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

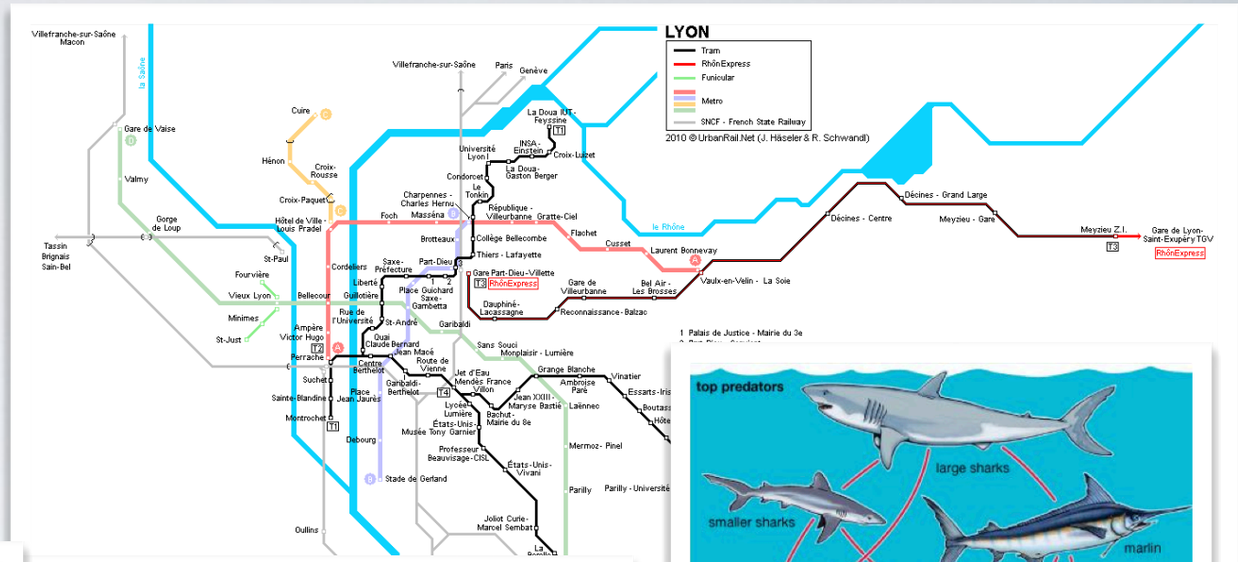
- Introduction à la science des réseaux: Décrire un graphe
- Introduction à Gephi: visualisation

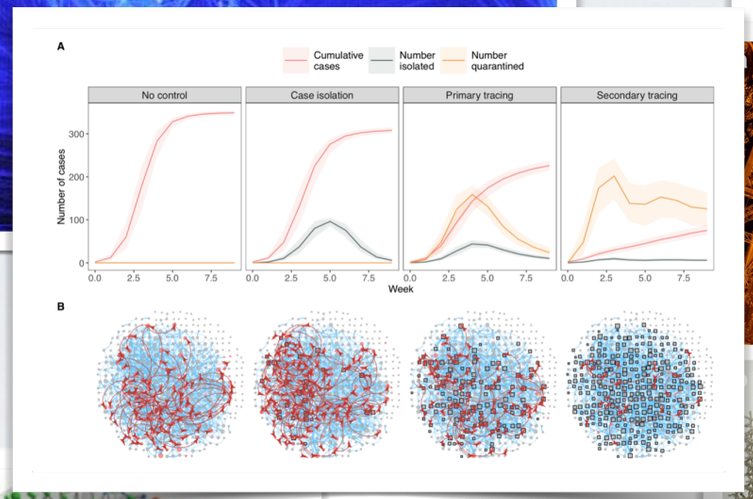
GEPHI

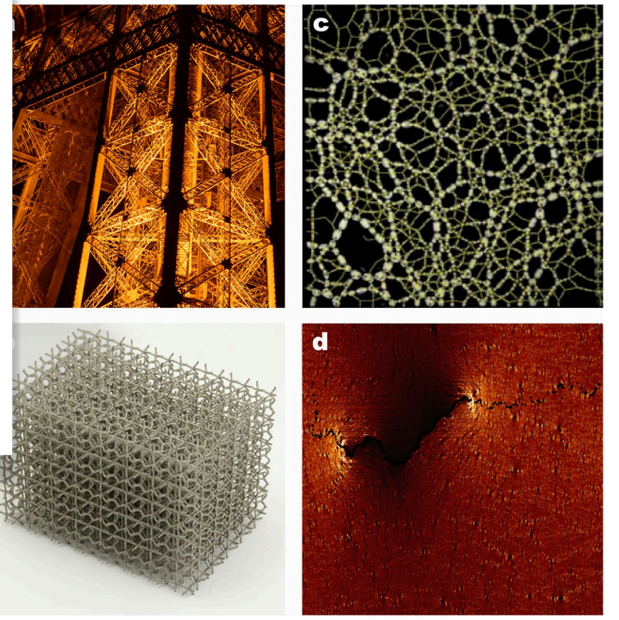
- Un logiciel pour visualiser des données sous forme de graphes, et pour des analyses réseaux simples

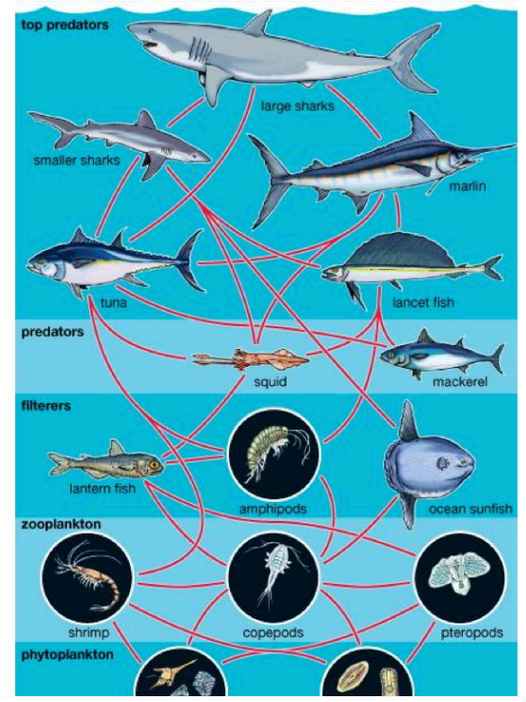
- Pour aller plus loins dans les analyses, programmation:
  - Python: Networkx, igraph, graph-tool, etc.

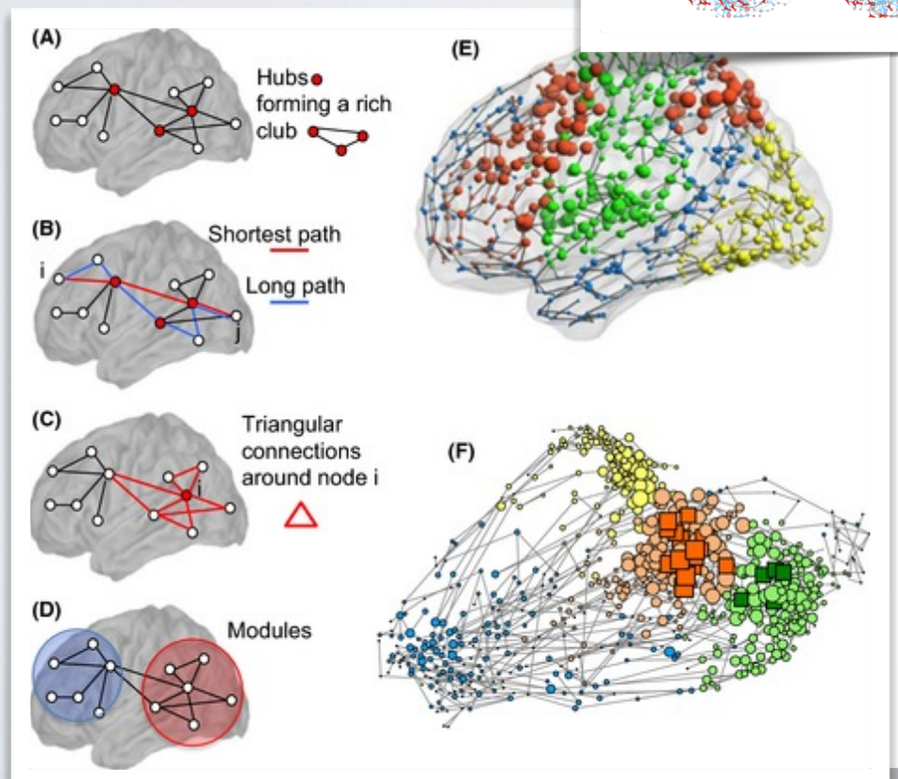
SCIENCE DES RÉSEAUX

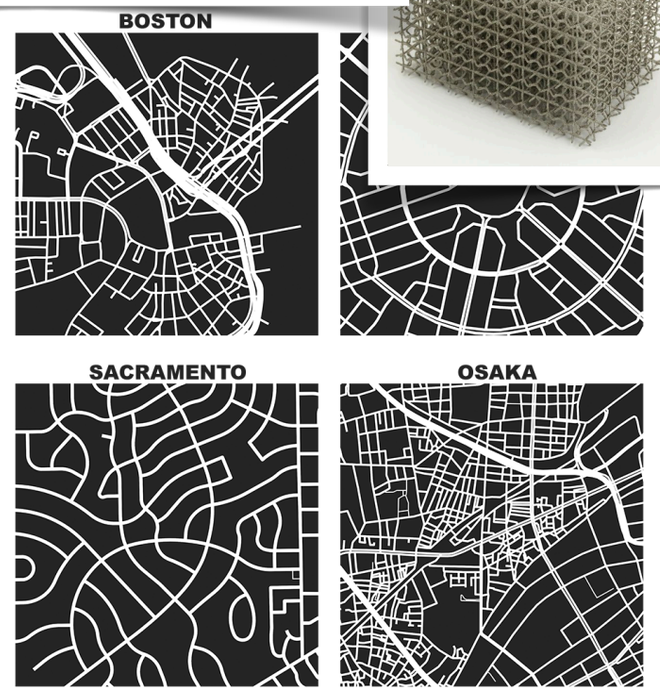


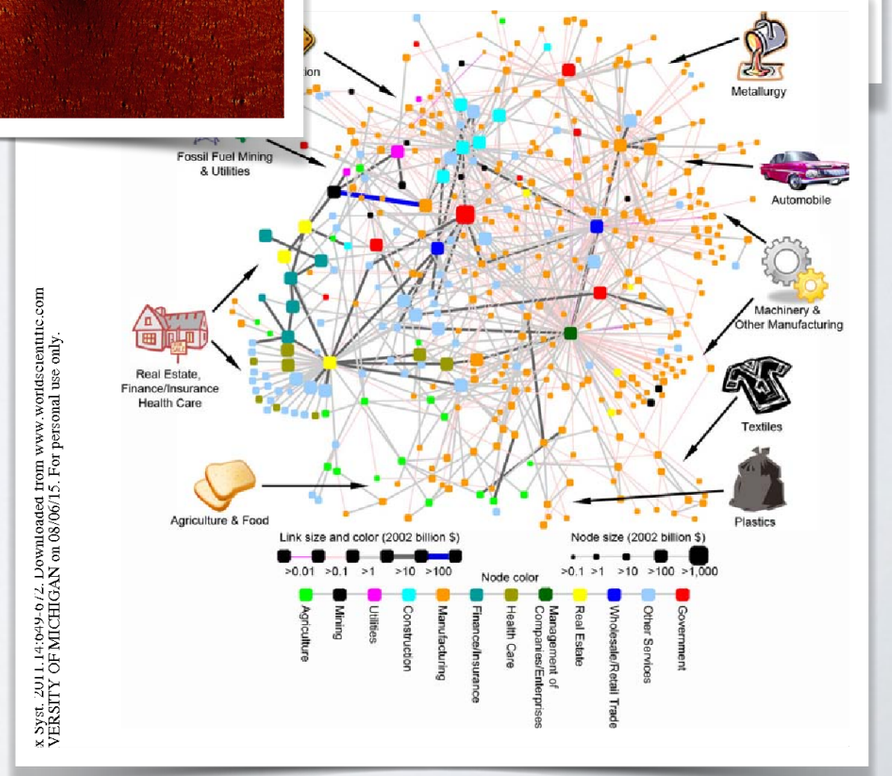


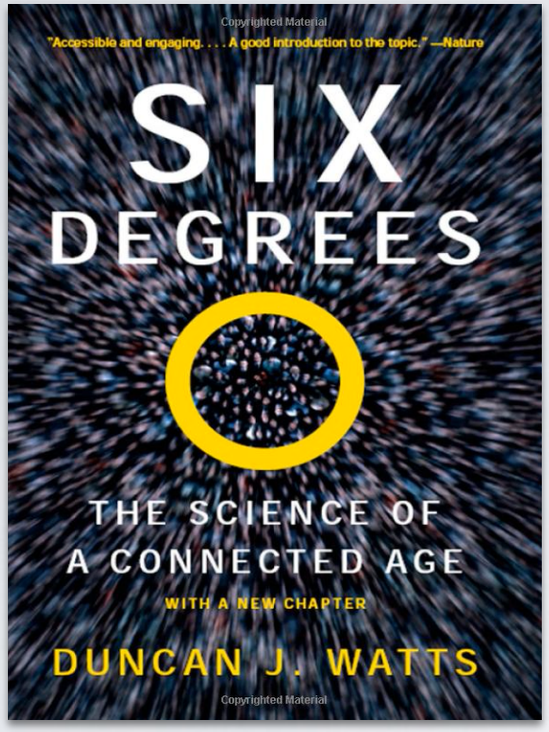


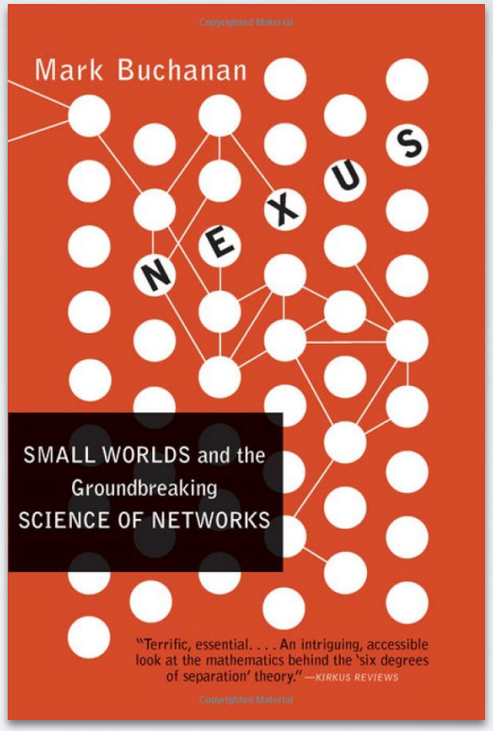


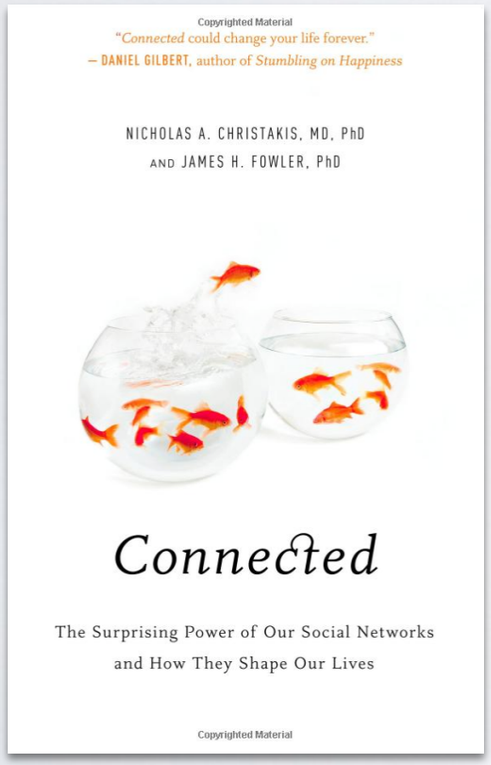


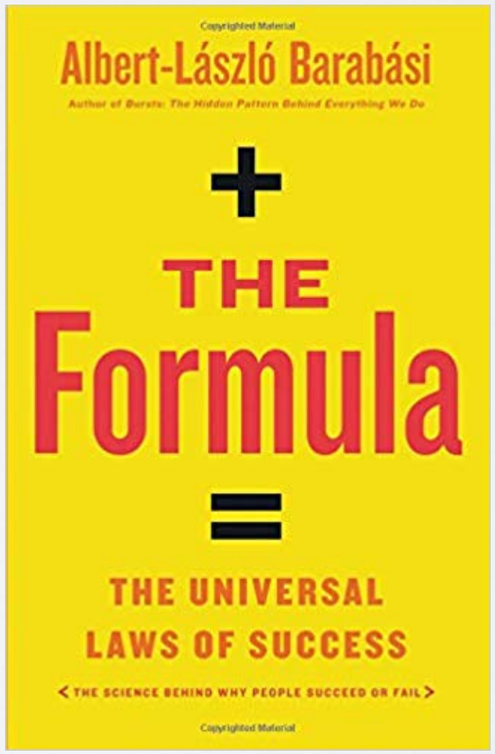








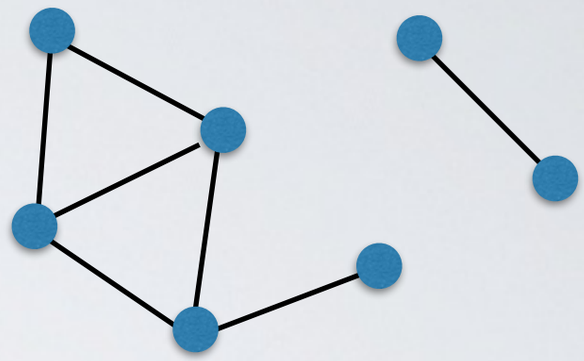




J'ai une copie que je peux prêter

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

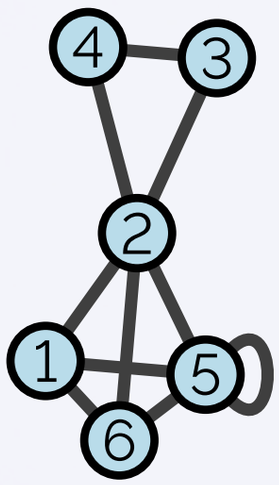
Réseaux : notation graphe

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.

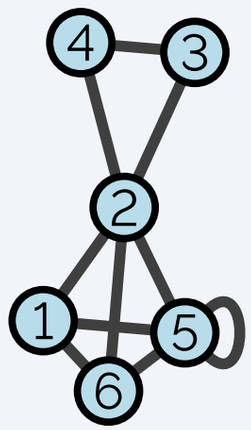
ADJACENCY MATRIX

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.

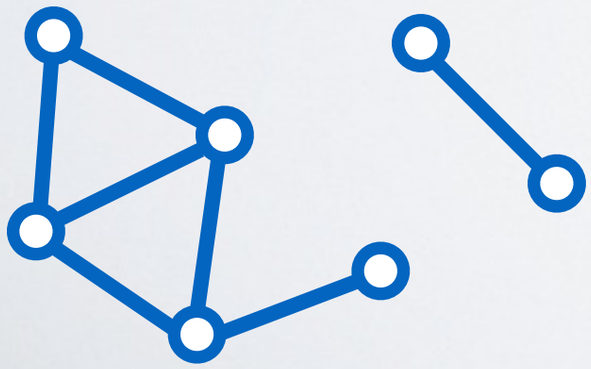
$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Types de Réseaux

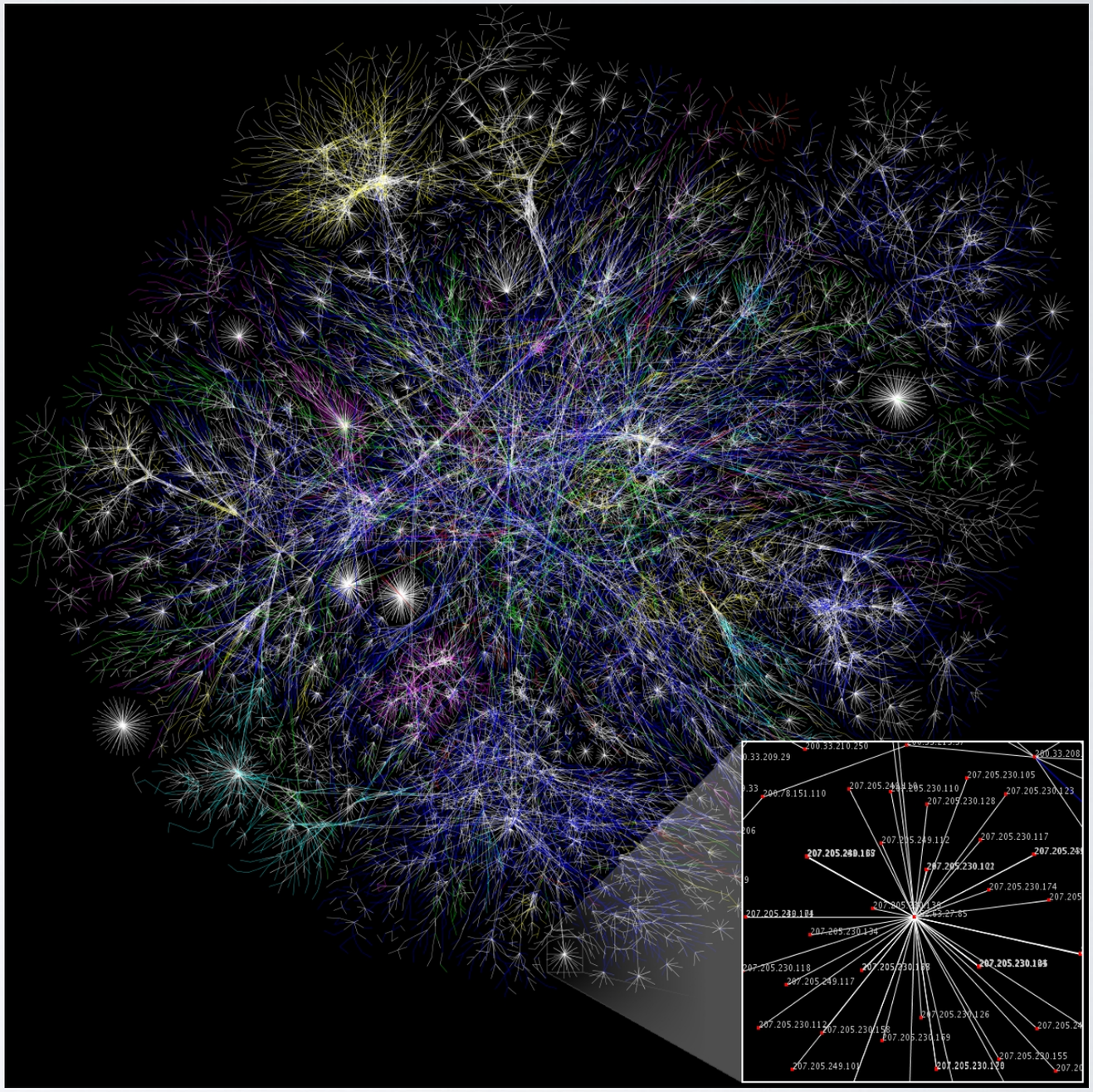
Non dirigés

$G=(V, E)$

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



Opte project

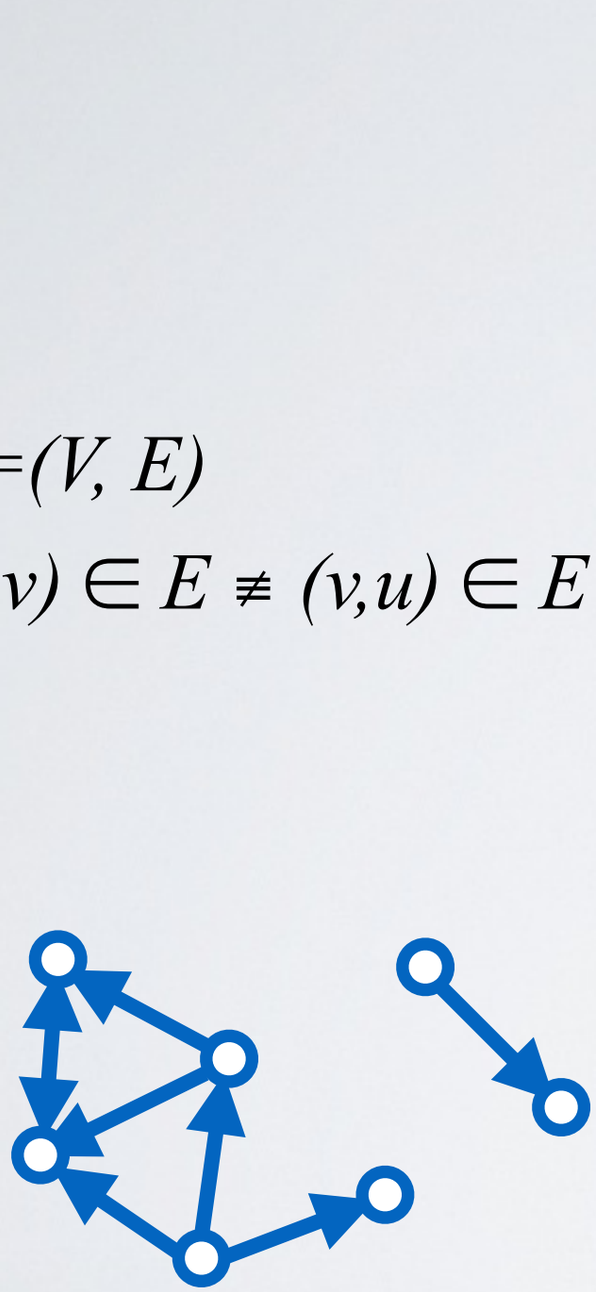


The Internet: Nodes - routers, Links - physical wires

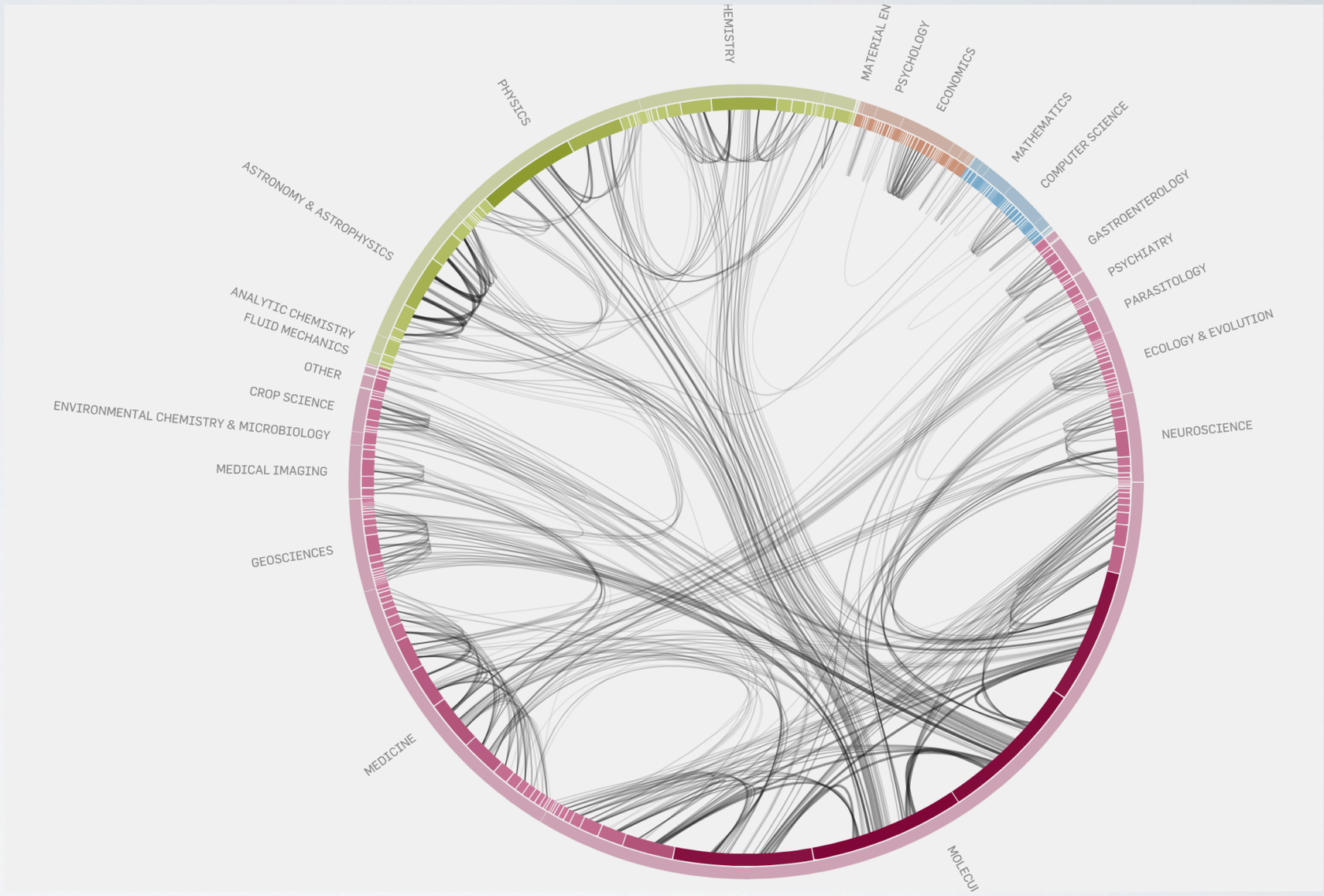
Dirigé

$$G=(V, E)$$

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



Moritz Stefaner, eigenfactor.com



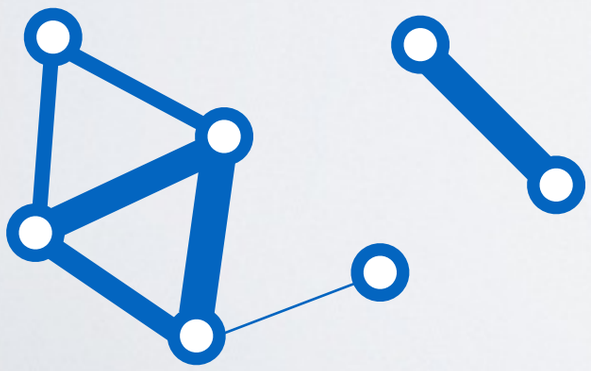
Citation network: Nodes - publications, Links - references

Réseaux pondérés

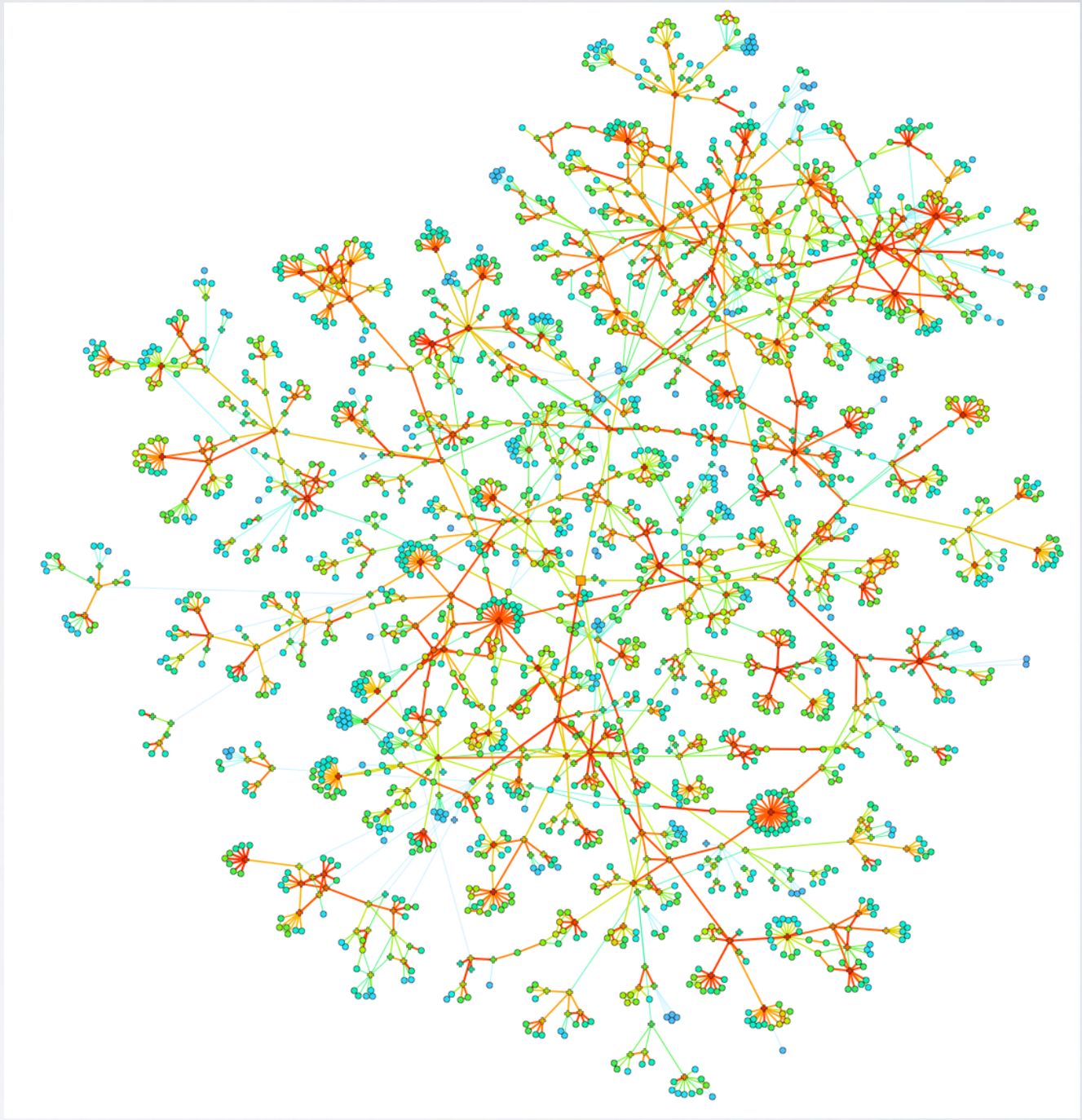
$G=(V, E, w)$

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

- La force des liens est représentée par un poids

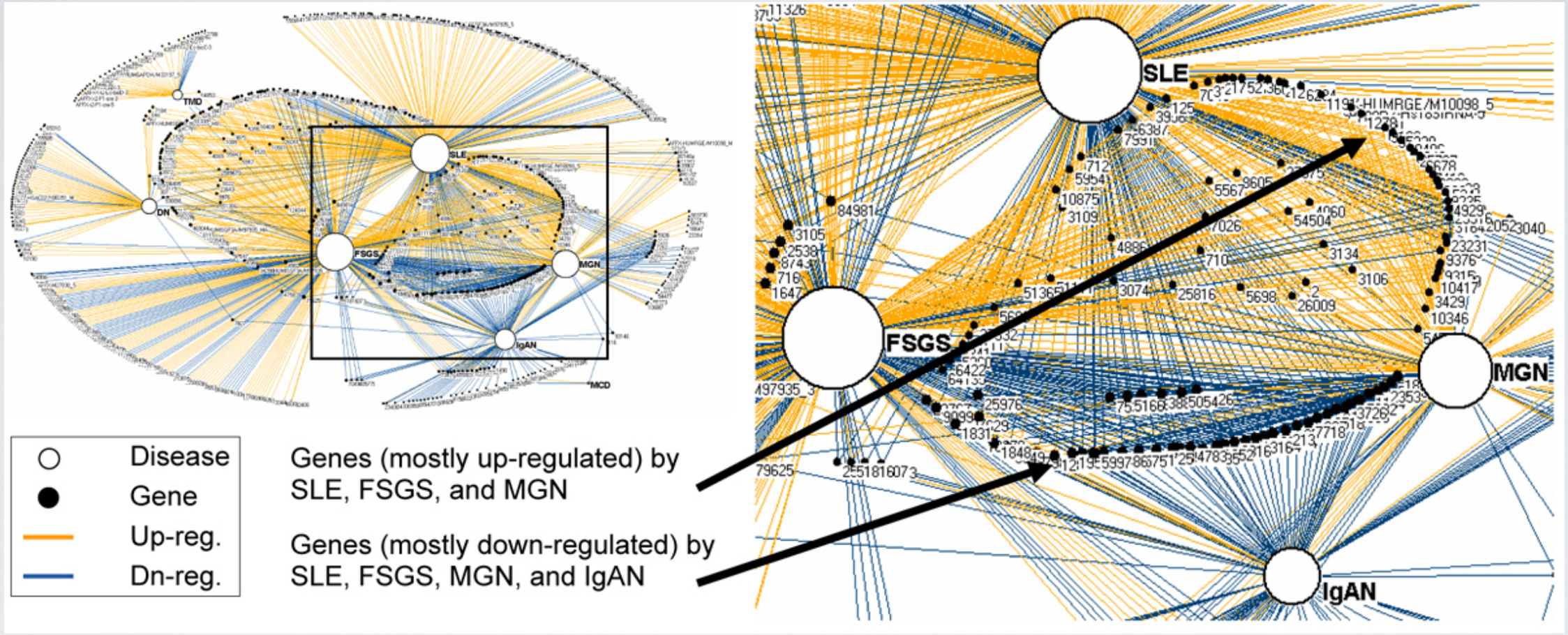


Onnela et.al. New Journal of Physics 9, 179 (2007).

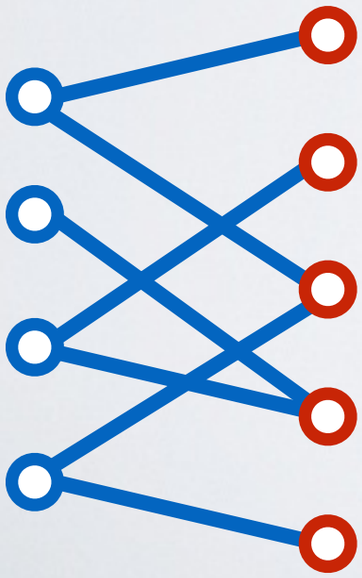


Social interaction network: Nodes - individuals
Links - social interactions

Réseaux bipartite



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



Gene-desease network:

Nodes - Desease (7)&Genes (747)

Links - gene-desease relationship

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

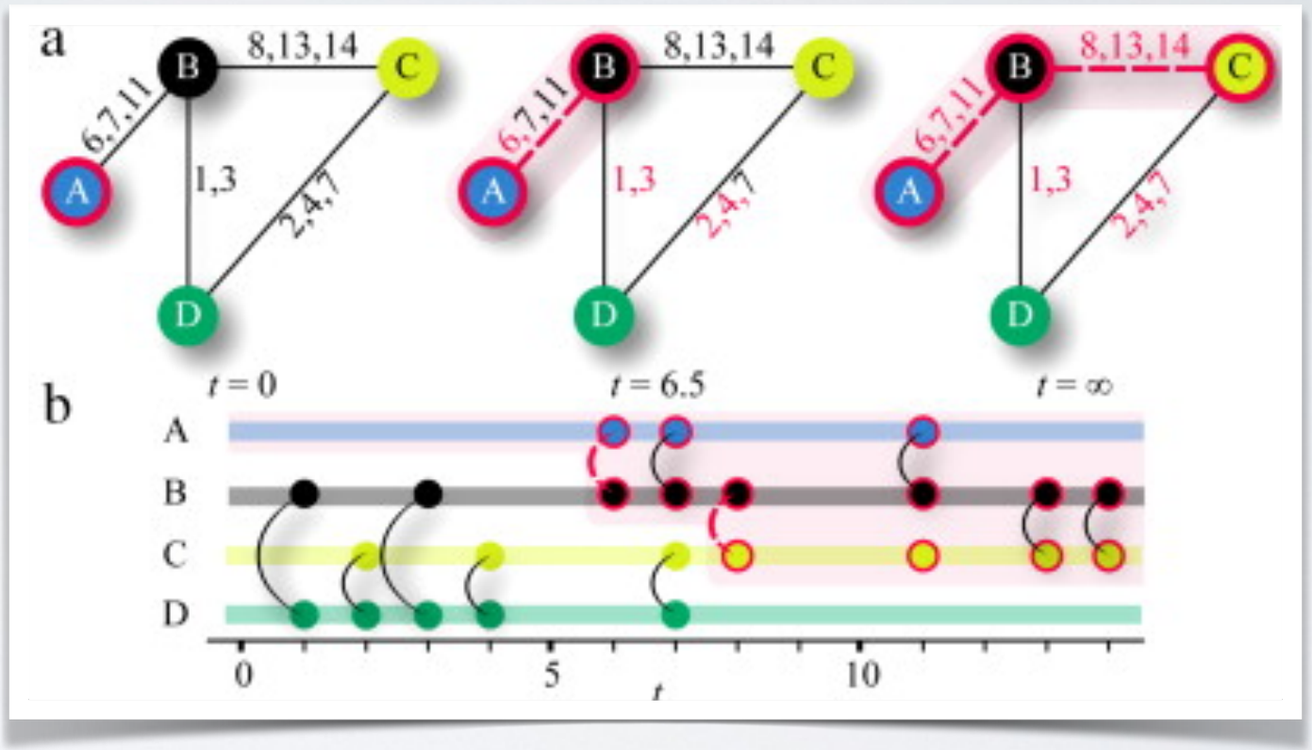
$$U \cap V = \varnothing$$

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

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)



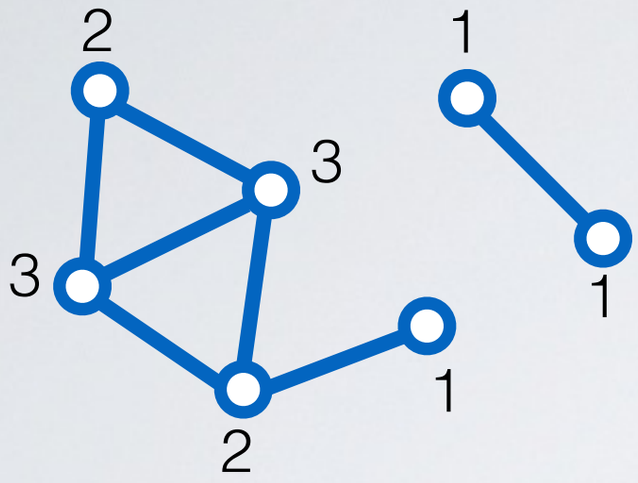
Mobile communication network
Nodes - individuals
Links - calls and SMS

Description des nœuds/liens

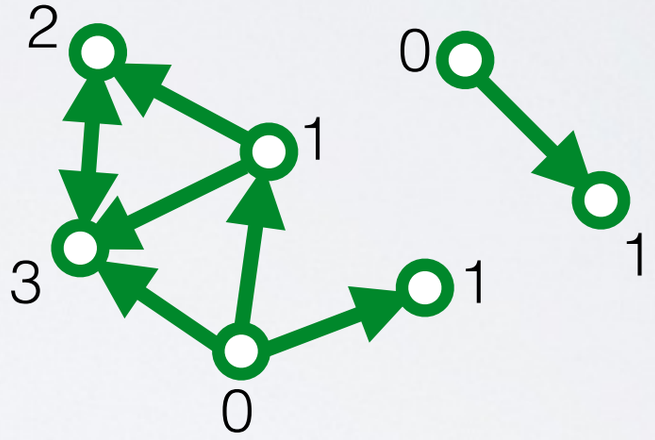
N_u	**Voisins** de u, nœuds qui partagent un lien avec u.
k_u	**Degré** de u, nombre de voisins $\|N_u\|$.
N_u^{out}	**Successeurs** de u, nœueds tels que $(u, v) \in E$ dans un graph dirigé
N_u^{in}	**Prédécesseurs** de u, nœuds tels que $(v, u) \in E$ dans un graphe dirigé
k_u^{out}	**Degré sortant**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

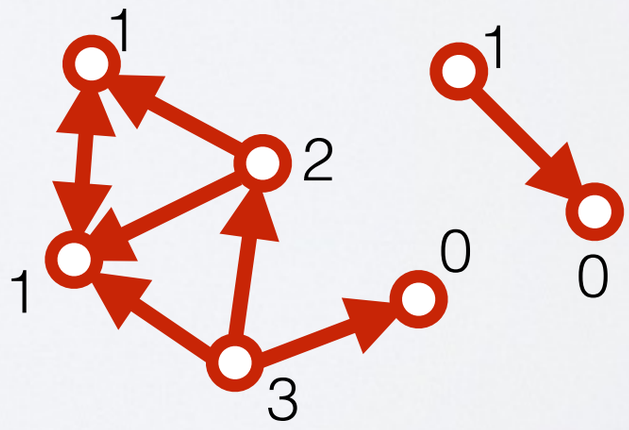
- Graphe non dirigé



- Graphe dirigé

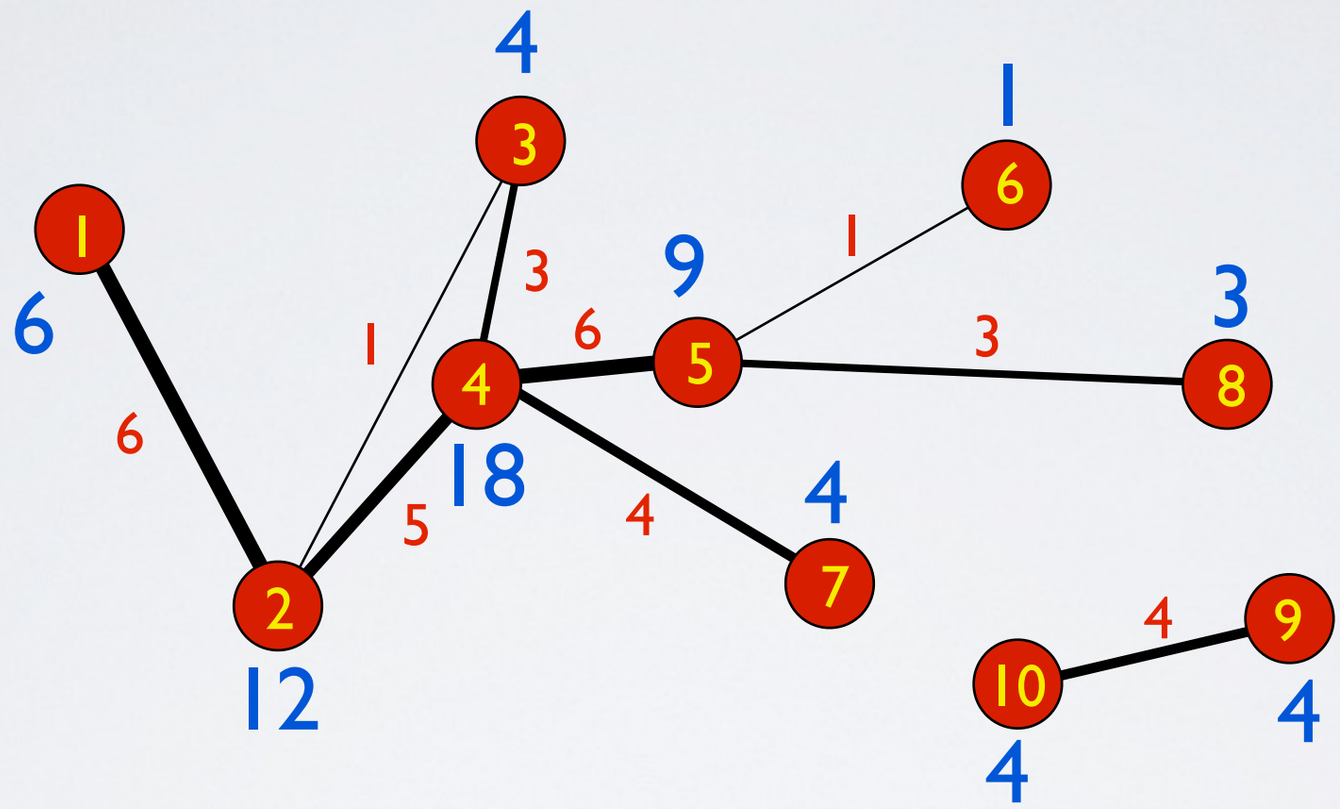


Degré entrant



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

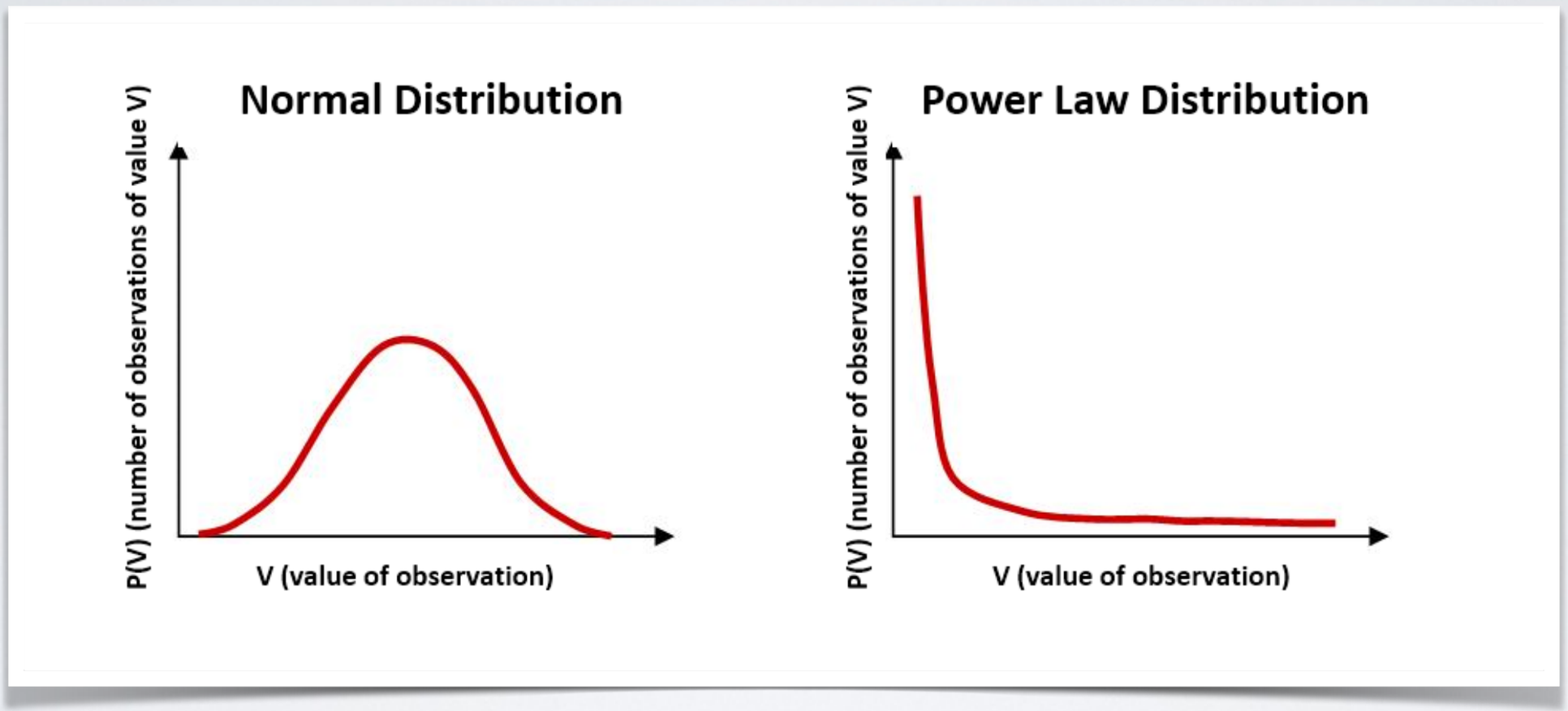
Description de réseaux - Nœuds/Liens

$\langle k \rangle$	**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)

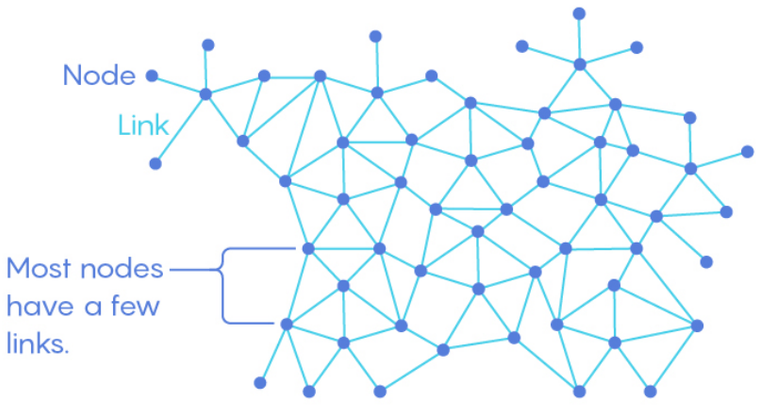
- Loi de puissance (**power law**)

To Be or Not to Be Scale-Free

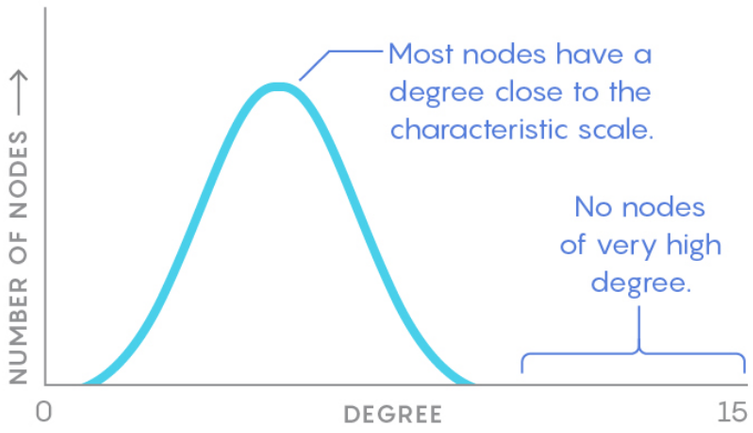
Scientists study complex networks by looking at the distribution of the number of links (or "degree") of each node.
Some experts see so-called scale-free networks everywhere. But a new study suggests greater diversity in real-world networks.

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.

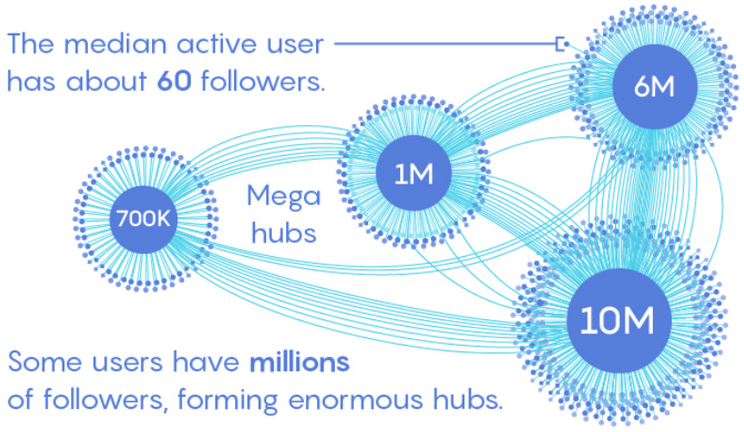


The distribution of degrees is shaped roughly like a bell curve that peaks at the network's "characteristic scale."

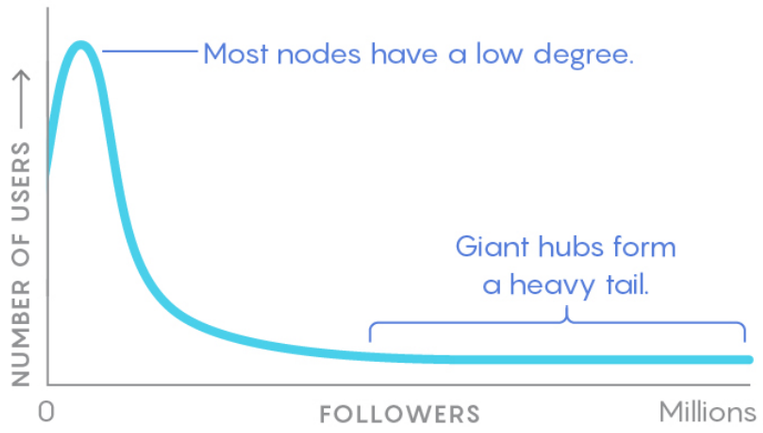


Twitter's Scale-Free Network

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

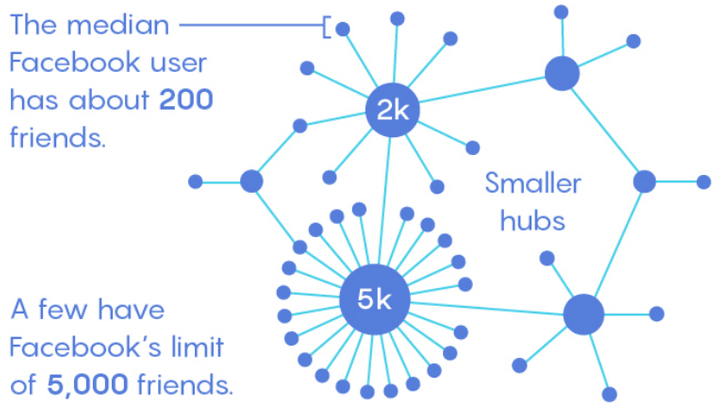


The degrees roughly follow a power law distribution that has a "heavy tail." The distribution has no characteristic scale, making it scale-free.

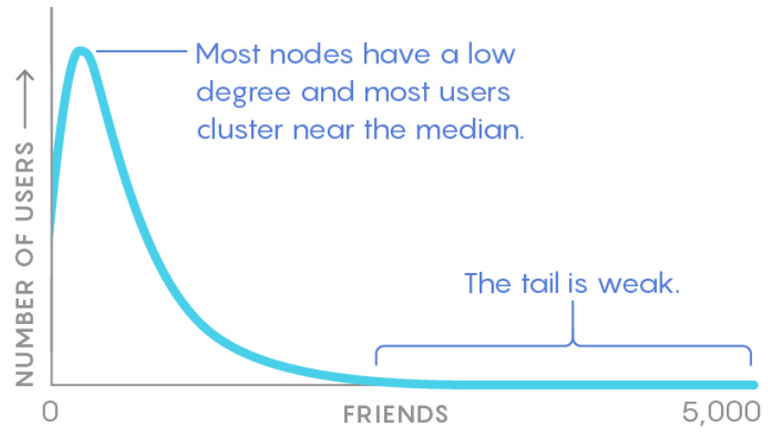


Facebook's In-Between Network

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

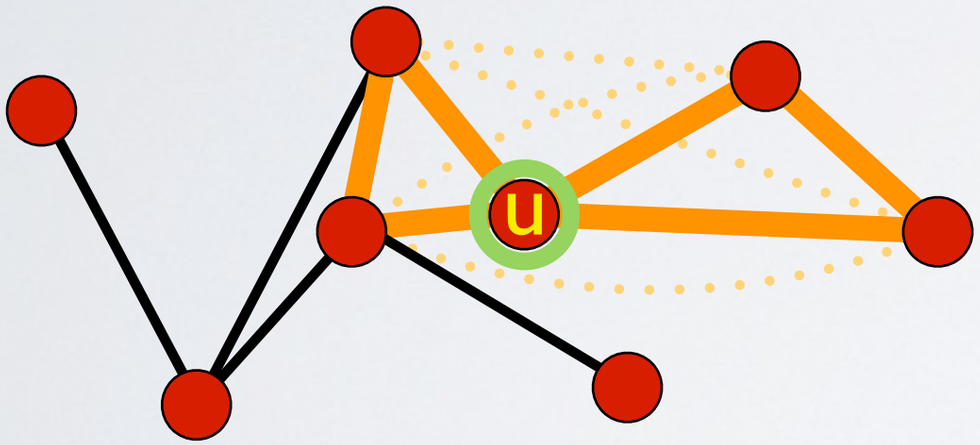
δ_u - **Triades de u:** nombre de triangles contenant le node u
Δ - **Nombre de triangles dans le graphe** $\Delta = \frac{1}{3}\sum_{u \in V} \delta_u$.

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

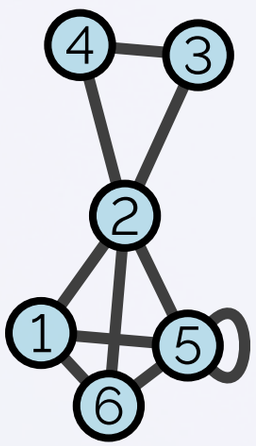
- CC Global:
  - Dans un réseau aléatoire, CC global = densité
    - =>Très petit pour des grands graphes
  - Facebook ego-networks: 0.6
  - Twitter lists: 0.56
  - California Road networks: 0.04

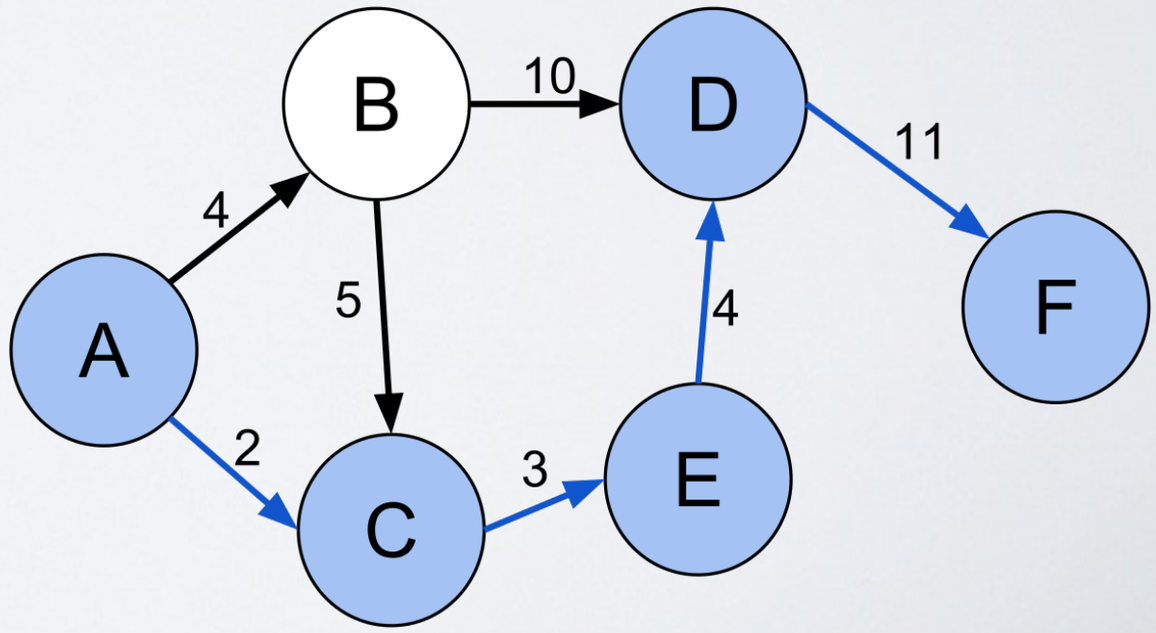
CHEMINS/MARCHES

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

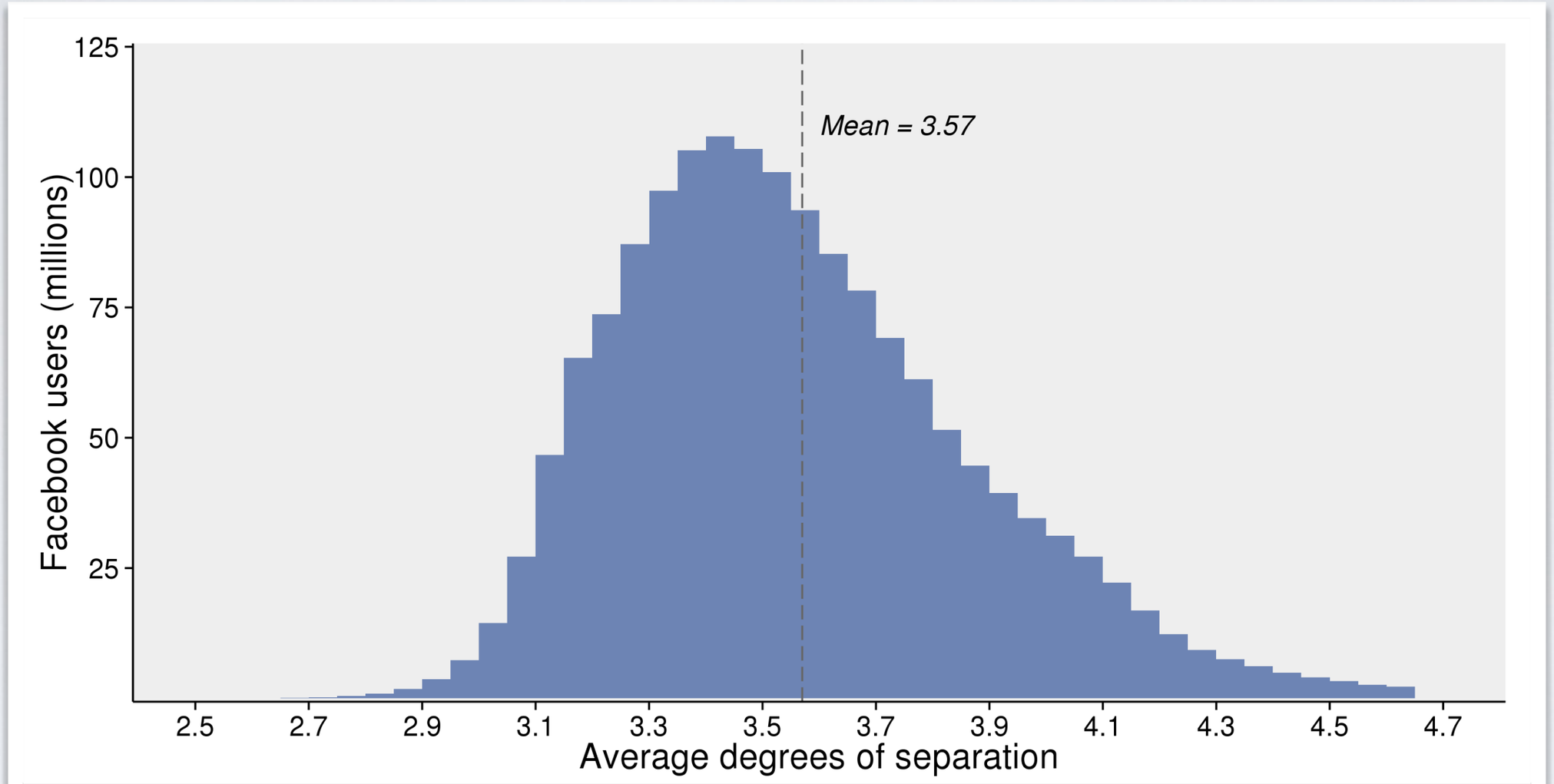
- Expériences du petit monde (60's)
  - Donner une enveloppe à des individus pris au hasard
  - Demander de l'envoyer à une personne qu'ils ne connaissent pas
    - A partir d'information (genre, age, métier)
  - Ils font transiter par des connaissances

- 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.

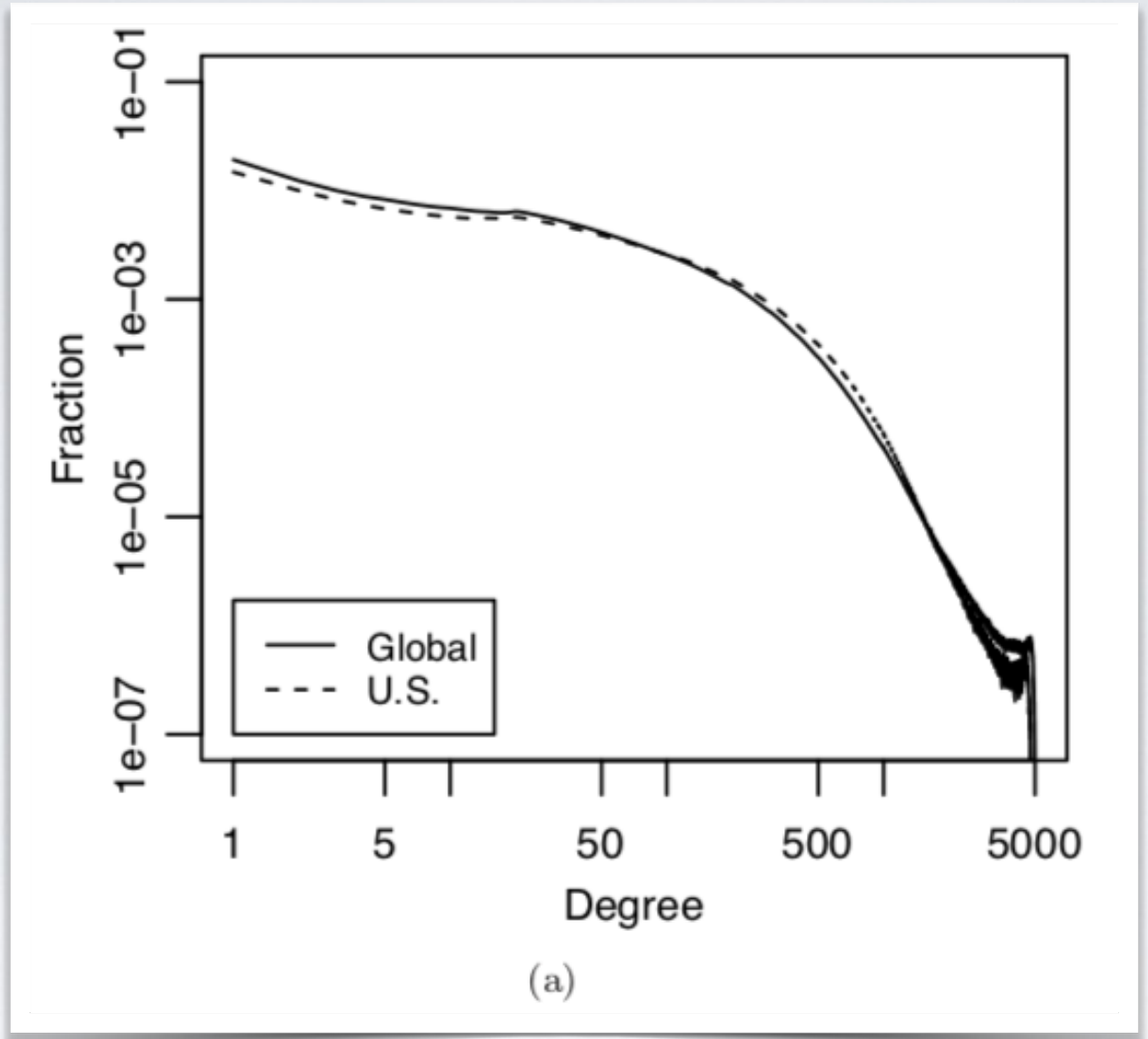
EXAMPLE D'ANALYSE DE GRAPHES

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

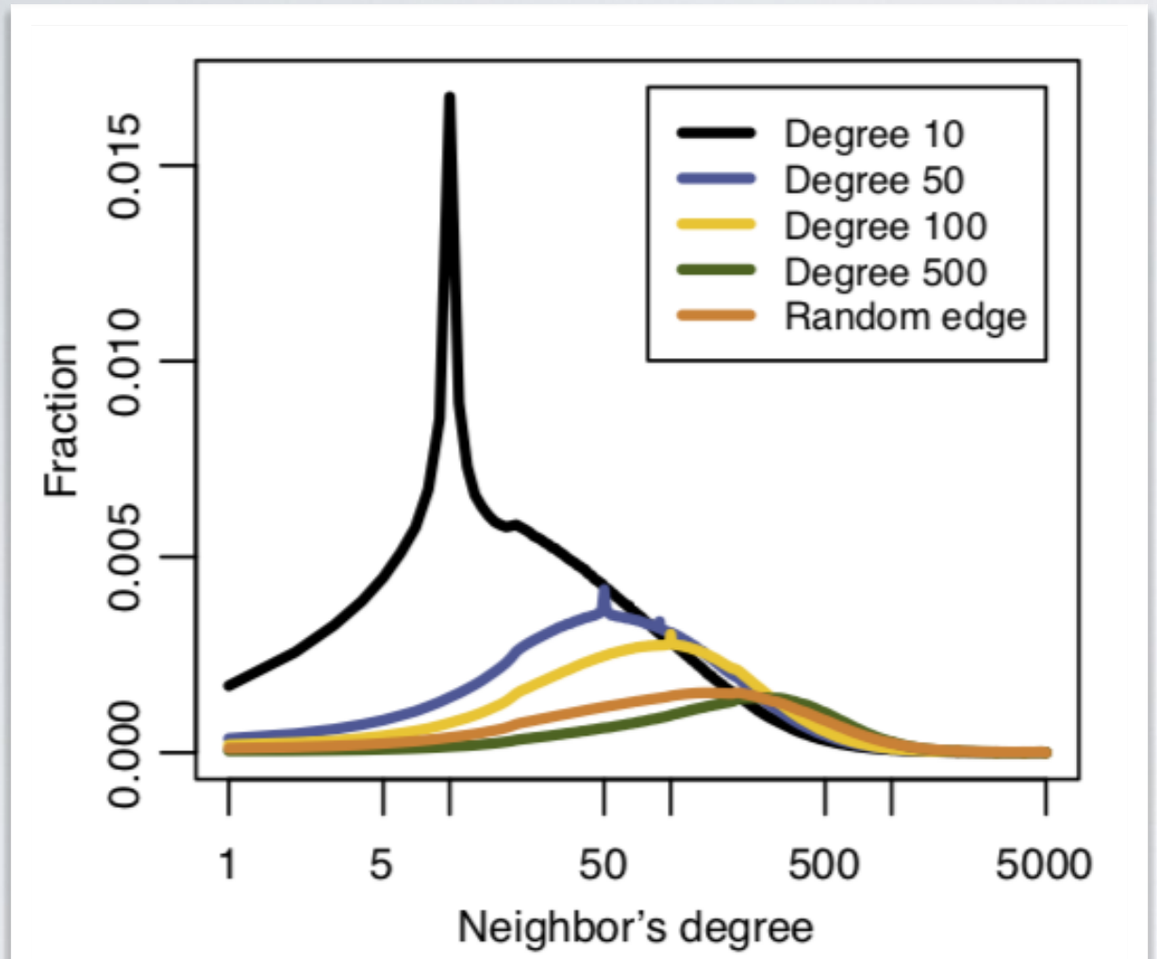
- Le réseau "d'amis" Facebook 2011

EXAMPLE D'ANALYSE DE GRAPHES

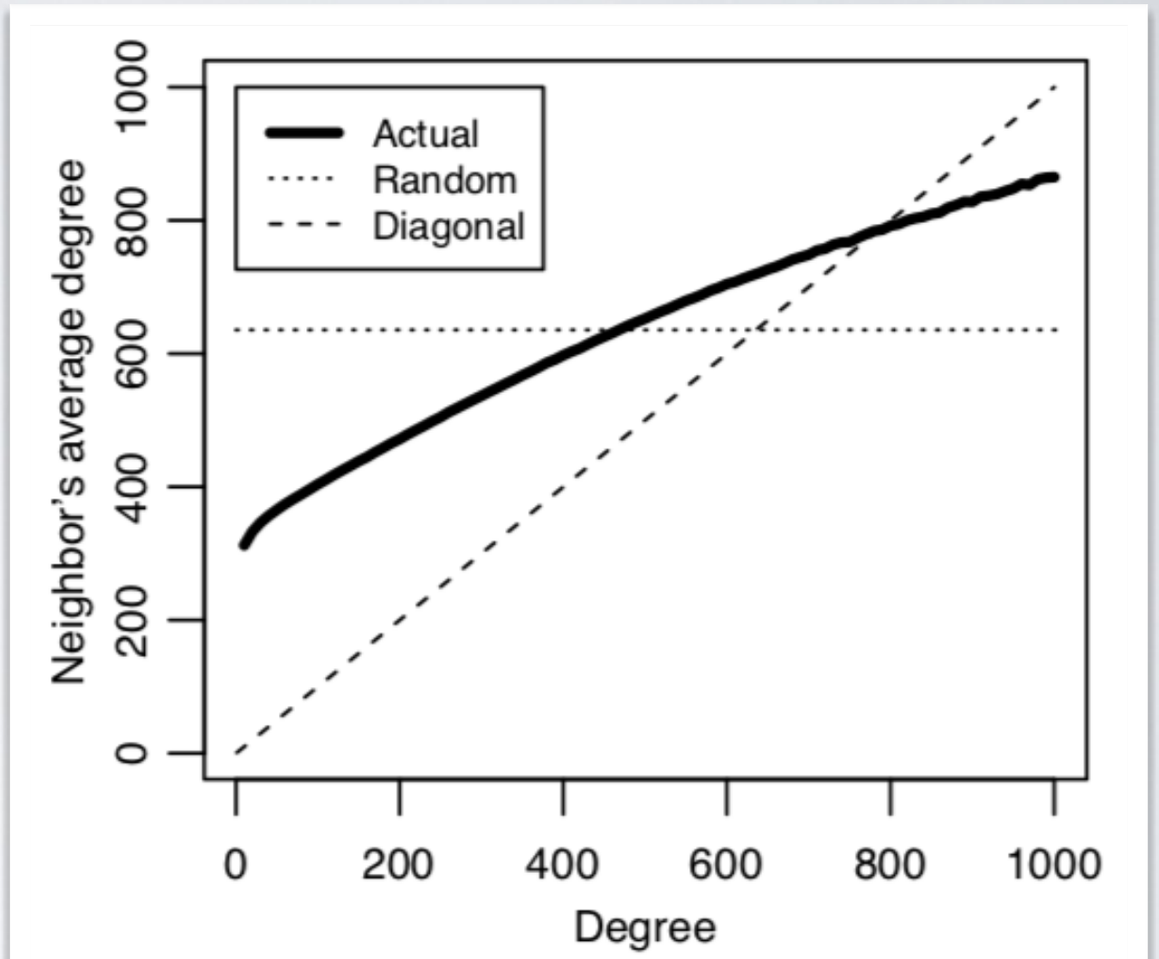
- 721M d'utilisateurs (nœuds) (actifs au cours des 28 derniers jours)
- 68 Milliards de liens
- Degré moyen: 190
- Degré médian: 99
- Composante connexe principale : 99.91%



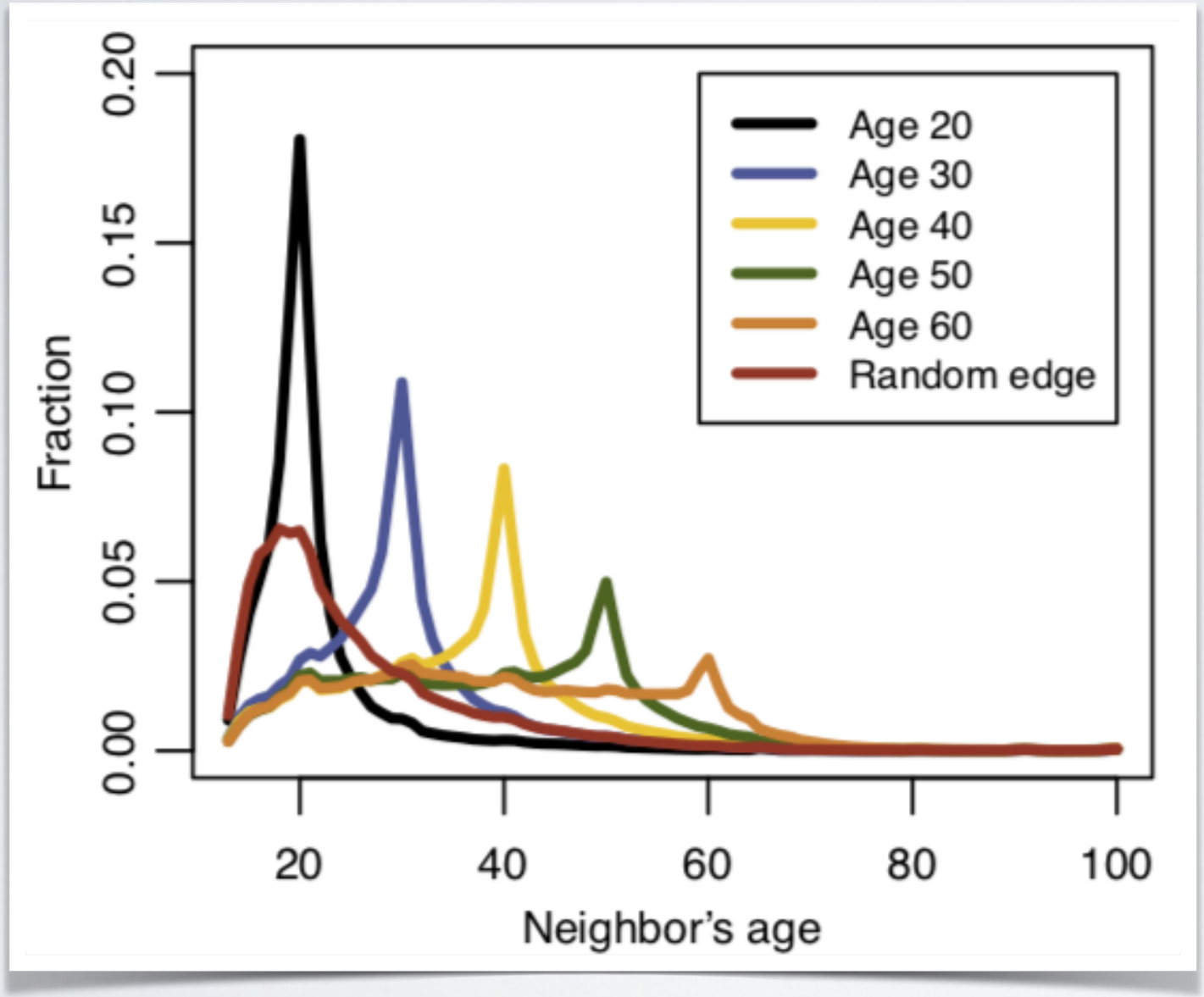
Distribution de degrés



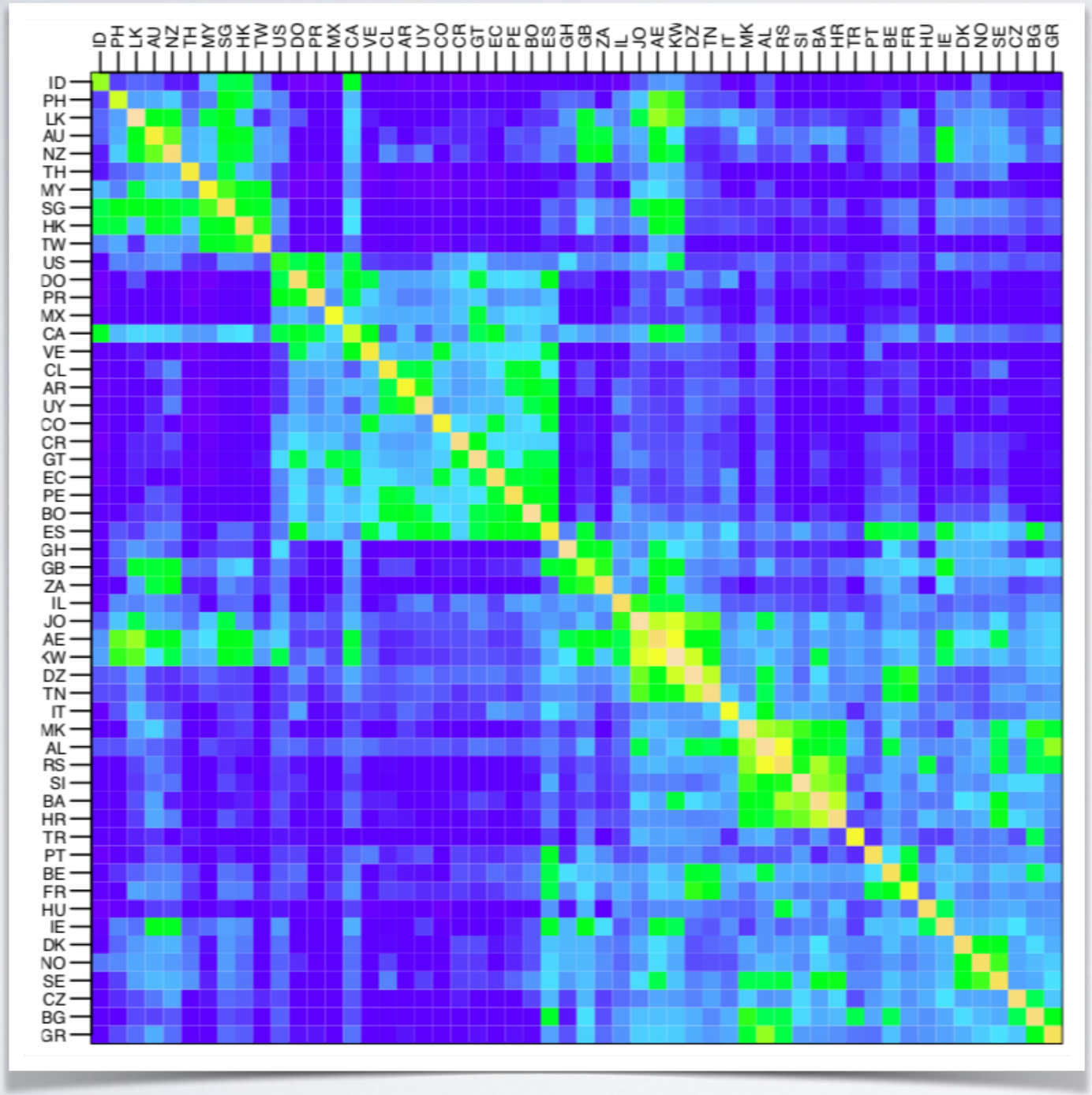
Beaucoup de mes amis ont le même Nombre d'amis que moi…



Mes amis ont plus d'amis que moi !



Homophily en fonction de l'âge



Similarité par pays

84.2% des liens sont à l'intérieur des pays