M1 CompuPhys 2022-2023

Introduction to Python Language

Network Analysis

Célestin Coquidé

PostDoc at UTINAM (Observatory building office n°13)

Table of contents :

Lectures and Practical Works

Lectures
26/10 1.5 h
28/10 1.5 h

Practicing
17/11 3h[1]
24/11 3h[2]

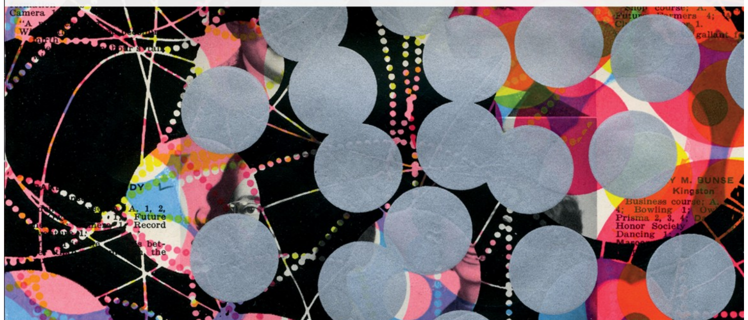
[1]: Handling of different packages + Exercises

[2]: Preparation for a graded homework

I Introduction

Data science

Harvard Business Review



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Forbes

Leadership / #IfIOnlyKnew

SEP 21, 2017 @ 03:45 PM 15,659

Becoming A Data Scientist: The Skills That Can Make You The Most Money

Karsten Strauss, FORBES STAFF

FULL BIO

listings. The ten skills that appeared most often as prerequisites for the job, and the percentage of job listings in which they appeared, were:

1. **Python** (72%)
2. **R** (64%)
3. **SQL** (51%)
4. **Hadoop** (39%)
5. **Java** (33%)

MENU nwi.com News Sports Business Photos Obits Classifieds 66°

URGENT

Data surge: Demand for data scientists set to rise in the coming years

Craig Guillot CTW Features Sep 17, 2017 Updated Sep 28, 2017

Le Monde.fr ÉDITION GLOBALE Rechercher

INTERNATIONAL POLITIQUE SOCIÉTÉ ÉCO CULTURE IDÉES PLANÈTE SPORT SCIENC

M CAMPUS

CAMPUS Lycée Universités Grandes écoles Le monde des étudiants O21 Etudier à l'étran

Le « data scientist », nouvelle coqueluche des recruteurs

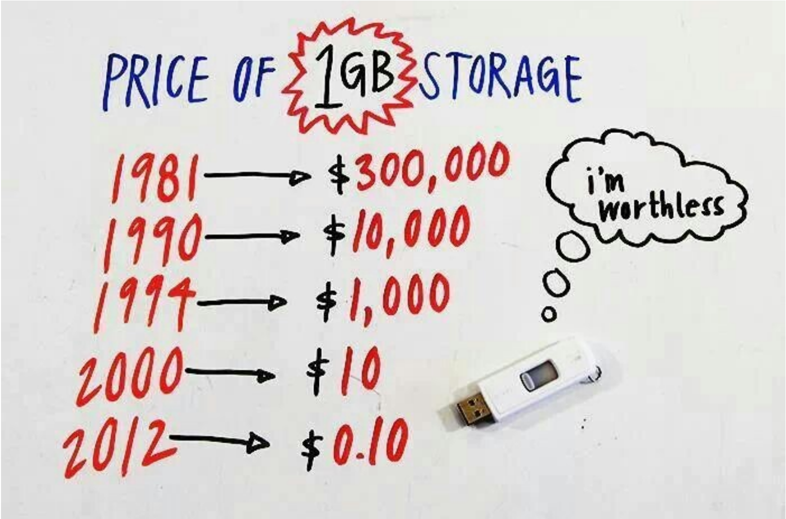
D'après notre baromètre en partenariat avec LinkedIn, les entreprises courtisent ces « scientifiques des données » pour leurs compétences en termes de transformation numérique

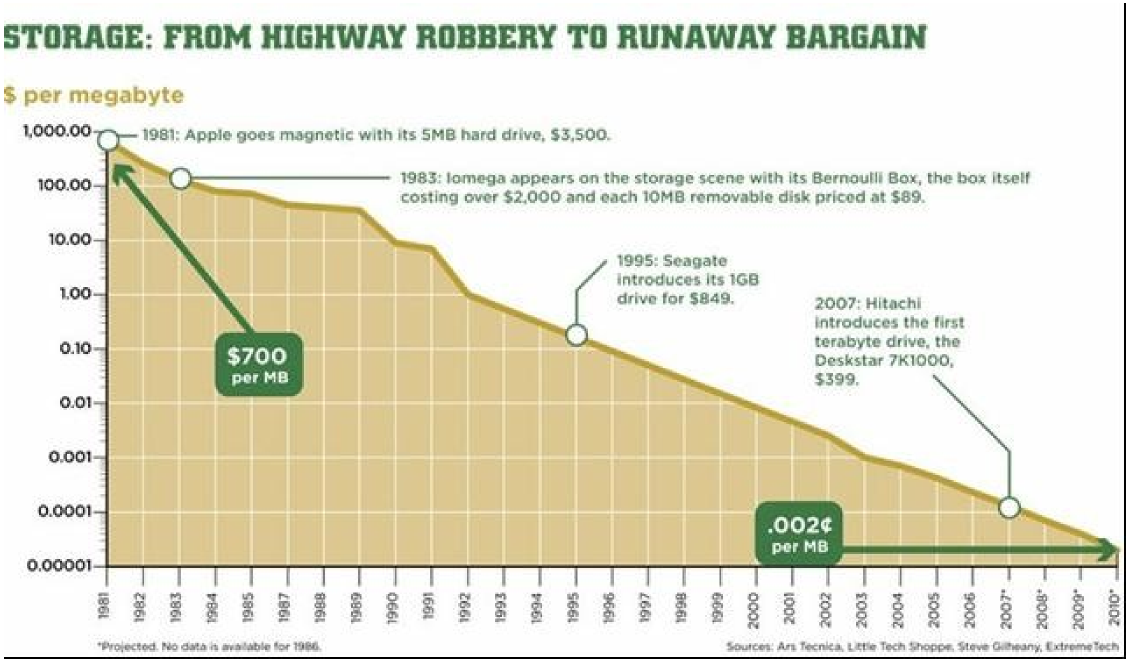
LE MONDE ECONOMIE | 12.05.2017 à 11h39 • Mis à jour le 16.05.2017 à 09h46 |

Data science

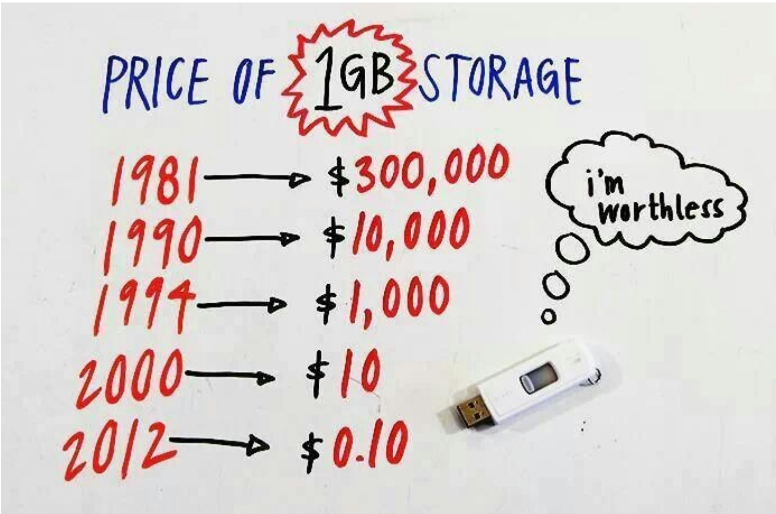
- A new discipline
- Few amount of book with a global and unified sight

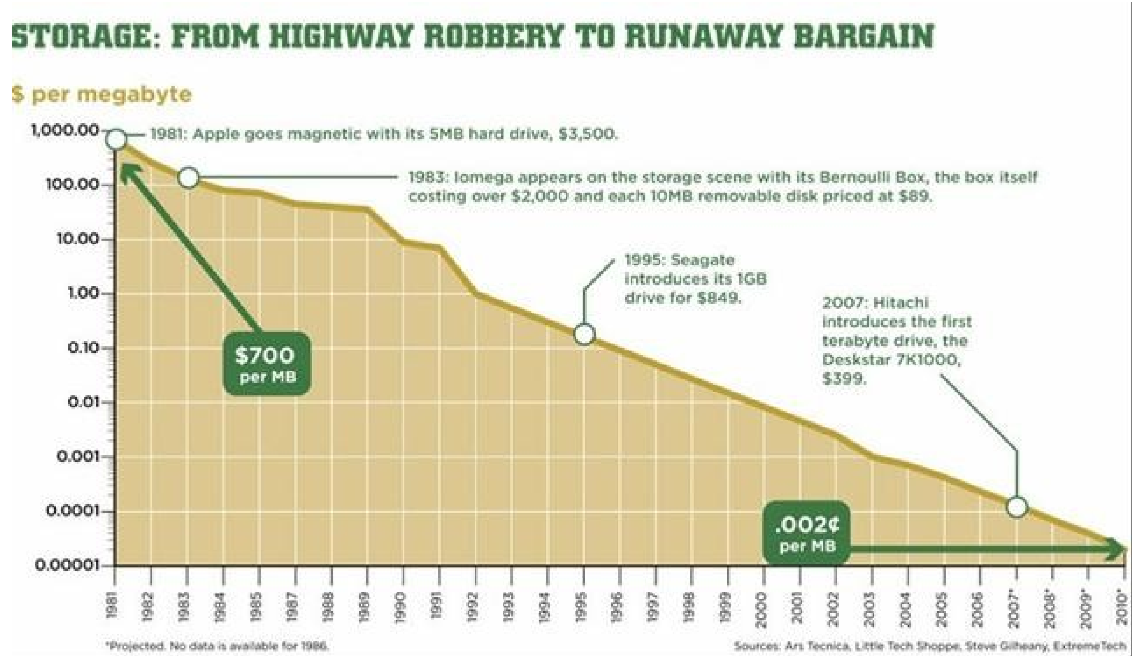
Evolutions in data

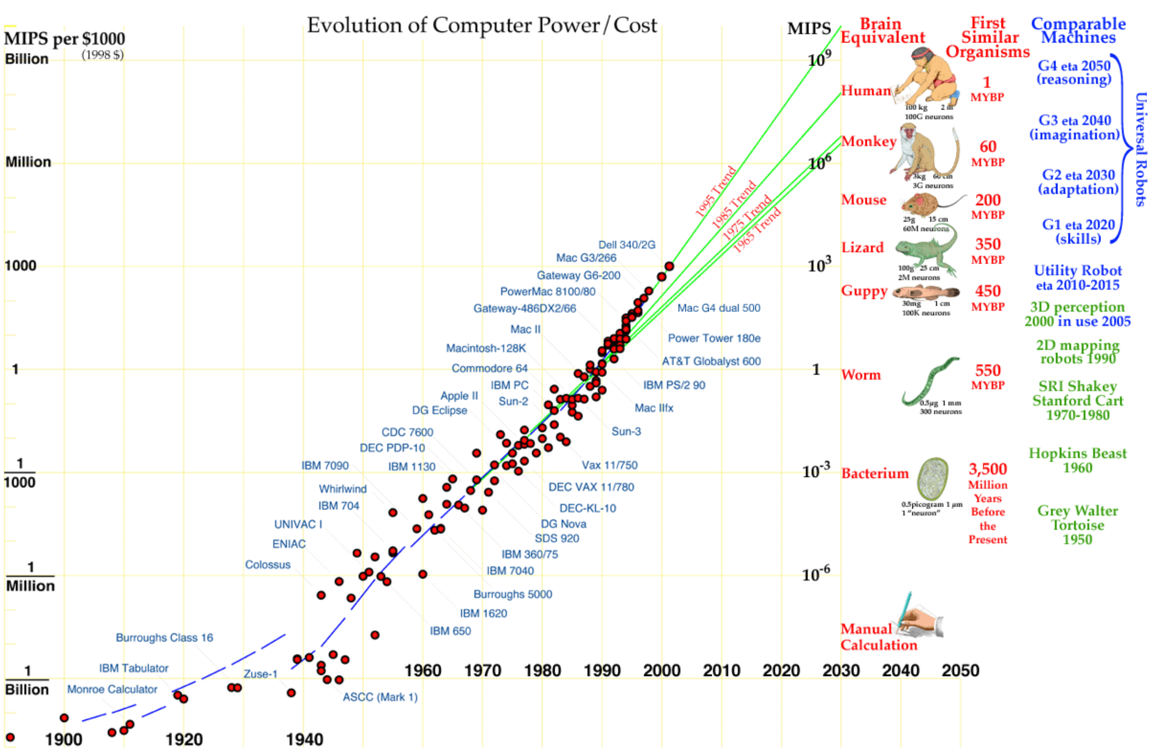


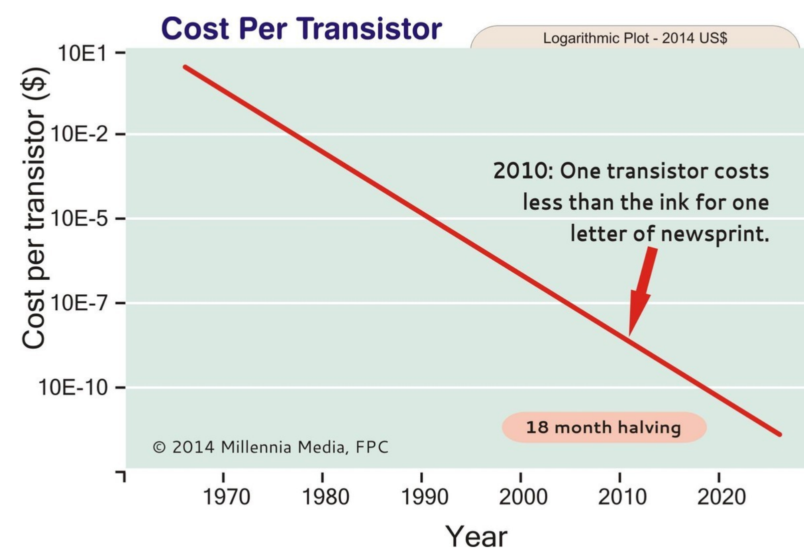


Evolutions in data

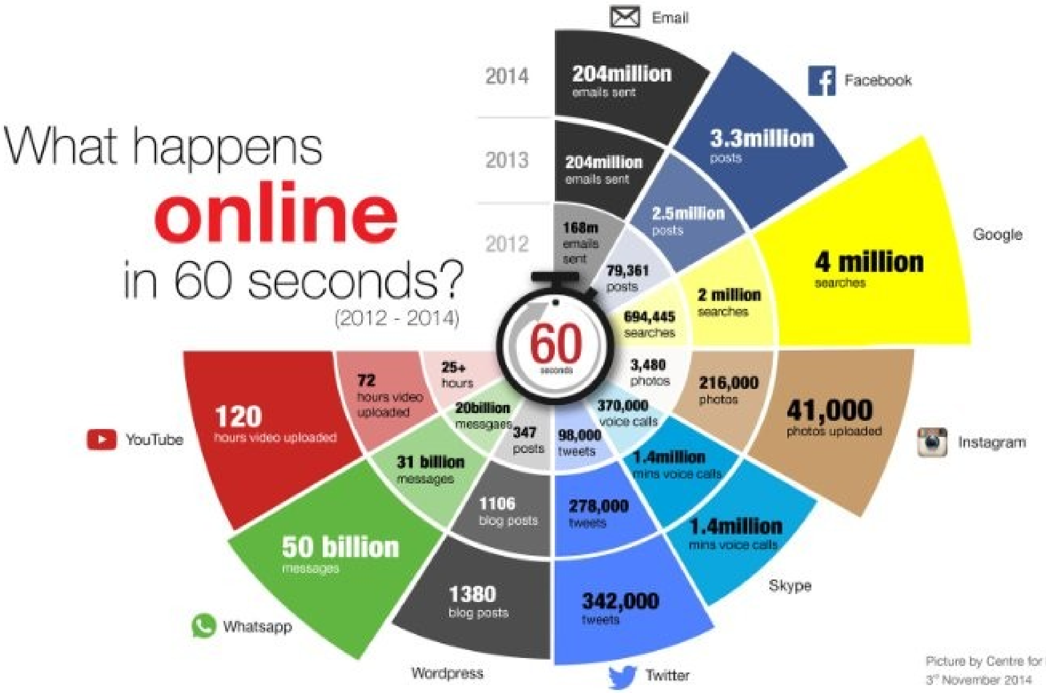


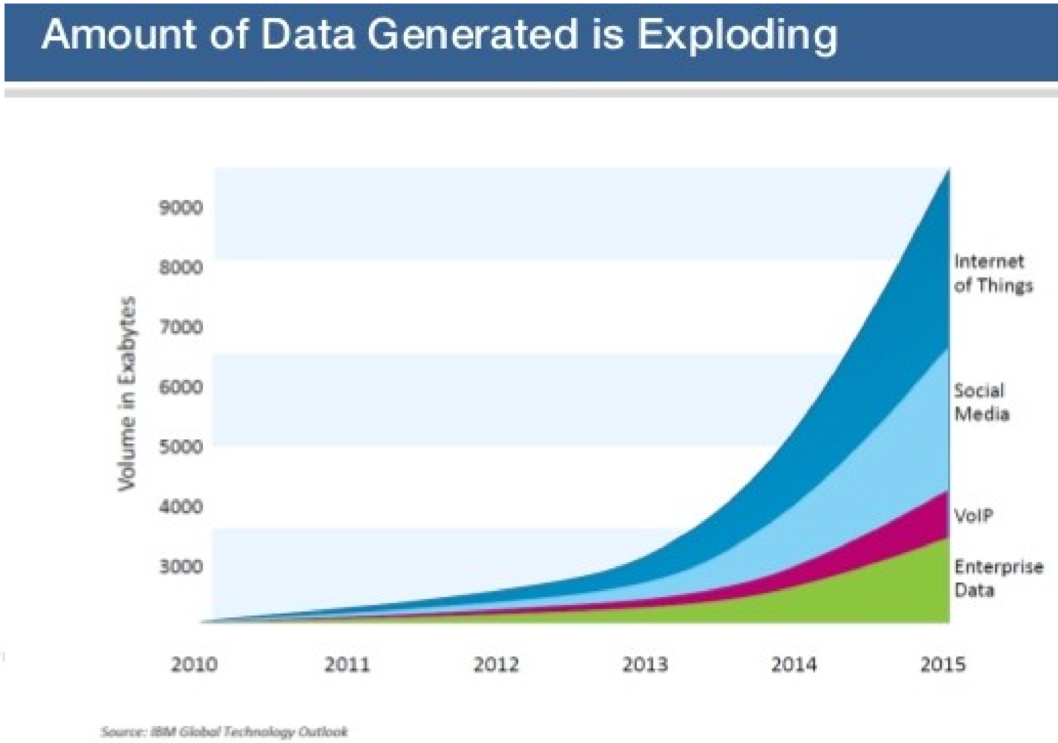


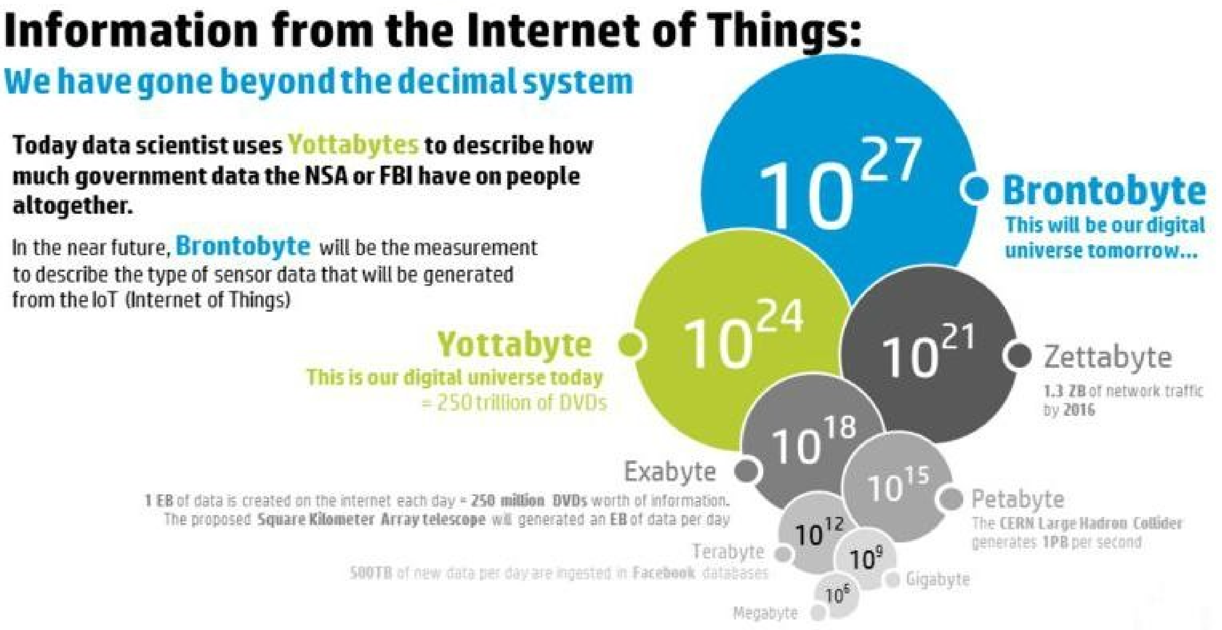




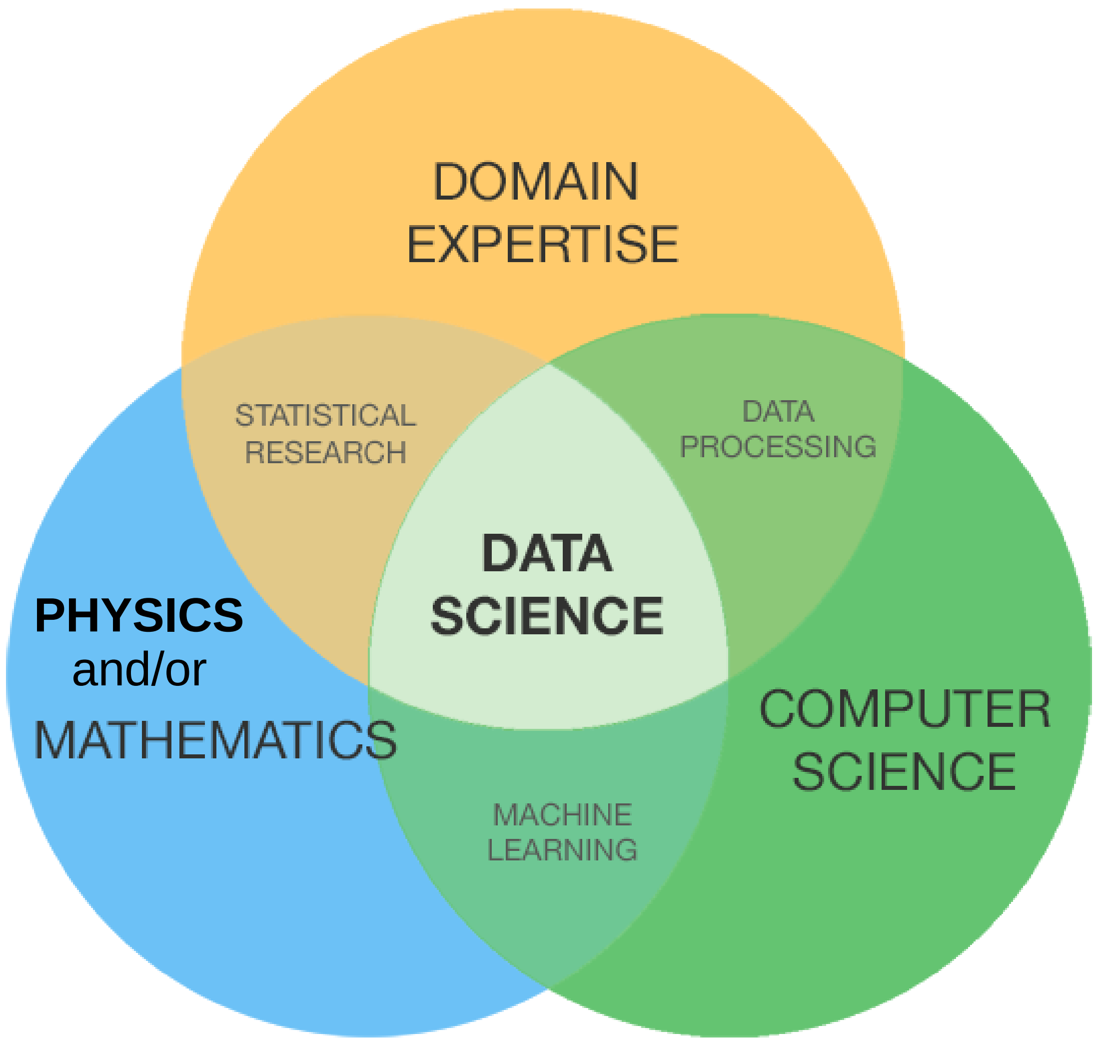
Evolutions in data







Multidisciplinary research field



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Different types of data

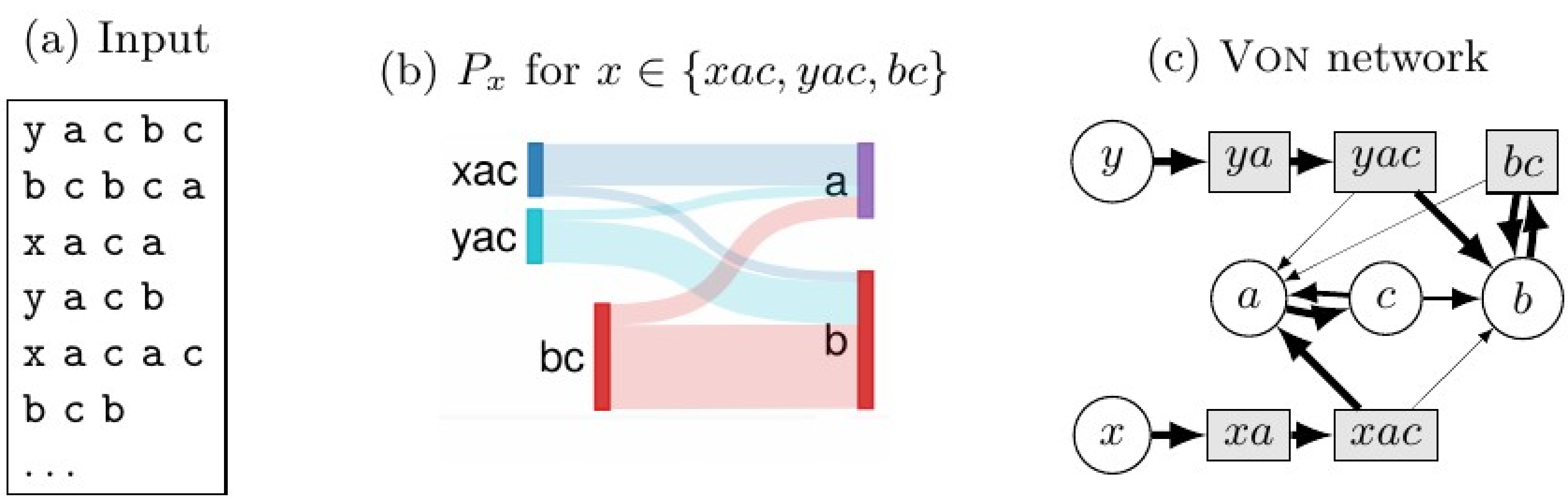
- Linear data
  ex: Correlation matrix between proteins

- Sequential data (implicit time)
  ex: Geographical itineraries

- Temporal data
  ex: Wall Street market

Different types of data

- Linear data
  ex: Correlation matrix between proteins
- Sequential data (implicit time)
  ex: Geographical itineraries
- Temporal data
  ex: Wall Street market

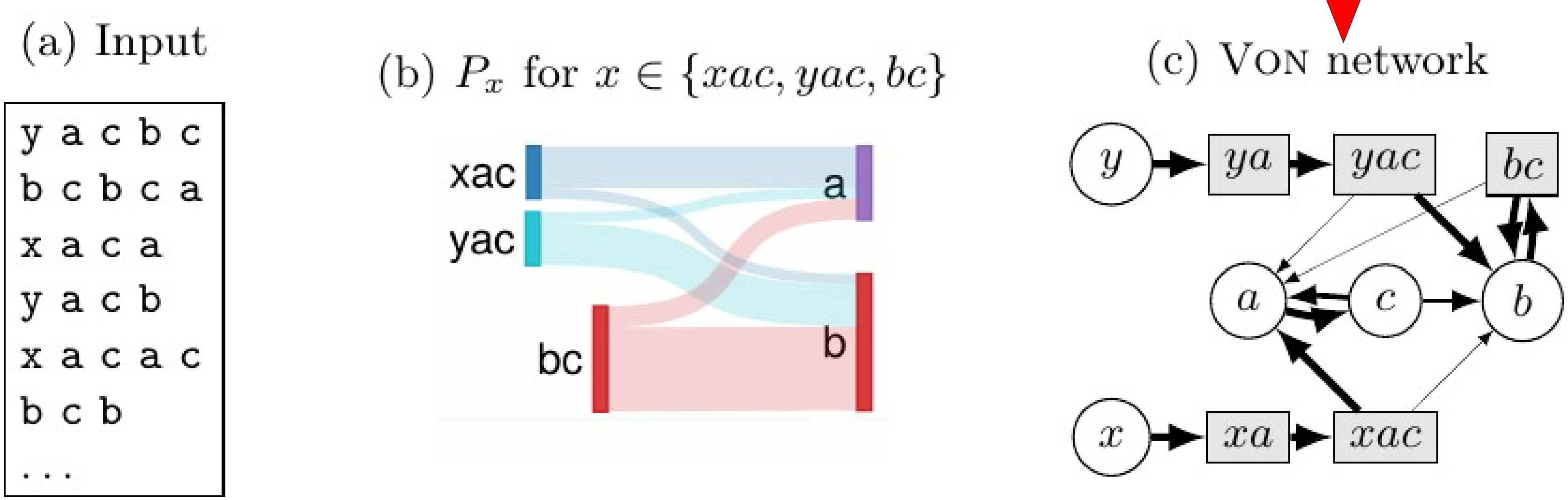


Example of sequential data and its analysis through variable-order Markov chains

Different types of data

- Linear data
  ex: Correlation matrix between proteins
- Sequential data (implicit time)
  ex: Geographical itineraries
- Temporal data
  ex: Wall Street market

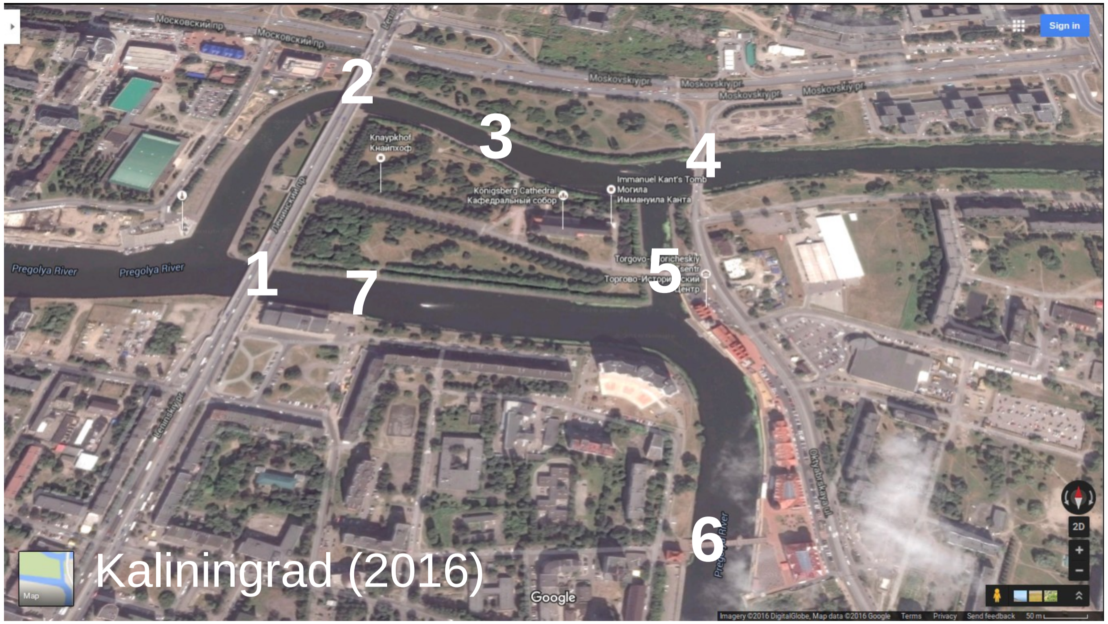
Network representation



Example of sequential data and its analysis through variable-order Markov chains

II Network Representations

Graph theory

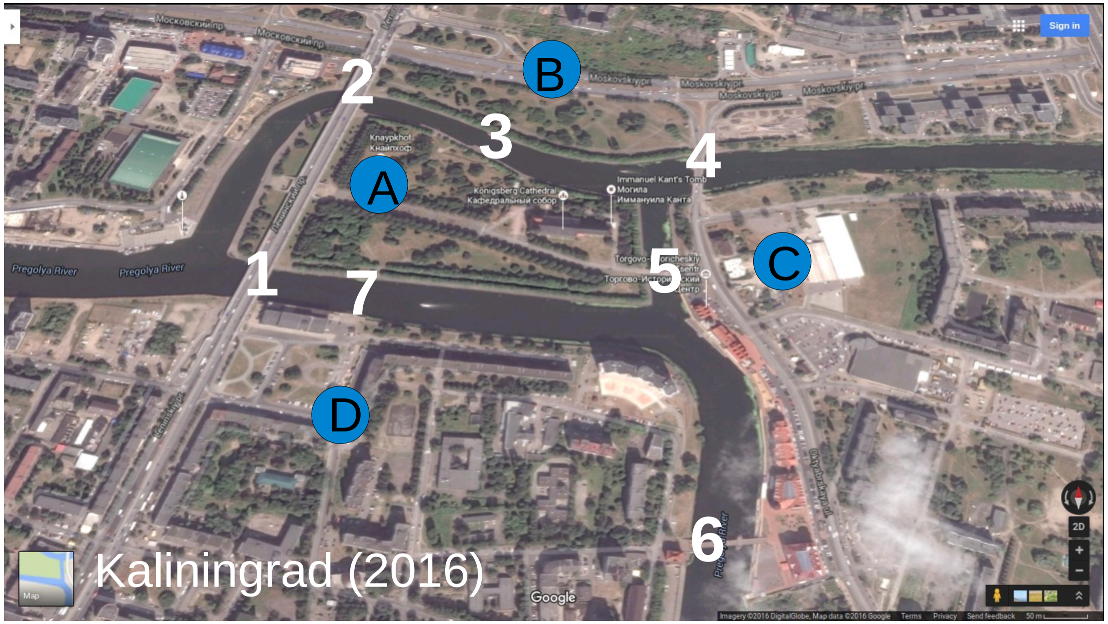


First paper on graph theory :

The Seven Bridges of Königsberg
Euler 1736

Is there a path connecting each territory passing by every bridge (but just once) ?

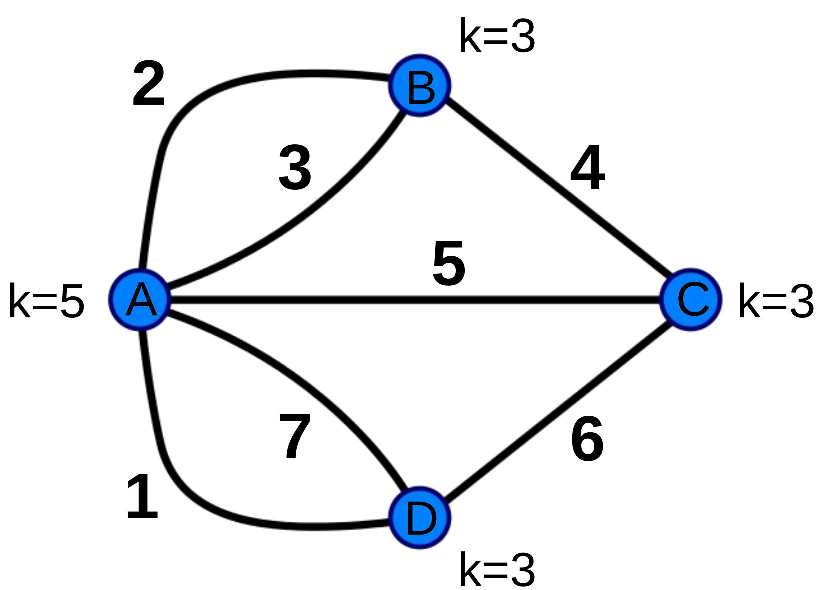
Graph theory



First paper on graph theory :

The Seven Bridges of Königsberg
Euler 1736

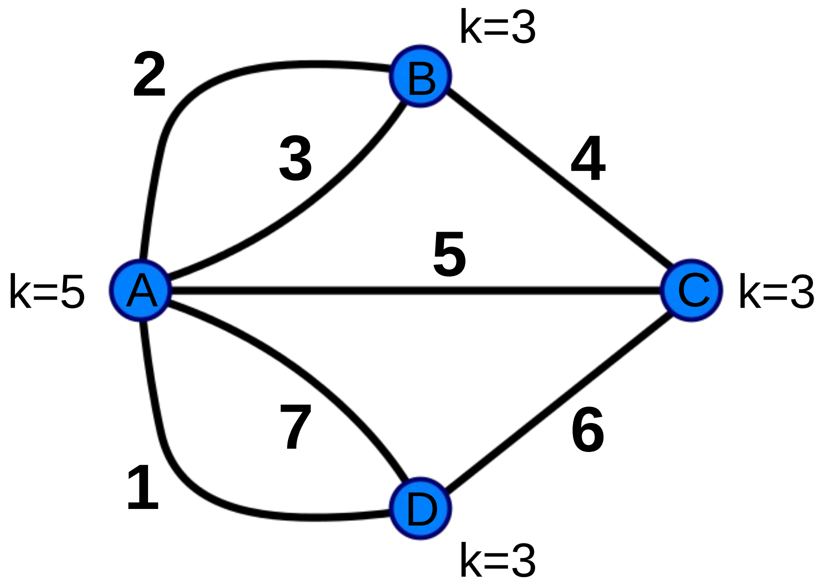
Is there a path connecting each territory passing by every bridge (but just once) ?
Answer : **No**



G=(V,E)
V : set of vertices (nodes)
E : set of edges (links)

k : vertex degree

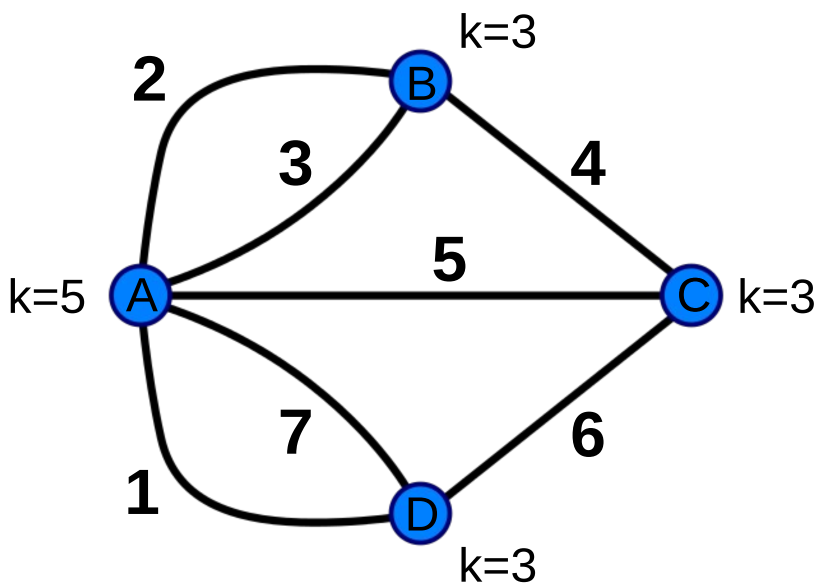
Network representation



Path = sequence of links

L_1(A,C) = [(A,D), (D,C)] Length 2
L_2(A,C) = [(A,C)] Length 1 = *shortest path*
L_3(A,C) = [(A,B), (B,C)] Length 2
L_4(A,C) = [(A,B), (B,A), (A,C)] Length 3
... ...

Network representation

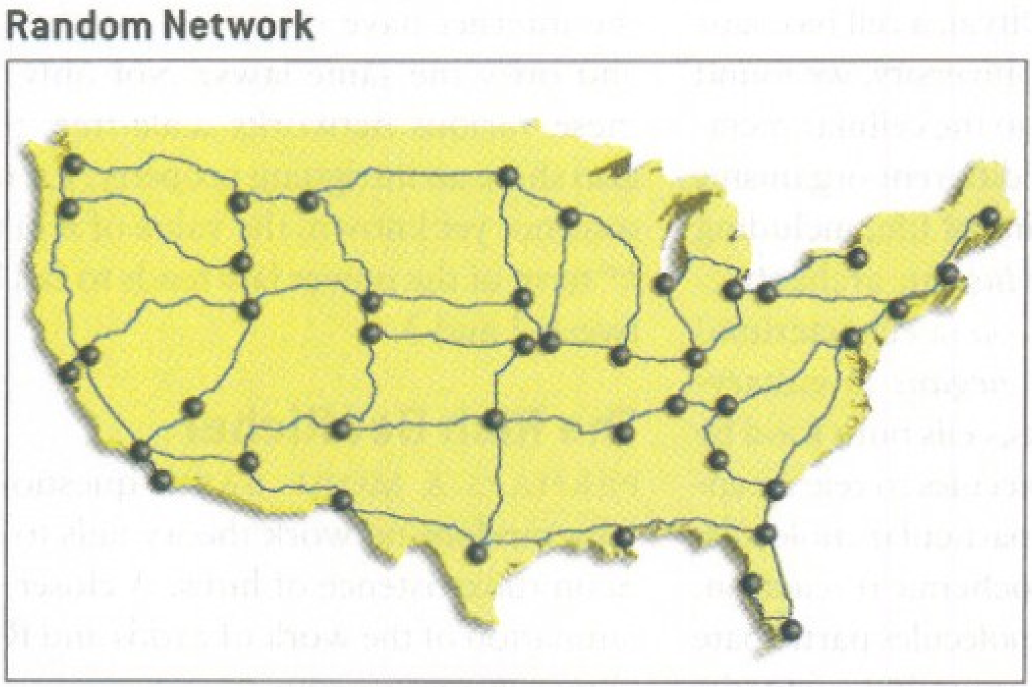


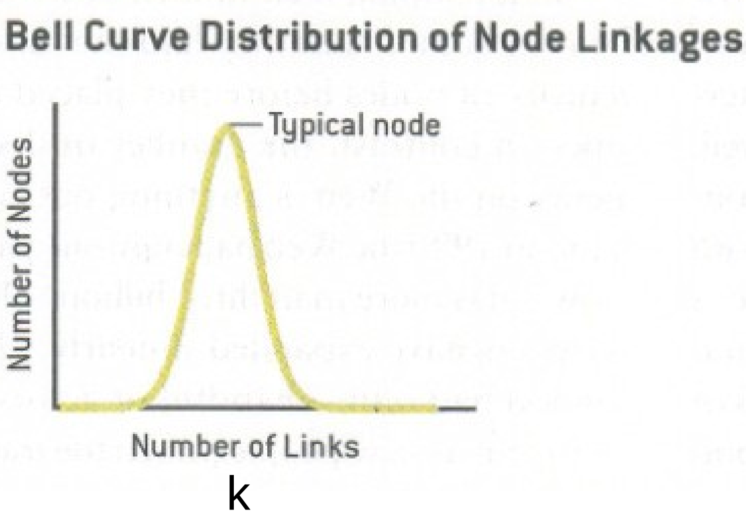
No
Eulerian path

Eulerian path = Sequence of links without duplicates representing a closed loop containing all nodes

A trivial model: Random networks

(Erdös-Rényi random graph model from '59)





Model 1 :
G=G(n,M)
Graphs with n vertices and M links

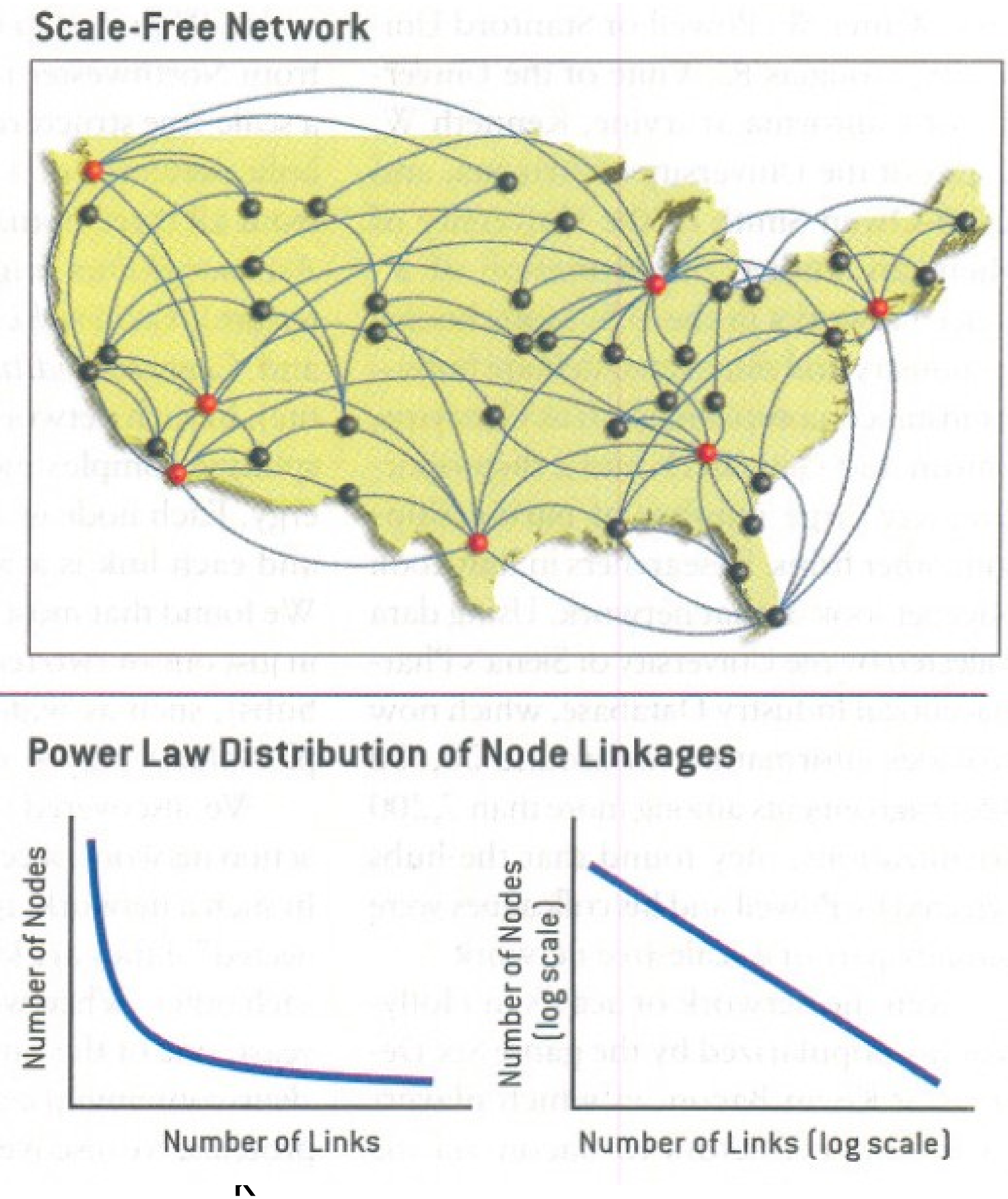
Model 2 :
G=G(n,p)
Graphs with n vertices and linking probability p.
Number of expected links = pn(n-1)/2

$n \rightarrow \infty$

Properties studied :
Connectedness
Graph diameter
Subgraphs
…

More complex model: Scale-free network

(Barabasi-Albert model from '99)



Preferential attachment :
$G=G(m_0,\vec{p})$
Initial graphs with m_0 vertices and preferential attachment vector $\vec{\mathbf{p}}$

$$p_i=\frac{k_i}{\sum_{j=1}^{m_0} k_j}$$

Where indexes i and j denote vertex "i" and "j" and *k* is a number of link.

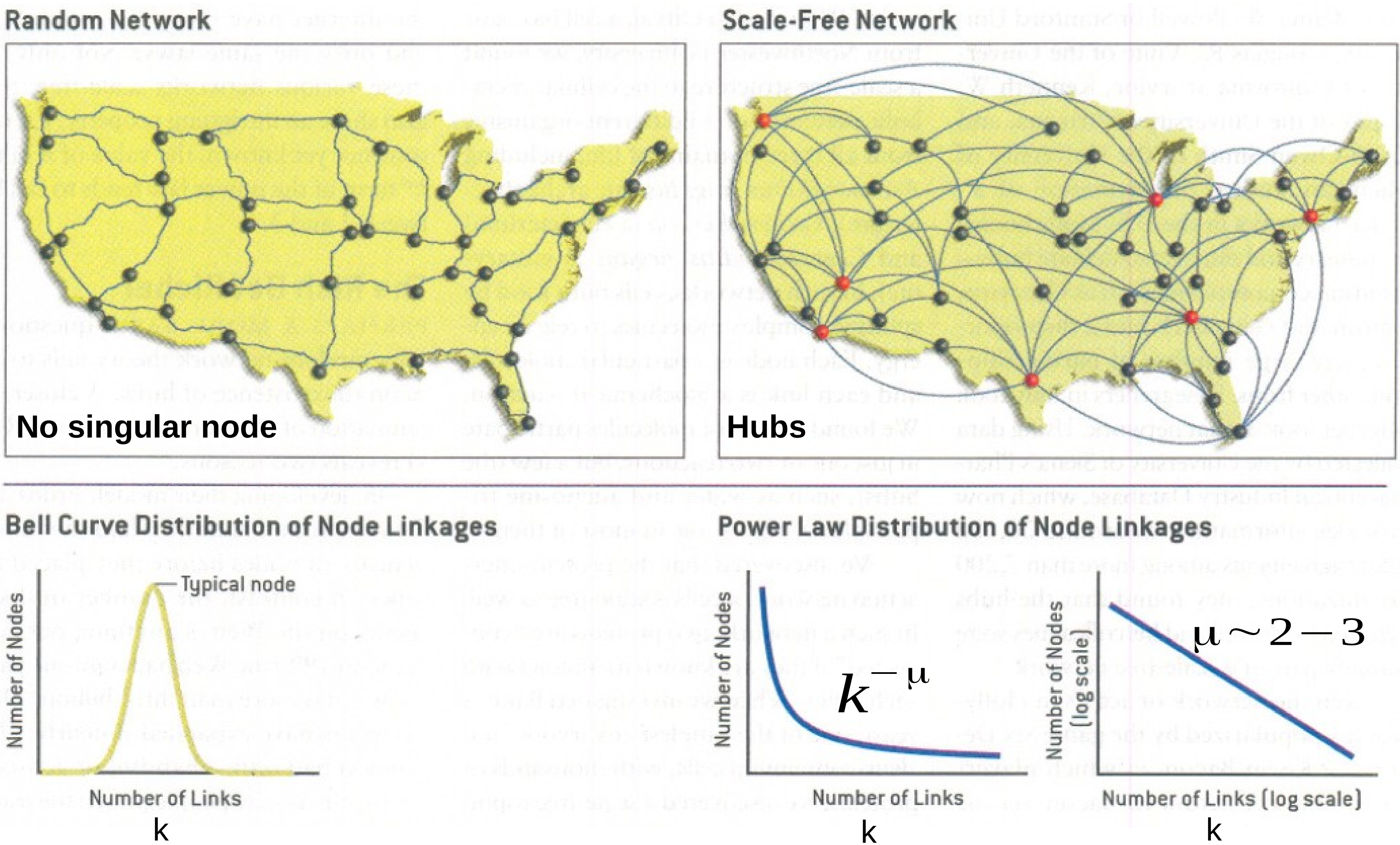
Properties studied :
Connectedness
Graph diameter
Subgraphs
…

Random graphs

(Erdös-Rényi random graph model from '59)

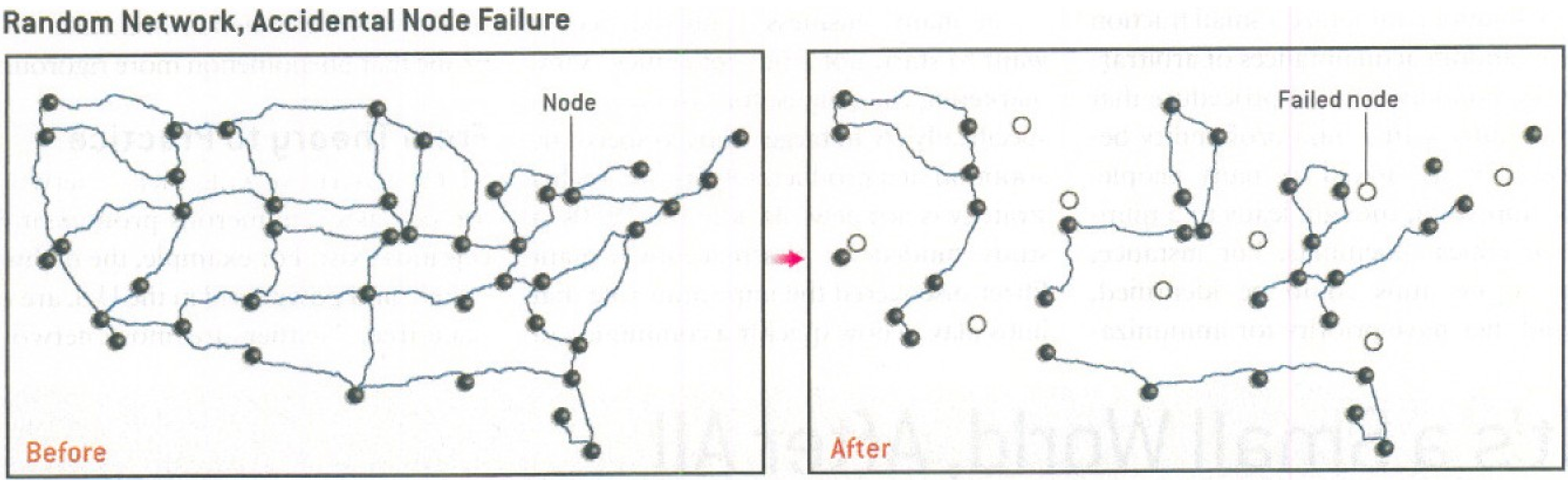
(Real) complex networks

(A.-L. Barabási, R. Albert '99)

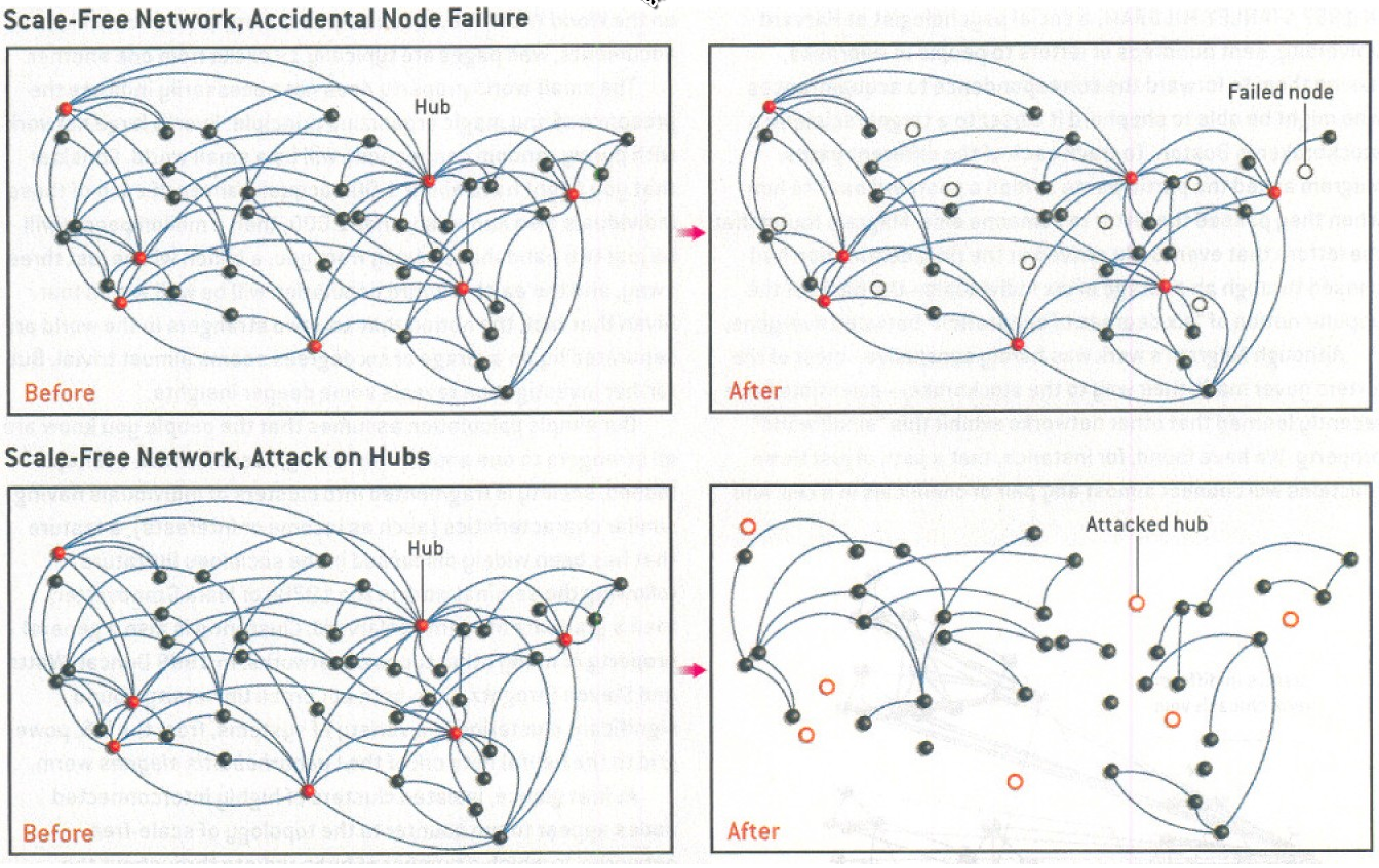


Random graphs

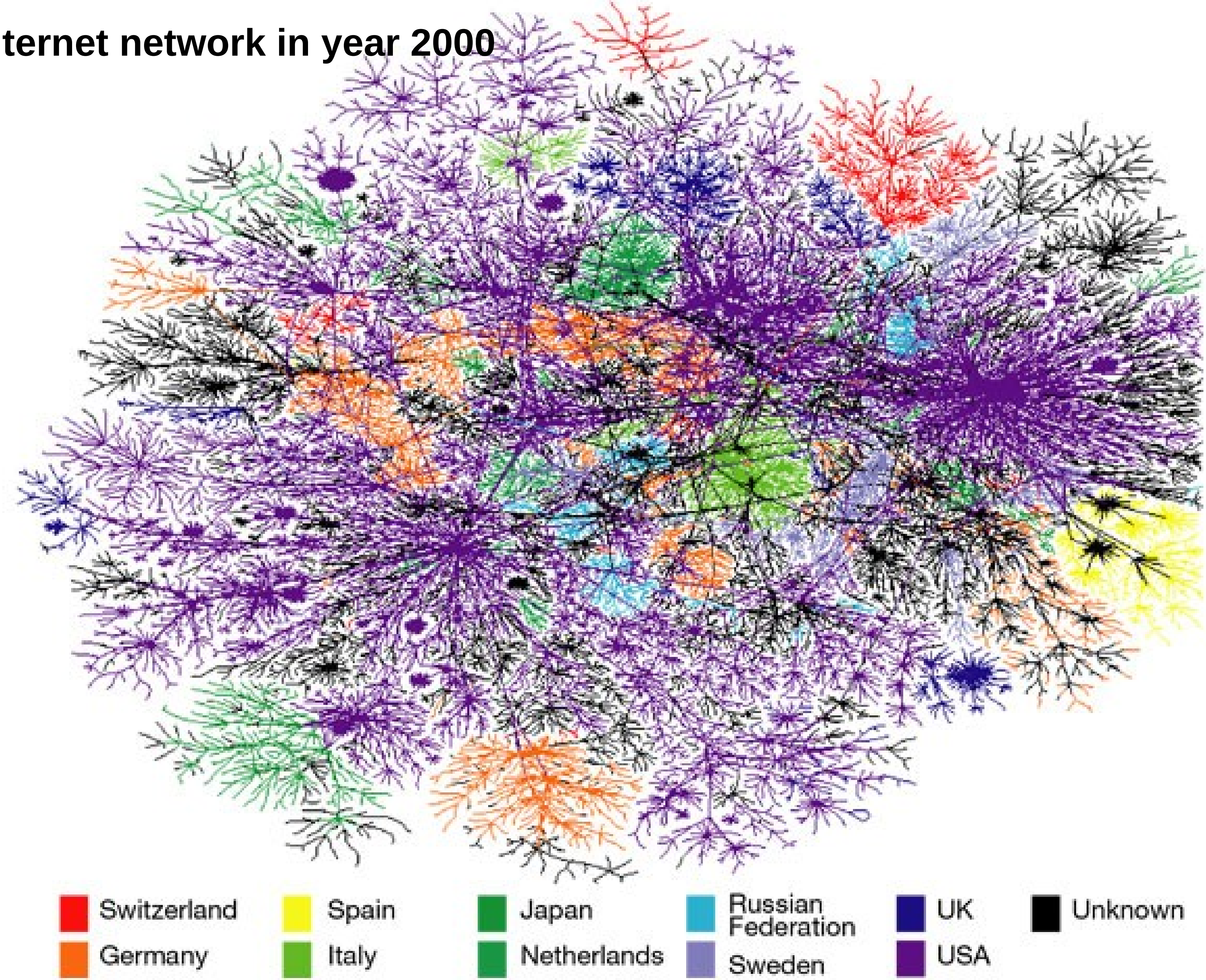
(Real) complex networks



Random graphs VS **(Real) complex networks**



Internet network in year 2000



From Hu, Nature 406, 353-354 (2000)

Real scale-free networks

Examples of Scale-Free Networks

NETWORK	NODES	LINKS
Cellular metabolism	Molecules involved in burning food for energy	Participation in the same biochemical reaction
Hollywood	Actors	Appearance in the same movie
Internet	Routers	Optical and other physical connections
Protein regulatory network	Proteins that help to regulate a cell's activities	Interactions among proteins
Research collaborations	Scientists	Co-authorship of papers
Sexual relationships	People	Sexual contact
World Wide Web	Web pages	URLs

Real scale-free networks

Examples of Scale-Free Networks

NETWORK	NODES	LINKS
Cellular metabolism	Molecules involved in burning food for energy	Participation in the same biochemical reaction
Hollywood	Actors	Appearance in the same movie
Internet	Routers	Optical and other physical connections
Protein regulatory network	Proteins that help to regulate a cell's activities	Interactions among proteins
Research collaborations	Scientists	Co-authorship of papers
Sexual relationships	People	Sexual contact
World Wide Web	Web pages	URLs

Construction property

"The rich get richer"
or
Matthew effect (Mt 25:29 NT)
or
Preferential attachment

Type of links:

Directed

Undirected

Weighted(D/U) 0.9

Scale-free networks

Examples of Scale-Free Networks

NETWORK	NODES	LINKS
Cellular metabolism	Molecules involved in burning food for energy	Participation in the same biochemical reaction
Hollywood	Actors	Appearance in the same movie
Internet	Routers	Optical and other physical connections
Protein regulatory network	Proteins that help to regulate a cell's activities	Interactions among proteins
Research collaborations	Scientists	Co-authorship of papers
Sexual relationships	People	Sexual contact
World Wide Web	Web pages	URLs

Construction property

"The rich get richer"
or
Matthew effect (Mt 25:29 NT)
or
Preferential attachment

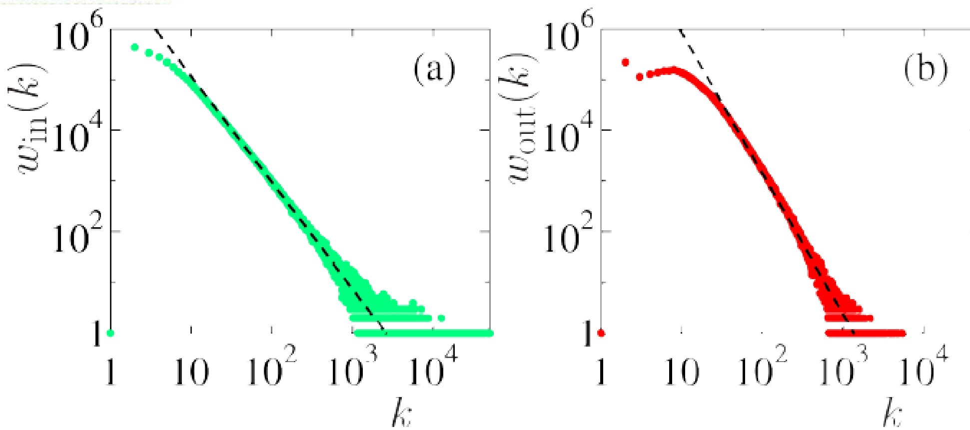
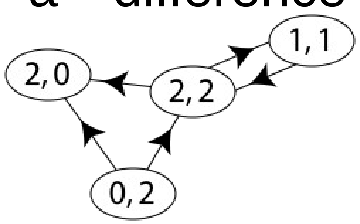


FIG. 2 (Color online) Distribution $w_{\rm in,out}(k)$ of number of ingoing (a) and outgoing (b) links k for $N = 3282257$ Wikipedia English articles (Aug 2009) of Fig. 1 with total number of links $N_\ell = 71012307$. The straight dashed fit line shows the slope with $\mu_{\rm in} = 2.09 \pm 0.04$ (a) and $\mu_{\rm out} = 2.76 \pm 0.06$ (b). After (Zhirov *et al.*, 2010).

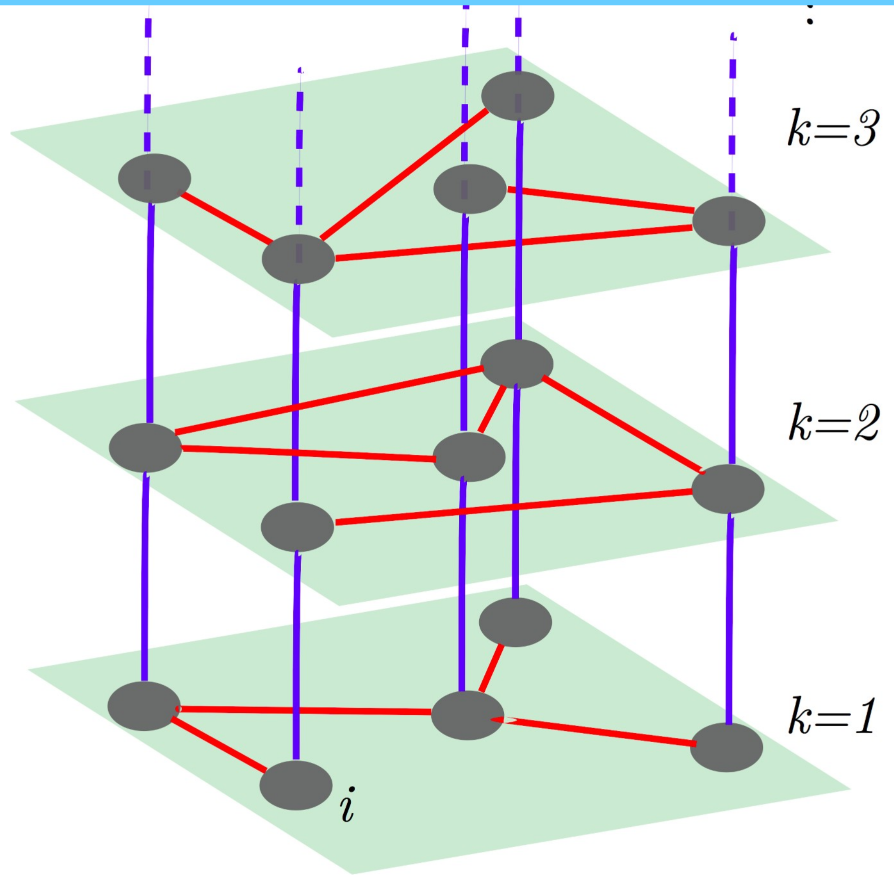
What about directionality ?
In fact any network can be considered as directed.

Now there is a difference between
node outdegree
And
node indegree

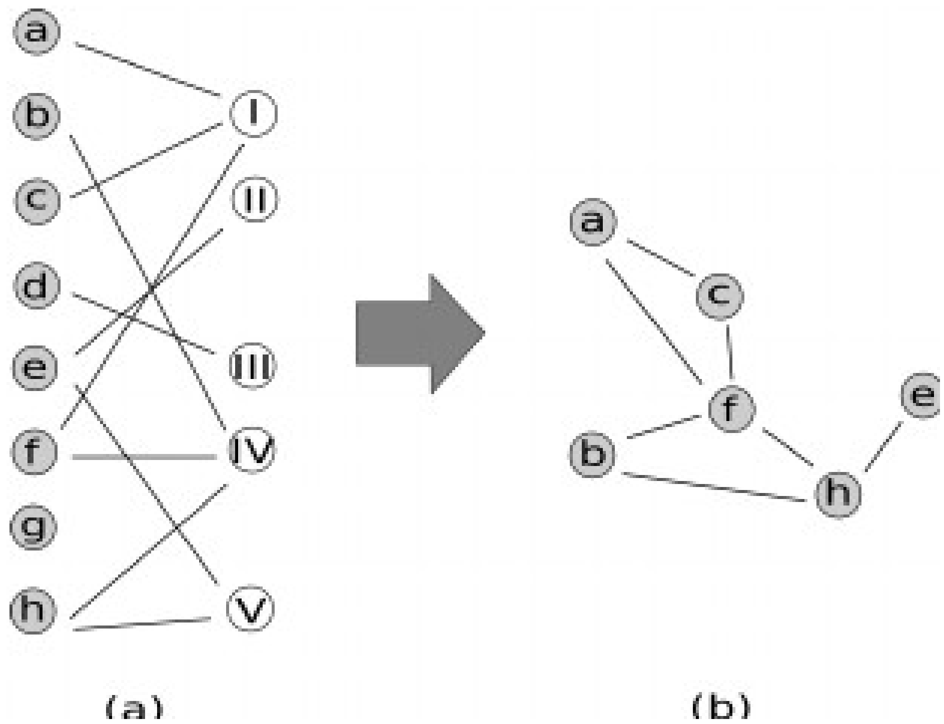


Large number of network models

Bipartite networks (ex: election network)
Multi-layer networks (ex: Wikipedia network)
Higher-order networks (Sequential data repr.)
Temporal networks (Time series)
Trees



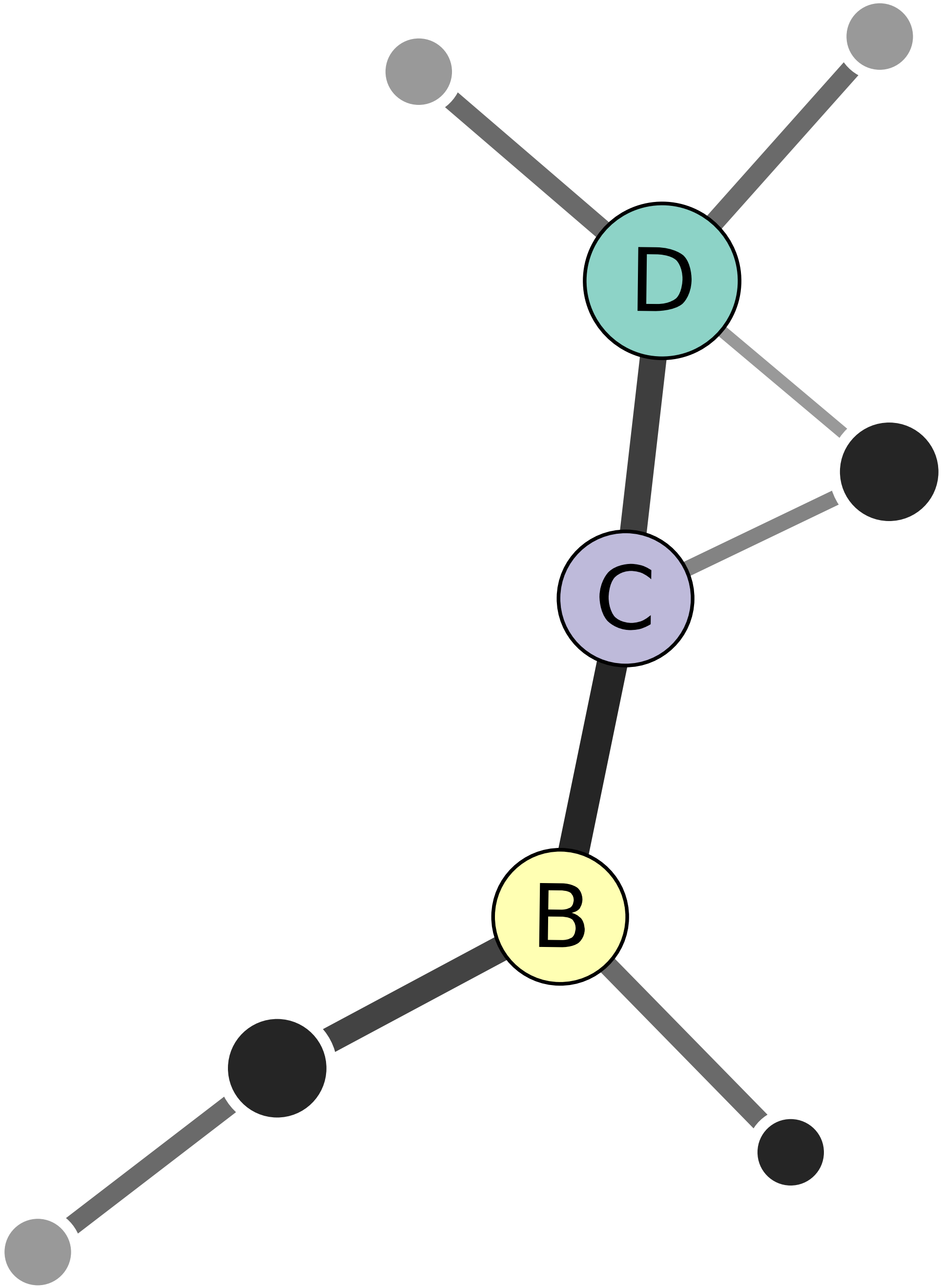
Multi-layer network



Bipartite network and its projection

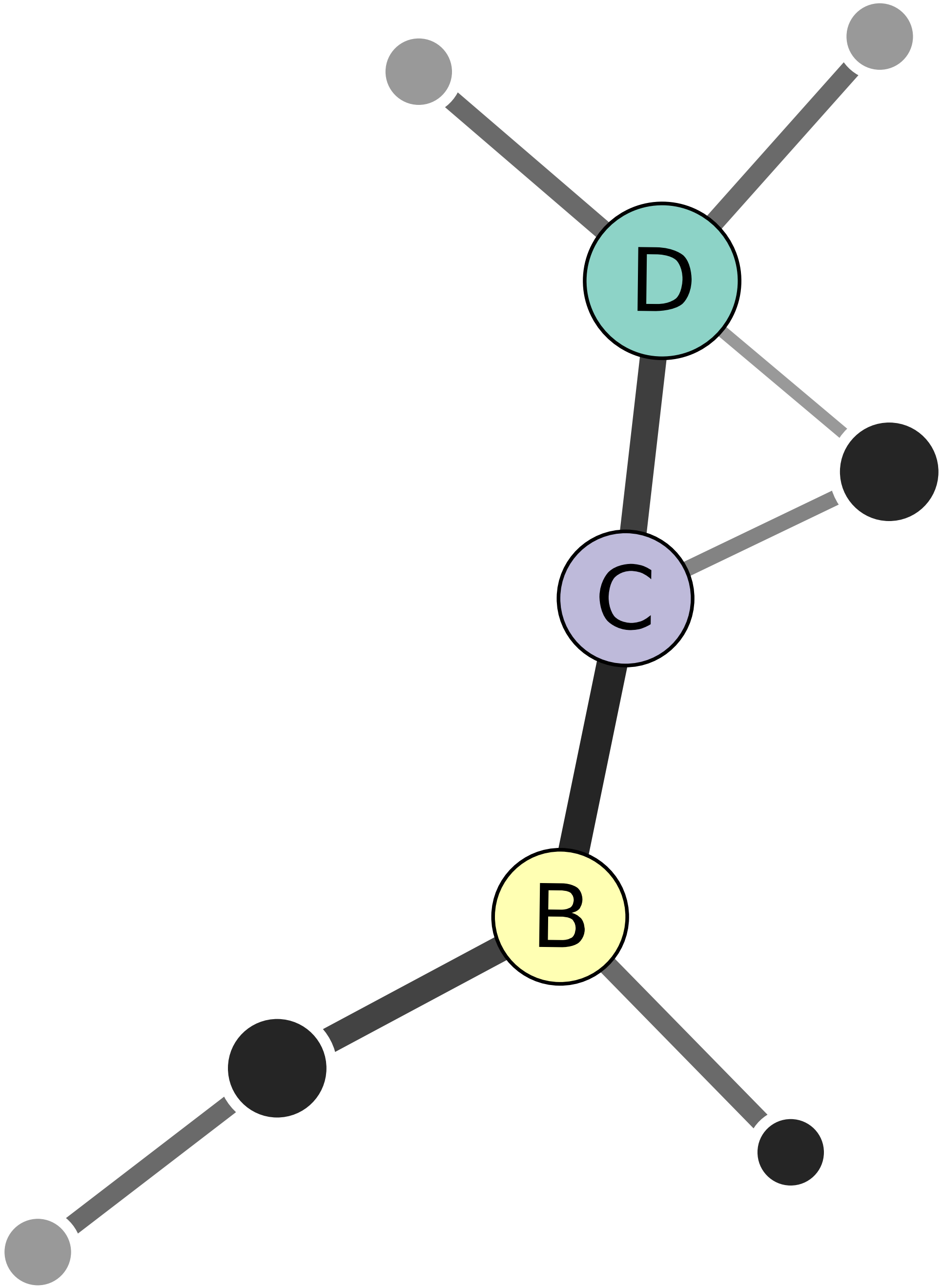
III Centrality Measures

What is a Centrality measure?



What is a Centrality measure?

Number of Link?
Efficiency of path?
Number of path?

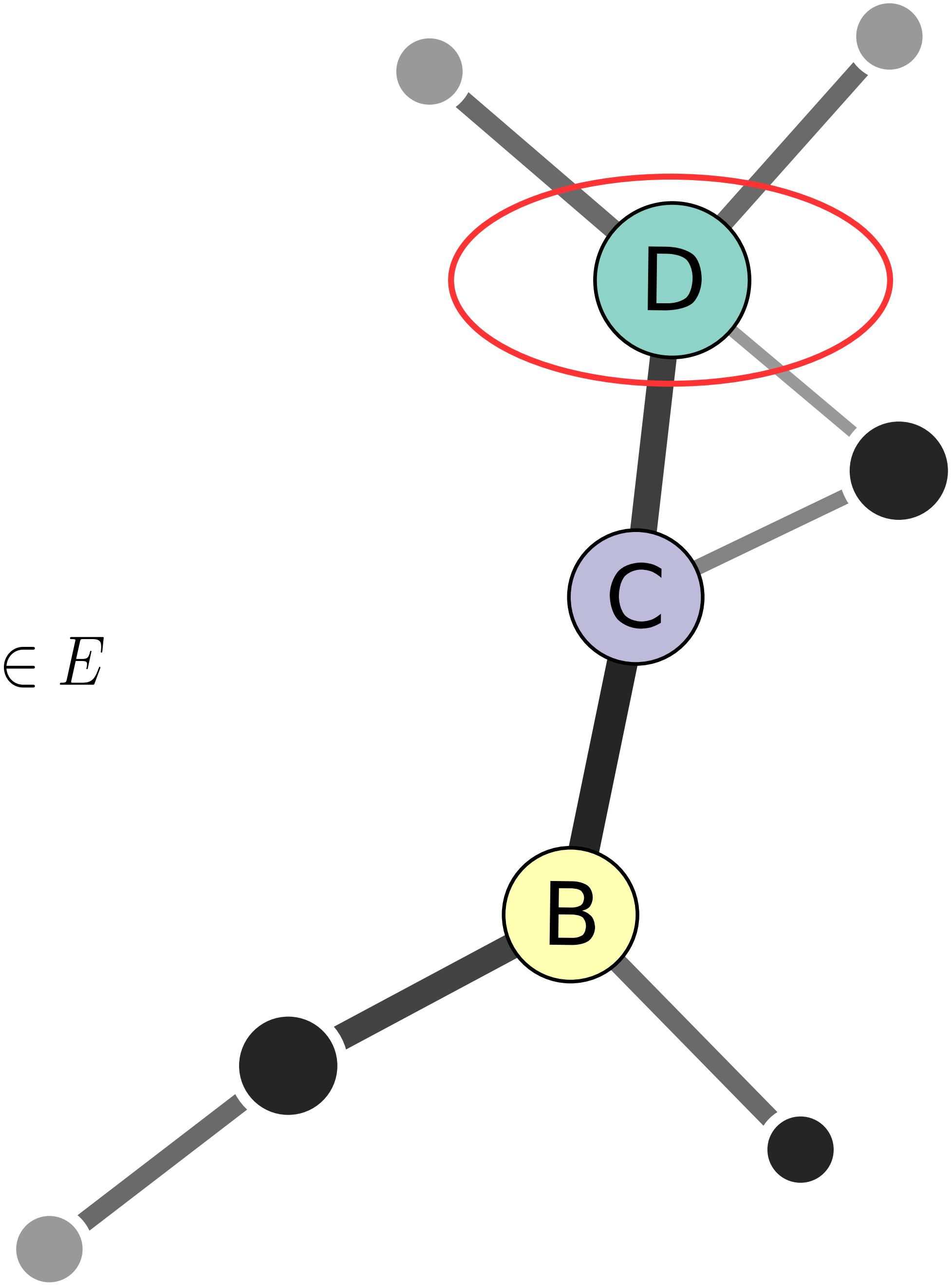


What is a Centrality measure?

Degree (K)

$$K(i) = \sum_{j \in V} x_{ij}$$

$$x_{ij} = \begin{cases} 1 & \text{if} \quad x_{ij} \in E \\ 0 & \text{else} \end{cases}$$

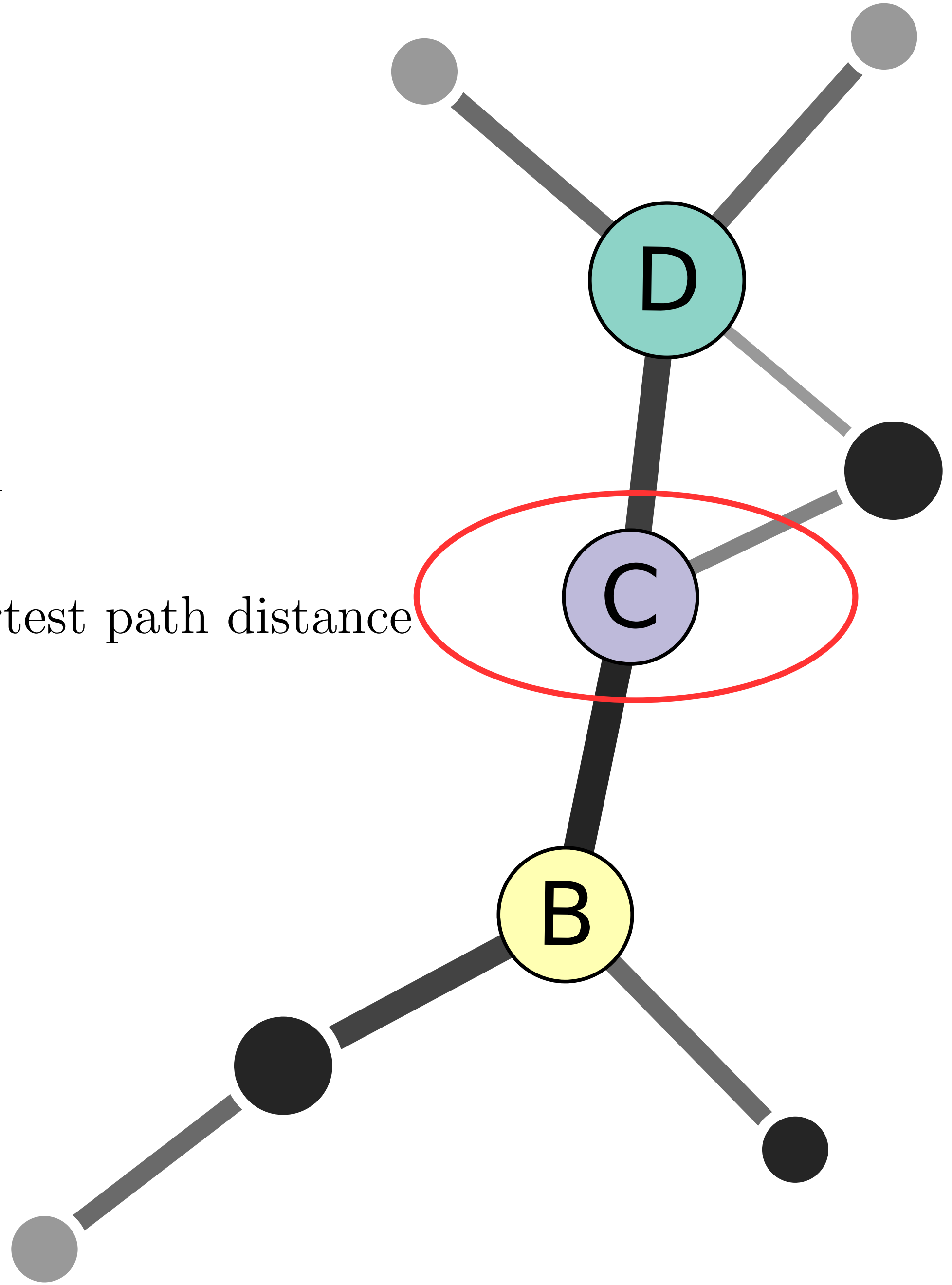


What is a Centrality measure?

Closeness

$$C(i) = (\sum_{j \in V} dist(i,j))^{-1}$$

where $dist(i,j)$ is the shortest path distance

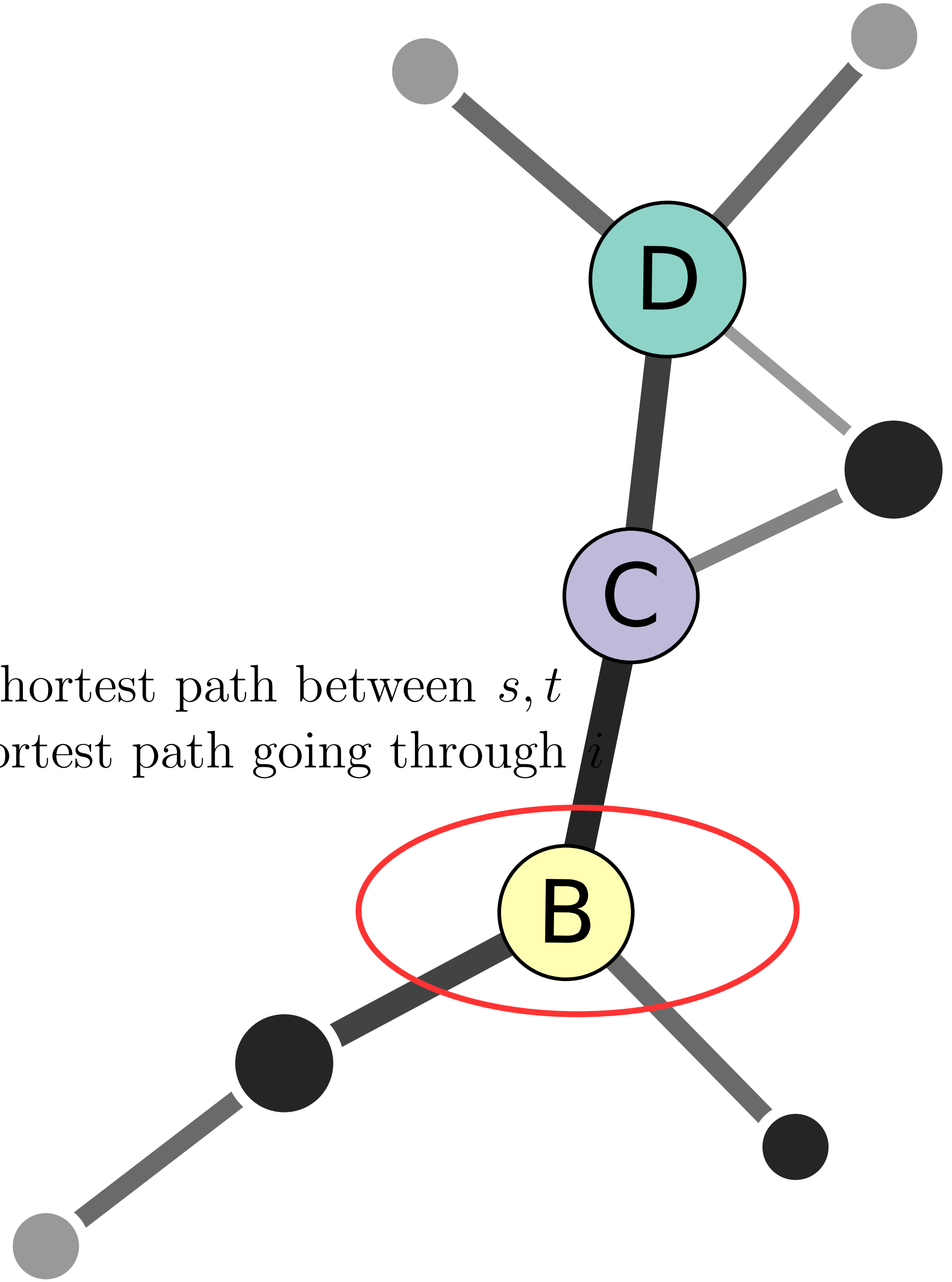


What is a Centrality measure?

Betweeness

$$B(i) = \sum_{s,t \in V} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

where σ_{st} is the number of shortest path between s, t and $\sigma_{st}(i)$ the number of shortest path going through i

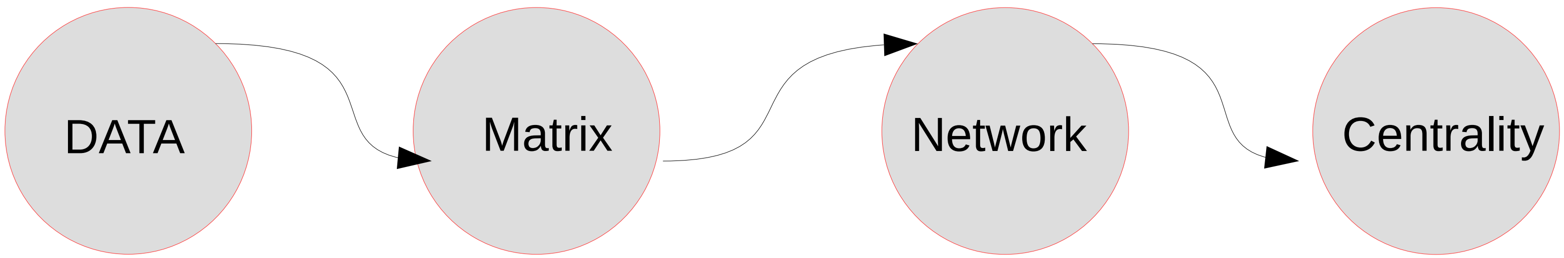


What is a Centrality measure?

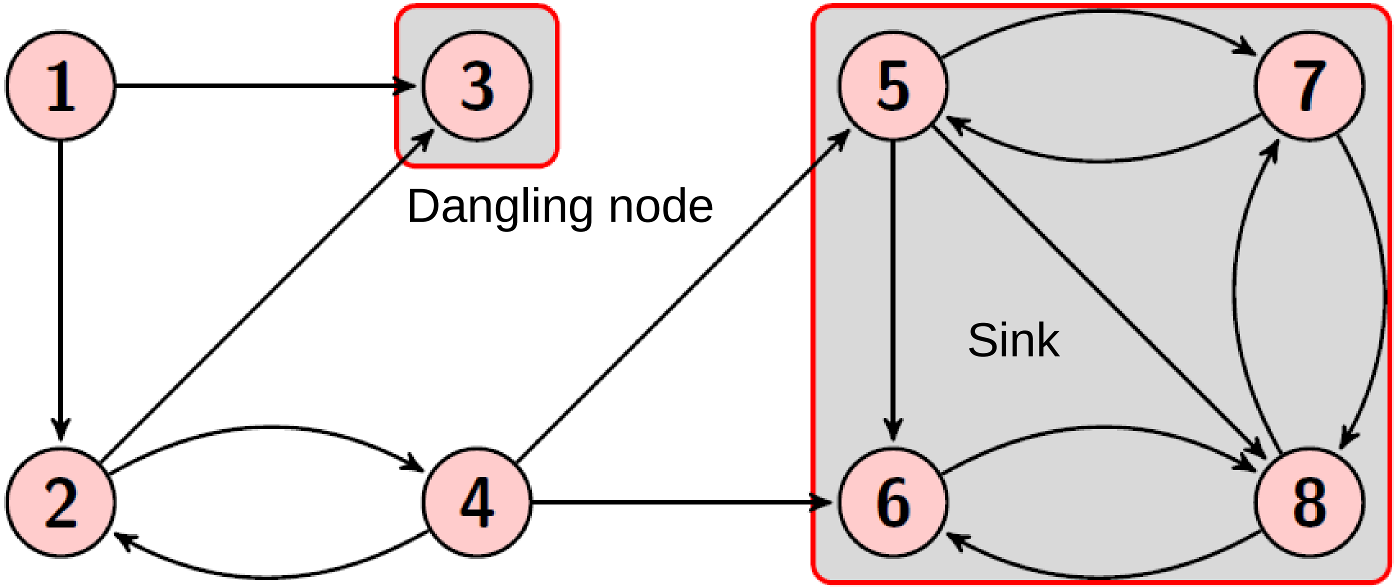
The method you use depends on what you are investigating

For directed network such that the World Wide Web, the eigenvector centrality is used

Ex: Google’s research engine algorithm



Modeling Random Walk on a network

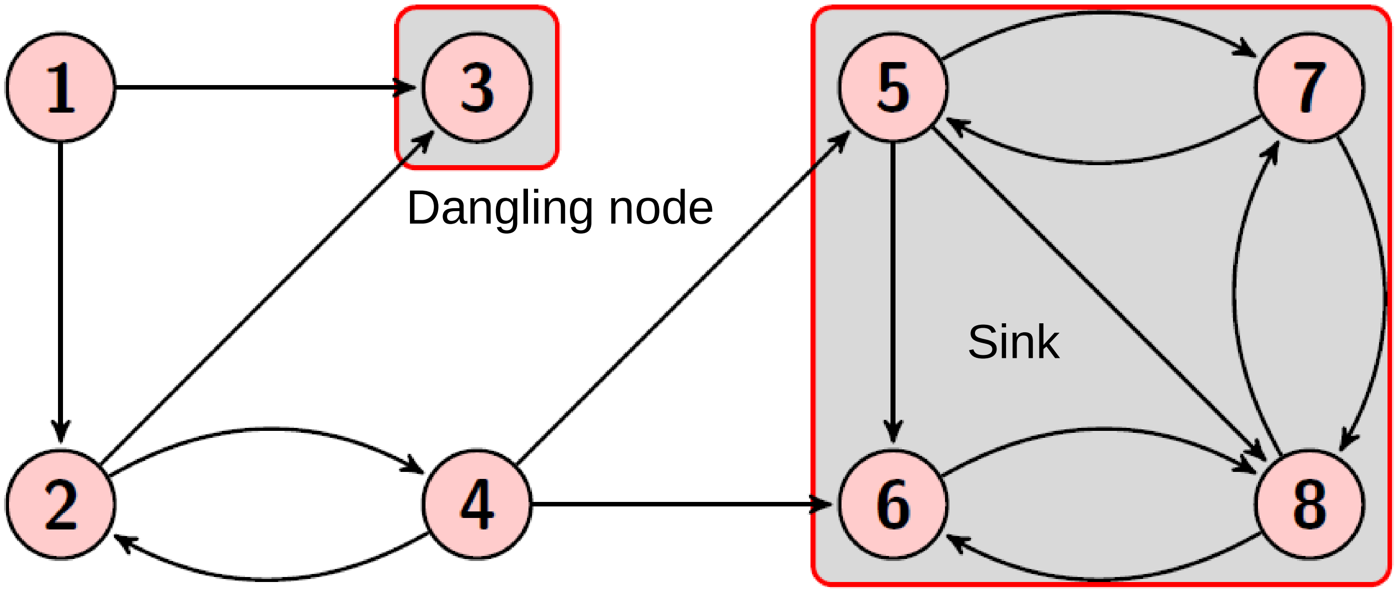


Adjacency matrix :

$$A_{ij} = \begin{cases} 1 & \text{if } j \to i \\ 0 & \text{otherwise} \end{cases}$$

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

The Stochastic Matrix

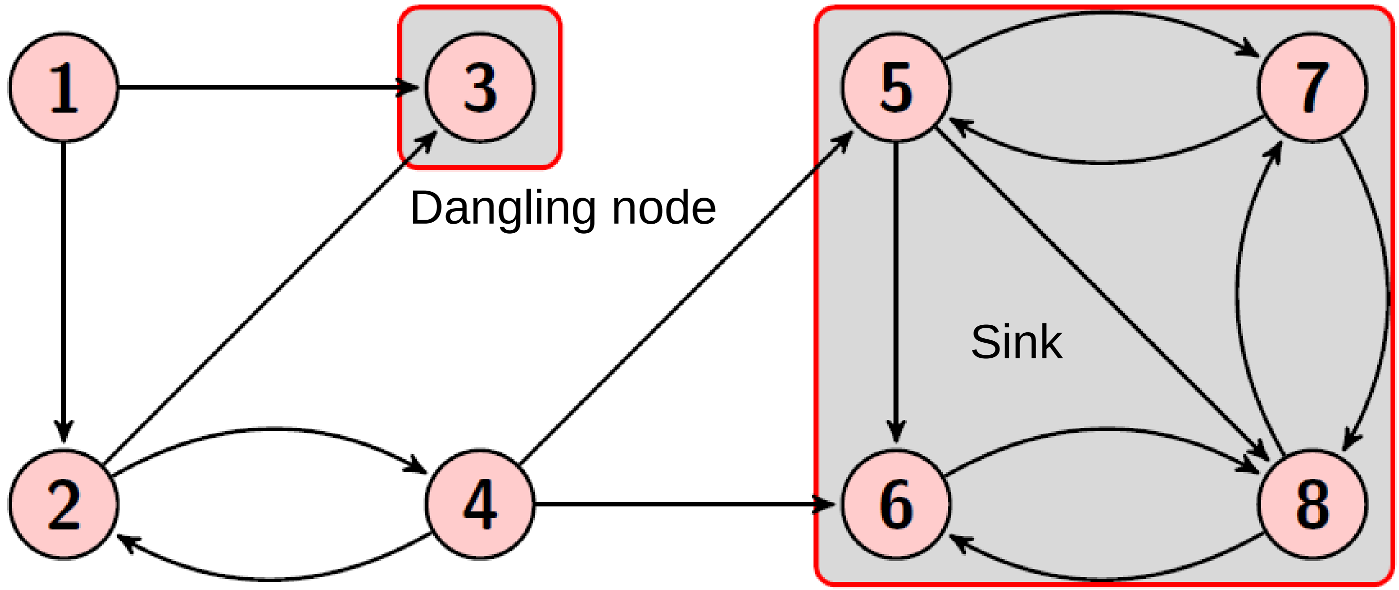


Stochastic matrix :

$$S_{ij} = \begin{cases} \dfrac{1}{N} & \text{if} j \text{ is a dangling node} \\ \dfrac{A_{ij}}{\sum_{i=1}^{N} A_{ij}} & \text{otherwise} \end{cases}$$

$$S = \begin{pmatrix} 0 & 0 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/8 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & 1/3 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/8 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 1/8 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 1/8 & 0 & 1/3 & 1 & 1/2 & 0 \end{pmatrix}$$

The Google Matrix



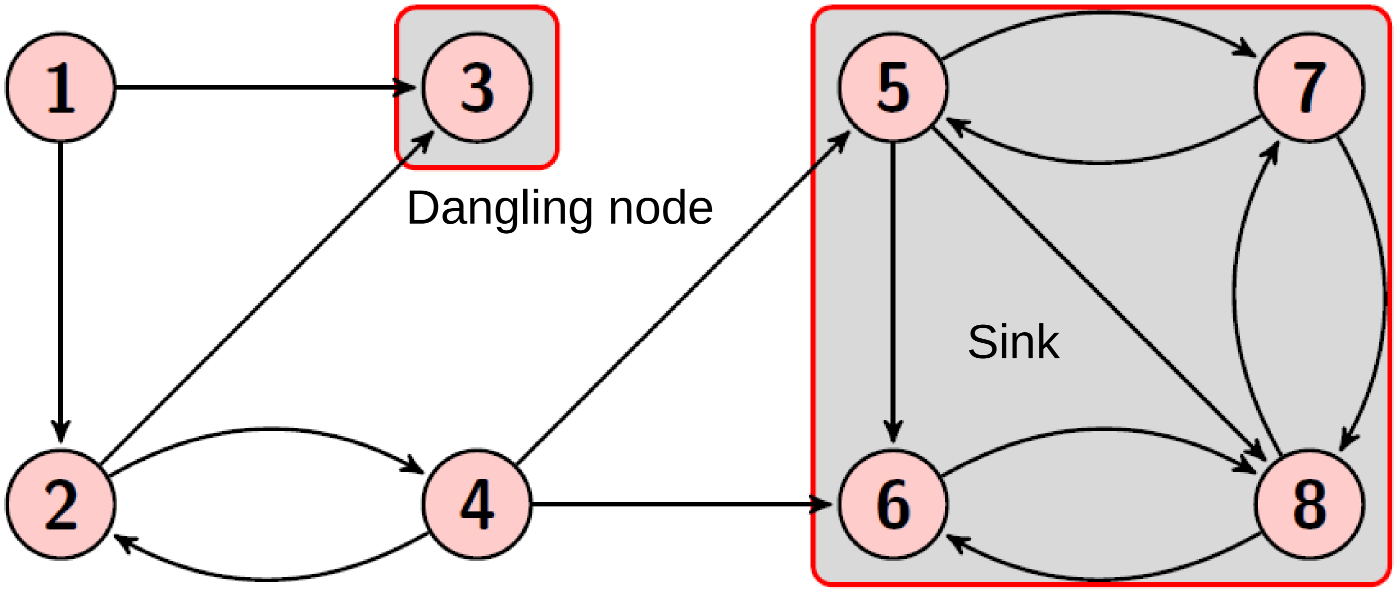
Google matrix :

$$G_{ij} = \alpha S_{ij} + (1-\alpha)/N$$

$\alpha \in [0, 1]$ is the damping factor

$$G = \begin{pmatrix} 1/40 & 1/40 & 1/8 & 1/40 & 1/40 & 1/40 & 1/40 & 1/40 \\ 17/40 & 1/40 & 1/8 & 7/24 & 1/40 & 1/40 & 1/40 & 1/40 \\ 17/40 & 17/40 & 1/8 & 1/40 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 17/40 & 1/8 & 1/40 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 1/8 & 7/24 & 1/40 & 1/40 & 17/40 & 1/40 \\ 1/40 & 1/40 & 1/8 & 7/24 & 7/24 & 1/40 & 1/40 & 17/40 \\ 1/40 & 1/40 & 1/8 & 1/40 & 7/24 & 1/40 & 1/40 & 17/40 \\ 1/40 & 1/40 & 1/8 & 1/40 & 7/24 & 33/40 & 17/40 & 1/40 \end{pmatrix}$$

The Google matrix



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

$$S = \begin{pmatrix} 0 & 0 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/8 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & 1/3 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/8 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 1/8 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 1/8 & 0 & 1/3 & 1 & 1/2 & 0 \end{pmatrix}$$

$$\alpha = 0.8$$

Fundamentals of Google research engine

$$G = \alpha S + \frac{1-\alpha}{N} e e^T \qquad\qquad e^T = (1, 1, \ldots, 1)$$

If the spectrum of the stochastic matrix S is $\{1, \lambda_1, \lambda_2, ..., \lambda_N\}$, then the spectrum of the Google matrix $G = \alpha S + (1-\alpha)\mathbf{e}\mathbf{v}^T$ is $\{1, \alpha\lambda_1, \alpha\lambda_2, ..., \alpha\lambda_N\}$, where $\mathbf{v}^T$ is a probability vector.

PageRank probability vector P

$$P = GP$$

P is the eigenvector associated to the largest eigenvalue, ie 1

For very large network (such as WWW), no way to directly diagonalize *G* → **Powermethod**

$$P = \lim_{n\to\infty} G^n v_0 \qquad \forall v_0$$

Interpretation

$$P_i = \sum_{j \in B_i} \frac{P_j}{k_{out}(j)}$$

The more a node is pointed by important node, the more it is important

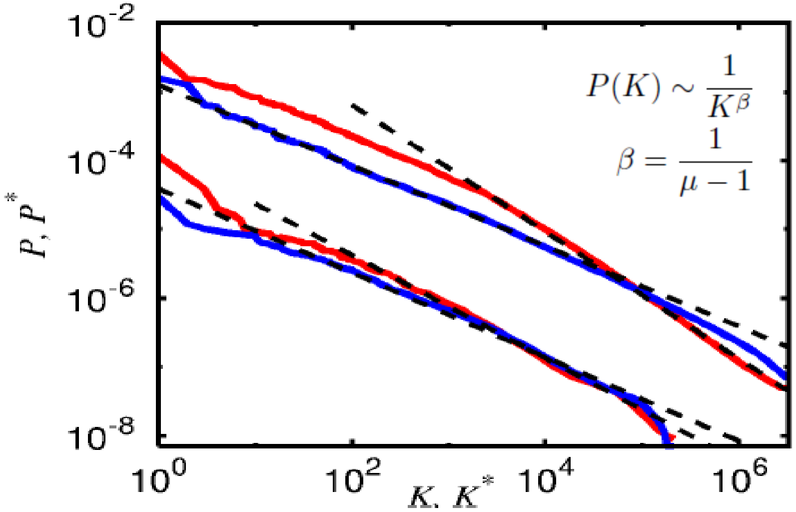
Algortithm at the hearth of Google TM search engine.

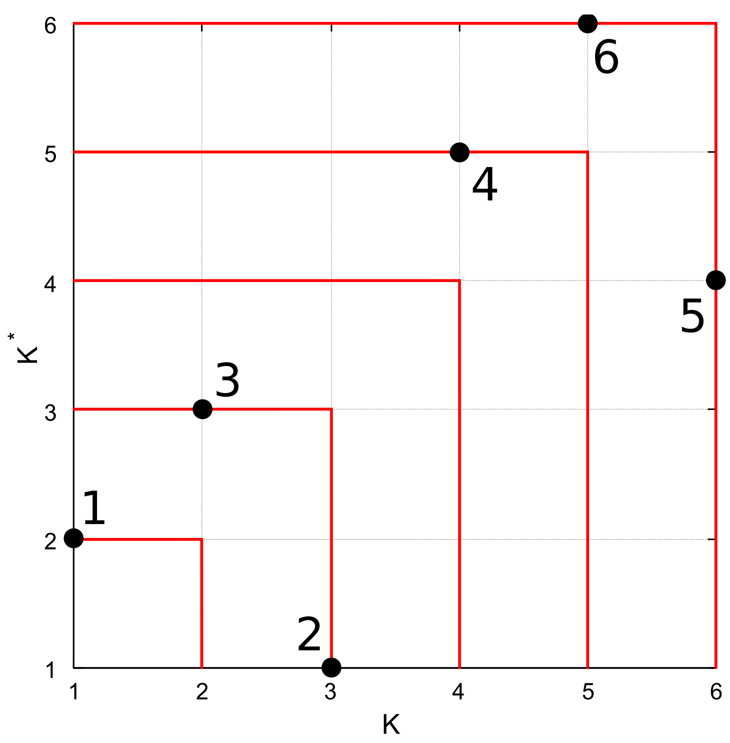
(Brin & Page cofounders)

PageRank – Measure of nodes centrality

PageRank
Measures the importance of a node as the property to be pointed by other important nodes (influent node).

CheiRank
Same as PageRank but considering the adjacency matrix of the inverted network.
Measures the importance of a node as the property to point to other important nodes = communicability





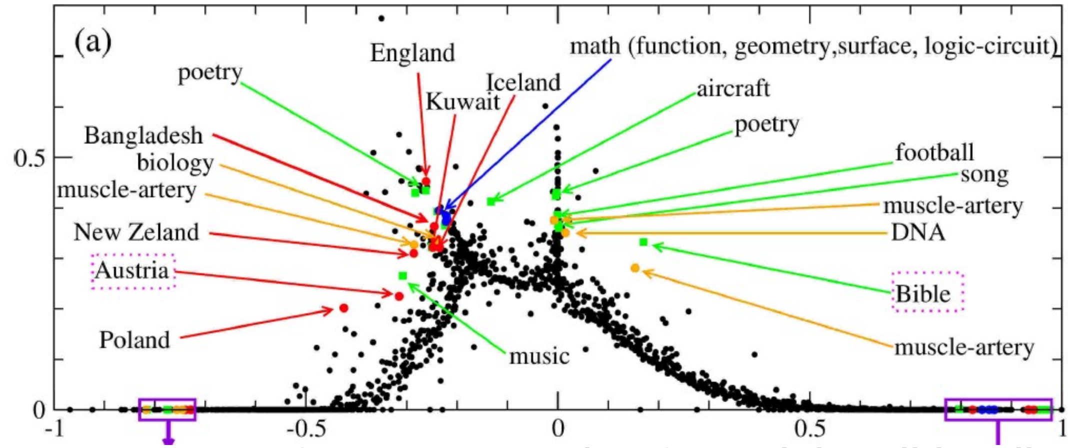
P : PageRank vector
K : PageRank
P* : CheiRank vector
K* : CheiRank

2DRank
Mix *PageRank* and CheiRank, measures the propension of a node to be pointed and to point towards other nodes.

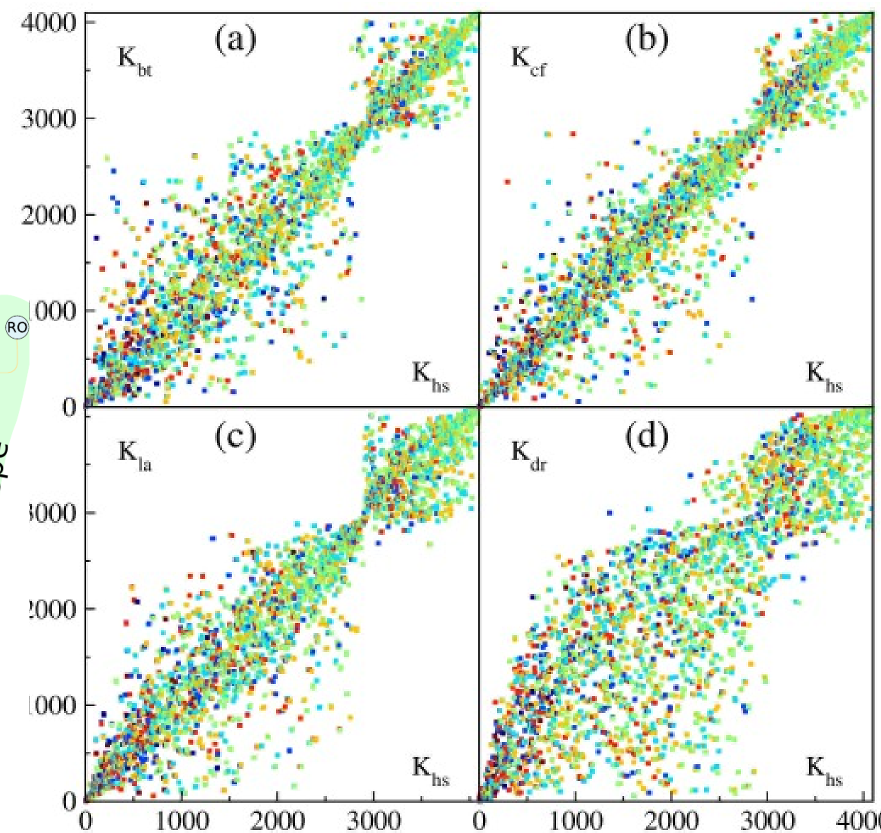
Plenty of applications in scientific research

Non exhaustive list of applications :

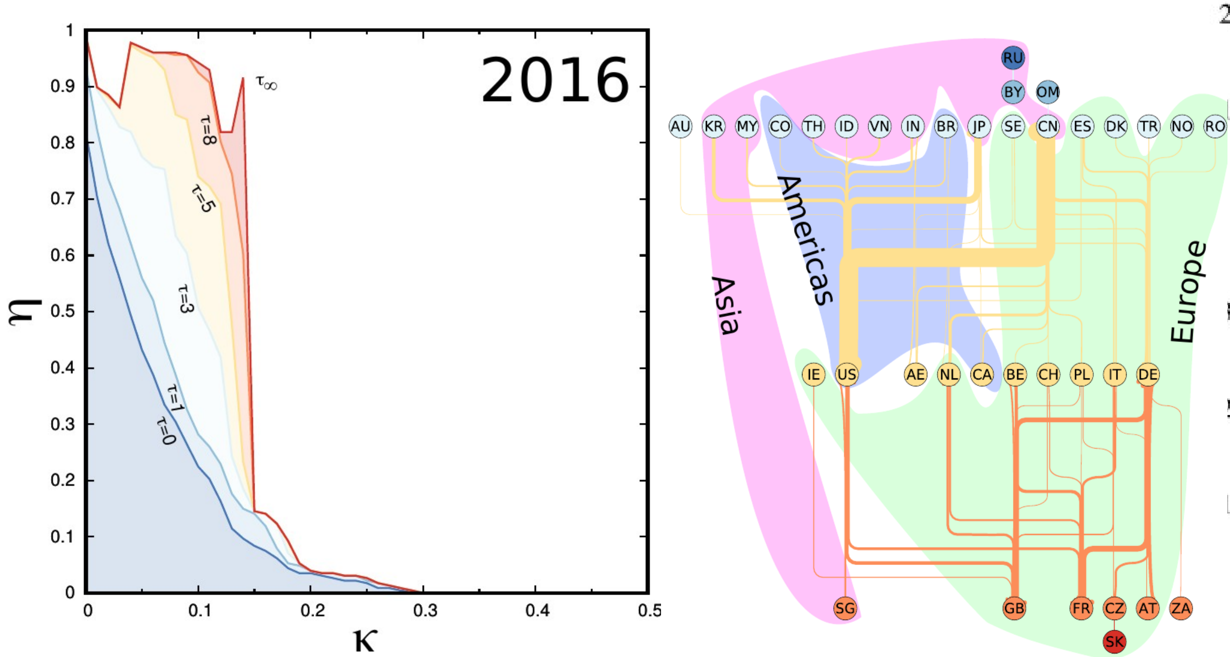
- Linux Kernel networks
- Wikipedia networks
- Twitter network
- World trade network
- Brain neural network
- DNA sequences
- Networks of Game Go moves
- Opinion formation on directed networks
- Contagion on networks
- ...



Spectrum of Google matrix of English Wikipedia



DNA sequence analysis



Crisis contagion on World trade network

IV Data and Network analysis with Python

Networkx and Pandas

NetworkX
Network Analysis in Python

Creation and modification of network objects
- From generative models
- From data

Algorithm ready to use
- Clustering
- Centrality

Network visualization

Documentation and guide available at https://networkx.org/

Installing with **`pip install networkx`**

Using with **`import networkx as nx`**

Networkx and Pandas

pandas

Data manipulation tools

    CSV to Python Dictionary

Data visualization and analysis tools

Documentations and guide available at https://pandas.pydata.org/

Installing with `pip install pandas`

Using with `import pandas as pd`

Numpy

Scientific computing for Python

Linear algebra

Documentations and guide available at http://numpy.org/

Installing with `pip install numpy`

Using with `import numpy as np`

Pyplot

Powerful Matplotlib object t generate graphics

PDF, SVG, JPG, GIF and other output formats

Documentations and guide available at https://matplotlib.org/

Installing with **pip install matplotlib**

Using with **import matplotlib.pyplot as plt**

Tutorial

Create a python script taking as input: *links.dat* and *nodes.dat*

Generate outputs A and B

A: Network representation in PDF
B: File consisting in the list of nodes with: degree, closeness and betweeness

In case of B, try you own function for degree, betweeness and closeness, before using networkx’s ones