

# Le graphe du Web : propriétés, modélisation et clustering

Fabien de Montgolfier

LIAFA, Université Paris 7, France

Projet Algo – 31 janvier 2005

- 1 Propriétés du graphe du Web
  - Le(s) graphe(s) du Web
  - L'effet «Petit Monde»
  - L'effet «scale free»
  - Les cliques biparties
  - Le nœud papillon
- 2 Génération de graphes du web aléatoires
  - Modèles de réseaux aléatoires
  - Modèle de crawl aléatoire
- 3 Clustering : recherche de cybercommunautés
  - Définition du problème
  - Trois algorithmes de partitionnement
  - Les modèles particuliers

# Définitions

## Le graphe du Web

- Sommets : pages Web ( $\simeq$  URL valides)
- Arêtes : hyperliens

## Crawler

Logiciel permettant de récupérer «le» graphe du Web

## Limitations des crawlers

- Bande passante
- Formats de fichiers
- Autorisations (confidentialité, copyright...)
- Évolution temporelle rapide
- Dynamicité (interaction observateur-donnée)
- Duplication
  - sémantique (copier/coller)
  - physique (URL non injectives)

On ne possède donc que des images partielles et biaisées du Web

# Intérêt du graphe du Web

## Utilité

- Intérêt commercial majeur : la mesure d'importance (PageRank de Google)
- Étude et compréhension du Web
  - structure hypertextuelle (sites logiques,...)
  - dynamique de l'Internet
  - Sociologie : communautés d'utilisateurs, blogs...
  - Fouille de données : détection de communautés d'intérêt

## Intérêt algorithmique

- On se ramène à de l'algorithmique de graphe
- Plus rapide que l'analyse sémantique
- Défi dû à la taille des données (8 058 044 651 pages chez google)

- 1 Propriétés du graphe du Web
  - Le(s) graphe(s) du Web
  - L'effet «Petit Monde»
  - L'effet «scale free»
  - Les cliques biparties
  - Le nœud papillon
- 2 Génération de graphes du web aléatoires
  - Modèles de réseaux aléatoires
  - Modèle de crawl aléatoire
- 3 Clustering : recherche de cybercommunautés
  - Définition du problème
  - Trois algorithmes de partitionnement
  - Les modèles particuliers

## Le taux de clustering (ou de transitivité)

Distance moyenne

$$\frac{\sum_{x \neq y} \text{dist}(x, y)}{C_n^2}$$

Clustering (transitivité)

$$\frac{\text{nombre de triangles}}{C_n^3}$$

## Les Petits Mondes [Watts Strogatz 98]

### Distance moyenne

Erdős-Rényi	Petit Monde	Régulier (grille hexa)
petite	petite	grande

### Clustering (transitivité)

Erdős-Rényi	Petit Monde	Régulier (grille hexa)
faible	grande	grande

Pour le Web :

- Distance moyenne 19 clics
- Transitivité : entre 0.1 et 0.7 (!)

# Les Petits Mondes [Watts Strogatz 98]

- Réseaux relationnels entre individus:
  - Criquets chanteurs
  - Relation sociale : distance moyenne 6 [Milgram 67]
  - Mathématiciens (nombre d'Erdős moyen = 5)
  - Acteurs (nombre de Kevin Bacon  $\leq 6$ )
  - Amitié, business, contacts sexuels, etc
- Réseaux dans l'espace :
  - Neurones (*Caenorhabditis elegans*)
  - Internet physique (routeurs et câbles réseau)
  - Réseaux routiers, électrique, ...
- Co-occurrence de mots dans une phrase
- Interaction protéine-protéine
- Participation au capital en bourse

## Distribution des degrés

[Barabási-Albert 99]

- Erdős-Rényi : loi de Poisson  $P(d = k) = e^{-D} \frac{D^k}{k!}$
- Scale-free : loi en puissance  $P(d = k) = C.k^{-\lambda}$
- Graphe régulier : constante  $P(d = k) = \delta_{k,D}$

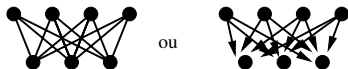
Valeurs de  $\lambda$  :

- **Web** (être pointé par) :  $2.1 \pm 0.1$
- Acteurs :  $2.3 \pm 0.1$
- citation (être cité par) : 3
- Réseau électrique américain : 4

Par ailleurs (et pour d'autres raisons) le nombre de liens d'une page Web suit aussi une loi de puissance de paramètre 2.71

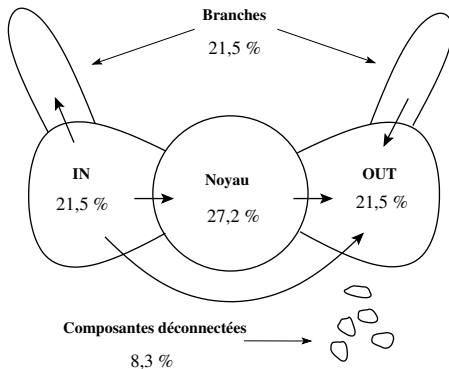
[Broder & al 2000]

## Cliques biparties



- Très nombreuses dans le Web
- Rares dans les Erdős-Rényi
- Nombreuses dans les réseaux d'interaction (relation fan / star)
- Nombreuses aussi dans les réseaux de citation (copier/coller)
- A priori plus caractéristiques de graphes orientés ou naturellement bipartis

## Le nœud papillon



[Kumar, Raghavan 99]

Forte présomption qu'il s'agisse d'un biais dû au crawl

- 1 Propriétés du graphe du Web
  - Le(s) graphe(s) du Web
  - L'effet «Petit Monde»
  - L'effet «scale free»
  - Les cliques biparties
  - Le nœud papillon
- 2 Génération de graphes du web aléatoires
  - Modèles de réseaux aléatoires
  - Modèle de crawl aléatoire
- 3 Clustering : recherche de cybercommunautés
  - Définition du problème
  - Trois algorithmes de partitionnement
  - Les modèles particuliers

## Modèles de génération statique

### Erdős-Rényi

une arête existe avec probabilité  $p$

Avantage : indépendance  $\Rightarrow$  étude facile

Inconvénient : indépendance  $\Rightarrow$  pas réaliste

### Modèles à distribution fixée

[Havel-Hakimi 55] graphe (non aléatoire) à distri. donné

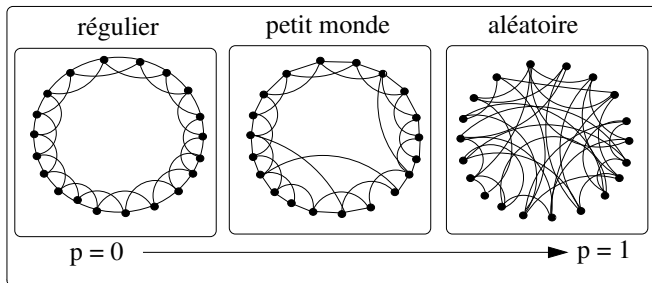
[Molloy-Reed 95] équiprobabilité, multi-graphe, pb connexité

[Aiello, Chung, Lu 02] étudient formellement les propriétés

[Gkantsidis & al 2003] équiprobabilité, graphe simple et connexe

## Modèles à recombinaison

Créer un graphe régulier puis rebrasser certains liens au hasard  
[Watts Strogatz 98] anneau  
[Kleinberg] grille



## Modèles dynamiques

On ajoute les sommets un par un en les liant à l'existant

### Copie de liens

[Kumar & al 99]

chaque nouveau sommet a un prototype

liens copiés du prototype ou choisis uniformément

### Attachement préférentiel

[Barabási-Albert 99] [Bollobás & al 2001]

[Dogorotsev Mendes 2002] meilleur clustering

Loi d'attachement proportionnel au degré de la cible

# Objectifs

Simuler le résultat d'un crawl, donc graphe (+ dates)

- faible distance moyenne
- fort coefficient de clusterisation
- degrés suivant une loi de puissance
- beaucoup de bipartis complets et de noyaux
- connexe a source unique
- avec beaucoup de sommets sans successeurs
- les sommets « importants » sont découverts tôt

## Notre modèle

Travail avec Toufik Bennouas (LIRMM, Montpellier)

### Distribution des degrés

- Nombre de liens sortants pré-tiré en  $k^{-2.71}$
- Nombre de liens entrants pré-tiré en  $k^{-2.1}$

### Files de sommets

- Ensemble des origines, initialement ne contient que la source
- Liste des extrémités, initialement contient les sommets répétés

# Algorithme

Simule un parcours en créant le graphe : à chaque instant

- 1 Extraire  $x$  de l'ensemble des origines
- 2 Si  $x$  déjà vu, ne rien faire. Sinon :
- 3 Extraire ses  $d(x)$  liens de la liste des extrémités
- 4 Les ajouter dans l'ensemble des origines

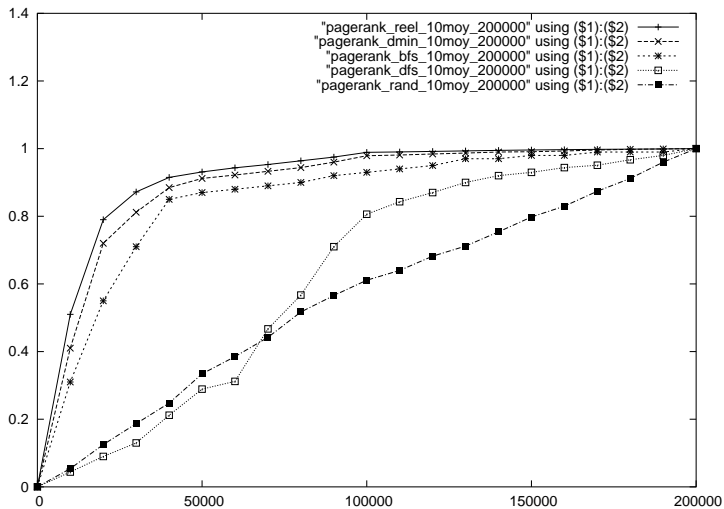
l'ensemble des origines peut être géré comme

- Une file : parcours en largeur (BFS)
- Une pile : parcours en profondeur (DFS)
- Un tas : parcours suivant le degré entrant max (DINMAX)
- pas de gestion (RAND)

# Résultats

- faible distance moyenne
- fort coefficient de clusterisation
- degrés suivant une loi de puissance
- beaucoup de bipartis complets et de noyaux
- connexe a source unique
- avec beaucoup de sommets sans successeurs
- les sommets « importants » sont découverts tôt

# capture des pages importantes



- 1 Propriétés du graphe du Web
  - Le(s) graphe(s) du Web
  - L'effet «Petit Monde»
  - L'effet «scale free»
  - Les cliques biparties
  - Le nœud papillon
- 2 Génération de graphes du web aléatoires
  - Modèles de réseaux aléatoires
  - Modèle de crawl aléatoire
- 3 Clustering : recherche de cybercommunautés
  - Définition du problème
  - Trois algorithmes de partitionnement
  - Les modèles particuliers

# Les communautés du Net

## Communautés explicites

- site (sous-arbre de l'arbre des URL, ou plus complexe)
- usagers qui se fédèrent

## Communautés implicites

- Pages relatives au même sujet
- Personnes partageant les mêmes centres d'intérêt

# Méthodes de détection

- Analyse sémantique
- Analyse hypertextuelle (sur le graphe du Web)
  - Approches locales
  - Approches globales

## Approches locales

Règles locales d'interprétation [Efe, Raghavan & al 2000]



Co-citation



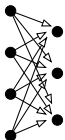
Choix social



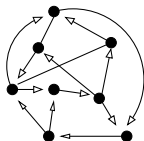
transitivité du vote

Marche moins bien pour les hyperliens que pour les citations d'articles !

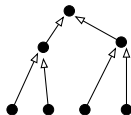
## Définition *a priori* de la forme d'une communauté



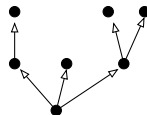
biparti complet



2-clan  
(diamètre <3)



arbre



arbre

## Approches globales

**postulat 1** : une page appartient à au plus une communauté

**postulat 2** : une page appartient à au moins une communauté

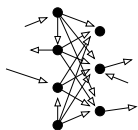
Ainsi, on recherche une **partition**

Analyse plus fine : approche **hiérarchique**. On recherche un arbre.

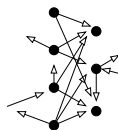
**postulat 3** : il y a exactement  $k$  communautés dans ce crawl

## La recherche de noyaux bipartis

[Kumar, Raghavan & al 1999], [Reddy, Kitsuregawa 2001] etc.



biparti complet



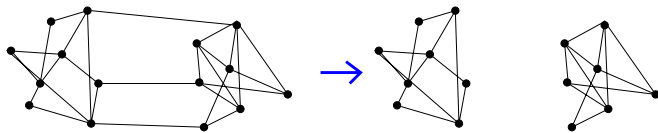
noyau biparti

- problème NP-complet !
- diverses simplifications et ruses algorithmiques

## Les algorithmes de coupe

[Flake, Lawrence & al 2000] postulent qu'une page d'une communauté doit pointer plus de pages de sa communauté que de pages extérieures.

d'où un algorithme décrémental :



## Optimisation d'une mesure de qualité

Algorithme : trouver la partition de qualité maximale.

Mesure de modularité [Newman] :

$$Q = \sum_i m_i - a_i^2$$

Mesure montpélierraine :

$$Q = \sum_i m_i - \bar{m}_i$$

$m_i$ : proba. qu'une arrête ait deux extrémités dans la communauté  $i$

$a_i$ : proba. qu'une arrête ait une extrémité dans la communauté  $i$

$\bar{m}_i$ : proba. qu'une arrête du complémentaire ait ses deux extrémités dans la communauté  $i$

# algorithme de Newman

- 1 Partir de la partition  $\{v_1\}\{v_2\}\dots\{v_n\}$
- 2 Tant que la qualité croît
  - 1 Déterminer les deux communautés dont la fusion augmentera le plus la qualité
  - 2 Les fusionner

Complexité en  $O(n^3)$

## Les modèles particuliers

- Une page Web = une **particule**
- Chaque particule est dans l'**espace** (3D infini, sphère...)
- Chaque particule a un **poids** : son PageRank
- Chaque particule bouge au cours du temps
- elle **interagit** avec les autres selon les hyperliens de la page correspondante

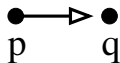
Algorithme:

- Placer les particules au hasard
- Laisser le système évoluer
- Arrêter quand la qualité cesse de croître
- Les particules sont maintenant regroupées en communautés

# Le modèle gravitationnel

Implémente les lois de Newton :

- accélération, vitesse et position suivent les lois physiques classique (discrétisée)
- attraction :



$$G \frac{\text{masse}(p) \cdot \text{masse}(q)}{\text{dist}(p, q)^2}$$

## Le modèle barycentrique

- L'**objectif** d'une particule est le barycentre des pages qu'elle pointe (que la page réfère)
- À chaque étape, une particule fait un **pas** vers son objectif

## Discussion

- Peu de paramètres à définir (taille du pas)
- Calcul rapide : quelques centaines de passes suffisent
- Résultats esthétiques
- Résultats instables
- Qualité obtenue plus faible que l'existant
- Pour le Web, manque de validation sémantique

## Conclusion

- On cherche à comprendre la structure du Web
- Pour cela il faut réussir à la reproduire
  - Beaucoup de possibilités : domaine en pleine expansion
- On cherche aussi à classifier les pages
  - problème mal formalisé mais domaine également à la mode
- Il faut enfin évaluer l'importance des pages
  - exposé de Fabien Mathieu