GEPHI
ANALYSE DE RÉSEAUX

QUI SUIS-JE

- Rémy Cazabet
- Maître de conférences
  - Université Lyon 1
  - LIRIS, DM2L Team (Data Mining & Machine Learning)
- Informatique => Science des réseaux
- Contact me: remy.cazabet@univ-lyon1.fr
- http://cazabetremy.fr

OBJECTIFS DU COURS

- Savoir comment une base de données bibliographique en accès libre, telle HAL, peut être interrogée en utilisant un language de programmation

- Savoir comment on peut passer de données brutes à des données modélisées sous la forme d'un graphe, et les bonnes questions à se poser

- Connaître les bases du domaine de la science des réseaux (Network Science), permettant de décrire et d'analyser des données représentées sous forme de graphes

- Savoir utiliser le logiciel libre Gephi pour

  - 1)Calculer des indicateurs d'analyse de réseaux et

  - 2)Produire des visualisations sous forme de réseaux de données de co-citations ou de collaborations.

AUJOURD'HUI

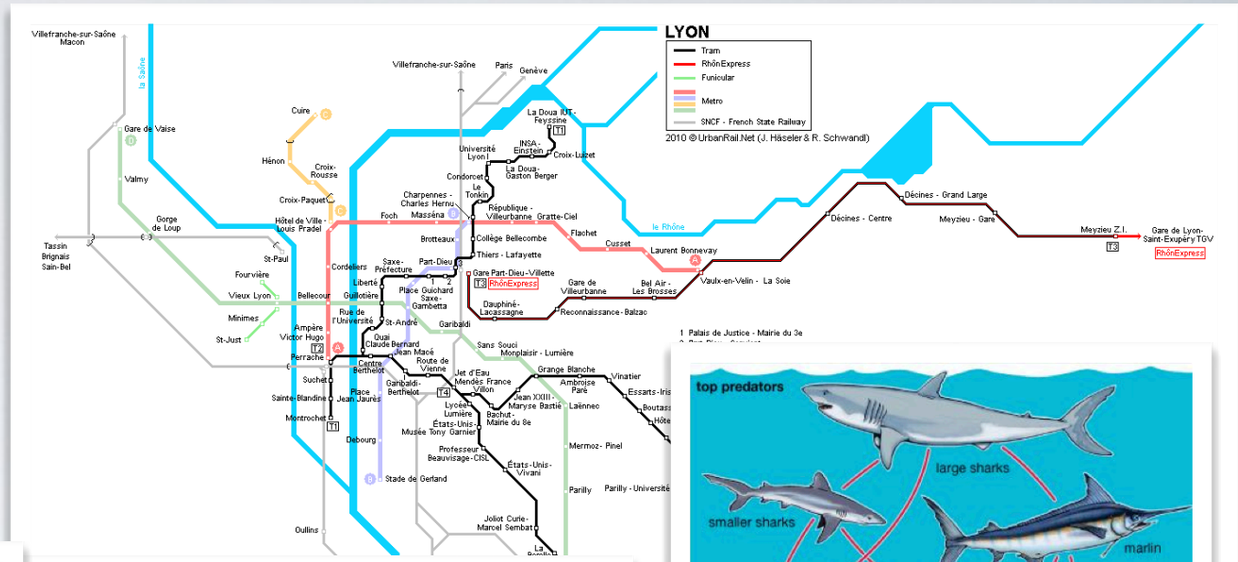
- Introduction à la science des réseaux: Décrire un graphe
- Introduction à Gephi: visualisation

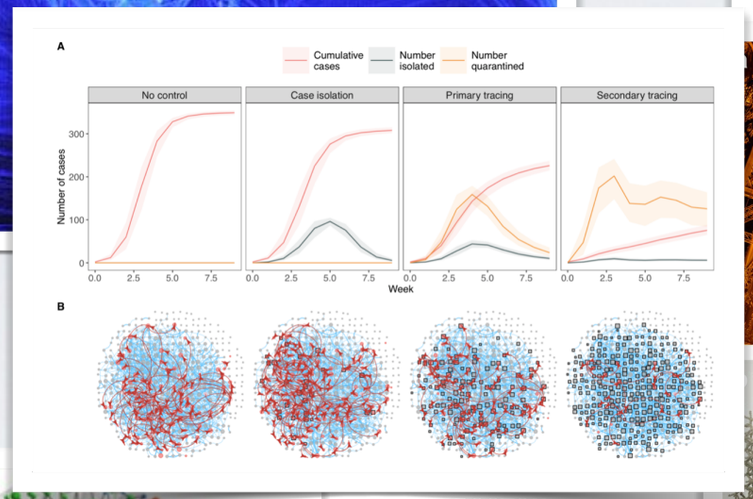
GEPHI

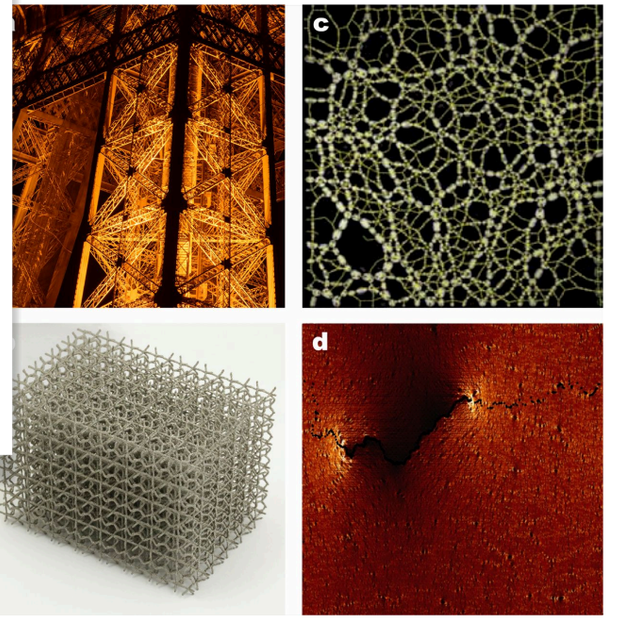
- Un logiciel pour visualiser des données sous forme de graphes, et pour des analyses réseaux simples

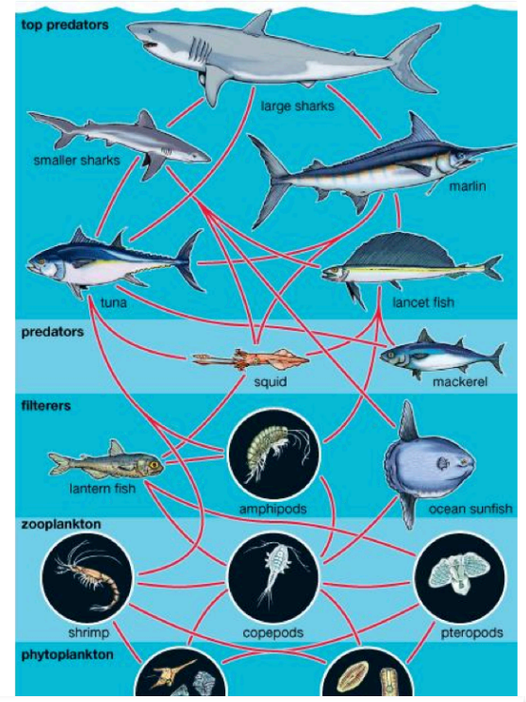
- Pour aller plus loins dans les analyses, programmation:
  - Python: Networkx, igraph, graph-tool, etc.

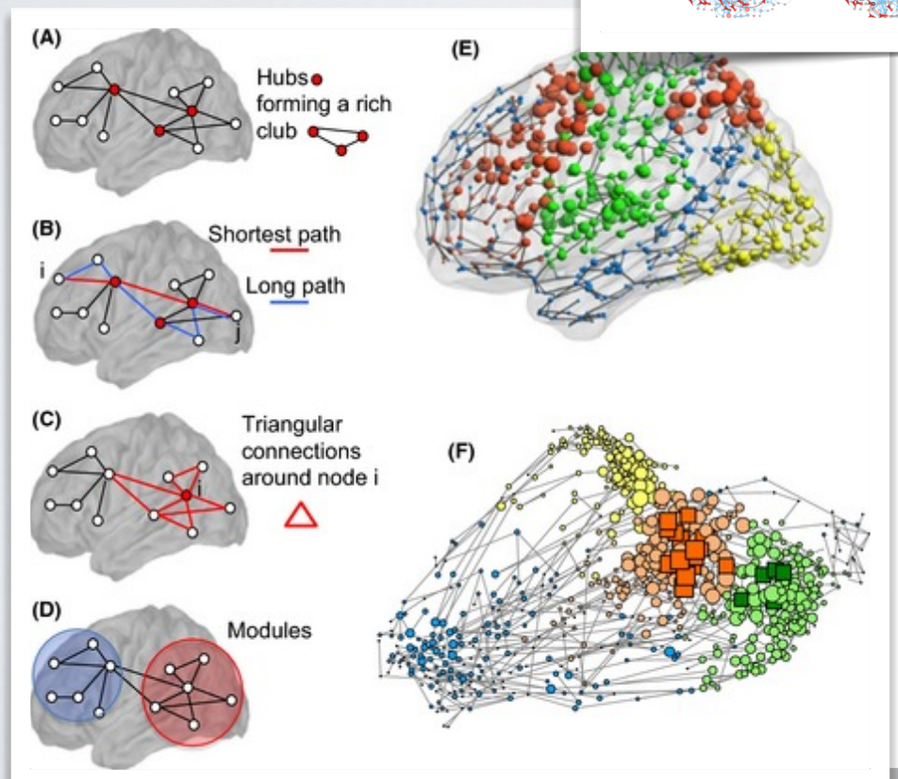
SCIENCE DES RÉSEAUX

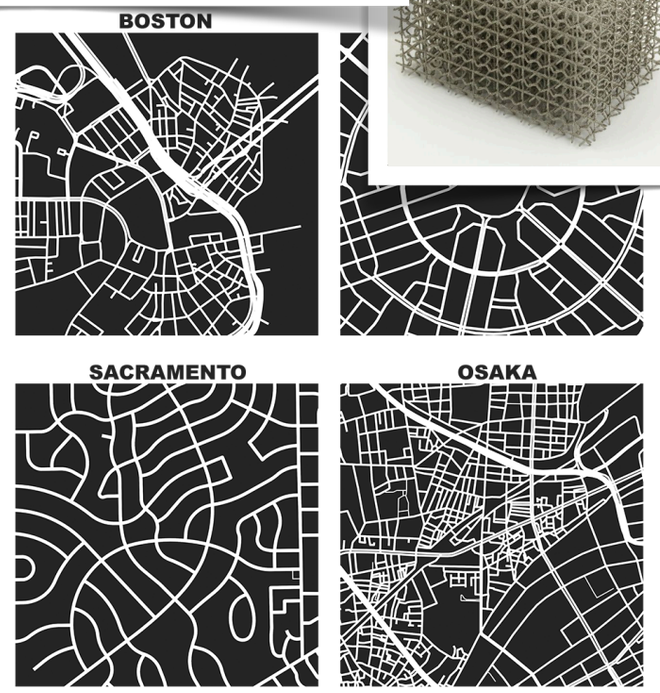


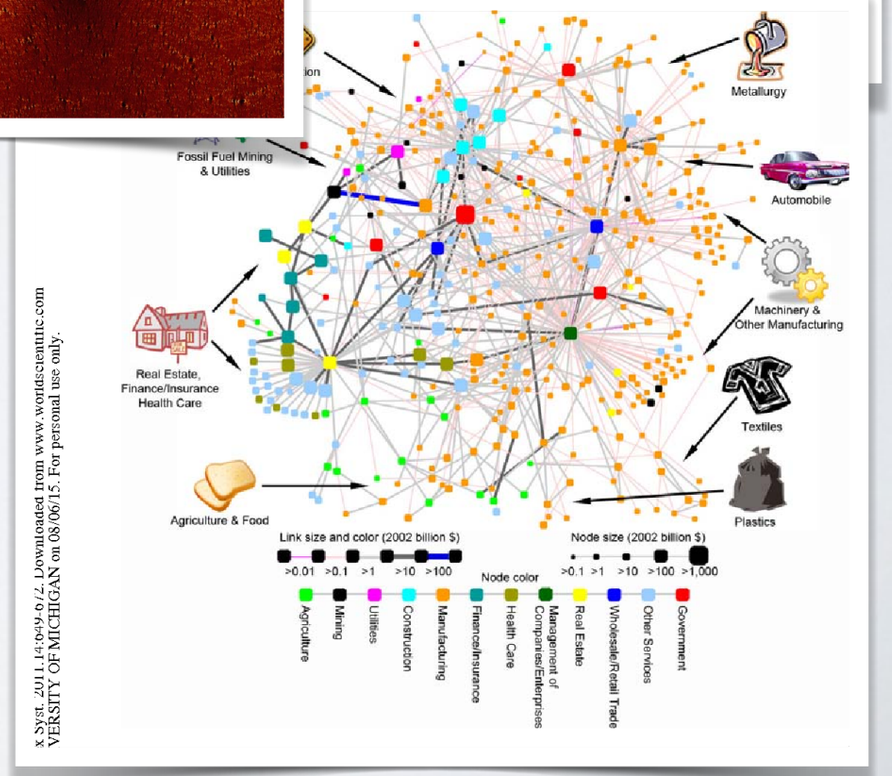


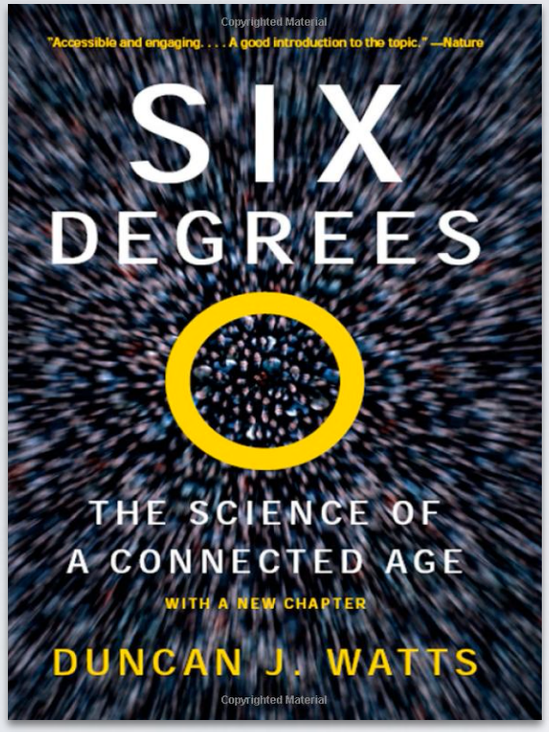


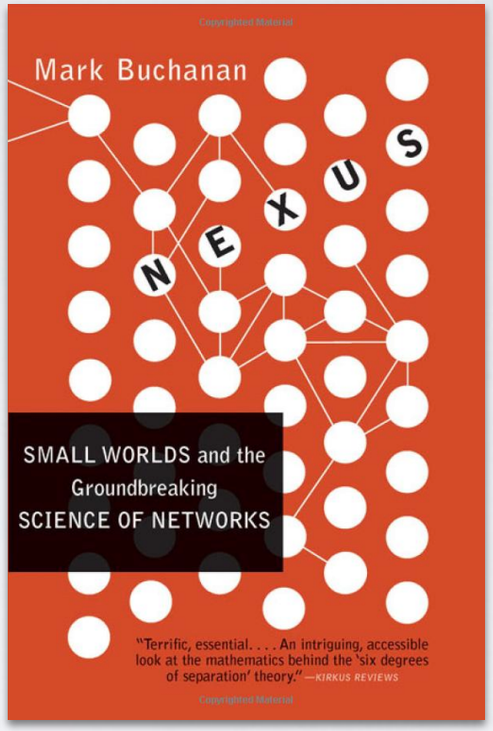


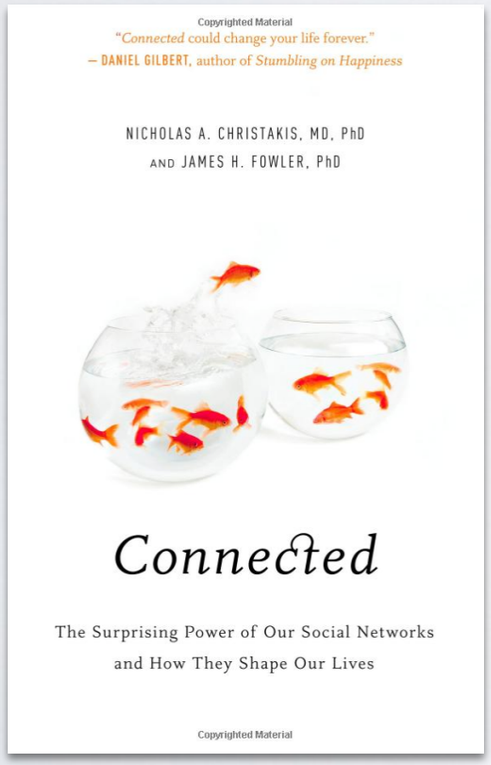


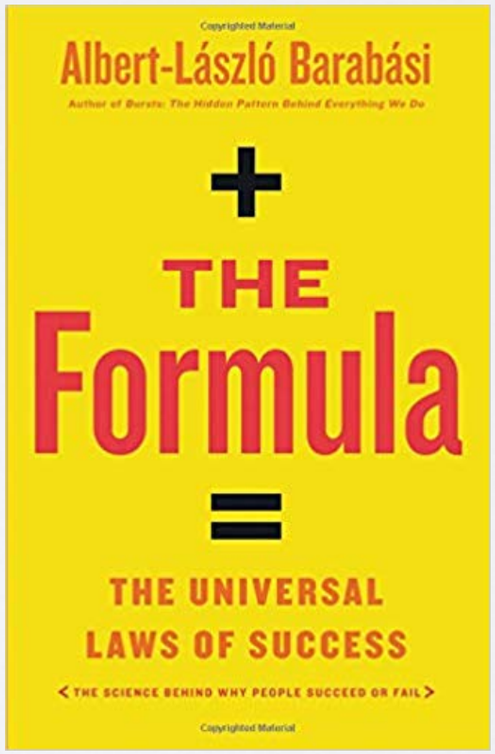








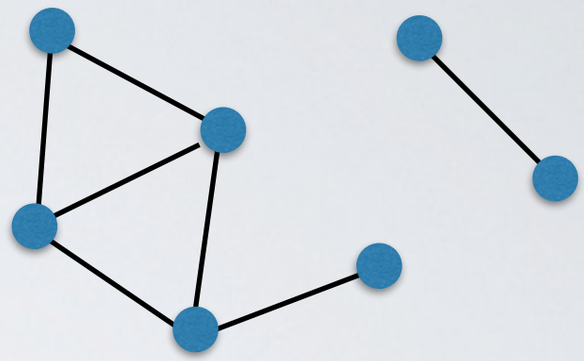




J'ai une copie que je peux prêter

GRAPHES ET RÉSEAUX

GRAPHS & NETWORKS



Réseaux : Objet réel

- www,
- Réseau social
- Réseau autoroutier
- Vocabulaire: (Réseau, nœud, lien)

Graphe : Représentation mathématique d'un réseau:
Vocabulaire: (Graphe, vertex, arête)

J'utilise les deux termes de manière interchangeable

Vertex	Lien
Personne	Amitié
neurone	synapse
Website	hyperlien
Auteur	co-écrit
gène	Régulation

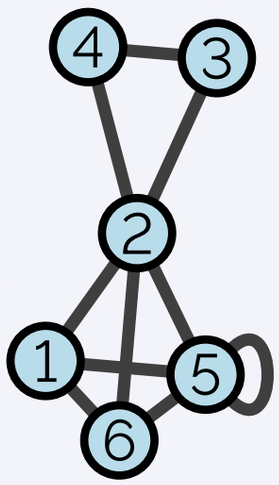
Réseaux : notation graphe

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

$$G = (V, E)$$

$$V = \{1, 2, 3, 4, 5, 6\}$$

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

LES GRAPHES 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.

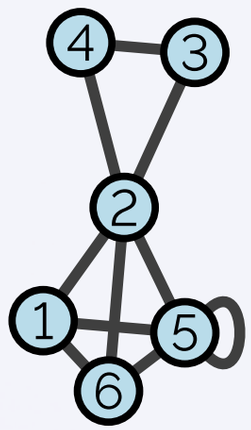
ADJACENCY MATRIX

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 **symétrique**: $A_{ij} = A_{ji}$ 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)

Graph



A - Adjacency Mat.

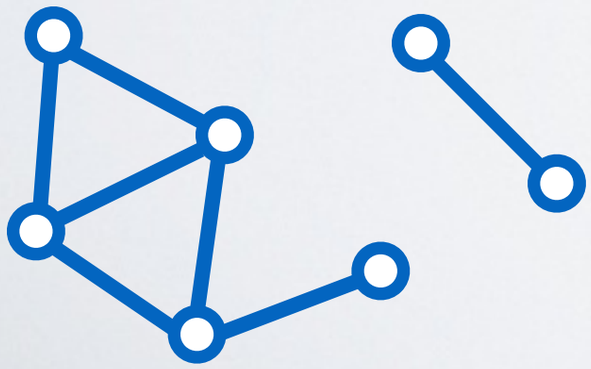
$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Types de Réseaux

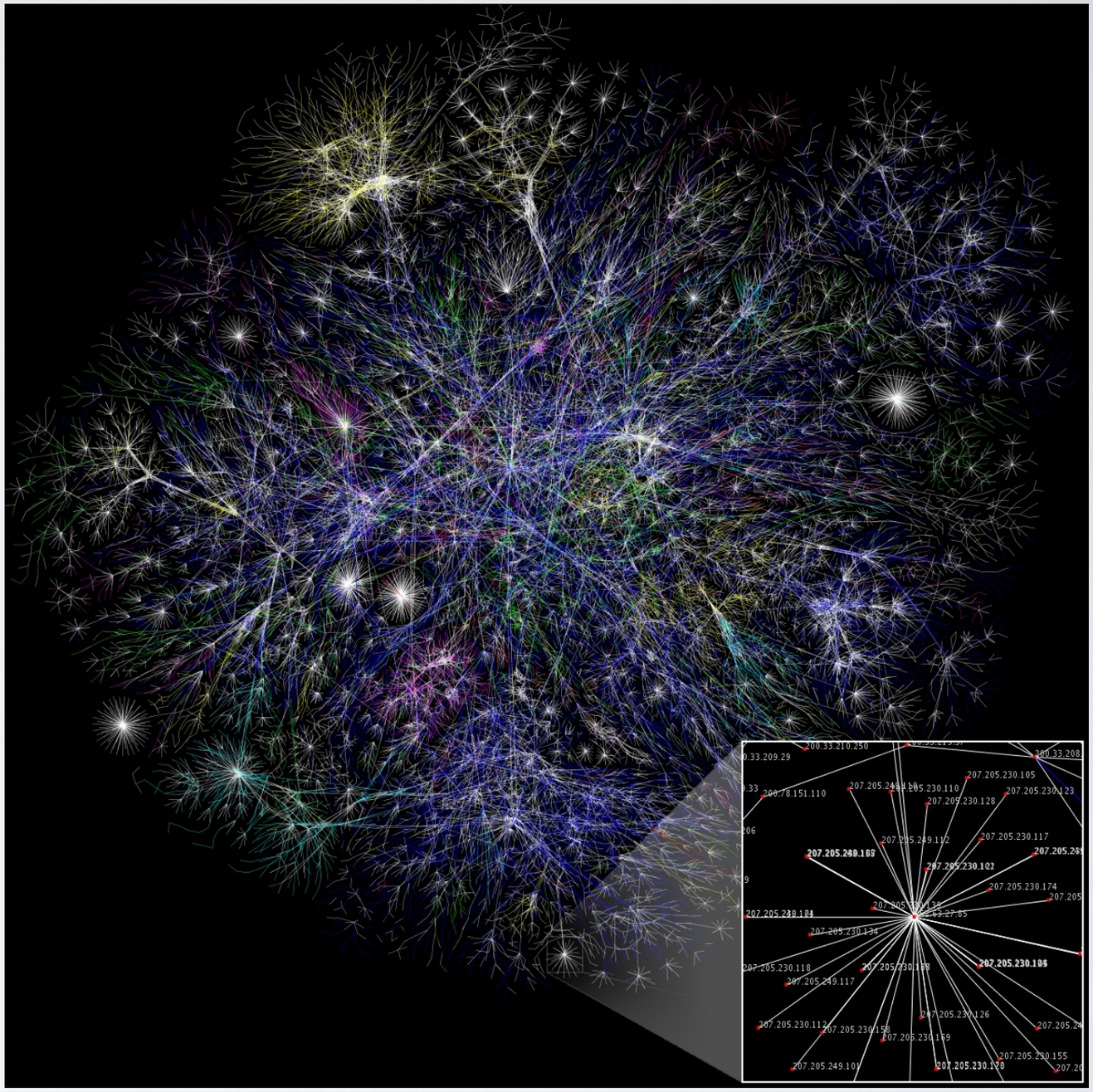
Non dirigés

$G=(V, E)$

$(u,v) \in E \equiv (v,u) \in E$



Opte project

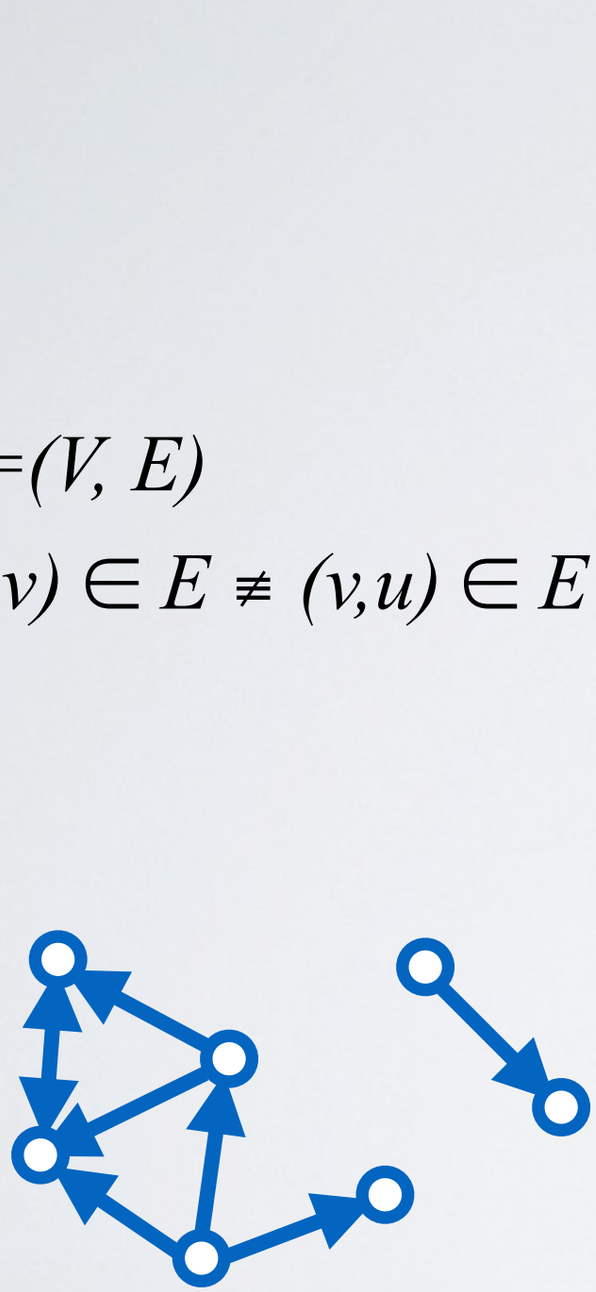


The Internet: Nodes - routers, Links - physical wires

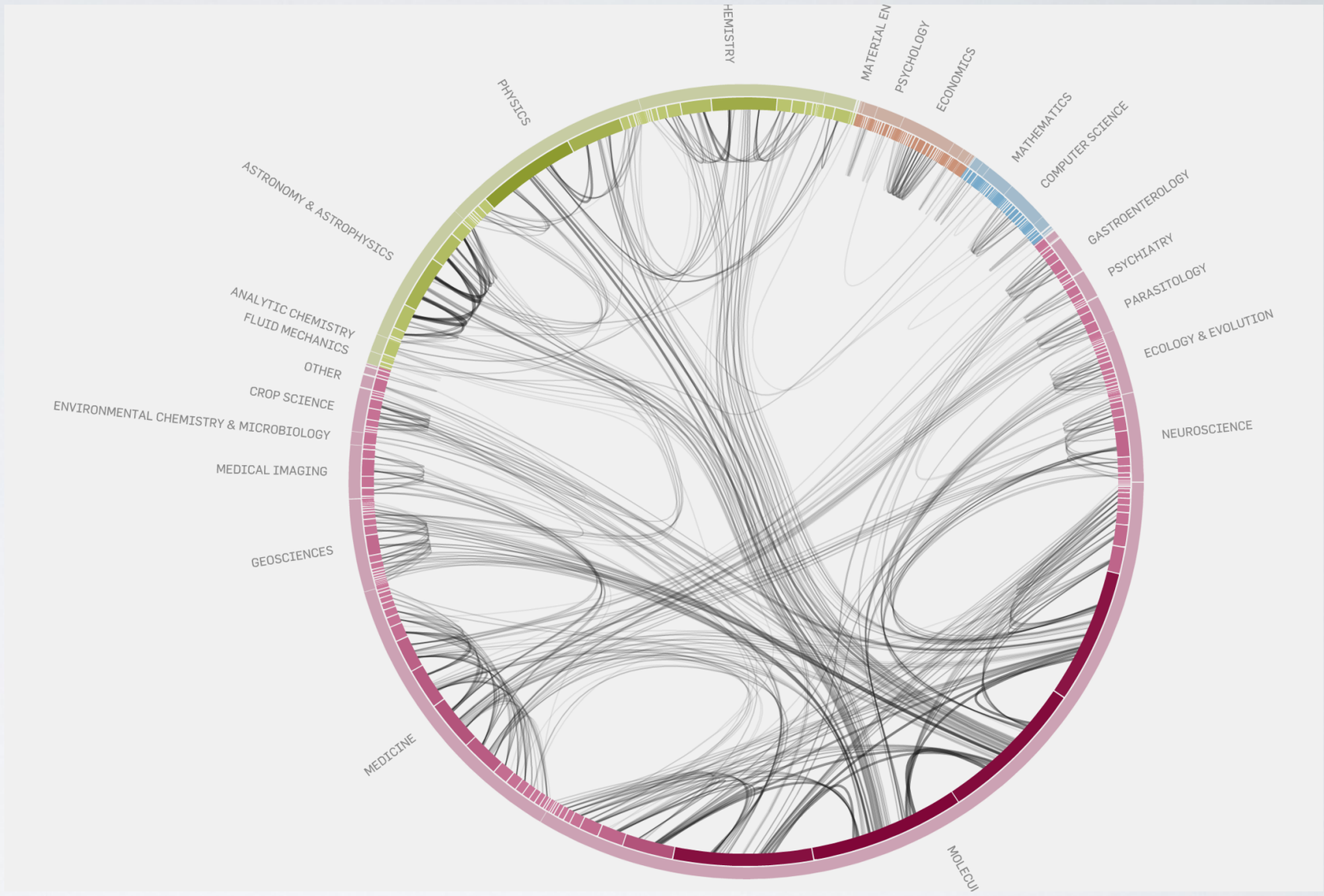
Dirigé

$$G=(V, E)$$

$$(u,v) \in E \not\equiv (v,u) \in E$$



Moritz Stefaner, eigenfactor.com



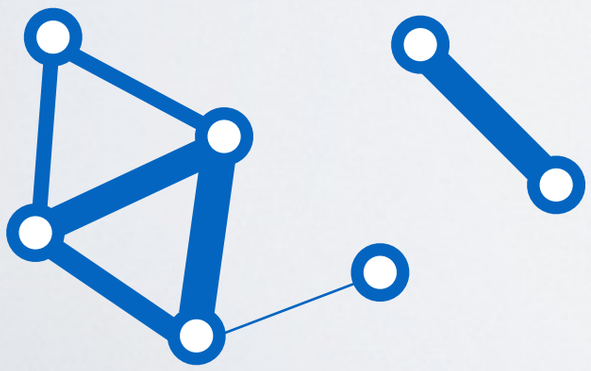
Citation network: Nodes - publications, Links - references

Réseaux pondérés

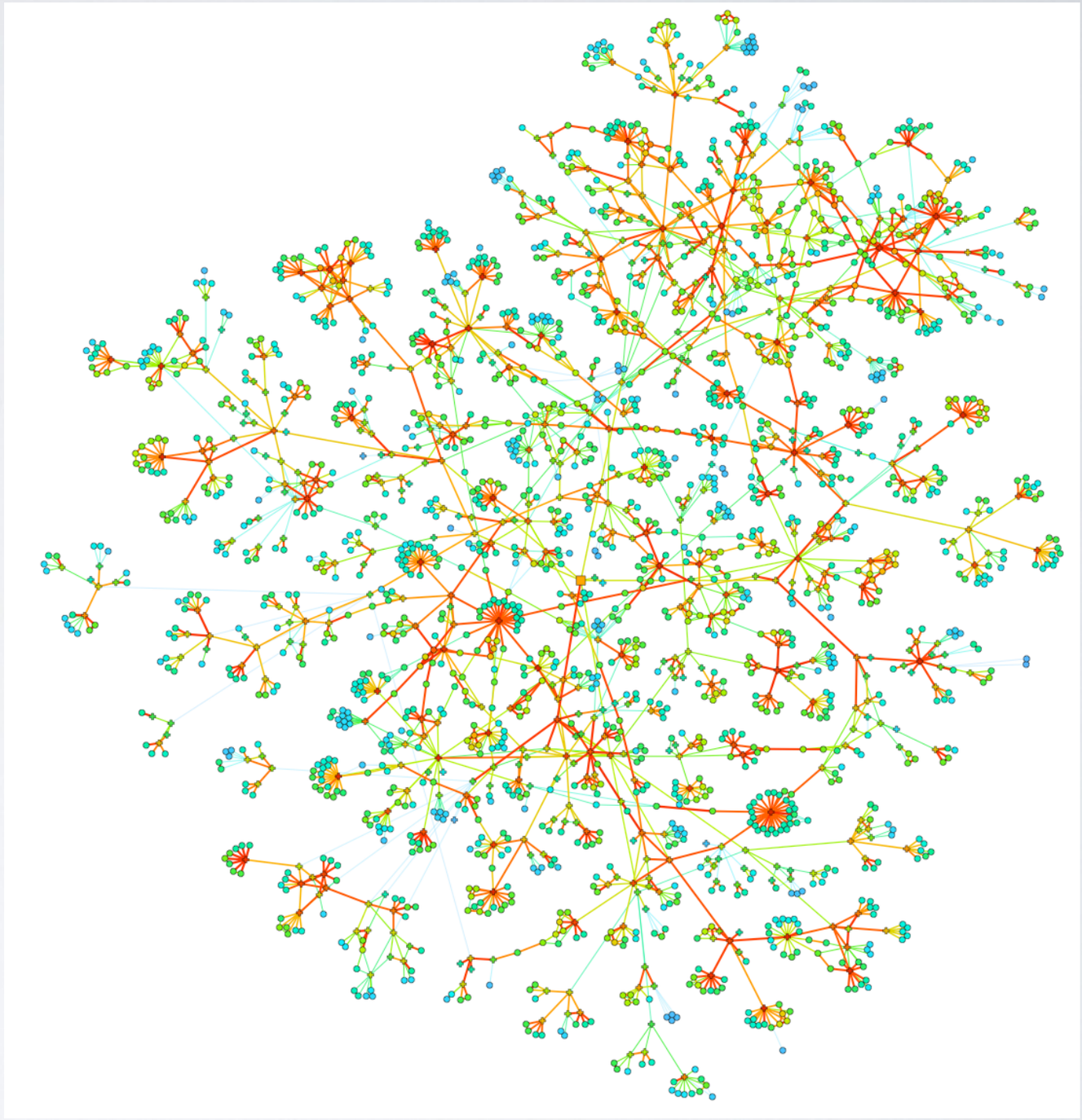
$G=(V, E, w)$

$w: (u,v) \in E \Rightarrow R$

- La force des liens est représentée par un poids

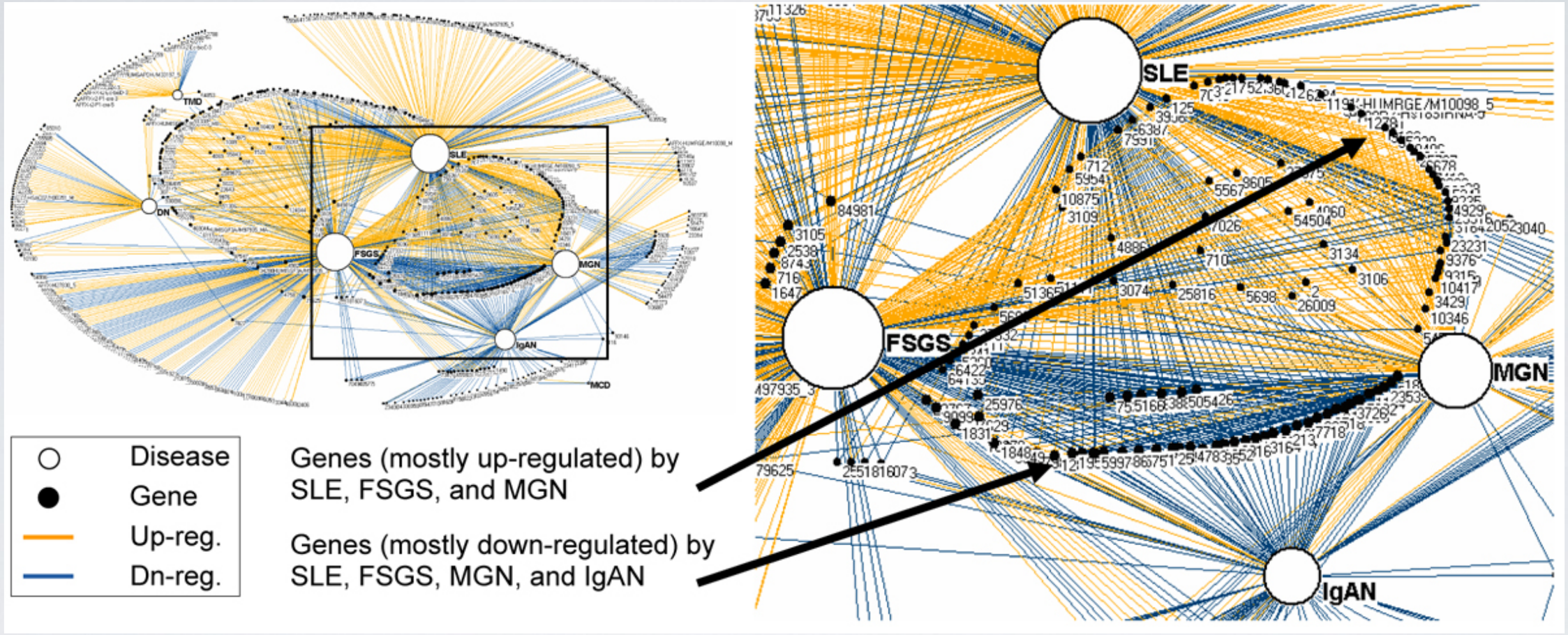


Onnela et.al. New Journal of Physics 9, 179 (2007).

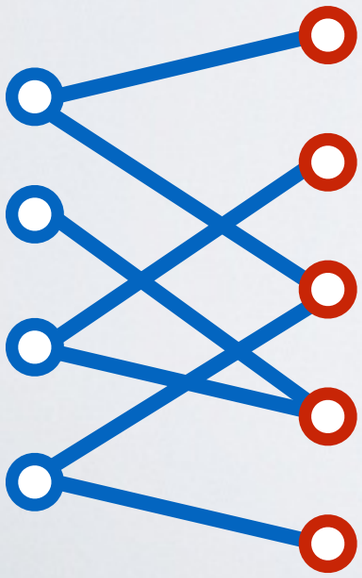


Social interaction network: Nodes - individuals
Links - social interactions

Réseaux bipartite



Bhavnani et.al. BMC Bioinformatics 2009, **10**(Suppl 9):S3



Gene-desease network:

Nodes - Desease (7)&Genes (747)

Links - gene-desease relationship

$$G=(U, V, E)$$

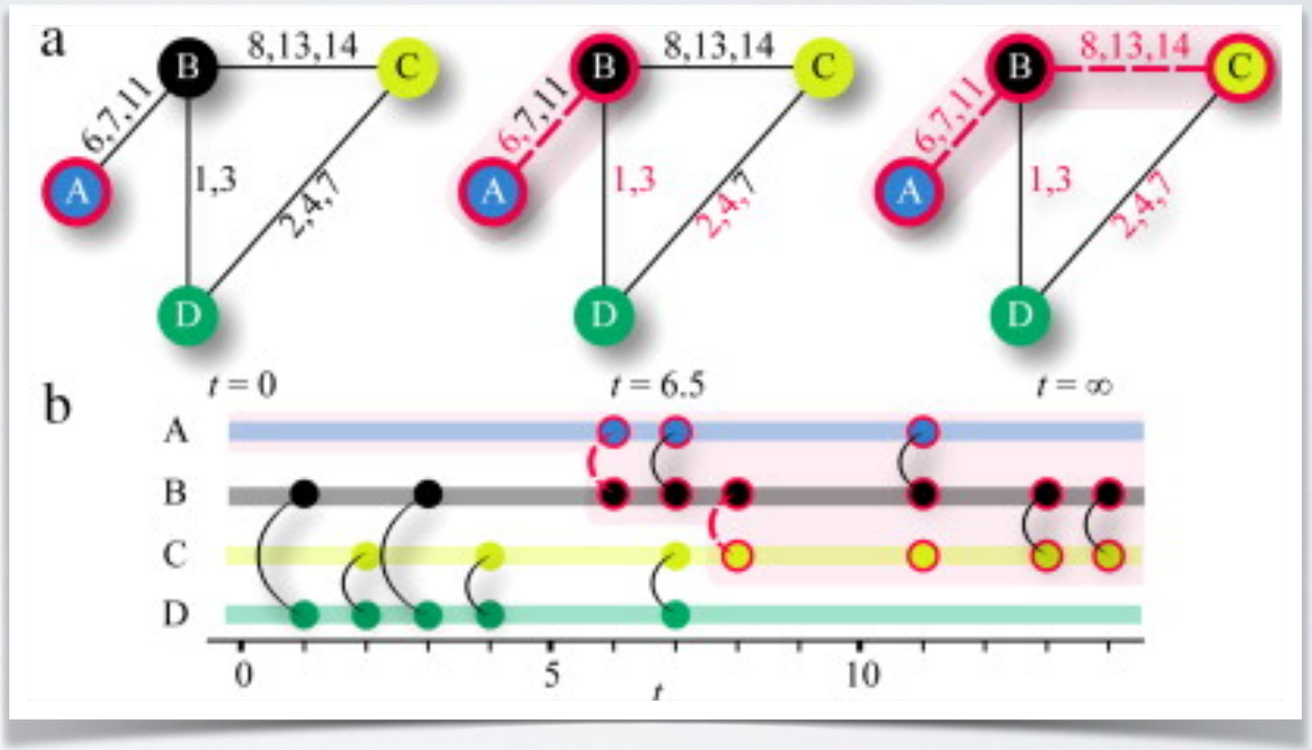
$$U \cap V = \varnothing$$

$$\forall (u,v) \in E,\ u \in U \text{ and } v \in V$$

Réseaux dynamiques

$G=(V, E_t), (u,v,t,d) \in E_t$

t - instant de l'interaction
d - durée de l'intéraction (u,v,t)



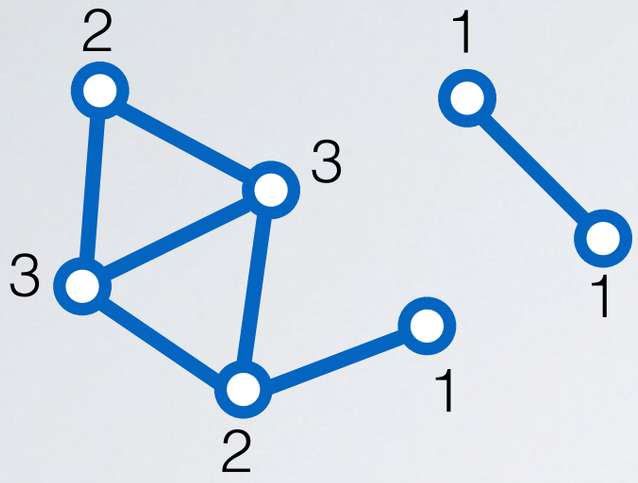
Mobile communication network
Nodes - individuals
Links - calls and SMS

Description des nœuds/liens

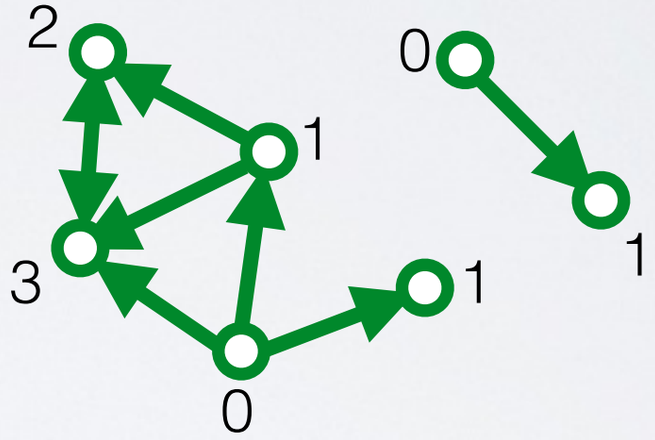
N_u	**Voisins** de u, nœuds qui partagent un lien avec u.
k_u	**Degré** de u, nombre de voisins $\|N_u\|$.
N_u^{out}	**Successeurs** de u, nœueds tels que $(u, v) \in E$ dans un graph dirigé
N_u^{in}	**Prédécesseurs** de u, nœuds tels que $(v, u) \in E$ dans un graphe dirigé
k_u^{out}	**Degré sortant** 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, $s_u = \sum_v w_{uv}$.

Degré des nœuds

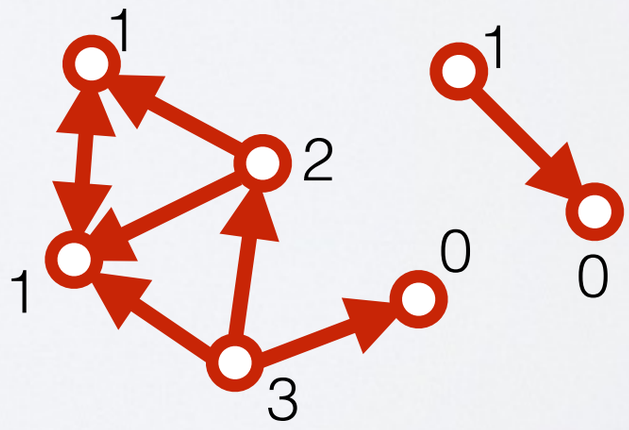
- Graphe non dirigé



- Graphe dirigé

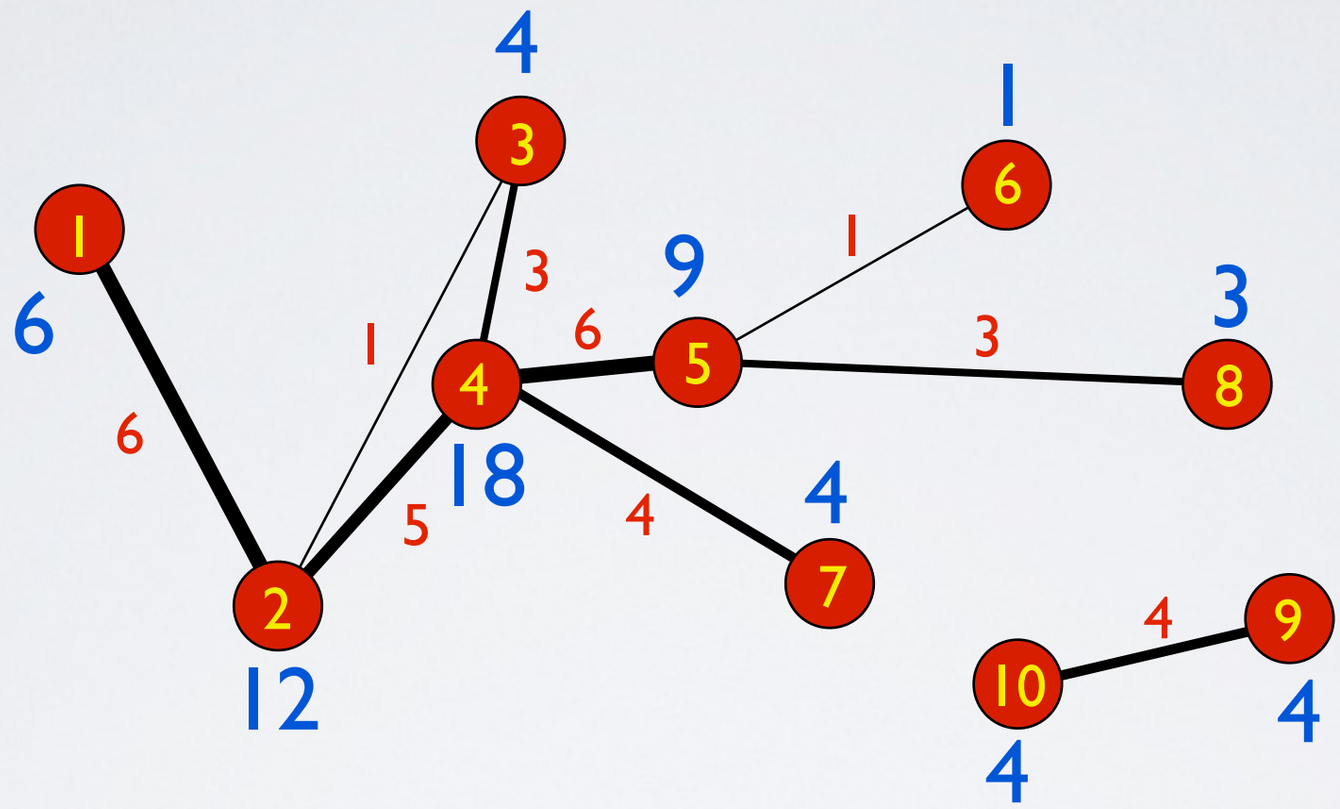


Degré entrant



Degré sortant

Degré pondéré : Force



DÉCRIRE DES GRAPHES

DÉCRIRE DES GRAPHES

- Si on nous donne un graphe, comment le décrire ?
- Comment comparer des graphes ?
- Que peut-on dire sur un graphe que l'on observe ?

TAILLE

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

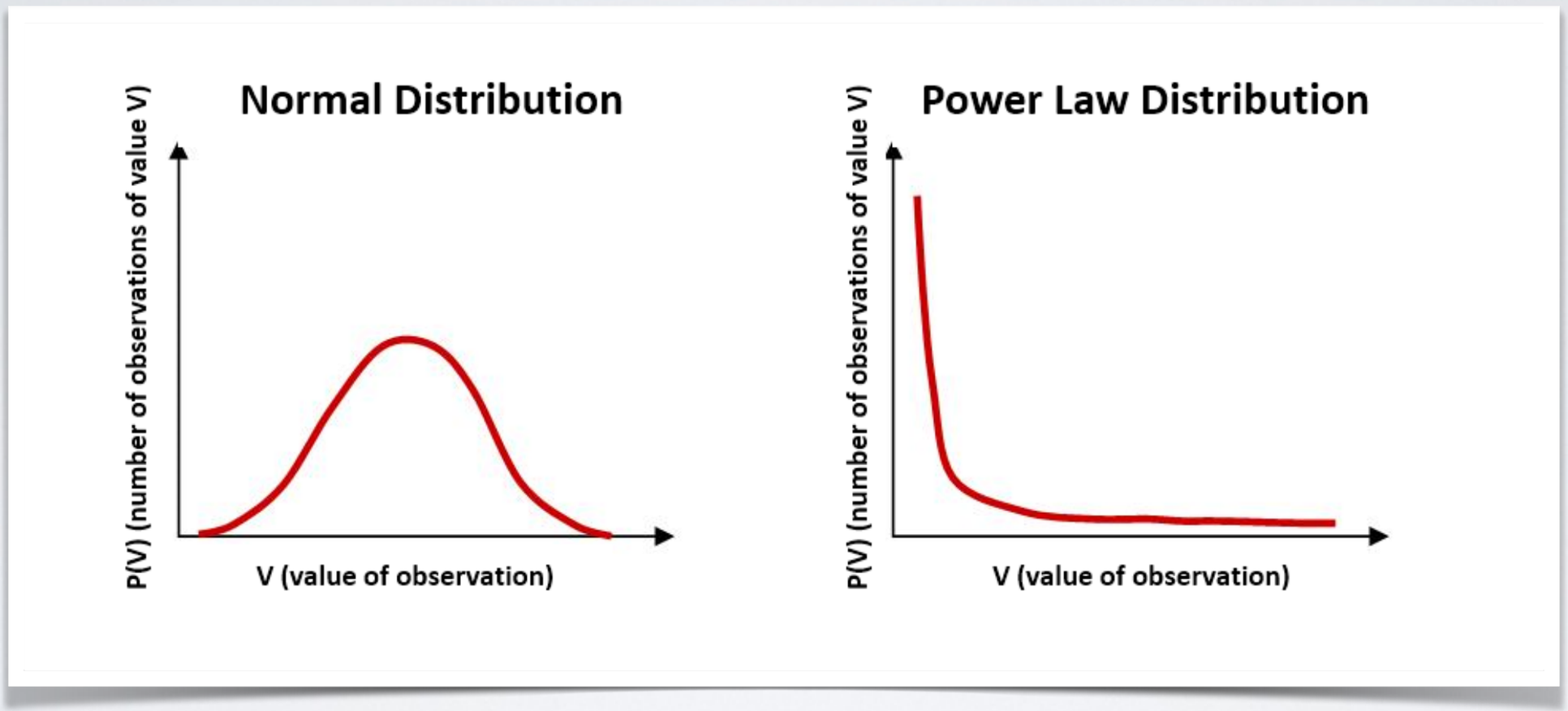
Description de réseaux - Nœuds/Liens

$\langle k \rangle$	**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. $$d = L/L_{\max}$$

	#nodes	#edges	Densité	Deg. Moyen
Wikipedia	2M	30M	$1.5x10^{-5}$	30
Twitter 2015	288M	60B	$1.4x10^{-6}$	416
Facebook	1.4B	400B	$4x10^{-9}$	570
Brain c.	280	6393	0,16	46
Roads Calif.	2M	2.7M	$6x10^{-7}$	2,7
Airport	3k	31k	0,007	21

Attention: Densité difficile à comparer entre des graphes de taille différente

DISTRIBUTION DE DEGRÉ



DISTRIBUTION DE DEGRÉ

- Dans un graphe complètement aléatoire (Erdos-Renyi), la distribution de degrés suit une loi normale (en fait, loi de Poisson) centrée sur le degré moyen.

- Dans les graphes réels, en général, pas le cas :
  - Grande majorité de nœuds de faible degré
  - Une faible quantité de nœuds de degré exceptionnel(Hubs)

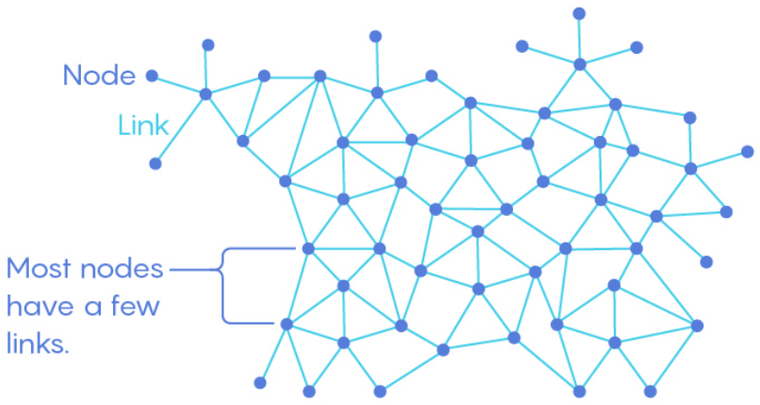
- Loi de puissance (**power law**)

To Be or Not to Be Scale-Free

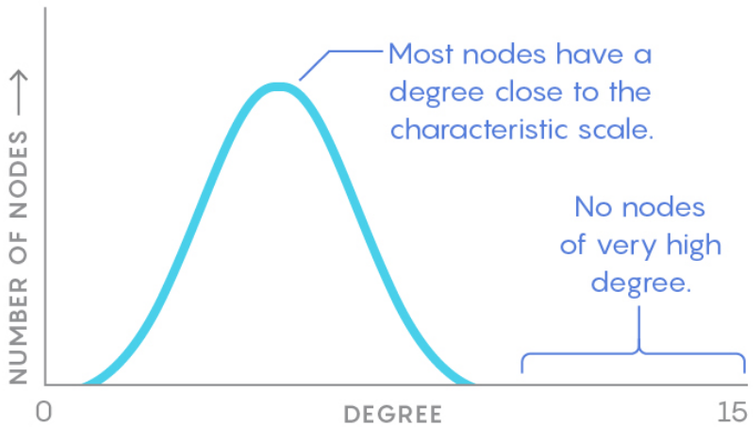
Scientists study complex networks by looking at the distribution of the number of links (or "degree") of each node.
Some experts see so-called scale-free networks everywhere. But a new study suggests greater diversity in real-world networks.

Random Network

Randomly connected networks have nodes with similar degrees. There are no (or virtually no) "hubs" — nodes with many times the average number of links.

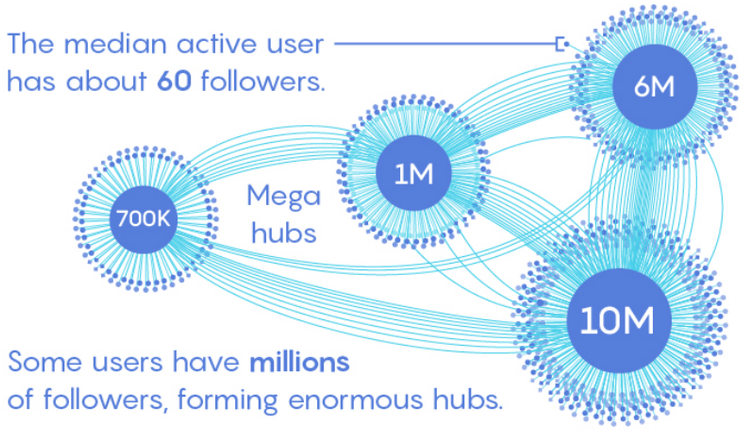


The distribution of degrees is shaped roughly like a bell curve that peaks at the network's "characteristic scale."

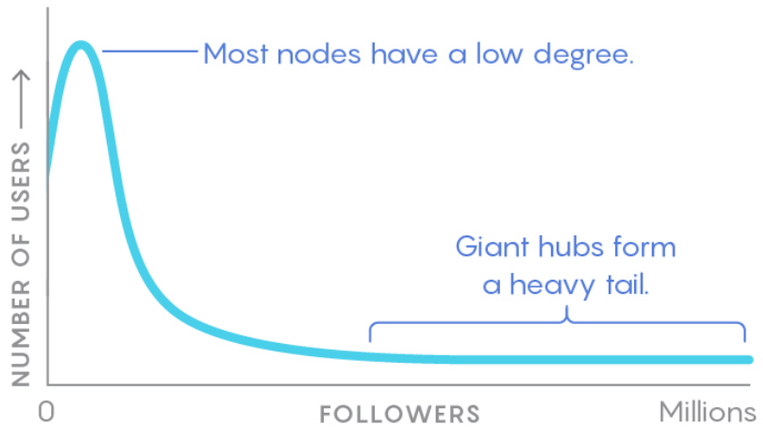


Twitter's Scale-Free Network

Most real-world networks of interest are not random. Some nonrandom networks have massive hubs with vastly higher degrees than other nodes.

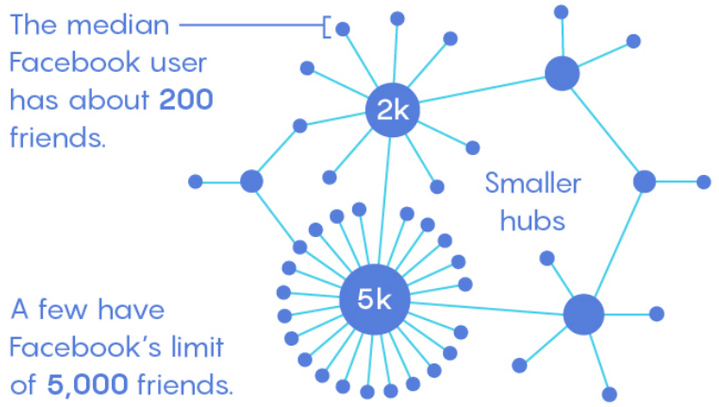


The degrees roughly follow a power law distribution that has a "heavy tail." The distribution has no characteristic scale, making it scale-free.

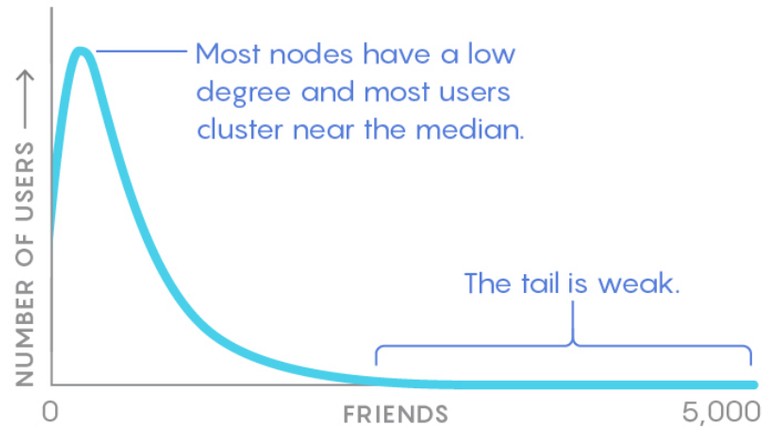


Facebook's In-Between Network

Researchers have found that most nonrandom networks are not strictly scale-free. Many have a weak heavy tail and a rough characteristic scale.



This network has fewer and smaller hubs than in a scale-free network. The distribution of nodes has a scale and does not follow a pure power law.



[Quanta magazine 2018]

DISTRIBUTION DE DEGRÉ

- Implications:
  - Le degré moyen n'est pas représentatif
  - Réseau "sans échelle" => Pas d'échelle caractéristique

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

COEFFICIENT DE CLUSTERING

- **Coefficient de clustering** ou **fermeture transitive**

- Les triangles sont considérés important dans un graphe
  - Réseau social: *les amis de mes amis sont mes amis*
  - Le nombre de triangle est très différent entre les réseaux aléatoires et les réseaux réels (en général)

CLUSTERING COEFFICIENT

Triangles

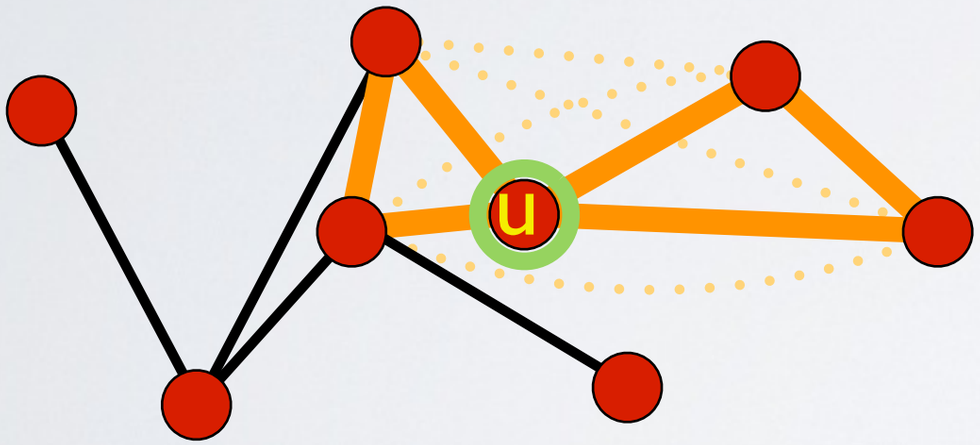
δ_u - **Triades de u:** nombre de triangles contenant le node u
Δ - **Nombre de triangles dans le graphe** $\Delta = \frac{1}{3}\sum_{u \in V} \delta_u$.

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. $\Delta^{\max} = \frac{1}{3}\sum_{u \in V} \delta^{\max}(u)$

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, $\frac{\delta_u}{\delta_u^{\max}}$



Liens: 2

Max liens: 4*3/2=6

C_u=2/6=1/3

Triangles=2

Triangles Possible = $\binom{4}{2}$=6

C_u=2/6=1/3

$\langle C \rangle$ - **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, $C^g = \frac{\Delta}{\Delta^{\max}}$

COEFFICIENT DE CLUSTERING

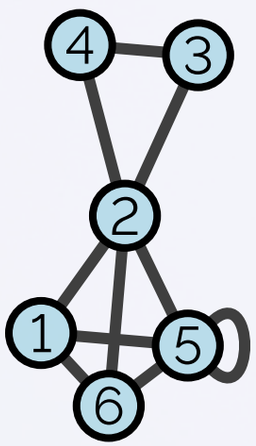
- CC Global:
  - Dans un réseau aléatoire, CC global = densité
    - =>Très petit pour des grands graphes
  - Facebook ego-networks: 0.6
  - Twitter lists: 0.56
  - California Road networks: 0.04

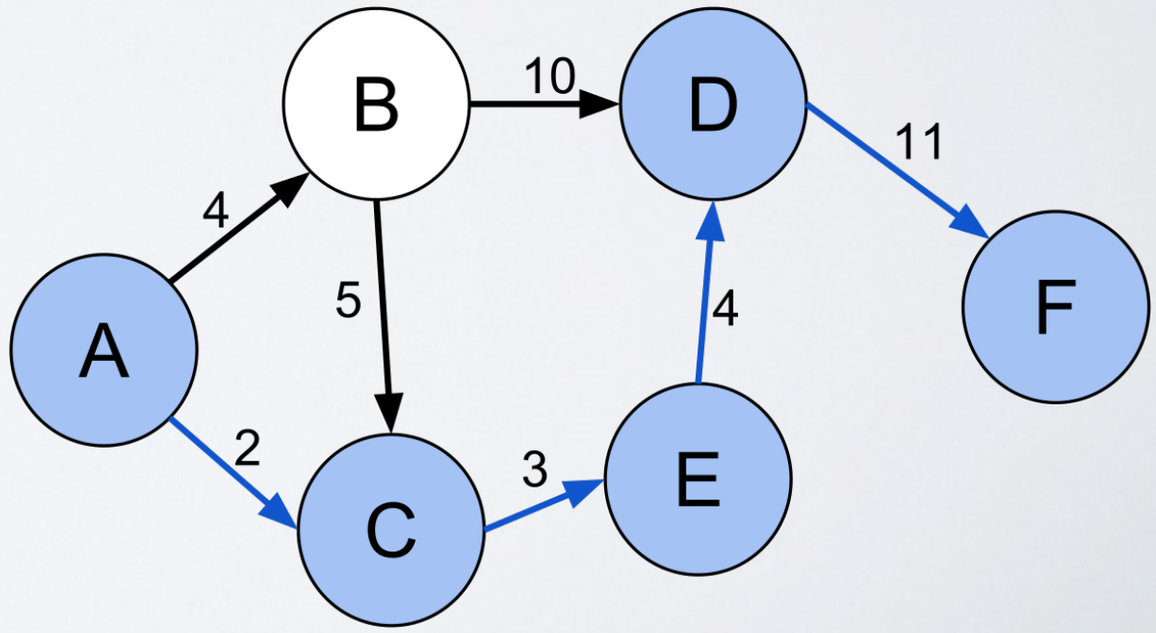
CHEMINS/MARCHES

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.

Graph





Description de réseaux - Chemins

$\ell_{\max}$
$\langle \ell \rangle$

Diametre: *distance* maximale entre 2 nœuds du réseau.
Distance moyenne, i.e., moyenne des distances entre toutes les paires de nœuds:

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

DISTANCE MOYENNE

- Les “6 degrés de séparation” (Expérience de Milgram)
  - Voir slides suivant

- Indique si le graphe est en “sac de nœuds”, ou s’il est étiré (“filaments”, moustaches…)

EXPÉRIENCE DE MILGRAM

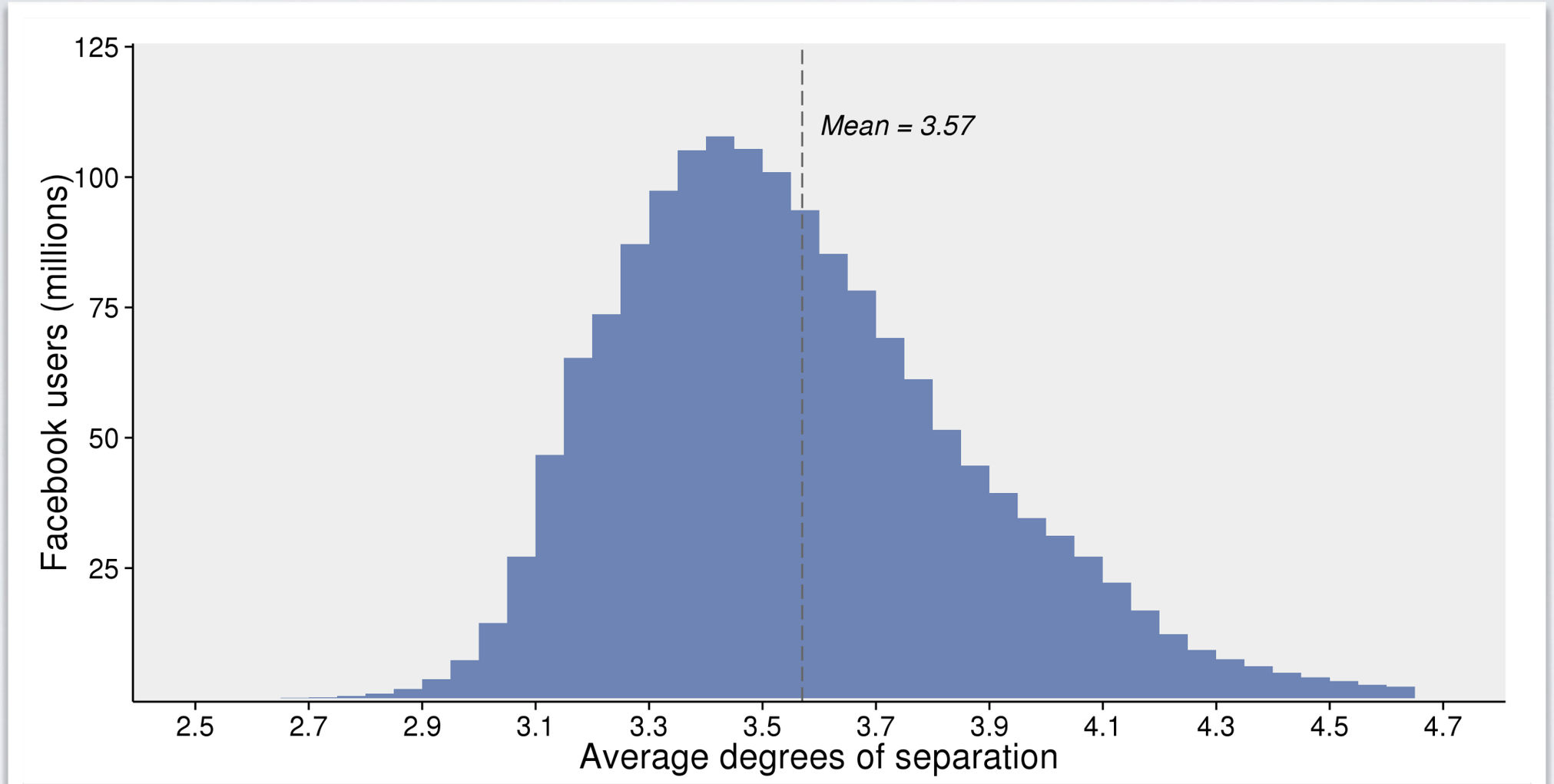
- Expériences du petit monde (60's)
  - Donner une enveloppe à des individus pris au hasard
  - Demander de l'envoyer à une personne qu'ils ne connaissent pas
    - A partir d'information (genre, age, métier)
  - Ils font transiter par des connaissances

- Résultats: en moyenne, 6 "sauts" avant d'arriver

EXPÉRIENCE DE MILGRAM

- Plusieurs critiques possibles
  - Certains courriers ne sont jamais arrivés
  - Nombre assez faible de participants
  - …
- Plus récemment, possibilité de tester sur de grands réseaux en ligne:
  - MSN messenger
  - Facebook
  - Etc.
  - …

EXPÉRIENCE DE MILGRAM



Facebook

SMALL WORLD

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.

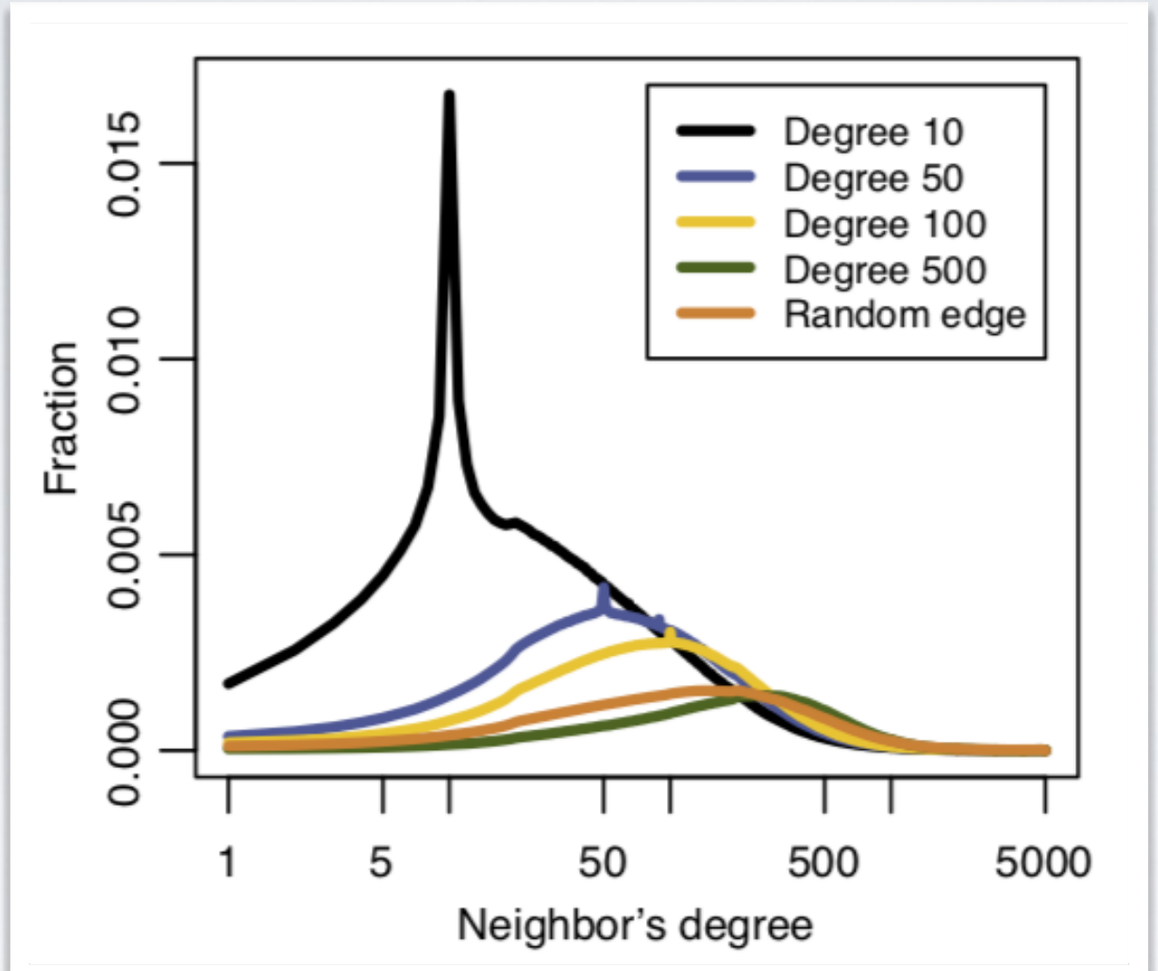
EXAMPLE D'ANALYSE DE GRAPHES

- Source: [The Anatomy of the Facebook Social Graph, Ugander et al. 2011]

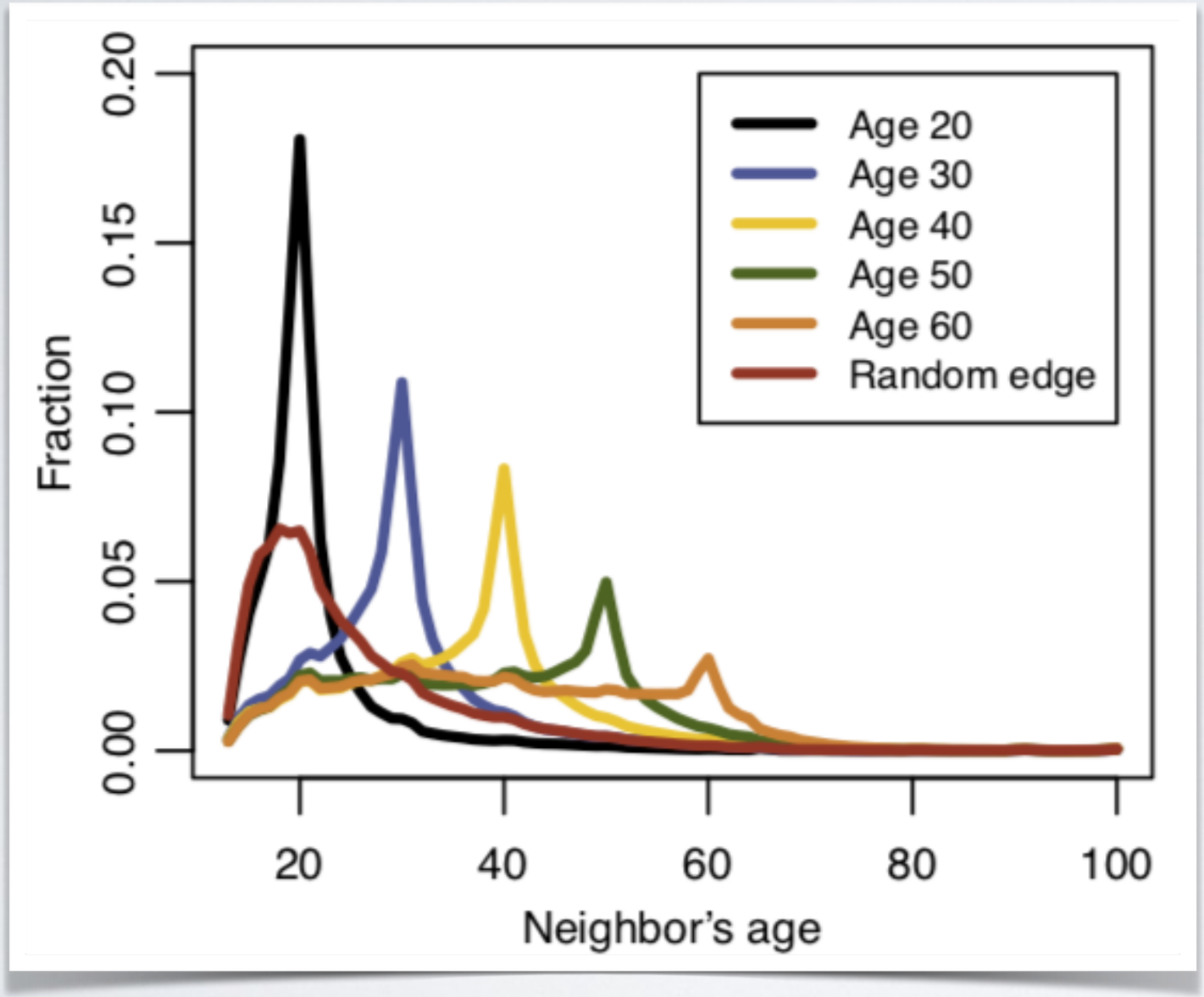
- Le réseau "d'amis" Facebook 2011

EXAMPLE D'ANALYSE DE GRAPHES

- 721M d'utilisateurs (nœuds) (actifs au cours des 28 derniers jours)
- 68 Milliards de liens
- Degré moyen: 190
- Degré médian: 99
- Composante connexe principale : 99.91%



Beaucoup de mes amis ont le même Nombre d'amis que moi…



Homophily en fonction de l'âge

Mesures de centralité

NŒUDS

- L'importance des nœuds se mesure avec des **centralités**.

- Attention: pas forcément être "au centre" du graphe

- Usage:
  - La plupart ont une interprétation claire
  - Certaines peuvent être utilisées par exemple pour du *Machine learning* (Prédire la réussite ou l'échec d'un politicien en fonction de sa position dans le réseau…)

DEGRÉ

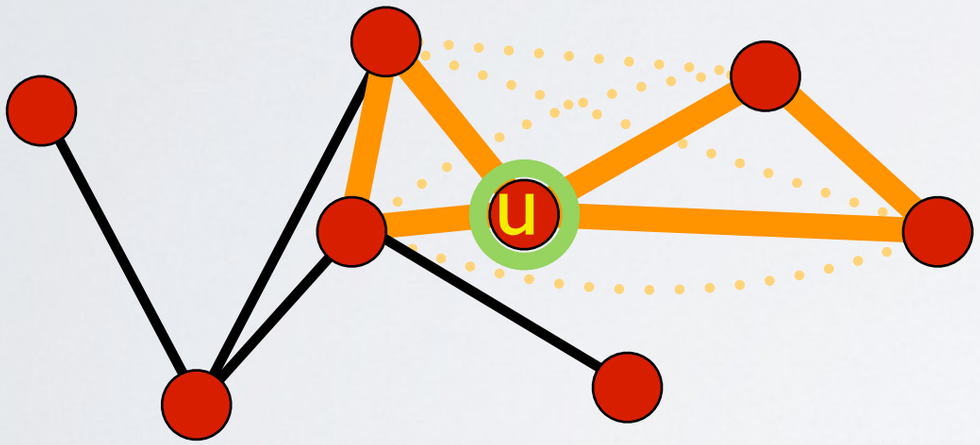
- **Degré**: Combien de voisins

- Souvent suffisant pour trouver les nœuds importants
  - Les personnages principaux d'une série sont ceux qui parlent le plus
  - Les aéroports les plus importants ont le plus de connexions
  - …

- Mais pas toujours
  - Les utilisateurs de Facebook avec le plus de contacts sont souvent des spammeurs
  - Les pages webs/Wikipedia avec le plus de liens sortants sont souvent de bêtes listes de pages. Les liens entrants sont facile à truquer via d'autres sites.
  - …

CLUSTERING COEFFICIENT

C_u - **Node clustering coefficient:** density of the subgraph induced by the neighborhood of u, $C_u = d(H(N_u)$. Also interpreted as the fraction of all possible triangles in N_u that exist, $\frac{\delta_u}{\delta_u^{\max}}$



Edges: 2

Max edges: 4*3/2=6

C_u=2/6=1/3

Triangles=2

Possible triangles= $\binom{4}{2}$ =6

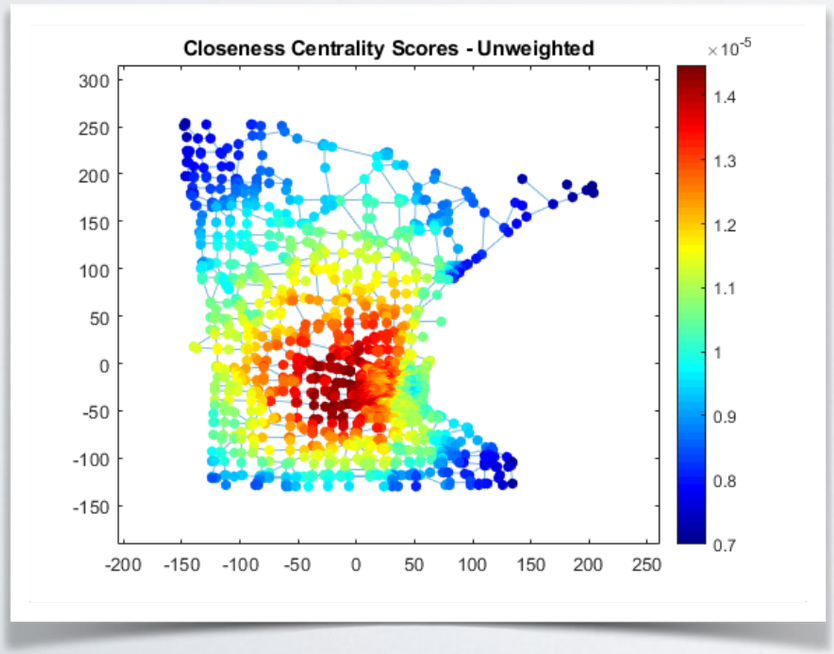
C_u=2/6=1/3

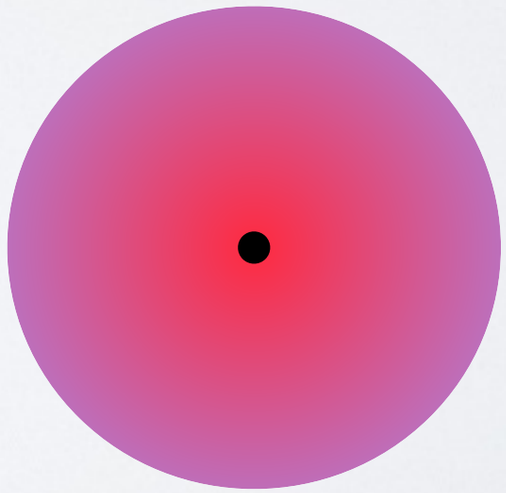
FARNESS, CLOSENESS

HARMONIC CENTRALITY

FARNESS, CLOSENESS

- Est-ce que le nœud est proche des autres, en moyenne.
- Parallèle avec le barycentre d'une figure:
  - Le barycentre d'une figure (centre d'une cercle) est le point le plus proche en moyenne de tous les autres points de la figure.



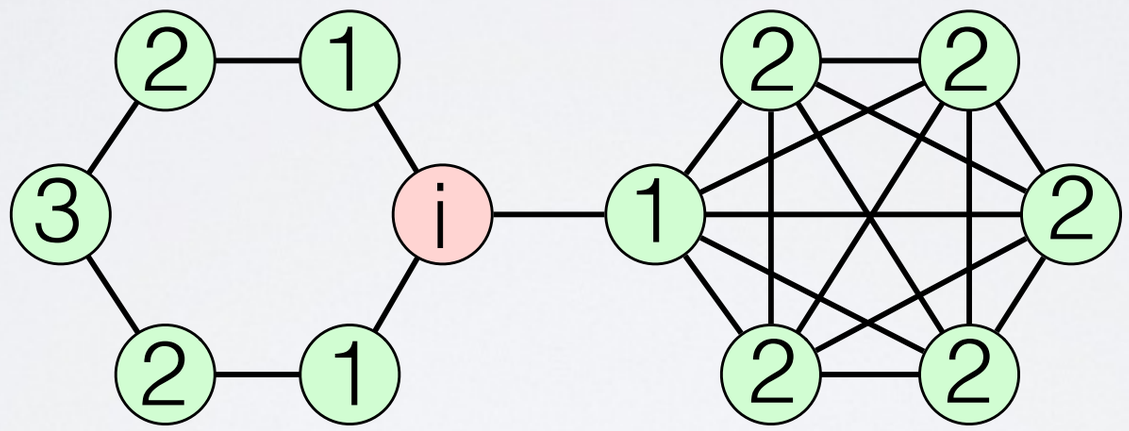


FARNESS, CLOSENESS

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 CENTRALITY



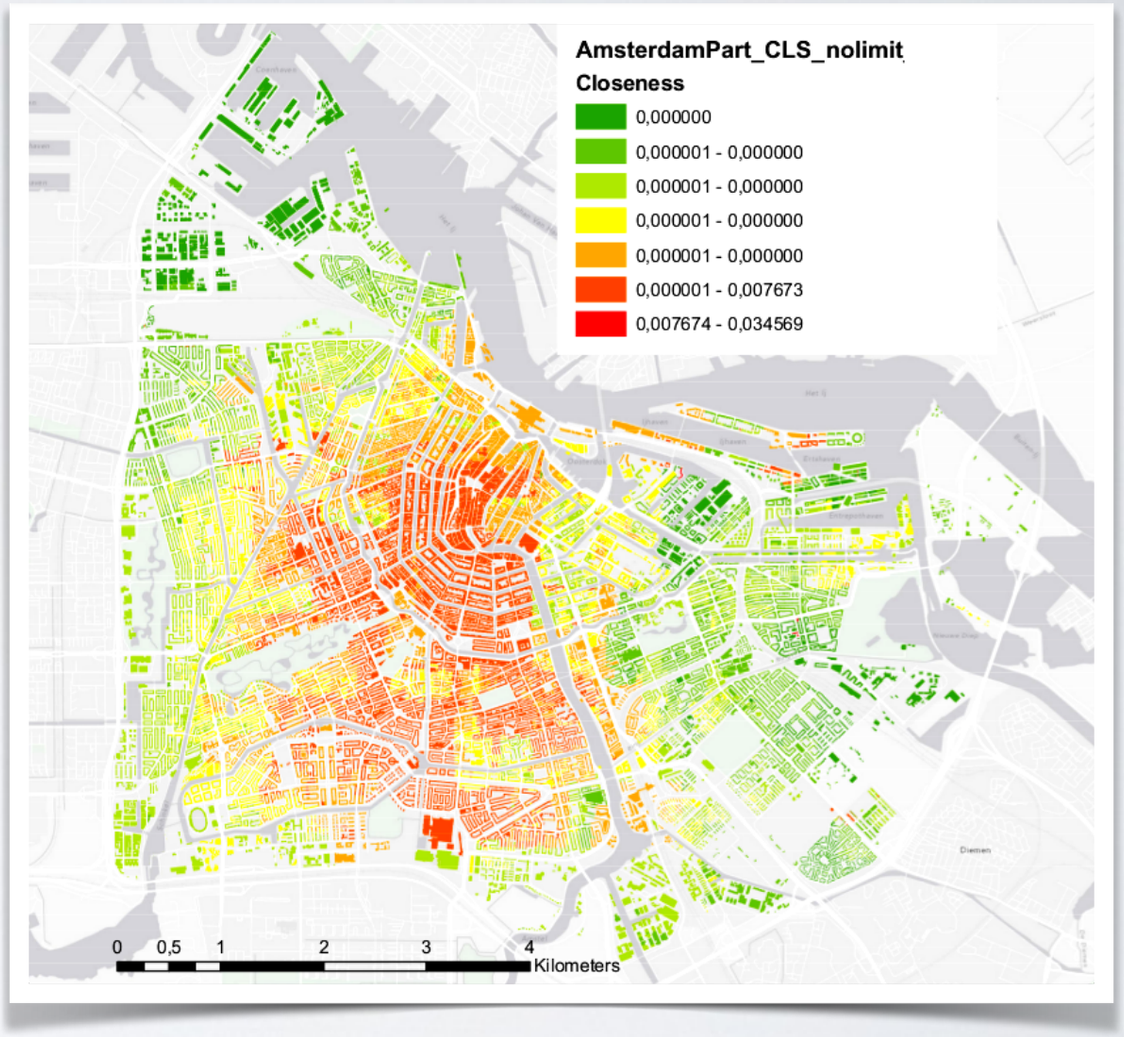
$$C_{cl}(i) = \frac{12 - 1}{(3 \times 1 + 7 \times 2 + 1 \times 3)} = \frac{11}{20} = 0.55$$

CLOSENESS CENTRALITY

Closeness: Inverse de la farness

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

I =tous les nœuds sont à distance I



BETWEENNESS CENTRALITY

Centralité d'intermédiarité

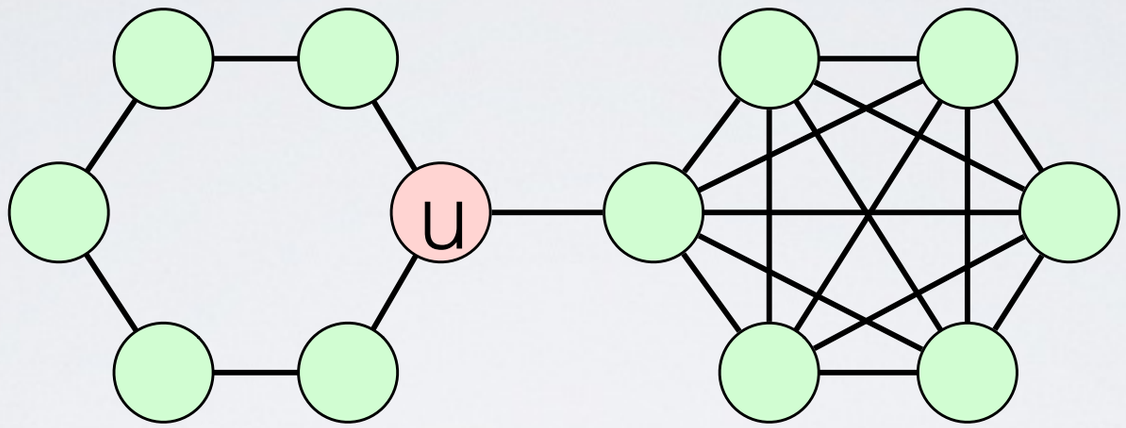
- Mesure à quel point le nœud joue le rôle d'un pont
- Betweenness de *u:* fraction de tous les plus courts chemins entre tous les nœuds qui passent par *u*

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

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: $C_B^{\text{norm}}(v) = \frac{C_B(v)}{(N-1)(N-2)}$.

Betweenness Centrality

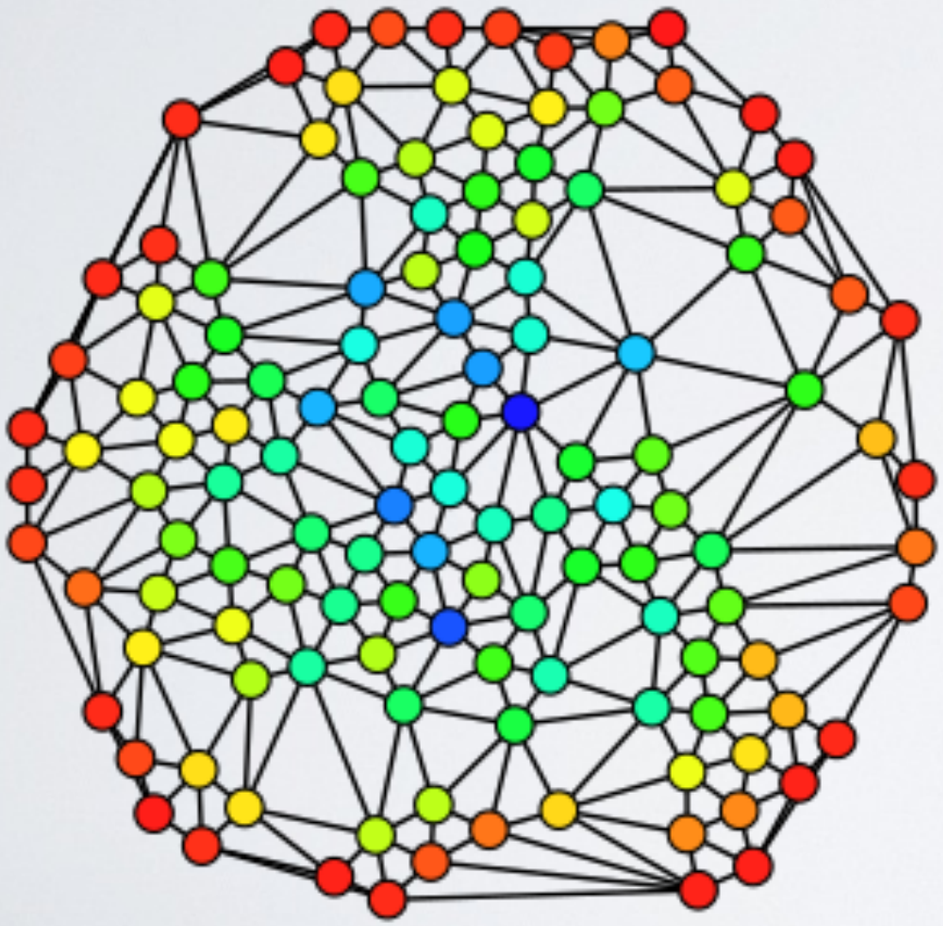
$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



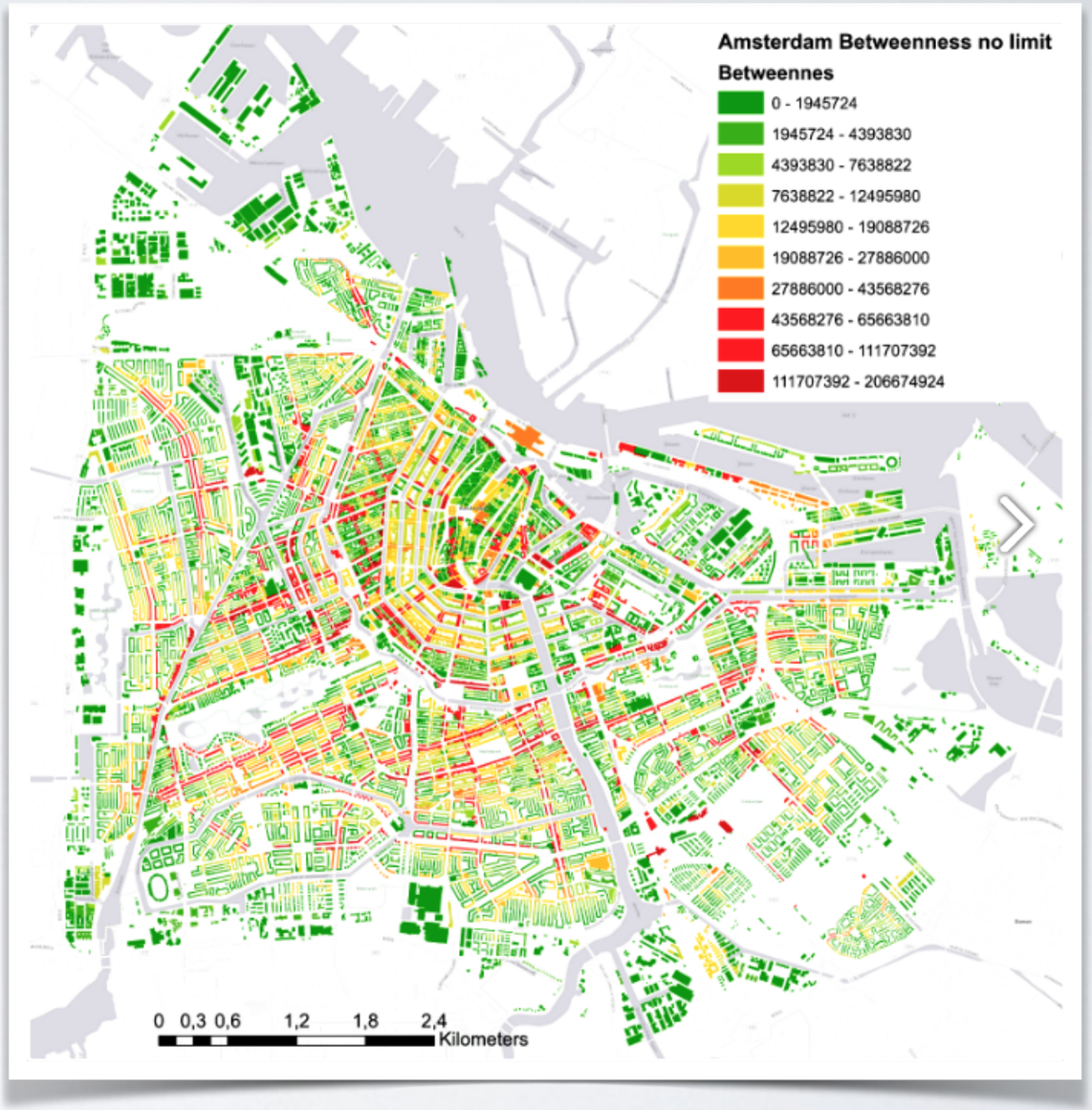
$$C_B(u) = 2\frac{5 * 6 + 1 + \frac{1}{2} + \frac{1}{2}}{11 * 10} = \frac{64}{110}$$

Peut être calculé de manière exacte (très couteux) ou approximative

BETWEENNESS CENTRALITY



Bleu valeur élevées



Rouge élevées

EDGE - BETWEENNESS

Intermédiarité des liens

Même définition que pour les nœuds.

Lien de plus forte betweenness dans le réseau ferroviaire Européen ?

DÉFINITIONS RÉCURSIVES

DÉFINITIONS RÉCURSIVES

- Importance récursive:
  - **Un nœud est important** s'il est connecté à (pointé par) ***des nœuds importants***

- Plusieurs centrales sont basées sur ce principe:
  - Centralité Eigenvectors ("valeurs propres")
  - PageRank
  - Hub et Autorités

DÉFINITIONS RÉCURSIVES

- Définissons l'objectif:
  - Chaque nœud à un score (centrality),
  - Si chaque nœud envoie son score à ses voisins, la somme (normalisée) des scores qu'il a reçu est égale à sa valeur de centralité

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

- Avec λ une constante de normalisation

DÉFINITIONS RÉCURSIVES

- Plus qu'à trouver une solution mathématique à ce problème
- Peut être résolue par la *power method (méthode des puissances itérées)*
  - 1) Tous les scores sont initialisés avec des valeurs aléatoires entre 0 et 1
  - 2)On applique la règle définie auparavant jusqu'à atteindre un point stable (les valeurs ne changent plus, donc objectif atteint)
  - Garantie de convergence vers un point stable
- Pourquoi est-ce que ça marche?
  - Théorème de Perron-Frobenius
  - => Vrai pour des graphes non-dirigés avec une seule composante connexe.

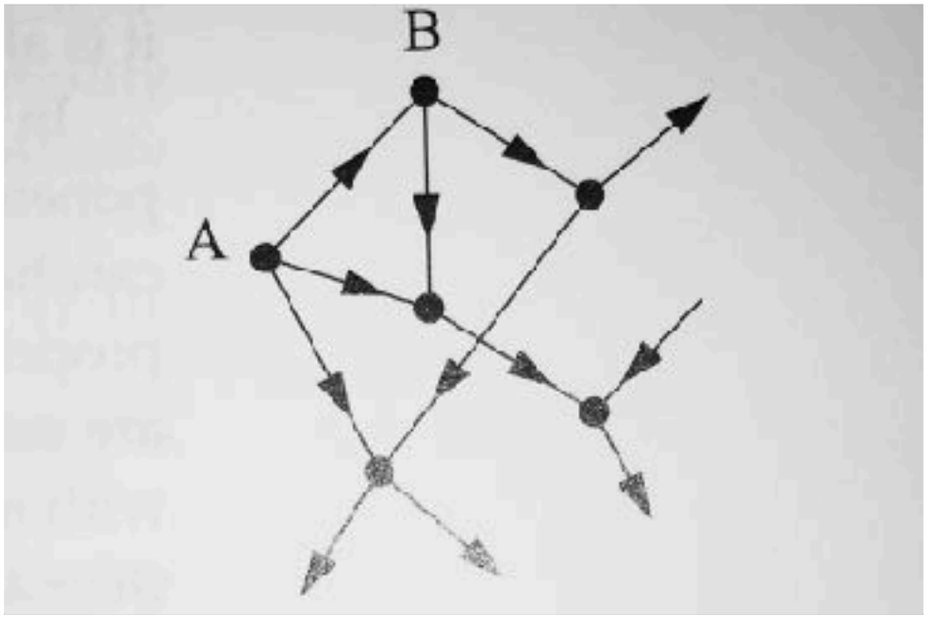
CENTRALITÉ EIGENVECTORS

- Ce que l'on viens de décrire : Centralité Eigenvector
- Un couple vecteur propre (x) et valeur propre (λ) est défini par la relation: $Ax = \lambda x$
  - x est un vecteur colonne de taille *n*, interprété comme les scores de nœuds
- Ce que dit le théorème Perron-Frobenius est que la méthodes des puissances va converger vers *le premier vecteur propre*, i.e., le vecteur propre associé à la valeur propre la plus élevée.

Eigenvector Centrality

Des problèmes avec les graphes dirigés:

- 2 ensembles de vecteurs propres (Gauche et Droit)
- On utilise les vecteurs propres droit : les nœuds envoient leurs poids dans le sens de la flèche.



Mais problème avec les nœuds source (degré entrant=0)
-Nœud A n'a que des liens sortants = sa centralité après la première itération

est 0

-Nœud B a des liens entrants et sortants, mais son lien entrant viens de A = Centralité de 0 au second tour

-etc.

Solution: Calcul seulement dans la plus grande composante connexe forte

Note: Les réseaux acycliques (e.g., réseau de citation) n'ont pas de composante connexe forte

PageRank Centrality

- Centralité Eigenvector généralisée aux graphes dirigés

PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

PageRank Centrality

- Eigenvector centrality generalised for directed networks

PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

PageRank Centrality

(Side notes)

-"We chose our system name, Google, because it
is a common spelling of googol, or 10^{100} and fits well with our goal of building very large-scale search "

-"[…] at the same time, search engines have migrated from the academic domain to the commercial. **Up until now most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art and to be advertising oriented (see Appendix A). With Google, we have a strong goal to push more development and understanding into the academic realm."**

-"[...], we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers."

PageRank Centrality

(Side notes)

Sergey Brin received his B.S. degree in mathematics and computer science from the University of Maryland at College Park in 1993. Currently, he is a Ph.D. candidate in computer science at Stanford University where he received his M.S. in 1995. He is a recipient of a National Science Foundation Graduate Fellowship. His research interests include search engines, information extraction from unstructured sources, and data mining of large text collections and scientific data.

Lawrence Page was born in East Lansing, Michigan, and received a B.S.E. in Computer Engineering at the University of Michigan Ann Arbor in 1995. He is currently a Ph.D. candidate in Computer Science at Stanford University. Some of his research interests include the link structure of the web, human computer interaction, search engines, scalability of information access interfaces, and personal data mining.

PAGERANK

- 2 améliorations principales:
  - Problème des nœuds source
    - => Ajout d'un petit gain constant à tous les nœuds ("téléportation")
  - Les nœuds de centrality forte donnent une centralité forte à tous leurs voisins (Même s'ils en ont énormément, et que certain n'ont pas d'autres entrées)
    - => Ce que chaque nœud "vaut" est divisé entre ses liens sortants (Normalisation par le degré)

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

With by convention $\beta = 1$ and α a parameter (usually 0.85) controlling the relative importance of β

PageRank - Marche aléatoire

Compréhension intuitive : Interprétation en tant que marche aléatoire

Marche aléatoire : On démarre d'un nœud pris au hasard, puis on "marche" au hasard dans le réseau, en suivant les liens. (On choisit un lien sortant au hasard)

Probabilité de téléportation : le paramètre α correspond à la probabilité de faire ce processus normalement, et $1\text{-}\alpha$ de sauter aléatoirement à n'importe quel nœud à la place.

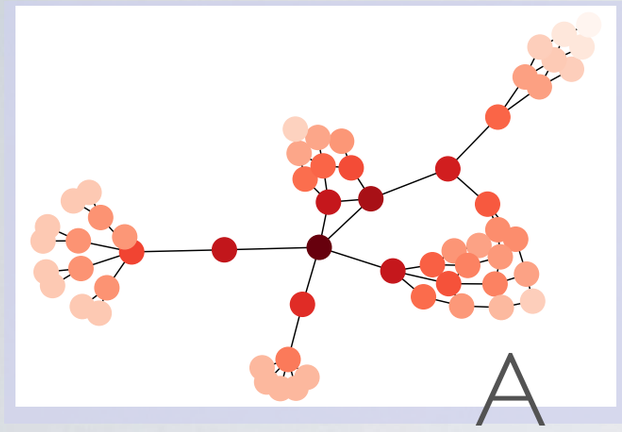
Pagerank score d'un nœud correspond à la probabilité que le marcheur aléatoire soit sur ce nœud après un nombre infini de déplacements.

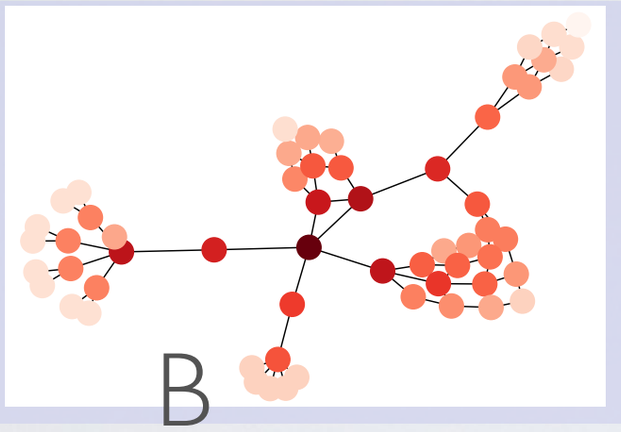
PAGERANK

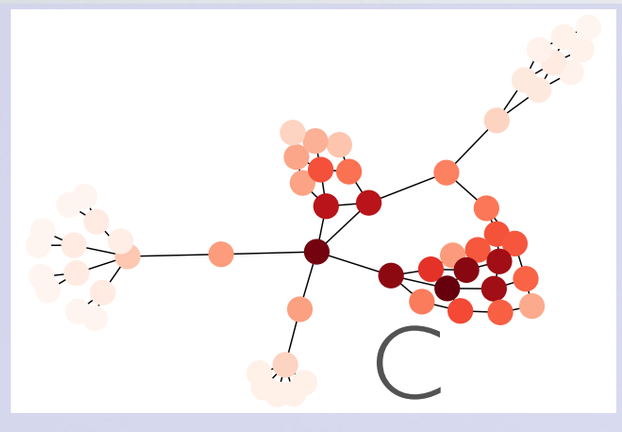
- Comment Google classe les résultats de nos recherches ?
- Calcul de pagerank (Power method)
- Filtre les pages contenant les mots recherchés
- Bien sûr aujourd'hui les méthodes sont plus complexes, mais non publiques:
  "Most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art" [Page, Brin, 1997]

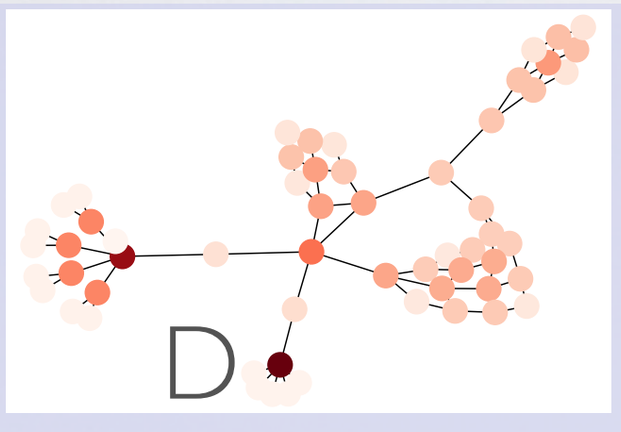
OTHERS

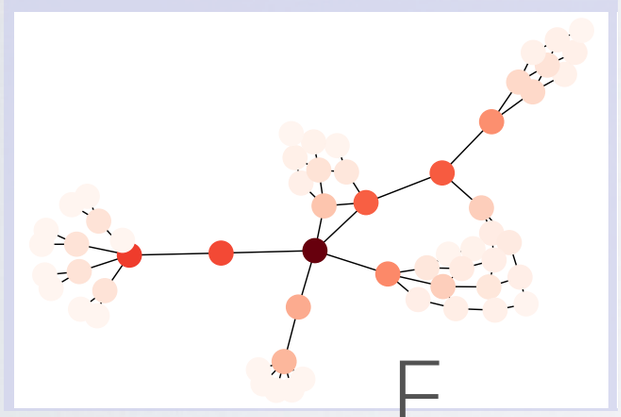
- Beaucoup d'autres centralités ont été proposées
- Le problème est souvent comment les interpréter

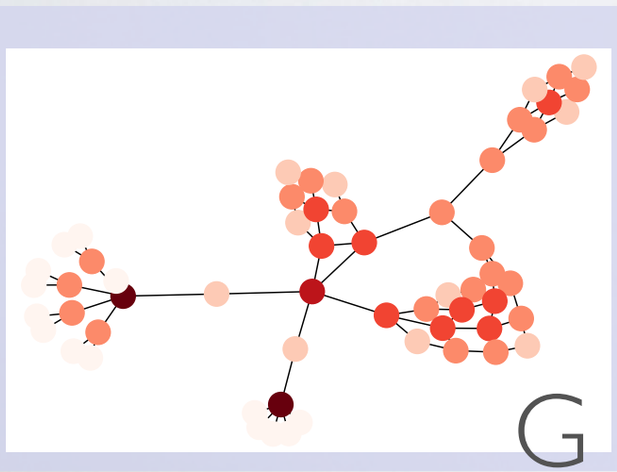


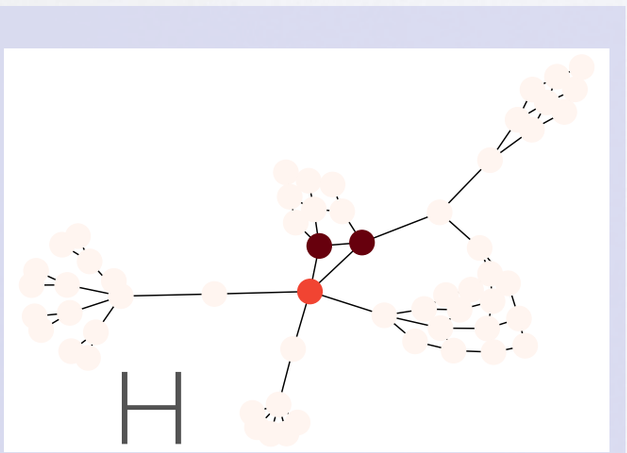






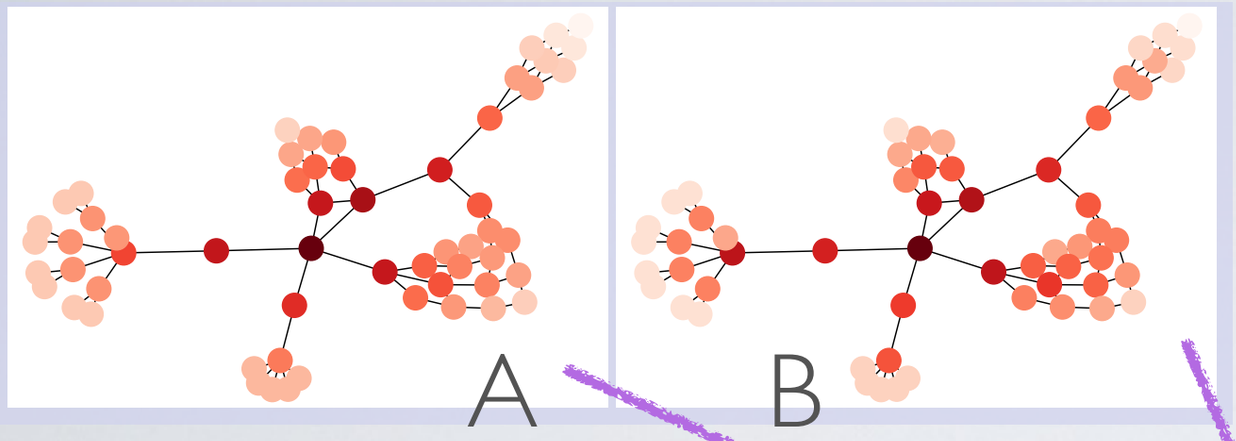


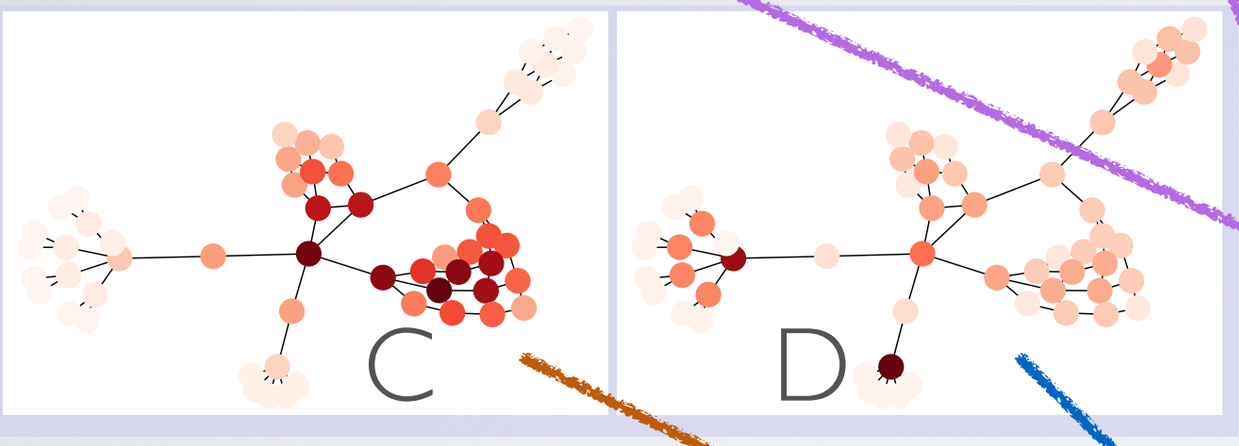


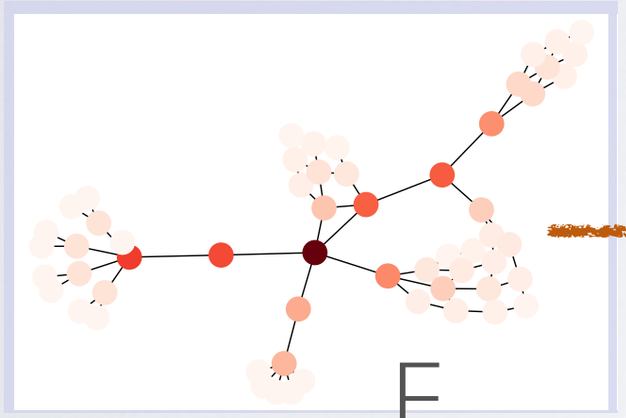


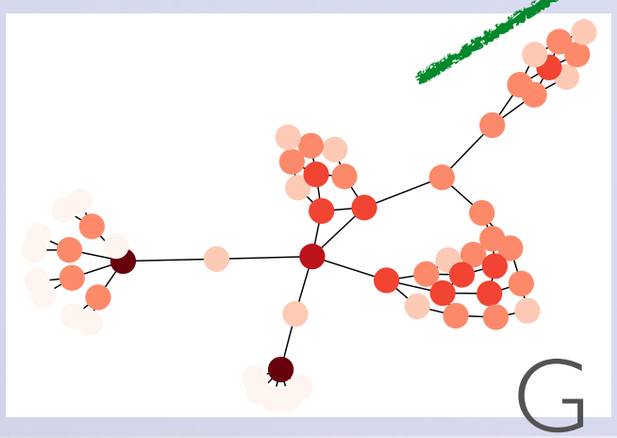
Qui est qui ?

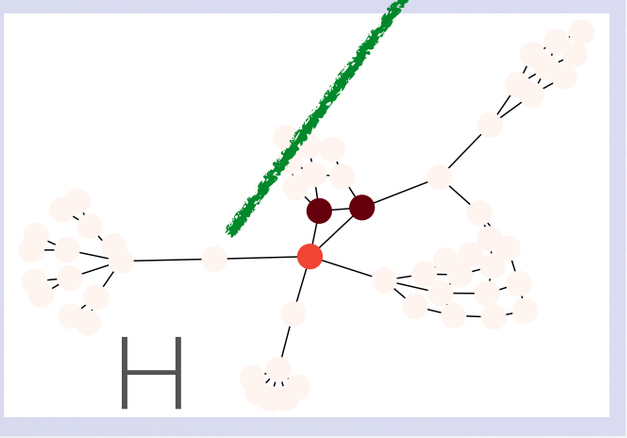
Degré
Clustering coefficient
Closeness
Harmonic Centrality
Betweenness
~~Katz~~
Eigenvector
PageRank











Qui est qui ?

Degré

Clustering coefficient

Closeness

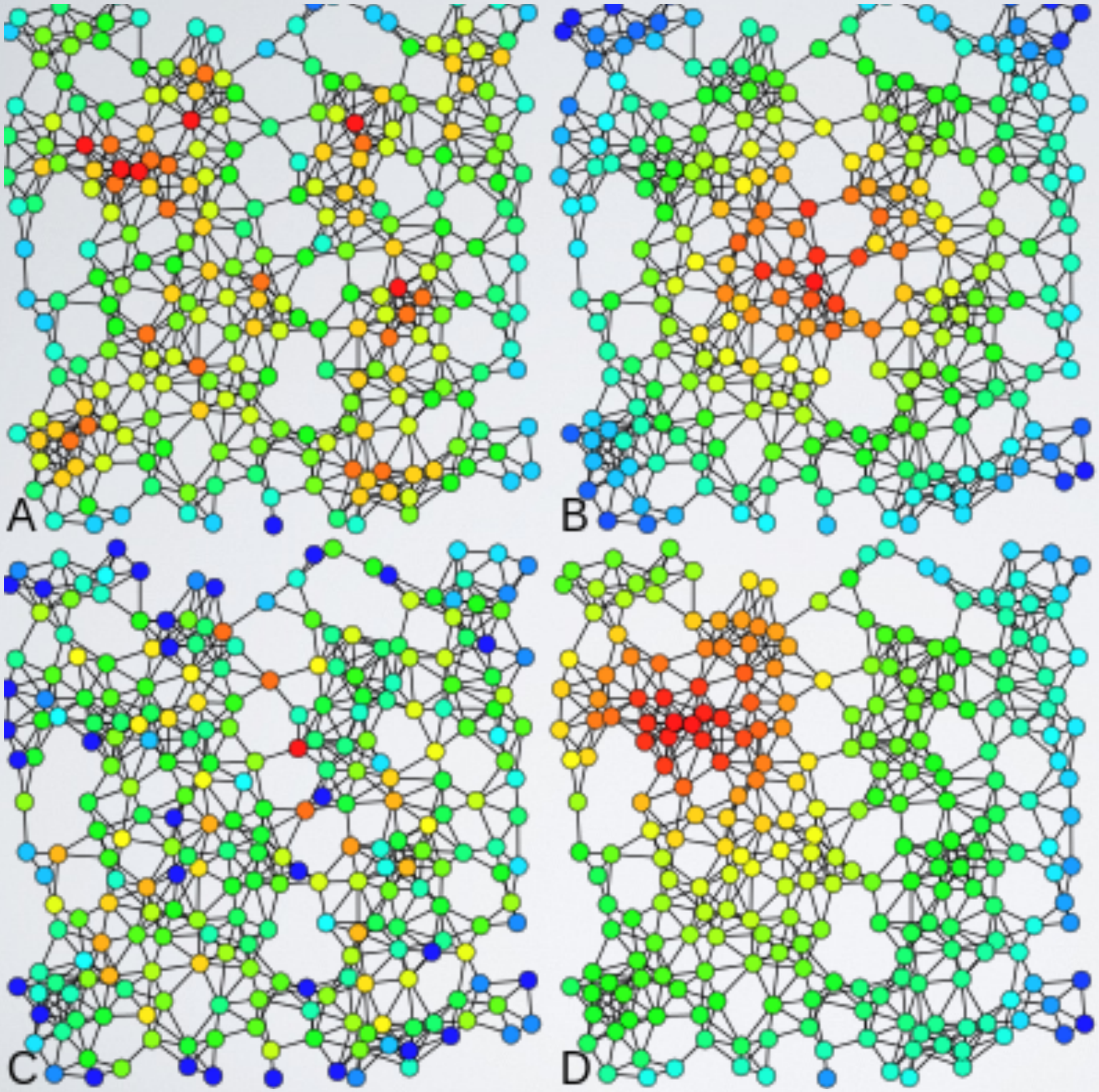
Harmonic Centrality

Betweenness

~~Katz~~

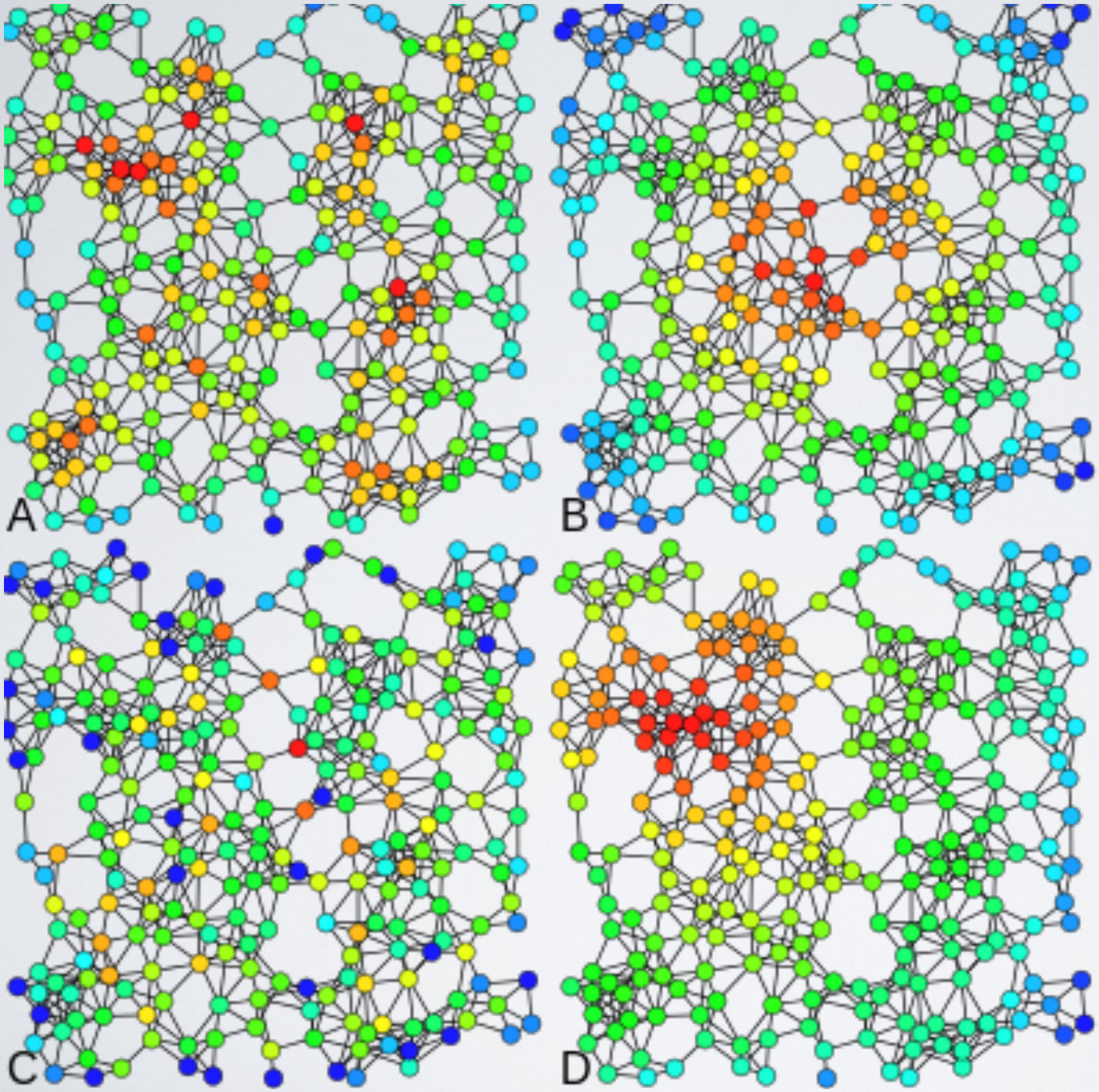
Eigenvector

PageRank



Autre essai :)

Degree
Betweenness
Closeness
Eigenvector



Autre essai :)

A: Degree
B:Closeness
C: Betweenness
D: Eigenvector