Science des Réseaux
Un résumé

UNIVERSITÉ DE LYON

Proposé par
Rémy Cazabet

1 Définition et Vocabulaire

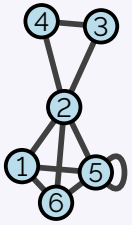
Réseaux : notation graphe

Notation graphe : $G = (V, E)$

V	Ensemble de nœuds/Vertex.
E	Ensemble de liens
$u \in V$	un nœud.
$(u, v) \in E$	un lien.

Réseaux : notation graphe

Graphe



Notation graphe

$$G = (V, E)$$
$$V = \{1, 2, 3, 4, 5, 6\}$$
$$E = \{(1, 2), (1, 6), (1, 5), (2, 4), (2, 3), (2, 5), (2, 6), (6, 5), (5, 5), (4, 3)\}$$

Types de réseaux

Réseaux simples: Les liens peuvent soit exister, soit ne pas exister entre les nœuds. Il n'y a pas de boucles (liens d'un nœud vers lui-même.
Graphe dirigé:Les liens ont une direction: $(u, v) \in V$ n'implique pas $(v, u) \in V$
Graphes pondérés: Un poid est associé à chaque lien, pour indiquer sa *force* par exemple.

D'autres types de graphes existent (multigraphes, multipartite, hypergraphes, etc.)

Compter les nœuds et les liens

N/n	**taille**: nombre de nœuds $\|V\|$.
L/m	nombre de liens $\|E\|$
L_{max}	Nombre maximal de liens

Réseaux non-dirigés: $\binom{N}{2} = N(N-1)/2$

Réseaux dirigés: $\binom{N}{2} = N(N-1)$

Description des nœuds/liens

N_u	**Voisins** de u, nœuds qui partagent un lien avec u.
k_u	**Degré** de u, nombre de voisins $\|N_u\|$.
N_u^{out}	**Successeurs** de u, nœueds tels que $(u, v) \in E$ dans un graph dirigé
N_u^{in}	**Prédécesseurs** de u, nœuds tels que $(v, u) \in E$ dans un graphe dirigé
k_u^{out}	**Degré sortant** de u, Nombre de liens dont u est l'origine $\|N_u^{out}\|$.
k_u^{in}	**Degré entrant** de u, nombre de liens qui ont pour destination $\|N_u^{in}\|$
$w_{u,v}$	**Poid** d'un lien (u, v).
s_u	**Force** de u, somme des poids des liens adjacents, $s_u = \sum_v w_{uv}$.

Description de réseaux - Nœuds/Liens

$\langle k \rangle$	**Degré moyen**: Les réseaux réels sont *clairsemé*(**sparse**), i.e., typiquement le degré est petit par rapport au nombre de nœuds: $\langle k \rangle \ll n$. Augmente lentement avec le nombre de nœuds, e.g., $d \sim \log(m)$

$$\langle k \rangle = \frac{2m}{n}$$

$d/d(G)$	**Densité**: Fraction des paires de nœuds connectées dans G.

$$d = L/L_{\max}$$

Chemins - Marches - Distance

Marche: Séquence de nœuds ou liens adjacents (e.g., **1.2.1.6.5** est une marche valide)
Chemin: Une marche dans laquelle tous les nœuds sont distincts.
Longueur d'un chemin: nombre de **liens** traversés par un chemin
Longueur pondérée d'un chemin: Somme des poids des liens sur un chemin
Plus court chemin: Le plus court chemin entre deux nœuds u, v est un chemin de *longueur* minimale. Souvent, il n'y en a pas qu'un seul.
Plus court chemin pondéré: Chemin de plus court *chemin pondéré*.
$\ell_{u,v}$**: Distance**: La distance entre les nœuds u, v est la longueur de plus court chemin entre eux.

Description de réseaux - Chemins

$\ell_{\max}$	**Diametre**: *distance* maximale entre 2 nœuds du réseau.
$\langle \ell \rangle$	**Distance moyenne**, i.e., moyenne des distances entre toutes les paires de nœuds:

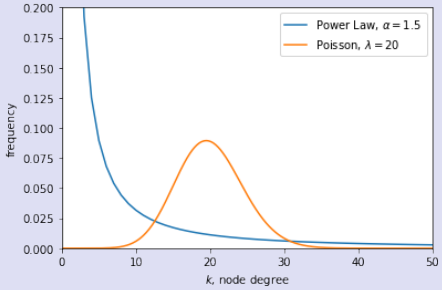
$$\langle \ell \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$$

Distribution de degré

La distribution des degrés des nœuds est considéré une propriété importante. On cherche principalement à savoir si elle suit l'une de ces deux formes :

- **Courbe en cloche** Distribution (Normale/Poisson/Binomiale)
- **Sans-échelle**(Scale-Free), aussi appelé *longue-queue*(long-tail) ou *Loi de puissance* (Power-law)

Une courbe en cloche à une *échelle*, une *valeur type*: la taille des êtres humains, par exemple: la plupart des personnes ont une taille assez proches de la valeur moyenne. Une distribution sans échelle est différente, un exemple est la richesse des individus: Une grande partie des valeurs sont faibles, mais quelques valeurs sont extrêmement élevées. La moyenne n'est pas du tout représentative de l'ensemble des valeurs.



sous-graphes

Sous-graphe $H(W)$ (Sous-graphe induit): ensemble des nœuds W du graphe $G = (V, E)$ et les liens qui les connectent dans G, i.e., sous-graphe $H(W) = (W, E'), W \subset V, (u, v) \in E' \iff u, v \in W \wedge (u, v) \in E$
Clique: sous-graphe de densité 1: $d = 1$
Triangle: clique de taille 3
Composante connexe: un sous-graphe tels que tous les nœuds sont connectés par un chemin, et pour lequel il n'y a pas de lien vers les autres nœuds de réseau.
Composante fortement connexe: Dans un graphe dirigé, une composante connexe si l'on prend en compte les directions des liens.
Composante faiblement connexe: Dans un graphe dirigé, une composante connexe si l'on ne prend pas en compte les directions des liens

Triangles

δ_u - **Triades de u:** nombre de triangles contenant le node u
Δ - **Nombre de triangles dans le graphe** $\Delta = \frac{1}{3}\sum_{u\in V}\delta_u$.

> Chaque **triangle** dans le graphe est compté comme une **triade** une fois par chacun des nœuds qui le compose.

$\delta_u^{\max}$ - **Potentiel de triangle de u:** Nombre maximal de triangles qui peuvent exister contenant u, étant donné son degré: $\delta_u^{\max} = \tau(u) = \binom{k_i}{2}$
$\Delta^{\max}$ - **Potentiel de triangle de G:** Nombre maximal de triangles qui peuvent exister dans le graphe, étant donné sa distribution de degré. $\Delta^{\max} = \frac{1}{3}\sum_{u\in V}\delta^{\max}(u)$

Coefficient de clustering

Le coefficient de clustering est une mesure de la fermeture transitive présente dans un graphe. La fermeture transitive est une notion qui viens de l'analyse de réseaux sociaux, souvent résumée par l'aphorisme: *Les amis de mes amis sont mes amis*. Plus les voisins des voisins d'un nœud n ont tendance à être des voisins de n, plus le coefficient de clustering est élevé.

C_u - **Clustering coefficient d'un nœud:** densité du sous-graphe induit par les voisins du nœud u, $C_u = d(H(N_u))$. Aussi interprété comme la fraction de tous les triangles possibles dans N_u qui existent, $\frac{\delta_u}{\delta_u^{\max}}$
$\langle C\rangle$ - **Coefficient de clustering moyen:** Moyenne des coefficients de clustering de tous les nœuds du graphe, $\bar{C} = \frac{1}{N}\sum_{u\in V} C_u$.

> Attention en interprétant cette valeur : les nœuds de faible degrés sont généralement majoritaires dans les graphes réels, et leur valeur de clustering C est très sensible, i.e., pour un nœud u de degré 2, $C_u \in 0, 1$, tandis que les nœuds de fort degré ont tendance à avoir des scores plus contrastés.

C^g - **Coefficient de clustering global:** Fraction de tous les triangles possibles qui existent dans le graphe, $C^g = \frac{\Delta}{\Delta^{\max}}$

Réseau petit monde

Un réseau est dit **petit monde**(Small world) lorsqu'il a certaines propriétés structurelles[a]. La définition n'a pas vraiment de définition quantitative, mais correspond aux propriétés suivantes:

- La distance moyenne doit être courte, i.e., de l'ordre de grandeur du log du nombre de nœud: $\langle\ell\rangle \approx \log(N)$
- Le coefficient de Clustering doit être grand, i.e., largement supérieur à celui d'un graphe aléatoire de propriétés équivalente , e.g., $C^g \gg d$, avec d la densité du graphe.

La propriété petit monde est considéré characteristique des *réseaux réels*, par opposition aux réseaux aléatoires. On associe cette propriété à certaines characteristiques du graphe (robustesse aux défaillances, circulation efficace de l'information, etc.).

Attention: dans certains contextes, *réseau petit monde* peut désigner simplement un réseau dont la distance moyenne est courte, indépendemment de son coefficient de clustering.

---

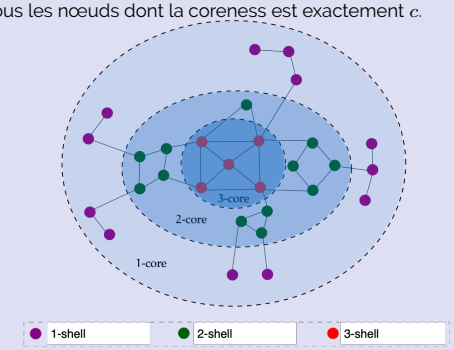
[a] Watts and Strogatz 1998.

cœurs et Shells

Beaucoup de réseaux réels sont connus pour avoir une structure dite en **cœur-périphérie**, i.e., il y a un cœur qui est densément connecté et une zone périphérique dont les nœuds sont faiblement connecté entre eux et avec le cœur.

---

k-cœur: le k-cœur (cœur d'ordre k) du graphe $G(V, E)$ est le plus grand sous-graphe $H(C)$ tel que tous ses nœuds ont au moins un degré k, i.e., $\forall u \in C, k_u^H \leq k$, avec k_u^H le degré du nœud u dans le sous-graphe H.
coreness: Un nœud u a une coreness k s'il appartient au k-cœur mais pas au $k+1$-cœur.
c-shell: Tous les nœuds dont la coreness est exactement c.



Vocabulaire

Singleton: ou nœud isolé, nœud de degré nul $k = 0$
Hub: nœud u de large degré, i.e., $k_u \gg \langle k\rangle$

Pont: nœud ou lien qui, s'il est enlevé, sépare le graphe en plusieurs composantes connexes. **Bout**: un bout est un demi-lien, i.e., le lien (u, v) a un bout connecté à u et un autre connecté à v.

Réseau complet: réseau où tous les liens possibles existe: $L = L_{max}$
Réseau clairsemé (sparse): réseau ayant peu de lien, $d \ll 1$, $L \ll L_{max}$
Graphe connecté: Graphe qui n'a qu'une seule composante connexe

2 Réseaux en tant que matrices

Les matrices en quelques mots

Les matrices sont des objets mathématiques qui sont des *tables* de nombres. La taille d'une matrice est exprimée comme $m \times n$, pour une matrice avec m lignes et n colonnes. **l'ordre (ligne/colonne) est important**.
M_{ij} représente l'élément sur la **ligne** i et **colonne** j.

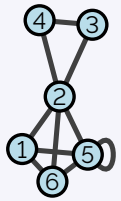
A - Matrice d'adjacence

La méthode la plus courante pour représenter un graphe par une matrice consiste à créer une matrice d'adjacence A. C'est une matrice carrée dont le nombre de lignes et de colonnes est égal au nombre de nœuds N du graph. Les nœuds du graphe sont numérotés de 1 à N, et il y a un lien entre les nœuds i et j si la valeur à la position A_{ij} n'est pas 0.

- Une valeur sur la diagonale représente une **boucle**
- si le graphe est **non dirigé**, la matrice est **symmétrique**: $A_{ij} = A_{ji}$ pour tout i, j.
- Dans un graphe **non pondéré**, les liens sont représentés par la valeur 1.
- Dans un graphe **pondéré**, la valeur A_{ij} représente le **poids** du lien (i, j)

Notation matricielle - Exemple

Graph



A - Adjacency Mat.

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Science des Réseaux
Un résumé

UNIVERSITÉ DE LYON

Proposé par
Rémy Cazabet

3 Indices structurels

Indices structurels de nœuds

Les indices structurels de nœuds, souvent appelés *centralité*, mesurent à quel point un nœud occupe un certain type de position dans la structure du graphe. Cette notion est parfois résumée comme *une mesure de l'importance des nœuds*, cependant *importance* et la notion d'être *central* sont des notions subjectives. Une centralité, malgré son nom, ne mesure donc pas forcément à quel point le nœud est central ou important, mais plutôt à quel point sa position est représentative du type de position mesuré par cette centralité.

Centralité de degré

La centralité de degré est l'une des plus utilisés. Elle peut souvent être interprétée comme une mesure de popularité, e.g., plus j'ai de relations, d'amis dans un réseau social, le plus *important* je suis dans ce réseau.

Centralité de proximité / Harmonique

La centralité de proximité (closeness) d'un nœud mesure à quel point il est proche de tous les autres nœuds. Pour interpréter ce score, un parallele peut être fait avec la position d'un point dans un cercle: le point qui est le plus proche de tous les autres points du cercle est son centre. Le nœud de plus grande closeness est l'équivalent pour ce graphe du centre pour un cercle. Formellement, le plus simple est de le définir comme l'inverser de la *farness*.

Farness: Distance moyenne à tous les nœuds du graphe.

$$\text{Farness}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \ell_{u,v}$$

Closeness: Inverse de la farness

$$\text{Closeness}(u) = \frac{N-1}{\sum_{v \in V \setminus u} \ell_{u,v}}$$

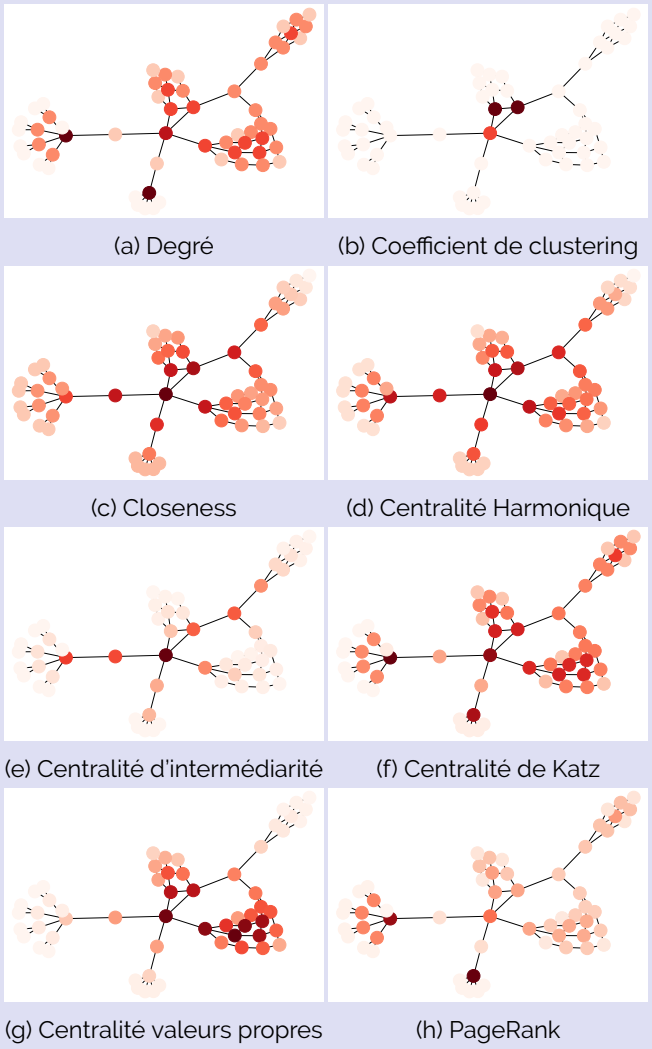
Centralité Harmonique: Une variante de la closeness définie comme la moyenne des inverses des distances à tous les autres nœuds (moyenne harmonique). Cette mesure est définie même sur des graphes non connexes, à condition de définir $\frac{1}{\infty} = 0$. Son interprétation est la même que la Closeness.

$$\text{Harmonic}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \frac{1}{\ell_{u,v}}$$

Coefficient de Clustering

Ce score, déjà défini, mesure la *fermeture transitive* d'un nœud. Un score élevé est souvent interprété comme un nœud qui appartient fortement à une et une seule communauté (les amis de mes amis sont mes amis, car nous appartenons tous au même groupe). Un score faible peut signifier que le nœud joue le rôle de pont: il n'y a pas de connections entre mes amis car ils appartiennent à des cercles sociaux différents.

Centralité - exemples



(a) Degré (b) Coefficient de clustering

(c) Closeness (d) Centralité Harmonique

(e) Centralité d'intermédiarité (f) Centralité de Katz

(g) Centralité valeurs propres (h) PageRank

Centralité de Katz

La centralité de Katz est considérée comme une mesure du potentiel d'influence du nœud. Pour un nœud u, elle est définie comme la somme, pour tous les marches de distance ℓ, du nombre de nœuds situés à une distance exactement ℓ de u, diminuté d'un facteur décroissant rapidement lorsque ℓ augmente. L'intuition est que, plus le nombre de nœuds qui peuvent être atteind en un faible nombre de sauts est grand, plus la valeur est élevée. Plus formellement, elle est définie comme:

$$C_{\text{Katz}}(u) = \sum_{\ell=1}^{\infty} \sum_{v=1}^{N} \alpha^{\ell} (A^{\ell})_{vu}$$

avec A^{ℓ}_{vu} le nombre de marches de longueur ℓ de v à u, et $\alpha < \frac{1}{\lambda_i}$ un paramètre plus petit que la plus grande valeur propre de A, Ce qui permet de calculer ce score en forme matriciel:

$$C_{\text{Katz}}(u) = ((I - \alpha A^T)^{-1} - I)\vec{1}$$

Notons que dnas un graphe dirigé, la centralité de Katz doit être interprété comme un mécanisme de *vote*: une centralité plus importante de u signifie que plus de nœuds peuvent atteindre u rapidement, et non que u peut atteindre de nombreux nœuds rapidement.

Centralité d'intermédiarité

La centralité d'intermédiarité (betweenness) mesure à quel point le nœud joue le rôle de pont, d'intermédiaire. Plus le score est haut, plus le nœud est essentiel au déplacement rapide dans le graphe.Plus formellement, la betweenness de u est définie comme la fractions de tous les plus courts chemins entre toutes les paires de nœuds du graphe (sauf u) qui passent par u. Par conséquent, si nous enlevons un nœud de betweenness élevé, de nombreux plus courts chemins vont devenir plus long, et donc la circulation dans le graphe sera plus difficile. Un cas extrême est celui d'un nœud qui est le seul point de passage entre deux groupes de nœuds: si on le retire, la circulation n'est plus du tout possible entre certains sous-graphes. Ces nœuds ont donc tendance à avoir un score de betweenness très élevé. Elle est définie comme:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

avec σ_{st} le nombre de plus court chemins entre s et t et $\sigma_{st}(v)$ le nombre de ces chemins qui passent par le nœud v.
La betweenness tend à augmenter avec la taille du graphe. Une version normalisée peut être obtenue en divisant par le nombre de paires de nœuds, pour un graphe dirigé: $C_B^{\text{norm}}(v) = \frac{C_B(v)}{(N-1)(N-2)}$.

Centralité Eigenvector

La centralité Eigenvector, ou centralité de vecteurs propres (eigenvector en anglais), est une définition récursive de l'importance: un nœud est important s'il est connecté à des nœuds importants. En pratique, elle est définie de la manière suivante: la centralité eigenvector C_u de chaque nœud u est telle que si chaque nœud *envoie* son score de centralité a ses voisins, alors la somme des scores reçus par chaque nœud est égale à λC_u (avec λ une constante de normalisation). Plus formellement,

$$C_u^{t+1} = \frac{1}{\lambda} \sum_{v \in N_u^{in}} C_v^t \quad (1)$$

Cette définition récursive peut être interprétée en termes de valeurs propres et vecteurs propres, d'où le nom. Un vecteur propre d'une matrice est défini par la relation $Ax = \lambda x$, avec x un vecteur propre, et λ la valeur propre correspondante. La centralité eigenvector est définie par le vecteur propre associé à la plus grande valeur propre, qui est la seule solution pour laquelle tous les élements du vecteur propre sont positifs. Une méthode simple pour calculer la centralité eigenvector est appelé la méthode des puissances itérées: des valeurs aléatoires sont attribuées à tous les nœuds, puis on répète l'équation 1. Au bout d'un certain nombre d'itération, il est prouvé que le résulta converge vers des valeurs fixes: la centralité eigenvecteur.

La centralité Eigenvector ne peut pas être calculée, en général, sur des graphes orientés, à cause de l'existence de *nœuds source*, i.e., $k^{in} = 0$. Ces nœuds ont, par définition, un score de centralité de 0 a $t+1$, et donc *envoient* une valeur de 0 a $t+2$, qui pourront de ce fait avoir maintenant un score de centralité de 0, qu'ils transmettront au prochain tour, et aisi de suite jusqu'à ce que tous ou une grande partie des nœuds finissent avec une centralité de 0.

Pagerank

La centralité Pagerank[a] est célèbre pour avoir été utilisée par Google pour classer les résultats de son moteur de recherche: Un score de Pagerank est calculé pour chaque page, sur le graph ou les nœuds sont des pages et les liens des liens hypertextes, puis, lorsque l'on recherche un ensemble de mots, toutes les pages qui contiennent ces mots sont retournées, classées par leur score PageRank. Aujourd'hui, les méthodes utilisées par Google sont plus complexes, notamment parce qu'elles personnalisent les résultats en fonction des utilisateurs.
Cette méthode est une variante de la centralité Eigenvector, permettant notamment de résoudre le problème des nœuds source.
Pagerank introduit deux nouveautés: 1) à chaque étape t, chaque nœud gagne une petite valeur constante, 2) Les valeurs envoyées par chaque nœud sont divisées également parmi tous ses successeurs (normalisation par le degré). L'équation 1 deviens donc:

$$C_u^{t+1} = \alpha \sum_{v \in N_u^{in}} \frac{C_v^t}{k_v^{out}} + \beta \quad (2)$$

Avec, par convention, $\beta = 1, \alpha \in [0, 1]$ un paramètre.
Pagerank peut aussi être exprimé comme le vecteur propre associé à la plus grande valeur propre d'une matrice appelé la **matrice google** G, définie telle que $G_{ij} = \alpha S_{ij} + (1-\alpha)/n$, avec S_{ij} la matrice d'adjacence normalisée par colonne.

[a] Page et al. 1999.

Indices structurels de liens

La position des liens dans le réseau peut aussi être décrite en utilisant des centralités de liens, généralement similaires à celles définies sur les nœuds.
Clustering C^e pour un lien (u, v) est définit comme la fraction des voisins d'au moins l'un des deux nœuds aux extrémités qui est voisin des deux simultanément.

$$C^e(u, v) = \frac{|N_u \cap N_v|}{|N_u \cup N_v| - 2}$$

Les liens de fort clustering sont dits *Intégratifs*, les liens de faible clustering sont dit dispersifs.
Intermediarité des liens est définie exactement comme la centralité d'intermédiarité des nœuds, mais en comptant le nombre de plus courts chemins qui passent par le lien au lieu du nœud, i.e.,

$$C_B(u, v) = \sum_{s \neq t \in V} \frac{\sigma_{st}(u, v)}{\sigma_{st}}$$

avec σ_{st} le nombre de plus courts chemins entre les nœuds s et t et $\sigma_{st}(u, v)$ le nombre de ces chemins qui passent par le lien (u, v).

References

[1] Lawrence Page et al. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[2] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *nature* 393.6684 (1998), pp. 440–442.