Introduction
à l'analyse des réseaux sociaux

Emmanuel Viennet

Université Paris 13, Sorbonne-Paris-Cité

emmanuel.viennet@univ-paris13.fr

Plan du cours

1. Introduction
2. Concepts de base (théorie des graphes)
3. Quelques outils logiciels
4. Propriétés des réseaux complexes
5. Communautés
6. Une approche simple pour l'analyse des réseaux sociaux
7. Quelques exemples d'applications

Crédits

Quelques éléments empruntés à:

- Lada Adamic (University of Michigan)
- Albert Lazlo Barabasí (Northeastern University)
- Françoise Fogelman-Soulié (Institut Mines-Télécoms)

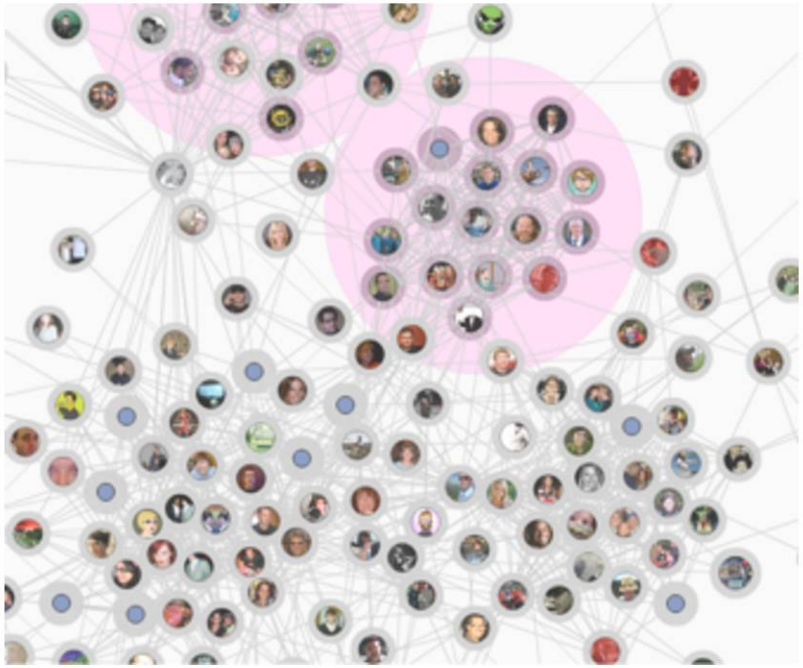
- Mes étudiants: The Anh Dang, Daniel Bernardes, Mamadou Diaby, Raphaël Fournier, Ivan Keller, Patricia Condé-Cespedes

Introduction
à l'analyse des réseaux sociaux

Partie 1
Introduction

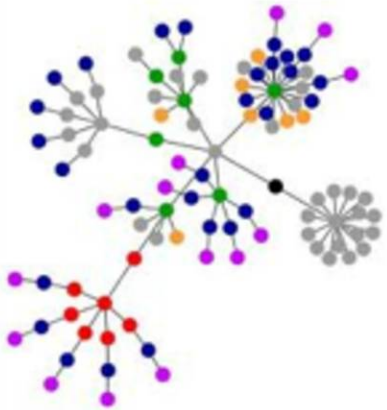
Réseaux sociaux ?

Données issues des réseaux sociaux ?



Text

```
[
  {
    "name": "apple",
    "type": "fruit"
  },
  {
    "name": "banana",
    "type": "fruit"
  },
  {
    "name": "cucumber",
    "type": "vegetable"
```



Analyse des réseaux sociaux (ou réseaux « complexes »)

Fouille de données

Sociologie

Probabilités/Statistiques

Apprentissage

Théorie des graphes

Link Analysis

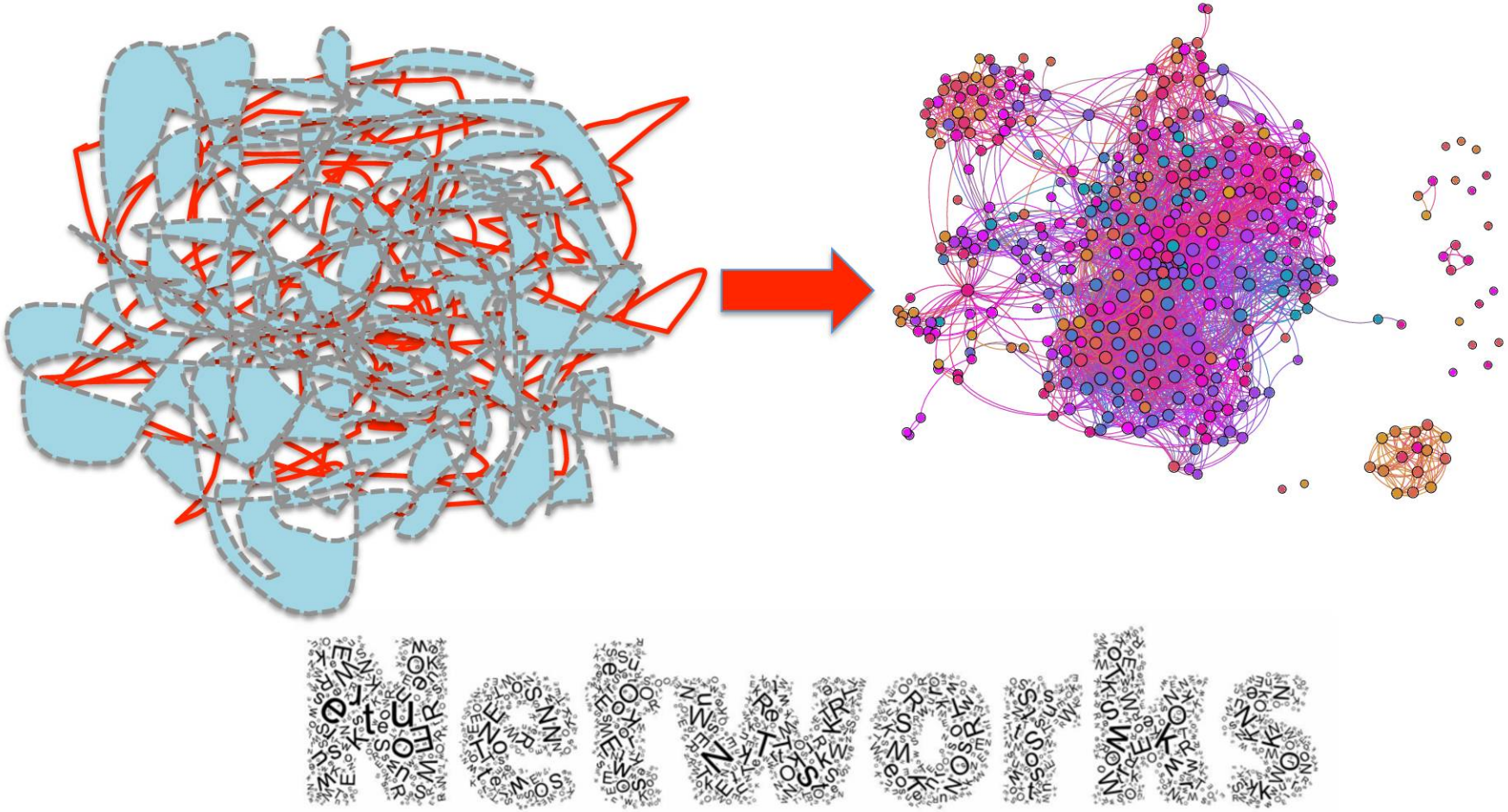
Algorithmique

Données textuelles (TAL)

Visualisation

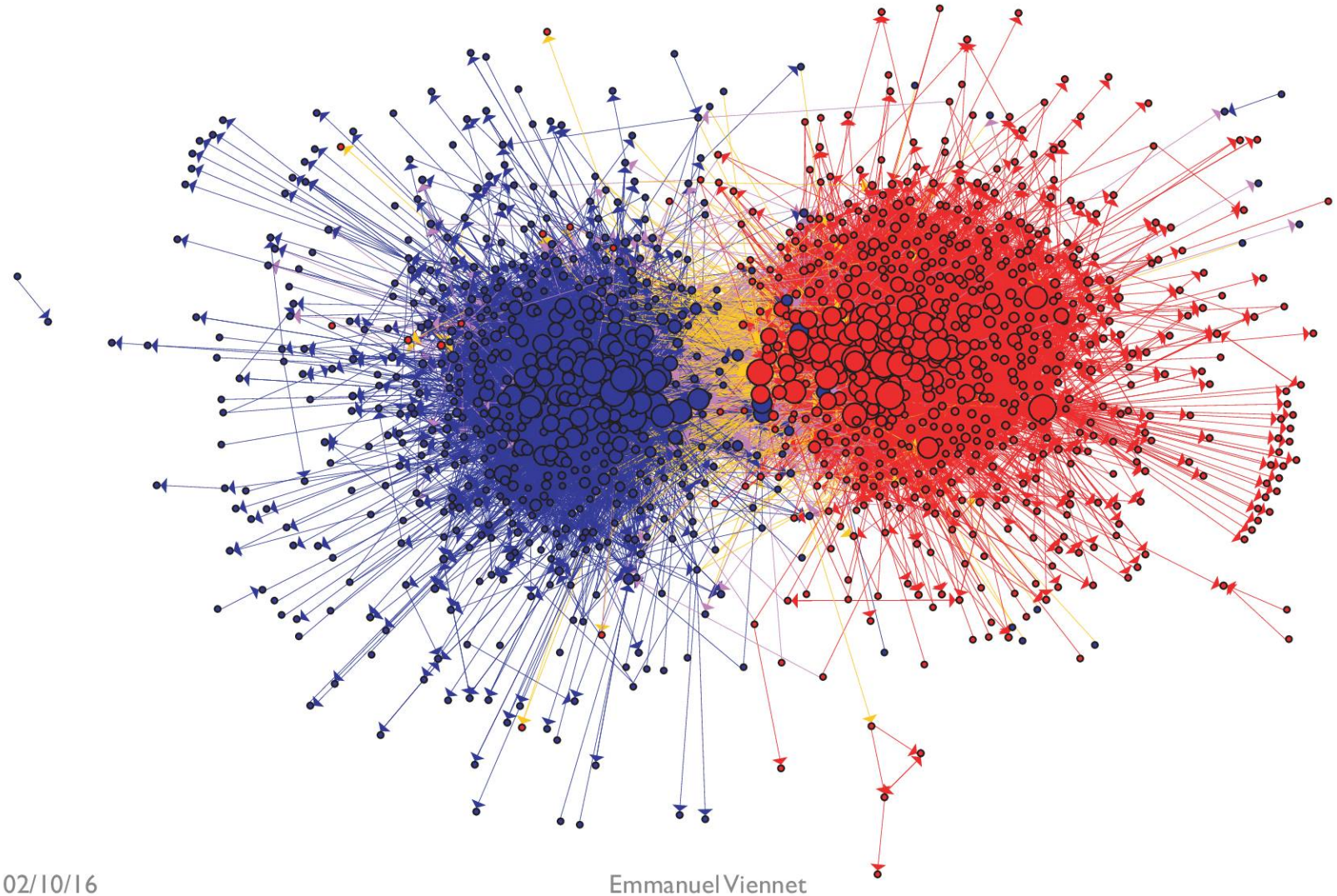
Multimédia

Les réseaux pour modéliser des systèmes complexes

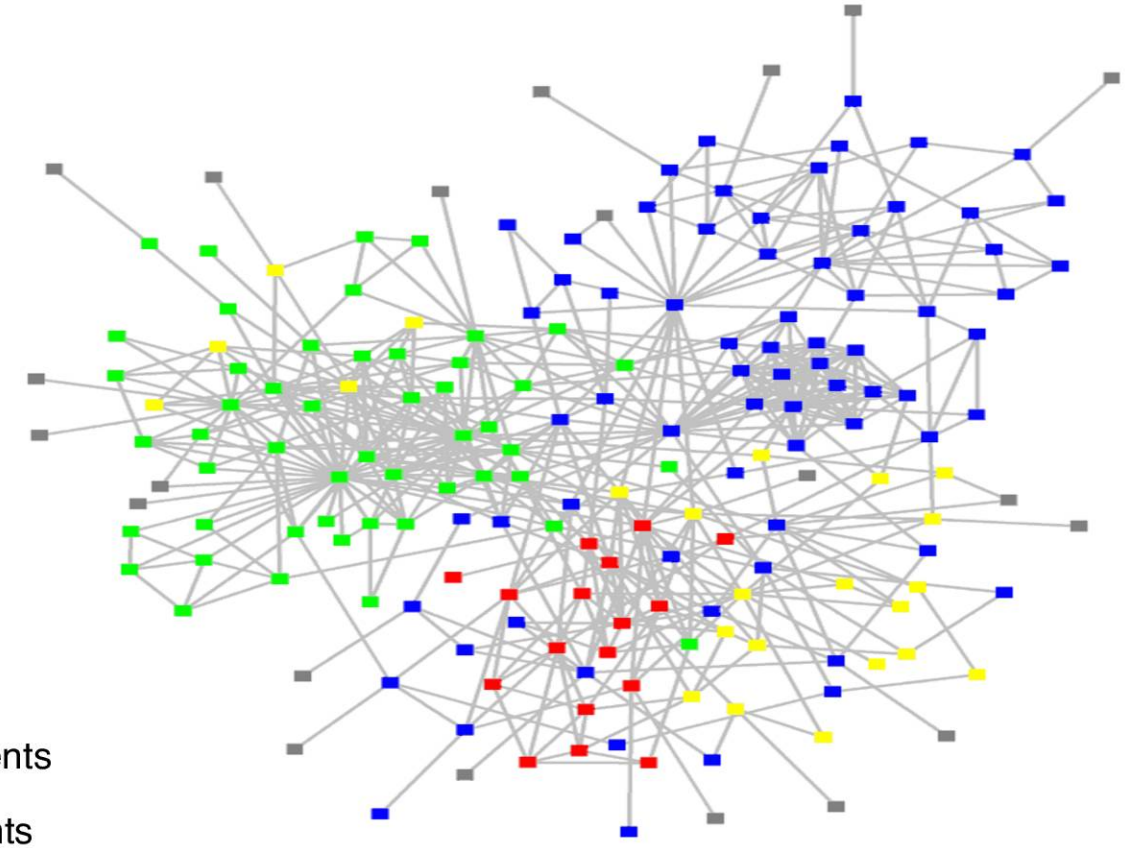


Credit: Lada Adamic

Exemple 1: blogs politiques



Exemple 2: Organisations, entreprises



: departments

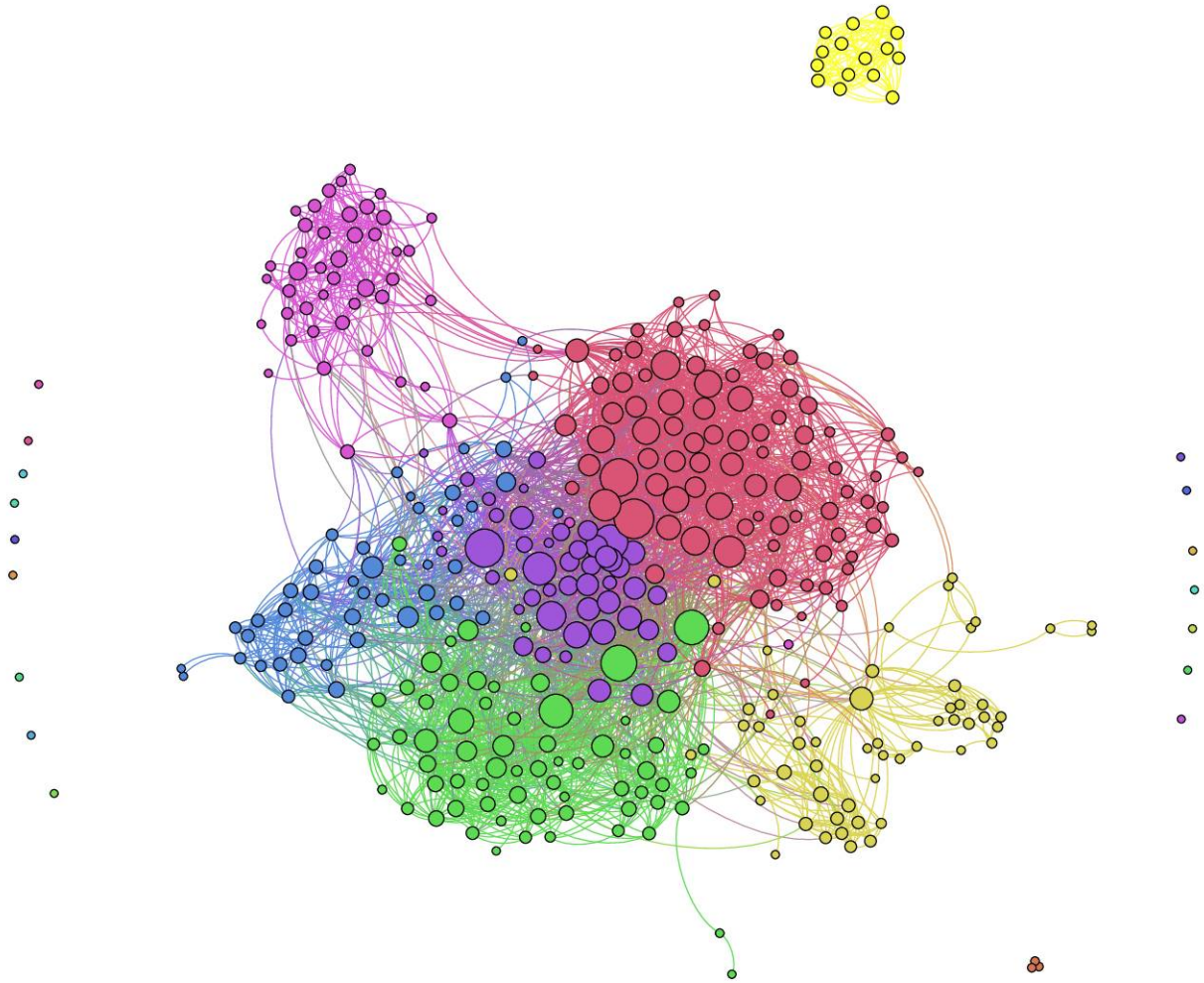
: consultants

: external experts

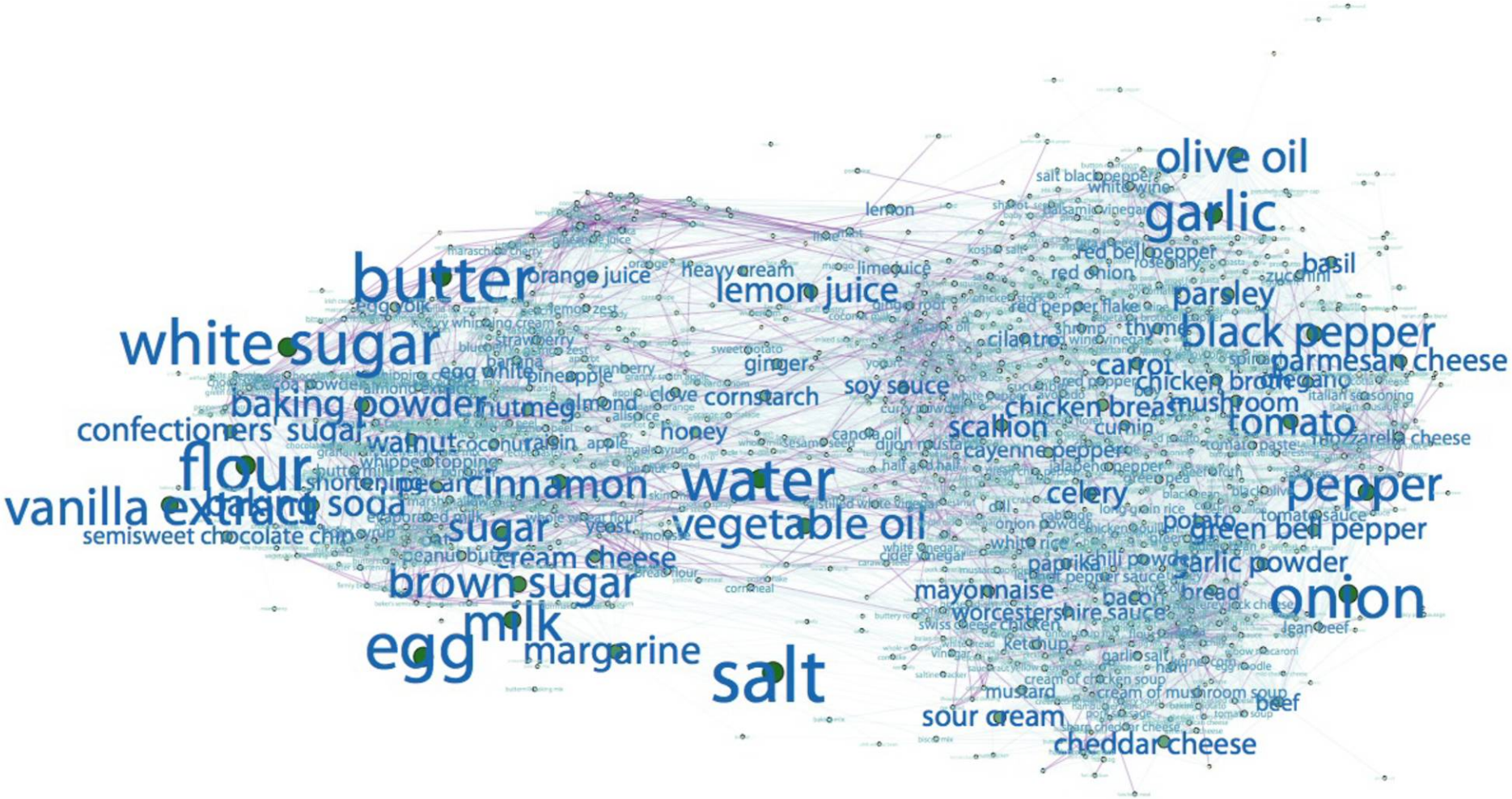
Graphe social de Facebook

Keith Shepherd's "Sunday Best". http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/

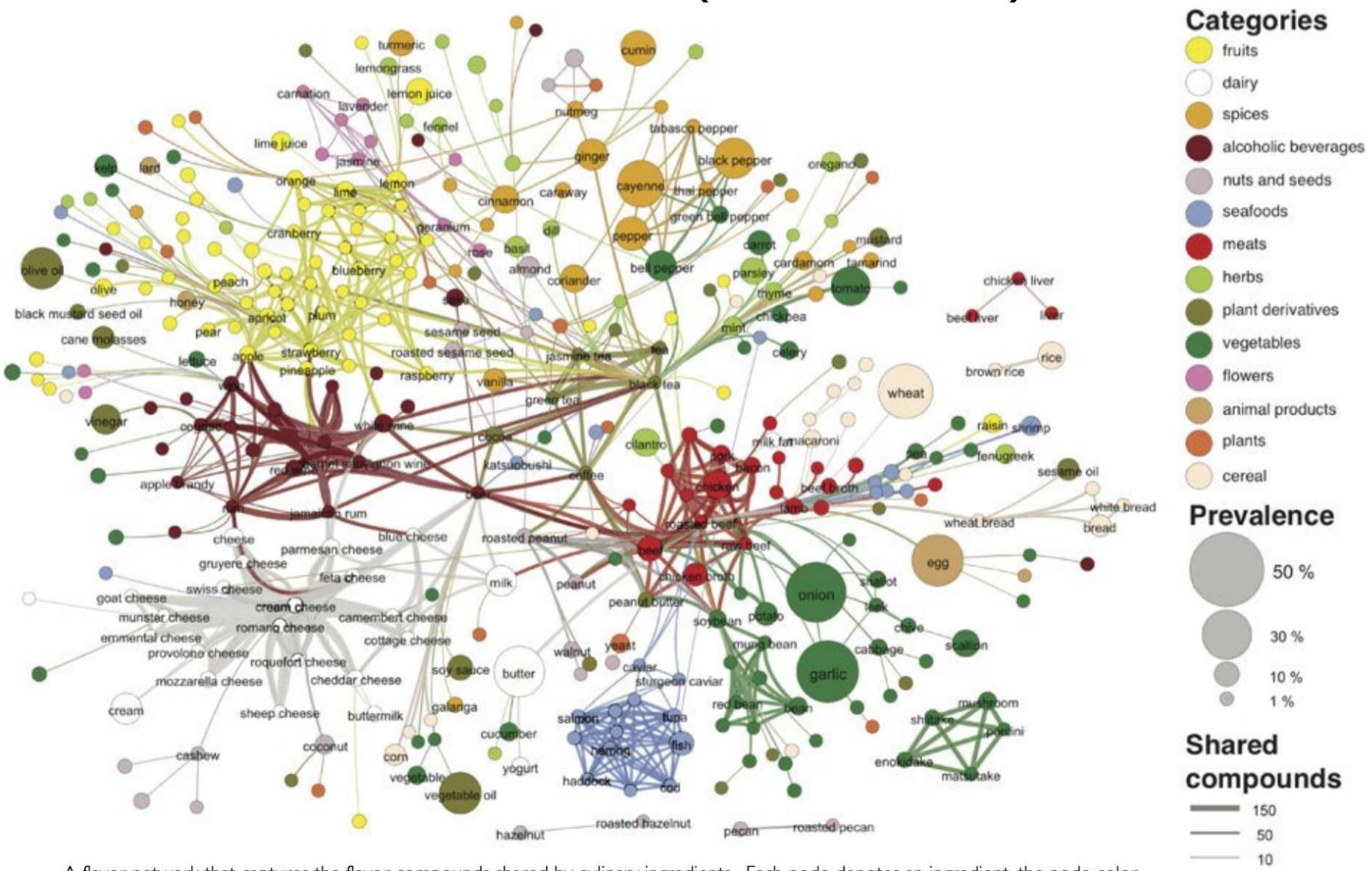
Graphe social de Facebook (local)



Réseau d'ingrédients (recettes)



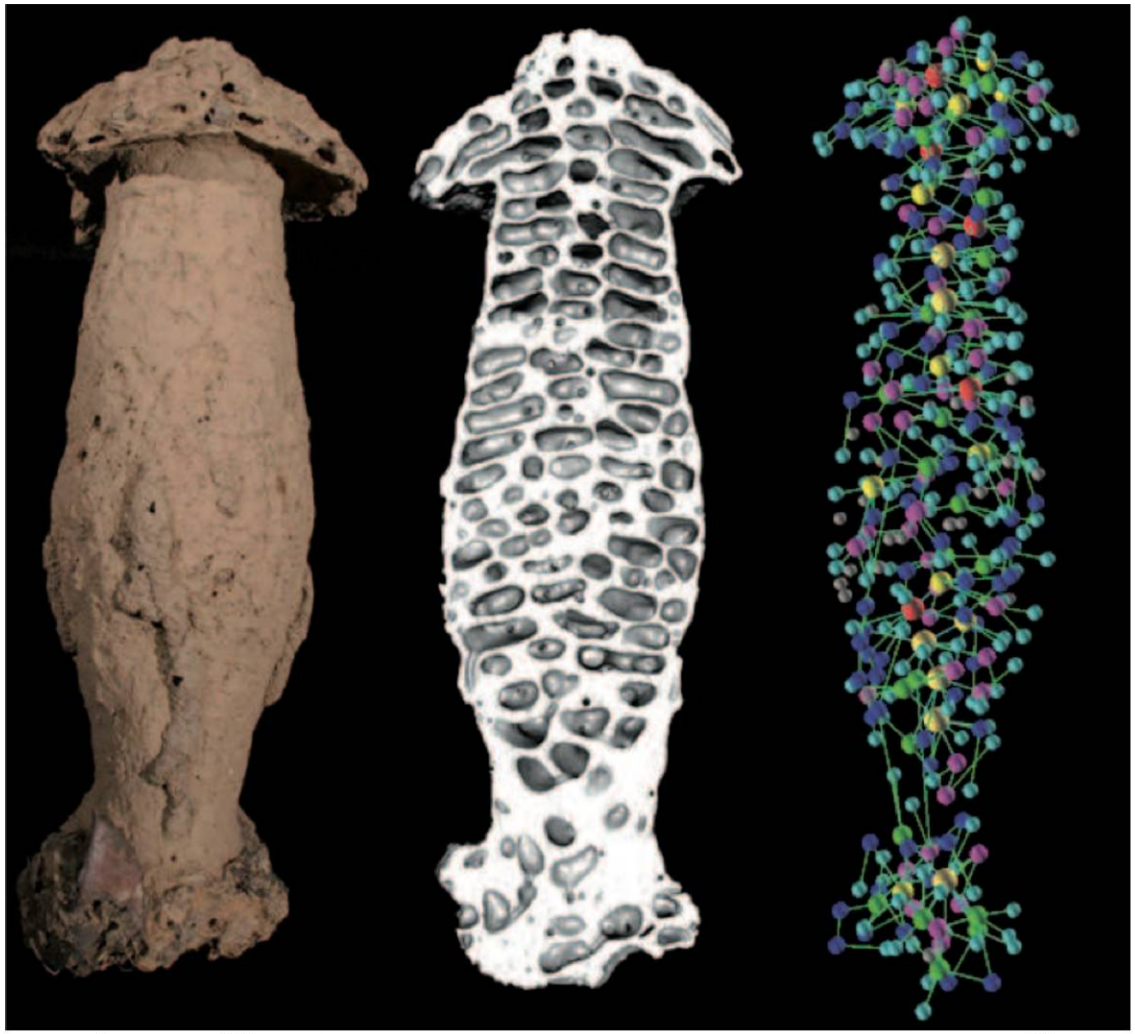
Saveurs (recettes)



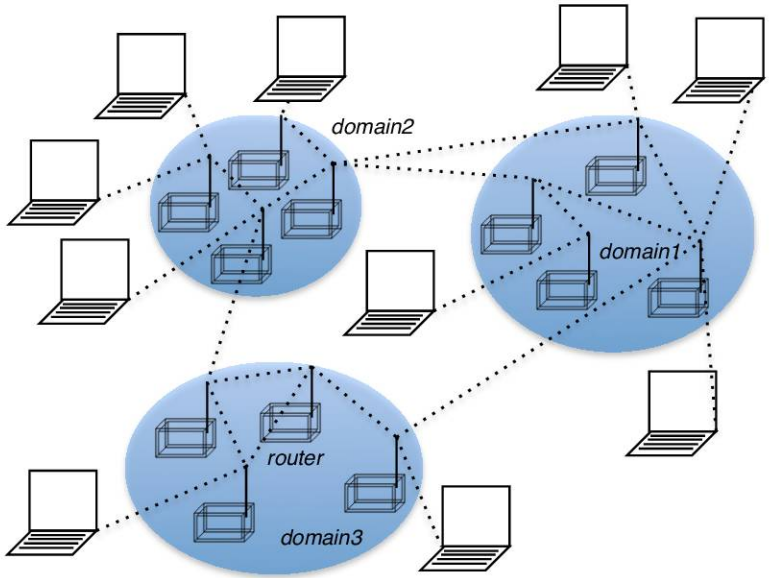
A flavor network that captures the flavor compounds shared by culinary ingredients. Each node denotes an ingredient, the node color indicates food category, and node size reflects the ingredient prevalence in recipes. Two ingredients are connected if they share a significant number of flavor compounds, link thickness representing the number of shared compounds between the two ingredients. (Barabasi et al 2012)

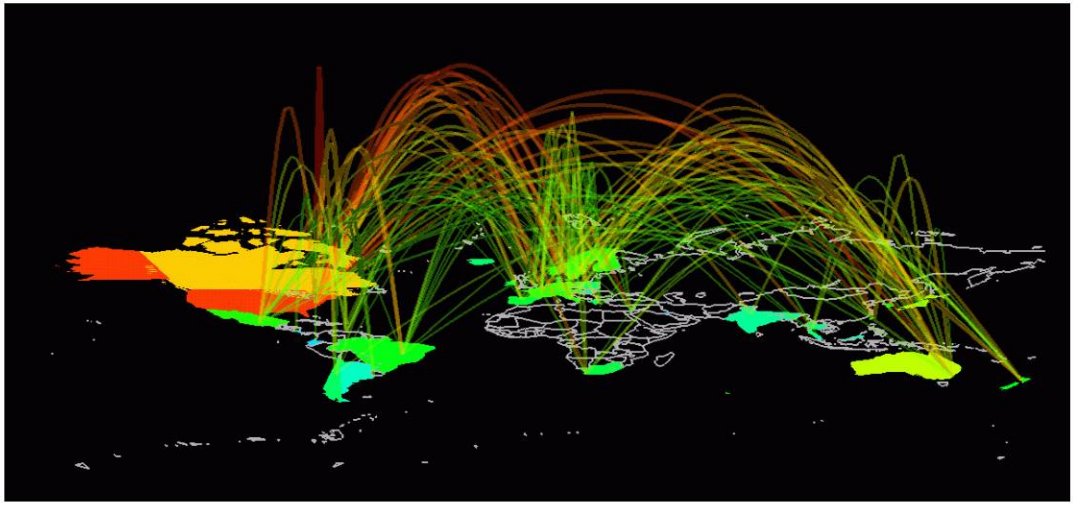
© Thomas Plotkowiak 2010

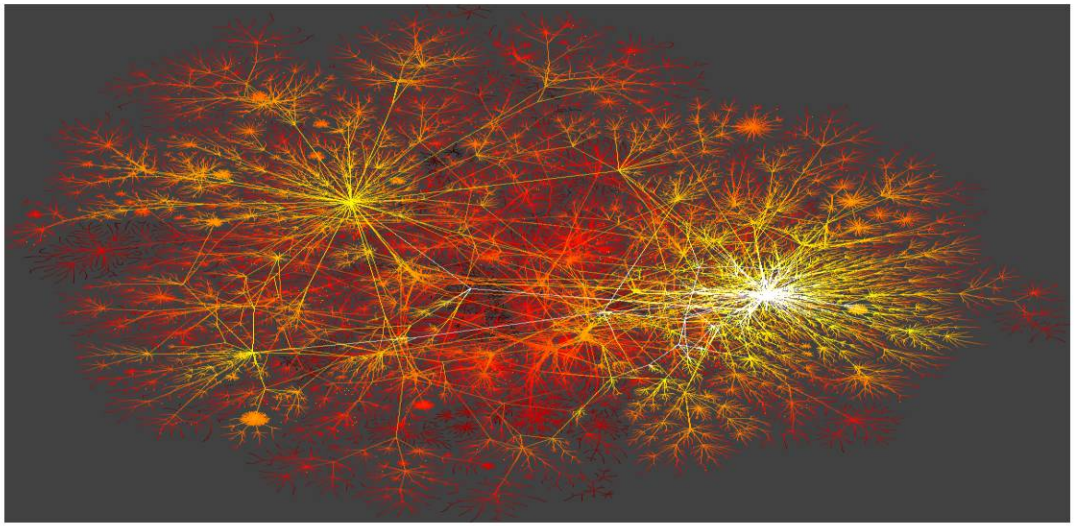
Modélisation de termitières



Internet

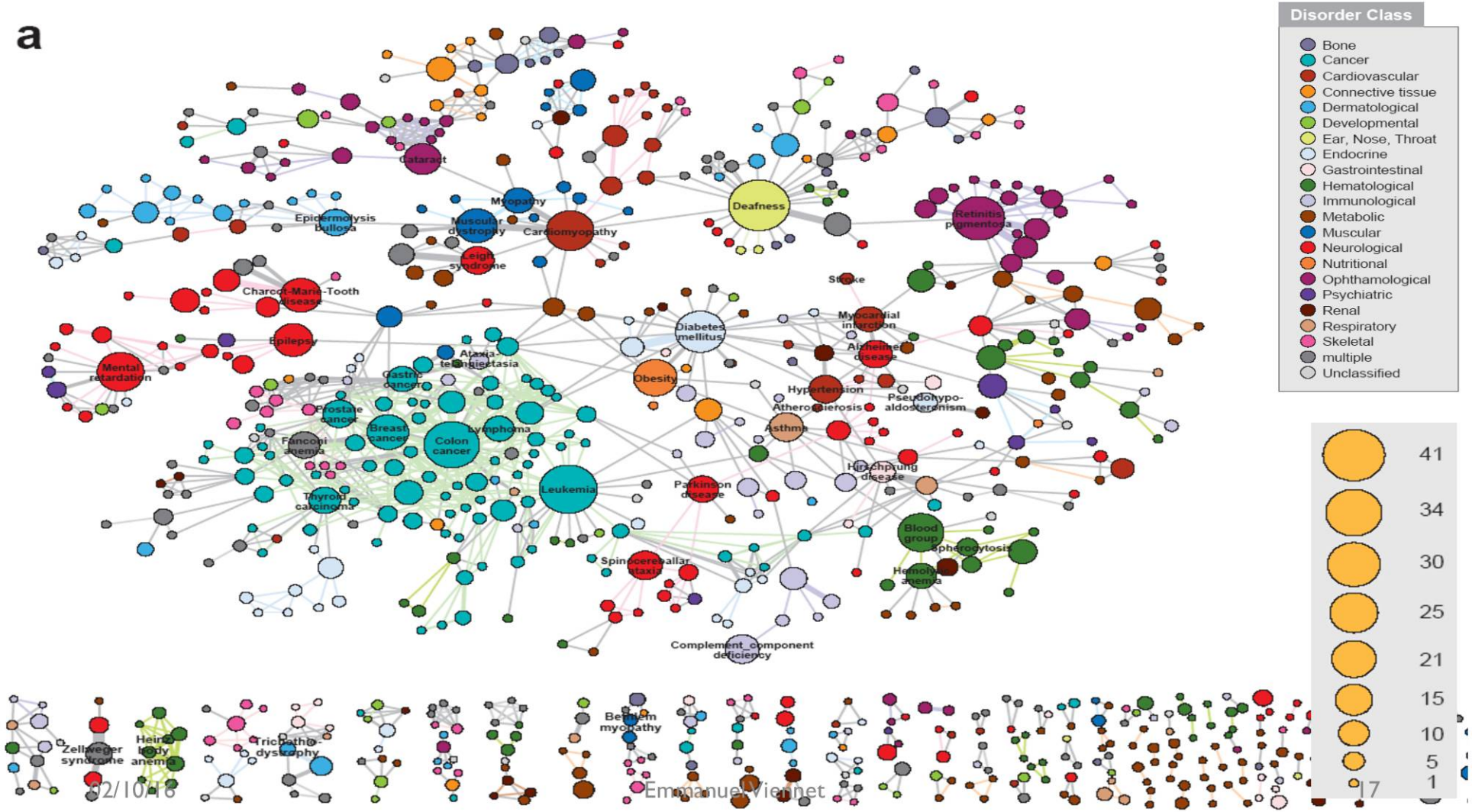






Santé publique

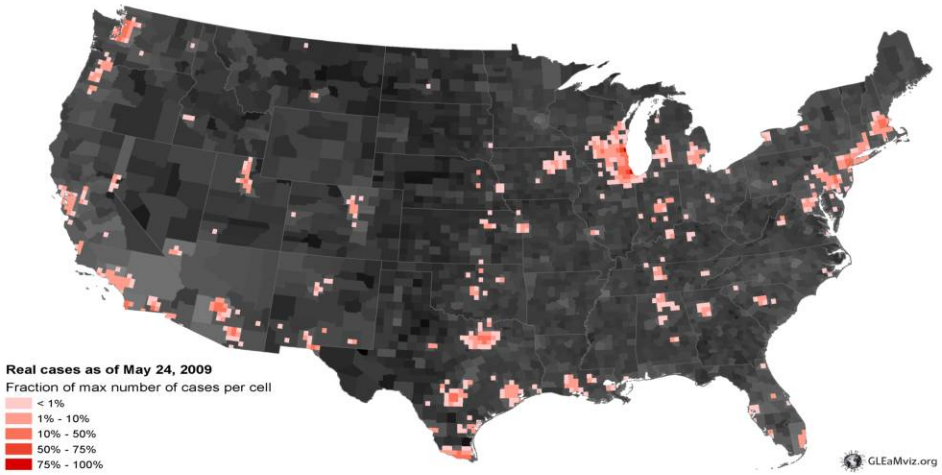
Réseau reliant des maladies



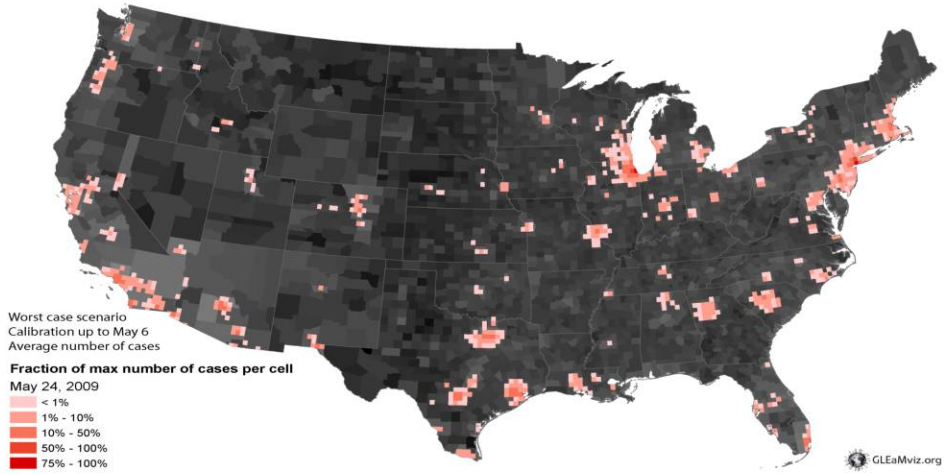
Prévision d'épidémies

H1N1 pandemia

Réel



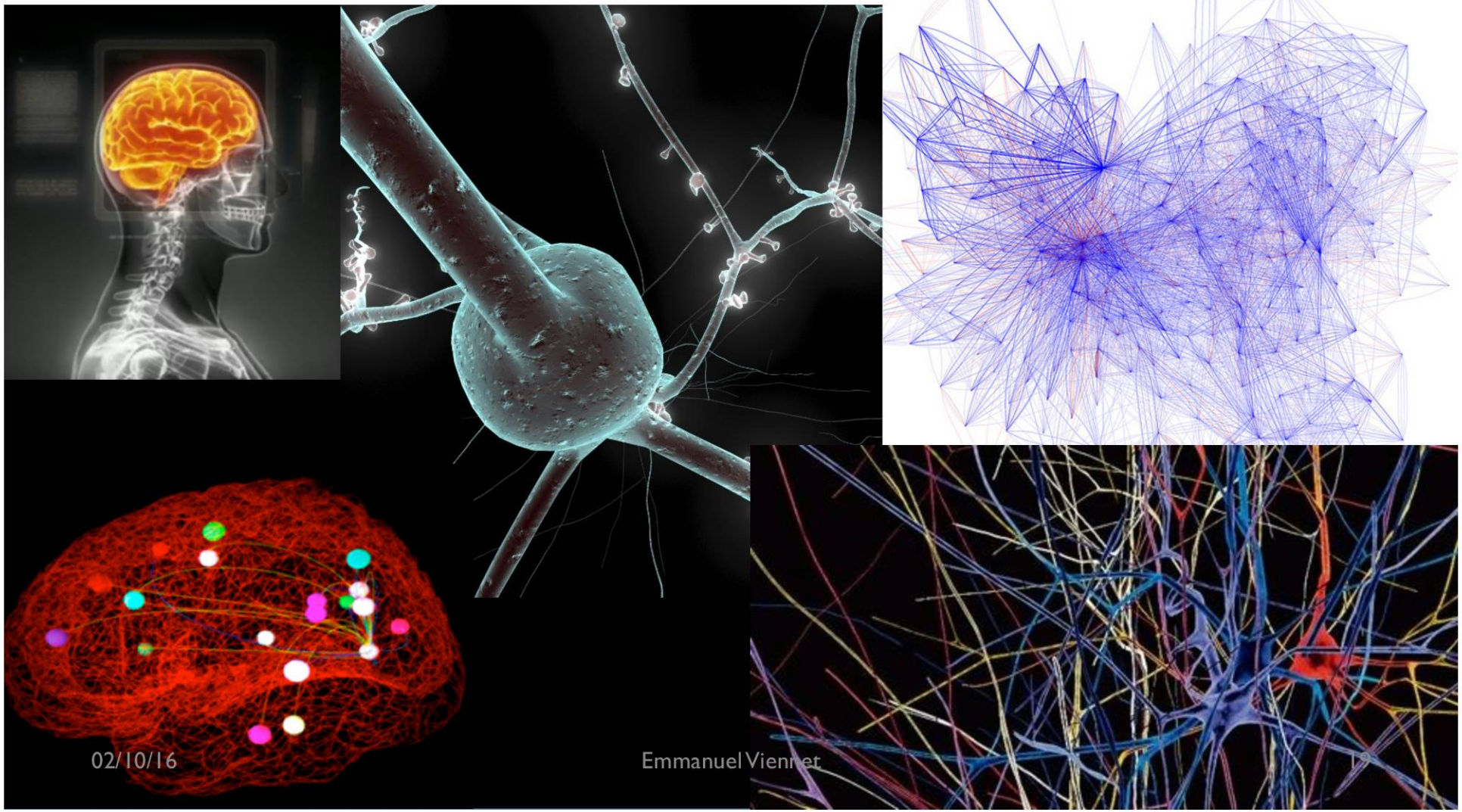
Prévu



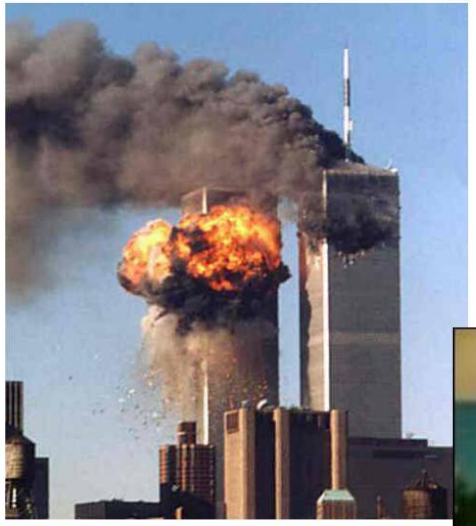
Modélisation du cerveau

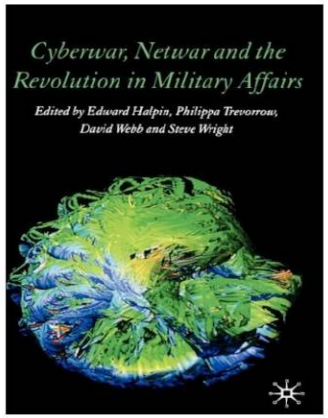
2013 Human Brain Project 10 ans, 1 milliard €

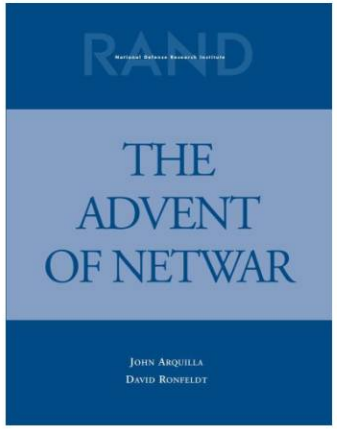
http://www.humanbrainproject.eu/



Contre-terrorisme







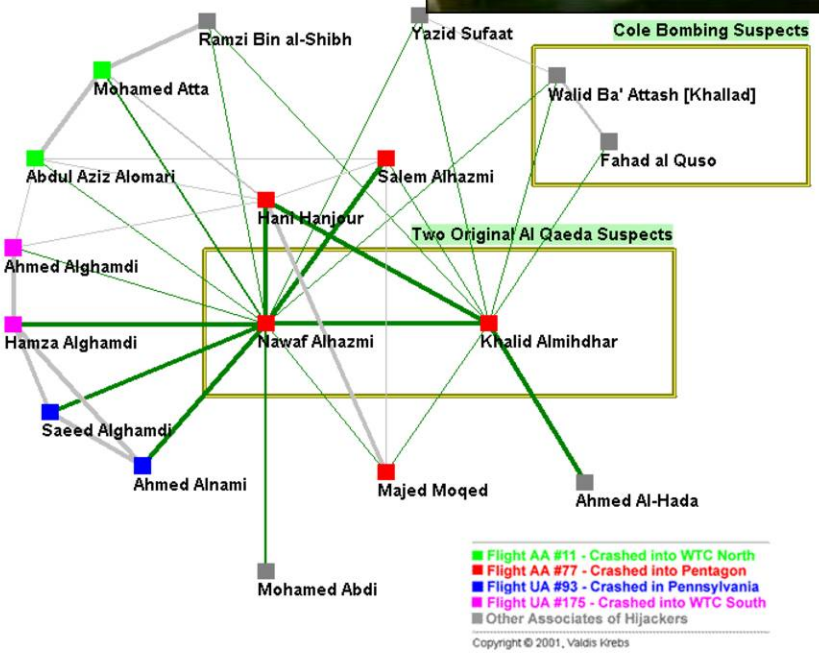
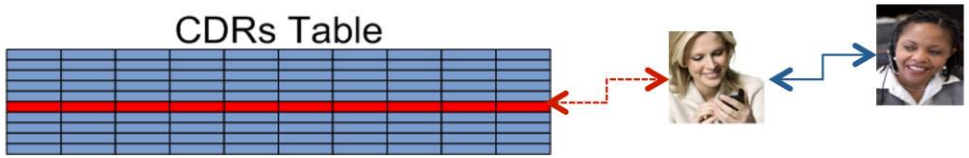


Figure 2 - All nodes within 1 step [direct link] of original suspects

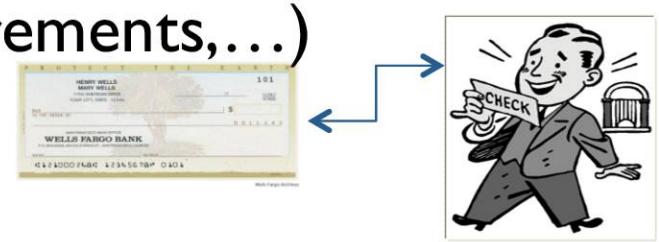
http://www.orgnet.com/prevent.html

Les données en réseau dans l'Industrie

- Télécom
  - données d'appels



- Banque
  - Transferts financiers (chèques, virements,…)
  - Transactions de cartes de crédit

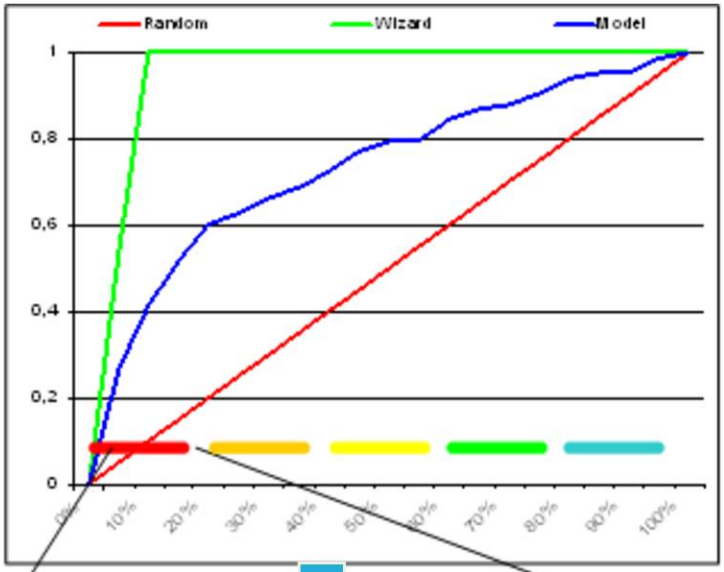


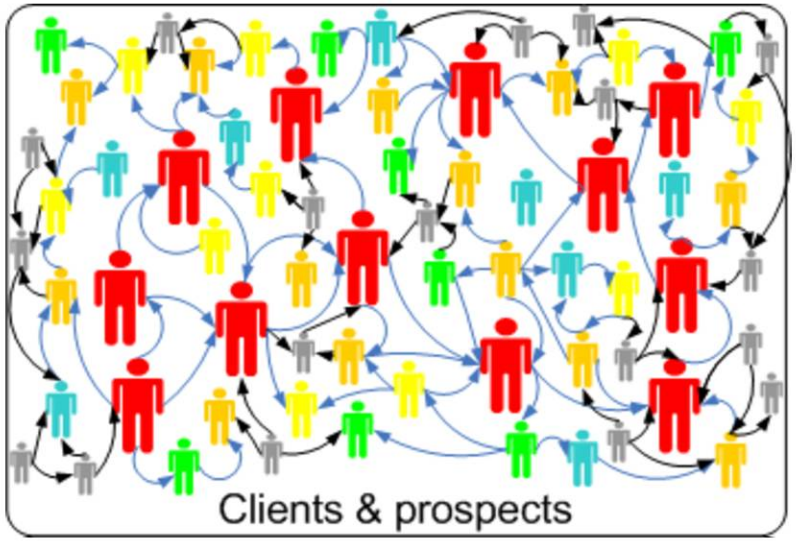
- Sites sociaux, blogs
  - Amis, followers
  - Posts et réponses…

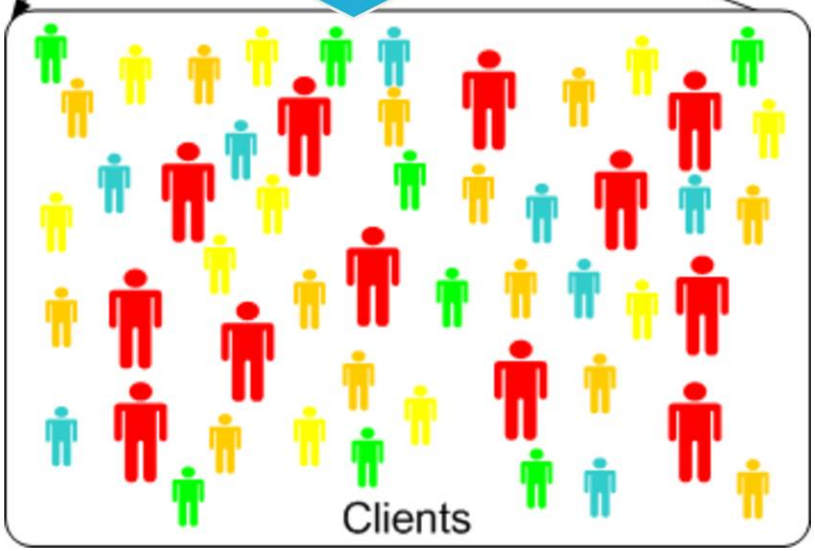
- Distribution
  - Clients achetant les produits

Les campagnes marketing

- Communiquer vers les clients
  - Cible définie par un modèle



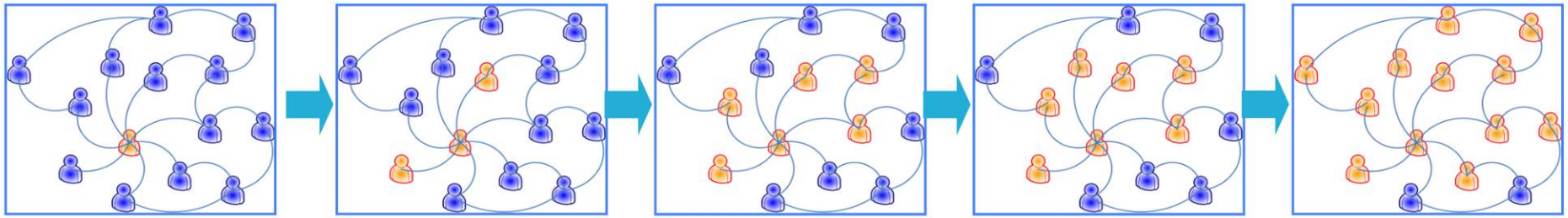




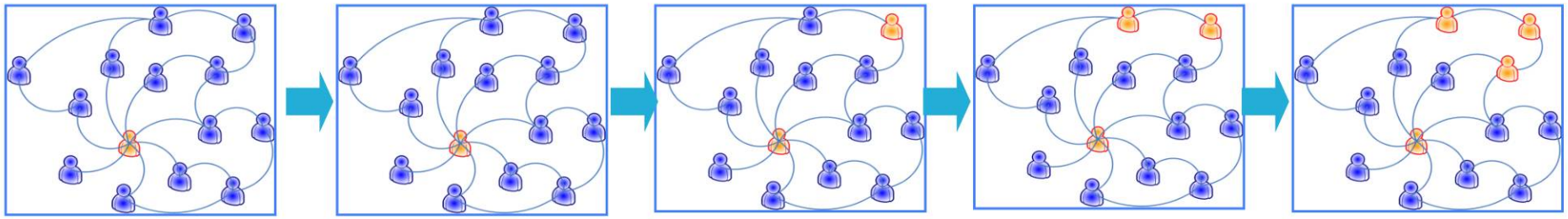
Les campagnes marketing

- **Les clients interagissant, un comportement peut être <u>contagieux</u>**

  Un client peut influencer ses amis : marketing viral



  … ou pas : le marketing viral ne fonctionne pas toujours



- **Réussir une campagne virale nécessite**
  - La compréhension des **rôles des nœuds**
  - La production d'un **modèle de propagation**

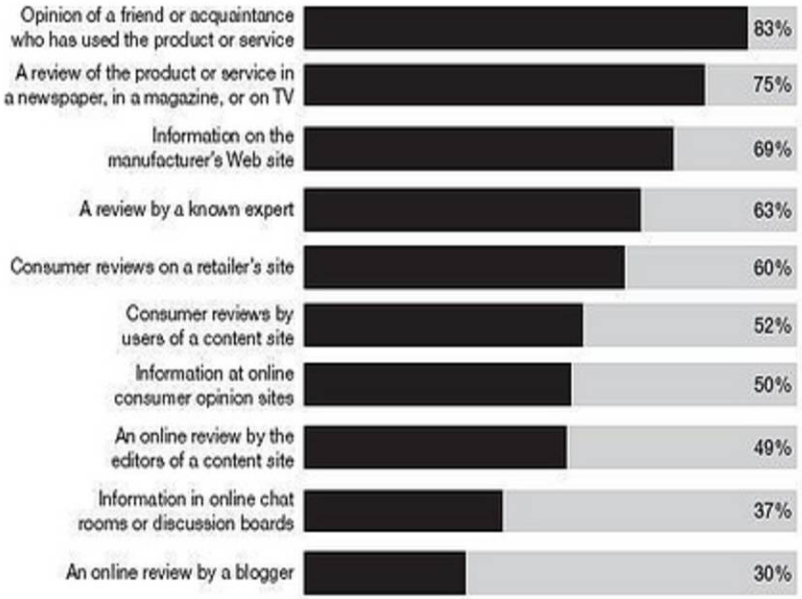
Les campagnes marketing

- **Les clients font confiance à leurs amis**

FIGURE 7-1

How much online North American consumers trust sources of information about products or services

Note that friends' opinions rank higher than reviews in a newspaper, in a magazine, or on TV. Note also that 60% trust reviews on a retailer's site—reviews from people whom they have never met.



Figures include respondents who answered 4 or 5 on a scale of 1 (do not trust) to 5 (trust completely).

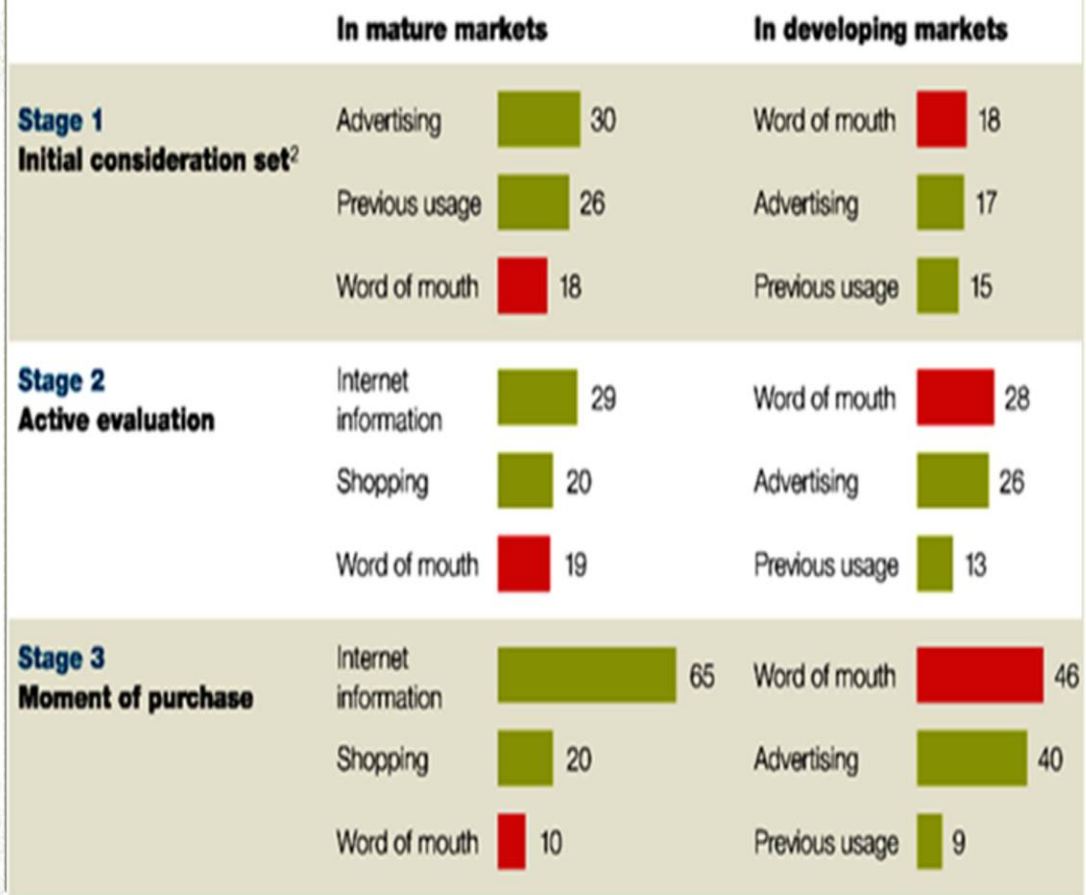
Base: Online North American consumers.

Source: Forrester's NACTAS Q3 2006 Media & Marketing Online Survey.

From *Groundswell: Winning in a World Transformed by Social Technologies* by Charlene Li and Josh Bernoff

Copyright 2008 Forrester Research, Inc.

Top 3 factors that influence whether a product is considered at each stage of the consumer decision journey, mobile-phone example, %[1]



https://www.mckinseyquarterly.com/Retail_Consumer_Goods/Strategy_Analysis/A_new_way_to_measure_word-of-mouth_marketing_2567

Les défis dans l'industrie

- **Les données de contact sont <u>massives</u>**
  - On ne peut pas tirer d'échantillon aléatoire

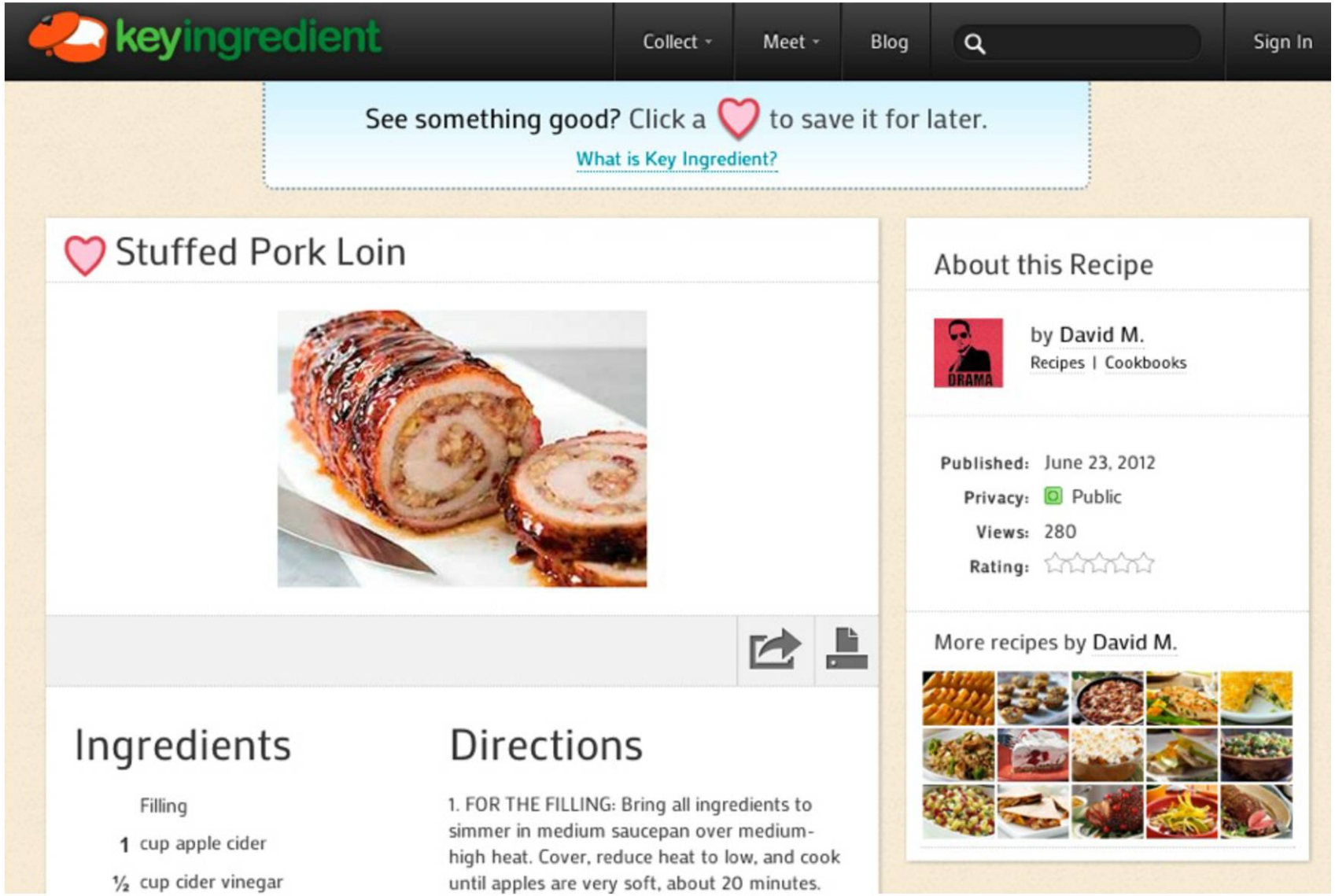
⇨ **Le passage à l'échelle est crucial («*scalabilité*»)**
L'évolution au cours du temps est très rapide

⇨ **Les algorithmes doivent s'exécuter rapidement (rafraichissement)**

- **Exemple en télécom (appels)**

	Lignes (M)	Nœuds (M)
Un jour	150	20
Une semaine	1 100	45
Un mois	4 360	70
Trois mois	13 080	90

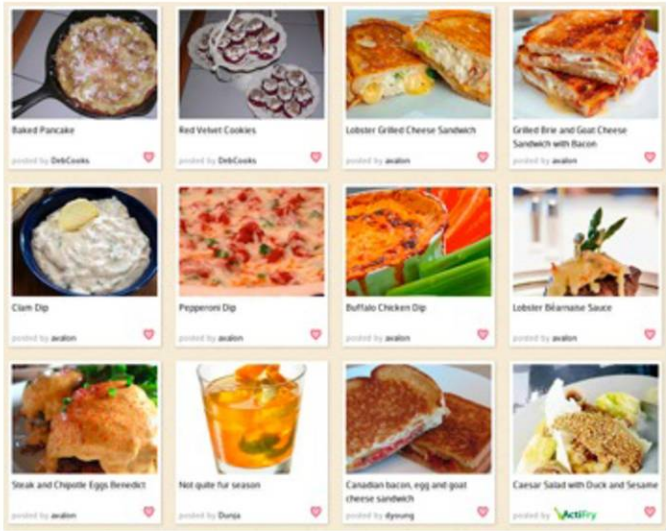
Contexte applicatif: un exemple



Contexte applicatif: un exemple

Utilisateurs

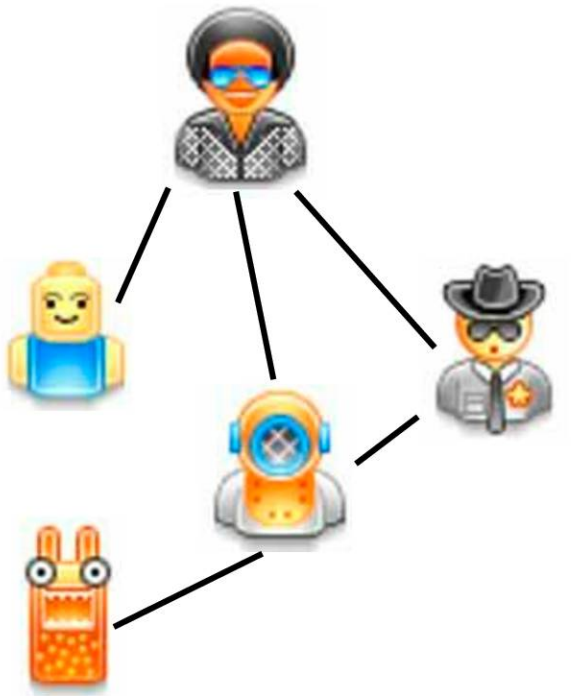
Recettes

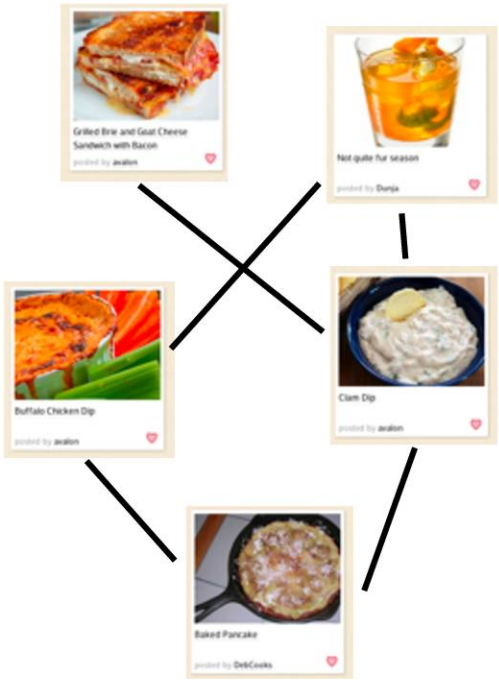


- Blogs (de l'utilisateur ou de la recette)
- Notes
- Tags
- Commentaires sur les recettes
- Similarité entre recettes (texte, image, ingrédients)

Contexte applicatif: un exemple

On peut construire de nombreux graphes...



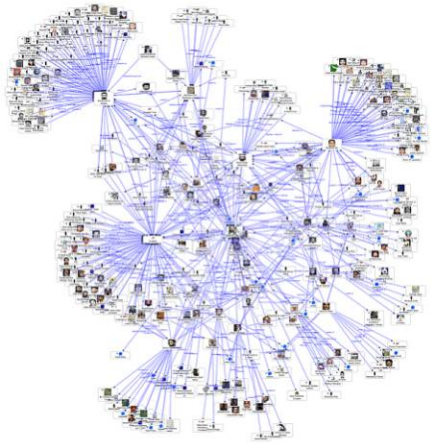


Unipartis ou bipartis, «explicites» ou «implicites»

Quelques questions pertinentes pour les entreprises:

Analyser le comportement des utilisateurs, mieux comprendre les clients

- Analyse de liens
  - Sécurité (finance, renseignement)
  - Fraude (télécoms, finance)
- Analyse de communautés
  - Segmentation
  - Animation de communautés
  - Groupes actifs / évolution



Exploiter le réseau pour diffuser

- Marketing viral
  - Détection d'influenceurs
  - Modèles de diffusion

Modèles prédictifs

- prévision de *churn* (attrition), x-sell/up-sell
- **Recommander**

Exemples de données réelles traitées par notre équipe

Appels téléphoniques	10 -100 M
Blogs	10 M
Achats (e-commerce)	10 M
Sites web	1000

Parcimonie (*sparseness*): nombre de liens *proportionnel* au nombre de nœuds

Jeux de données disponibles en ligne

- la collection de Mark Newman
  `http://www-personal.umich.edu/~mejn/netdata`
- celle de Jure Leskovec : *Stanford Large Network Dataset Collection* (SNAP)
  `http://snap.stanford.edu/data`
- les compétitions Kaggle, qui ne concernent pas seulement des graphes :
  `https://www.kaggle.com/datasets`
- la collection KONECT à l'université de Koblenz :
  `http://konect.uni-koblenz.de/networks`
- et enfin, *The Colorado Index of Complex Networks* (ICON), récemment mis en ligne, qui recense plus de 3100 réseaux, avec un moteur de recherche permettant de choisir les caractéristiques désirées :
  `https://icon.colorado.edu`.

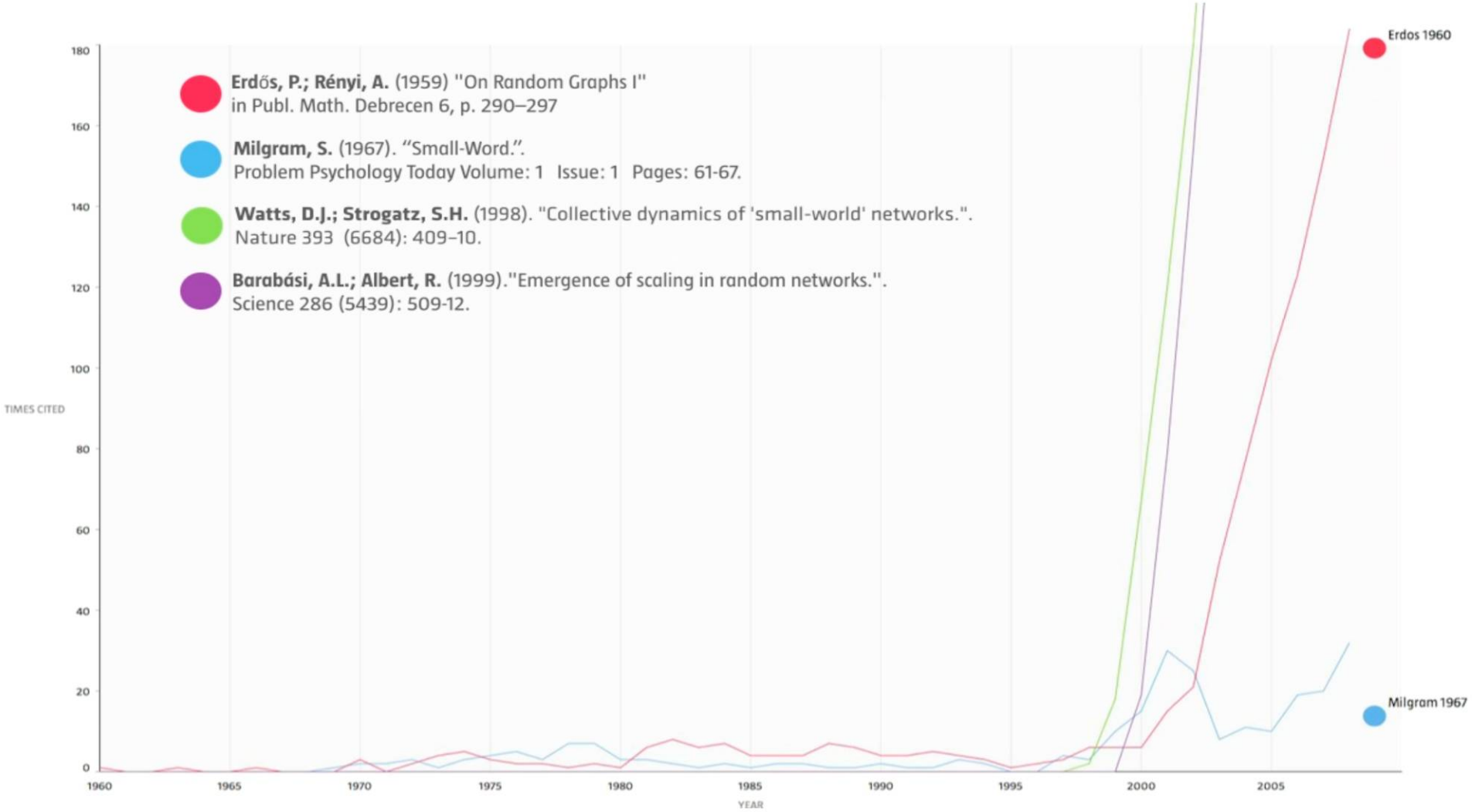
Histoire de l'analyse des réseaux

- 1735 Théorie des graphes (Euler)

- 1930 Réseaux sociaux (Moreno)
- 1950 Erdos-Rény
- 1960-70 Réseaux de télécom., Internet, biologie, écologie
- 1990 Web (Barabasi-Albert)
- 2000 Web Social (« 2.0 »)
  Data mining, algorithmes pour les grands graphes

- Présent:
  réseaux omniprésents,
  applications industrielles

Une science du 21ème siècle



Pourquoi maintenant ?

Données disponibles

C elegans neural wiring diagram 1990
Movie Actor Network, 1998
World Wide Web, 1999
Citation Network, 1998
Metabolic Network, 2000
Réseaux sociaux en ligne (Web 2.0)
Bases de données "big data" avec liens

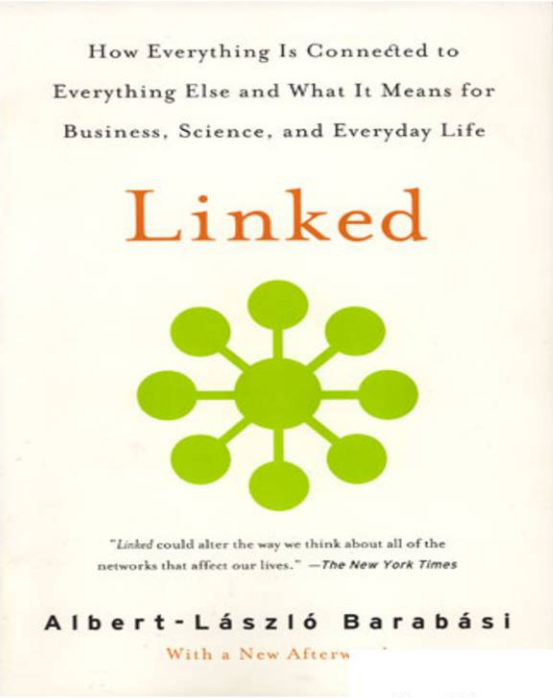
Universalité

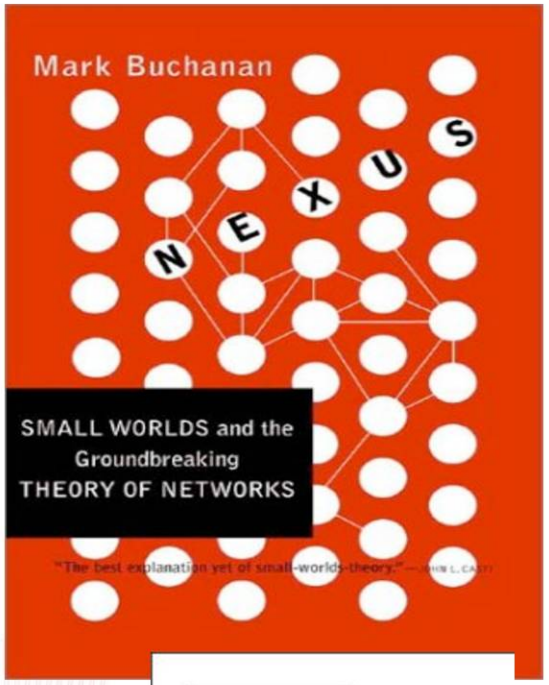
Les réseaux observés dans différents domaines de la science, de la nature et des techniques partagent souvent les même caractéristiques. Pourquoi ?

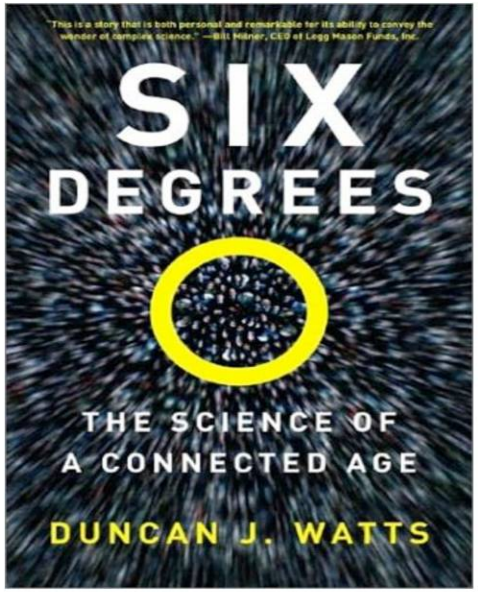
Besoin (urgent) d'appréhender la complexité

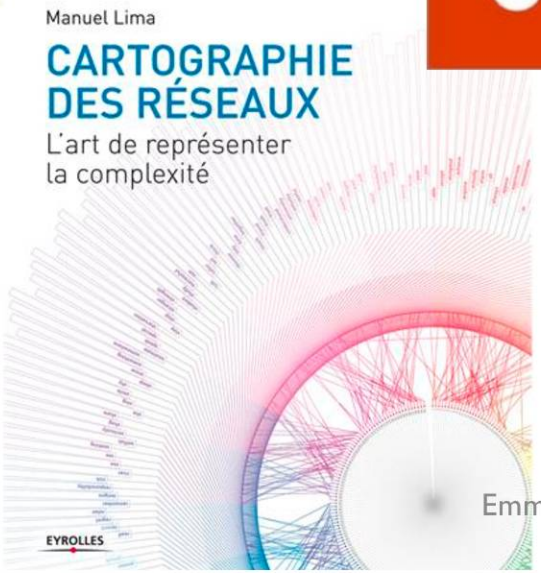
Les réseaux offrent des outils puissants pour modéliser et comprendre les systèmes complexes.

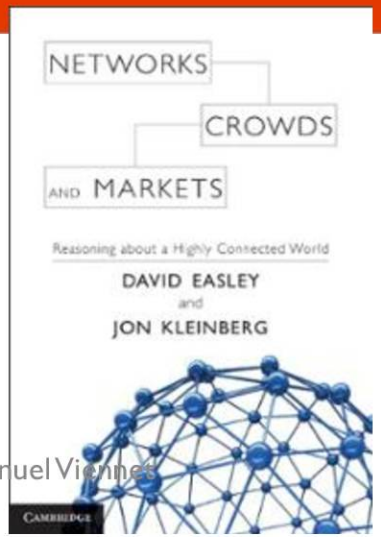
Ouvrages généraux

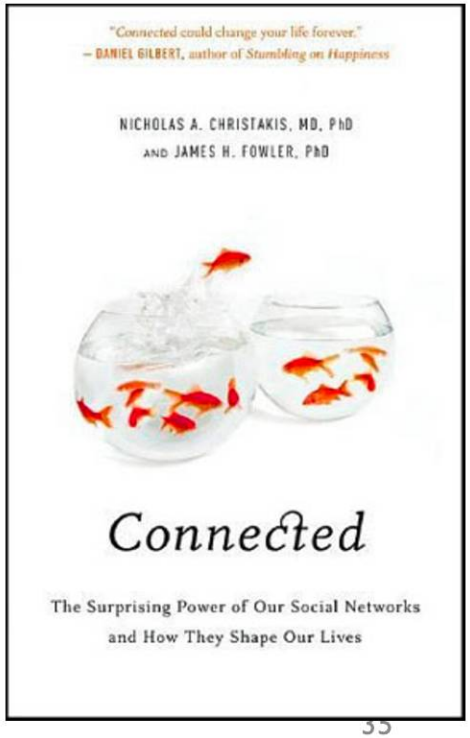




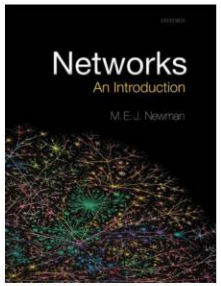




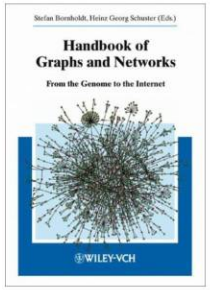




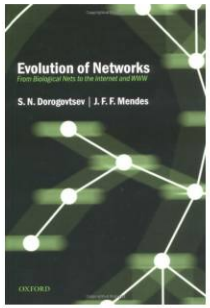
Ouvrages spécialisés (1)



M. Newman. Networks, an introduction. (Oxford University Press, 2010).

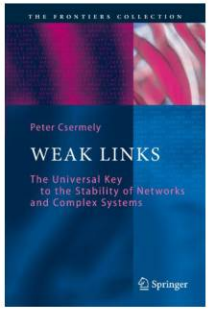


Handbook of Graphs and Networks: From the Genome to the Internet (Wiley-VCH, 2003).

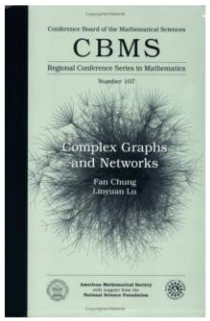


S. N. Dorogovtsev and J. F. F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW (Oxford University Press, 2003).

A.-L. Barabasi: Network Science (8/2016). http://barabasi.com/networksciencebook/

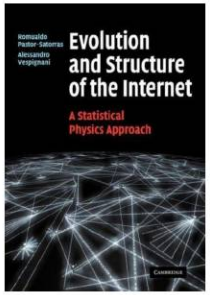


P. Csermely, Weak Links: The Universal Key to the Stability of Networks and Complex Systems (The Frontiers Collection) (Springer, 2006).

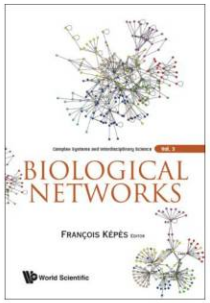


L. L. F. Chung, Complex Graphs and Networks (CBMS Regional Conference Series in Mathematics) (American Mathematical Society, 2006).

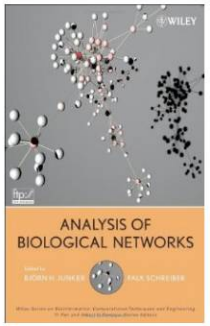
Ouvrages spécialisés (2)



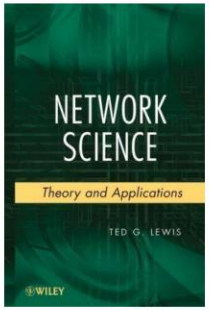
R. Pastor-Satorras, A. Vespignani, Evolution and Structure of the Internet: A Statistical Physics Approach (Cambridge University Press, 2007).



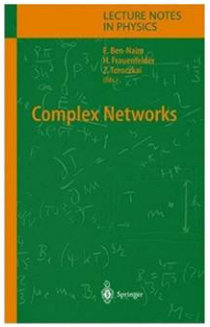
F. Kopos, Biological Networks (Complex Systems and Interdisciplinary Science) (World Scientic Publishing Company, 2007).



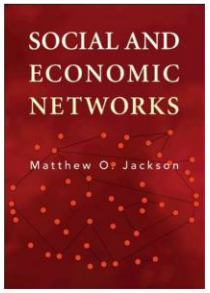
B. H. Junker, F. Schreiber, Analysis of Biological Networks (Wiley Series in Bioinformatics) (Wiley-Interscience, 2008).



T. G. Lewis, Network Science: Theory and Applications (Wiley, 2009).

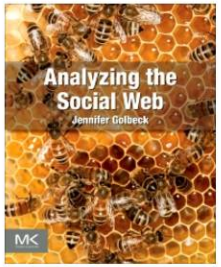


E. Ben Naim, H. Frauenfelder, Z.Torotzai, Complex Networks (Lecture Notes in Physics) (Springer, 2010).

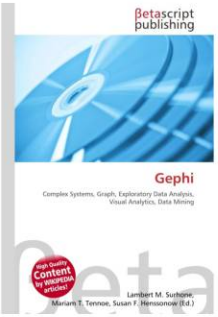


M. O. Jackson, Social and Economic Networks (Princeton University Press, 2010).

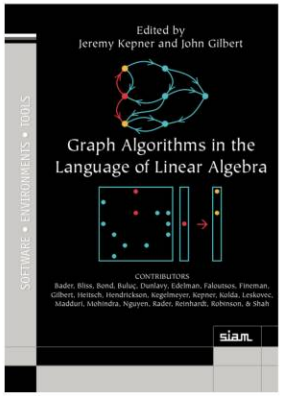
Ouvrages spécialisés (3)



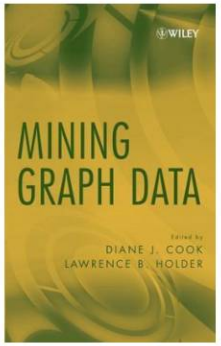
J. Golbeck. Analyzing the Social Web. Morgan Kaufmann, avril 2013.



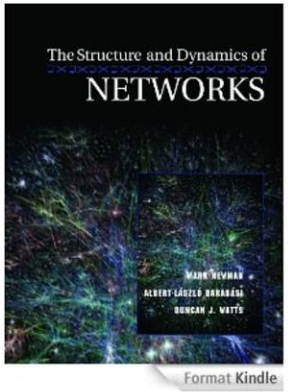
M. Surhone. Complex Systems, Graph, Exploratory Analisis, Visual Analytics, Data Mining (VDM Publishing House, 2010)).



Kepner & Gilbert
Graph Algorithms in the Language of Linear Algebra
SIAM, 2011



D. J. Cook, B. Holder (eds). Mining Graph Data. Wiley, 2006.



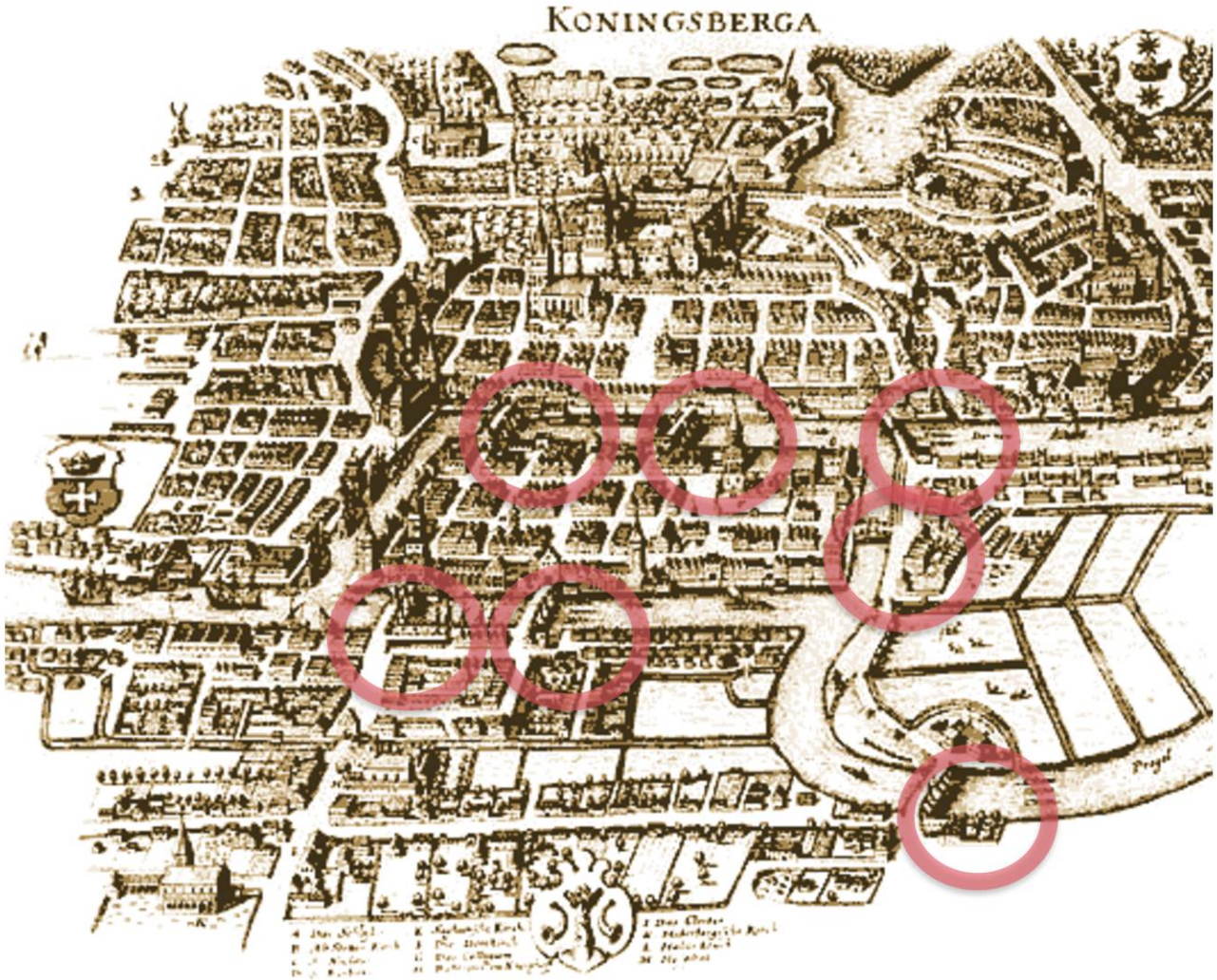
Albert-László Barabási, Mark Newman, Duncan J. Watts
The Structure and Dynamics of Networks
Feb. 2013

Introduction
à l'analyse des réseaux sociaux

Partie 2
Notions sur les graphes

Les 7 ponts de Kœnigsberg

(Kaliningrad)



Traverser chaque pont sans passer deux fois par le même?

rivière Pregel

Pregel: A System for Large-Scale Graph Processing

Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn,
Naty Leiser, and Grzegorz Czajkowski
Google, Inc.
{malewicz,austern,ajcbik,dehnert,ilan,naty,gczaj}@google.com

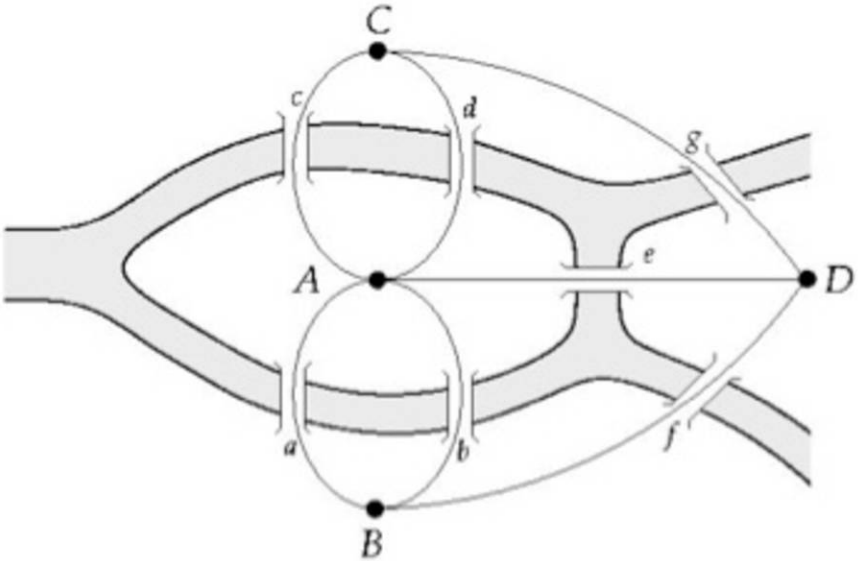
ABSTRACT

Many practical computing problems concern large graphs. Standard examples include the Web graph and various social networks. The scale of these graphs—in some cases billions of vertices, trillions of edges—poses challenges to their efficient processing. In this paper we present a computational model suitable for this task. Programs are expressed as a sequence of iterations, in each of which a vertex can receive messages sent in the previous iteration, send messages to other vertices, and modify its own state and that of its outgoing edges or mutate graph topology. This vertex-centric approach is flexible enough to express a broad set of

disease outbreaks, or citation relationships among published scientific work—have been processed for decades. Frequently applied algorithms include shortest paths computations, different flavors of clustering, and variations on the page rank theme. There are many other graph computing problems of practical value, *e.g.*, minimum cut and connected components.

Efficient processing of large graphs is challenging. Graph algorithms often exhibit poor locality of memory access, very little work per vertex, and a changing degree of parallelism over the course of execution [31, 39]. Distribution over many machines exacerbates the locality issue, and increases the

Les 7 ponts de Königsberg

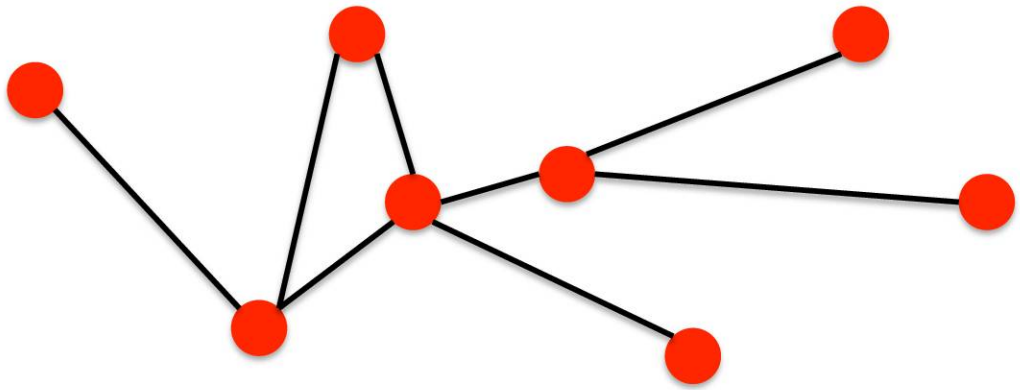


Traverser chaque pont sans passer deux fois par le même?

1735: théorème de **Leonhard Euler:**

Existence chemin si et seulement si présence de 0 ou 2 nœuds de degrés impairs

Graphe



- Nœuds (ou sommets): ensemble V
- Relations: arêtes (ou arcs): ensemble E
- Graphe: (V, E)

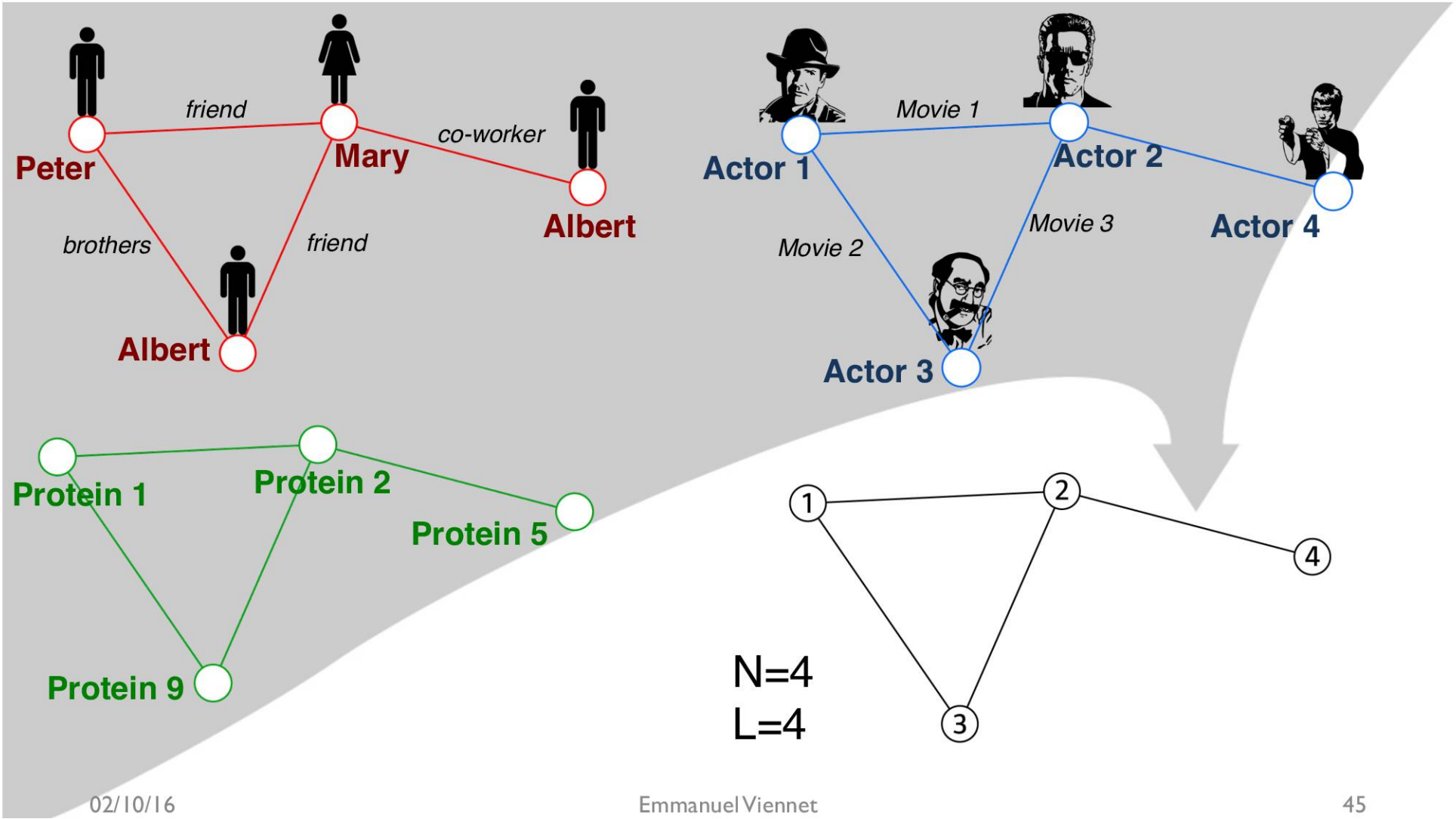
Graphes versus Réseaux

- **Réseau**: système réel
  - Web
  - Réseau Social
  - Réseau métabolique, etc.

- **Graphe**: représentation mathématique du système

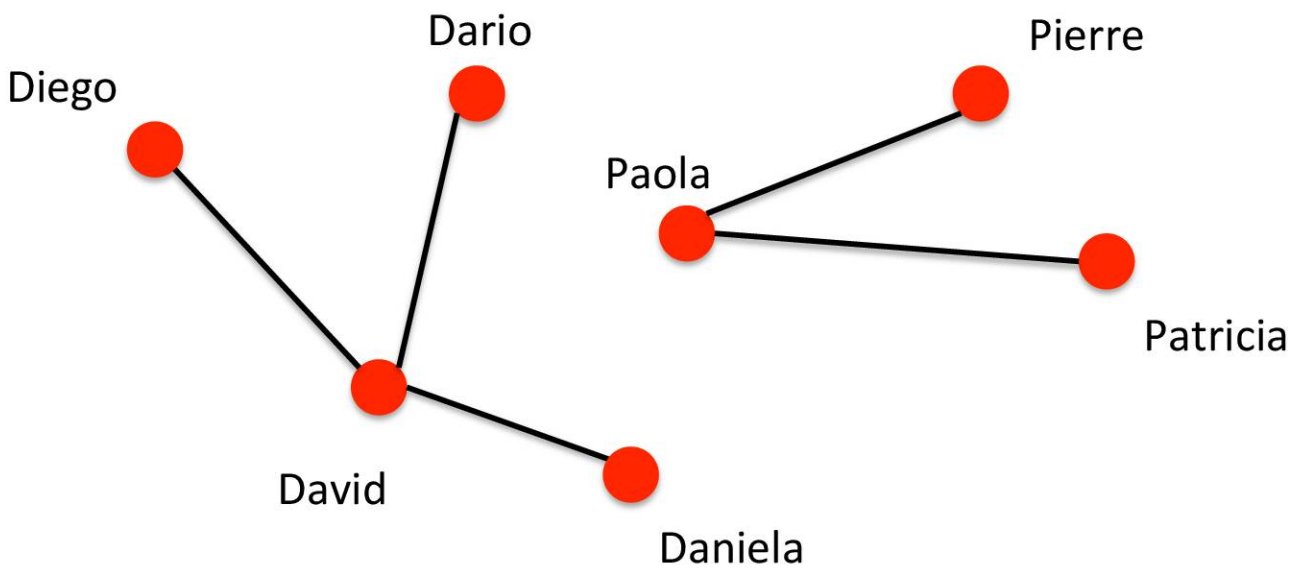
En pratique, ces termes sont souvent interchangeables.

Graphes: un langage unique pour des phénomènes variés



Choisir la bonne représentation

Si nous connectons les personnes dont le prénom commence par la même lettre, on obtient aussi un graphe, mais qui n'a pas de sens…



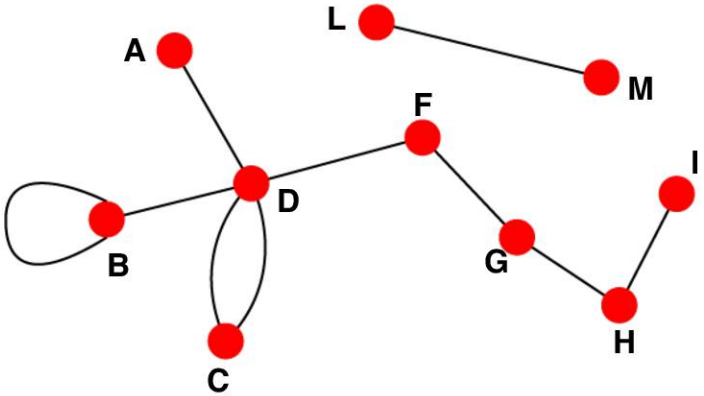
Arêtes (ou arcs, ou liens)

- Orientées ou pas, pondérées ou pas

Non orienté

Liens: symétriques

Graphe:

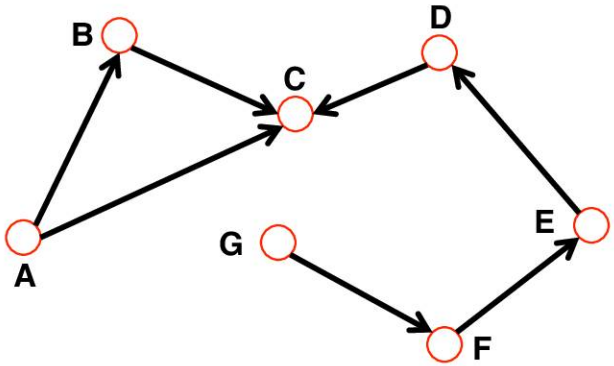


Exemples de liens non orientés:
Interactions (ex: protéines)
Co-auteurs
Projections de graphes bipartites

Dirigé

Liens: dirigés (*arcs*).

Digraph = directed graph:



Exemples de liens dirigés:
URLs du Web
Appels téléphoniques
Réactions métaboliques

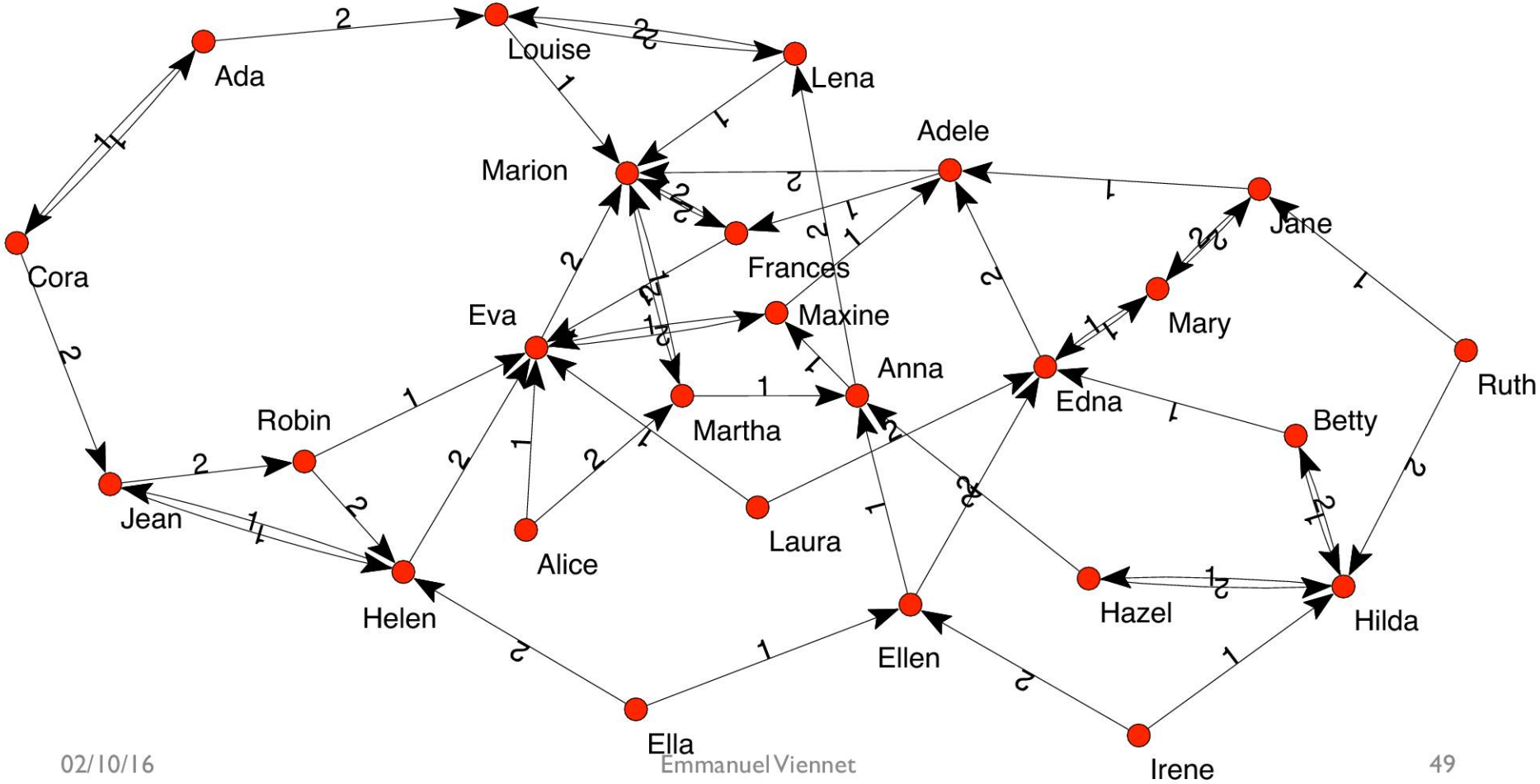
Attributs des arêtes

- Pondération (nombre réel): nombre de message échangés, distance, …
- Type de lien: ami, collègue, …
- Rang (ordre): meilleure amie, seconde amie, …
- Message: texte, données quelconques

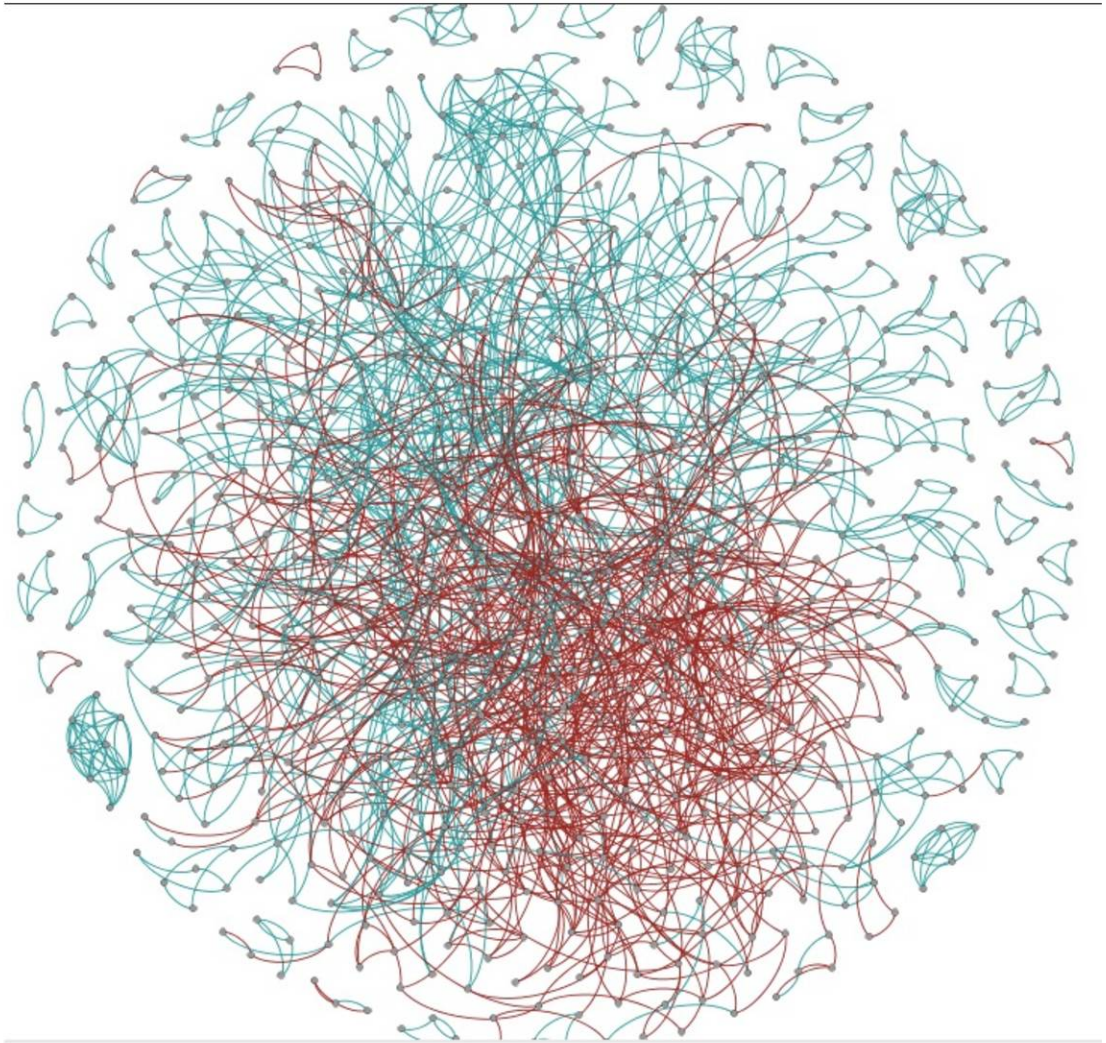
Ou bien: propriété structurelle du graphe, comme l'intermédiarité (voir plus loin).

Graphe dirigé

- girls' school dormitory dining-table partners, 1st and 2nd choices
  (Moreno, *The sociometry reader*, 1960)



Graphe avec poids positifs ou négatifs



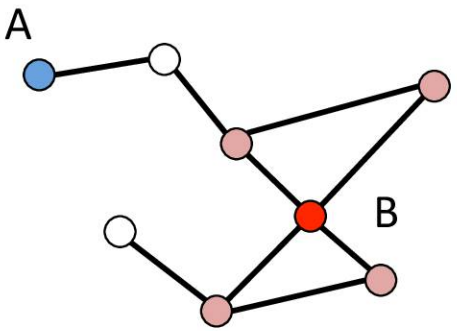
Exemple:

confiance/défiance entre personnes

sample of positive & negative ratings from Epinions network

Degré d'un nœud

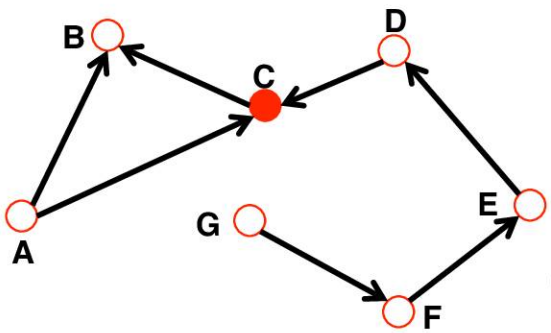
Non dirigé



Degré = nombre d'arcs liés à ce nœud

$$k_A = 1 \qquad k_B = 4$$

Dirigé



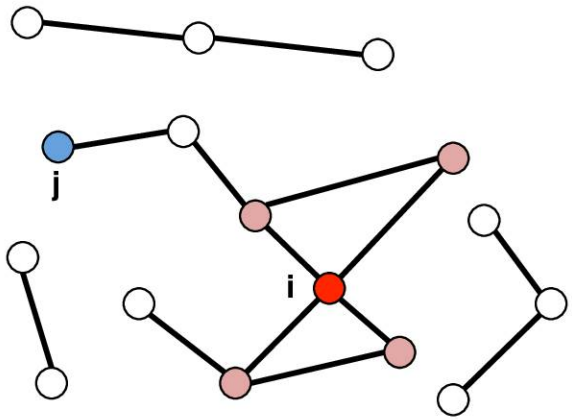
On peut définir le degré entrant (in-degree) et le degré sortant (out-degree). Le degré total est leur somme.

Source: nœud avec $k^{in} = 0$

Collecteur: nœud avec $k^{out} = 0$.

Degré moyen

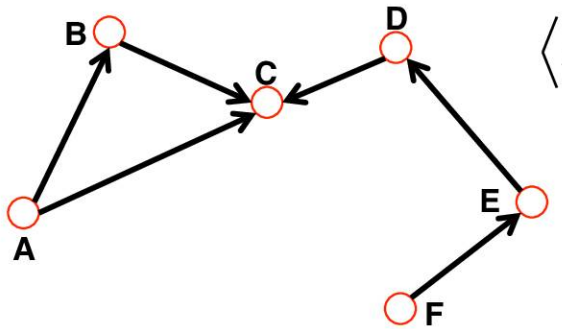
Non dirigé



$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$

N = nombre de nœuds, L = nombre de liens

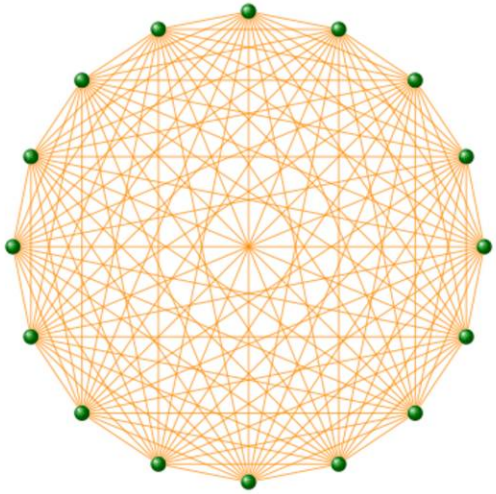
Dirigé



$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

$$\langle k \rangle \equiv \frac{L}{N}$$

Graphe complet



Un graphe de *N* nœuds peut compter au maximum L_{max} arêtes

$$L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$

Un graphe de degré $L=L_{max}$ est dit *complet*,

Son degré moyen est **<**k**>=***N*-1

Les réseaux réels sont parcimonieux (*sparse*)

La plupart des réseaux complexes donnent des graphes avec

$$L \ll L_{max}$$

ou

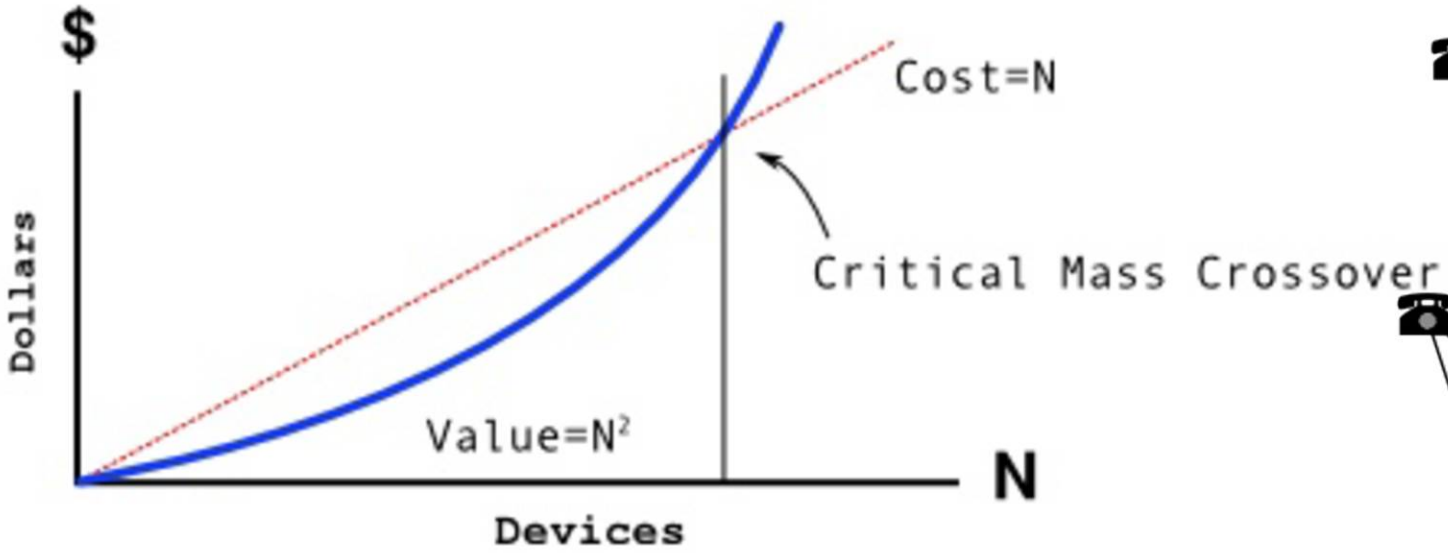
$$\langle k \rangle \ll N-1$$

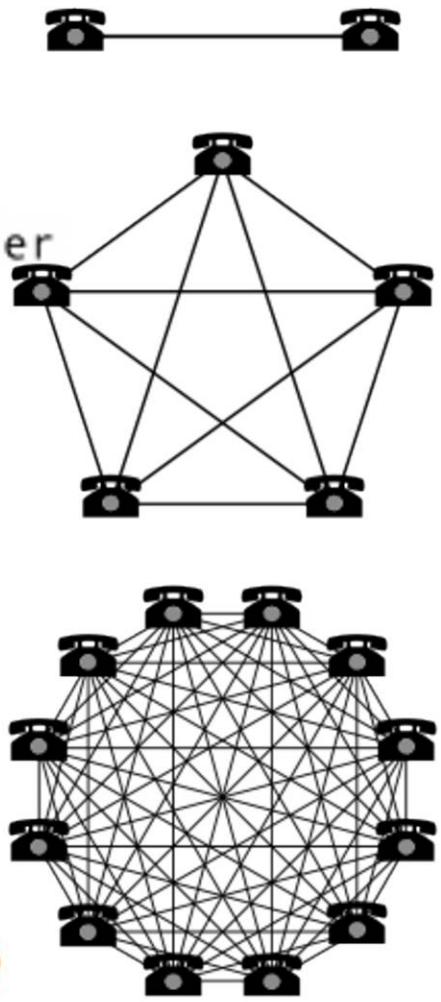
Ce qui revient à dire que le nombre de voisins d'un nœud ne dépend pas de la taille du graphe

WWW (ND Sample):	N=325 729;	L=1.4 10^6	$L_{max}=10^{12}$	<k>=4.51
Protein (*S. Cerevisiae*):	N= 1 870;	L=4 470	$L_{max}=10^7$	<k>=2.39
Coauthorship (Math):	N= 70 975;	L=2 10^5	$L_{max}=3\ 10^{10}$	<k>=3.9
Movie Actors:	N=212 250;	L=6 10^6	$L_{max}=1.8\ 10^{13}$	<k>=28.78

(Source:Albert, Barabasi, RMP2002)

“Loi” de Metcalfe (« valeur » d’un réseau)



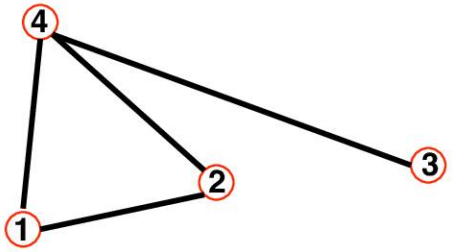


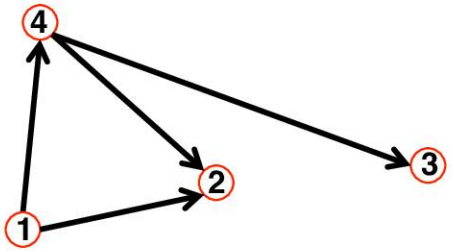
Nombre maximal d’arcs dans un graphe:

$$L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$

Hypothèse fausse, que l’on a liée à la bulle Internet de 2000

Matrice d'ajacence



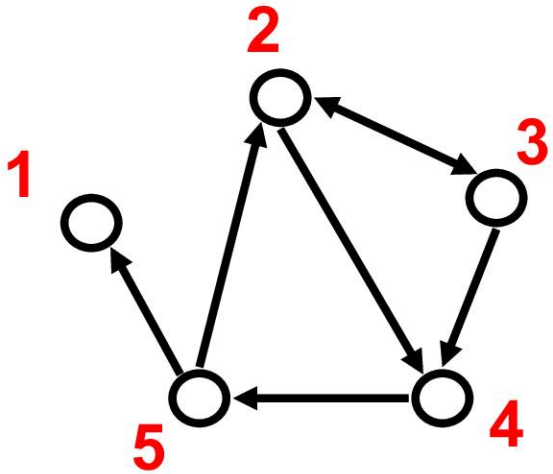


A_{ij}=1 s'il y a un lien entre les nœuds *I* et *j*
A_{ij}=0 si les nœuds *i* et *j* ne sont pas voisins

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Pour le graphe dirigé (à droite), la matrice n'est pas symétrique.

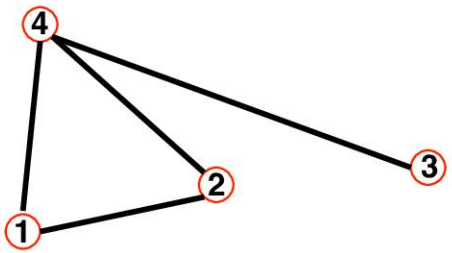
Matrice d'adjacence exemple



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

Matrice d'adjacence et degrés

Non dirigé



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

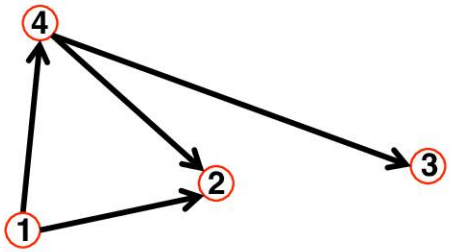
$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

$$k_i = \sum_{j=1}^{N} A_{ij}$$

$$k_j = \sum_{i=1}^{N} A_{ij}$$

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{ij}^{N} A_{ij}$$

Dirigé



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

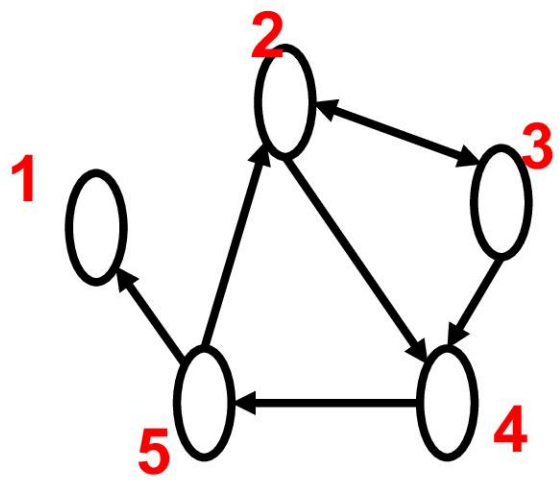
$$k_j^{in} = \sum_{i=1}^{N} A_{ij}$$

$$k_i^{out} = \sum_{i=1}^{N} A_{ij}$$

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j}^{N} A_{ij}$$

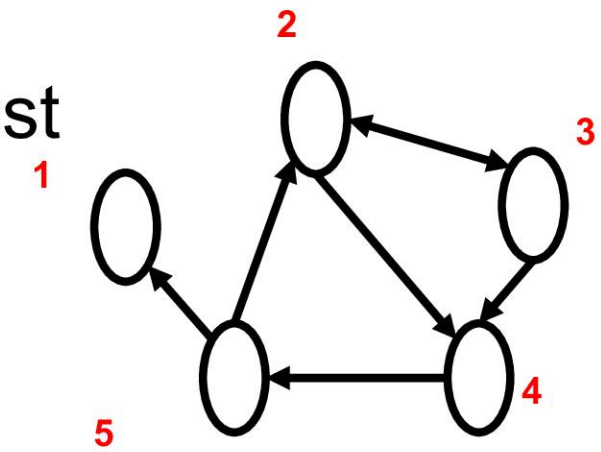
Représentation en liste de liens

- Liste d'adjacence (*edge list*)
  - 2, 3
  - 2, 4
  - 3, 2
  - 3, 4
  - 4, 5
  - 5, 2
  - 5, 1



Représentation en liste de voisins

- Plus commode si le graphe est
  - Grand
  - *Sparse*
- Donne immédiatement la liste des voisins d'un nœud



- 1:
- 2: 3 4
- 3: 2 4
- 4: 5
- 5: 1 2

Introduction
à l'analyse des réseaux sociaux

Partie 3
Outils logiciels

Logiciels
pour l'analyse des réseaux sociaux

NetLogo

Gephi

Tulip

NetworkX

igraph library

Pajek

NODEXL Network Graphs
The Social Media Research Foundation

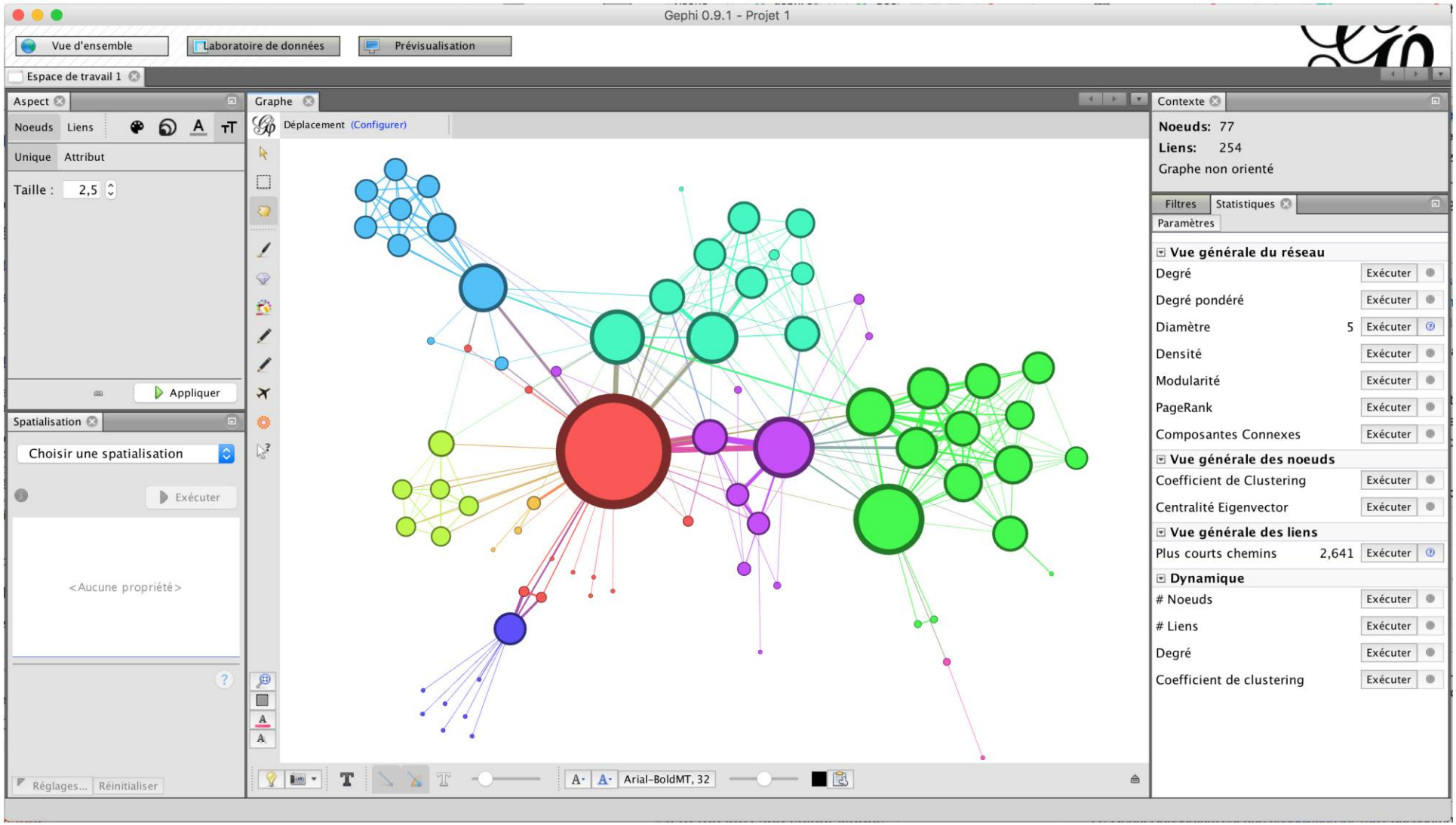
Cytoscape

UCINET

http://gephi.org

- *Open Source*, développé par une équipe française
- Ecrit en Java (NetBeans)
- Linux / Windows / Mac
- Analyse et visualisation de réseaux

Gephi


Gephi 0.9.1 - Projet 1
Vue d'ensemble
Laboratoire de données
Prévisualisation
Espace de travail 1
Aspect
Noeuds
Liens
Unique
Attribut
Taille :
2,5
Appliquer
Spatialisation
Choisir une spatialisation
Exécuter
<Aucune propriété>
Réglages...
Réinitialiser
Graphe
Déplacement (Configurer)
Arial-BoldMT, 32
Contexte
Noeuds: 77
Liens: 254
Graphe non orienté
Filtres
Statistiques
Paramètres
Vue générale du réseau
Degré
Degré pondéré
Diamètre 5
Densité
Modularité
PageRank
Composantes Connexes
Vue générale des noeuds
Coefficient de Clustering
Centralité Eigenvector
Vue générale des liens
Plus courts chemins 2,641
Dynamique
Noeuds
Liens
Degré
Coefficient de clustering
Exécuter

LaBRI

Inria

http://tulip.labri.fr

Tulip

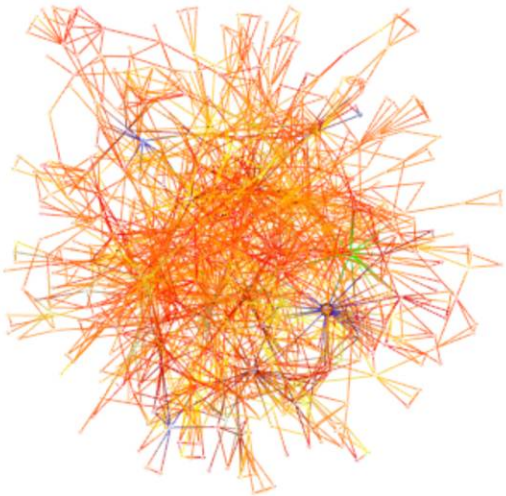
- *Open Source*, développé au LABRI (Université de Bordeaux), INRIA
- Ecrit en C++
- Linux / Windows / Mac
- Analyse et visualisation de réseaux
- Facilite le développement de prototypes, architecture modulaire.

Tulip

File Edit Algorithm Graph View Options Windows Help

unnamed_1 /home/antoine/data/pkrgraphoctober_anonymized.tlp.gz

Node Link Diagram view : /home/antoine



Python Script view : /home/antoine/data/pkrgraphoctober_anon

```
#To cancel the modifications performed by the script
#on the current graph, click on the undo button.

from tulip import *

#the processGraph function must be defined
#to run the script on the current graph
def processGraph(graph) :
    #insert your script code here
    alpha = 100
    viewColor = graph.getColorProperty("viewColor")
    itN = graph.getNodes()
    while itN.hasNext():
        node = itN.next()
        print node.id
        nodeColor = viewColor.getNodeValue(node)
        nodeColor.setA(alpha)
        viewColor.setNodeValue(node, nodeColor)
```

Code python

Check script

Run script

```
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
```

nodes:859, edges:2127

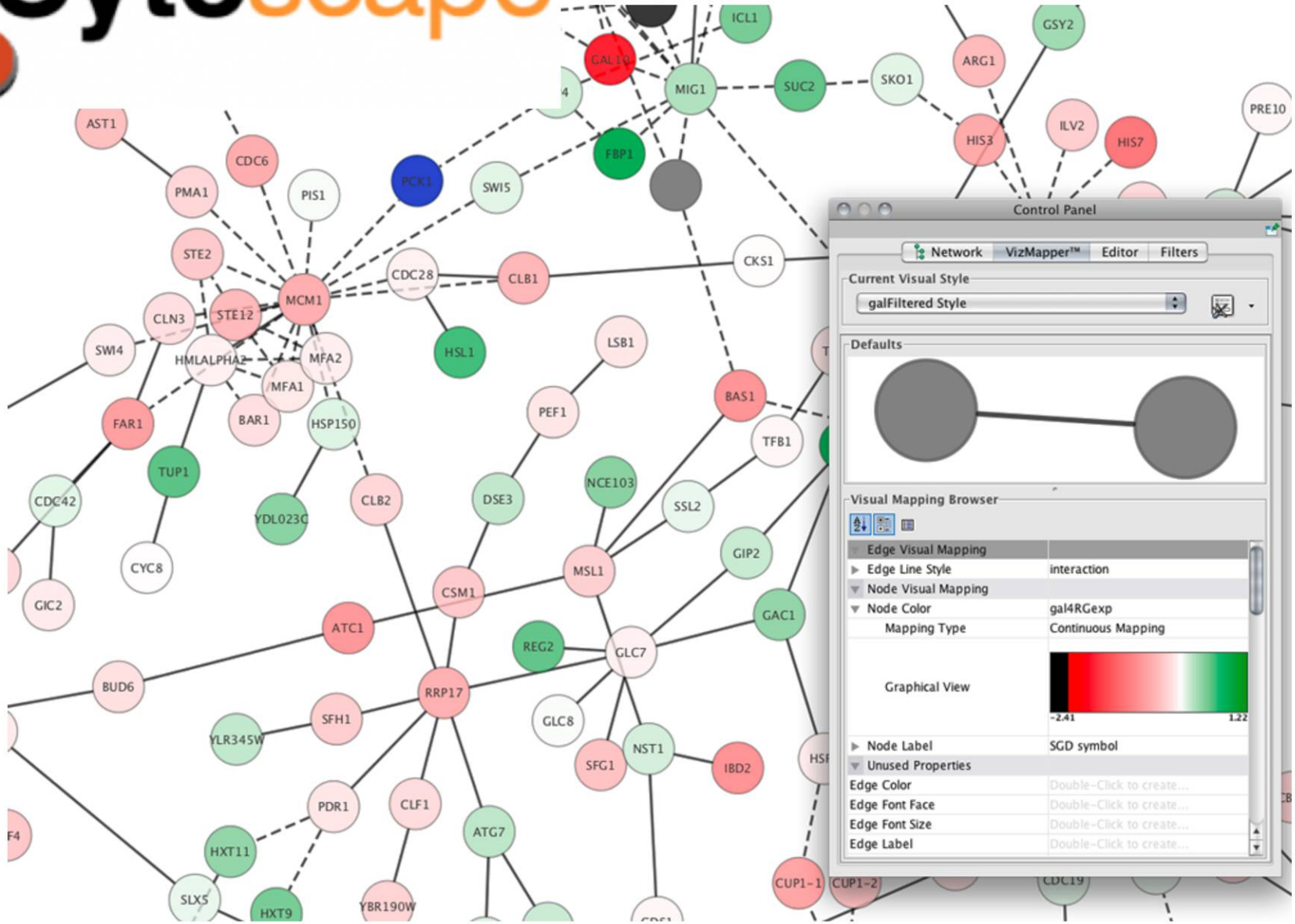
Social Power Chart of Chinese Provinces

area of circles shows mentions in news
width of connection shows mentions of the pair in coverage
data source: Xinhua News
by **xiaoji** / www.xiaoji-chen.com

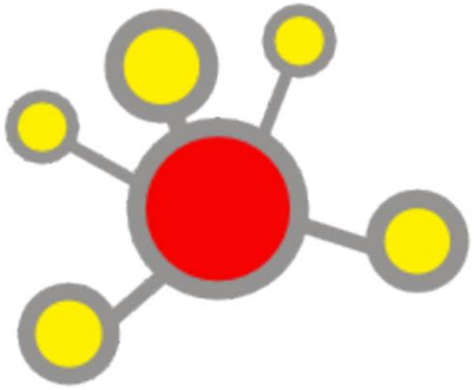
http://www.cytoscape.org

- *Open Source*, développé par l'Institute of Systems Biology (Seattle)
- Ecrit en Java
- Linux / Windows / Mac
- Spécialisé pour les réseaux d'interaction biologiques

Cytoscape


Control Panel
Network
VizMapper™
Editor
Filters
Current Visual Style
galFiltered Style
Defaults
Visual Mapping Browser
Edge Visual Mapping
Edge Line Style
interaction
Node Visual Mapping
Node Color
gal4RGexp
Mapping Type
Continuous Mapping
Graphical View
-2.41
1.22
Node Label
SGD symbol
Unused Properties
Edge Color
Edge Font Face
Edge Font Size
Edge Label
Double-Click to create...
MCM1
MIG1
RRP17
GLC7
MSL1
CSM1
CDC28
CLB1
STE12
STE2
CLN3
SWI4
HMLALPHA2
MFA2
MFA1
BAR1
HSP150
FAR1
TUP1
CDC42
CYC8
GIC2
YDL023C
CLB2
DSE3
PEF1
LSB1
HSL1
PCK1
SWI5
FBP1
CDC6
PMA1
PIS1
AST1
ICL1
GAL10
SUC2
SKO1
ARG1
GSY2
HIS3
ILV2
HIS7
PRE10
CKS1
BAS1
TFB1
NCE103
SSL2
GIP2
GAC1
REG2
ATC1
BUD6
SFH1
YLR345W
GLC8
NST1
SFG1
IBD2
PDR1
CLF1
ATG7
HXT11
SLX5
HXT9
YBR190W
CUP1-1
CUP1-2
CDC19

http://igraph.sourceforge.net/



igraph library

- Bibliothèque *Open Source*
- Langage C
- Linux / Windows / Mac
- Extension de Python ou R

http://networkx.github.io

NetworkX

python

- Module Python en *Open Source*
- Ecrit en Python, parties en C/ Fortran
- Linux / Windows / Mac
- Nombreux algorithmes sur les graphes

Modules pour R

- statnet *(modélisation probabiliste, ERGM, espaces latents)*
- igraph
- bipartite (réseaux écologiques)
- tnet (graphes pondérés)

http://pajek.imfm.si

Pajek

- Facile à utiliser *(menus)*
- Windows seulement
- Gratuit mais pas libre

http://ccl.northwestern.edu/netlogo/

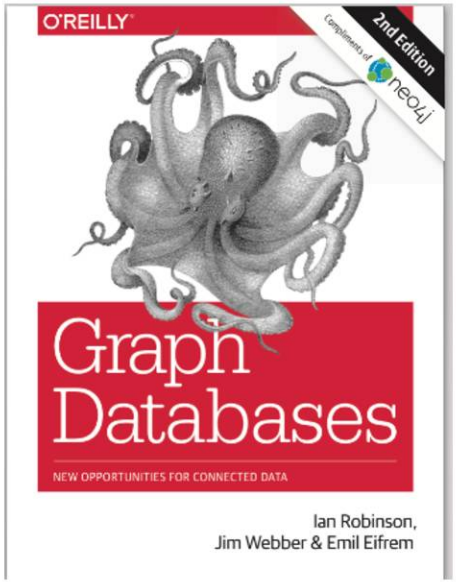
- *Open Source*, développé par CCL (Northwestern University)
- Linux / Windows / Mac (Java)
- Environnement de simulation multi-agent (pour la pédagogie)

Bases graphes

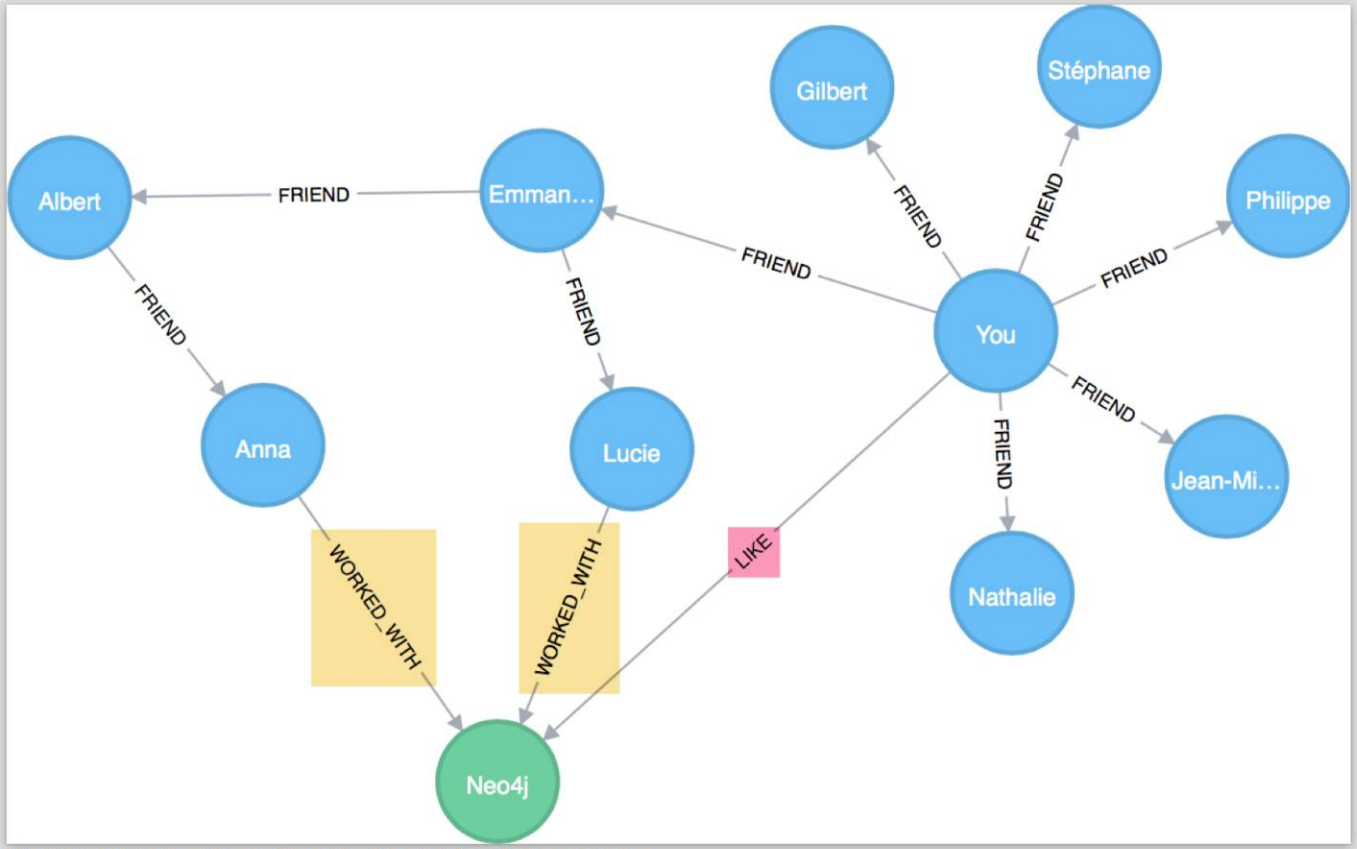
Données: bases relationnelles SQL
+ accès uniforme, outils rodés
- peu adapté aux données massives hétérogènes
- pas naturel de gérer des données liées (graphes)

Bases NoSQL (MongoDB, Reddis, Cassandra, Hbase)
+ rapides, adaptés aux très gros volumes (et plus)
- pas naturel de gérer des données liées (graphes)

Bases graphes (Neo4j, HyperGraphDB, InfoGrid)
+ rapides, adaptés aux très gros volumes
- Langage de requête adapté aux graphes



La base graphe Neo4J (http://neo4j.com) permet de déclarer des classes de nœuds porteurs d'attributs, et des liens modélisant des relations ou interactions. Les liens et les nœuds sont typés et associés à des attributs.



Voici comment trouver dans ce réseau social vos amis :

```
MATCH (you {name:"You"})-[:FRIEND]->(yourFriends)
RETURN you, yourFriends
```

Ou, plus compliqué, trouver les chemins les plus courts dans le graphe vers des amis qui aient travaillé avec l'outil Neo4j :

```
MATCH (you {name:"You"})
MATCH (expert)-[:WORKED_WITH]->(db:Database {name:"Neo4j"})
MATCH path = shortestPath( (you)-[:FRIEND*..5]-(expert) )
RETURN db,expert,path
```

Introduction
à l'analyse des réseaux sociaux

Partie 4

Caractéristiques des réseaux complexes

Caractéristiques de base des graphes (issus de réseaux complexes)

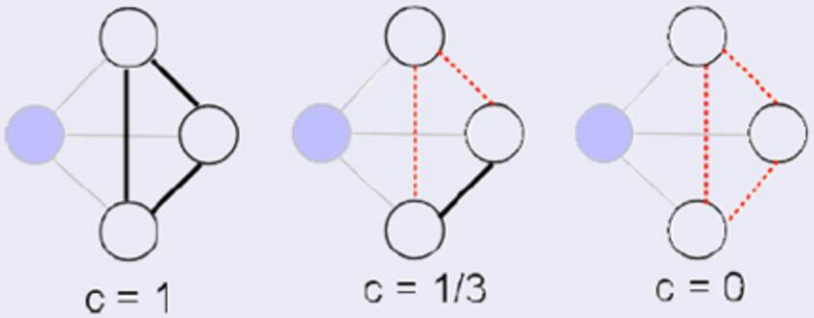
- Coefficient de clustering
- Petit monde
- Distribution des degrés
- Composantes connexes

Coefficient de clustering

Combien de mes amis sont aussi amis entre eux ?

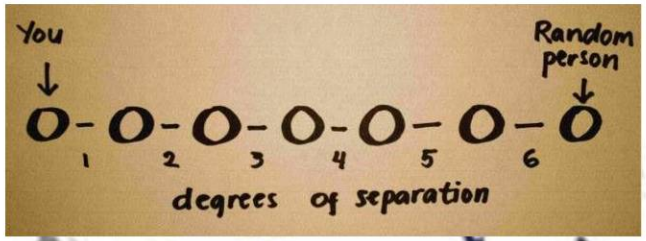
Coefficient de clustering

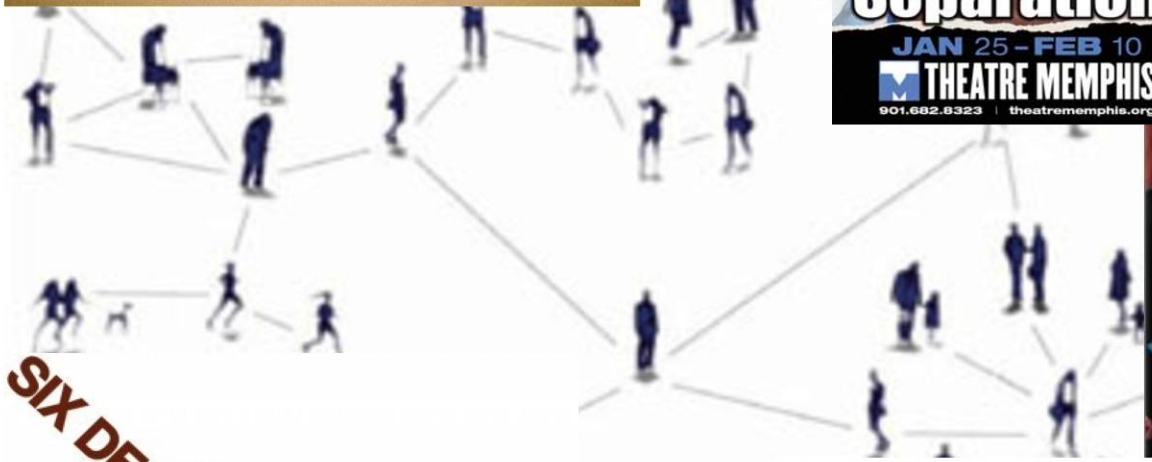
Lié au nombre de voisins d'un nœud qui sont eux mêmes reliés entre eux (triangles) (Watts et Strogatz, 1998)



Effet « petit monde »

Les 6 degrés de séparation





E-mail Study Corroborates Six Degrees of Separation

By Dan Cho

ces are, you don¿t personally know any
ralian policemen, Estonian archival
ctors or Norwegian army
inarians. But you could probably get in
with one of these distant individuals
gh a friend, or a friend of a friend, or a
d of your friend¿s friend. The notion
very person on the planet is separated

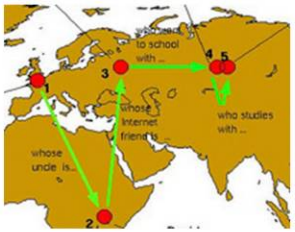
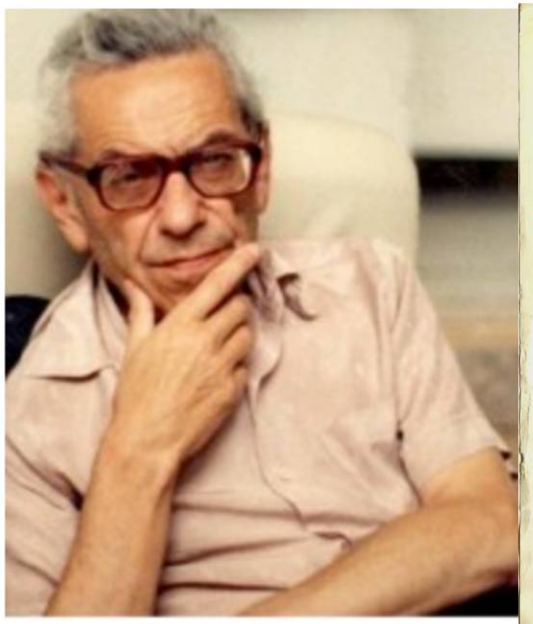
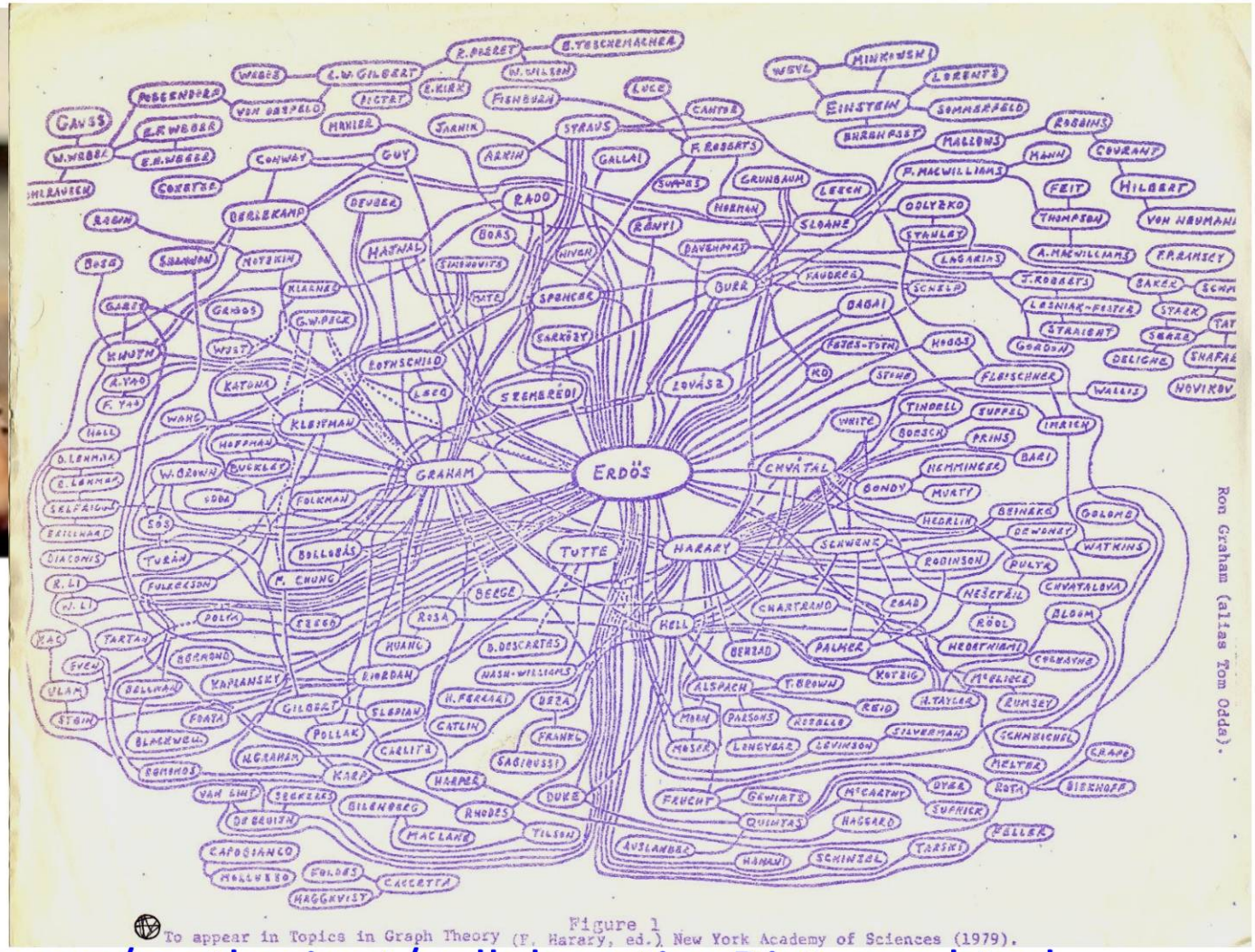


Image: COURTESY OF DUNCAN J. WATTS

Le nombre d'Erdös

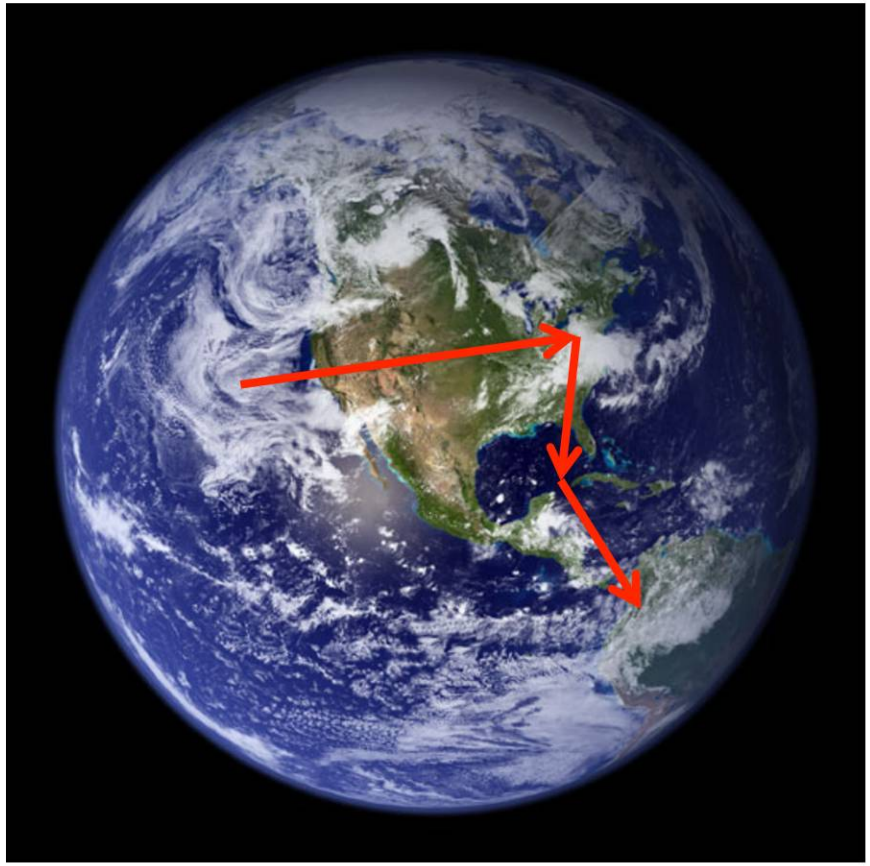




http://www.ams.org/mathscinet/collaborationDistance.html

Expérience de Milgram

Milgram 1960: courrier postal



Avec e-mail : 2003
Dodds, Muhamad, Watts,
Science 301, (2003)

•18 destinations
•13 pays différents

• Plus de 60 000 participants
• 24 163 chaînes de messages
(384 atteignirent leur but)

=> **Longueur chemin moyen: 4.0**

De nombreux réseaux complexes observés sont de "petits mondes"

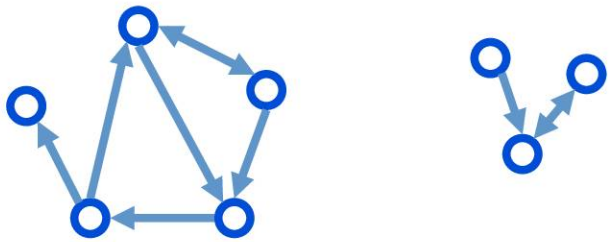
Longueur moyenne du plus court chemin

$$l_{\text{network}} \approx \ln(N)$$

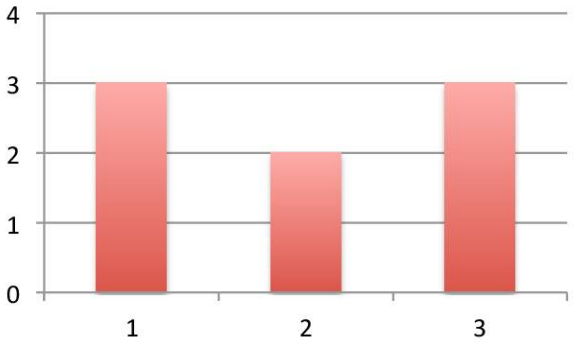
- réseau neuronal du vers C. elegans,
- réseaux sémantiques en langage naturel,
- graphe de collaborations d'acteurs
- réseaux d'ingrédients, etc.

Distribution des degrés

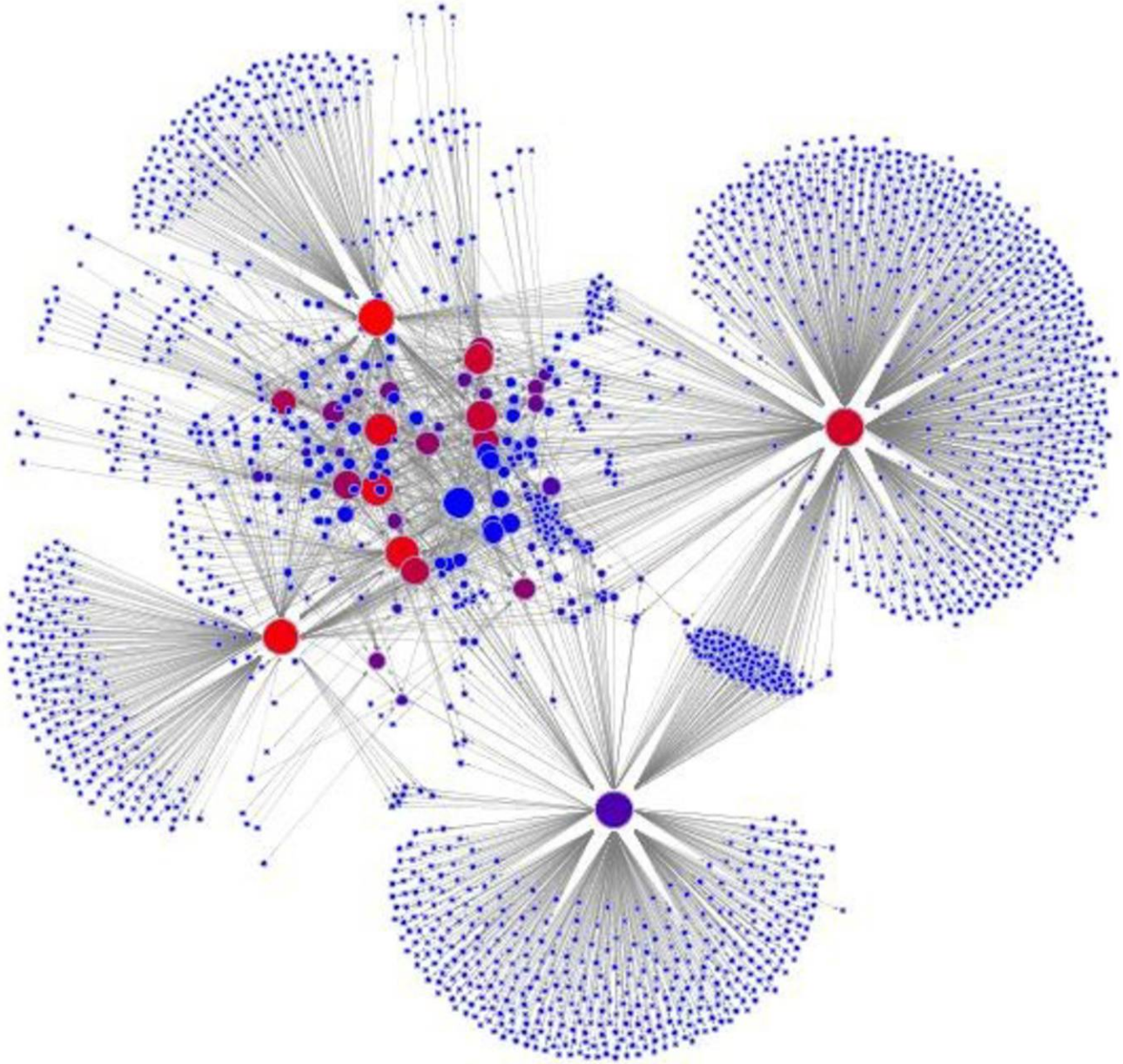
- Liste des degrés de chaque nœud, triée par ordre décroissant
  - [3, 3, 3, 2, 2, 1, 1, 1]



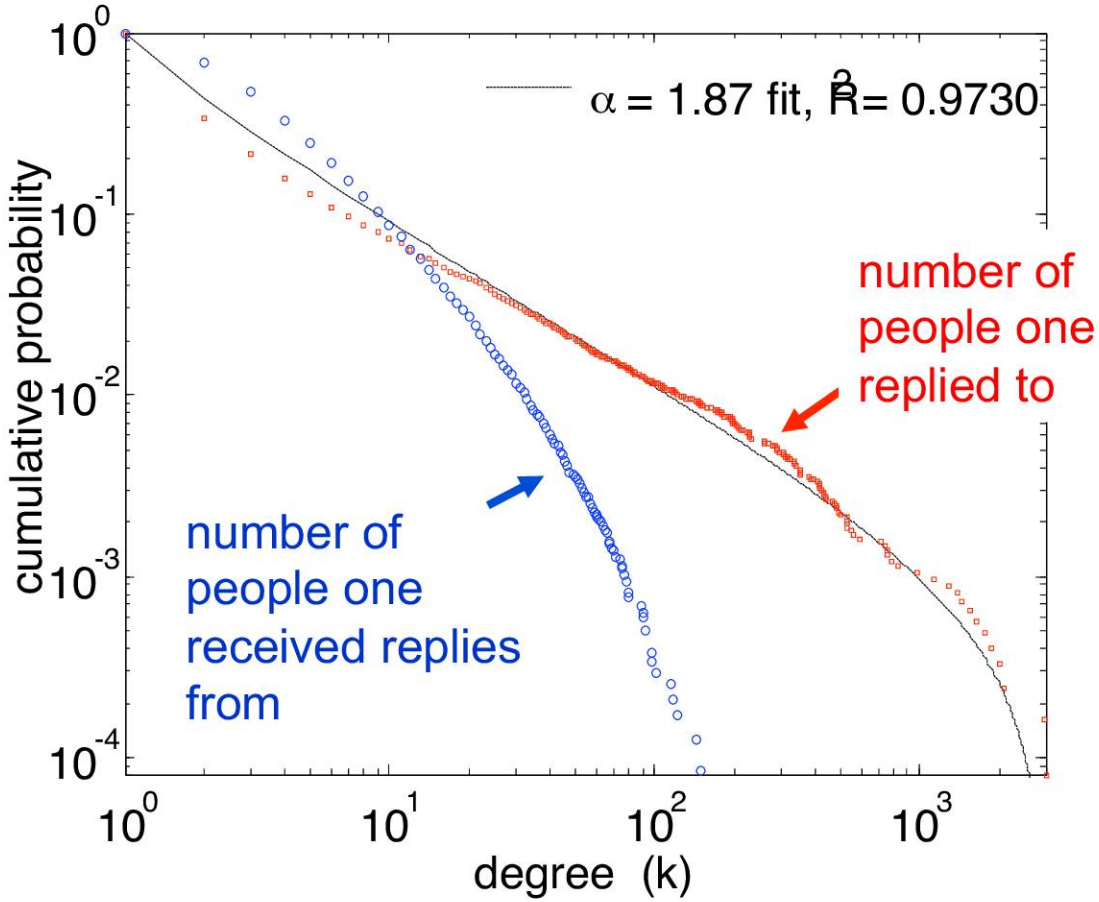
- Distribution des degrés: fréquence de chaque valeur
  - [(3,3) (2,2) (1,3)]



Exemple: forum de discussion



Distribution des degrés dans les réseaux complexes



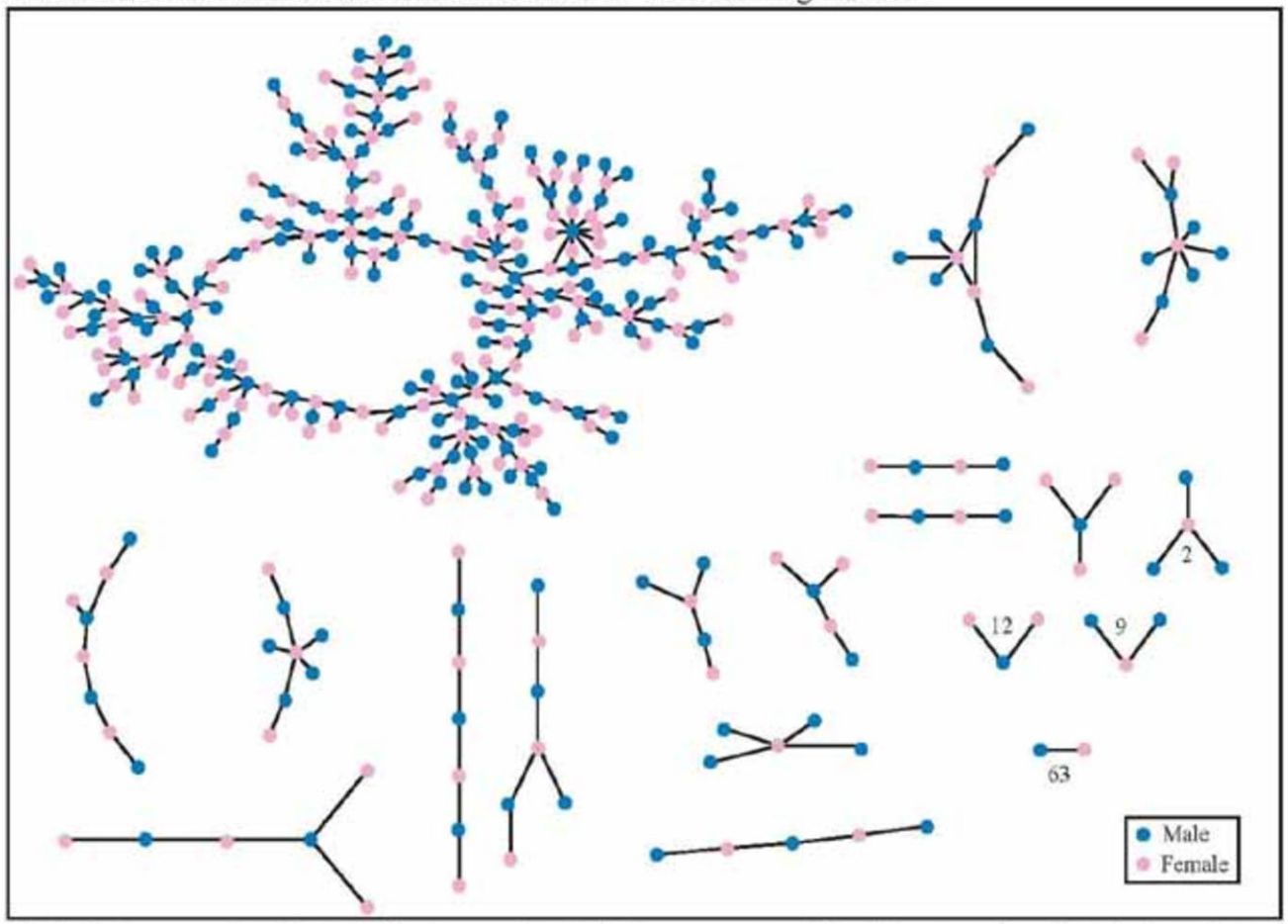
– 'answer people' may reply to thousands of others

– 'question people' are also uneven in the number of repliers to their posts, but to a lesser extent

=> participation très variable

Distribution des degrés
Exemple 2: "réseau amoureux"

The Structure of Romantic and Sexual Relations at "Jefferson High School"



Each circle represents a student and lines connecting students represent romantic relations occuring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else)

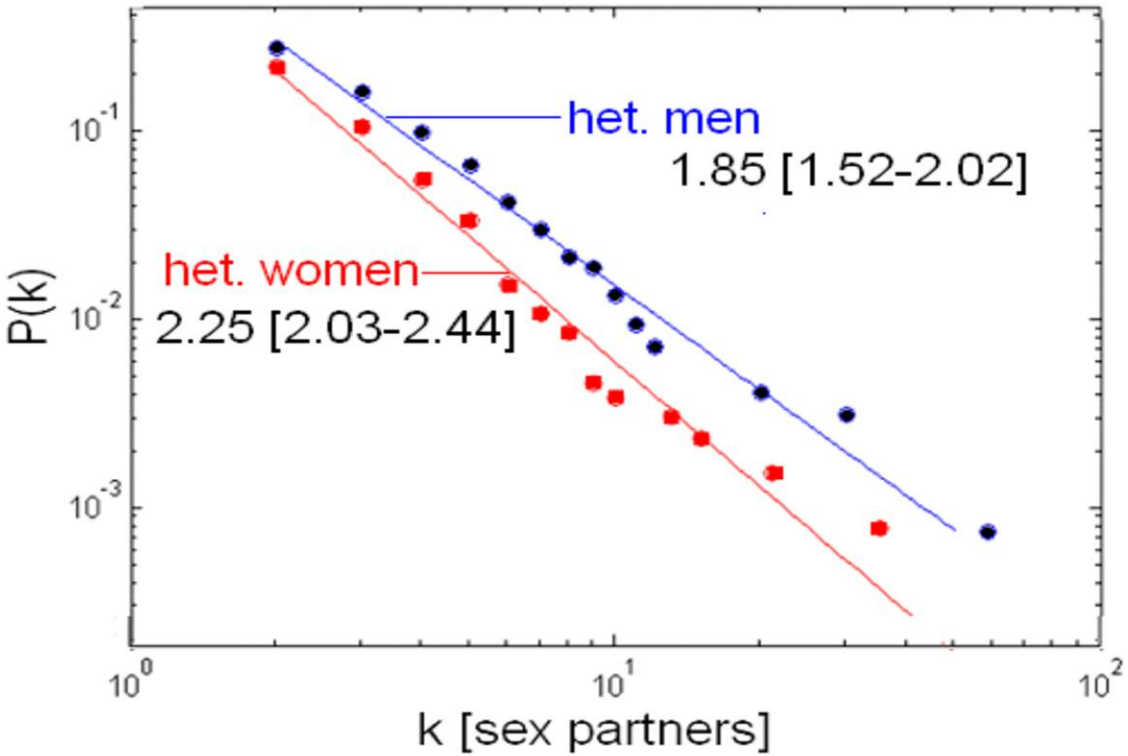
Relations "amoureuses" entre étudiant(e)s dans une université US

"sexual network"

American Journal of Sociology, Vol. 100, No. 1. "Chains of affection: The structure of adolescent romantic and sexual networks" Bearman PS, Moody J, Stovel K. (2004)

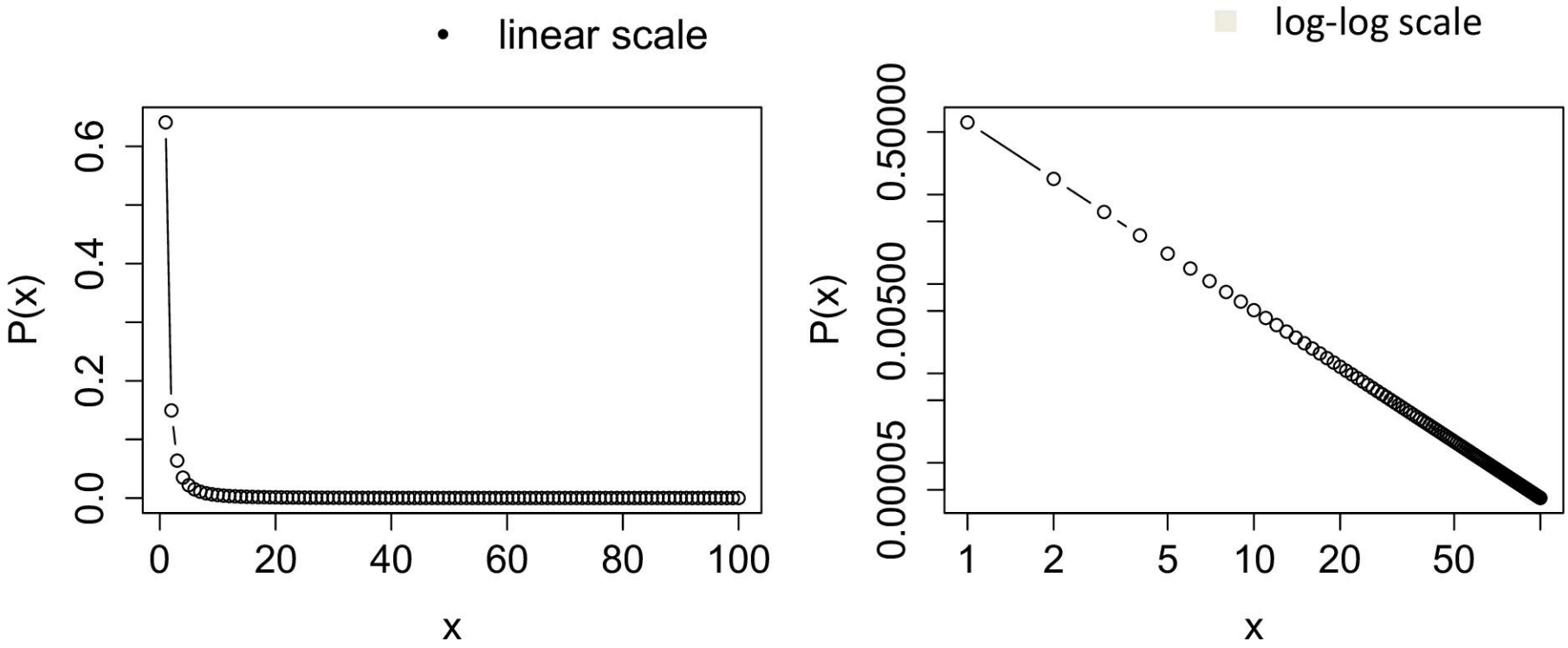
Distribution des degrés
Exemple 2: “réseau amoureux”

- Réseau ”sexuel”
- Fortes variations du nombre de partenaires



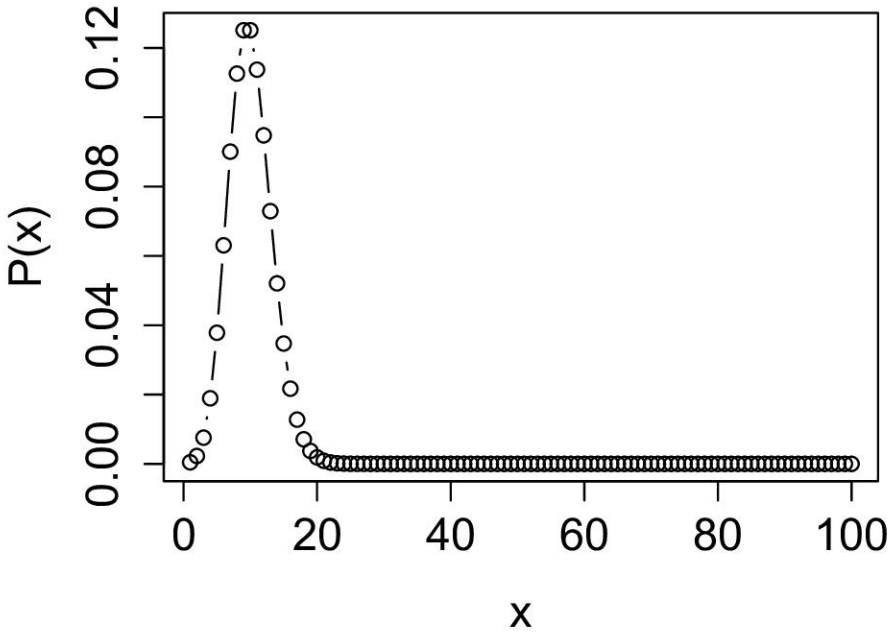
Distribution en loi de puissance

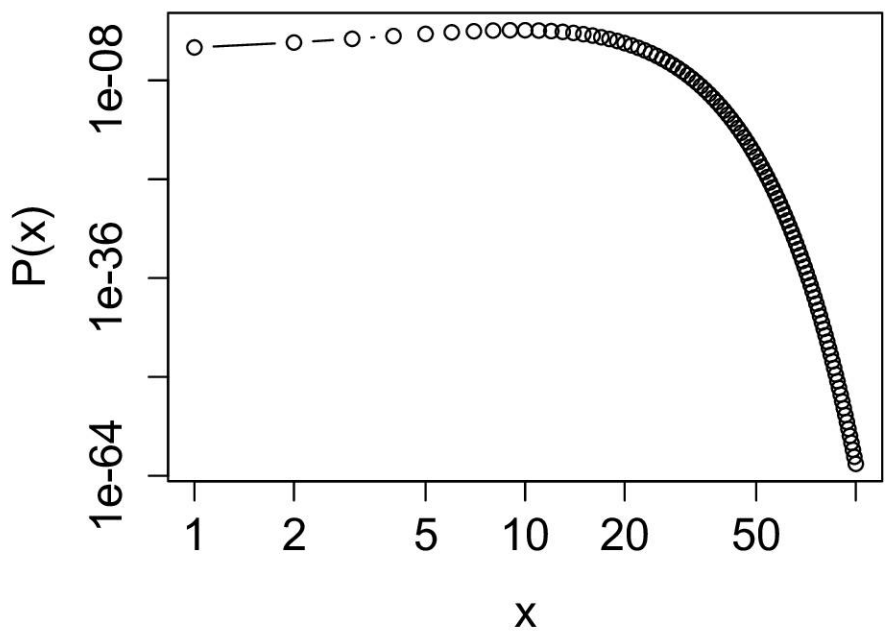
$$P(\deg = x) \propto x^{-\alpha}$$



- Très asymétrique
- Linéaire avec des axes log-log

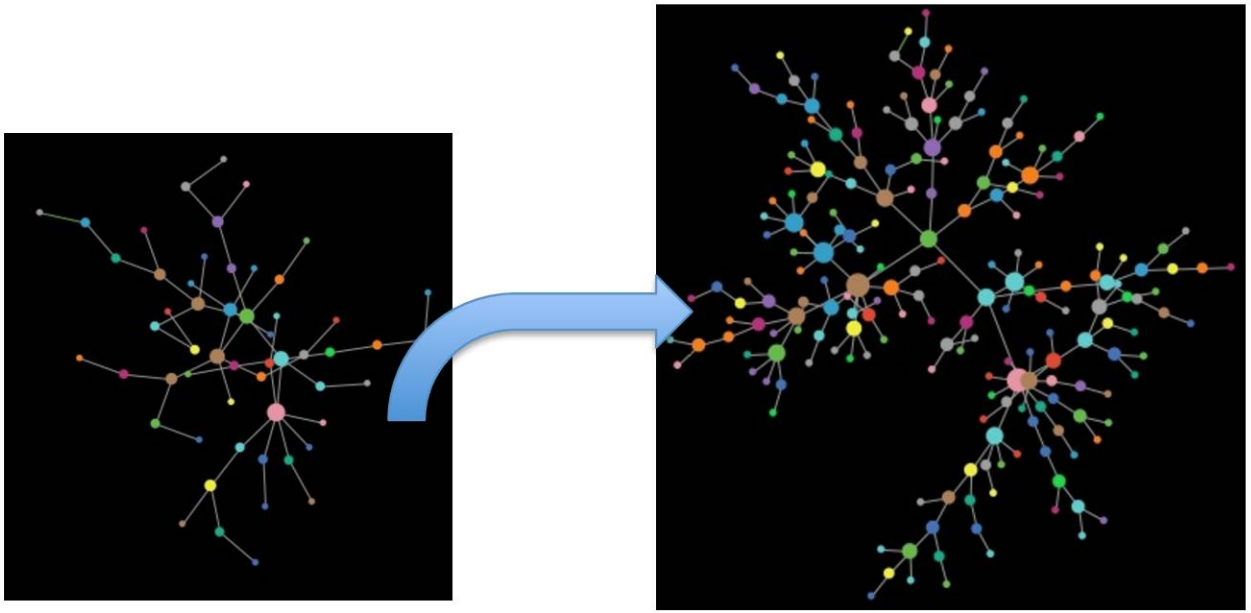
Distribution *de Poisson*
(attendue sur un graphe aléatoire)





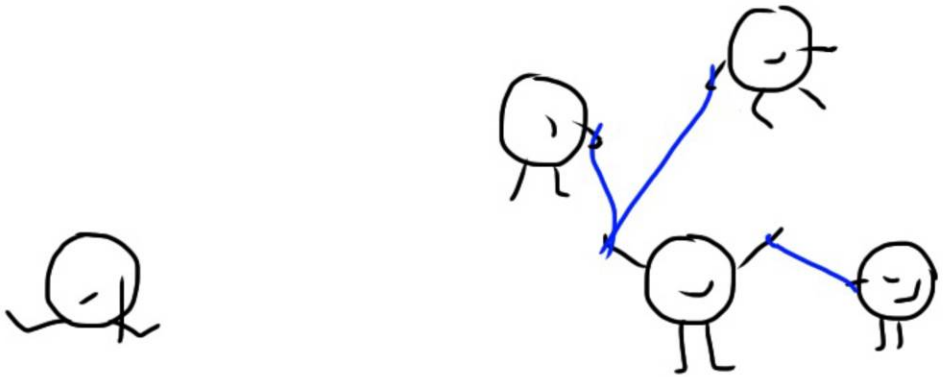
2 ingrédients pour obtenir une distribution *en loi de puissance*

1) Les nœuds apparaissent progressivement (croissance)

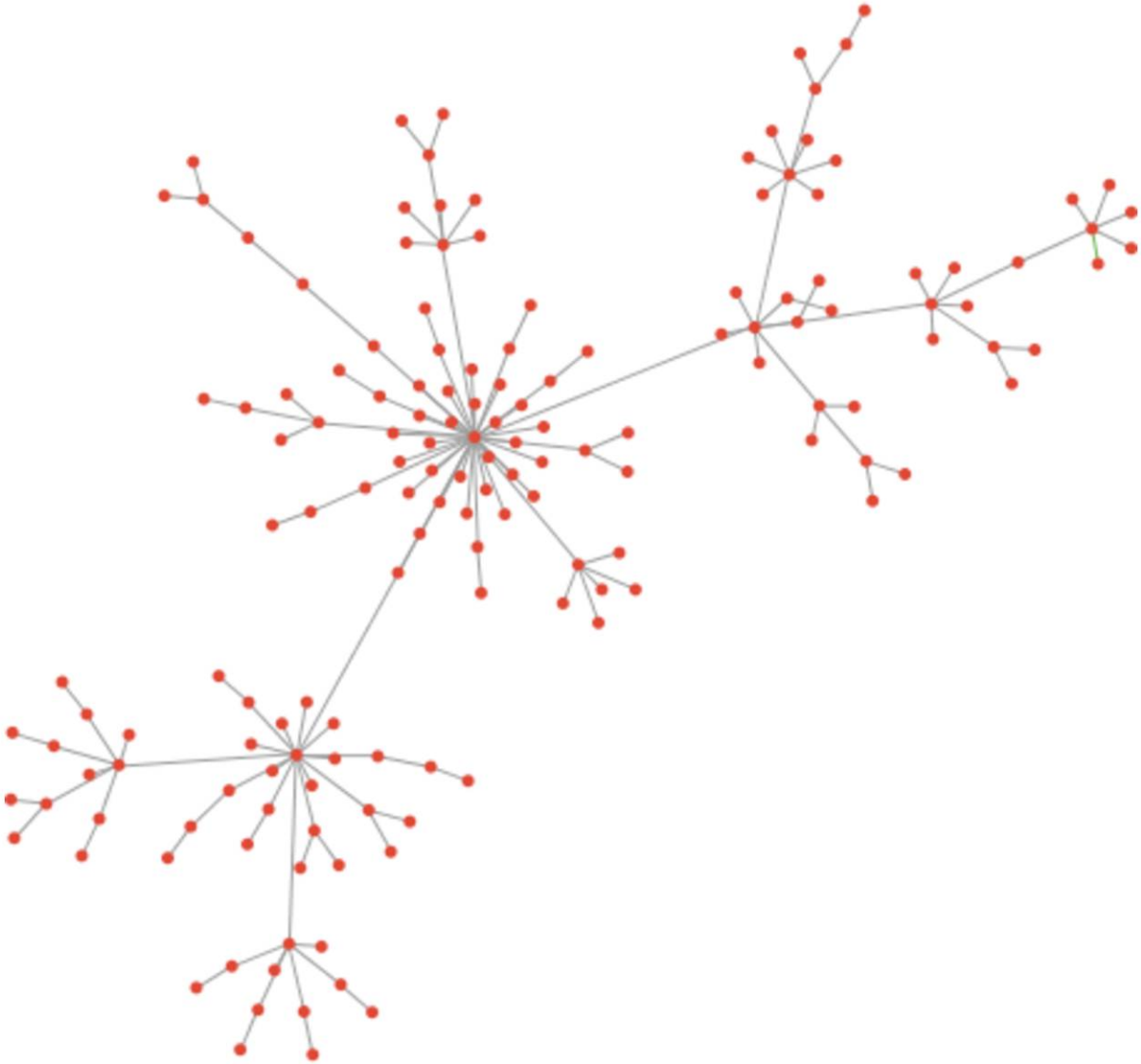


2 ingrédients pour obtenir une distribution *en loi de puissance*

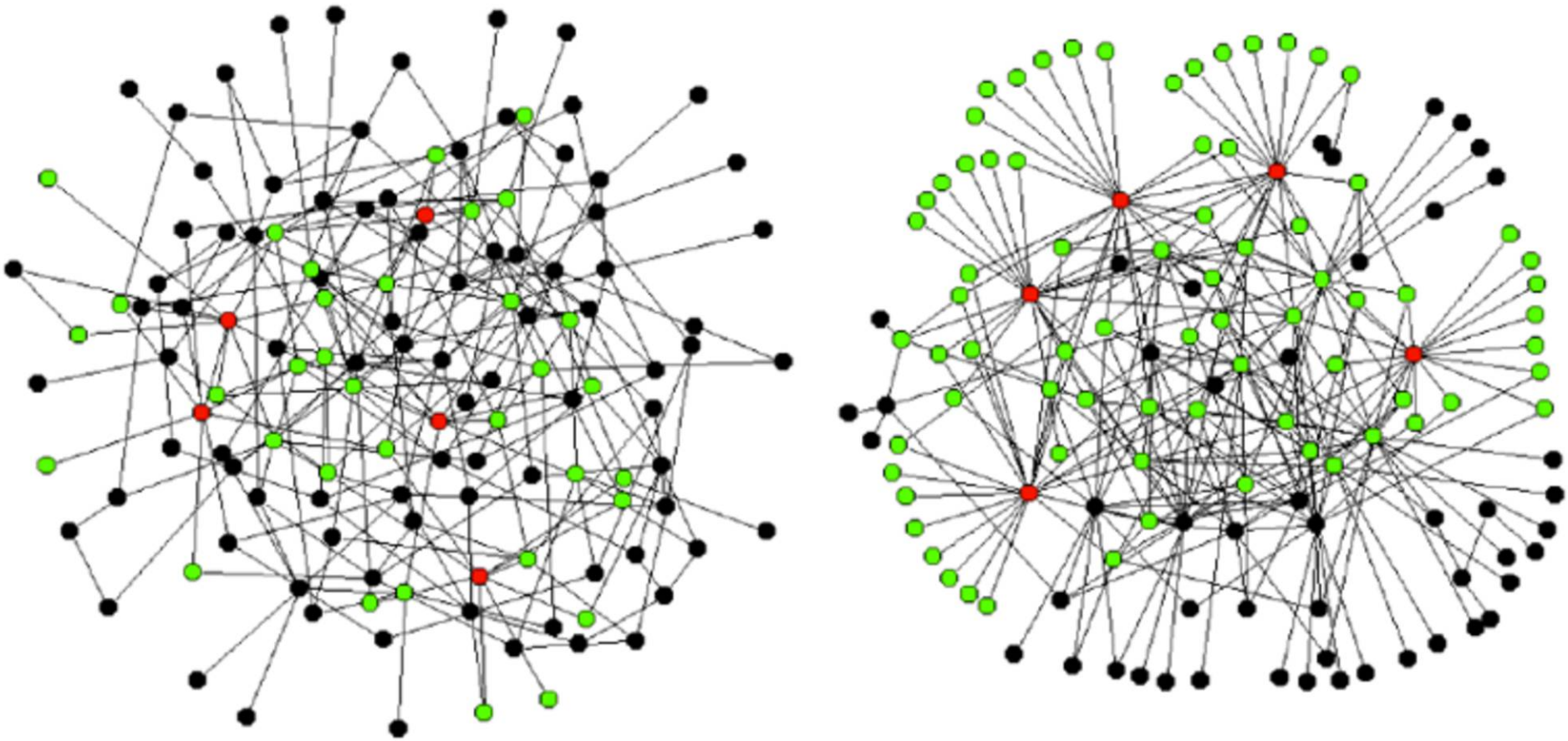
2) Les nœuds préfèrent s'attacher à des nœuds qui ont déjà de nombreuses connexions (*attachement préférentiel*)



Croissance: modèle «attachement préférentiel»



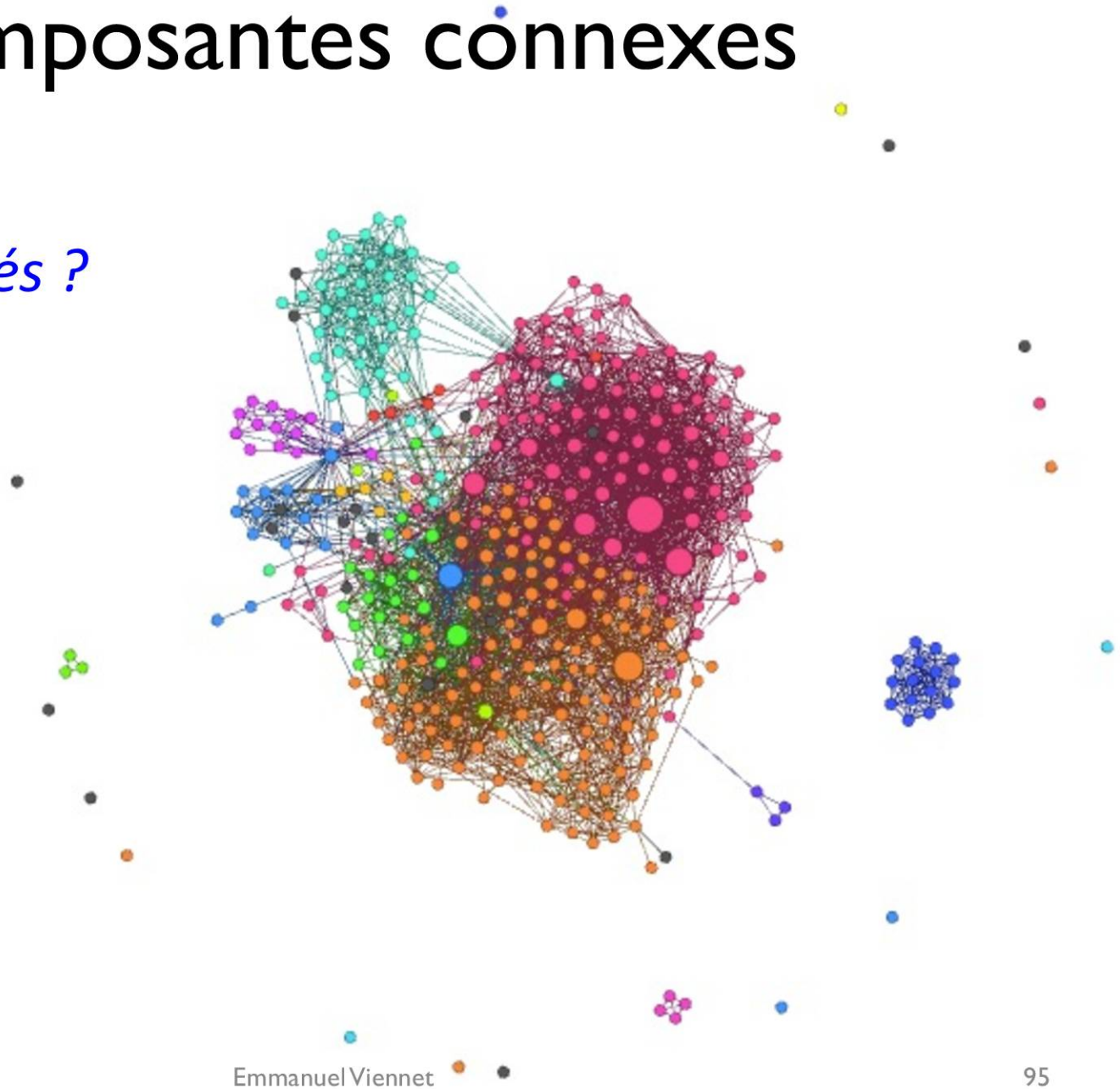
$$P(\text{degré d'un nœud } = k) \propto k^{-\gamma}$$



(Albert et al, 2000)

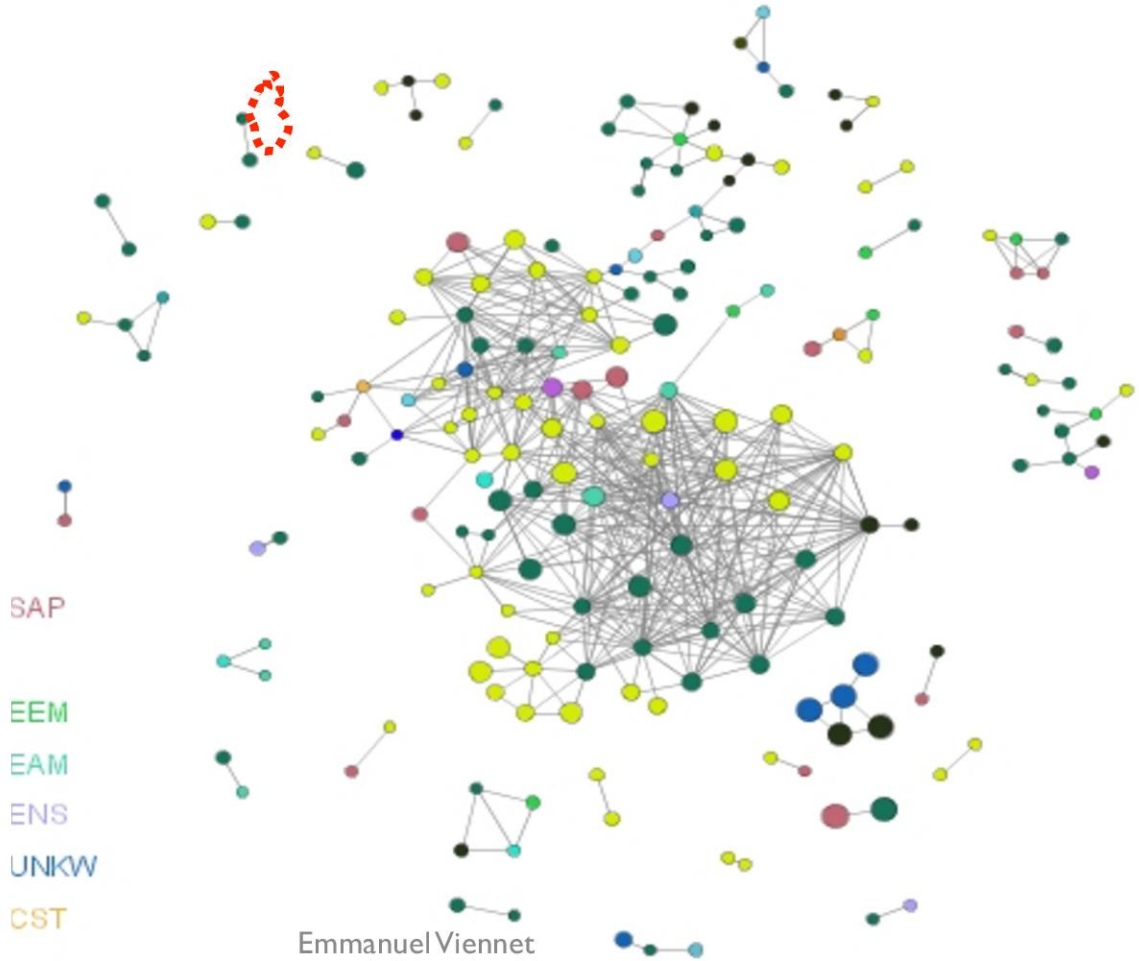
Composantes connexes

Tous connectés ?



Composante géante

- Si la plus grande composante connexe contient l'essentiel du graphe, on l'appelle la **composante géante**



Mesurer l'importance d'un nœud

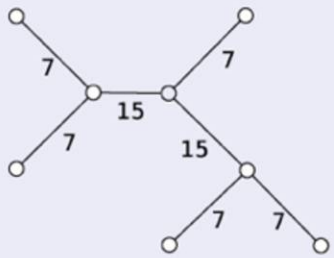
Importance ou « centralité »

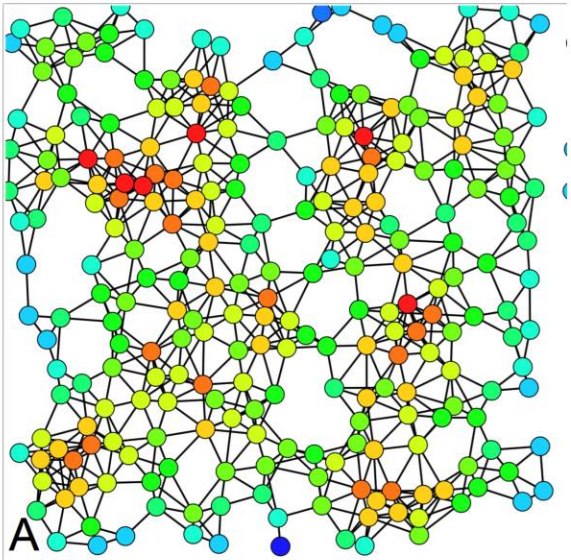
Peut être mesurée par:

- Le degré
- La centralité de proximité (distance moyenne aux autres nœuds)
- Intermédiarité (associée à un nœud ou une arête)

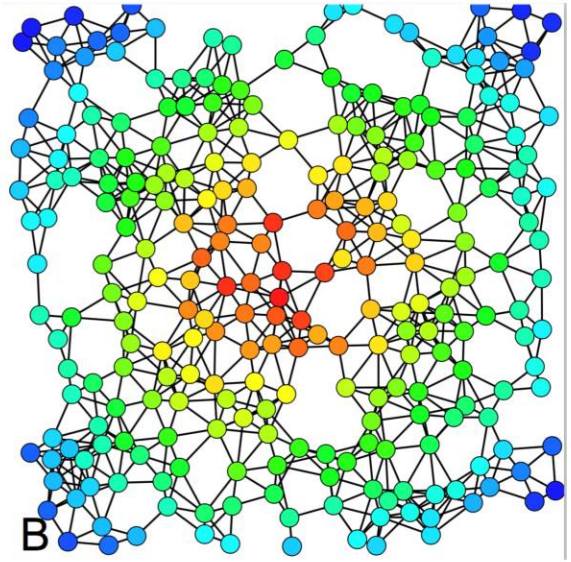
Intermédiarité

Nombre de plus courts chemins passant par une arête (Newman 2004)

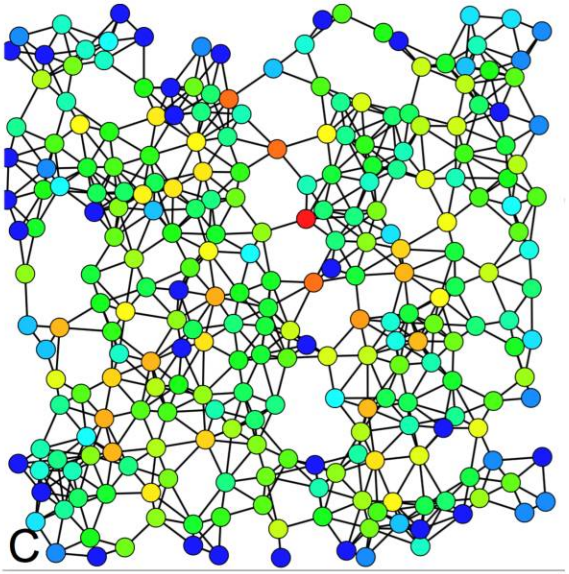




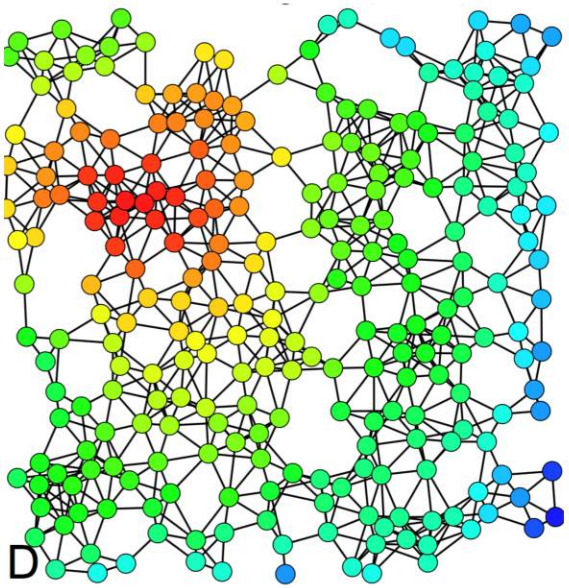
Centralité de degré



Centralité de proximité

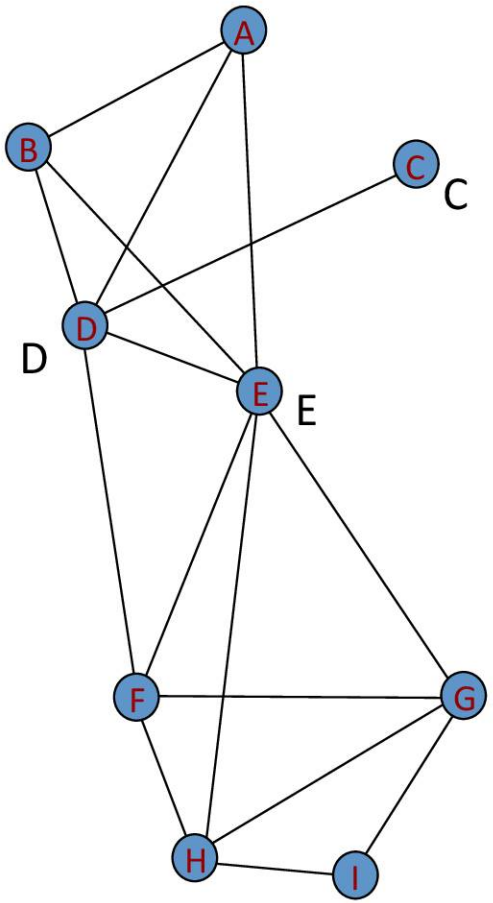


Centralité d'intermédiarité



Centralité spectrale

Rôles sociaux des nœuds



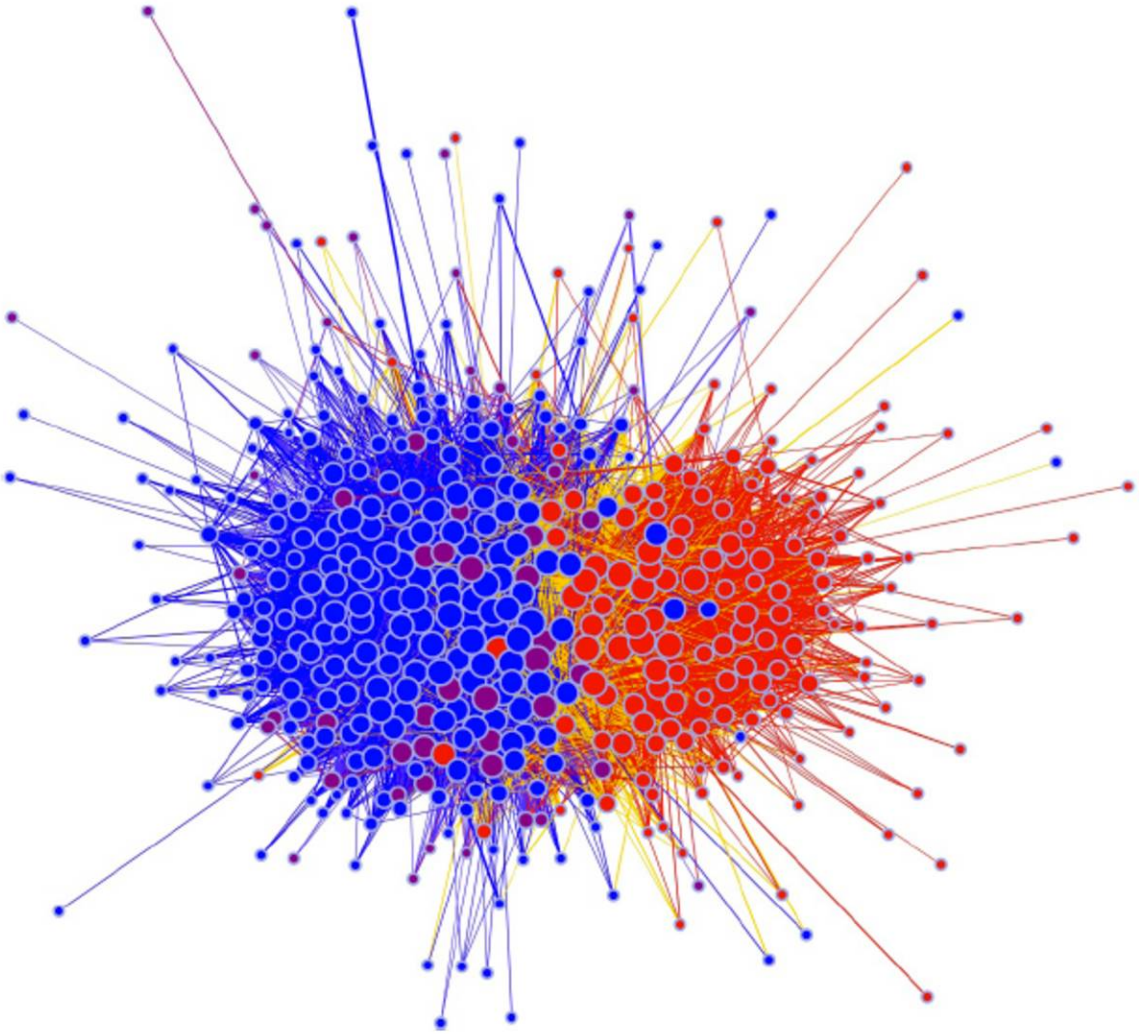
- **Rôles simples (structurels)**
  - Périphérique : faible centralité C C
  - Connecteur/central: haute centralité D D
  - Passeur : haute intermédiarité E E
- **Rôles basés sur les actions**
  - Auteur, commentateur, lecteur
  - Opinion (ex. text mining)

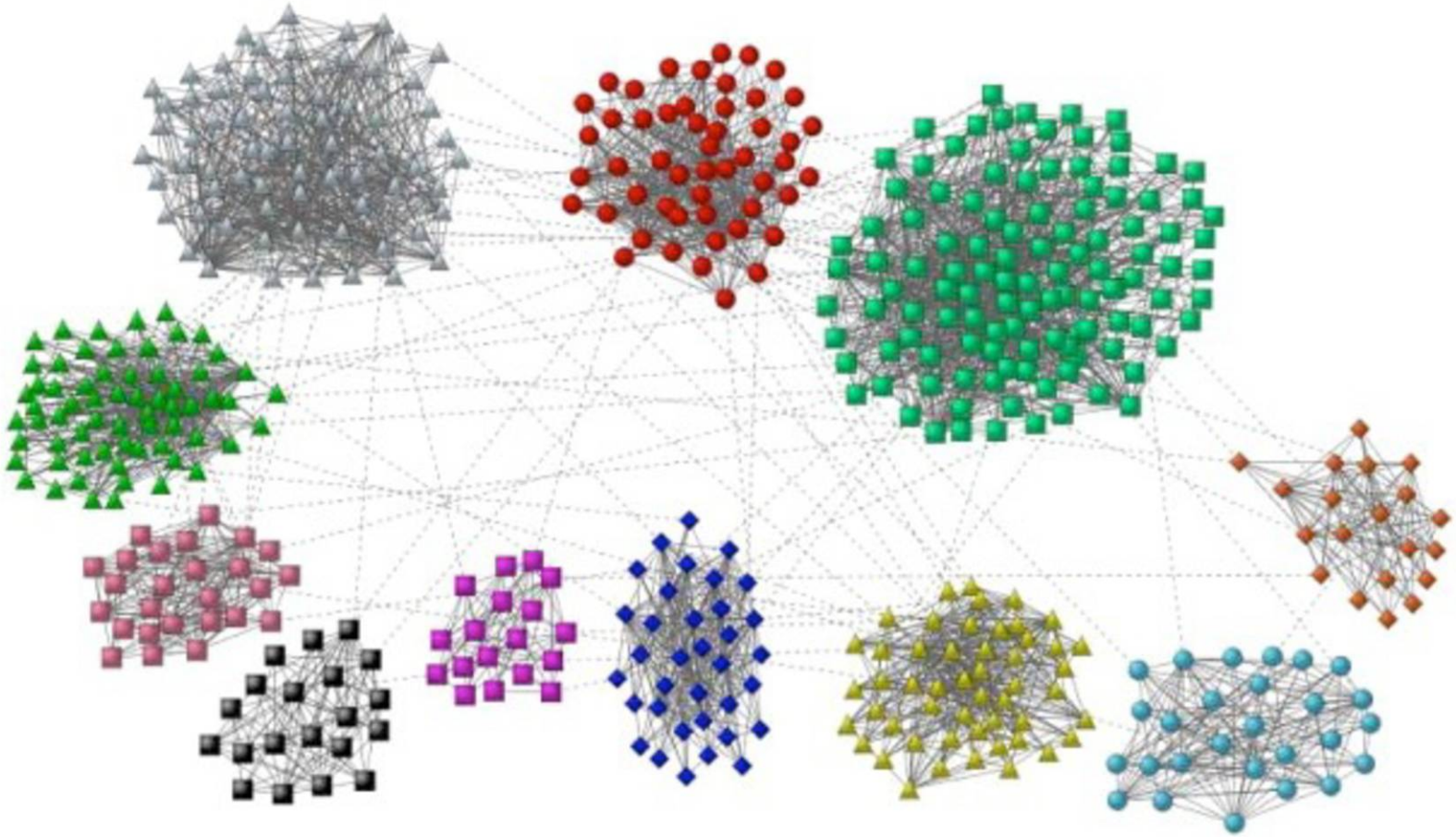
Introduction
à l'analyse des réseaux sociaux

Partie 5

Communautés dans les réseaux

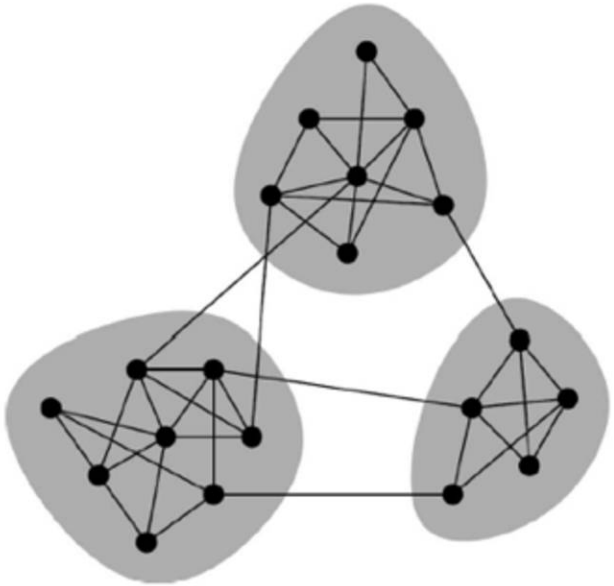
Détection de communautés dans les réseaux

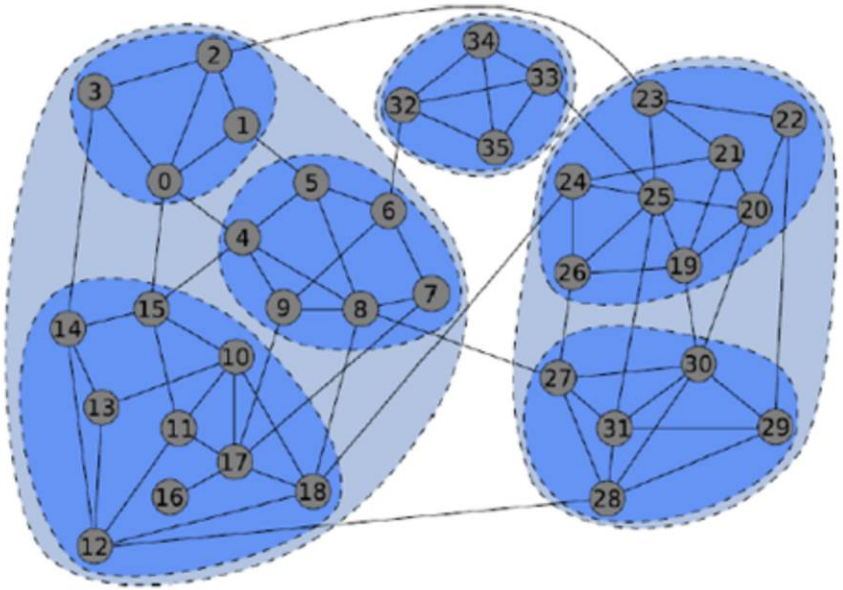




A. Lancichinetti, S. Fortunato (2009)

Communautés





(P. Pons, 2007)

- Recherche de communautés = partitionnement du graphe en N
- Identification = recherche d'une communauté autour d'un nœud donné

⚠ On ne considère que les liens (la «structure» du réseau)

Modularité d'une partition

(Newman 2004)

Comparaison des densités d'arêtes avec un *null model*:

Modularité d'une partition d'un graphe

$$Q = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right]$$

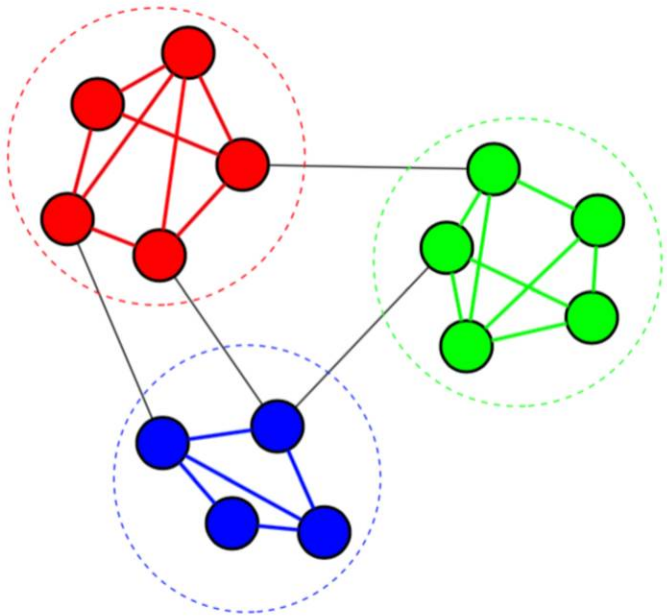
n_c : nombre of communautés (parties)

l_c : nombre d'arêtes *internes* dans c

d_c : somme des degrés des nœuds de c

L : nombre total d'arêtes

$Q \in [-1, 1]$



Modularité d'une partition

(Newman 2004)

Comparaison des densités d'arêtes avec un *null model*:

Modularité d'une partition d'un graphe

$$Q = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{d_c}{2L} \right)^2 \right]$$

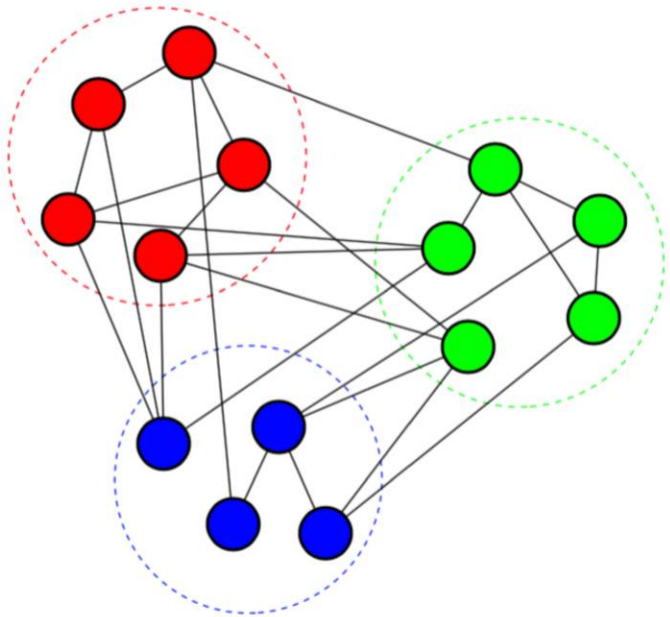
n_c : nombre of communautés (parties)

l_c : nombre d'arêtes *internes* dans c

d_c : somme des degrés des nœuds de c

L : nombre total d'arêtes

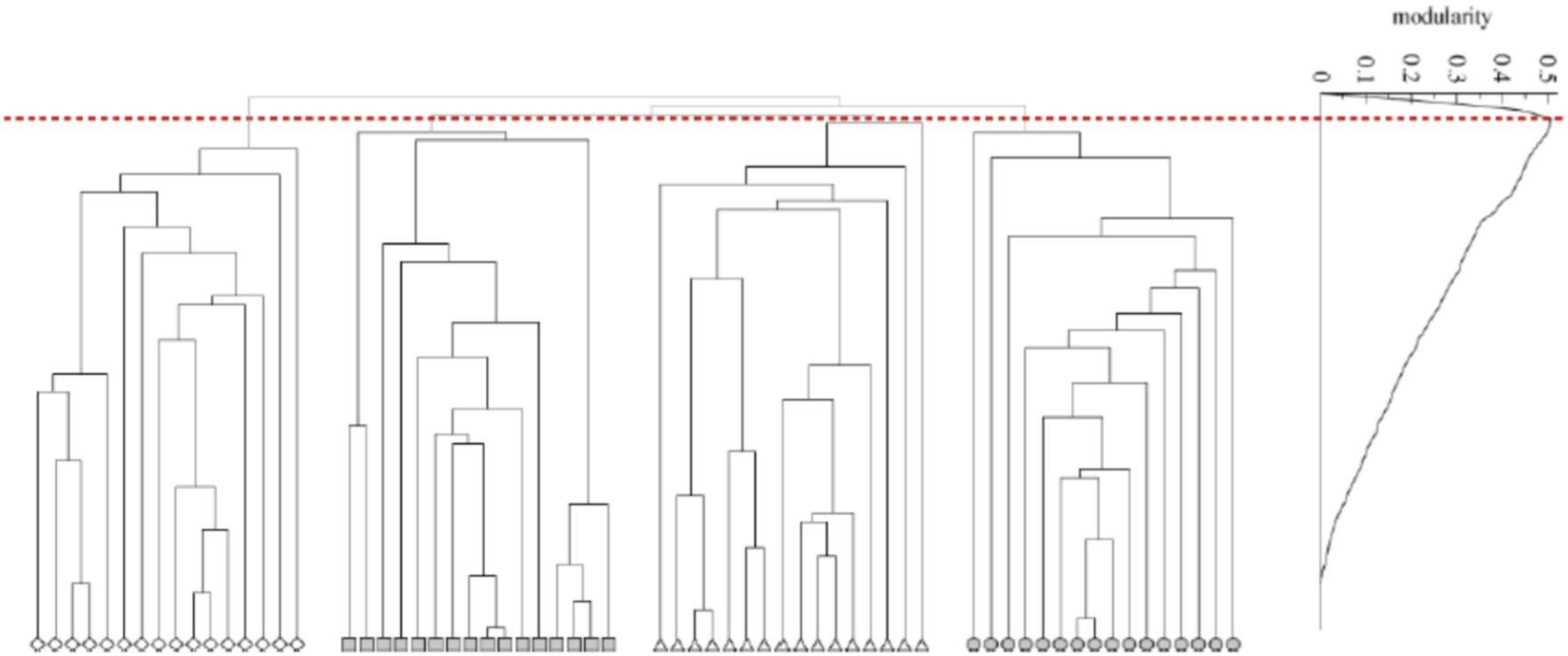
$Q \in [-1, 1]$



Recherche de communautés maximisant Q

Première proposition: Newman & Girvan (2004)

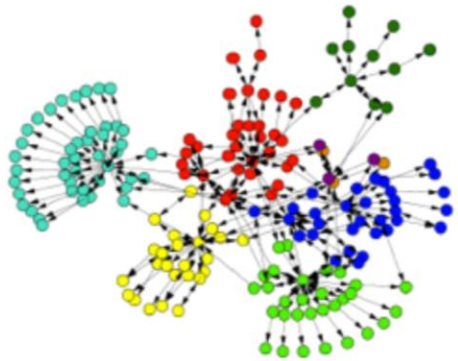
- Répéter:
  1. calculer l'intermédiarité des arètes
  2. couper l'arète la plus importante
- jusqu'à isoler tous les nœuds (méthode séparative)



Recherche de communautés maximisant Q
Progrès rapides dans la décennie 2000

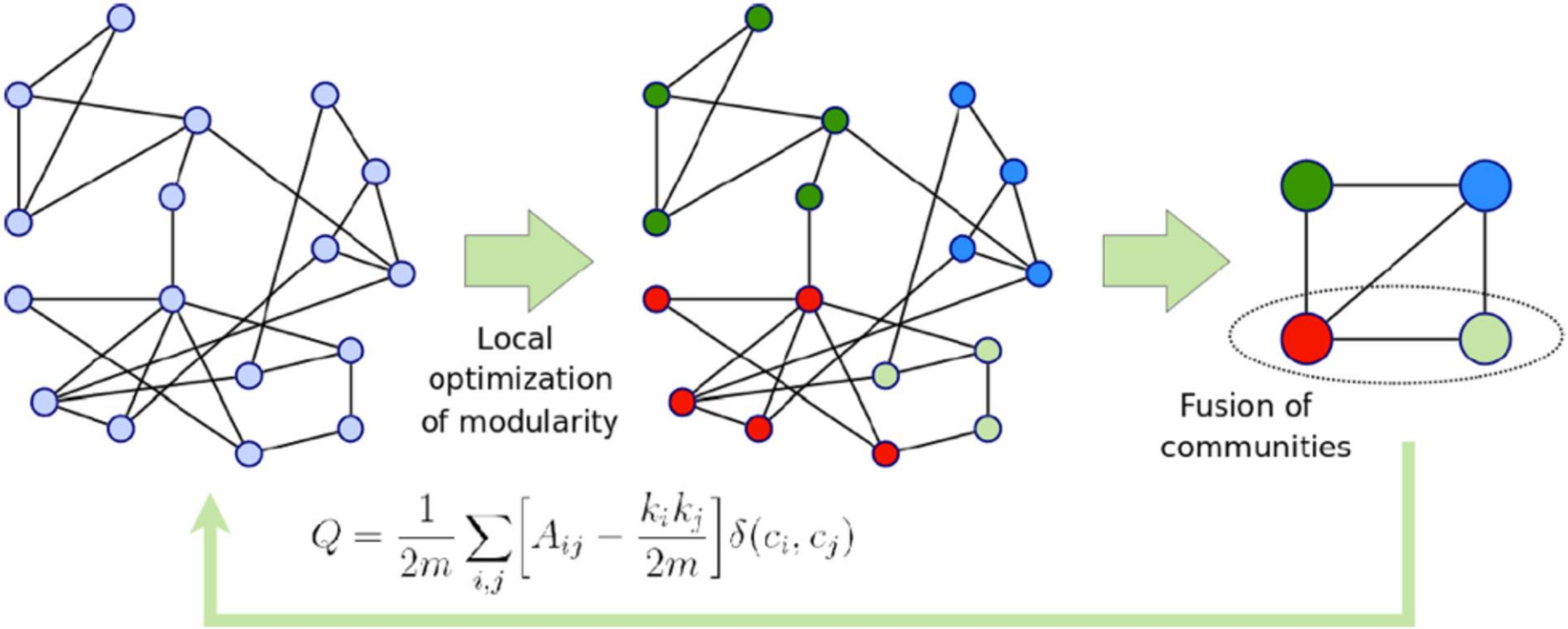
Pour un grand réseau parcimonieux de n nœuds:

Newman & Girvan	2004	$O(n^3)$
Newman	2004	$O(n^2)$
Wakita & Tsurumi	2007	$O(n \log^2 n)$
Blondel et al. (Louvain)	2008	quasi-linéaire



$\rightsquigarrow$ moins de 5 minutes pour 1 million de nœuds, ou 40 minutes pour 23 millions

Recherche de communautés: méthode de Louvain

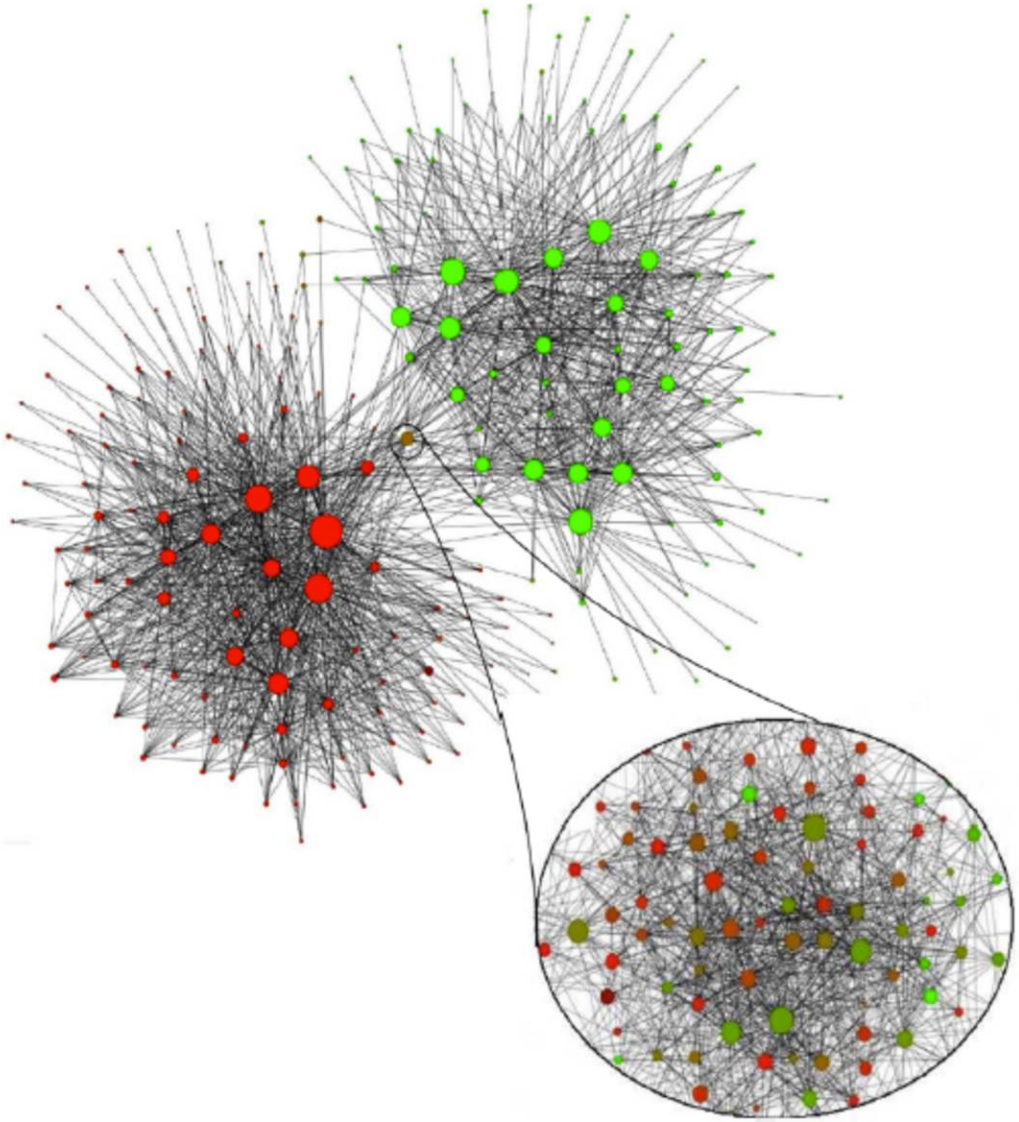


$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

Optimisation locale, on tente d'associer un nœud à la même communauté que l'un de ses voisins.

Blondel et al., Fast unfolding of communites in large networks, 2008

Exemple 2: réseau téléphonique en Belgique



2 millions de clients

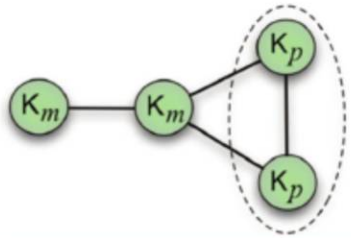
Ici chaque cercle est une communauté avec > 100 clients

Les couleurs représentent le langage parlé dans la communauté (français en rouge, flamand en vert)

D'après Blondel, et al.

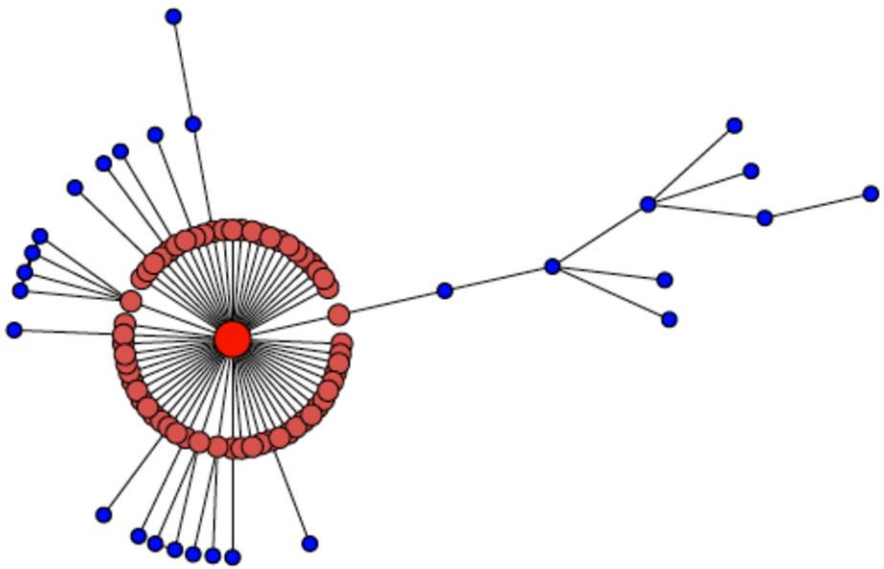
Limitations de la modularité

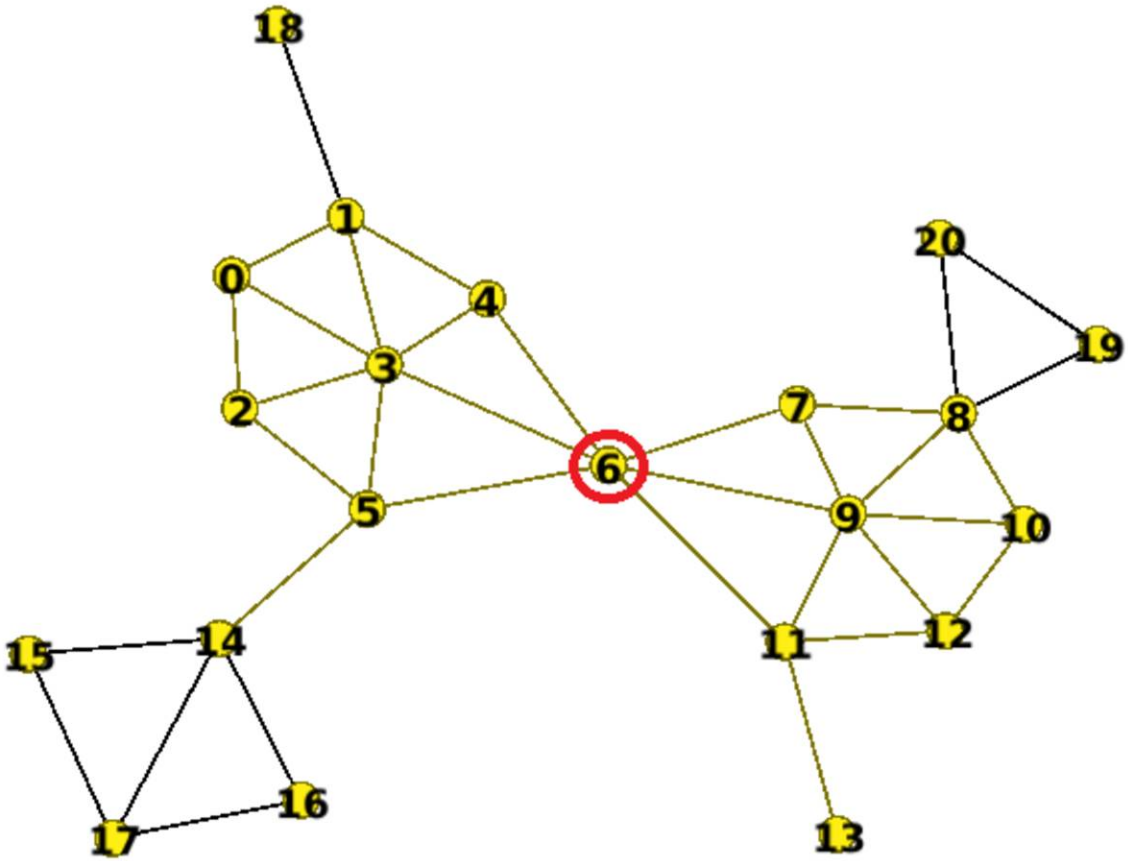
- **Stabilité**: de nombreuses partitions donnent des valeurs proches de la modularité Q
- Absence d'un critère fiable indiquant **l'existence** de communautés significatives (voir contre-exemple détaillé dans le chapitre 9)
- **Limite de résolution**: impossible de détecter de petites communautés



- S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, 2010.
- A. Lancichinetti and S. Fortunato, "Limits of modularity maximization in community detection," *Physical Review E*, vol. 84, no. 6, 2011.
- Ivan Keller and Emmanuel Viennet. "A characterization of the modular structure of complex networks based on consensual communities". In *IEEE SITIS 2012, Complex Networks and their Applications*, 2012.

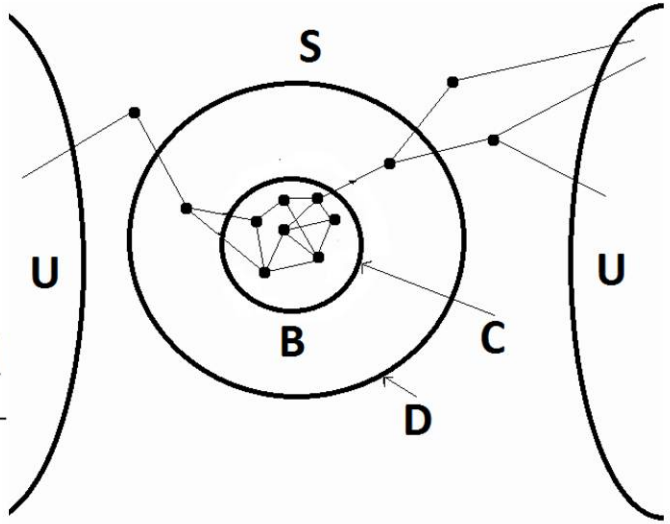
Communautés locales (ou egocentrées)




18
1
0
4
3
2
5
6
7
8
9
10
11
12
13
14
15
16
17
20
19

Algorithme de détection de la communauté locale d'un nœud

(Ngonmang, Tchuente & Viennet 2012)



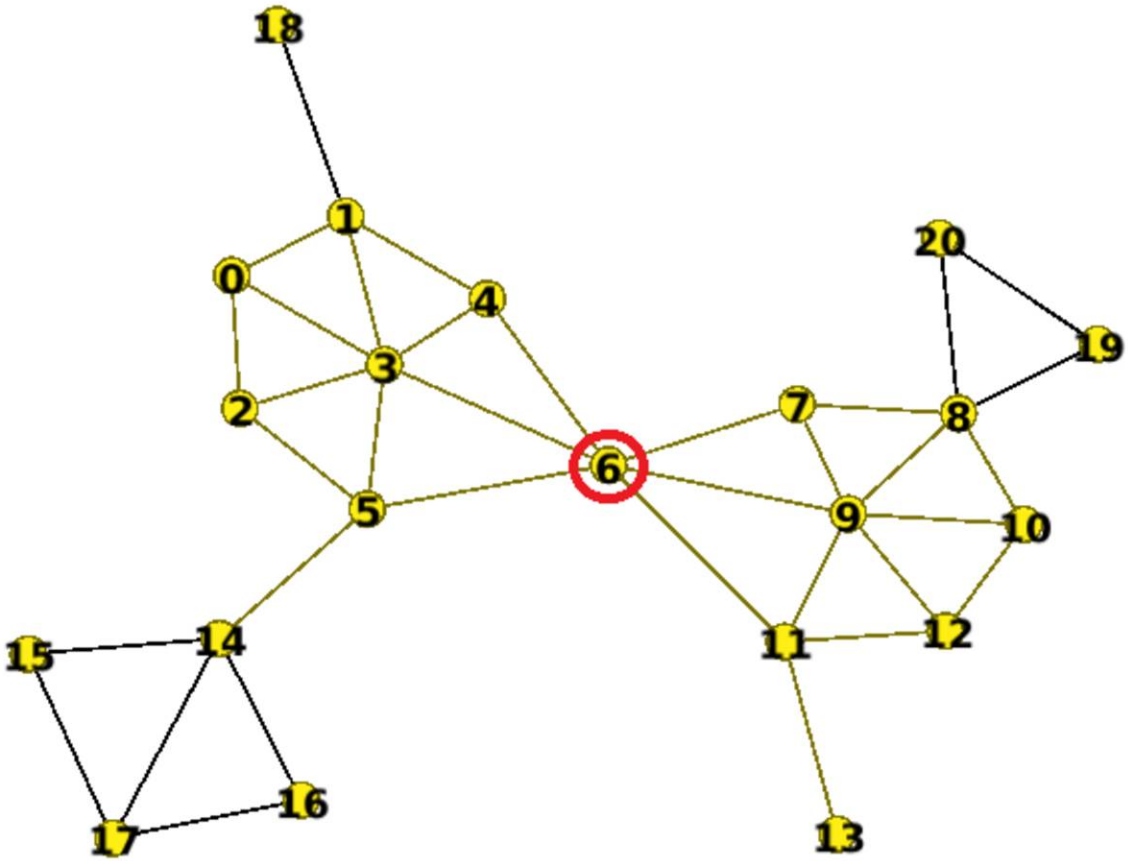
$$Q = \frac{Q_{in}}{Q_{ex}}$$

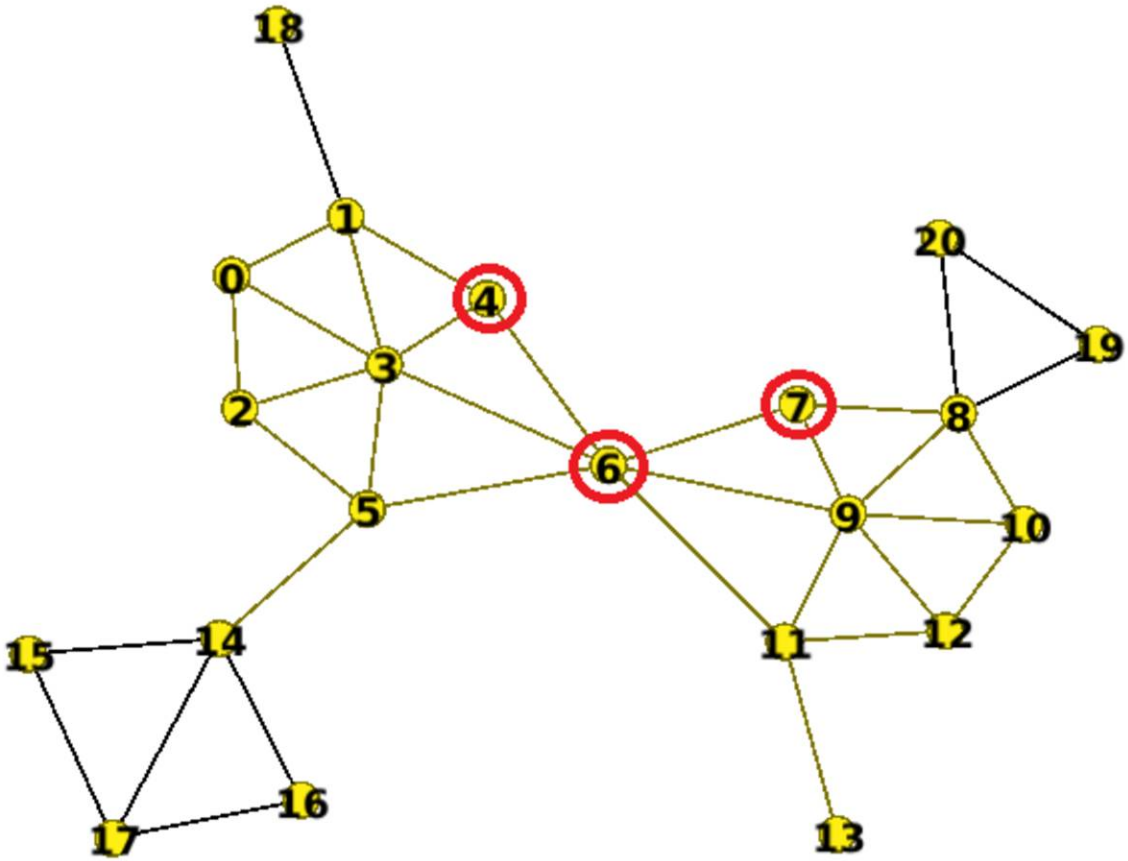
où Q_{in} est la contribution des liens internes :

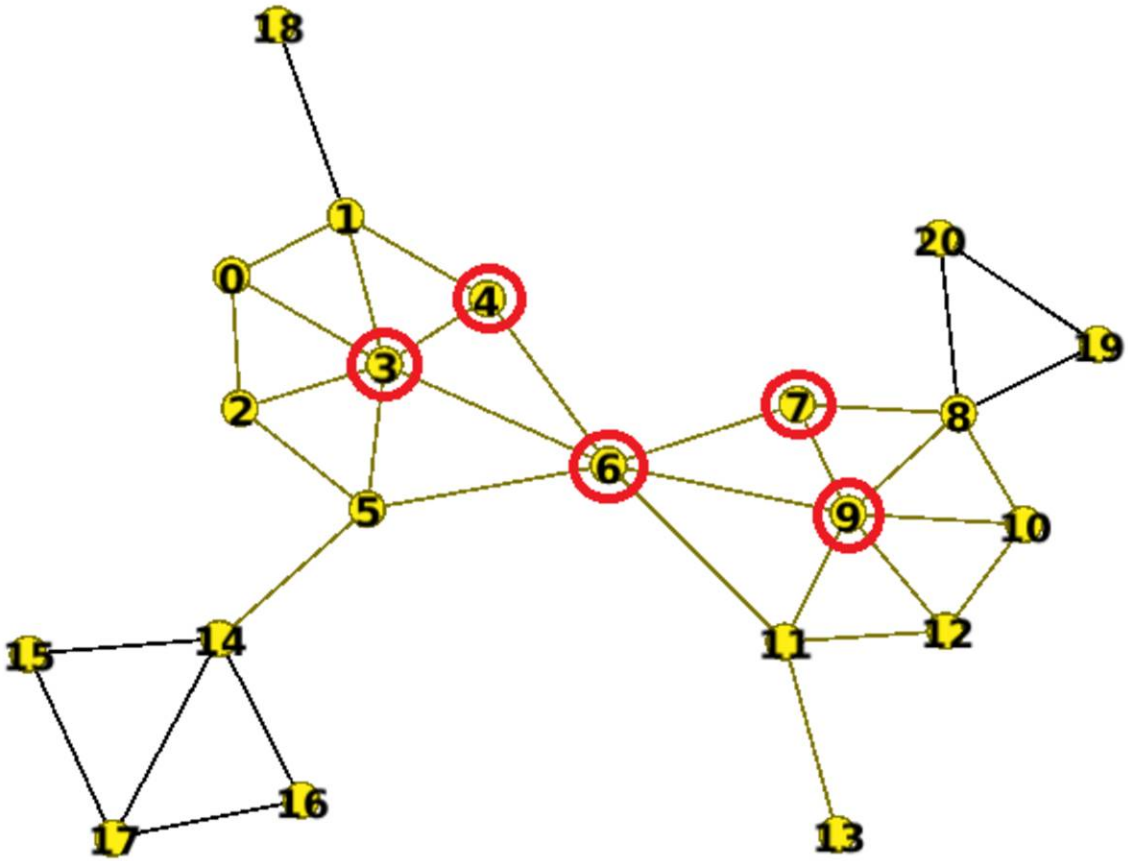
$$Q_{in} = \frac{\sum_{i \in D} \frac{|\Gamma(i) \cap D|}{(1+d_i)}}{D}$$

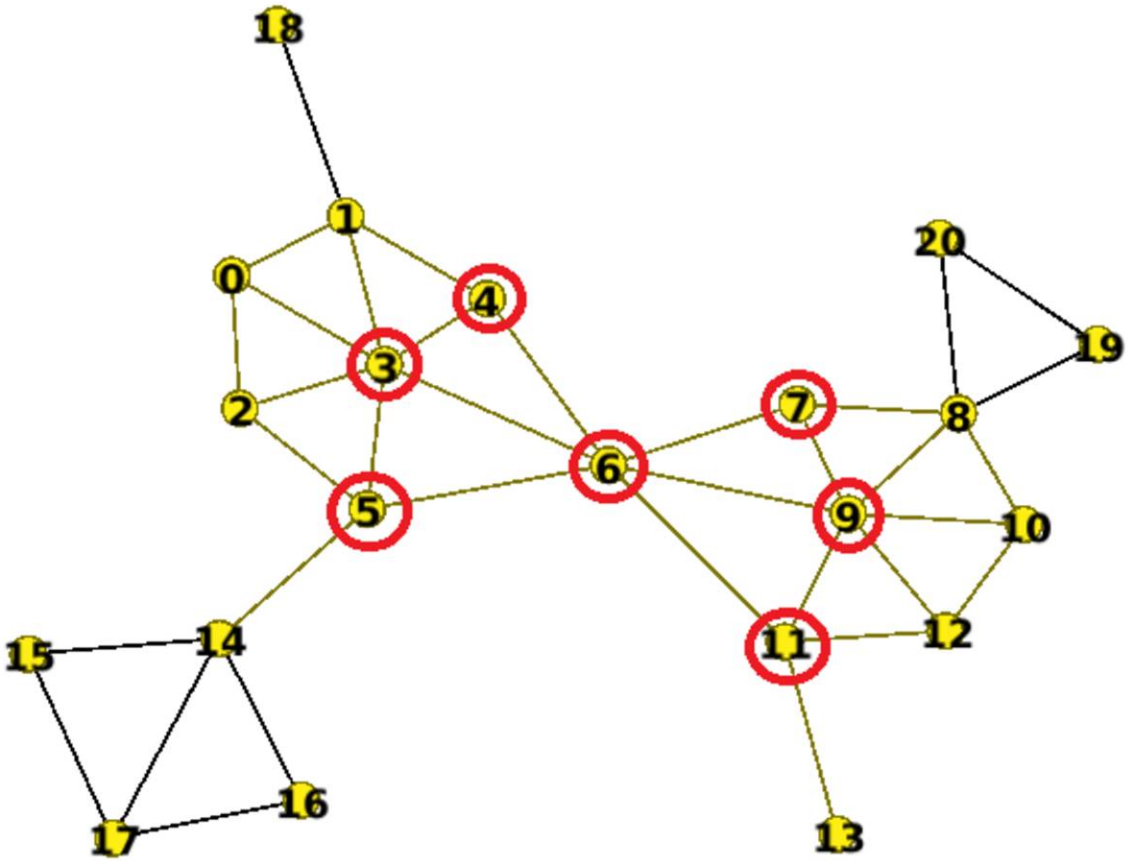
et Q_{ex} celle des liens externes :

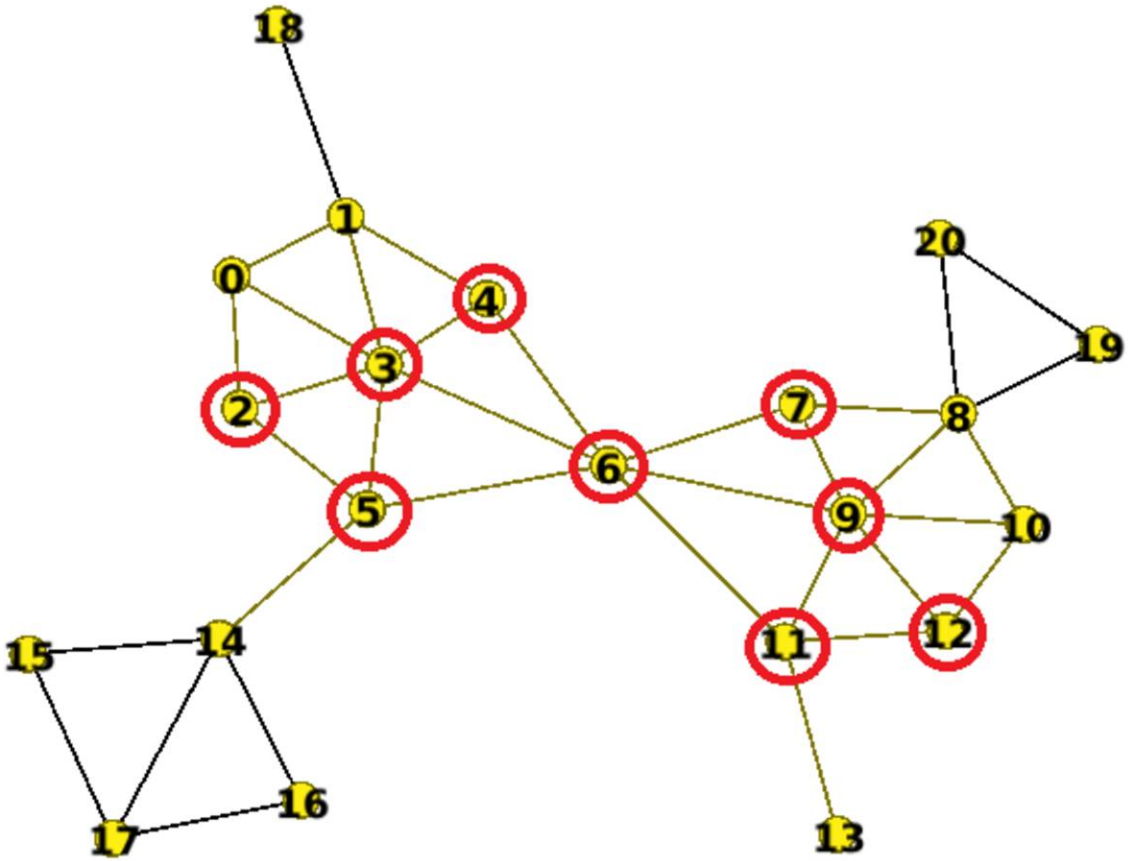
$$Q_{ex} = \frac{\sum_{i \in D} |\Gamma(i) \cap S|(1+d_i)}{D}$$

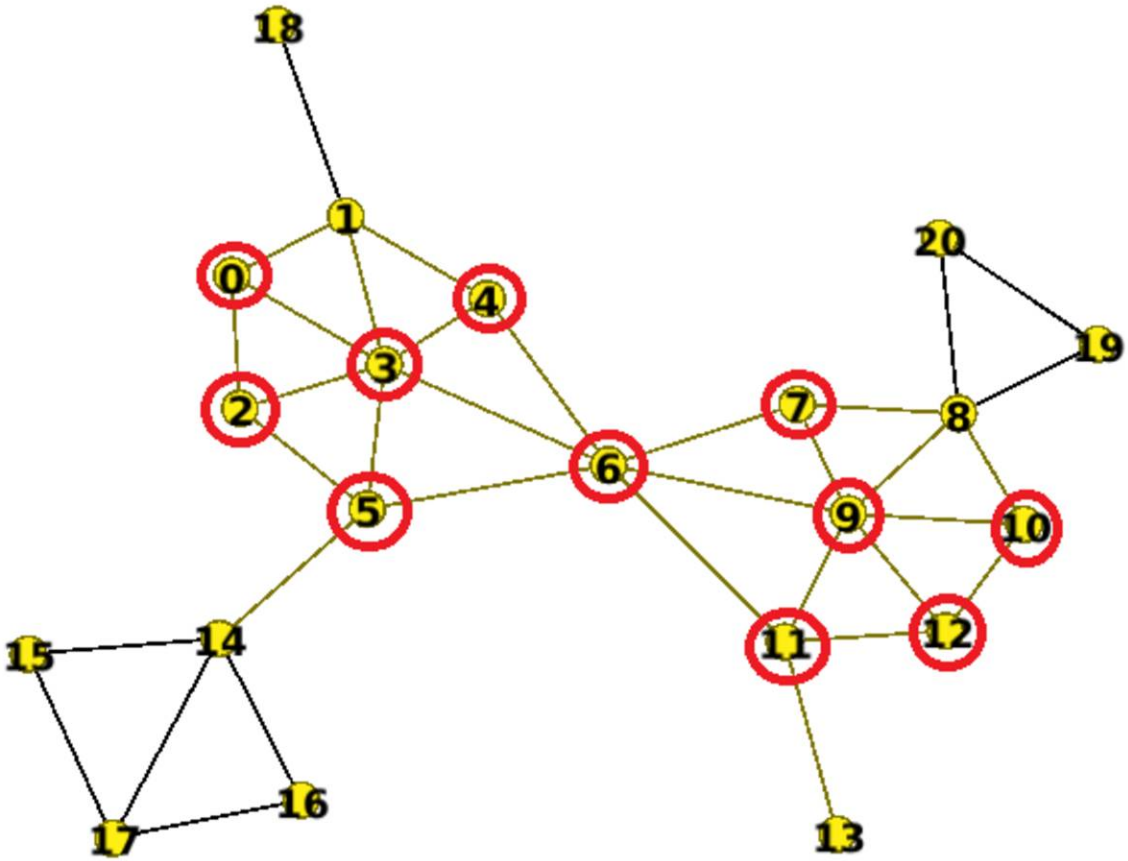

18
1
0
4
3
2
5
6
7
8
9
10
11
12
13
14
15
16
17
20
19

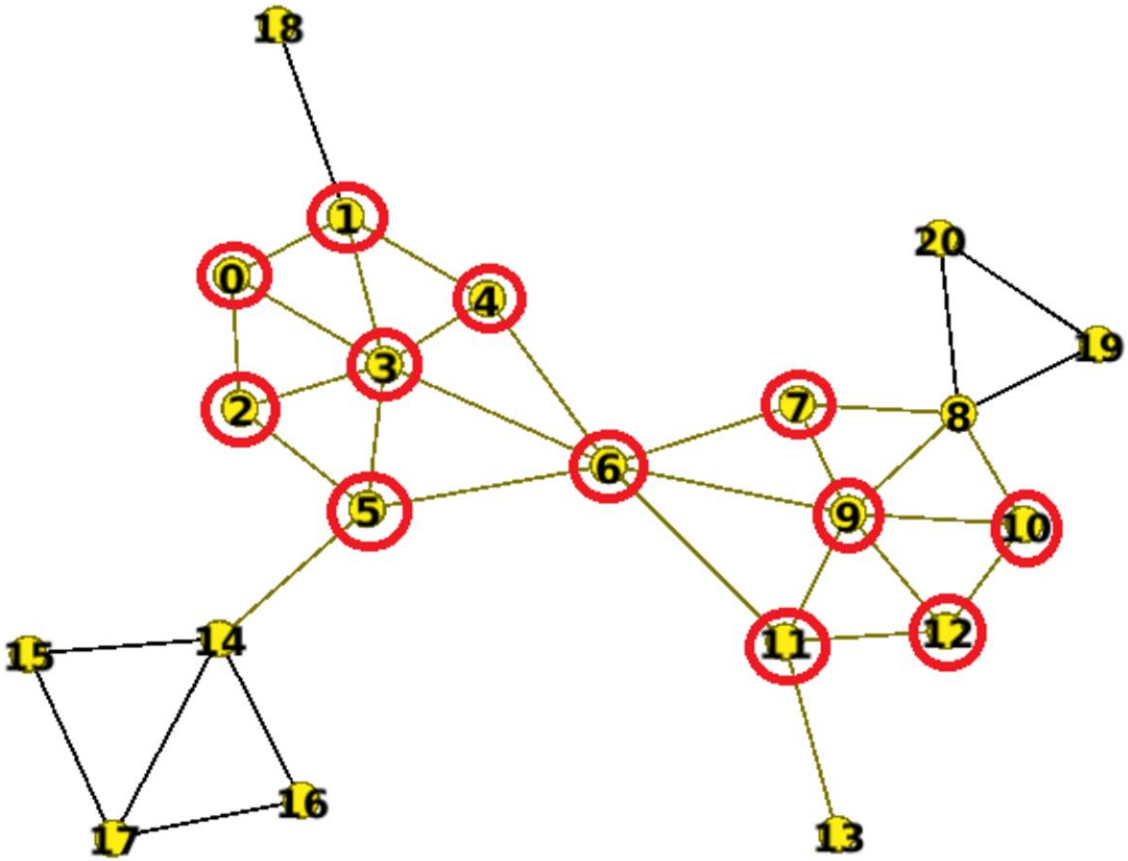

18
1
0
4
3
2
5
6
7
8
9
10
11
12
13
14
15
16
17
20
19

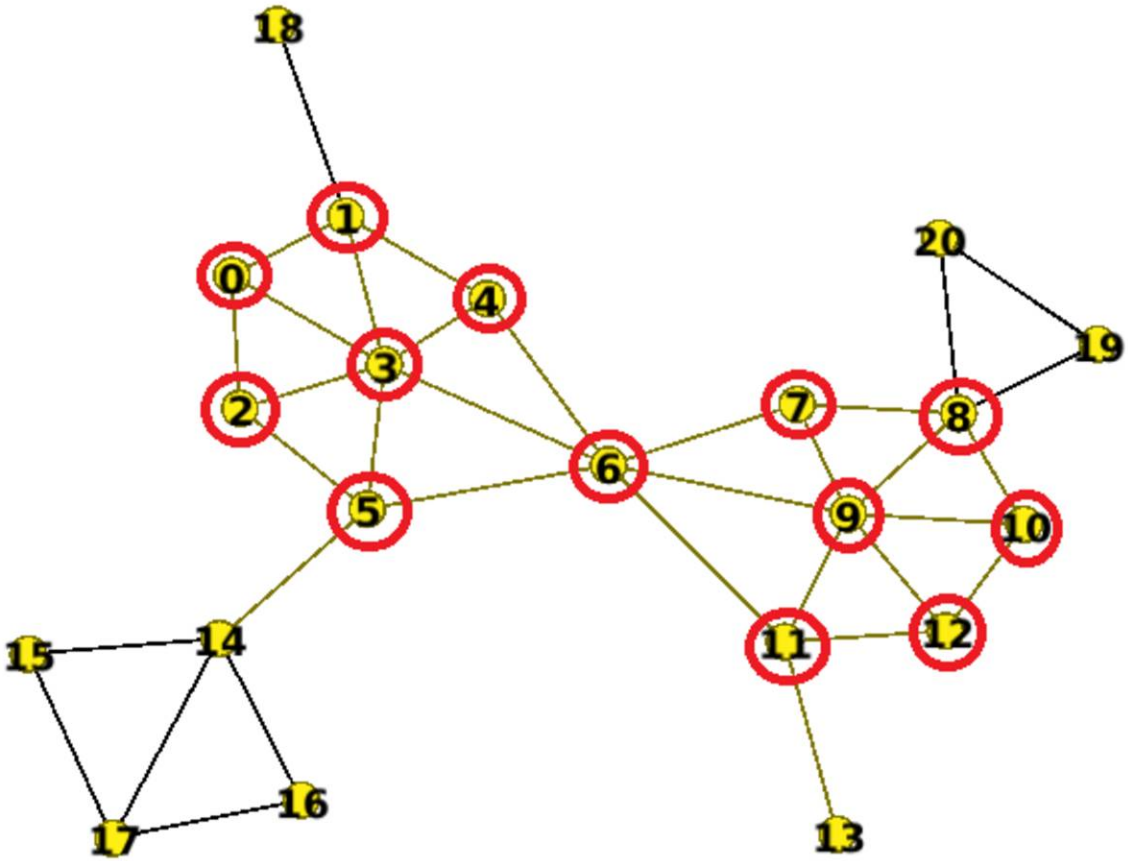

18
1
0
4
3
2
5
6
7
9
8
20
19
10
12
11
13
15
14
16
17


18
1
0
4
3
2
5
6
7
8
9
10
11
12
13
14
15
16
17
20
19

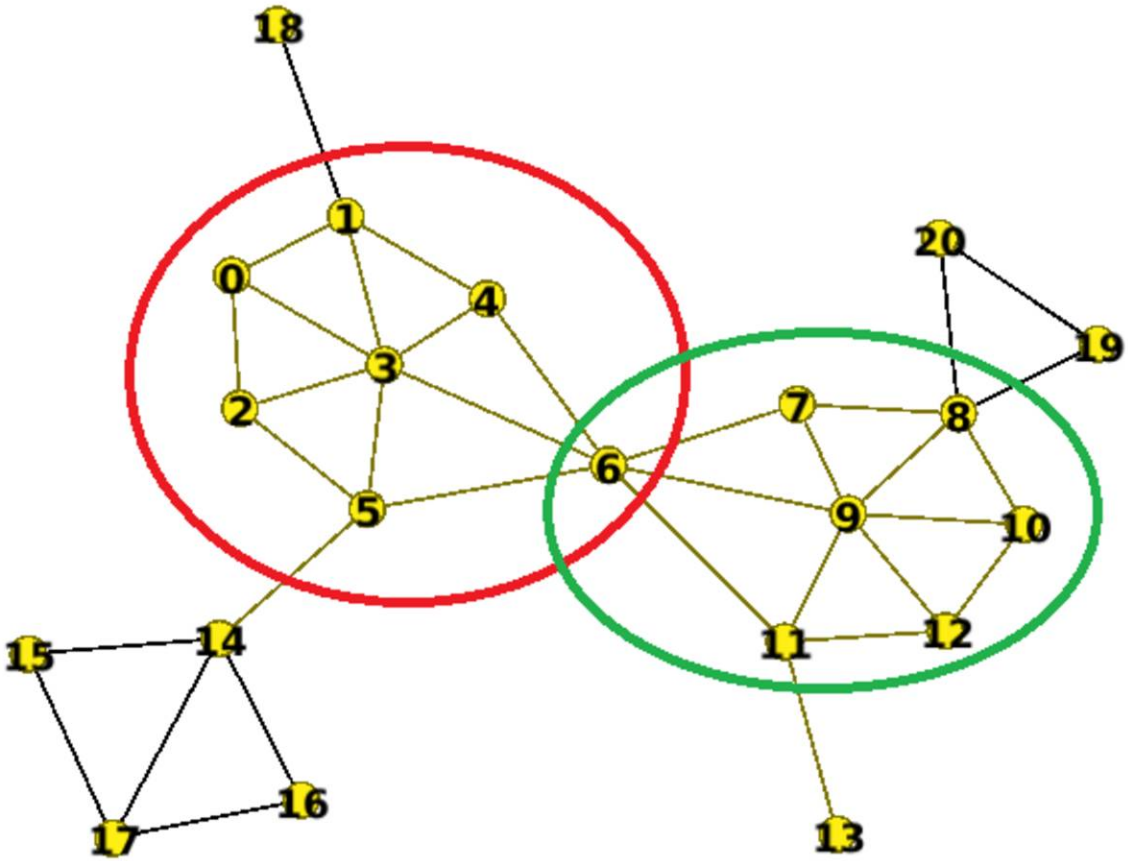

18
1
0
4
3
2
5
6
7
8
9
10
11
12
13
14
15
16
17
20
19


18
1
0
4
3
2
5
6
7
8
9
10
11
12
13
14
15
16
17
20
19

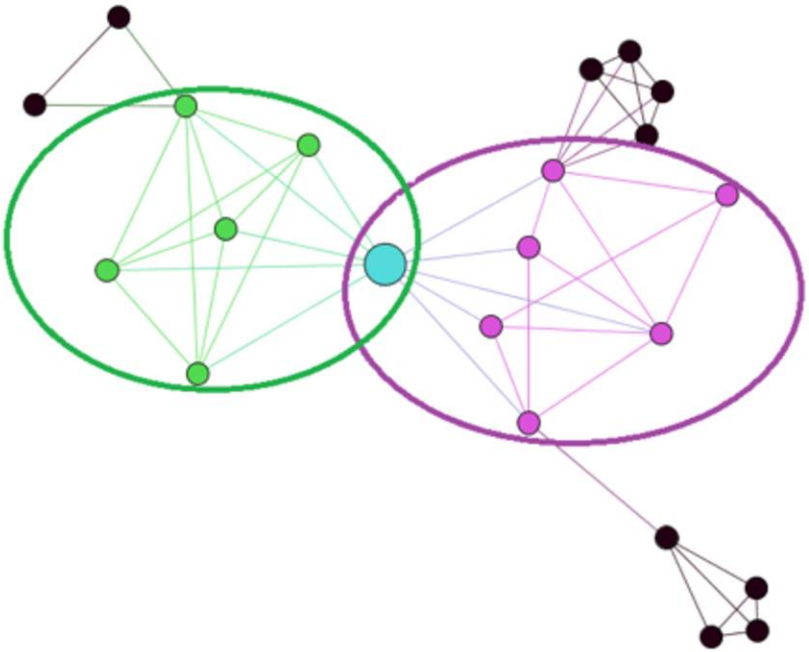

18
1
0
4
3
2
5
6
7
8
9
10
11
12
13
14
15
16
17
20
19

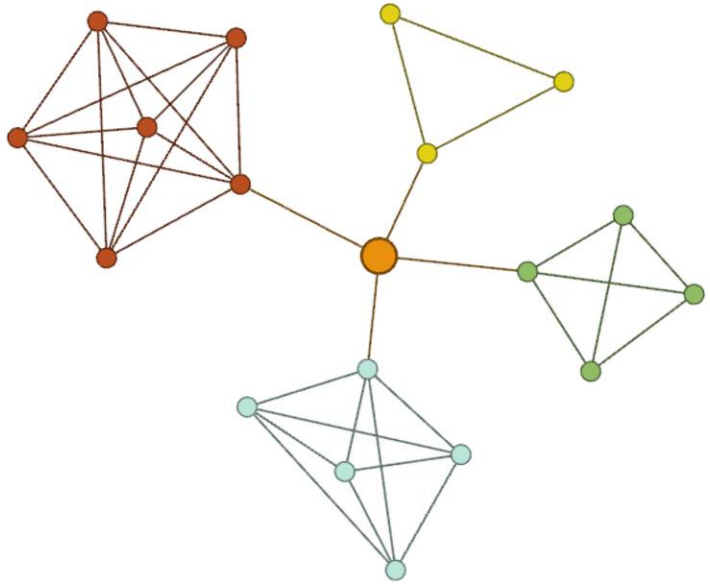

18
1
0
4
3
2
5
6
7
8
9
10
11
12
13
14
15
16
17
20
19

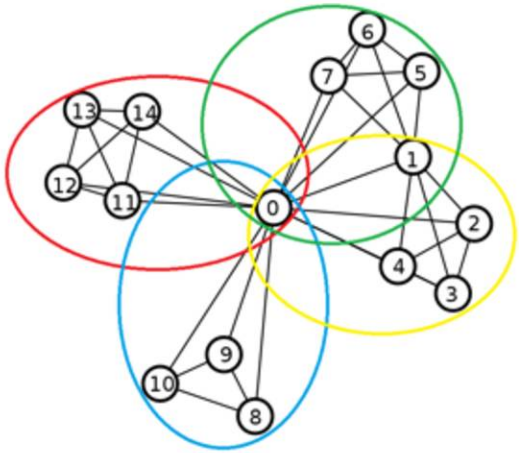
Ici 2 communautés locales recouvrantes autour du nœud 6



Autres exemples:







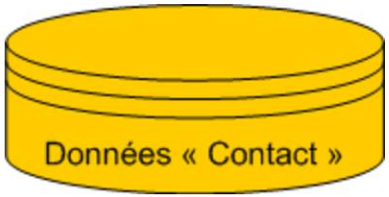
Introduction
à l'analyse des réseaux sociaux

Partie 6

Une approche simple pour la modélisation prédictive à partir de données en réseau:

extraction de variables « sociales »

Les données: des transactions ou contacts



- **Télécoms**

Nœud 1	Nœud 2	Date	Type	Durée	...
01 35 24 42 18	06 75 36 21 13	7 janvier 2010	Voix	00:04:22	
06 75 36 21 13	06 21 43 27 11	10 janvier 2010	SMS	00:00:10	
...					

- **Banque**

Nœud 1	Nœud 2	Date	Type	Montant	...
20041 10020 0058741005T 15	18206 00210 54872667002 17	2 janvier 2010	INT	345,12 €	

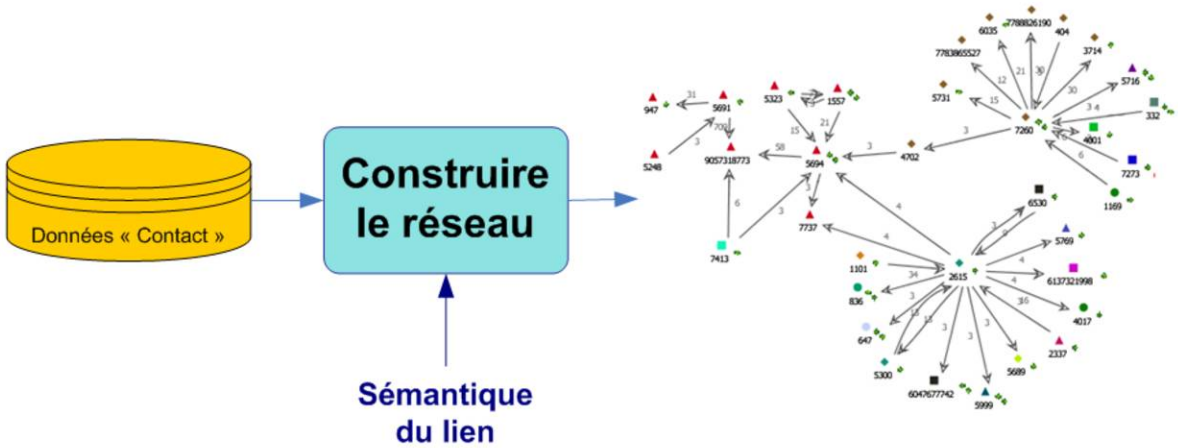
- **Site social**

Nœud 1	Nœud 2	Date	Type	Défi	...
Nathalie1369	badbad08	10 janvier 2010	Réponse	1	

Définition des liens : la « sémantique » du lien

Chaque règle donne un nouveau graphe

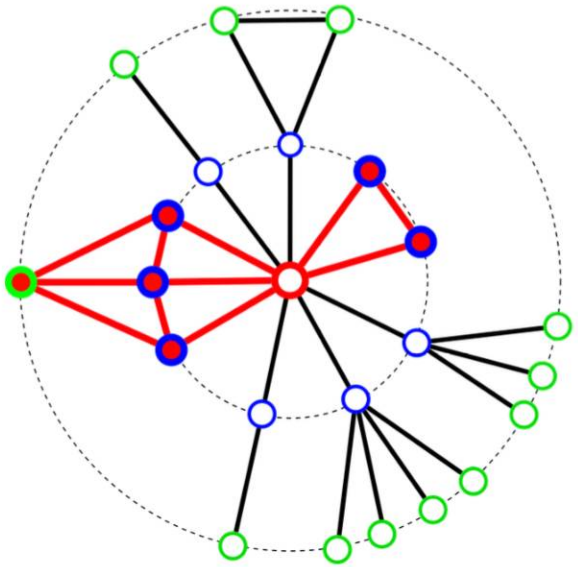
- **Exemples**
  - Telecom
    - Lien entre 2 numéros s'ils se sont appelés **au moins 4 fois**
  - Banque
    - Lien entre 2 comptes si le montant transféré est **supérieur à 500 €**
  - Site social
    - Lien entre 2 « amis » s'ils ont communiqué **au moins 3 fois**

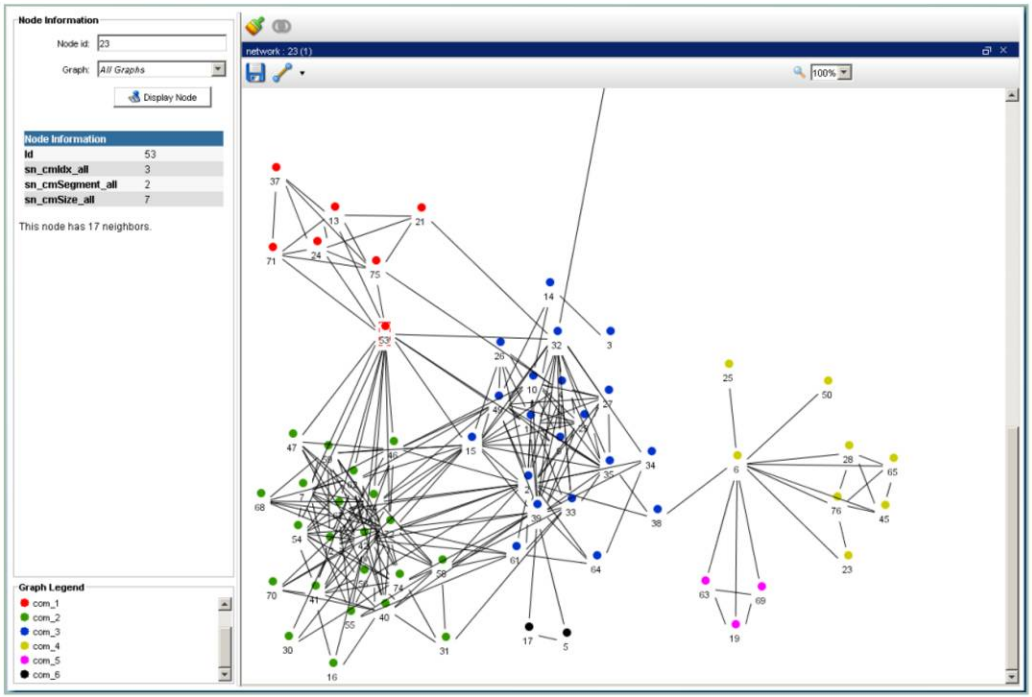


Le processus d'analyse de réseaux sociaux

Calcul des communautés

Et définition de voisinages pour le calcul d'agrégats



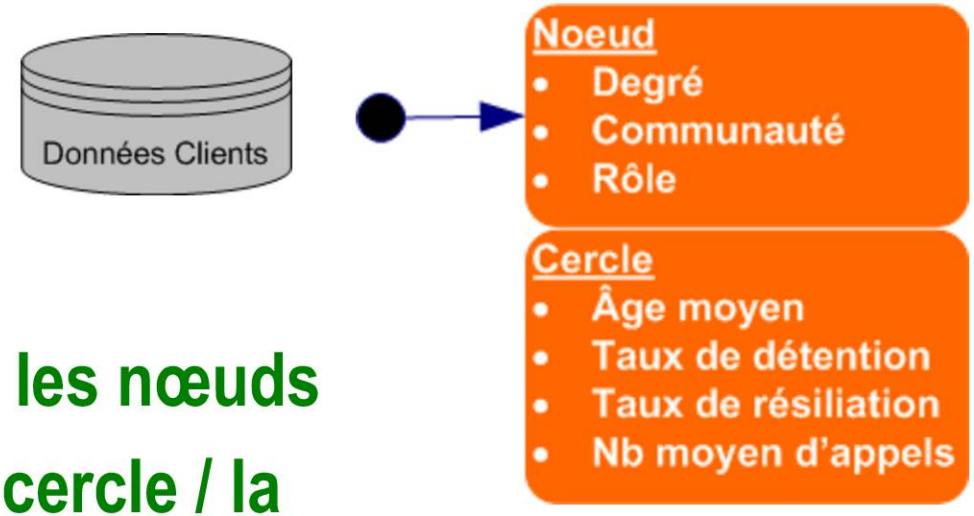


Le processus d'analyse de réseaux sociaux

Extraction des attributs sociaux

Pour chaque nœud

- Diverses métriques calculées dans le réseau



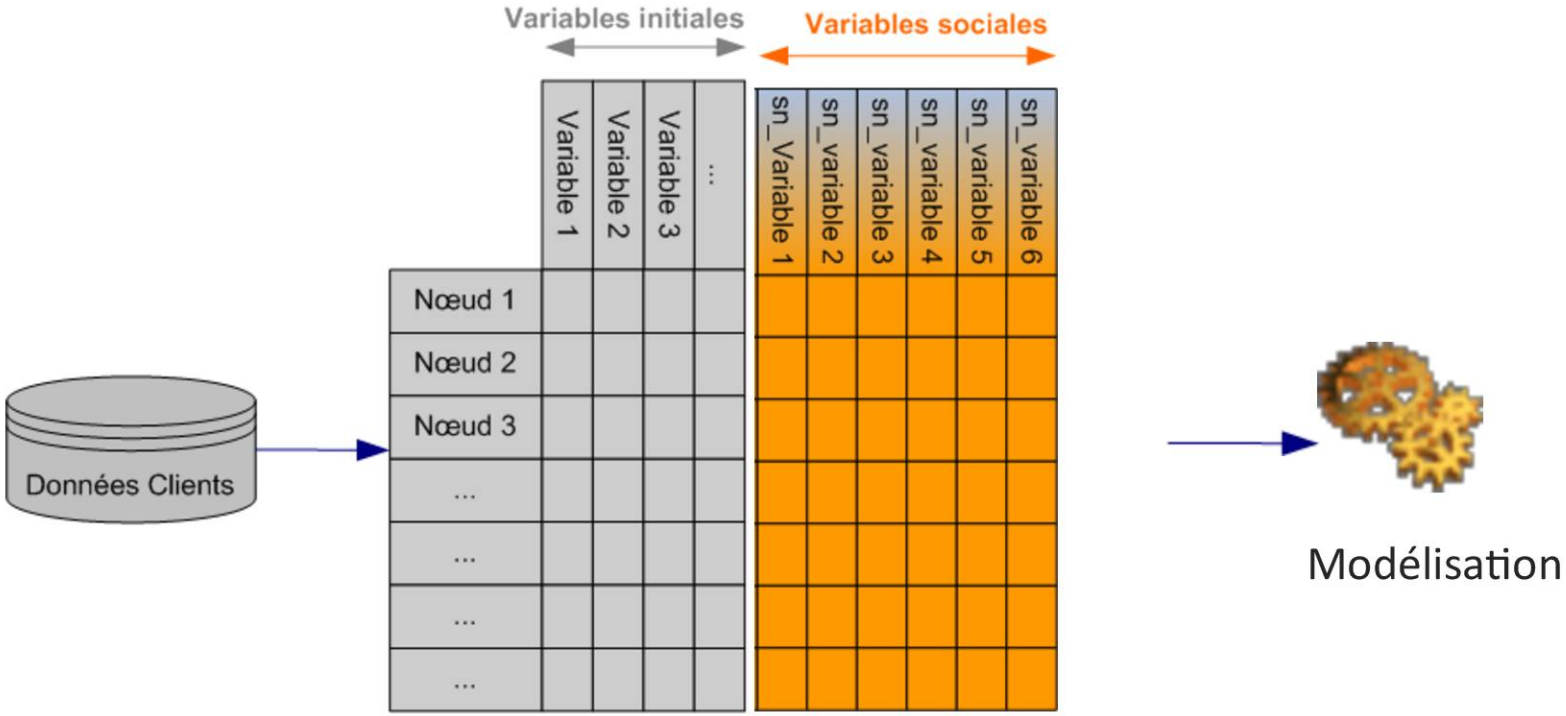
Si on dispose d'attributs sur les nœuds

- Attributs calculés dans le cercle / la communauté

Le processus d'analyse de réseaux sociaux

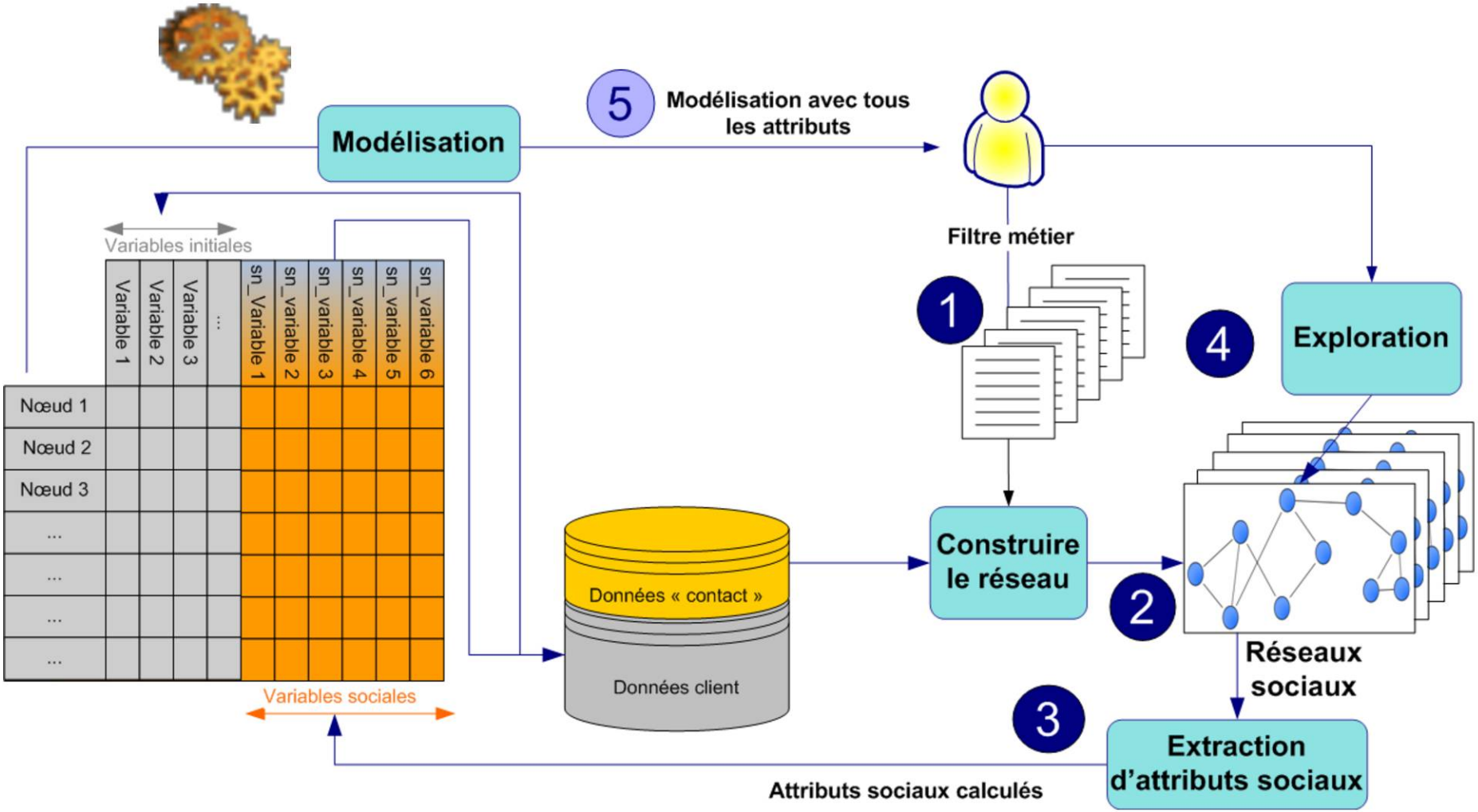
Analyse prédictive

- **On peut ensuite faire des modèles prédictifs sur l'ensemble des attributs**
  - Données clients + attributs sociaux



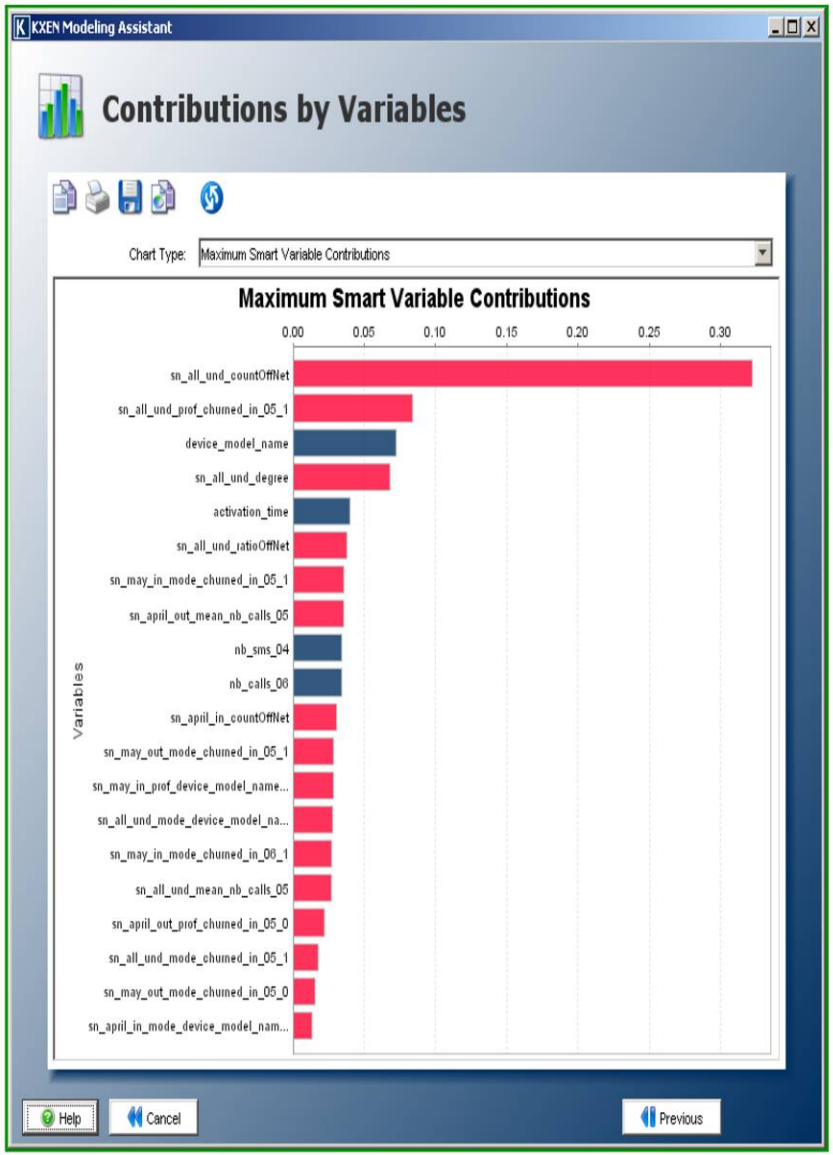
Le processus d'analyse de réseaux sociaux

- **Construction (itérative) d'un modèle prédictif**



L'apport des variables sociales

- **Variables très significatives**
- **Le modèle est (presque) toujours meilleur**



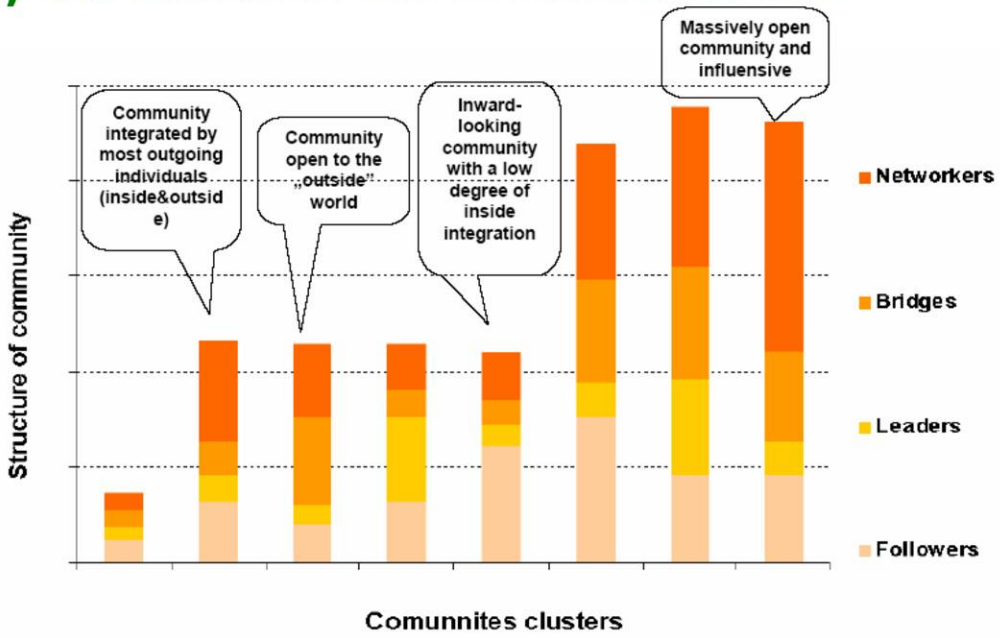
Introduction
à l'analyse des réseaux sociaux

Partie 7

Quelques exemples d'applications

Un opérateur télécom européen

- **Cet opérateur veut réduire son taux de résiliation**
- **On commence par construire le réseau et réaliser une segmentation par extraction de communautés**
  - 20 Millions de nœuds : 2 heures
- **Puis on extrait les attributs sociaux**
- **La performance (AUC) du modèle de résiliation est augmentée de 47%**



Apprentissage et test du modèle d'attrition (*churn*)

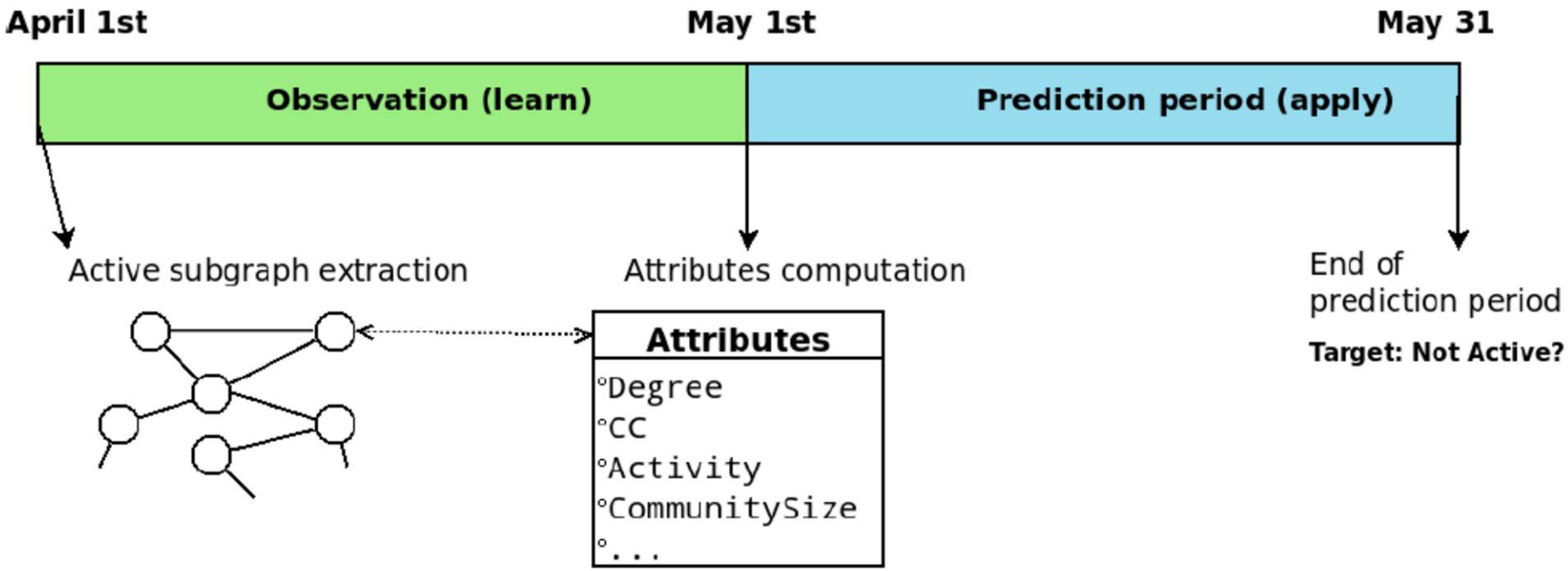
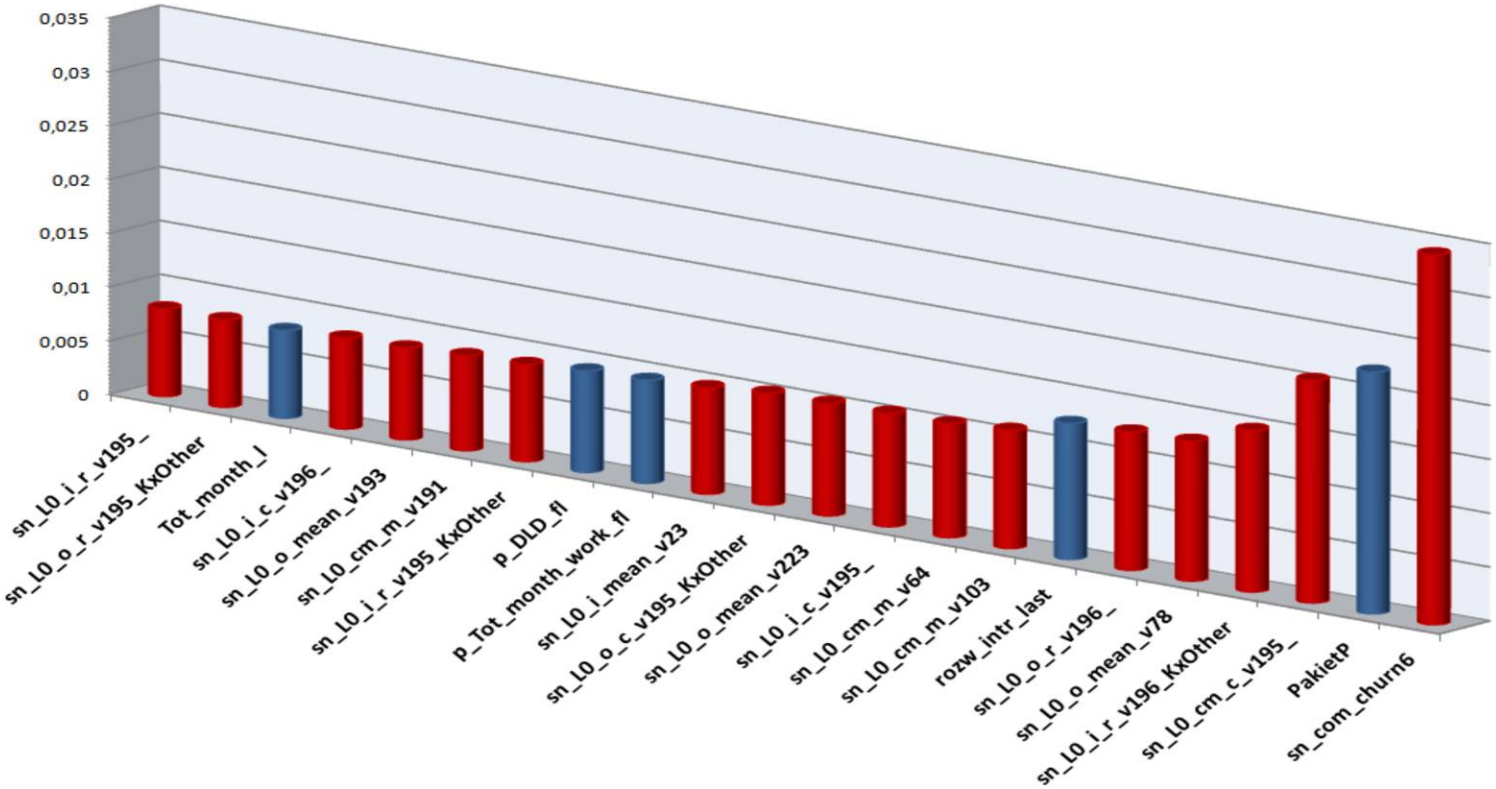


Fig. 4. Experimental setup used for churn prediction.

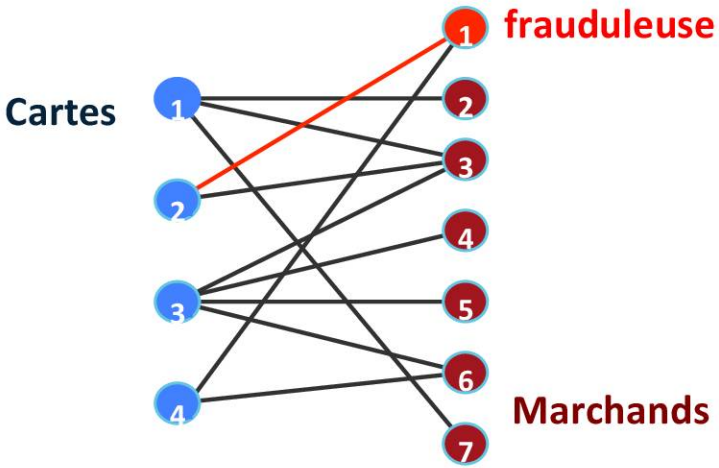
Un opérateur télécom européen

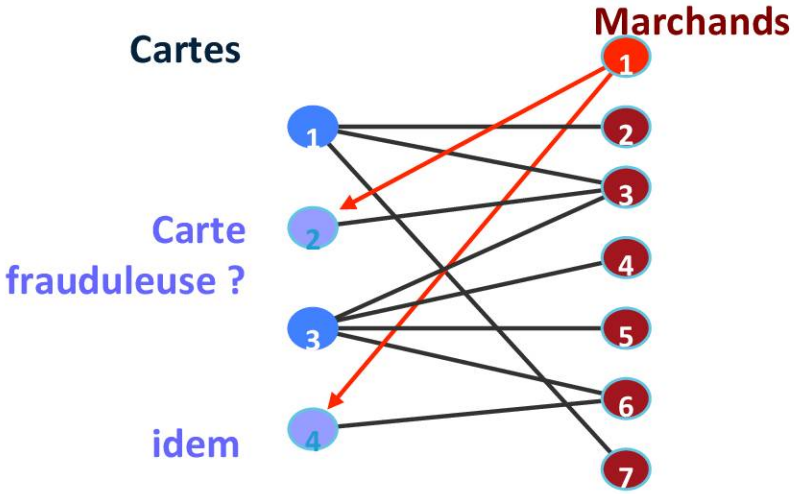
- **Top : # de «churners» dans la communauté les 6 derniers mois**
- **15 des 20 variables les plus importantes sont des <u>variables sociales</u>**



Fraude à la carte bancaire

- **Objectif**
  - Détecter la fraude à la carte bancaire
- **Méthode**
  - Construire un graphe bi-partite
    - Cartes Marchands
  - Analyser les communautés
  - Détecter les « patterns » de fraude
  - Variables sociales pour créer un modèle de détection de la fraude

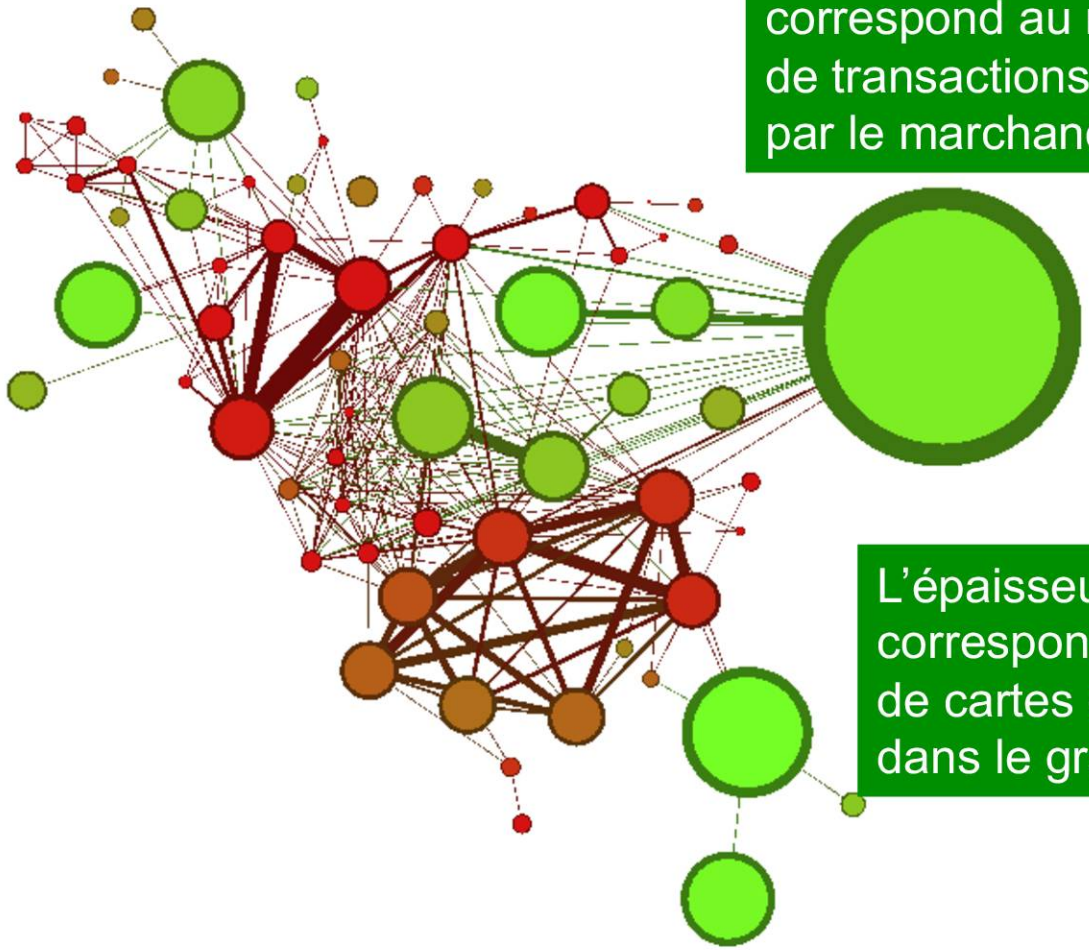




Fraude à la carte bancaire

- **Zoom sur une communauté**

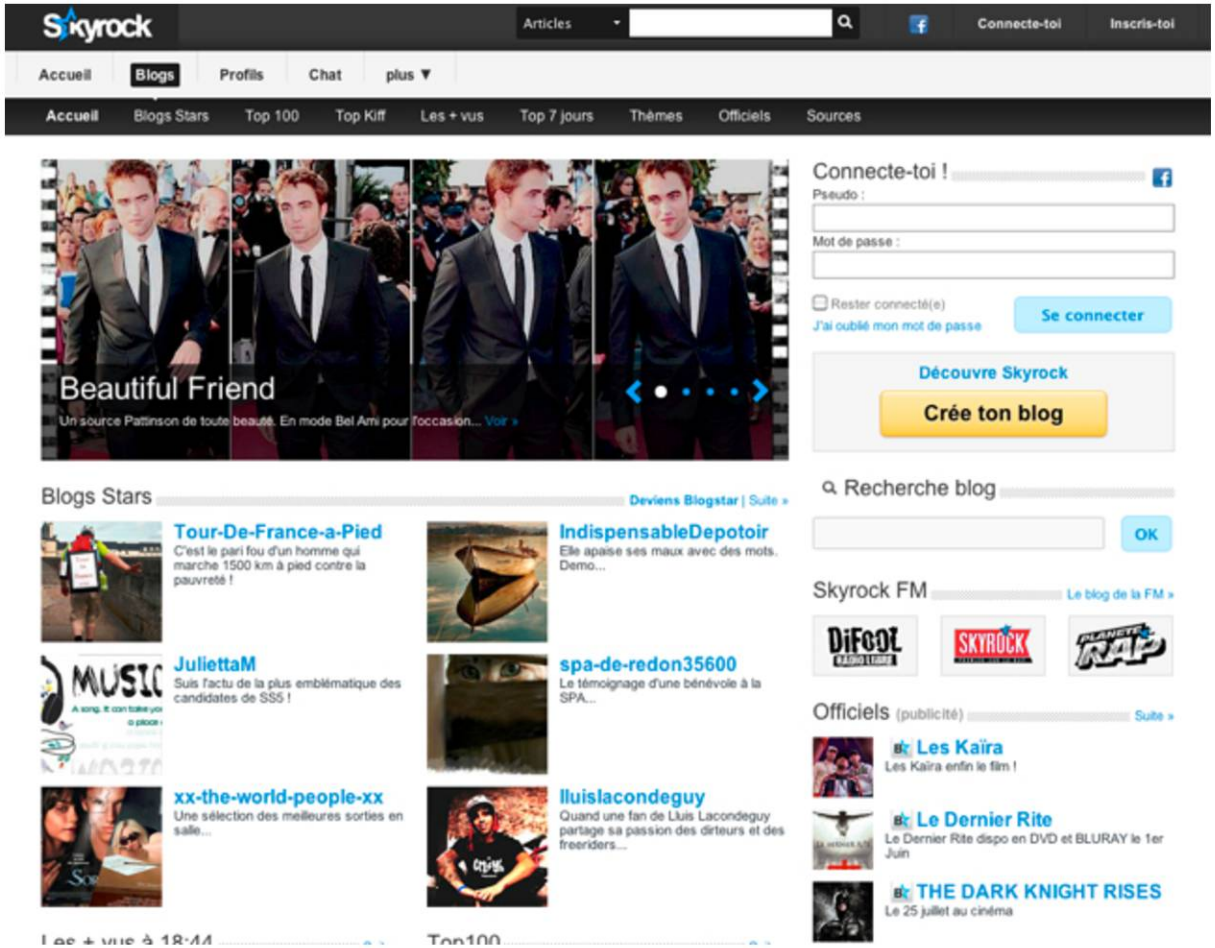
La taille d'un nœud correspond au nombre de transactions traitées par le marchand



L'épaisseur des liens correspond au nombre de cartes en commun dans le graphe bipartite

Rouge : fraude
Vert : pas fraude

Réseau social en ligne: SKYROCK Free People Network



Projet ANR ExDEUSS / CEDRES (2012)

Blog | Profil

100% lluis lacondeguy-fan source français et bientôt Español

SOURCE

lluislacondeguy

Description :

voici mon blog sur mon rider dirt slopestyle mtb (et bmx) préférée , il est espagnol , talentueux et cool ^^ . je l ai découvert il y a 5 ans est maintenant je suis une vrai fans donc voila je vous fait découvrir ma passion =) j'éspére que vous aimerait les news, les images et que vous me laisseré un pti commentaire de temps en temps ^^ , bonne visite !

aquí está mi blog en mi jinete tierra mtb slopestyle (y bmx) preferida, que es el español, con talento y ^ ^ fresco. He descubierto hace unos 5 años

bienvenue ! bienvenida!

hello tout le monde !!!

Les raisons pour lesquels j'ai crée ce blog sont diverses et varié , la première est que je suis totalement admiratif des dirteur et freerider en général , je trouve que toute les figures qu'il font sont totalement bluffante . J aime particulierment Lluis car il me semble etre un gars super simple et extra doué et j'adore sa ! donc voila de

Infos

Création : **28/07/2009 à 14:01**
Mise à jour : **Hier à 12:08**
76 articles
80 commentaires
199 amis
16 favoris
405 kiffs

Son morceau préféré

RIZZLE KICKS
Traveller's Chant -Rizzle Kicks (Stereo Typical)

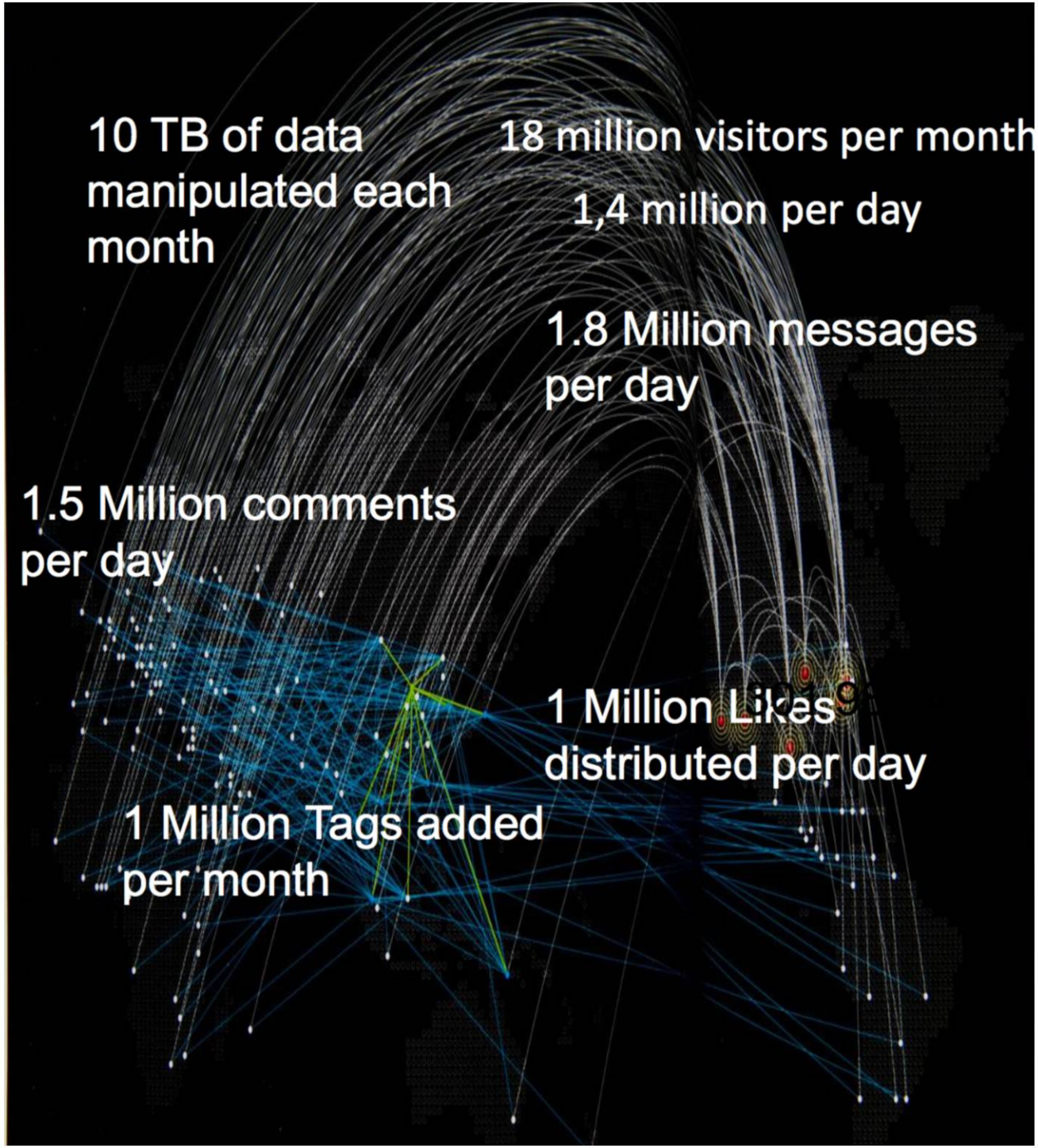
Ajouter à mon blog

Skyrock Music

Tags

... 2012 andreu lacondeguy artbmx autriche bienve aquado alba facebook page fans FISE 2011 lacondeguy lluis lluis lacondeguy master of dirt new world disorder patrick quirnez photoshoot sam reynolds sebas romero photography ti ellis

» Suite

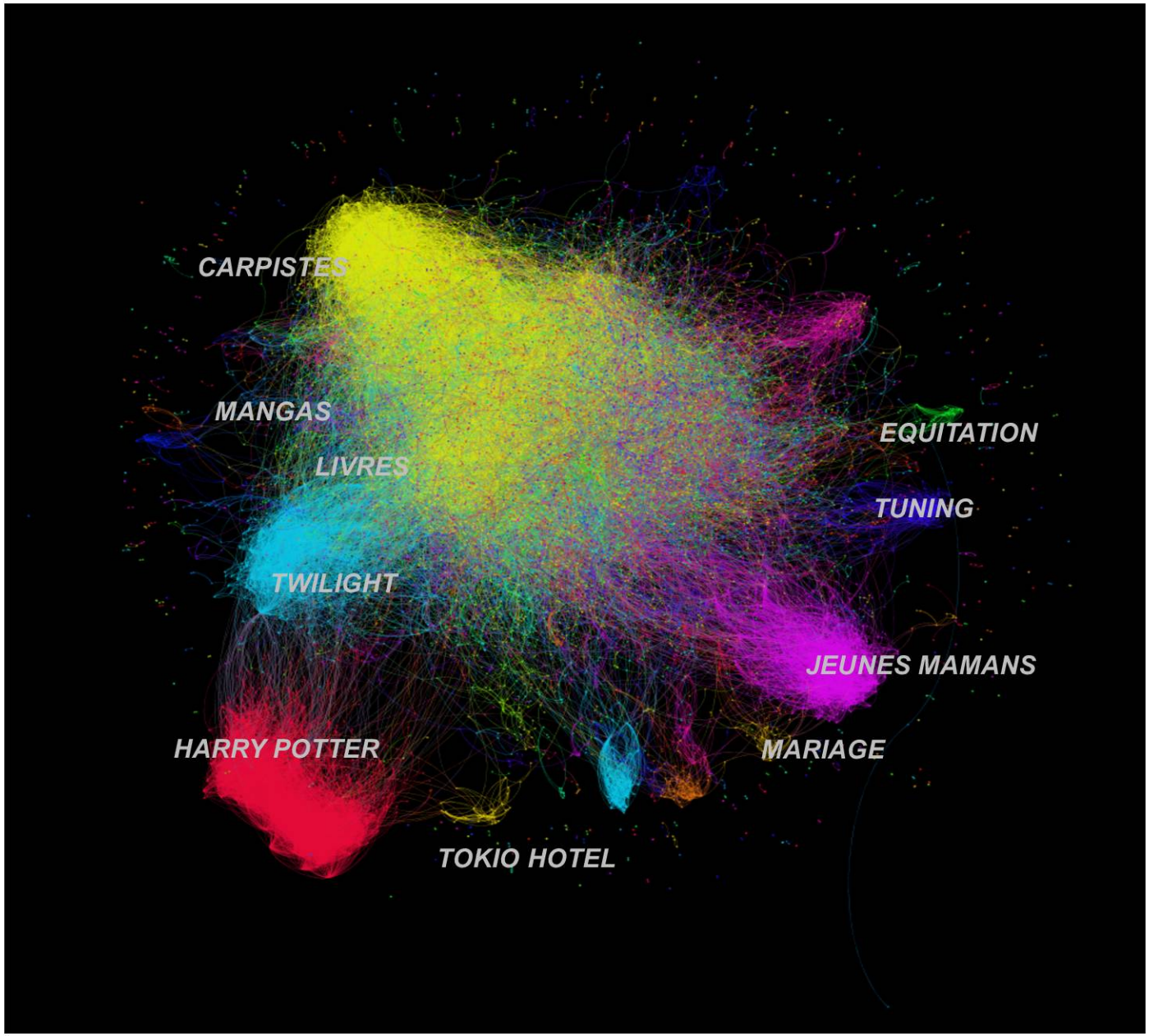


SKYROCK
Free People Network

Comment valoriser le site ?

- Cibler les publicités
  - Augmenter le taux de clicks
- Identifier les « influenceurs »
  - Les récompenser
    - Visibilité, cadeaux …
- Augmenter les interactions et l'activité
  - Recommandation d'amis
  - Animation thématique
  - Prévision de l'attrition (*churn*)

SKYROCK
Free People Network



Plus de 150 000 communautés

Ici étiquetées par l'analyse des pseudos des utilisateurs

Exemple: détection de thèmes à partir des *tags*

On crée un réseau *bipartite* utilisateurs x *tags*

Projeter côté étiquettes:

- Deux étiquettes sont connectées si elles apparaissent ensembles sur un utilisateur
- Similarité entre étiquettes

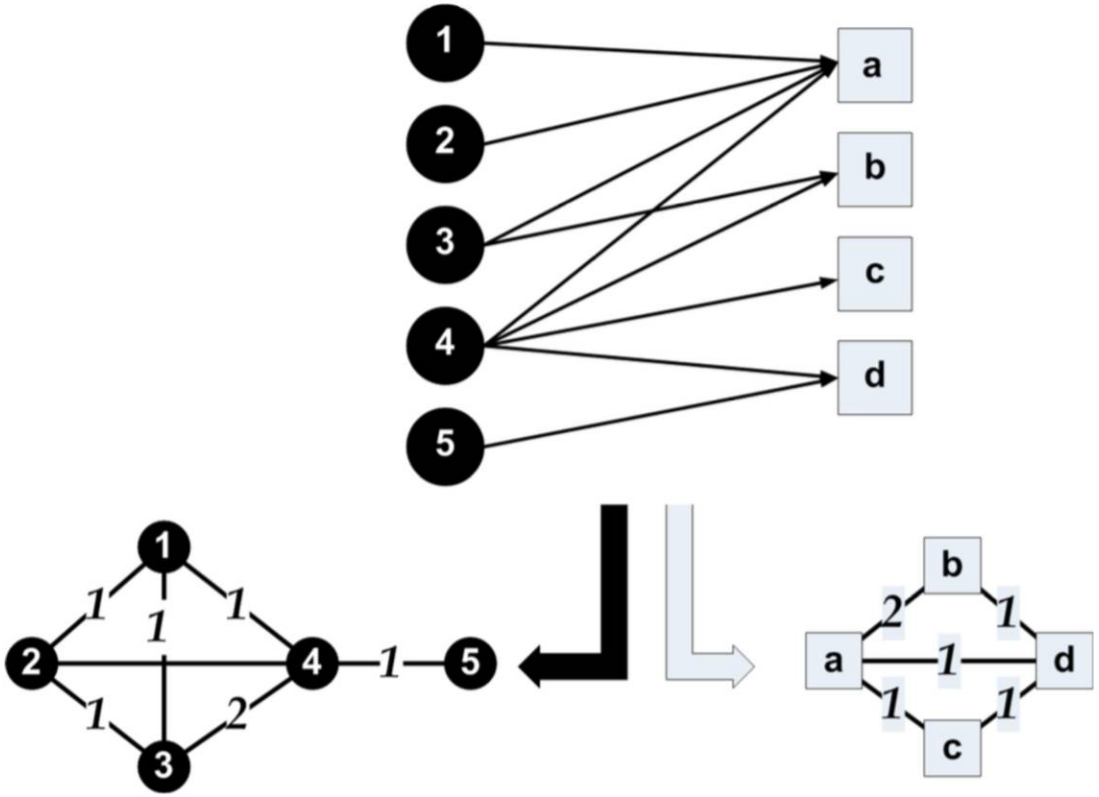
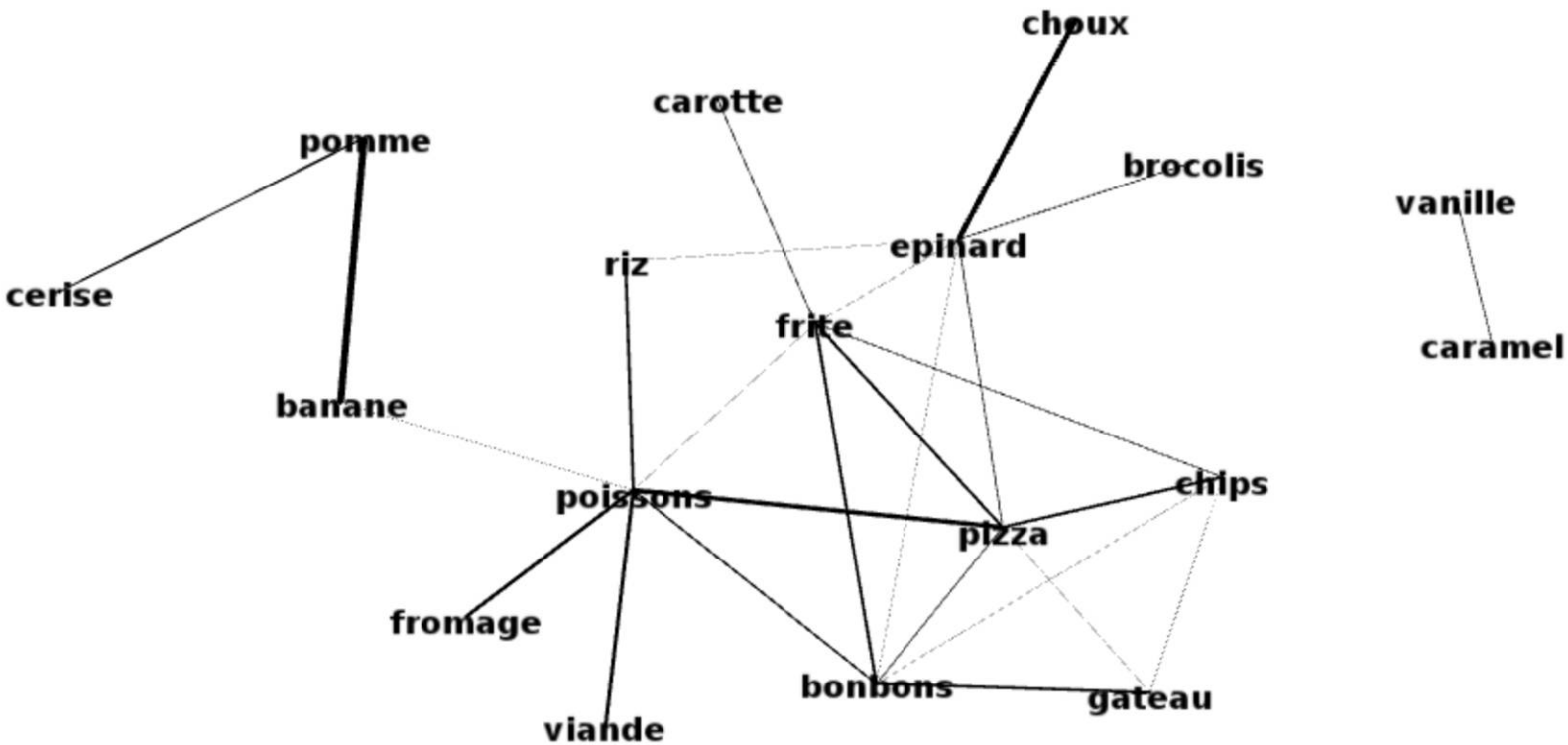


FIGURE 3. Un réseau bipartite (en haut) et les deux réseaux unipartites projetés (en bas), où on a indiqué le nombre k ou k' de voisins en commun.

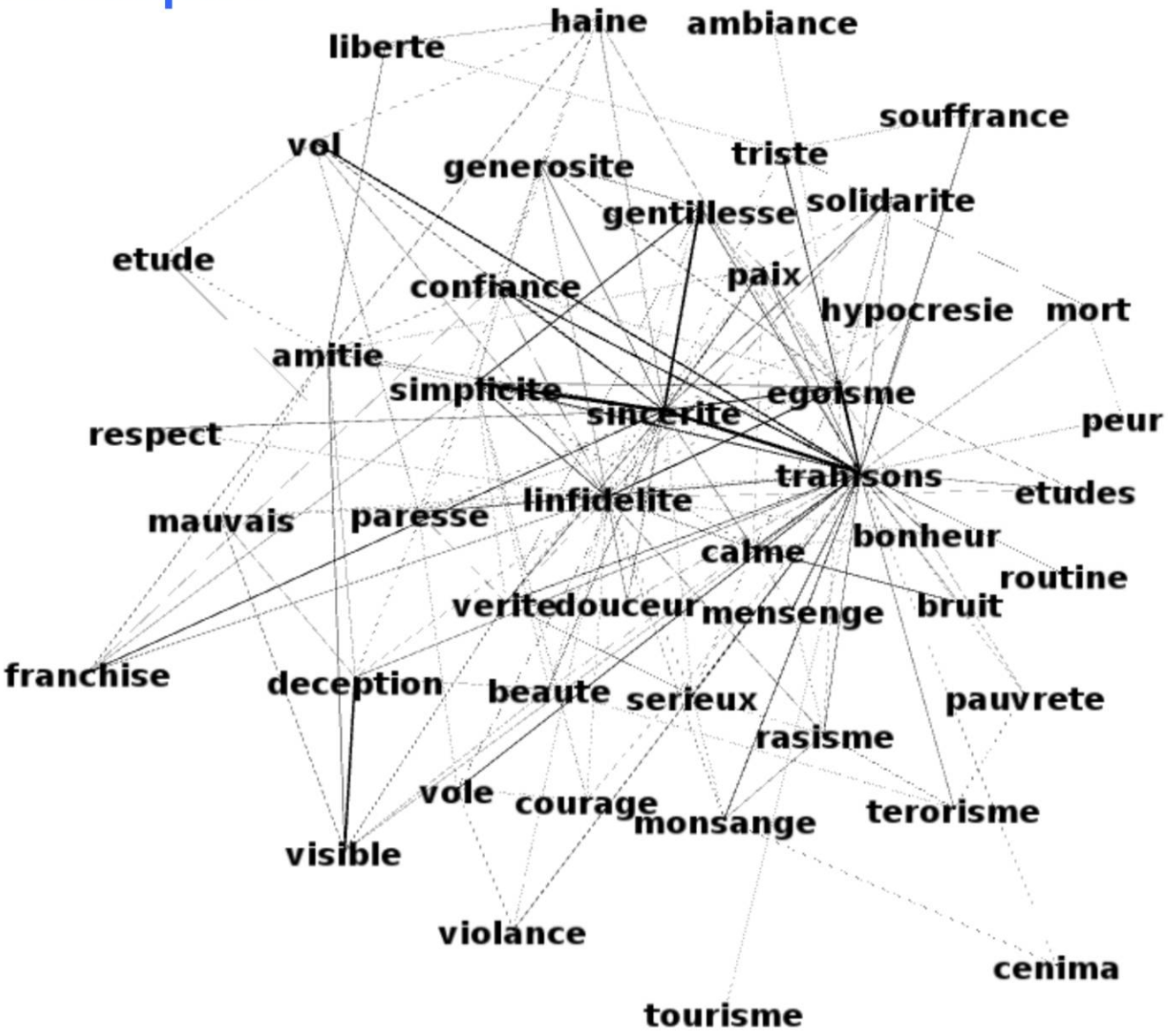
⚠️ Attention à l'implémentation des calculs si très grands graphes !

Communauté de tags skyrock: un exemple



[Dang & Viennet 2012]

Autre exemple:

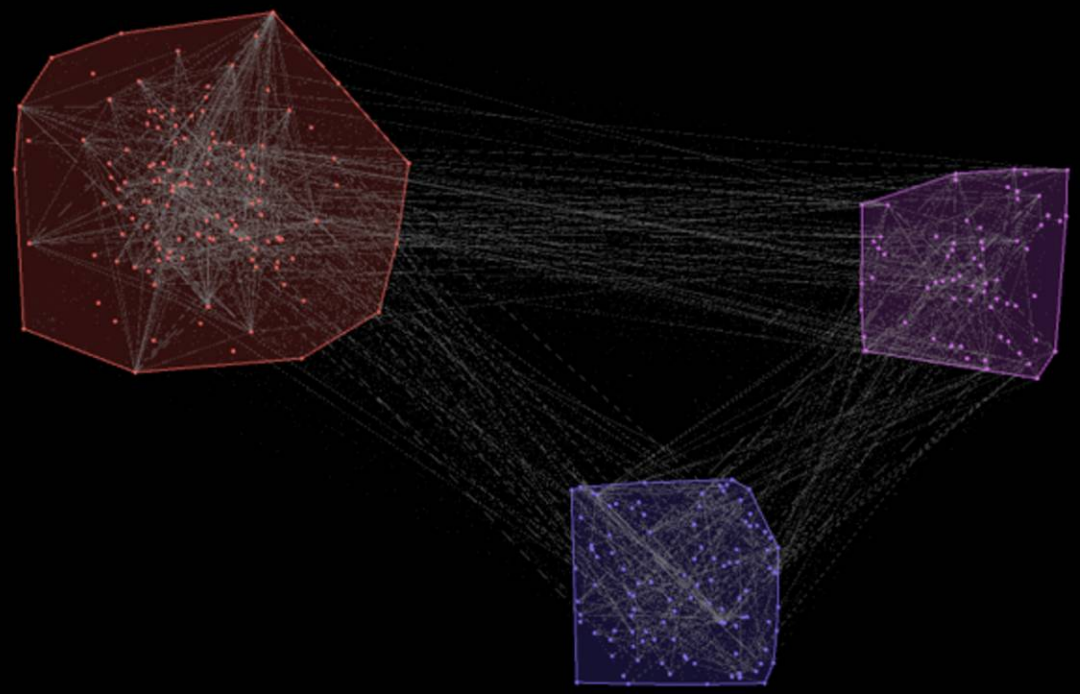


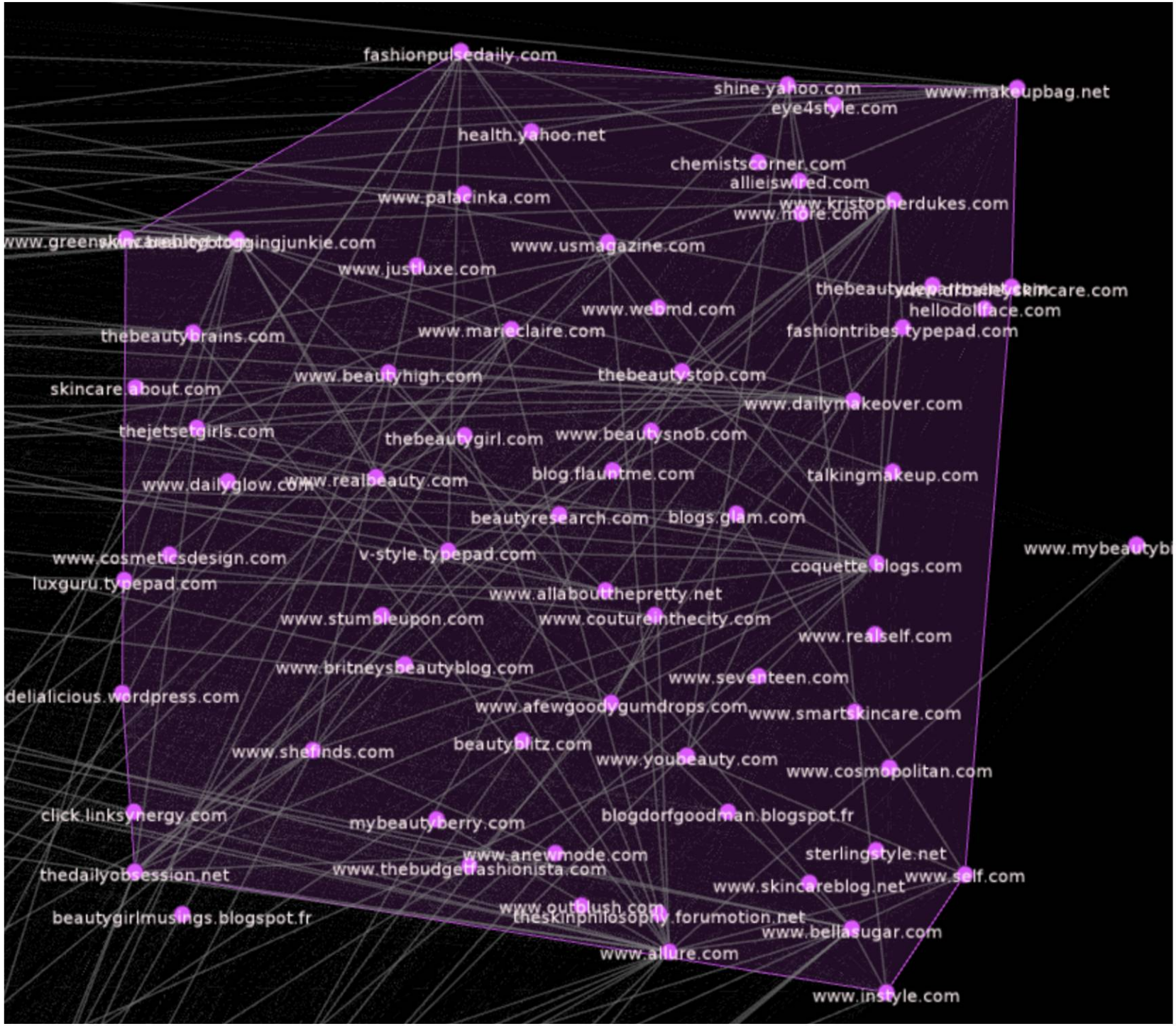
Une autre application: réseaux de sites web (affiliés à des marques)

heaven conseil

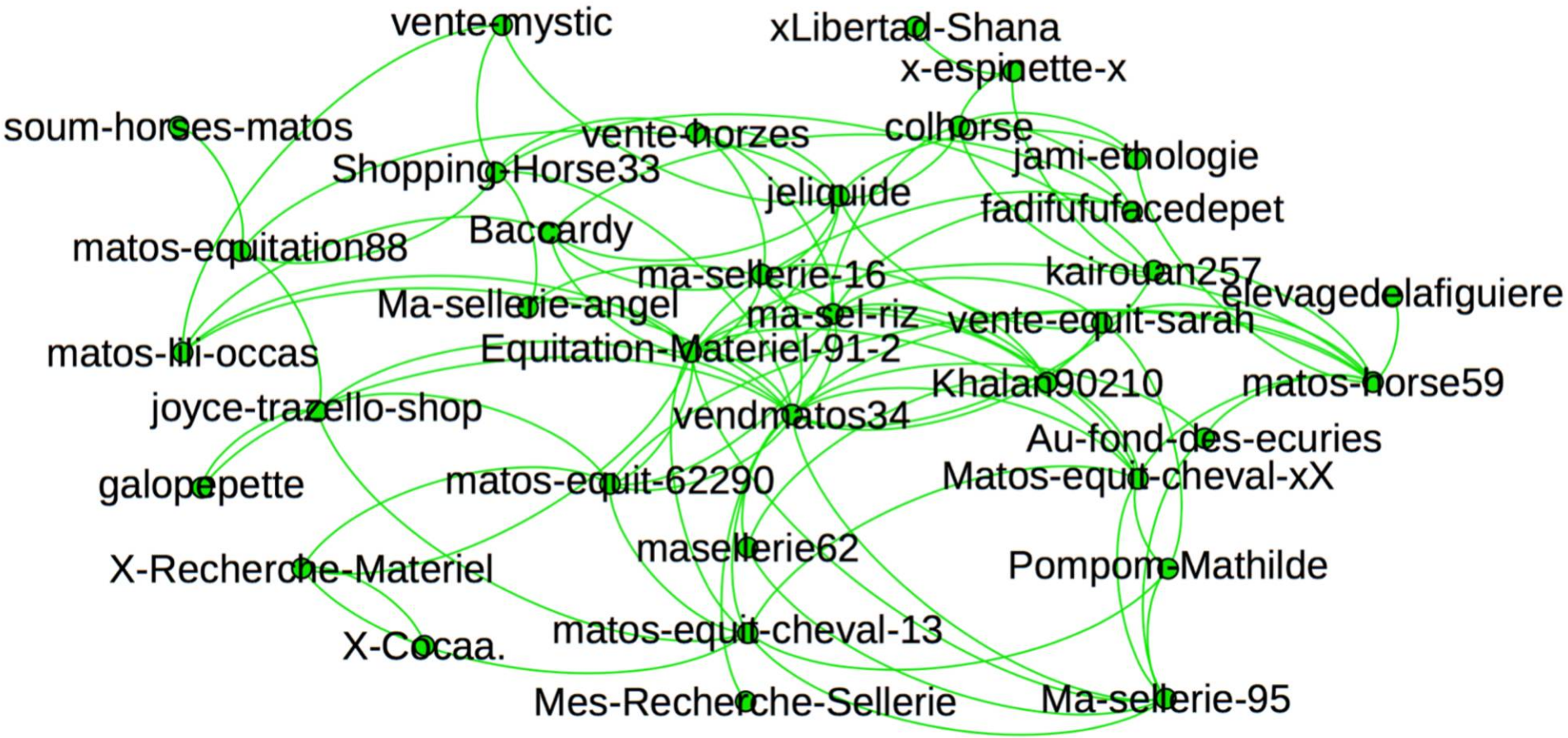
Liens (html) entre sites, et tags (sur les pages des blogs)

=> graphe **avec attributs**, que l'on segmente




fashionpulsedaily.com
shine.yahoo.com
www.makeupbag.net
eye4style.com
health.yahoo.net
chemistscorner.com
allieiswired.com
www.palacinka.com
www.kristopherdukes.com
www.more.com
www.usmagazine.com
www.justluxe.com
www.webmd.com
hellodollface.com
thebeautybrains.com
www.marieclaire.com
fashiontribes.typepad.com
www.beautyhigh.com
thebeautystop.com
skincare.about.com
www.dailymakeover.com
thejetsetgirls.com
thebeautygirl.com
www.beautysnob.com
www.dailyglow.com
www.realbeauty.com
blog.flauntme.com
talkingmakeup.com
beautyresearch.com
blogs.glam.com
www.cosmeticsdesign.com
v-style.typepad.com
coquette.blogs.com
luxguru.typepad.com
www.allaboutthepretty.net
www.stumbleupon.com
www.coutureinthecity.com
www.realself.com
www.britneysbeautyblog.com
www.seventeen.com
delialicious.wordpress.com
www.afewgoodygumdrops.com
www.smartskincare.com
www.shefinds.com
beautyblitz.com
www.youbeauty.com
www.cosmopolitan.com
click.linksynergy.com
mybeautyberry.com
blogdorfgoodman.blogspot.fr
www.anewmode.com
sterlingstyle.net
thedailyobsession.net
www.thebudgetfashionista.com
www.self.com
www.skincareblog.net
www.outblush.com
beautygirlmusings.blogspot.fr
theskinphilosophy.forumotion.net
www.bellasugar.com
www.allure.com
www.instyle.com

Les communautés d'amis Skyrock : «équitation»



Les communautés d'amis Skyrock

FIGURE 6. Tags associés à deux communautés d'intérêt autour de la marque Adidas.

- visualisation du réseau
- identification des communautés
- qualification des communautés
- meilleure compréhension de nos utilisateurs
- capacité à identifier des sujets émergents

En résumé:

- Données en réseaux (non iiid !) présentes partout
- Les réseaux complexes partagent des propriétés structurelles étonnantes;
- Nombreuses applications industrielles, impliquant des données hétérogènes de très grandes tailles
- Il existe des algorithmes d'analyse quasi-linéaires
- L'extraction de « variables sociales » permet de construire des modèles simples et performants.

Suites possibles de cette introduction

- Prise en compte de la dynamique
- Modèles de propagation
- Recommandation sociale (vendredi matin)