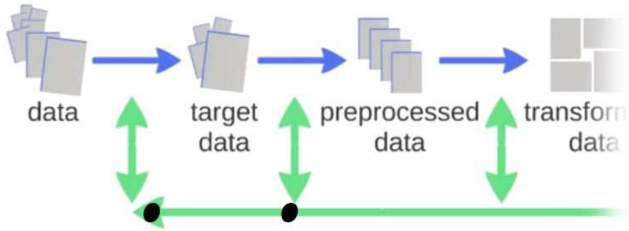
Master Data Science

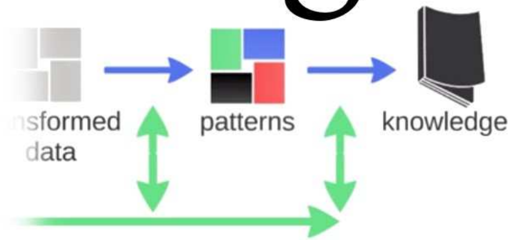
Pattern Mining
&
Social Network Analysis

Pascale Kuntz

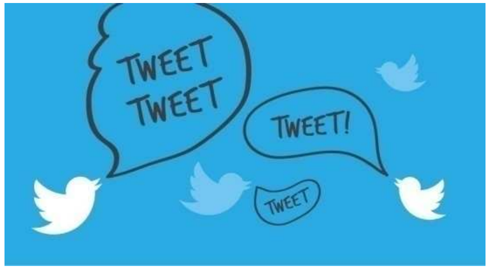
2020-2021

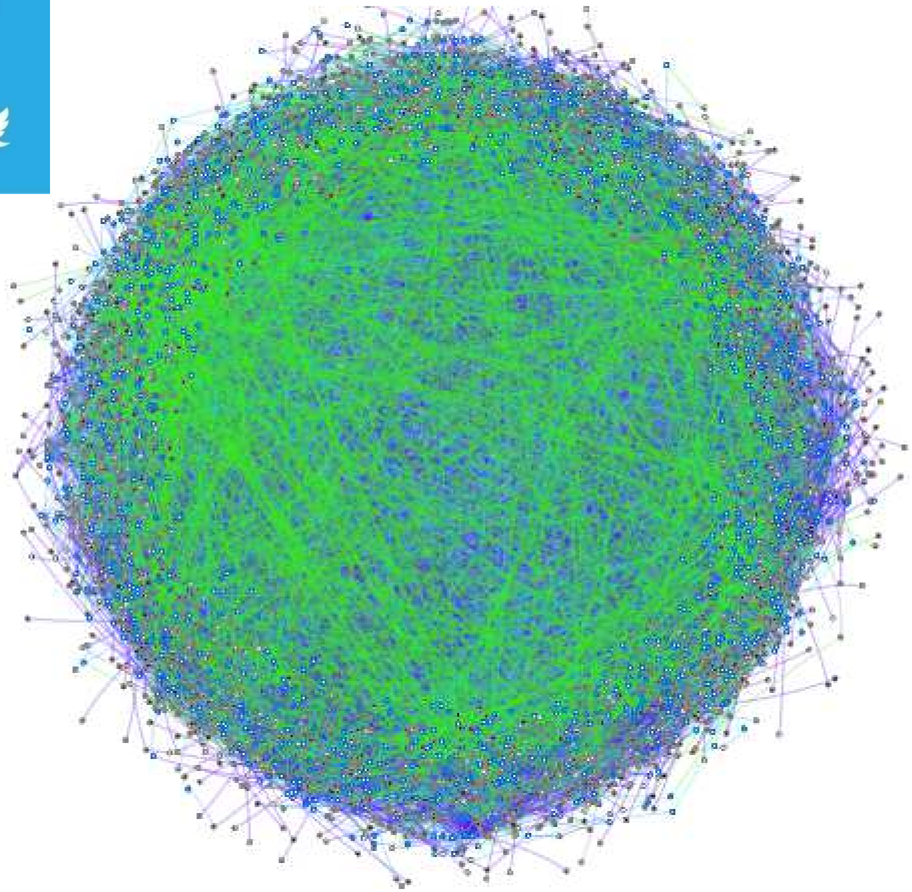
How data mining works





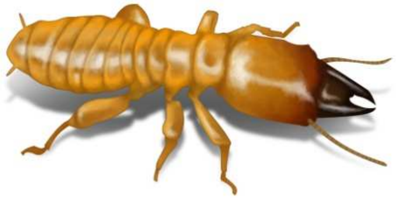
- Learning
- **Discovering**

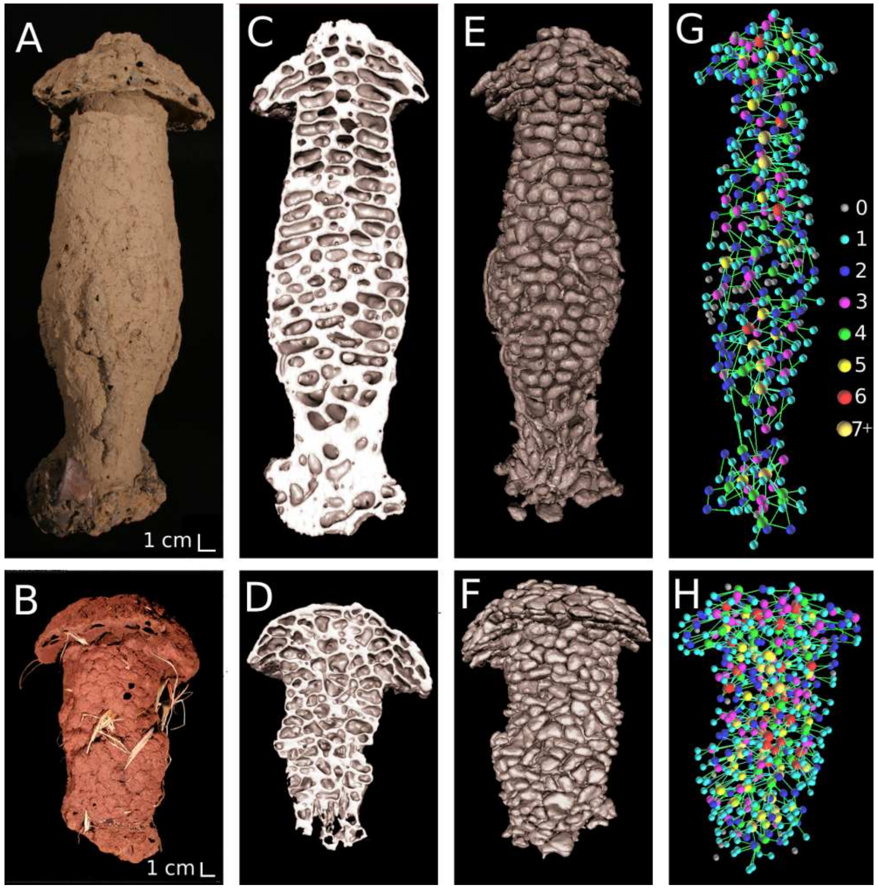




a Twitter conversation graph with #macron

How to discover hidden structures ?

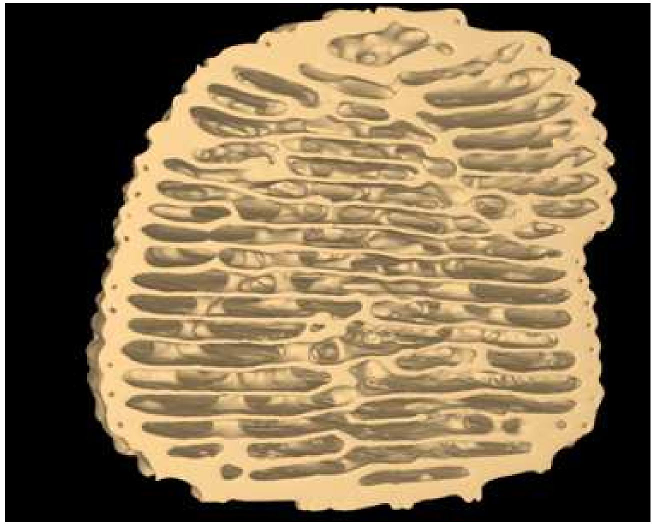


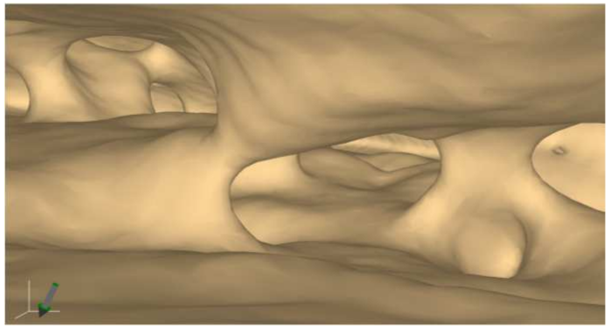


networks built by termites

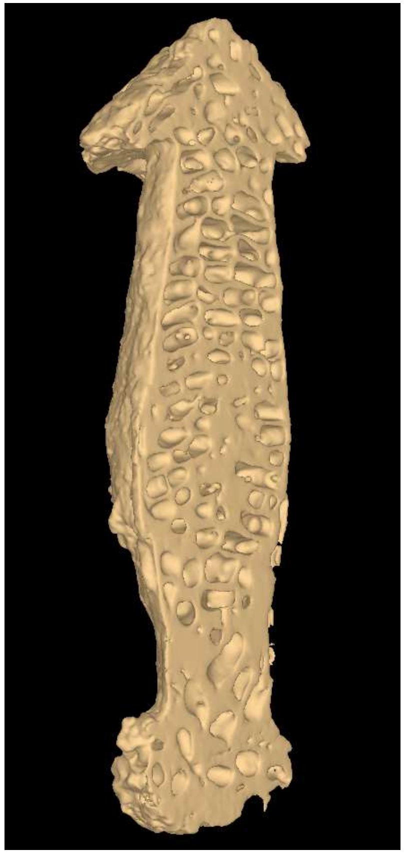
From real life data to networks 1

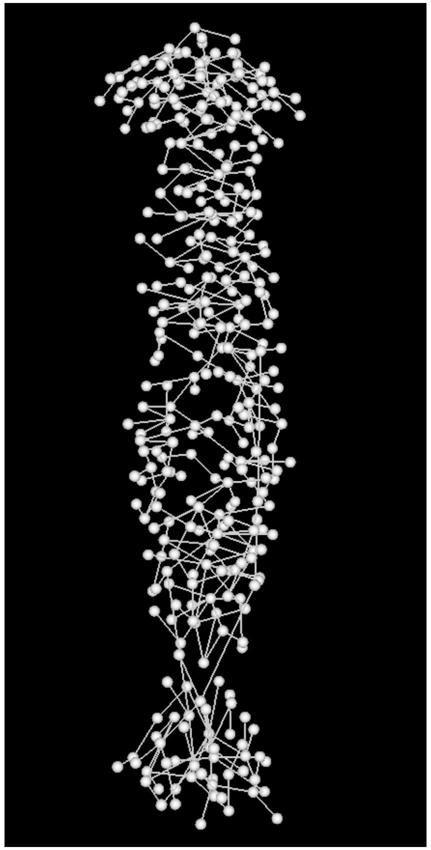
X-ray scanner





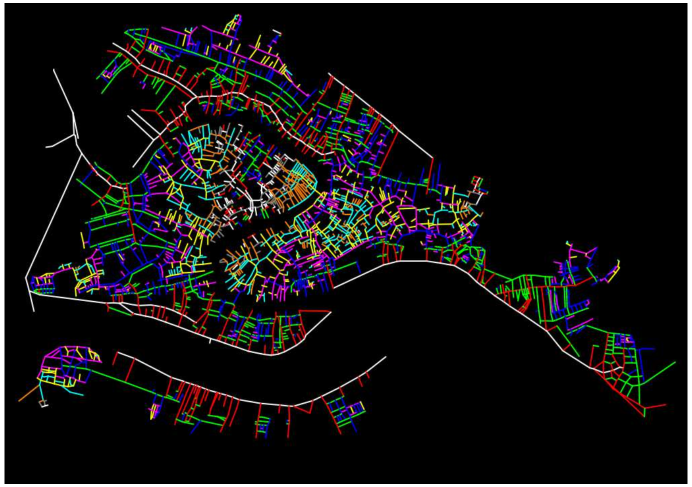
From real life data to networks 2



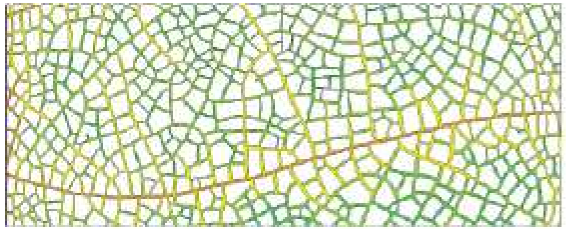


Only connections to adjacent chambers are possible

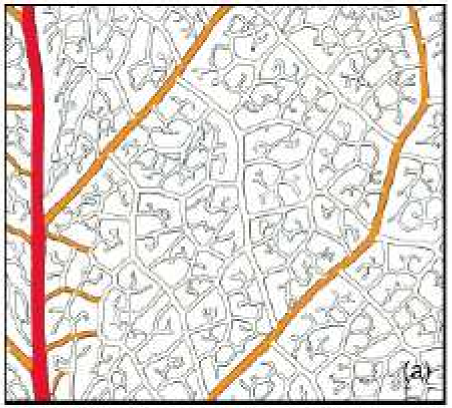
How to discover hidden structures ?



street pattern of Venice



crack patterns formed in a ceramic glaze



vein pattern of a leaf

How to characterize such patterns ?

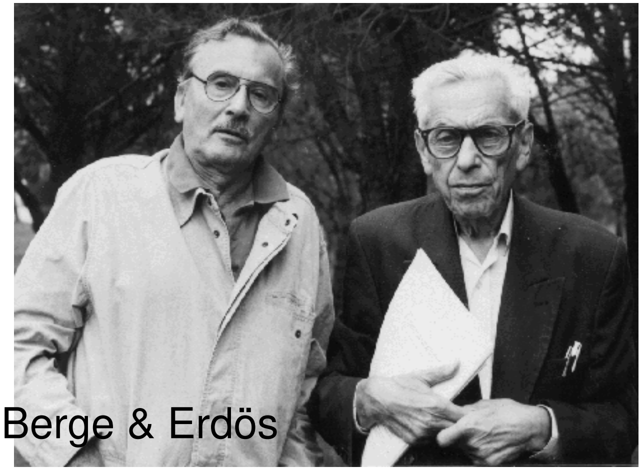
Schedule

- Introduction : Graphs (lecture)
- Graph visualization for structure discovery (lecture)
- Graph decomposition (lecture)
- Two additional autonomous sessions

- Additional lectures on Pattern Mining with F. Guillet

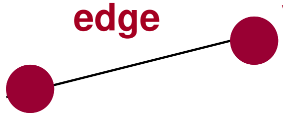
- Exam (date to be confirmed)

A graph



Berge & Erdös

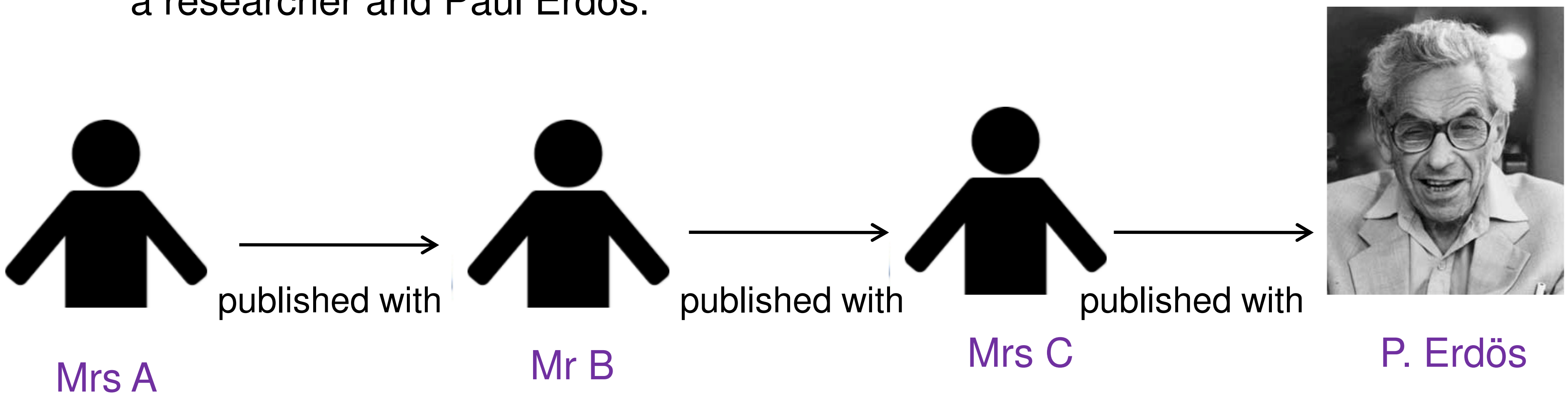
« One must imagine things we call vertices,
and for each vertex pair either an edge which joins them, or a non-edge which lets them free:
this is a **graph** according to Berge»
Rosenstiehl, M&SH, 2002



C. Berge (1958). *La théorie des graphes et ses applications*, Dunod, Paris (english edition Wiley 1961)

Erdös number

Measure of the collaborative distance in authoring academic papers between a researcher and Paul Erdös.



Erdös number = 3

https://oakland.edu/enp/

A graph

$G = (X, E)$

$X = \{\text{vertices}\}$ $\text{card}(X) = n$

E **symmetrical binary relationship on** $X \times X$: edges $\text{card}(E) = m$

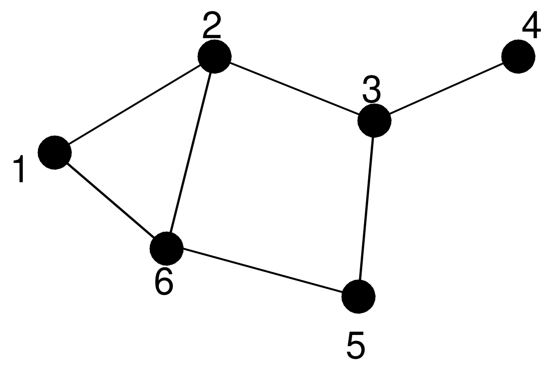
graph G

$\{1 ; 2\}$
$\{1 ; 6\}$
$\{2 ; 3\}$
$\{2 ; 6\}$
$\{3 ; 4\}$
$\{3 ; 5\}$
$\{5 ; 6\}$

A graph

graph G

{1 ; 2}
{1 ; 6}
{2 ; 3}
{2 ; 6}
{3 ; 4}
{3 ; 5}
{5 ; 6}



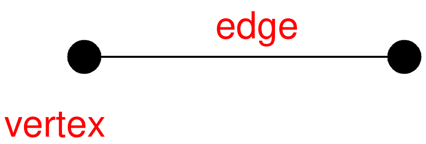
a possible drawing of G

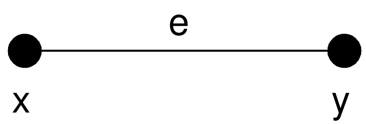
But a graph is not a drawing
It’s a combinatorial object

Graph

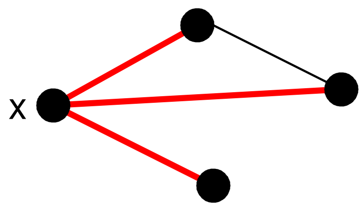
$G = (X, E)$

E **symmetrical** relationship on $X \times X$





edge e with extremities x and y

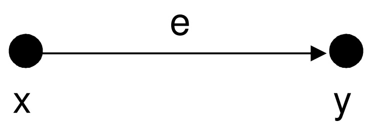


degree of x : $\delta(x) = \mathbf{3}$

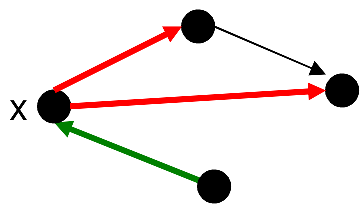
Digraph (directed)

$G = (X, U)$

U **non symmetrical** relationship on $X \times X$

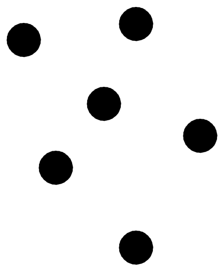


x : **origin** of the directed edge e
y : **extremity** of the directed edge e



in-degree of x : $\delta+(x) = \mathbf{2}$
out-degree of x : $\delta-(x) = \mathbf{1}$

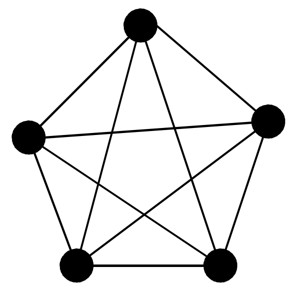
Special graphs 1



discrete graph

n

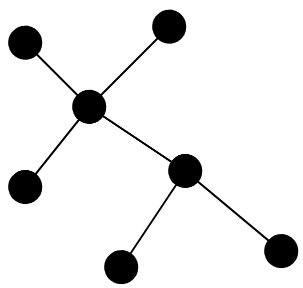
m = 0



complete graph

n

m = ?

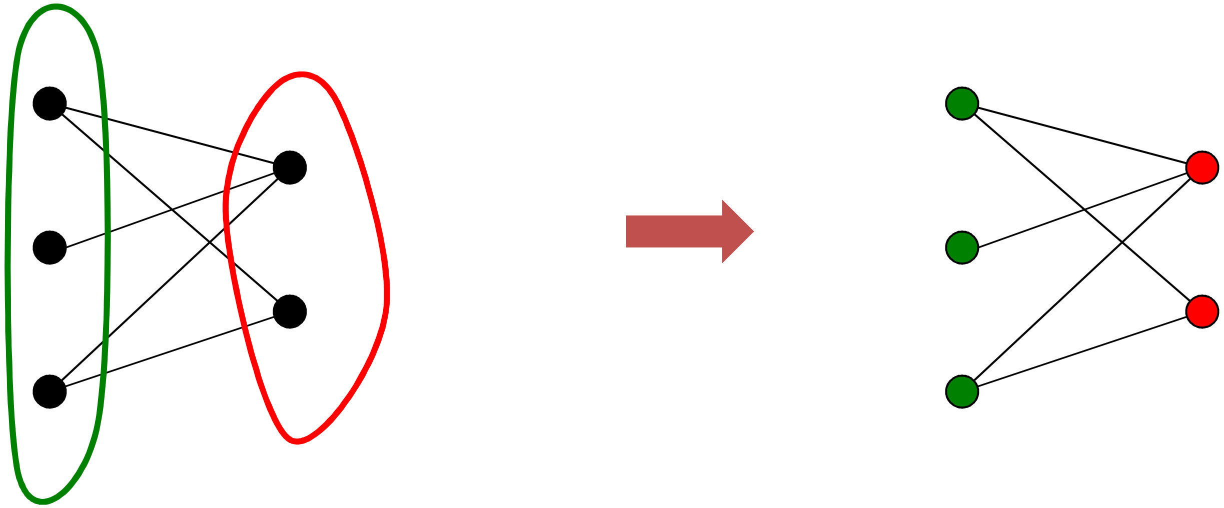


tree

n

m = ?

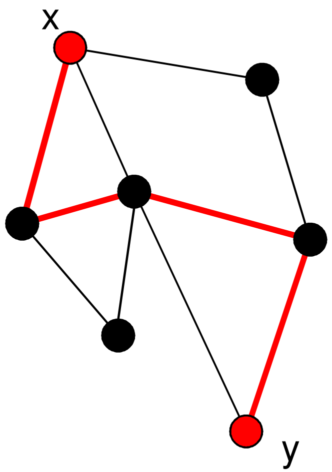
Special graphs 2



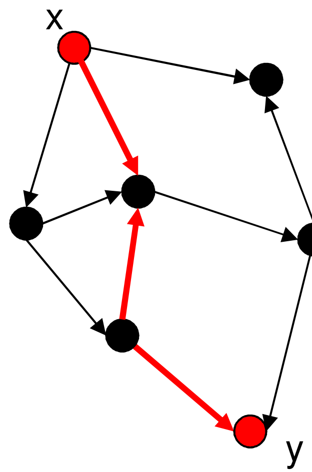
partition in 2 classes

bipartite graph
2-coloration

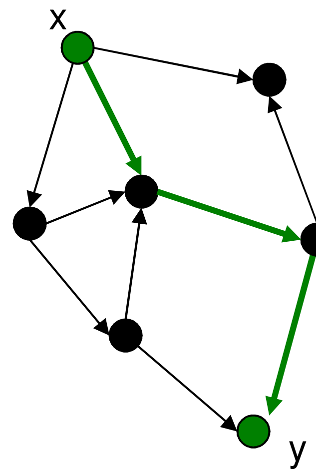
Walking on a graph



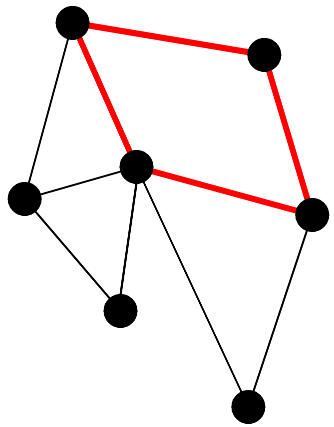
a chain



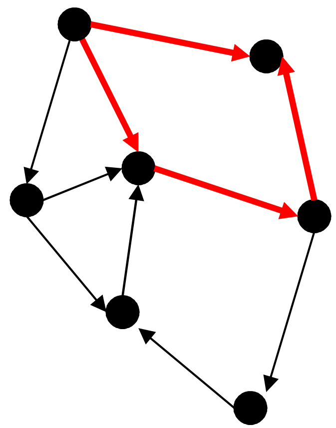
a chain



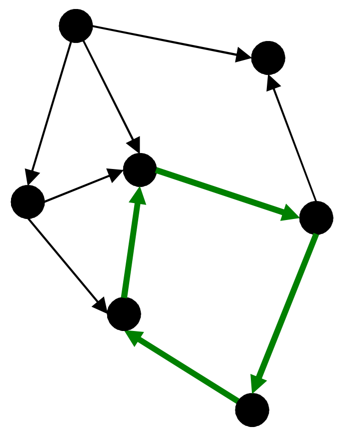
a path



a cycle



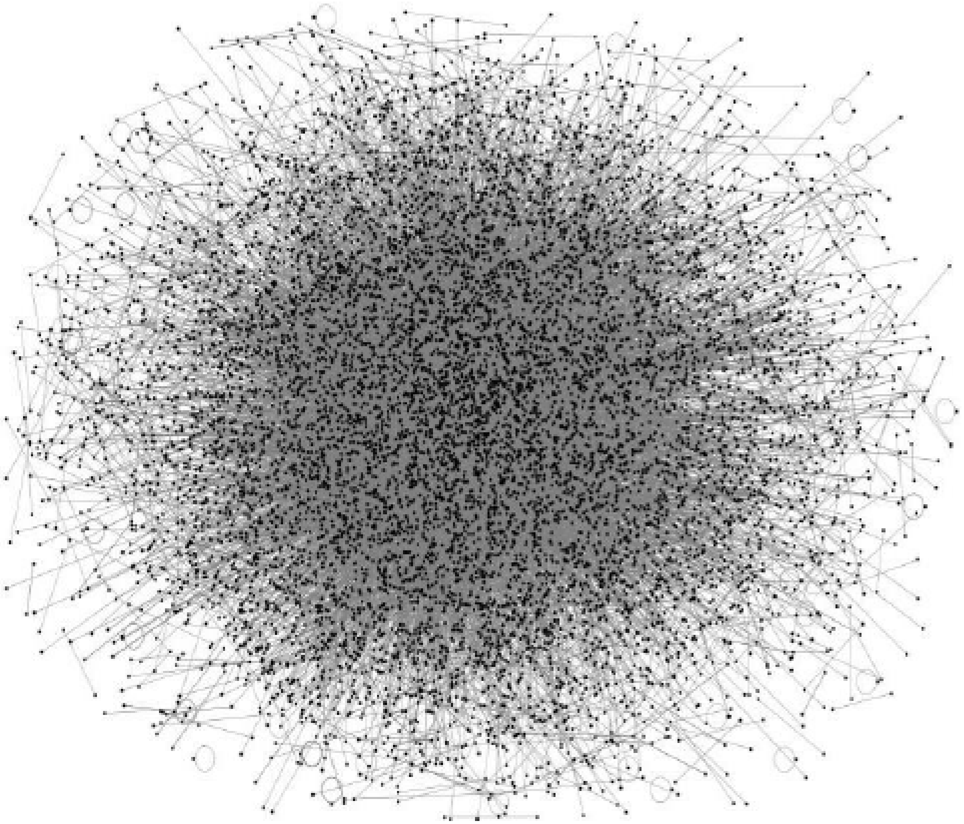
a cycle



a circuit

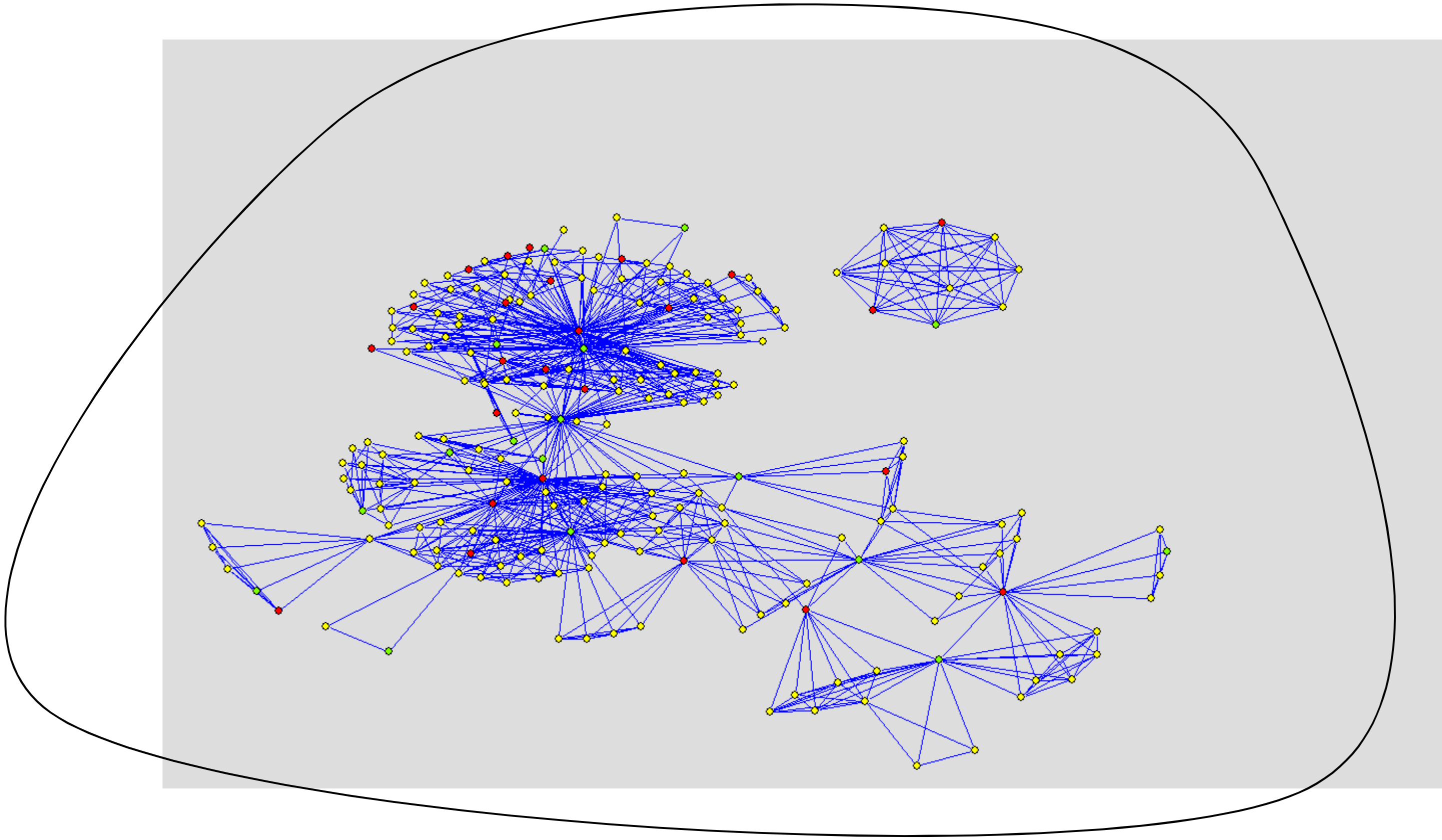
without orientation

with orientation



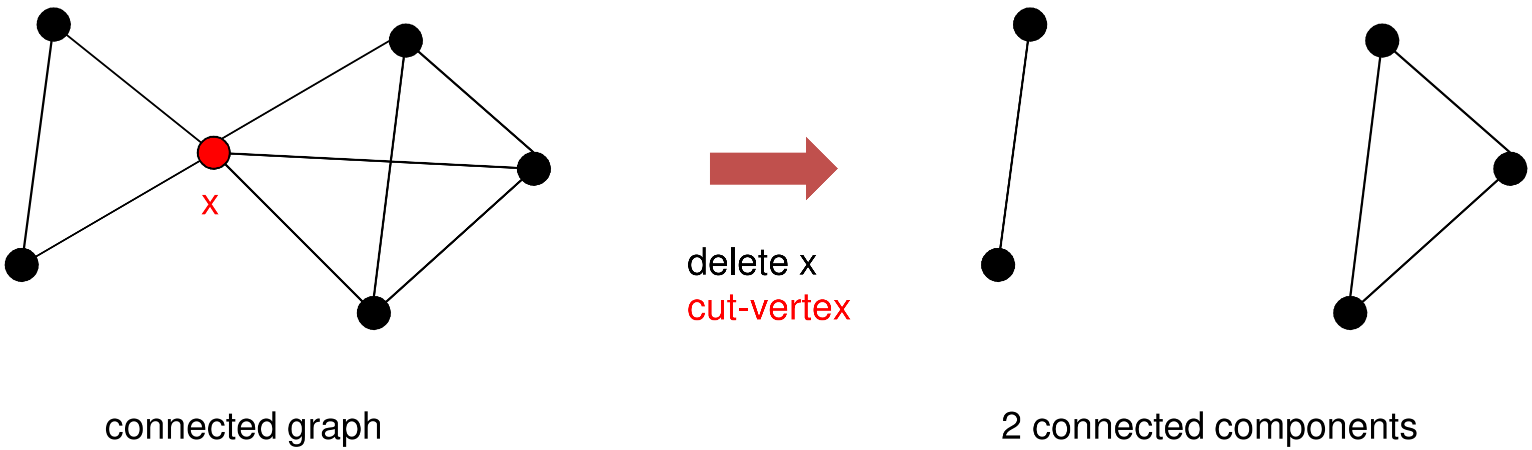
?

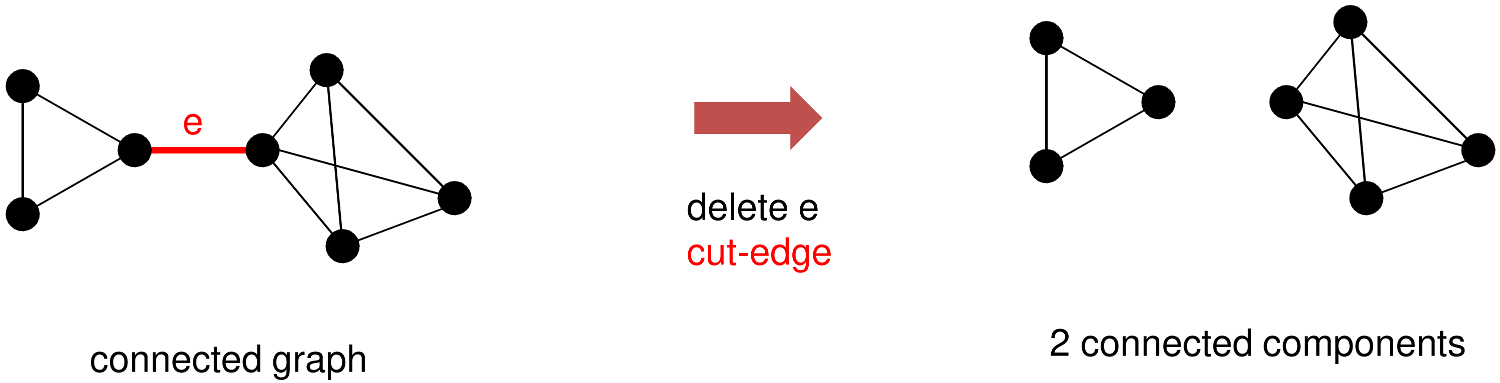
Connectivity



1 graph
2 connected components

Special vertices and edges

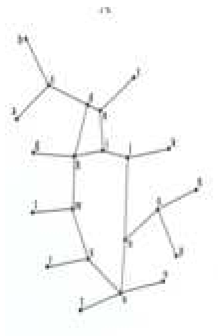




Towards a relational econometry

History

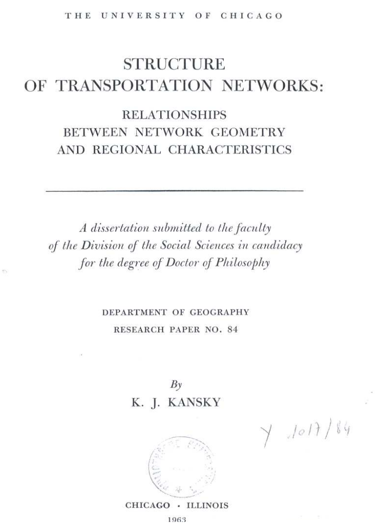
Railroad network
of Sardonia



Graphic simplification
of the network

"The communication routes determine throughout the whole territories they cut across a set of patterns of irregular outlines to which the term **network** has been very justly given"

Lalanne L. (1863). An essay on a theory of railway systems based on the observation of facts and the essential laws governing population grouping. CRAS, Paris



THE UNIVERSITY OF CHICAGO

STRUCTURE
OF TRANSPORTATION NETWORKS:

RELATIONSHIPS
BETWEEN NETWORK GEOMETRY
AND REGIONAL CHARACTERISTICS

A dissertation submitted to the faculty of the Division of the Social Sciences in candidacy for the degree of Doctor of Philosophy

DEPARTMENT OF GEOGRAPHY
RESEARCH PAPER NO. 84

By
K. J. KANSKY

CHICAGO · ILLINOIS
1963

"Graph theoretic measures (…) were developed in harmony with several rules of the theory of measurements.
The measures were designed so that:
(a) the indices express relationships between the numbers and the objects of properties to which they are assigned;
(b) the measures establish a metrical order among different transportation networks and among particular properties of these networks;
(c) the same individual resulting values express the same state in different transportation systems "

Kansky, 1963

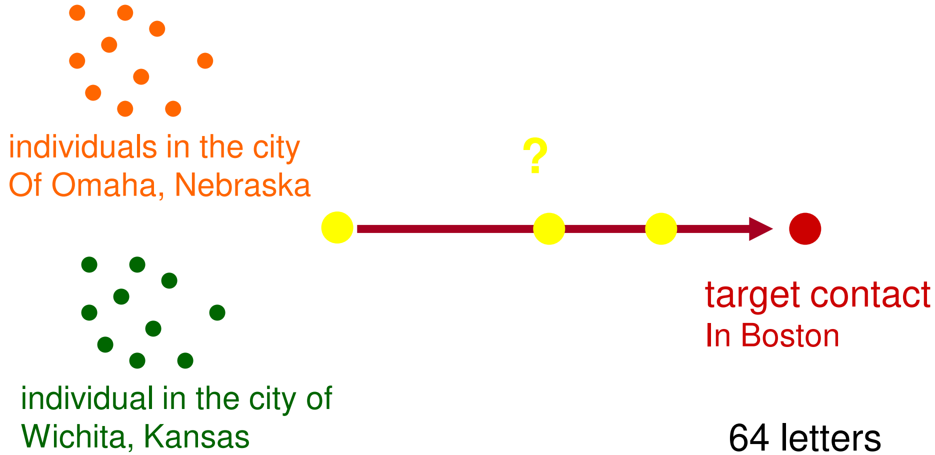
Towards a relational econometry

History : social psychology

Letter path from Nebraska to Boston

Stanley Milgram (1967)

Average path lenght for social networks of people in US



64 letters

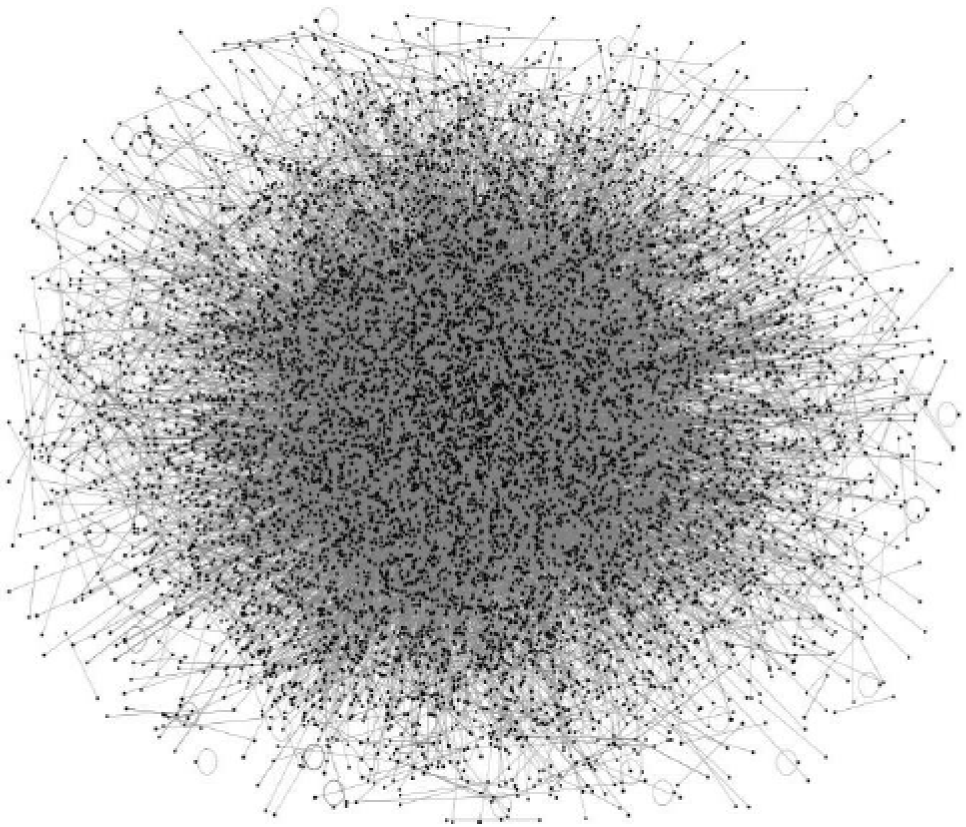
If the person did not personally known the target, then he/she has to think to a friend or relative who was more likely to know the target. Then he/she forward the letter to this relative and so on …
When the letter reached the contact in Boston , the researches countered the number of times it has been forwarded from person to person

path average : 5.5

« people in US are separated by about 6 people on average »

facebook 4.74
novembre 2011, *Anatomy of Facebook*

Relational econometry

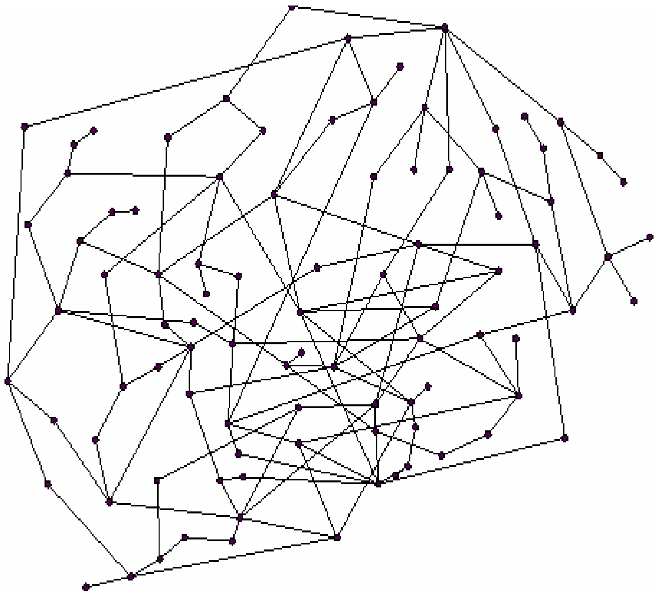


- Degree distribution
- Degree correlation
- Local and global connectivity
- Centrality
- Robustness

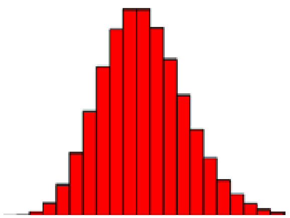
etc …

Degree distribution 1

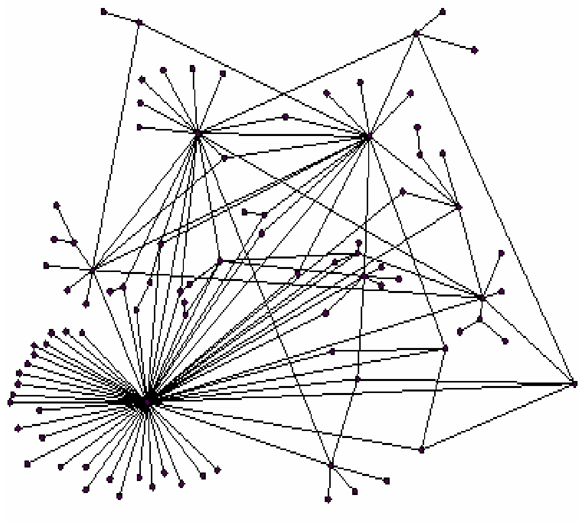
- P(k) : frequency of vertices of degree k



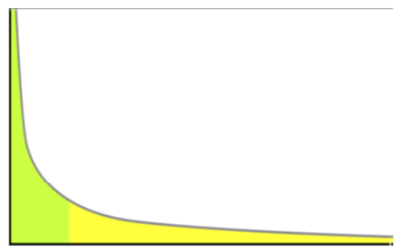
Random network



Binomial distribution
asymptotically Poisson distribution



Scale-free network



$P(k)=k^{-\alpha}$

Power-law distribution

scale free $f(ax) = cf(x)$

Random graph

Erdös-Renyi model G(n,p)

Connect vertices randomly : include each edge in the graph with probability p
independently from the other edges

As p increases from 0 to 1, the model becomes more and more likely to include graphs with more edges

Properties.A graph G(n,p) has on average $\binom{n}{2}p$ edges.
The distribution of the degree of any particular vertex is *binomial* and this distribution is *Poisson* for large n and np = const.

Degree distribution 2

Examples of power law distributions $P(k)=k^{-\alpha}$

- Internet (Autonomous Systems) (Faloutsos et al. 99)
- Web graph (Broder et al 00)
- Actor collaborations (Barabasi-Alpert 00)
- Online social networks (Leskovec et al 07)

Scale free networks (scale free $f(ax) = cf(x)$)

- Great variability of connectivity ~~(average as typical element)~~
- Many vertices with few connections and few vertices highly connected
- «The scale-free topology is evidence of organizing principles acting at each stage of the network formation» (Barabasi 05)

Power laws 1

History : economy

Vilfredo Pareto

Cours d'économie politique
Lausane, 1864

Schedule D — Année 1893-94.

x £	N GREAT BRITAIN	N IRELAND
150	400 648	17 717
200	234 185	9 365
300	121 996	4 592
400	74 041	2 684
500	54 419	1 898
600	42 072	1 428
700	34 269	1 104
800	29 311	940
900	25 033	771
1000	22 896	684
2000	9 880	271
3000	6 069	142
4000	4 161	88
5000	3 081	68
10000	1 104	22

Taxpayer declaration
for income tax

x : a given income
N : number of taxpayers with an income greater than x

$$\log N = \log A - \alpha \log x$$

Power laws 2

« We are faced with a universal law, as universal at least as the Laplace-Gauss law. (…) The main explanation of this universality is given by a theory developed by Paul Levy in probability calculus: the theory of stable laws (…) »
(M. Barbut in *La mesure des inégalités* 2007)

Where does the privilege according to the « normal law » come from ?

1- Central limit theorem
2- Stability for the addition: the sum of gaussian independent random variables is still a gaussian variable

Question ?
Which distributions are stable for the addition and which ones can approximatively be the average of independent random variables ?

Answer (Paul Levy *Theory of the addition of random variables* 1936)
Besides the Laplace-Gauss law there is a family of laws which are all asymptotically paretian.

Degree correlation 1

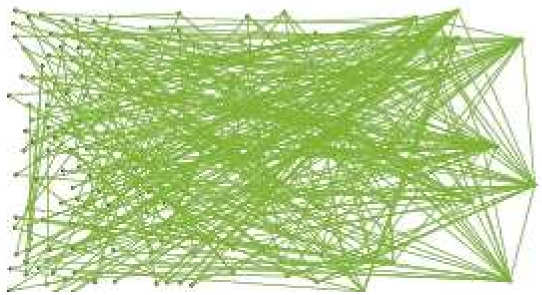
Assortativity

Pearson correlation coefficient r of the degrees between pairs of linked vertices

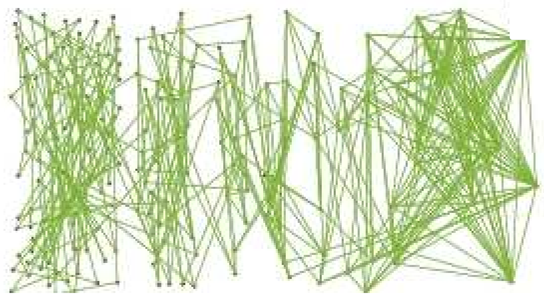
when r > 0 : relationships between vertices of similar degree
when r < 0 : relationships between vertices of different degree
when r = 1 : the network is said to have perfect assortative mixing patterns
when r = 0 : the network is non assortative

	network	n	r
real-world networks	physics coauthorship[a]	52 909	0.363
	biology coauthorship[a]	1 520 251	0.127
	mathematics coauthorship[b]	253 339	0.120
	film actor collaborations[c]	449 913	0.208
	company directors[d]	7 673	0.276
	Internet[e]	10 697	−0.189
	World-Wide Web[f]	269 504	−0.065
	protein interactions[g]	2 115	−0.156
	neural network[h]	307	−0.163
	food web[i]	92	−0.276
models	random graph[u]		0
	Callaway *et al.*[v]		$\delta/(1+2\delta)$
	Barabási and Albert[w]		0

Neyman, 2002



r = 0



r = 0.43

Degree correlation 2

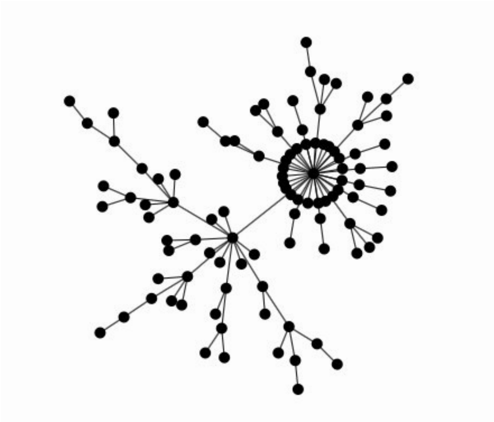
The Mattew effect … or « the rich get richer »

Gospel according to St Matthew : *For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath.*

Preferential attachment (Price 65, Albert-Barabasi 99)

probability of linking a new vertex to vertex i is proportional to its degree δ_i

A vertex that acquires more connections than an other will increase its connectivity at a higher rate.
Thus, an initial difference in the connectivity between two nodes will increase further as the network grows



Preferential attachment

Barabasi-Albert model

Generating random scale-free network using a preferential attachment

Algorithm

•Buid an initial connected network with n_0 vertices

•Add new vertices one at a time. Each new vertex is connected to m < m_0 existing vertices with a probabililty that is proportional to the number of links that the existing vertices already have

Probability p_i that the new vertex is connected to vertex i : $p_i = \frac{\delta_i}{\sum_j \delta_j}$

where δ_i is the degree of vertex i and j is the number of pre-existing vertices

Properties.

•The new vertices have a « preference » to attach themselves to the already heavily linked vertices.

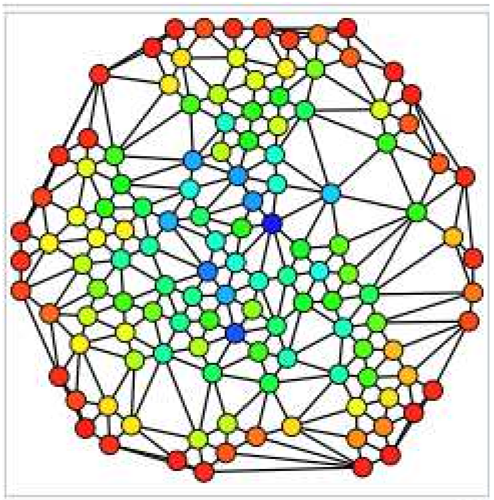
•The degree distribution is scale free.

Centrality

Betweeness centrality of a vertex : the number of times a vertex acts as a bridge along the shortest path between two other vertices

$$C_B(i) = \sum_{k \neq l \neq i} \frac{S_{kl}(i)}{S_{kl}}$$

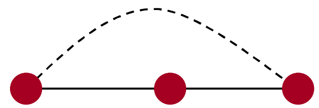
where S_{kl} is the total number of shortest paths from vertex k to vertex l and $S_{kl}(i)$ is the number of those paths that pass through v



Hue (from red = 0 to blue = max) shows the vertex betweeness

Connectivity 1

Local connectivity



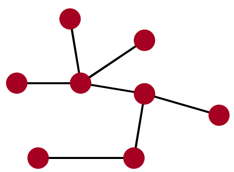
The friends of my friends are my friends.

Clustering coefficient C = density of cycles of length 3

Global connectivity

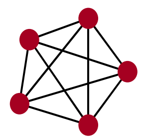
Diameter : maximum of the shortest paths (min number of vertices)
Shortest path average $\langle L \rangle$

tree



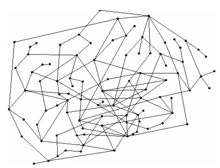
$C = 0$

complete graph



$C = 1$
$\langle L \rangle = 0$

random graph

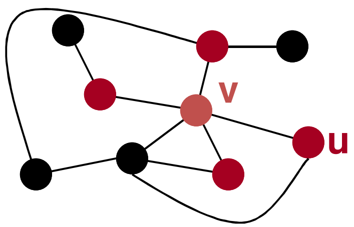


$C \approx \langle\delta\rangle/n$
$\langle L \rangle = \log(n)/\log(\langle\delta\rangle)$
$\langle\delta\rangle$: mean degree

Connectivity 2

Global connectivity

Efficiency Efficiency of the information exchange between two vertices: inversely proportional to the shortest path distance between them



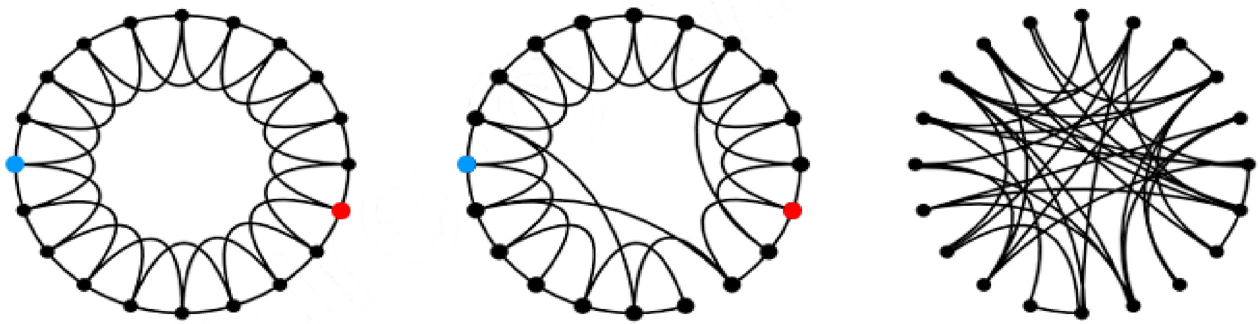
Local efficiency E(v): average of $1/d_{uv}$ on the neighbours u

Global efficiency E: average of E(v)

Small world graphs 1

The simplest way of formulating the small world problem is “what is the probability that two any people, selected arbitrary from a large population (…) will know each other?” Travers & Milgram (1969)

P=0 P=1 increasing randomness



regular graph | **small world graph** (Watts-Strogatz 98) | random graph

high lobal connectivity
high glocal connectivity

high local connectivity
low global connectivity

low local connectivity
low global connectivity

Look **locally** like regular graphs
globally like random graphs

Small world graphs 2

Examples:

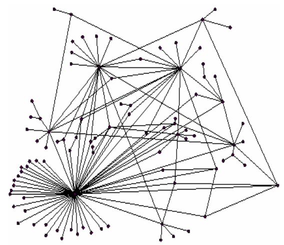
- networks of connected proteins
- networks of brain neurons
- word co-occurrence networks
- telephone call graphs
- electric power grids
- etc etc

Robustness

Robustness: effects of vertex deletion

Two types of removals :

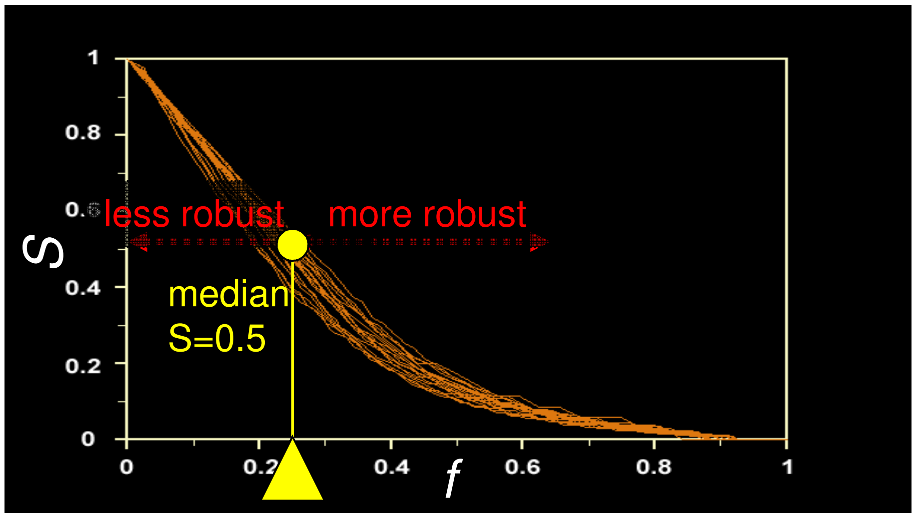
- Random removals
- Selective removals (targeted attacks)



Scale free network

High resilience to random removals
High vulnerability to selective removals

%size of the largest connected component S(i)/n



Fraction of disconnected vertices : f = i/n (i=1, 2, …,n)