chapitre 1

Propriétés structurelles des GRI

Section 1:

degrés

Fabien de Montgolfier
fm@irif.fr

21 janvier 2022

Plan du cours (rappel)

1. **Description**
   Comprendre les propriétés structurelles des GRI
   en informatique, sociologie, biologie, physique, linguistique...
   - **en TP** : Calcul de divers paramètres.
     Big data → structures de données et algorithmes efficaces
2. **Modélisation**
   Classifier ces réseaux d'interaction.
   Modèles aléatoires.
   Si même propriétés que le réel, alors même génération ?
   - **en TP** : Génération aléatoire
3. **Deux applications**
   Dissémination (contagion)
   P2P (DHT ; diffusion temps réel)

Propriétés communes des réseaux Pourquoi ces propriétés ?

En plus de la **grande taille** on a (presque toujours) :

1. Distribution des **degrés** : peu de riches beaucoup de pauvres
2. Distribution des **distances** : tout le monde proche de tout le monde (effet *small world*)
3. Existence d'un **cœur** : les riches se connaissent, pas les pauvres
4. Composante connexe **géante**
5. **Transitivité** forte (= *coefficient de clustering*) : les amis de mes amis sont mes amis
6. **Navigabilité** (pas toujours) : on peut atteindre une cible de proche en proche. Ex : routage IP.
7. existence de **communautés** (pas toujours non plus) : sous-graphes denses ayant un sens.

Degré moyen $\overline{d}$

Notation : n est le nombre de sommets, m le nombre d'arêtes

Degré

$d(x)$ = Nombre d'arête incidentes à un sommet x
En **orienté** : somme des degrés entrant et sortant

$$d(x) = d^+(x) + d^-(x)$$

Degré moyen

$$\overline{d} = \frac{1}{n}\sum_x d(x) \quad \text{donc} \quad \overline{d} = \frac{2m}{n}$$

$$0 \leq \overline{d} < n$$

Dans un GRI le degré moyen est petit

Valeurs empiriques

Dans un GRI le degré moyen est petit

Données de `https://snap.stanford.edu/data/`

n	m	$\overline{d}$	nom
1632803	30622564	37,51	Pokec online social network
134833	1380293	20,47	Gemsec Facebook dataset
65608366	1806067135	55,06	Friendster online social network
3072441	117185083	76,28	Orkut online social network
36692	183831	10,02	Email communication network from Enron
2394385	5021410	4,19	Wikipedia talk (communication) network
1791489	28511807	31,83	Wikipedia hyperlinks
3774768	16518948	8,75	Citation network among US Patents
685230	7600595	22,18	Web graph of Berkeley and Stanford
403394	3387388	16,79	Amazon product co-purchasing from 2003
1965206	2766607	2,82	Road network of California
1696415	11095298	13,08	Internet topology graph from traceroutes
17069982	476553560	55,84	Tweets collected between June-Dec 2009

Graphes denses et clairsemés (*sparse graphs*)

Intuition : Un graphe est clairsemé s'il a peu d'arêtes.
Sinon il est dense

Tentative de définition

*$m = O(n)$ quel est le problème ?s'applique à une **classe** de graphes non à **un** graphe*

Definition

Une classe de graphes $\mathcal{C}$ est *clairsemée* si

$$\forall G \in \mathcal{C} \text{ on a } m = O(n)$$

Rappel : cela implique qu'il existe des constantes n_0 et k telles que

$$n > n_0 \implies m \leq kn$$

Abusivement, on dira qu'**un graphe** de cette classe est clairsemé

Remarques

- On trouve parfois la définition que $\overline{d} = O(\log n)$. Il est dur de distinguer si $\overline{d} = O(1)$ ou si $\overline{d} = O(\log n)$ dans des cas particuliers car pour un graphe réel à $n = 1\,000\,000\,000$ on a $\log_{10}(n) = 9$...
- Un graphe est clairsemé $\iff$ son complément est dense
- Les structures de données à utiliser ne sont pas les mêmes !
  Un graphe dense se représente par une **matrice d'adjacence** et un graphe clairsemé par une **liste d'adjacence** (ou variante en $O(m)$ en espace)
- Par exemple pour un graphe du Web on sait stocker en 5,5 bits par hyperlien [Randall et al. 2001] ou même en **3,08 bits par hyperlien** [Boldi Vigna 2008]

Exemples

Graphes complets

Un graphe complet (une clique) est dense car $\overline{d} = n - 1$
et donc $m = \frac{n(n-1)}{2}$

Arbres

Les arbres sont clairsemés car $m = n - 1$ donc $\overline{d} < 2$

Graphes connexes

Si un graphe est connexe alors $m \geq n - 1$ donc $\overline{d} \geq 2 - \frac{2}{n}$
et donc $\overline{d} \geq 2 - o(1)$

Le métro de Londres

Un graphe planaire peut être dessiné sans croisement des arêtes

Graphes planaires

Définition

Un graphe est **planaire** si on peut le dessiner (sur un plan, donc) sans croisement des arêtes

Formule d'Euler

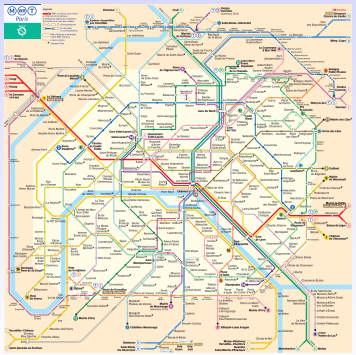
Si un graphe planaire possède f **faces** alors n - m + f = 2
Exemple pour un cube 8 sommets - 12 arêtes + 6 faces = 2

Degré moyen

Chaque face F a au moins 3 arêtes (triangle) donc $m(F) \geq 3$.
$2m = \sum_F m(F) \geq 3f$ car chaque arête appartient à 2 faces
$6 = 3n - 3m + 3f \leq 3n - m$ et $\overline{d} = \frac{2m}{n} \leq \frac{2(3n-6)}{n} \leq 6 - \frac{12}{n}$

Exercice : trouver une classe de graphes planaires de $\overline{d}$ maximal

Le métro de Paris est quasiment planaire... et donc clairsemé



Moralité

Le degré moyen d'un GRI peut être contraint

- par la nature **physique** du réseau (les réseau de distribution sont quasi-planaires) ou pas...
- Par des contraintes **économiques** (tendre vers un arbre) ou pas...
- Par des contraintes **humaines** (on ne peut pas connaître tout le monde dans un réseau social) ou pas...

mais reste petit.

En effet un même acteur (un nœud) ne peut avoir qu'un nombre limités d'interactions.

Les réseaux sans facteur d'échelle (*scale-free*)

Notons $P_G(k)$ la proportion (ou probabilité) de sommets de degré k d'un graphe G

Définition

Une famille de graphes $\mathcal{G}$ est dite **sans facteur d'échelle** (*scale-free*) s'il existe un réel $\gamma > 0$ tel que :

$$\forall G \in \mathcal{G},\ P_G(k) = k^{-\gamma}$$

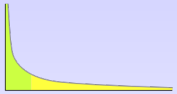
$2 \leq \gamma \leq 3$ dans la plupart des applications

Analyse

Plus le degré augmente, moins il y a de sommets. Cela décroit en puissance $-\gamma$: on parle donc de loi de puissance (ou *power law*) pour désigner les degrés des sommets

Loi de puissance

$P_G(k) = k^{-\gamma}$ avec usuellement $2 \leq \gamma \leq 3$



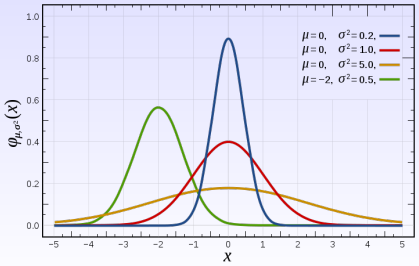
On retrouve des lois de puissance **partout** :

- Dans la **richesse** (peu de riches, beaucoup de pauvres). Le nombres de gens ayant une fortune de F€ est en $F^{-\gamma}$ (et donc ceux ayant une fortune d'*au plus* F€ est en $F^{-\gamma+1}$)
- Dans la **taille des villes** (peu de métropoles, beaucoup de villages) Le nombres de villes de h habitants est en $h^{-\gamma}$ (et donc le nombre de villes d'*au plus* h habitants est en $h^{-\gamma+1}$)
- Dans l'**audience** des chaînes Youtube ou des comptes Twitter
- Dans votre nombre d'**amis Facebook**

Pourquoi ? et est-ce bien vrai ?...

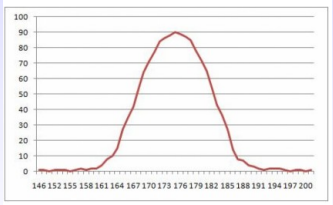
Loi de puissance *vs* loi normale

Pourtant la pluspart des phénomènes physiques ne suivent pas une loi de puissance mais une **loi normale** dont la distribution suit une **gaussienne** (courbe en cloche)



Lois normales

Théorème central limite,
ou limite de lois binomiales.
On y reviendra (graphes aléatoires)



Taille des hommes adultes en France

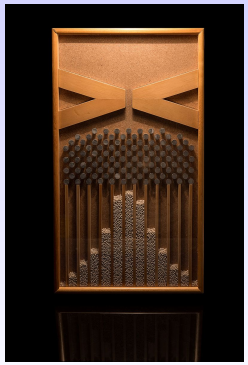


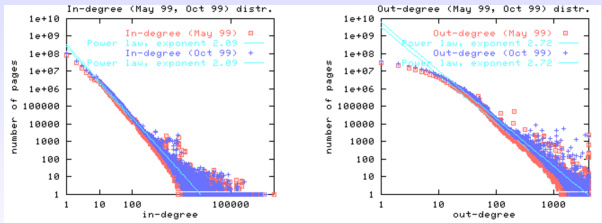
Planche de Galton

Invariance d'échelle et power laws

- $P_G(k) = k^{-\gamma}$ implique
  $P_G(ak) = (ak)^{-\gamma} = (a^{-\gamma})k^{-\gamma}$ donc proportionnel à $P_G(k)$.
- $\log(P_G(k)) = -\gamma \log(k) + cte$
  On retrouve l'équation d'une droite dont la pente est $-\gamma$
- Cela permet de calculer γ en pratique : on passe au log, on regarde la courbe de la distribution et on fait une régression linéaire (tous les logiciels de statistiques savent faire ça, par exemple `gnuplot`)

De vieux résultats sur le graphe du Web

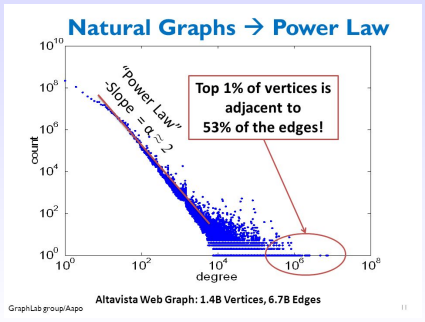
Kumar et Raghavan [1999]



Figures 3 and 4: In- and out-degree distributions show a remarkable similarity over two crawls

Que vaut l'exposant ? Plein de valeurs différentes publiées !

Cette diapo ci a aussi une vingtaine d'années...



chapitre 1 Propriétés structurelles des GRI Section 1:degrés
└Loi de puissance et réseau Scale-free

Discussion **théorique** sur les lois de puissance

Approximation théorique

En fait cela ne peut pas être *exactement* $P_G(k) = k^{-\gamma}$: pour $k = 1$ cela donne $P_G(1) = 1$. 100% des sommets devraient avoir degré 1... En fait $\sum_{k=1}^{k=\infty} k^{-\gamma} = \zeta(\gamma)$ où ζ est la fonction zêta de Riemann avec $\zeta(2) = \frac{\pi^2}{6} \simeq 1,645$ et $\zeta(3) \simeq 1,202$. Il faudrait donc pour être exact écrire

$$P_G(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

Problème du zéro

Pour $k = 0$ on voit que la probabilité d'un sommet isolé n'est pas définie. Or des gens sans ami sur Facebook ou sans le moindre euro en poche, ça existe... Et surtout des pages Web sans lien sortant !

chapitre 1 Propriétés structurelles des GRI Section 1:degrés
└Loi de puissance et réseau Scale-free

Discussion **pratique** sur les lois de puissance

Mais surtout

Il s'agit d'une modélisation théorique d'un phénomène réel : ça ne colle jamais aussi bien, surtout pour les petites valeurs de k !

Décroissance

La vraie question à se poser est : est-ce que quand k devient grand $P_G(k)$ se comporte asymptotiquement en $O(k^{-\gamma})$?

chapitre 1 Propriétés structurelles des GRI Section 1:degrés
└Loi de puissance et réseau Scale-free

La taille de la queue est importante

Queue lourde

Une distribution est à queue lourde (*heavy-tailed*) quand $P_G(k) = \omega(e^{-k})$: elle est asymptotiquement supérieure à une exponentielle (elle décroît moins vite qu'une exponentielle).

Queue épaisse

Les distributions où $P_G(k) = O(k^{-\gamma})$ sont dites à queue épaisse (*fat-tailed*). Les queues épaisses sont lourdes. Mais il y a d'autres queues lourdes : distribution de Pareto, de Lévy, log-normales...

Queue légère

Au contraire la gaussienne et la loi de Poisson (cf cours graphes aléatoires) ont une décroissance en $o(e^{-k})$ (exponentielle)

chapitre 1 Propriétés structurelles des GRI Section 1:degrés
└Loi de puissance et réseau Scale-free

Mouton noir ?

Idée reçue

D'après S. Vigna [2016] les degrés de Facebook ne suivent **pas** une loi de puissance

De nombreuses définitions sur les réseaux complexes sont produites par des physiciens ...

Ils généralisent beaucoup trop vite ! Par exemple pour eux tous les GRI ont une distribution des degrés en loi de puissance.

chapitre 1 Propriétés structurelles des GRI Section 1:degrés
└Loi de puissance et réseau Scale-free

Conclusion

- Cette définition (loi de puissance $P_G(k) = k^{-\gamma}$) permet d'établir des propriétés générales des GRI.
- Cette définition devrait s'appliquer à une **classe** de graphes mais on peut faire des statistiques sur un seul graphe.
- Mais souvent elle ne formalise pas plus que l'idée intuitive : il y a peu de sommets de grand degré dans un graphe «réel», et beaucoup de petit degré
- Aucun consensus sur la valeur de γ pour divers graphes *beaucoup* étudiés comme le graphe du Web

chapitre 1 Propriétés structurelles des GRI Section 1:degrés
└Densité locale et communautés

Densité locale

- Un graphe clairsemé peut-il contenir des cliques aussi grandes que l'on veut ?
- Oui !
- Il suffit de prendre un graphe G_n constitué d'une clique de taille $\sqrt{n}$ prolongée par une chaîne de taille $n - \sqrt{n}$.
- $m = \frac{(\sqrt{n})(\sqrt{n}-1)}{2} + (n - \sqrt{n}) \leq \frac{3}{2}n$
- Comme $\overline{d} < 1.5$, cette famille de graphes est bien clairsemée !

chapitre 1 Propriétés structurelles des GRI Section 1:degrés
└Densité locale et communautés

Moralité

Nous aimerions une définition impliquant une régularité interne : tout **sous-graphe** est clairsemé.

Rappel : clairsemé

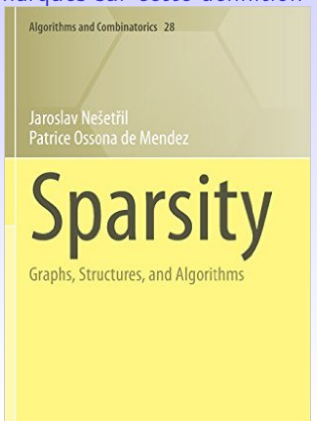
Une classe de graphes $\mathcal{G}$ est clairsemée si
$\exists k \in \mathbb{N}$ tel que $\forall G \in \mathcal{G}$ on a $m(G) \leq k.n(G)$

Clairsemé partout

Une classe de graphes $\mathcal{G}$ est **clairsemée partout** si
$\exists k \in N$ tel que $\forall G \in \mathcal{G}$ et $\forall H \subseteq G$ on a $m(H) \leq k.n(H)$

chapitre 1 Propriétés structurelles des GRI Section 1:degrés
└Densité locale et communautés

Remarques sur cette définition



Grands réseaux d'interaction construits

- Certains GRI sont construits (plus ou moins) humainement, c'est-à-dire qu'il y a une volonté humaine d'insérer un nœud à un endroit donné
  Exemples : réseau des amis dans Facebook, graphe des échanges dans Twitter, graphe du Web (quoique la majorité des liens sont compilés...) etc.
- Chaque individu appartient (plus ou moins consciemment) à une ou plusieurs **communauté** et se connecte préférentiellement aux gens de la même communauté.
  On suppose que chaque individu appartient un nombré borné de communautés
- Il y a donc $O(n)$ communautés (qui se chevauchent) et on peut donc peut envisager d'écrire un algorithme qui les énumère toutes

Qu'est-ce qu'une communauté ?

Bonne question ! Autant de réponse que de socio-informaticiens

- Parfois on cherche à déterminer les **rôles** que jouent les individus dans les réseaux d'interaction.
- parfois il s'agit de trouver des **similarités** entre individus
- On simplifie beaucoup (d'un point de vue algorithmique) le problème en disant que chaque individu appartient à une communauté. On veut alors calculer une **partition**.
- Une chose assez consensuelle dans le cadre de la recherche de communautés : les sous-ensembles denses donnent un première approximation.
- Nous recherchons donc des sous graphes maximaux denses.
- En anglais une communauté s'appelle un *cluster* et le calcul de communautés le *clustering*. On reverra ce mot avec un autre sens : *coefficient de clustering*.

Densité et communautés

Degré moyen d'un sous-graphe H de G

$\boxed{\overline{d}(H) = \frac{2m(H)}{n(H)}}$ Varie de 0 (peu dense) à $n(H) - 1$ (dense)

Densité d'un sous-graphe H de G

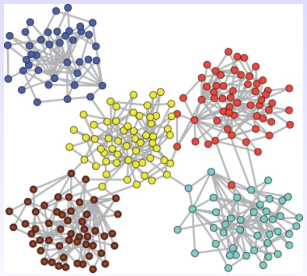
$\boxed{\rho(H) = \frac{2m(H)}{n(H)(n(H)-1)}}$ Varie de 0 (peu dense) à 1 (dense)

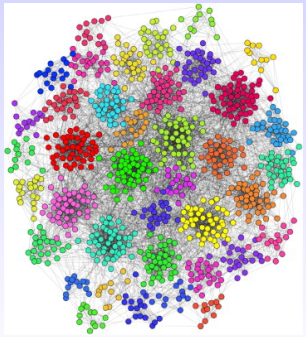
Problème (à formulation floue) du calcul des communautés

Trouver une **partition** (ou un recouvrement) des sommets en c sous-graphes (c fourni à l'avance ou pas) tels leur densité (Moyenne ? Minimale ?) soit maximale

unicité très peu garantie ! Dépend fortement de la définition et de l'algorithme qui l'implémente

On voudrait ça :





Communauté la plus dense

Degré moyen d'un sous-graphe H de G

$\boxed{\overline{d}(H) = \frac{2m(H)}{n(H)}}$ Varie de 0 (peu dense) à $n(H) - 1$ (dense)

MAD subgraph (ou densest subgraph)

Sous-graphe H de degré moyen maximum.

Algorithme

Trouver un MAD subgraph est NP-complet !

MAD

Algorithme 2-approché

$G_0 := G$;
Pour $i := 1$ à n :

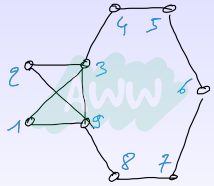
- Soit u un sommet de degré minimum,
- $G_i := G_{i-1} - u$.

Retourner G_i tel que $\frac{2m(G_i)}{n(G_i)}$ est maximum.

Si $deg_G(u) < \frac{\overline{d}(G)}{2}$ alors $\overline{d}(G - u) > \overline{d}(G)$.
Pour le sous-graph optimum, $degmin(H) \geq \frac{\overline{d}(H)}{2}$.
Quid du graphe G_i où le premier sommet u de H est éliminé ?

MAD : exercice

Trouver un graphe et une exécution de l'algorithme MAD 2-approché qui soit sous optimale.



V entier vérifie $m/n = 1.22...$ Mais pour $\{1, 2, 3, 9\}$ c'est 1.25

Conclusion

- Les GRI ont un petit degré moyen : ils sont clairsemés
- Leurs degrés suivent souvent une loi de puissance ou, au moins, une loi à queue lourde : peu de sommets de fort degré, beaucoup de sommets de petit degré
- Les zones de forte densité locale permettent de calculer des communautés. On y reviendra en parlant de centralité.

Article recommandé :
The Graph Structure in the Web – Analyzed on Different Aggregation Levels, Meusel et al. 2015
`https://webscience-journal.net/webscience/article/download/11/75`