chapitre 1

Propriétés structurelles des GRI

Section 3

transitivité, sociologie et grillons

Fabien de Montgolfier
fm@irif.fr

4 février 2022

Propriétés communes des réseaux

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. Éxistence 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.

La composante géante

Définition informelle

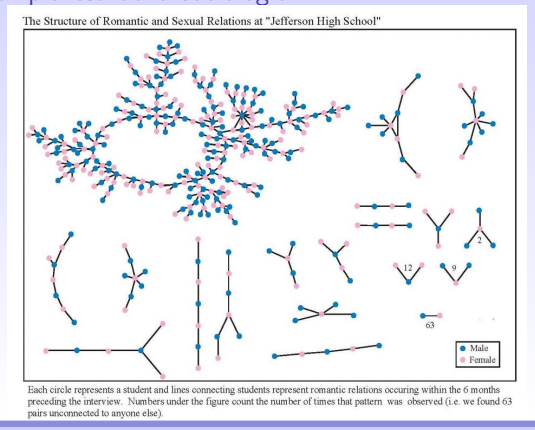
On dit qu'**un graphe** a une composante connexe géante quand il a plusieurs composantes connexes mais qui sont toutes très petites sauf une, comprenant une grande majorité des sommets du graphe

Définition formelle

Une **famille de graphes** issus d'un modèle probabiliste possède une composante connexe géante quand, *avec forte probabilité*, une composante d'une instance a taille $O(n)$ et toutes les autres $o(n)$.

On reviendra à la deuxième définition en parlant des graphes aléatoires $\mathcal{G}(n,p)$ d'Erdős-Renyi

Un exemple issu de la sociologie



La composante fortement connexe géante (anglais : *GSSC* ou *LSSC*)

n	SCC	%	nom
1632803	1304537	79,9%	Pokec online social network
65608366	65608366	100%	Friendster online social network
3072441	3072441	100%	Orkut online social network
36692	33696	91,8%	Email communication network from Enron
2394385	111881	4,7%	Wikipedia talk (communication) network
1791489	1791489	100%	Wikipedia hyperlinks
3774768	1	0%	Citation network among US Patents
685230	334857	48,9%	Web graph of Berkeley and Stanford
403394	395234	98%	Amazon product co-purchasing from 2003
1965206	1957027	99,6%	Road network of California
1696415	1694616	99,9%	Internet topology graph from traceroutes

Transitivité et coefficient de clustering

- Une relation R est transitive si xRy et yRz impliquent xRz.
- Si R est l'amitié, cela donne
  les amis de mes amis sont mes amis
- Bien sûr l'amitié n'est pas réellement transitive sinon deux humains seraient toujours amis (sauf quelques ermites)
- Un graphe est une relation.
- Le **coefficient de clustering** mesure à quel point un graphe est transitif.
- Va de 1 (le graphe est transitif au sens strict : formé de cliques disjointes) à 0 (graphe sans triangle)
- Deux définitions pour les graphes non-orientés (aucune ne s'applique aux graphes orientés)

Attention, **deux** définitions du coefficient de clustering

- $tri(G)$ et $tri(x)$: le nombre de triangles de G / touchant x
- Clustering Local = densité du voisinage

$$cluL(x) = \frac{2 * tri(x)}{deg(x)(deg(x)-1)}$$

- **Clustering Local Moyen** $cluL(x) = \frac{1}{n}\sum_{x\in G} cluL(x)$
- Un $\mathbb{V}$: deux arêtes incidentes. Il se referme en triangle ou pas.
- $nv(G)$ est le nombre de V de G
- **Clustering global** proportion des $\mathbb{V}$ qui se ferment en triangle

$$cluG(G) = \frac{3 * tri(G)}{nv(G)}$$

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

n	cluL	cluG	nom
1632803	0.1094	0.01611	Pokec online social network
65608366	0.1623	0.005859	Friendster online social network
3072441	0.1666	0.01414	Orkut online social network
36692	0.4970	0.03015	Email communication network from Enron
2394385	0.0526	0.001112	Wikipedia talk (communication) network
1791489	0.2746	0.00165	Wikipedia hyperlinks
3774768	0.0757	0.02343	Citation network among US Patents
685230	0.5967	0.002746	Web graph of Berkeley and Stanford
403394	0.4177	0.06206	Amazon product co-purchasing from 2003
1965206	0.0464	0.02097	Road network of California
1696415	0.2581	0.001802	Internet topology graph from traceroutes

À retenir :

Les réseaux d'interaction sont transitifs !

Oui, les amis de mes amis sont (probablement) mes amis

La mesure de transitivité s'appelle **coefficient de clustering**

- Il y a deux définitions : local moyen et global
- Malgré ce nom, rien à voir avec la présence de communautés (*clusters*)

L'anneau de Watts et Strogatz

Une publication célèbre

D. J. Watts et S. H. Strogatz, *Collective dynamics of "small-world" networks*, Nature, 1998.

Une motivation surprenante

Strogatz voulait comprendre comment des grillons chantent à l'unison dans un champs

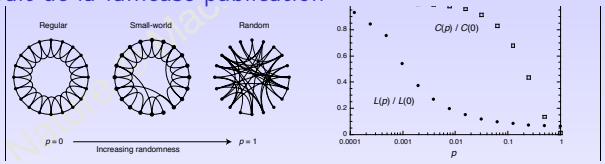
Définition

Chaque sommet a degré k (typiquement, 4). Un sommet est relié

- à un voisin avec une probabilité $1 - p$
- à un sommet choisi aléatoirement avec une probabilité p

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Anneau de Watts et Strogatz

Extrait de la fameuse publication



- À droite on voit bien leur motivation : avoir des distances moyennes courtes (L) **et** un clustering élevé (C)
- Ils se comparent aux graphes aléatoires $\mathcal{G}(n, p)$d'Erdős-Renyi
- Pourquoi partir d'un anneau ? Quel phénomène réaliste est ainsi modélisé ? Mystère !
- Il existe depuis des modèles bien plus crédibles et reproduisant bien plus de propriétés des GRI. On verra l'attachement préférentiel et la grille de Kleinberg.

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Anneau de Watts et Strogatz

Conclusion

- Watts et Strogatz ont inventé le concept de *small world* et le coefficient de clustering, et ont donné une grande publicité à ces notions.
- Avec un modèle beaucoup trop simpliste mais peu importe !
- On savait depuis longtemps que les distances moyennes étaient courtes.
- Avec les graphes aléatoires $\mathcal{G}(n, p)$ d'Erdős-Renyi on verra que il n'est pas surprenant que les distances moyennes soient courtes, mais que le clustering, lui, révèle un mécanisme non «bêtement» aléatoire à l'œuvre.

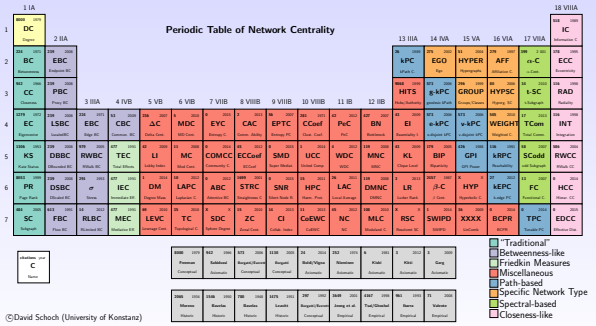
- Comment mesurer l'importance d'une personne dans un réseau social ?
- Ou l'importance d'une page (ou d'un site) Web ? Qui mettre en premier résultat d'une recherche Google ?
- On veut donc une mesure numérique de l'importance : qui est central ?
- On a déjà vu que Kevin Bacon est le centre de l'univers (du cinéma)... Mais que (presque) tout le monde aussi !

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

- Il existe de très nombreuses définitions
- On peut donc se demander lesquelles sont pertinentes ? Selon quels critères ?
- Ainsi, les chimistes, pour essayer de trouver dans leurs molécules de chimie organique les atomes importants dit "centraux", ont défini plus d'une centaine de mesures de centralités adaptées... Et il s'agit de petits graphes !

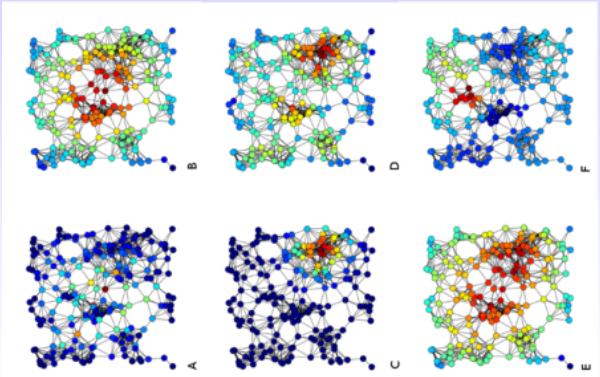
chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Table de http://schochastics.net/sna/periodic.html



chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Exemples de centralités sur un même graphe



chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Critères de choix d'une mesure en particulier

1. Calculer quelque chose de pertinent (paramètre qui dépend du réseau social étudié).
2. Robustesse (résistant à la suppression d'arêtes ou de sommets)
3. Être calculable efficacement (temps linéaire en la taille du graphe)

- Pour une application donnée, il peut y avoir une notion de centralité, qui tienne vraiment la route. Par exemple PageRank pour le graphe du Web.
- Examinons en détail les idées sous-jacentes aux principales définitions de centralité et essayons de comprendre ce qu'elles mesurent.

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Deux grandes approches

1. Les flots : mesurer quelque chose qui **circule** dans le réseau
2. La cohésion : mesurer les sommets qui **maintiennent** le réseau.

Hélas ces deux notions ne sont pas totalement indépendantes.

Variations

- Introduire de la dynamique (par exemple dans Twitter)
- L'influence réelle peut être assez différente de la centralité (par exemple sur Twitter, les partisans russes de Trump)

Tentative basique : le degré

- Facile à calculer et robuste
- Si graphe orienté (citations, Web) : degré entrant
- Mesure l'influence en nombre de connexions (exemples *like* Facebook, nombre de followers dans Twitter, ...)
- Altavista, premier moteur de recherche (de 1995 à 2000) à base de graphe du Web, fonctionne (en partie) ainsi. Facile de booster l'importance de sa page par des *link farm*
- PageRank, utilisé par Google [2000] en est une amélioration (il faut être pointé par des pages de fort PageRank, car on additionne les PageRank des pages entrantes, au lieu de compter 1 par page entrante) (on va en reparler)

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Seconde tentative : par les distances

Sommets dont l'eccentricité est minimale.

- C'est la notion «naturelle» de centralité (barycentre)
- On calcule l'ensemble des centres du graphe
- Calcul en temps $\Omega(n^2)$ bof
- Moins robuste (ajouter un long chemin sortant d'un graphe va déplacer tous les centres)
- Et beaucoup de centres souvent (exemple Kevin Bacon)

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Centralités par déconnection

- Un sommet p est un **point d'articulation** si $G - p$ a davantage de composantes connexes que G
- Ces points d'articulation ont la meilleur centralité. Le reste du graphe est formé de composantes biconnexes.
- On peut continuer : les paires de sommets séparant le graphe, restent les composantes triconnexes (il y a un joli algo linéaire)
- Puis on obtient des sommets de moins bonne centralité, par k croissant : les k-séparateurs. Restent les composantes $k+1$-connexes. Pour $k > 3$ on utilise des algorithmes de flots
- Problème, séparer un sommet du reste du monde joue le même rôle que couper le graphe en deux parties égales. La *betweenness* améliore cela

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Betweenness (intermédiarité)

Définition

- $npcc(s,t)$: nombre de plus courts chemins entre deux sommets s et t
- $npcc(s,v,t)$: nombre de pcc entre s et t qui passent par v

$$bet(s,v,t) = \frac{npcc(s,v,t)}{npcc(s,t)} \quad \in [0,1]$$

$$bet(v) = Moyenne\{bet(s,v,t)\} \quad \in [0,1]$$

$$bet(v) = \frac{1}{(n-1)(n-2)} \sum_{s \neq v \neq t \neq s} bet(s,v,t)$$

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Betweenness (intermédiarité)

- Donc plus il y a de plus courts chemins qui passent par v plus sa betweeness augmente
- Il existe une version non normalisée (pas une moyenne mais une somme)
- Un sommet isolé ou pendant (feuille) a betweenness 0
- Le centre d'une étoile a betweenness 1
- Peu robuste
- Coûteux à calculer $O(nm)$ par l'algorithme de Brandes
- Il peut y avoir un nombre exponentiel de plus courts chemins entre deux points donnés, c'est pour cela qu'on normalise
- Parfois (souvent) l'implémentation est **fausse** : calcule pour toute paire $\{y,z\}$ **un** plus court chemin et augmente $bet(v)$ pour tout v le long de ce chemin.

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Centralités à base de flots

- Mesure l'influence en nombre de choses qui passent par les arêtes ou sommets (information, messages par exemple, click sur un hyperlien ...)
- Liens avec le *Gossip Problem* du distribué : cf cours sur la diffusion
- Coûteux à calculer

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Centralité à base de marches aléatoires

- On procède à des marches aléatoires dans le graphe et l'on calcule la probabilité qu'un sommet donné y apparaisse
- Variations sur la marche utilisée
- Facile à calculer, parfois par calcul matriciel, relativement robuste

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

Un résultat célèbre : PageRank de Google

chapitre 1 Propriétés structurelles des GRI Section 3 transitivité, sociologie et grillons
└Centralités

PageRank

But : prédire l'audience (popularité) d'une page

- Les résultats d'une recherche google sont classés par audience
- Non, Google ne connait pas **tous** les clics sur **toutes** les pages !
- Mais ils connaissent le graphe du Web, par *crawling*

Le surfeur aléatoire

- Il passe de page en page. Si plusieurs liens sortant, il en prend un au hasard. Si aucun, il repart de n'importe où.
- Le Pagerank d'une page est le nombre de passages du surfeur sur cette page au bout d'un temps trèèès long
- *cf* cours MAAIN

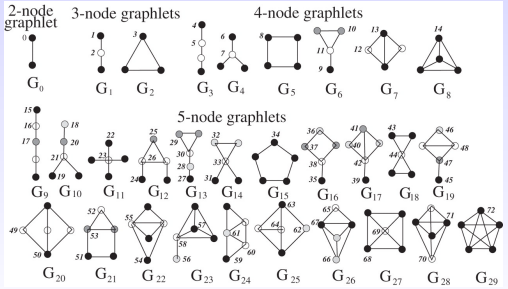
Autres centralités

- La centralité de Katz sorte de degré généralisé : on compte aussi les sommets à distance k mais avec un facteur d'amortissement
- Centralité par les valeurs propres de la matrice (variations sur PageRank, encore)
- ...

Les distributions des sous-graphes de petites tailles

- La méthode des *Graphlets* (pas de bonne traduction...)
- Construisons l'histogramme des sous graphes de petites tailles de notre réseau (i.e. calculons le nombre de sous-graphes de notre réseau isomorphes à un P_3, triangle, P_4, C_4)
- C'est une généralisation du calcul de nombre de triangles ou des coefficients de clustering
- Pour des raisons de temps de temps de calcul, on se limitera aux sous-graphes de taille 5, 6 ou 7

Graphlets



Ces histogrammes «caractérisent» les réseaux

- Trivialement, l'histogramme des graphlets caractérise le réseau, car G est un graphlet de G.
- Mais c'est également vrai si on limite leur taille
- En particulier, ils permettent de séparer les graphes géométriques des graphes sans facteur d'échelle.
- Cette notion s'étend aux graphes étiquetés (les arêtes étant colorées application à la biologie)
- On peut faire des comparaisons d'histogrammes de deux graphes et comparer avec l'histogramme des graphlets d'un graphe aléatoire.
- Cette méthode est depuis très utilisée dans l'analyse des grands réseaux d'interactions.
- Les sociologues (réseaux égo-centrés) et quelques bio-informaticiens aiment bien cet outil

Rôles

Une définition sociologique naturelle (à la mode depuis 1970)

Informellement le **rôle** d'un humain est une étiquette : prof, étudiant, policier, boulanger...

Dans un graphe deux sommets ont le même rôle quand leurs voisins ont le même rôle

Beaucoup de publications de sociologie pure avec de petits graphes étudiés «à la main»

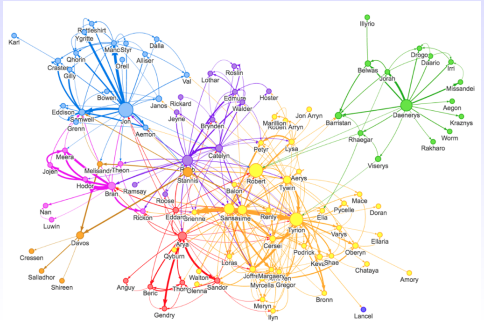
Mais un calcul difficile

On peut toujours donner le même rôle à tout le monde... Trivial

C'est NP-complet d'après Fiala et Paulusma [2005] de trouver des «bons» rôles

Liens faibles, liens forts

On part d'une définition «naturelle» de sociologie : lien fort avec son conjoint, faible avec son facteur. Comment la formaliser afin de la rendre opérationnelle et de la calculer sur des réseaux sociaux ?



Liens faibles, liens forts

- Un isthme (arête déconnectant le graphe) est un lien faible (mais important pour la structure, ce qui est contradictoire !)
- On peut appliquer la notion de betweenness aux arêtes (nombre de plus courts chemins passant par une arête)
- Ou encore, si en enlevant l'arête xy alors x et y se retrouvent à plus de k (k constante fixée) on dit que le lien est faible et fort sinon. Pour les isthmes on prend $k = \infty$
- Autre idée, à base de complétions en triangles (par transitivité donc) mais cela dépend de l'ordre des complétions.
- En général on renvoie le problème au calcul de **communautés** (*clusters*) : tous les liens entre les communautés sont faibles et les liens internes forts.
- Le temps de calcul est linéaire pour les isthmes, au moins $O(nm)$ sinon

Liens faibles, liens forts et applications

On a vu

Que le calcul de communauté définit des liens internes (forts) et externes (faibles)

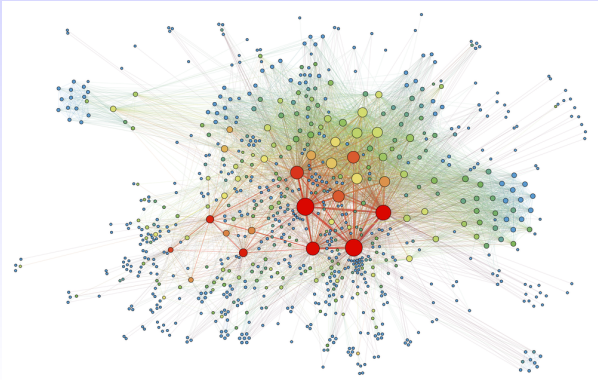
À l'inverse

La détection des liens faibles permet en retour de calculer des communautés : On enlève les liens faibles et les composantes connexes restantes sont les communautés !

En dessin de graphe

On veut dessiner les liens forts plus courts que les liens faibles. Et plus foncés. Utilisé en général avec des poids sur les sommets également

Dessin basé sur les forces (Force-directed algorithms)



Bien d'autres applications en sciences sociales

Claude
Lévi-Strauss
(1908–2009)
Anthropologue
Fondateur du
structuralisme