p < \frac{1}{n}$ il y a beaucoup de petites composantes qui sont surtout des arbres
- si $p > \frac{1}{n}$ une **composante géante** émerge
- si $p = \frac{1}{n}$ alors la plus grande composante a (avec forte proba.) taille $\Theta(n^{2/3})$ [Bollobás 2001]
- si $p > \frac{\ln n}{n}$ alors le graphe est (avec forte proba.) connexe
- NB : <u>avec forte probabilité</u> veut dire avec probabilité $> 1 - 1/n$ quand n tend vers ∞. Tandis que <u>presque sûrement</u> veut dire avec probabilité 1. Ce qui ne veut pas dire <u>toujours</u> !

Conclusion

Un modèle peu réaliste

$\mathcal{G}(n, p)$et $\mathcal{G}(n, m)$ne modélisent pas bien les réseaux d'interaction

On peut tout calculer sur les graphes aléatoires $\mathcal{G}(n, p)$

à condition d'être bon en calcul...
Certains en abusent !

Il existe d'autres modèles aléatoires de réseaux

Nous avons déjà vu l'anneau de Watts et Strogatz, nous verrons ~~la semaine prochaine~~ **maintenant** l'attachement préférentiel de Barabási-Albert et la grille de Kleinberg.

Generating Random Regular Graphs J. H. Kim & Van H. Vu [2003] chapitre 2

Modèles de GRI

Section 3

Attachement préférentiel & grille de Kleinberg

Fabien de Montgolfier
fm@irif.fr

18 fevrier 2022

Rappel sur $\mathcal{G}(n, p)$

Le modèle $\mathcal{G}(n, p)$

Il possède 4/9 des propriétés des GRI réels

- degré moyen $\overline{d} = p(n-1)$ mais degrés pas en power-law
- distances moyennes et diamètre logarithmique si p petit
- pas de cœur mais comp. connexe géante (selon valeur de p)
- pas de transitivité, ni de navigabilité, ni de communautés

Comment modéliser mieux

Faire un modèle de création d'arêtes non-indépendantes

- simple
- et qui ressemble à la réalité

Attachement préférentiel : motivation

Motivation

Taille médiane des communes françaises : 450 habitants. Ici à Paris 2 220 445 habitants. Pourquoi ?
Patrimoine médian d'une personne sur cette planète : 3945$. Pourquoi ?

Attachement préférentiel : motivation

Motivation

Taille médiane des communes françaises : 450 habitants. Ici à Paris 2 220 445 habitants. Pourquoi ? **Les gens s'établissent plus dans les (rares) grandes villes que dans les (nombreux) villages**
Patrimoine médian d'une personne sur cette planète : 3945$.
Pourquoi ? **Il est plus facile à un riche d'augmenter sa fortune de 1$ qu'à un pauvre**

rich get richer

Attachement préférentiel : motivation

Motivation

Taille médiane des communes françaises : 450 habitants. Ici à Paris 2 220 445 habitants. Pourquoi ? **Les gens s'établissent plus dans les (rares) grandes villes que dans les (nombreux) villages**
Patrimoine médian d'une personne sur cette planète : 3945$. Pourquoi ? **Il est plus facile à un riche d'augmenter sa fortune de 1$ qu'à un pauvre**

rich get richer

Origine du modèle d'attachement préférentiel

- Introduit par Derek de Solla Price en 1970 pour modéliser le graphe des citations entre articles scientifiques.
- Retrouvé et nommé par Albert-Laszlo Barabási [2001]. Très cité !

Attachement préférentiel : description

Le modèle possède trois paramètres : $d \leq n_0 \leq n$

- On part d'une clique de n_0 sommets tous reliés entre eux
- A chaque étape i on ajoute un sommet x_i qui est connecté à d anciens sommets avec un probabilité p_i proportionnelle à son degré dans le graphe précédent. Ainsi pour $j < i$ l'arête $x_j x_i$ est créée avec probabilité

$$\frac{deg(x_j)}{\Sigma_{x \in G} deg(x)}$$

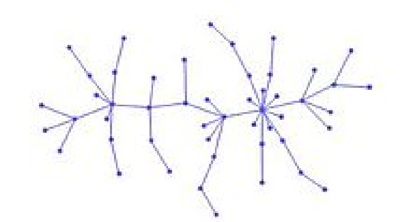
- i va de $n_0 + 1$ à n. On ajoute donc $n - n_0$ sommets et chaque sommet ajouté a degré d. Le degré moyen du graphe si $n \gg n_0$ est $\overline{d} \simeq 2d$

Propriétés de l'attachement préférentiel

On obtient une **power-law** de paramètre $\gamma = 3$ (noter que le degré minimum est d ce qui est inhabituel...) Le graphe est bien sûr **connexe** et a un **cœur** de sommets bien connectés : ceux de petit numéro.

À part les communautés et la navigalibité (propriétés facultatives) on a toutes les propriétés des GRI (pour la navigabilité, il faudrait rajouter une distance...)

Si $d = 1$:
arbres scale-free



Une variante du modèle avec **ajout d'arête**

Ce modèle a deux paramètres : le nombre final de sommets n et une probabilité p. À chaque étape

- avec proba p on ajoute un sommet relié à **un** autre, tiré selon son degré
- et avec proba $1-p$: on tire deux nœuds (selon leur degré) non reliés et on ajoute une arête entre eux

Théorème : degrés en power-law avec $\gamma = 1 + \frac{2}{2-p}$

Une troisième variante du modèle avec **densification**

Avin, Lotker, Nahum, Peleg [2015]

- avec proba $p(t)$ on ajoute un sommet relié à **un** autre, tiré selon son degré
- et avec proba $1 - p(t)$: on tire deux nœuds (selon leur degré) non reliés et on ajoute une arête entre eux

$p(t) \rightarrow 0$ avec le temps $t \rightarrow \infty$

Comparaison entre ajout d'arête et densification

Au cours du temps si $p(t)$ diminue la densité augmente (puisqu'on ajoute moins de sommets, on ajoute plus d'arêtes) ce qui est effectivement observé dans certains réseaux réel.

La densification produit un cœur plus petit (de taille $o(n)$ et non $\Theta(n)$ comme dans les deux précédents modèles). Plus précisément

	p constant	$p(t)$ variable $\to 0$
nb arêtes $m(t) \simeq$	$n(t)/p$	$n(t)\ln(n(t))$
taille cœur $\simeq$	$\alpha n(t)$ $\alpha =$cste	$n(t)^{\beta}$ $\beta =$cste
power law	$\gamma = 1 + \frac{2}{2-p}$	$\gamma = 2$

D'autres modèles encore

Sujet à la mode donc beaucoup de variantes, de moins en moins simples

- Orienté
- Ajout des sommets par paquets et non un par un
- Disparition de sommets ou d'arêtes
- Préférence selon d'autres paramètres que le degré : PageRank, centralité...

Conclusion sur l'attachement préférentiel

- Moralité : «Rich get Richer» ou encore «ça tombe toujours sur les mêmes»
- C'est un modèle purement topologique : il n'utilise pas de connaissance sur les auteurs, le contenu des articles...
- Le modèle est assez bon pour les citations scientifiques car la base de données est monotone croissante (mais il n'y a pas de thématiques)
- Un peu moins bon pour le graphe du Web, car pas de dynamique
- Avec l'attachement préférentiel il y a bien un **cœur** dense (plus fort degré et meilleure connectivité interne)... composé des plus **anciens** nœuds du réseau. Est-ce ainsi dans la vraie vie ?

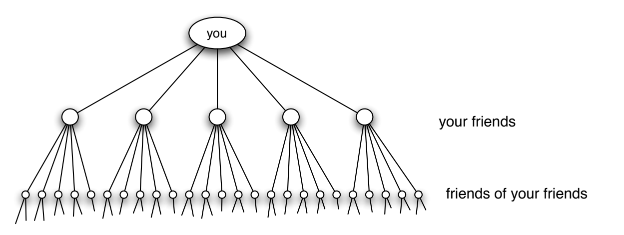
Le livre de Kleinberg

Il est disponible en ligne !

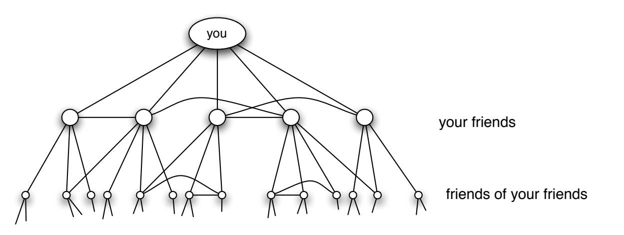
https://www.cs.cornell.edu/home/kleinber/networks-book/

Il contient des tas de choses passionnantes

Par exemple le chapitre 20 «The Small-World Phenomenon» contient les figures suivantes :



(a) *Pure exponential growth produces a small world*



(b) *Triadic closure reduces the growth rate*

Figure 20.1: Social networks expand to reach many people in only a few steps.

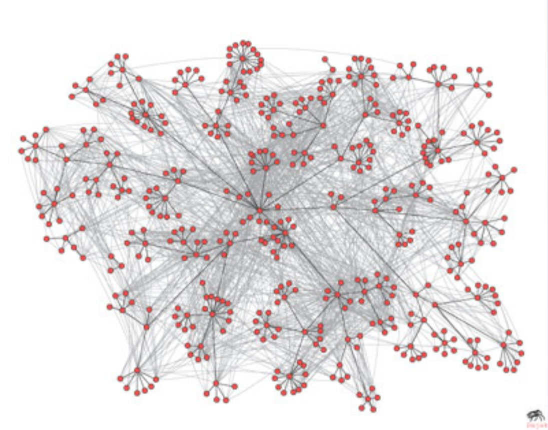


Figure 20.12: The pattern of e-mail communication among 436 employees of Hewlett Packard Research Lab is superimposed on the official organizational hierarchy, showing how network links span different social foci [6]. (Image from http://www-personal.umich.edu/ ladamic/img/hplabsemailhierarchy.jpg)

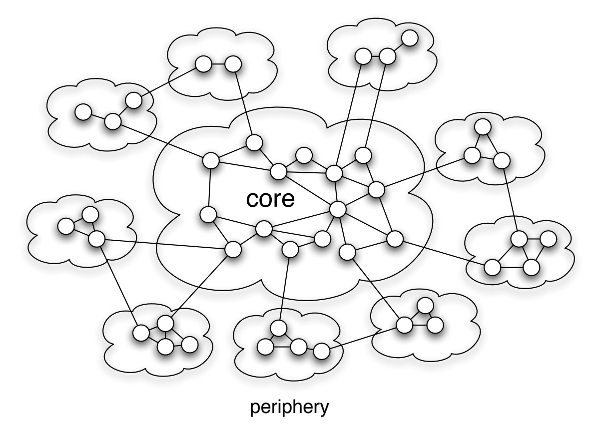


Figure 20.13: The core-periphery structure of social networks.

20.6 Core-Periphery Structures and Difficulties in Decentralized Search

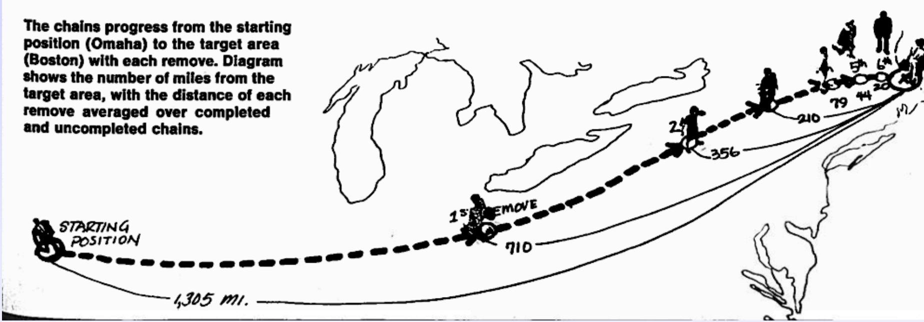


Figure 20.4: An image from Milgram's original article in *Psychology Today*, showing a "composite" of the successful paths converging on the target person. Each intermediate step is positioned at the average distance of all chains that completed that number of steps. (Image from [297].)

Revenons au routage glouton

Le problème

Atteindre une cible dont on connaît les coordonnées

Le modèle

Un graphe. Un nœud ne connaît que ses voisins. Décisions locales

Revenons au routage glouton

Le problème

Atteindre une cible dont on connaît les coordonnées

Le modèle

Un graphe. Un nœud ne connaît que ses voisins. Décisions locales

La solution

Envoyer le paquet au voisin **le plus proche** de la cible

Différence avec le routage ordinaire

Le routage glouton transmet le paquet au voisin le plus proche

N'est-ce pas ce que font les algos vus en M1 : RIP, OSPF... ?

Différence avec le routage ordinaire

Le routage glouton transmet le paquet au voisin le plus proche

N'est-ce pas ce que font les algos vus en M1 : RIP, OSPF... ?

non ! C'est un cas particulier. Ils calculent une distance **dans le graphe** et le paquet suit un plus court chemin.
C'est un routage **optimal** !

Attention à la définition de «le plus proche»

Dans le routage glouton cela veut dire celui dont **les coordonnées** sont les plus proches de celles de la cible

Métaphore routière

- Je suis en voiture
- Je connaît les coordonnées de ma destination (latitude, longitude) et les miennes
- Je sais donc dans quelle direction aller.
- Dois-je aller dans cette direction ?

Métaphore routière

- Je suis en voiture
- Je connaît les coordonnées de ma destination (latitude, longitude) et les miennes
- Je sais donc dans quelle direction aller.
- Dois-je aller dans cette direction ? non !
- Si je dois aller au Nord, aller au Sud prendre l'autoroute peut être pertinent
- De plus les **impasses** font échouer le routage routier glouton

Discussion sur le routage glouton

Il faut des coordonnées

Système à **une** dimension (anneau de Watts et Strogatz), **deux** dimensions (la planète Terre, la grille de Kleinberg) ou **multidimensionnel** (le monde social)

On peut échouer

- Impasses routières
- La pluspart des lettres de Milgram n'arrivent pas

On peut être très mauvais

L'absence de vue globale peut conduire à de mauvaise décision (prendre le chemin de terre au Nord au lieu de l'autoroute au Sud)

Navigabilité

Définition

Un réseau est **navigable** si le routage glouton réussit entre deux sommets quelconques avec probabilité «raisonnable»

Proximité

Il faut de plus une notion de **proximité** pour pouvoir router

- soit des coordonnées
- soit une évaluation empirique de la distance à la cible (comme dans l'expérience de Milgram)

Discussion sur le routage glouton

Inconvénient

- Pas de garantie de performance
- Pas même de garantie de correction ni terminaison

Avantages

- Un processus **décentralisé** (décisions locales)
- Fonctionne avec une connaissance très partielle (que les voisins)

Il existe des variantes : Decentralized Search, Myopic Search...

Motivation

Kleinberg montre que les anneaux de W-S sont navigables

En plus d'avoir de petites distances moyennes, le routage glouton trouve des chemins courts

Le système de coordonnées est simplement le numéro des sommets

Motivation

Kleinberg montre que les anneaux de W-S sont navigables

En plus d'avoir de petites distances moyennes, le routage glouton trouve des chemins courts

Le système de coordonnées est simplement le numéro des sommets

mais Kleinberg le trouve trop compliqué

Que garder pour la navigabilité ?

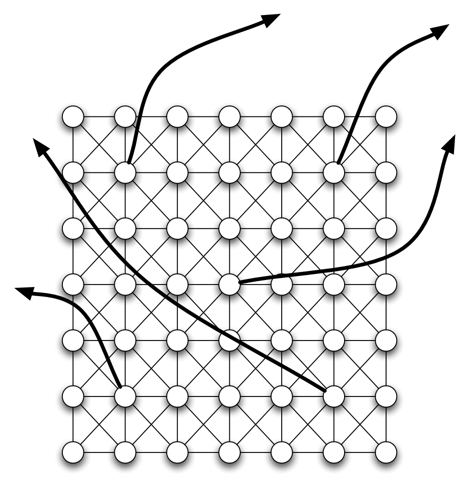
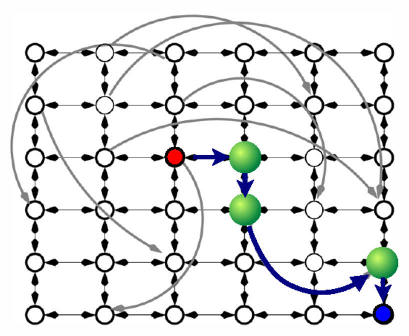


Figure 20.3: The general conclusions of the Watts-Strogatz model still follow even if only a small fraction of the nodes on the grid each have a *single* random link.

La grille de Kleinberg

- $n \times n$ sommets sur une grille, degrés sortant **5**
- 4 voisins les plus proches
- Un **lien long** de u à v avec proba en $\frac{1}{dist(u,v)^2}$



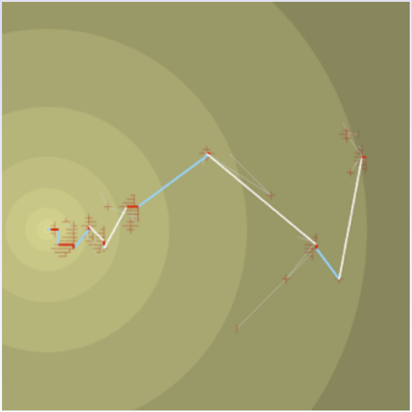
Théorème de navigalibité de Kleinberg

Routage glouton

Pour atteindre une cible de coordonnées (x, y) connues, propager au voisin le plus proche

Nombre de sauts

- Sans lien long $O(\sqrt{n})$ car grille
- Avec liens longs $O(\log^2(n))$



Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.

Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.
- si on suit $\log_2(n)$ bons liens on tombera sur la cible (recherche dichotomique). Or on est à distance D de la cible.

Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.
- si on suit $\log_2(n)$ bons liens on tombera sur la cible (recherche dichotomique). Or on est à distance D de la cible.
- Quelle probabilité que le lien long de (x, y) soit bon ?

Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.
- si on suit $\log_2(n)$ bons liens on tombera sur la cible (recherche dichotomique). Or on est à distance D de la cible.
- Quelle probabilité que le lien long de (x, y) soit bon ?
- Doit tomber dans le disque de diamètre D et de centre la cible

Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.
- si on suit $\log_2(n)$ bons liens on tombera sur la cible (recherche dichotomique). Or on est à distance D de la cible.
- Quelle probabilité que le lien long de (x, y) soit bon ?
- Doit tomber dans le disque de diamètre D et de centre la cible
- Ce disque contient $D^2/2$ points. La proba que chacun soit l'extremité du lien long est c/D^2 à un facteur $(1/2)^2$ à $(3/2)^2$ près.

Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.
- si on suit $\log_2(n)$ bons liens on tombera sur la cible (recherche dichotomique). Or on est à distance D de la cible.
- Quelle probabilité que le lien long de (x, y) soit bon ?
- Doit tomber dans le disque de diamètre D et de centre la cible
- Ce disque contient $D^2/2$ points. La proba que chacun soit l'extremité du lien long est c/D^2 à un facteur $(1/2)^2$ à $(3/2)^2$ près.
- Donc la proba que le lien long tombe dans le disque cible est en $O(c\frac{D^2}{D^2})$, soit $O(c)$.

Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.
- si on suit $\log_2(n)$ bons liens on tombera sur la cible (recherche dichotomique). Or on est à distance D de la cible.
- Quelle probabilité que le lien long de (x, y) soit bon ?
- Doit tomber dans le disque de diamètre D et de centre la cible
- Ce disque contient $D^2/2$ points. La proba que chacun soit l'extremité du lien long est c/D^2 à un facteur $(1/2)^2$ à $(3/2)^2$ près.
- Donc la proba que le lien long tombe dans le disque cible est en $O(c\frac{D^2}{D^2})$, soit $O(c)$.
- c est la constante de normalisation telle que $\sum_D 4D\frac{c}{D^2} = 1$ soit $c = \frac{1}{4\log\sqrt{n}} = \frac{1}{2\log n}$.

Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.
- si on suit $\log_2(n)$ bons liens on tombera sur la cible (recherche dichotomique). Or on est à distance D de la cible.
- Quelle probabilité que le lien long de (x, y) soit bon ?
- Doit tomber dans le disque de diamètre D et de centre la cible
- Ce disque contient $D^2/2$ points. La proba que chacun soit l'extremité du lien long est c/D^2 à un facteur $(1/2)^2$ à $(3/2)^2$ près.
- Donc la proba que le lien long tombe dans le disque cible est en $O(c\frac{D^2}{D^2})$, soit $O(c)$.
- c est la constante de normalisation telle que $\sum_D 4D\frac{c}{D^2} = 1$ soit $c = \frac{1}{4\log\sqrt{n}} = \frac{1}{2\log n}$.
- Le facteur $\log n$ supplémentaire est comme dans une épreuve de Bernoulli : il faut essayer $O(\log n)$ fois pour que ça marche

Intuition de démonstration

- un **bon lien** divise la distance à la cible par ≥ 2.
- si on suit $\log_2(n)$ bons liens on tombera sur la cible (recherche dichotomique). Or on est à distance D de la cible.
- Quelle probabilité que le lien long de (x, y) soit bon ?
- Doit tomber dans le disque de diamètre D et de centre la cible
- Ce disque contient $D^2/2$ points. La proba que chacun soit l'extremité du lien long est c/D^2 à un facteur $(1/2)^2$ à $(3/2)^2$ près.
- Donc la proba que le lien long tombe dans le disque cible est en $O(c\frac{D^2}{D^2})$, soit $O(c)$.
- c est la constante de normalisation telle que $\sum_D 4D\frac{c}{D^2} = 1$ soit $c = \frac{1}{4\log\sqrt{n}} = \frac{1}{2\log n}$.
- Le facteur $\log n$ supplémentaire est comme dans une épreuve de Bernoulli : il faut essayer $O(\log n)$ fois pour que ça marche
- Donc on réussit en $O(\log^2(n))$ essais en moyenne

Le gros soucis de la dimension

- La proba d'atterrir dans le disque cible doit être constante

Le gros soucis de la dimension

- La proba d'atterrir dans le disque cible doit être constante
- Donc les liens long doivent être en $\frac{1}{d^2}$

Le gros soucis de la dimension

- La proba d'atterrir dans le disque cible doit être constante
- Donc les liens long doivent être en $\frac{1}{d^2}$
- Kleinberg prouve que pour des liens longs en $\frac{1}{d^\alpha}$, **tout algorithme décentralisé** de routage
  - Si $0 < \alpha < 2$ doit utiliser $\Omega(n^{(2-\alpha)/3})$ sauts
  - Et si $\alpha > 2$ doit utiliser $\Omega(n^{(\alpha-2)/(\alpha-1)})$ sauts

Le gros soucis de la dimension

- La proba d'atterrir dans le disque cible doit être constante
- Donc les liens long doivent être en $\frac{1}{d^2}$
- Kleinberg prouve que pour des liens longs en $\frac{1}{d^\alpha}$,
  tout algorithme décentralisé de routage
  - Si $0 < \alpha < 2$ doit utiliser $\Omega(n^{(2-\alpha)/3})$ sauts
  - Et si $\alpha > 2$ doit utiliser $\Omega(n^{(\alpha-2)/(\alpha-1)})$ sauts
- le routage glouton trouve des chemins de longueur logarithmique que pour $\alpha = 2$, polynomiale sinon !

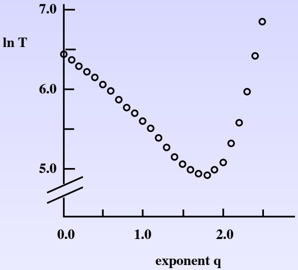


Figure 20.6: Simulation of decentralized search in the grid-based model with clustering exponent q. Each point is the average of 1000 runs on (a slight variant of) a grid with 400 million nodes. The delivery time is best in the vicinity of exponent $q = 2$, as expected; but even with this number of nodes, the delivery time is comparable over the range between 1.5 and 2 [248].

Discussion

- En pratique, le routage glouton est très efficace, alors qu'il n'y a pas de raison que les liens longs «de la vraie vie» suivent une loi en $1/d^2$: le modèle est trop simpliste. Notre monde n'est pas une grille...
- Mais Kleinberg a l'avantage de fournir un cadre où l'on peut **prouver** formellement l'efficacité du routage glouton et **comprendre** l'utilité des liens longs dans la topologie

Conclusion sur le chapitre 2 : Modèles

1. Les **graphes aléatoires de Erdős-Rényi** contribue à expliquer que le diamètre des GRI soit petit (attention à la planarité !)

Conclusion sur le chapitre 2 : Modèles

1. Les **graphes aléatoires de Erdős-Rényi** contribue à expliquer que le diamètre des GRI soit petit (attention à la planarité !)
2. L'**attachement préférentiel de Barabási** contribue à expliquer la distribution des degrés (en power-law)

Conclusion sur le chapitre 2 : Modèles

1. Les **graphes aléatoires de Erdős-Rényi** contribue à expliquer que le diamètre des GRI soit petit (attention à la planarité !)
2. L'**attachement préférentiel de Barabási** contribue à expliquer la distribution des degrés (en power-law)
3. La **grille de Kleinberg** contribue à expliquer la navigabilité (pourquoi le routage glouton marche)

Conclusion sur le chapitre 2 : Modèles

1. Les **graphes aléatoires de Erdős-Rényi** contribue à expliquer que le diamètre des GRI soit petit (attention à la planarité !)
2. L'**attachement préférentiel de Barabási** contribue à expliquer la distribution des degrés (en power-law)
3. La **grille de Kleinberg** contribue à expliquer la navigabilité (pourquoi le routage glouton marche)
4. Les **anneaux de Watts et Strogatz** n'expliquent pas grand'chose mais ont popularisé l'étude du Petit Monde

Conclusion sur le chapitre 2 : Modèles

1. Les **graphes aléatoires de Erdős-Rényi** contribue à expliquer que le diamètre des GRI soit petit (attention à la planarité !)
2. L'**attachement préférentiel de Barabási** contribue à expliquer la distribution des degrés (en power-law)
3. La **grille de Kleinberg** contribue à expliquer la navigabilité (pourquoi le routage glouton marche)
4. Les **anneaux de Watts et Strogatz** n'expliquent pas grand'chose mais ont popularisé l'étude du Petit Monde
5. D'innombrables autres modèles existent pour tenter d'expliquer davantage de phénomènes