Analyse de réseaux 101

(Avec Python) – Émilien Schultz

Rappel sur l’analyse de réseau

Données relationnelles : **tout peut être un réseau**

- étape centrale de modélisation

Ensuite, l’enjeu est de se doter des outils pour

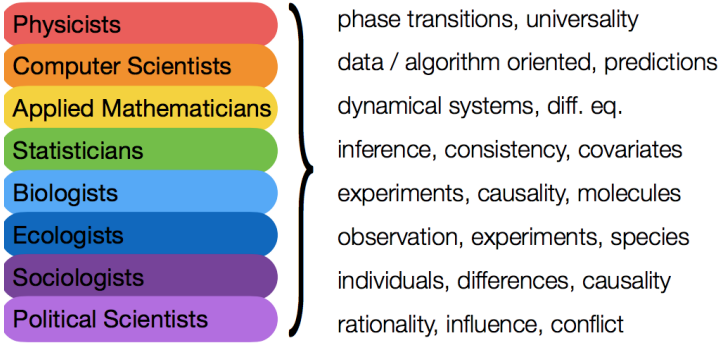
1/ construire le réseau

2/ le manipuler

3/ obtenir des résultats

Un domaine interdisciplinaire

domain	*network*	*vertex*	*edge*
biological	metabolic network	metabolite	metabolic reaction
	protein-interaction network	protein	bonding
	gene regulatory network	gene	regulatory effect
	drug interactions	drug	*in vivo* health interaction
	connectome	neuron	synapse
	physiology	muscles and bones	physical attachment
	pollination network	plants and pollinators	pollination
	food web	species	predation or resource transfer
social	friendship network (offline)	person	friendship, trust, etc.
	friendship network (online)	account	"friendship," follow, etc.
	proximity network	person	physical proximity
	sexual network	person	intercourse
	coauthorships	authors	collaboration
	fictional	character	co-appearance
	animal behavior	animals	interaction
economic	hiring network	workers and jobs	hired into
	international trade	country	trade flow
	purchasing	users and items	purchased
	board of directors	directors and boards	sits on
	inventions	inventors and patents	authored
information	software	function	function call
	World Wide Web	web page	hyperlink
	documents	article, patent, legal case	citation
	artifacts	item, document, concept	relatedness or similarity
	language	word	adjacency in text
technological	Internet (1)	computer	IP network adjacency
	Internet (2)	autonomous system (AS)	GBP connection
	digital circuits	logic gates	wire
	power grid	generating or relay station	transmission line
transportation	rail system	rail station	railroad tracks
	road network (1)	intersection	pavement
	road network (2)	named road	intersection
	airport network	airport	non-stop flight



Exemple des collaborations scientifiques

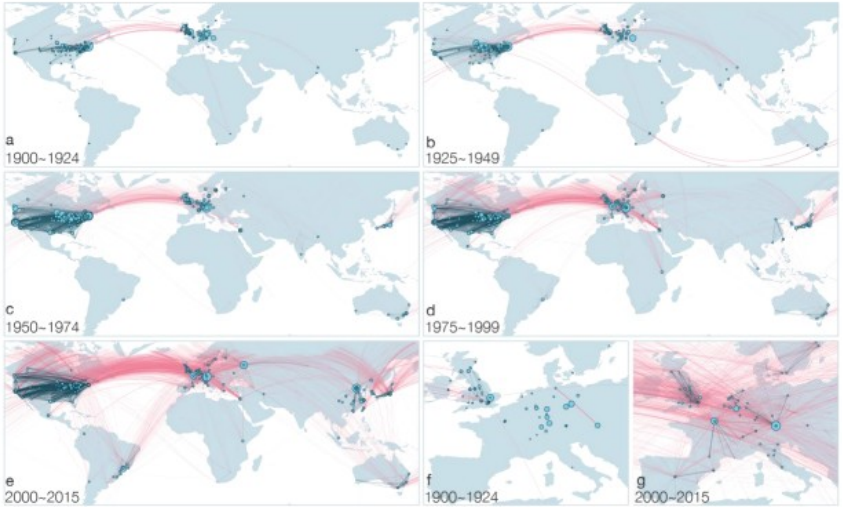


Figure 2.1.2 **A brief history of globalization of science collaborations**. The figure shows two types of collaborations: across institutions within the same country (green) and across institutions of different countries (blue). Between 1900–1924, collaborations across different institutions were prominent only i the US; international collaborations were mainly between the U.S. and the U.K. However, both types of collaborations were relatively weak. Between 1925–1949, international collaborations started to form between India and the U.K., as well as between Australia and the U.S. Due to World War II, and collaborations in Europe shrank during this period. Meanwhile, collaborative relationships in America were rapidly developing in the Midwest. Between 1950–1974, Israel and Japan started to engage in international teamwork. At the same time, the West Coast and the Southern United States became hubs scientific collaboration. Between 1975–1999, Africa began to develop relationships with Europe. Surprisingly, within-institution collaborations in the U.S. decreased relative to those in Europe, althoug the absolute number of collaborations grew substantially over time for all countries. In the 21st century, more and more countries have risen to the international stage to collaborate extensively with others. Aft Dong 2017 [13].

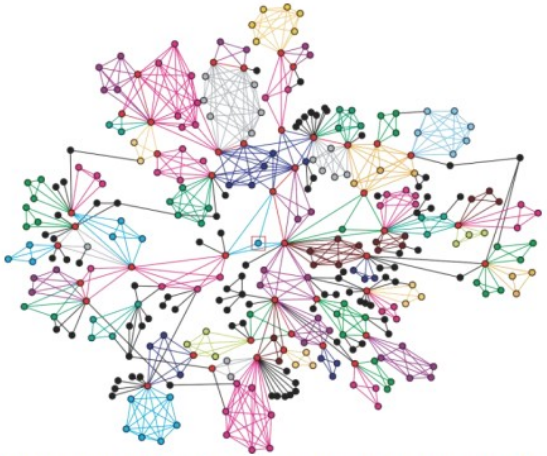
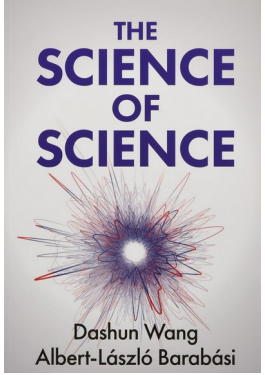


Figure 2.3.1 **Co-authorship network.** The figure shows the local structure of the co-authorship network between physicists in the vicinity of a randomly chosen individual (marked by a red frame). The network is constructed based on papers from Cornell University's archive server (cond-mat), the precursor of the widely used arXiv, containing at that time over 30,000 authors. Each node is a scientist, and links document collaborative relationships in the form of co-authored publications. The color-coded communities represent groups of collaborators that belong to locally densely interconnected parts within the network. Black nodes/edges mark those scientists that do not belong to any community. After Palla *et al.* [34].



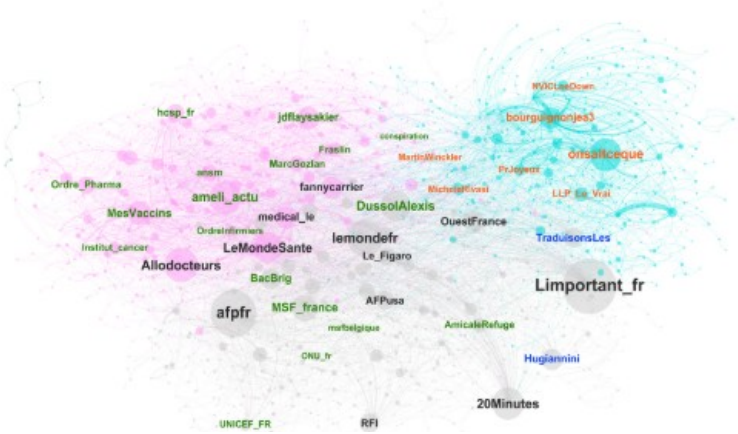
https://github.com/marcessz/Twenty-Years-of-Network-Science

Exemples : médias sociaux en santé publique

Asymmetric participation of defenders and critics of vaccines to debates on French-speaking Twitter

Floriana Gargiulo[1,3], Florian Cafiero[1], Paul Guille-Escuret[1,2], Valérie Seror[2] & Jeremy K. Ward[1,2,3*]

For more than a decade, doubt about vaccines has become an increasingly important global issue. Polarization of opinions on this matter, especially through social media, has been repeatedly observed, but details about the balance of forces are left unclear. In this paper, we analyse the flow of information on vaccines on the French-speaking realm of Twitter between 2016 and 2017. Two major asymmetries appear. Rather than opposing themselves on each vaccine, defenders and critics focus on different vaccines and vaccine-related topics. Pro-vaccine accounts focus on hopes for new groundbreaking vaccines and on ongoing outbreaks of vaccine-preventable illnesses. Vaccine critics concentrate their posts on a limited number of "controversial" vaccines and adjuvants. Furthermore, vaccine-critical accounts display greater craft and energy, using a wider variety of sources, and a more coordinated set of hashtags. This double asymmetry can have serious consequences. Despite the presence of a large number of pro-vaccine accounts, some arguments raised by efficiently organized and very active vaccine-critical activists are left unanswered.



Id	0
Allodocteurs	MEDIA
ameli_actu	PRO
LeMondeSante	MEDIA
MesVaccins	PRO
jdllaysakier	PRO
fannycarrier	MEDIA
medical_le	MEDIA
hcsp_fr	PRO
Ordre_Pharma	PRO
ansm	PRO
MarcGozlan	PRO
moniteur_news	MEDIA
Fraslin	PRO
Topsantecom	MEDIA
Anthony_ANEPF	PRO
Institut_cancer	PRO
leQPH_fr	MEDIA
LeFigaro_Sante	MEDIA
APMinfos	MEDIA
OrdreInfirmiers	PRO

Id	0
Limportant_fr	MEDIA
afpfr	MEDIA
lemondefr	MEDIA
20Minutes	MEDIA
DussolAlexis	PRO
MSF_france	PRO
OuestFrance	MEDIA
Hugiannini	OUT
BacBrig	PRO
TraduisonsLes	OUT
RFI	MEDIA
AFPusa	MEDIA
Le_Figaro	MEDIA
LesEchos	MEDIA
franceinfo	MEDIA
lemonde_sante	MEDIA
AmicaleRefuge	PRO
exingg	MEDIA
radiookapi	MEDIA
RFIAfrique	MEDIA

Id	0
onsaitceque	ANTI
bourguignonjea3	ANTI
LLP_Le_Vrai	ANTI
MicheleRivasi	ANTI
PrJoyeux	ANTI
vaccinssansalu	ANTI
NVICLoeDown	ANTI
monde_libre	Not_classified
mapoulebranchet	ANTI
lenfantmarie2	ANTI
FamillesSantePr	ANTI
WildstrikeW	Not_classified
dangersante	ANTI
NicoleDelegine	ANTI
thierrysoucoar	ANTI
FX_Olivador	ANTI
carcajou76865	ANTI
libreactus	Not_classified
LaLaReneFrench75	ANTI
Mains_Propres	ANTI

Figure 3. Layers of the retweet network for four categories (ADJUVANTS, SEASONAL FLU, AWAITED and OTHER vaccines) and 3 controversial subjects (HPV, Measles, Hepatitis B). The color of the nodes corresponds to the classification of the users: "ANTI" (orange), "PRO" (green), "MEDIA" (dark grey), "NEUTRAL" (cyan), "un-classifiables" (dark blue) and not coded (light grey).

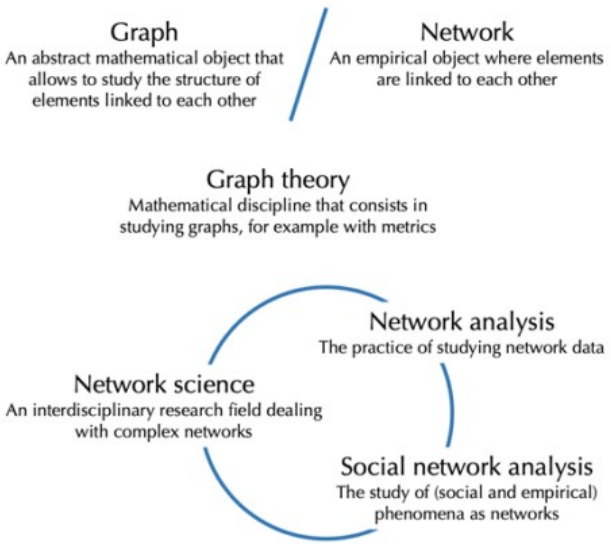
Théorie des graphs 101 à l’infini

Les outils conceptuels liés aux réseaux peuvent aller très loin.

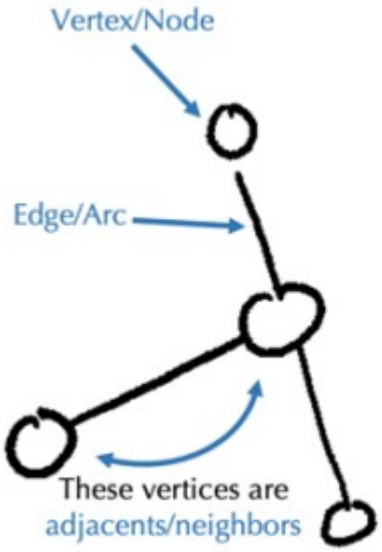
Pas le lieu ici de faire un cours spécifique sur l’analyse de réseaux

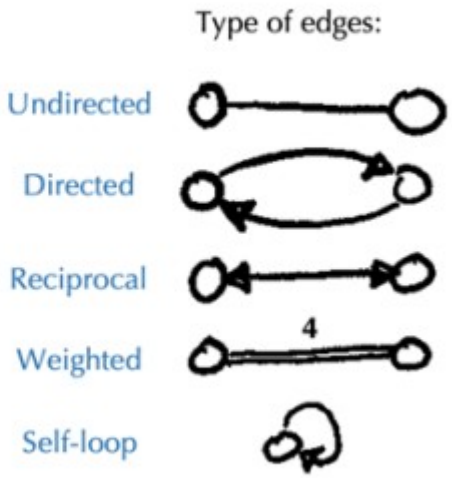
Tout un pan de théorie : les **graphs**

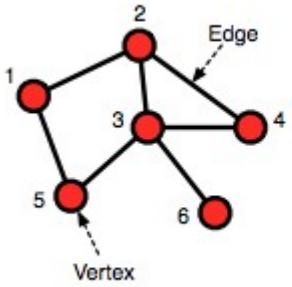
Mais il faut quelques notions : noeuds, liens

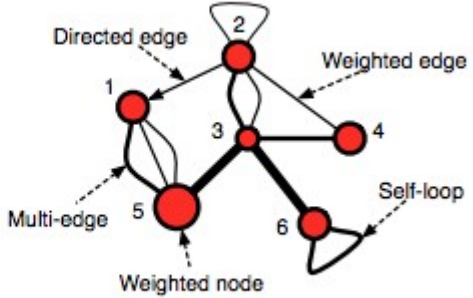


Les types de relations

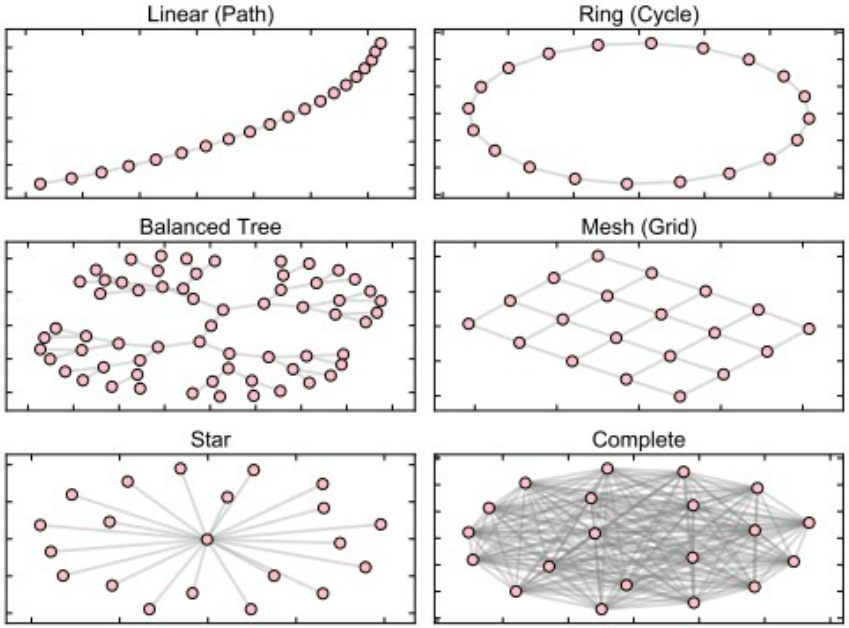


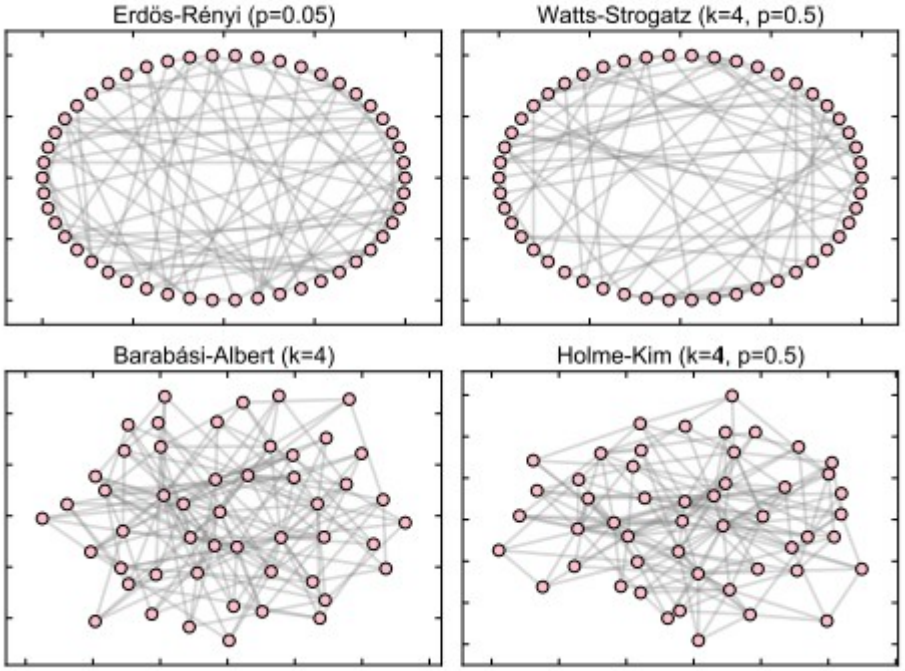






Des réseaux “de synthèse” générables





Des mesures à l’échelle générale du graph

- Type du réseau
- Nombre de noeuds, liens
- Nombre de composantes
- Diamètre
- Densité
- Distribution des degrés

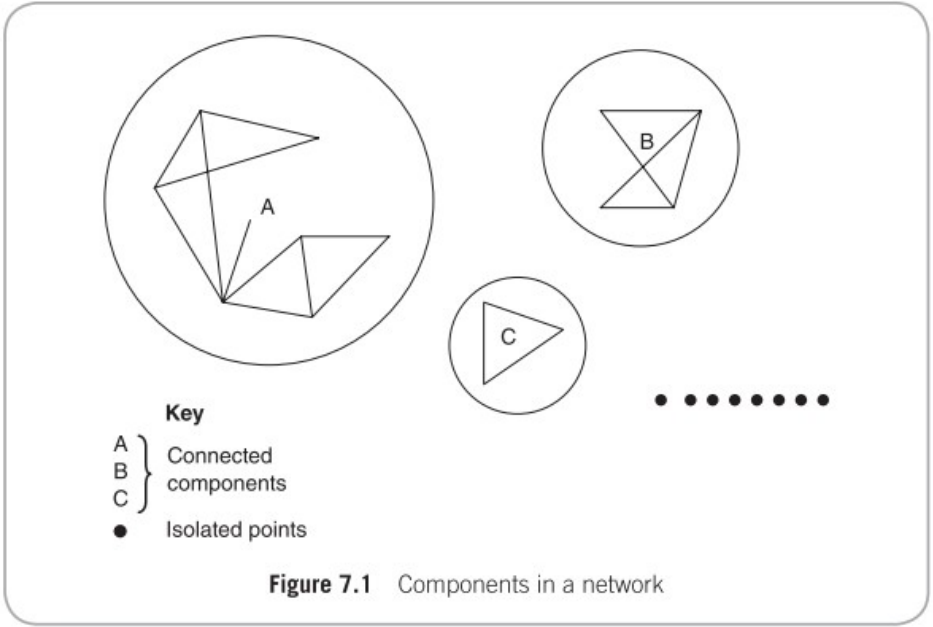
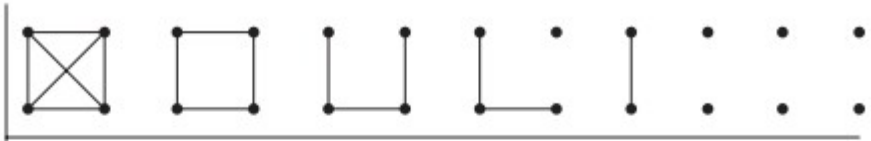


Figure 7.1 Components in a network

Des mesures à l’échelle du noeuds/liens

- degré
- nombre de voisins
- diverses dimensions ajoutées

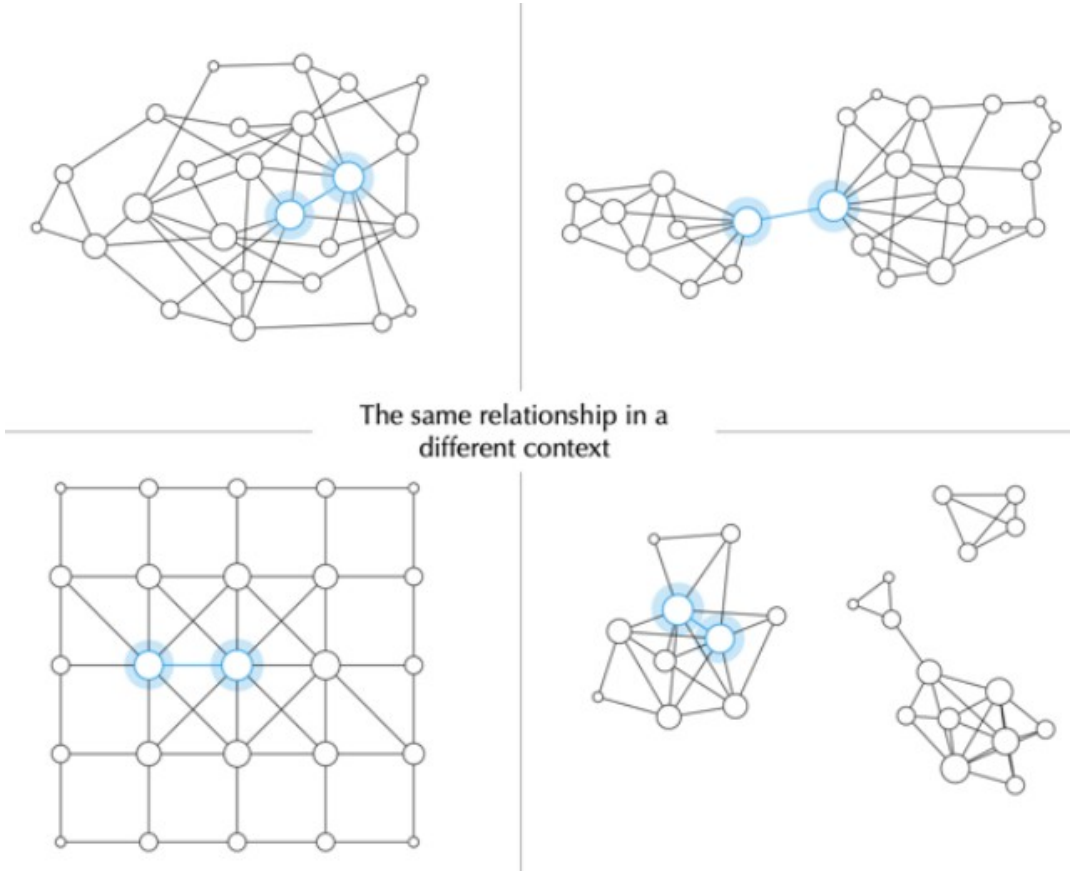
Exemples



No. of connected points	4	4	4	3	2	0
Inclusiveness	1.0	1.0	1.0	0.7	0.5	0
Sum of degrees	12	8	6	4	2	0
No. of lines	6	4	3	2	1	0
Density	1.0	0.7	0.5	0.3	0.1	0

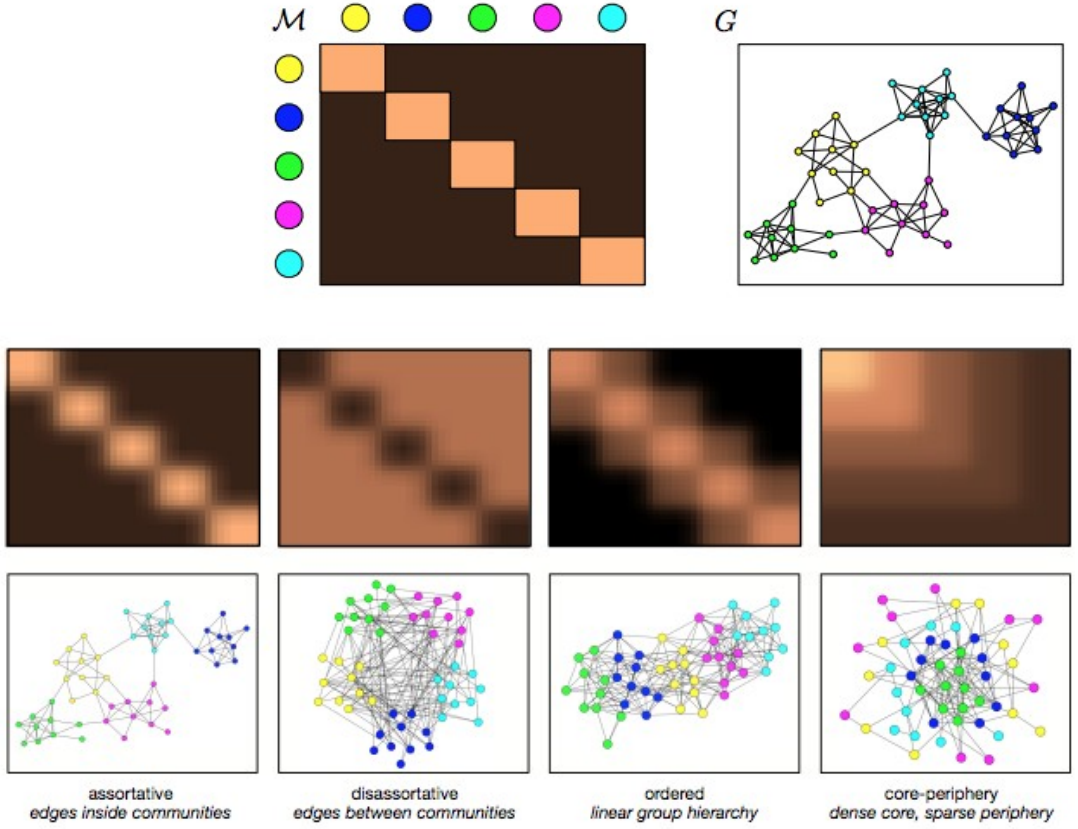
Figure 5.4 Density comparisons

Des propriétés locales : importance du contexte



Les métriques les plus utilisées

- Betweeness centrality
- Closeness centrality
- Eigenvector centrality
- Noyau/couche
- Coefficient de clustering
- Clustering



L’intérêt pour les sciences sociales

Des caractéristiques nativement relationnelles (capital social)

Des structures explicatives (circulation de l’information)

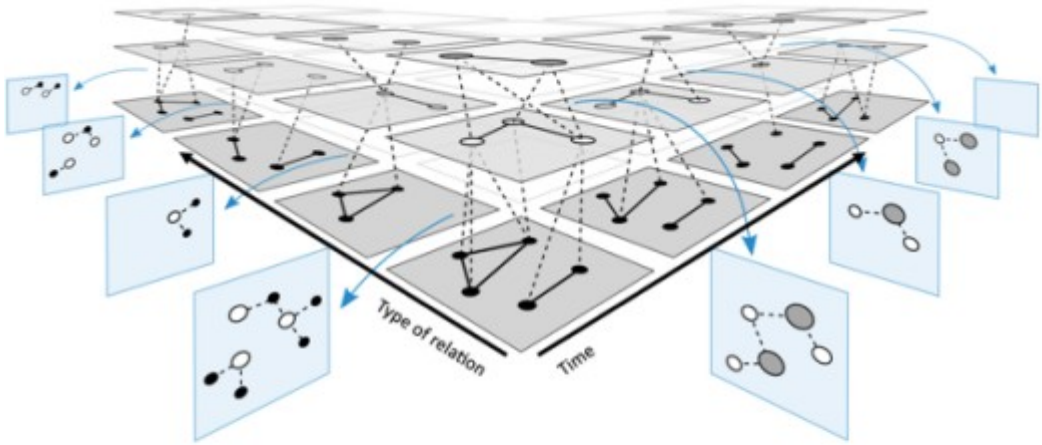
Des stratégies d’exploration

Network property	Examples of social interpretation
Local topology	Structural equivalence: if two actors have similar connections to other actors, they are similar or equivalent.
	Triadic closure: two friends of an actor eventually become friends.
	Balance theory: a friend of friend is a friend, a friend of a foe is a foe, and so on.
Degree and eigenvector centrality	Social capital: an actor produces common good for the friends.
	Influence: an actor causes a change in behavior in the friends.
Closeness centrality	Influence: see above.
	Information dissemination/diffusion: how good are actors in broadcasting or sharing information?
Betweenness centrality	Information dissemination: see above.
	Brokerage: how good are actors in serving as “go-betweens”?
Community structure	Homophily (cognitive balance): “birds of a feather flock together.”
	Knowledge preservation: actors in tightly knit communities preserve knowledge.
	Complex contagion: a gang of interconnected infected actors is a source of contagion.
Degree distribution	Small world (six degrees of separation): any two actors on average are connected by six “handshakes.”
	Friendship paradox: “my friends have more friends than I do.”
Network dynamics	Preferential attachment (Pareto principle): “the [actors] rich [in friends] get richer.”

Pas de limite de complexité

Des réseaux potentiellement très complexes.

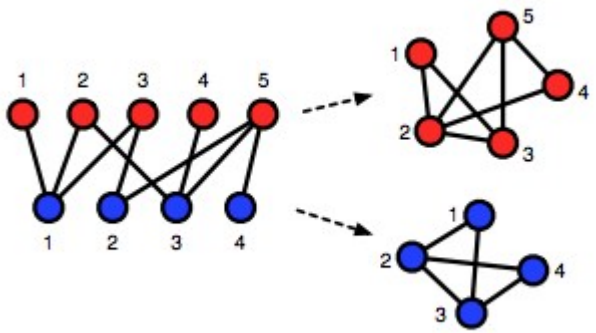
Une partie importante peut être juste de circuler dans le réseau pour identifier certaines caractéristiques



■ Knudsen et al. (2019) « Unifying the Framework of Multi-Layer Network and Visual Analytics », *Visual Analytics of Multilayer Networks Across Disciplines*, Dagstuhl Reports, 9, 2, 19-23.

De multiples transformations

- Filtrer
  - Retenir la composante principale
  - Enlever les noeuds faiblement connectés
- Projeter
  - Réseau bipartite vers un réseau simple



(i) Incidence matrix

Companies \ Directors	A	B	C	D	E
1	1	1	1	1	0
2	1	1	1	0	1
3	0	1	1	1	0
4	0	0	1	0	1

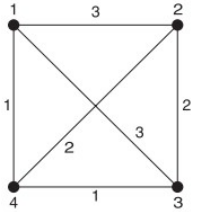
(ii) Adjacency matrix: companies-by-companies

	1	2	3	4
1	–	3	3	1
2	3	–	2	2
3	3	2	–	1
4	1	2	1	–

(iii) Adjacency matrix: directors-by-directors

	A	B	C	D	E
A	–	2	2	1	1
B	2	–	3	2	1
C	2	3	–	2	2
D	1	2	2	–	0
E	1	1	2	0	–

(iv) Sociogram: companies



(v) Sociogram: directors

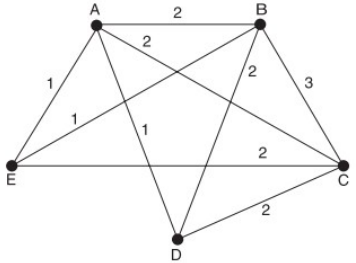


Figure 4.5 Matrices for interlocking directorships

La question des visualisations

- Un réseau est une structure théorique
- Mais qui peut être visualisée
- Mais ça veut dire quoi visualiser un réseau ?

Original Research Article

BIG DATA & SOCIETY

What do we see when we look at networks: Visual network analysis, relational ambiguity, and force-directed layouts

Big Data & Society
January–June: 1–16

DOI: 10.1177/20539517211018488
journals.sagepub.com/home/bds
SAGE

Tommaso Venturini[1], Mathieu Jacomy[2] and Pablo Jensen[3,4]

Abstract

It is increasingly common in natural and social sciences to rely on network visualizations to explore relational datasets and illustrate findings. Such practices have been around long enough to prove that scholars find it useful to project networks in a two-dimensional space and to use their visual qualities as proxies for their topological features. Yet these practices remain based on intuition, and the foundations and limits of this type of exploration are still implicit. To fill this lack of formalization, this paper offers explicit documentation for the kind of *visual network analysis* encouraged by *force-directed layouts*. Using the example of a network of Jazz performers, band and record labels extracted from Wikipedia, the paper provides guidelines on how to make networks readable and how to interpret their visual features. It discusses how the inherent ambiguity of network visualizations can be exploited for exploratory data analysis. Acknowledging that vagueness is a feature of many relational datasets in the humanities and social sciences, the paper contends that visual ambiguity, if properly interpreted, can be an asset for the analysis. Finally, we propose two attempts to distinguish the ambiguity inherited from the represented phenomenon from the distortions coming from fitting a multidimensional object in a two-dimensional space. We discuss why these attempts are only partially successful, and we propose further steps towards a metric of spatialization quality.

Le réseau :
un moment dans l'analyse

Garimella K., Morales G.D.F., Gionis A., Mathioudakis M., 2018, « Quantifying Controversy on Social Media », ACM Transactions on Social Computing, 1, 1, p. 1-27.

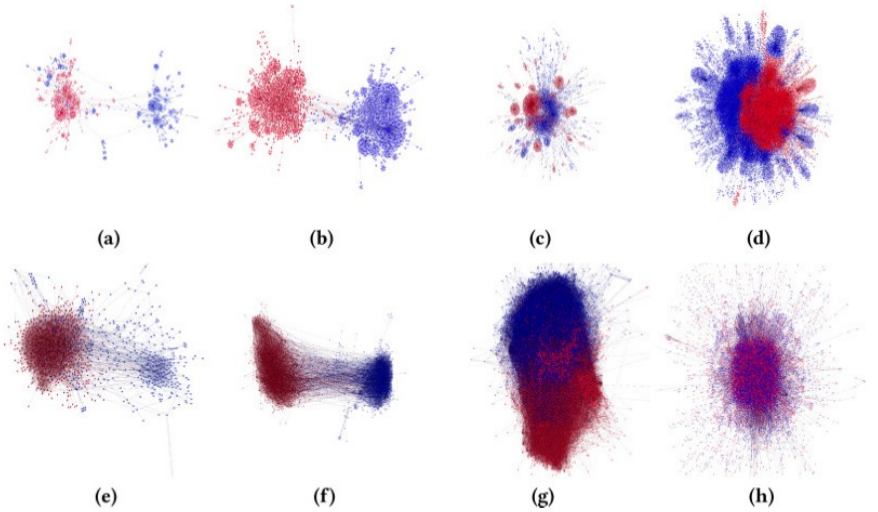


Fig. 3. Sample conversation graphs with retweet (top) and follow (bottom) aspects (visualized using the force-directed layout algorithm in Gephi). The left side is controversial, ((a) and (e)) #beefban and ((b) and (f)) #russia_march, while the right side is non-controversial, ((c) and (g)) #sxsw and ((d) and (h)) #germanwings. Only the largest connected component is shown.

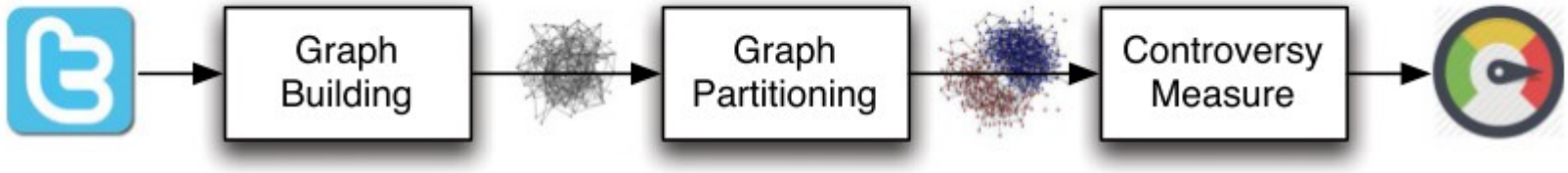


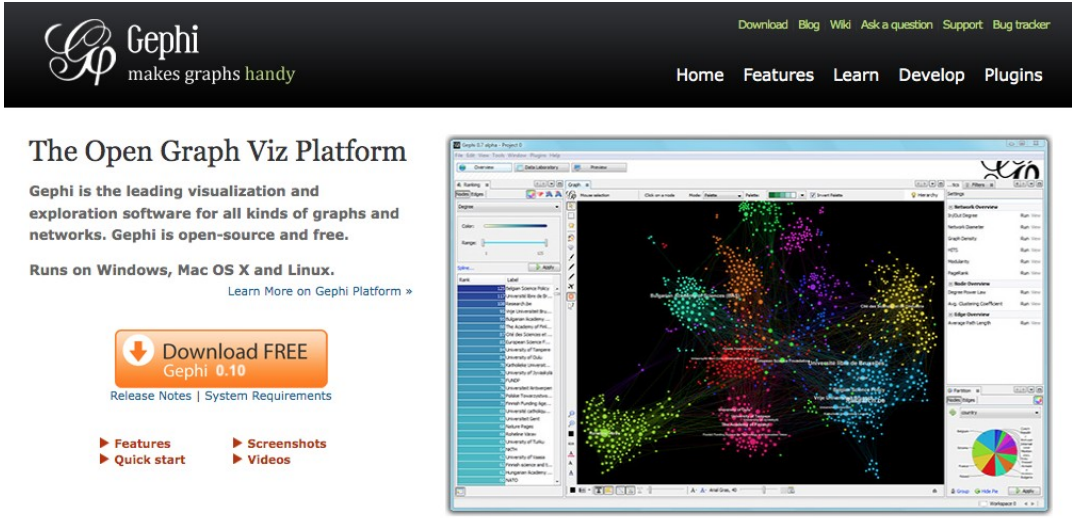
Fig. 1. Block diagram of the pipeline for computing controversy scores.

Workflow complet analyse de réseaux en SHS

- Collecter des données
- Les décrire et les mettre en forme
- Construire un réseau
- Le visualiser / l’analyser
- Produire des résultats finalisés

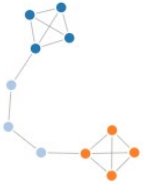
Pour mettre en oeuvre

- NodeXL (with Excel)
- Pajek (old)
- Gephi (state of the art)
- Python / R (a possibility)



NetworkX
Network Analysis in Python

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.



Software for complex networks

- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source 3-clause BSD license
- Well tested with over 90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

Pourquoi Python

Construire une chaîne complète de traitement de la collecte à l’analyse

Permettre la reproductibilité des étapes (éviter le clic)

Recourir aux outils existants facilitant la programmation scientifique (Notebooks)

Intégrer plus facilement les étapes de machine learning

Proposition pour l’analyse

- Utiliser les notebooks permettant l’intégration de la collecte à la visualisation
  - Garder potentiellement GEPHI pour les traitements finaux
- Utiliser Python pour la souplesse de son écriture
  - Pandas pour traiter les tableaux
  - Networkx pour la manipulation de structure relationnelle
  - Ipysigma pour la visualisation exploratoire
- Pas à pas avec un exemple : les publications scientifiques

Le réseau : un format à part entière

Un objet particulier (le graph)

Des opérations assez standardisées

Intrégrer l’ensemble dans un processus de travail

Contact

Mailing list
Issue tracker
Source

Releases

Stable (notes)

3.0 — January 2023
download | doc | pdf

Latest (notes)

3.1 development
github | doc | pdf

Archive

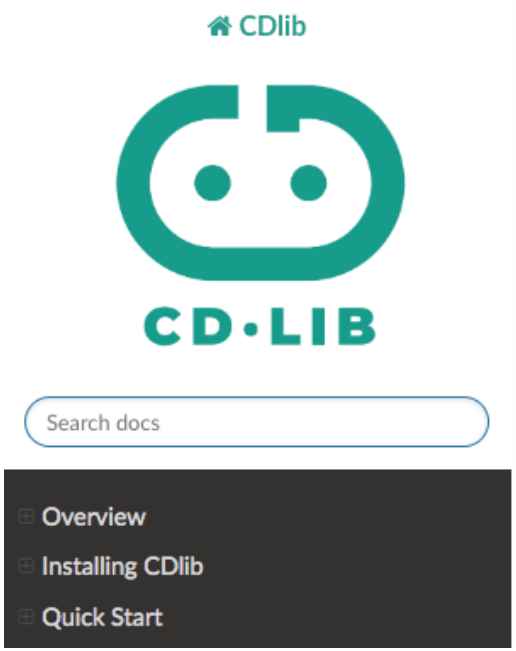
NetworkX
Network Analysis in Python

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

Software for complex networks

- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source 3-clause BSD license
- Well tested with over 90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

Et plus si affinité



Docs » CDlib - Community Discovery Library

Edit on GitHub

CDlib - Community Discovery Library

`CDlib` is a Python software package that allows to extract, compare and evaluate communities from complex networks.

The library provides a standardized input/output for several existing Community Discovery algorithms. The implementations of all CD algorithms are inherited from existing projects, each one of them acknowledged in the dedicated method reference page.

If you would like to test `CDlib` functionalities without installing it on your machine consider using the preconfigured Jupyter Hub instances offered by the H2020 SoBigData++ research project.

Bibliographie

- Introduction to Social Network Analysis: Basics and Historical Specificities. HNR+ResHist Conference 2021, Historical Network Research, 2021, Luxembourg, Luxembourg. ⟨10.5281/zenodo.5083036⟩. ⟨halshs-03351755⟩
  - Video: https://www.martingrandjean.ch/introduction-to-social-network-analysis/
- A complete course : https://aaronclauset.github.io/courses/5352/
- Scott J., 2017, Social network analysis.
- Zinoviev D., 2018, Complex Network Analysis in Python. Recognize → Construct → Visualize → Analyze → Interpret.
- Database of networks : https://icon.colorado.edu/#!/ ; http://snap.stanford.edu/data/index.html