Chapter 5 :

Réseaux complexes / Complex networks

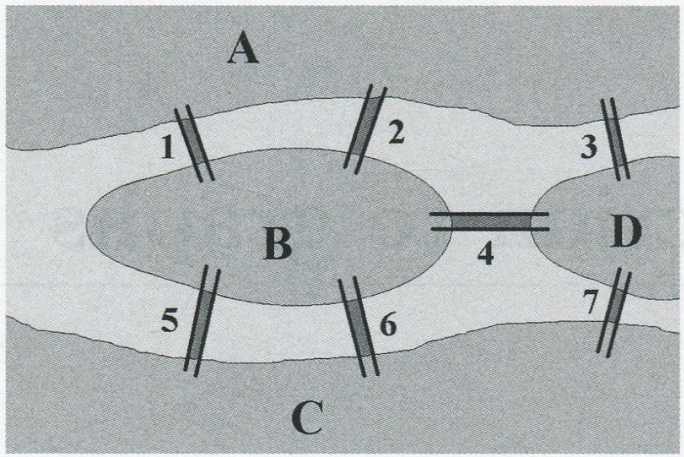
G. Caldarelli, Scale-Free Network, Oxford University Press, 2007
M.E.J. Newman, Network, an Introduction, Oxford University Press, 2010

Overview

- Motivation to use graphs in science
- Definitions and mathematical description
- Dynamical processes on a graph
- Properties of complex networks
- Centrality measure
- Assortativity and modularity
- Communities
- Algorithms to generate complex networks

5.1 Motivations and examples

- A graph (network) is a mathematical abstraction describing a relation between entities.
- Origin : Euler and the Königberg's bridges problem
- Is it possible to find a closed tour that uses all the bridges, once and only once ?



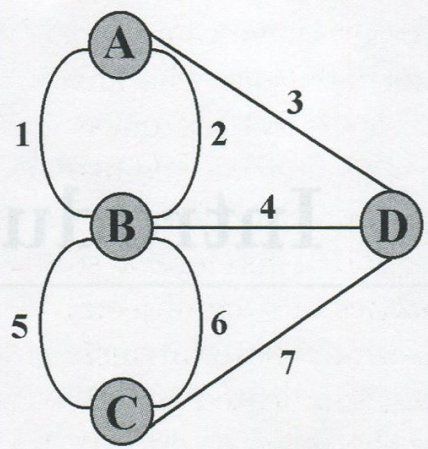


Fig. 1.2 On the left a very schematic map of the town centre of Königsberg at the time of Euler (1736). On the right the resulting graph. Here the various parts of the city (**A, B, C, D**) are stylized as vertices. The various bridges are represented by the edges of the graph.

(Image G. Caldarelli)

Motivations

- Relevant approach for most scientific domains. Already used a lot in social sciences.
- **Complex network :** study of large graphs, to determine the presence of «patterns», and global properties induced by the connections.
- From the structure of a graph, one expects to learn or infer the properties of the system it describes.
- For instance : robustness to failure and attacks

Motivations

- One can consider a graph in a static way, for its connection topology (e.g. friendship network)
- We can also consider that quantities associated with the nodes of the graph evolve over time : we define that as **dynamic networks**
- For instance : propagation of epidemics, economical models.
- Generalization of a Cellular Automata to a non-regular lattice.
- Both the state of the nodes and the graph topology can evolve.

The Milgram's small-world experiment 1960

(See Newmann, paragraphe 3.6)

- What is the typical distance between the members of a social network ?
- This is the number of persons through which one need to go to reach anyone in the network.
- Milgram tried to estimate this distance in a real social network (before Internet)
- He sent 96 letters to persons randomly chosen in the phone book of the city of Omaha, Nebraska.
- Each letter was containing a «passport» with the logo of the University of Harvard.

The Milgram's small-world experiment

- Each recipient of the letter was asked to transfer this passport to one of Milgram's colleague, in Boston.
- The name and address of the colleague was indicated, but the goal was to send the passport through a chain of friends.
- At each step, the intermediate person was asked to add his/her name in the passport, before sending it to the next person, until it reaches Milgram's colleague, from acquaintance to acquaintance.

The six degrees of separation

- This process was a way to measure the length of the route from Omaha to Boston.
- 18 of the 96 passports that were sent arrived to destination. This was considered as a success.
- The average distance of these 18 routes was of **5.9** steps. (It varied between 2 and 10).
- This created the popular belief that there are **six degree of separation** between any individual on Earth.

Examples of graphs/networks

Food chain

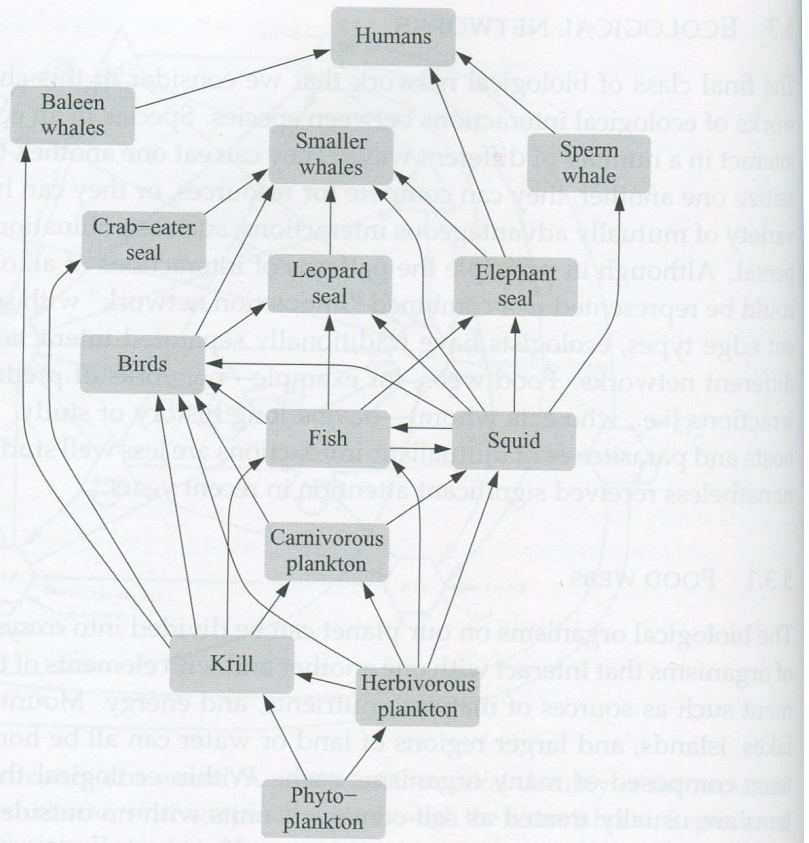


Figure 5.8: A food web of species in Antarctica. Vertices in a food web represent species or sometimes, as with some of the vertices in this diagram, groups of related species, such as fish or birds. Directed edges represent predator–prey interactions and run in the direction of energy flow, i.e., from prey to predator.

Food chain

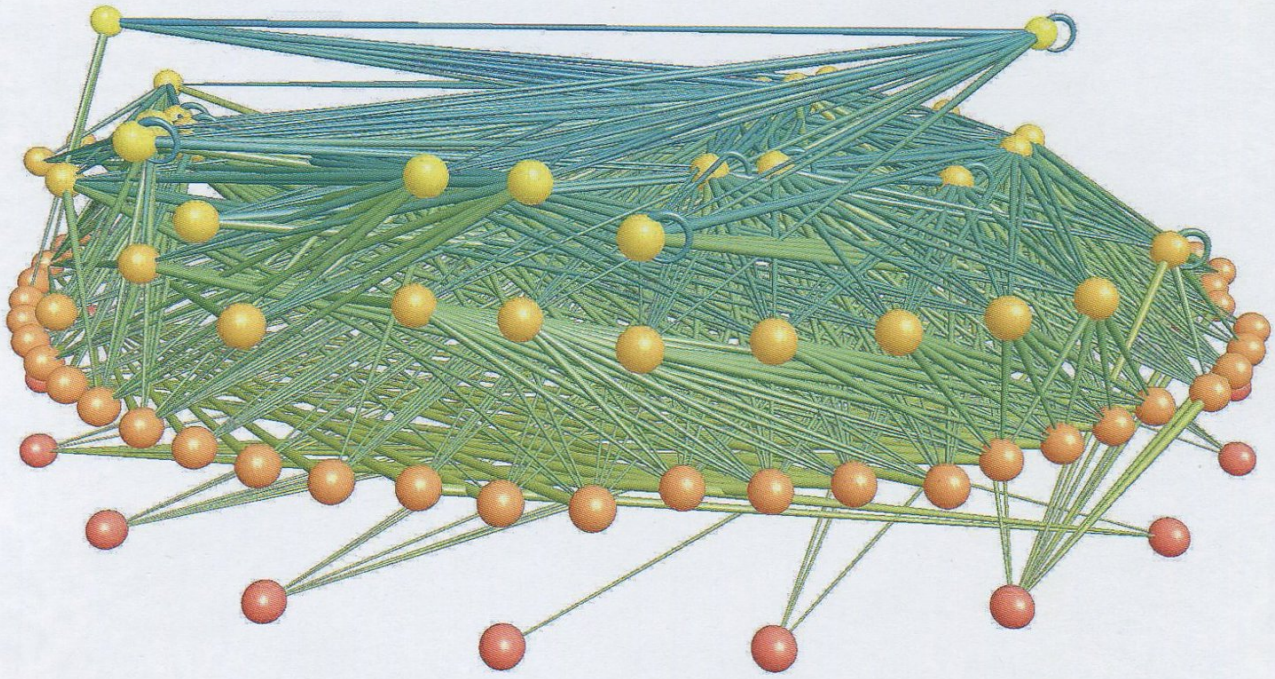


Plate II: The food web of Little Rock Lake, Wisconsin. This elegant picture summarizes the known predatory interactions between species in a freshwater lake in the northern United States. The vertices represent the species and the edges run between predator–prey species pairs. The vertical position of the vertices represents, roughly speaking, the trophic level of the corresponding species. The figure was created by Richard Williams and Neo Martinez [210].

Internet

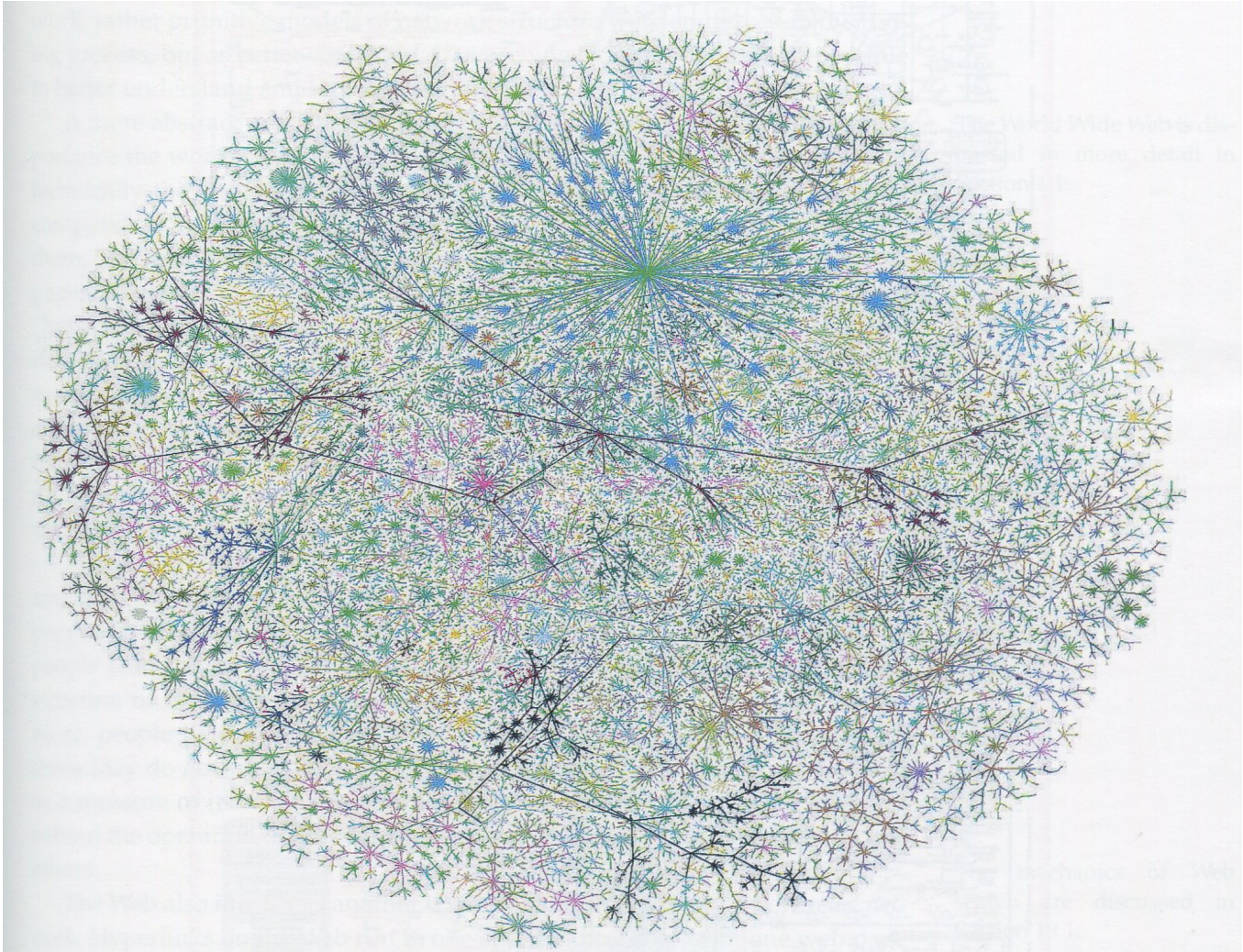


Plate III: The structure of the Internet at the level of autonomous systems. The vertices in this network representation of the Internet are autonomous systems and the edges show the routes taken by data traveling between them. This figure is different from Plate I, which shows the network at the level of class C subnets. The picture was created by Hal Burch and Bill Cheswick. Patent(s) pending and Copyright Lumeta Corporation 2009. Reproduced with permission.

Pipelines

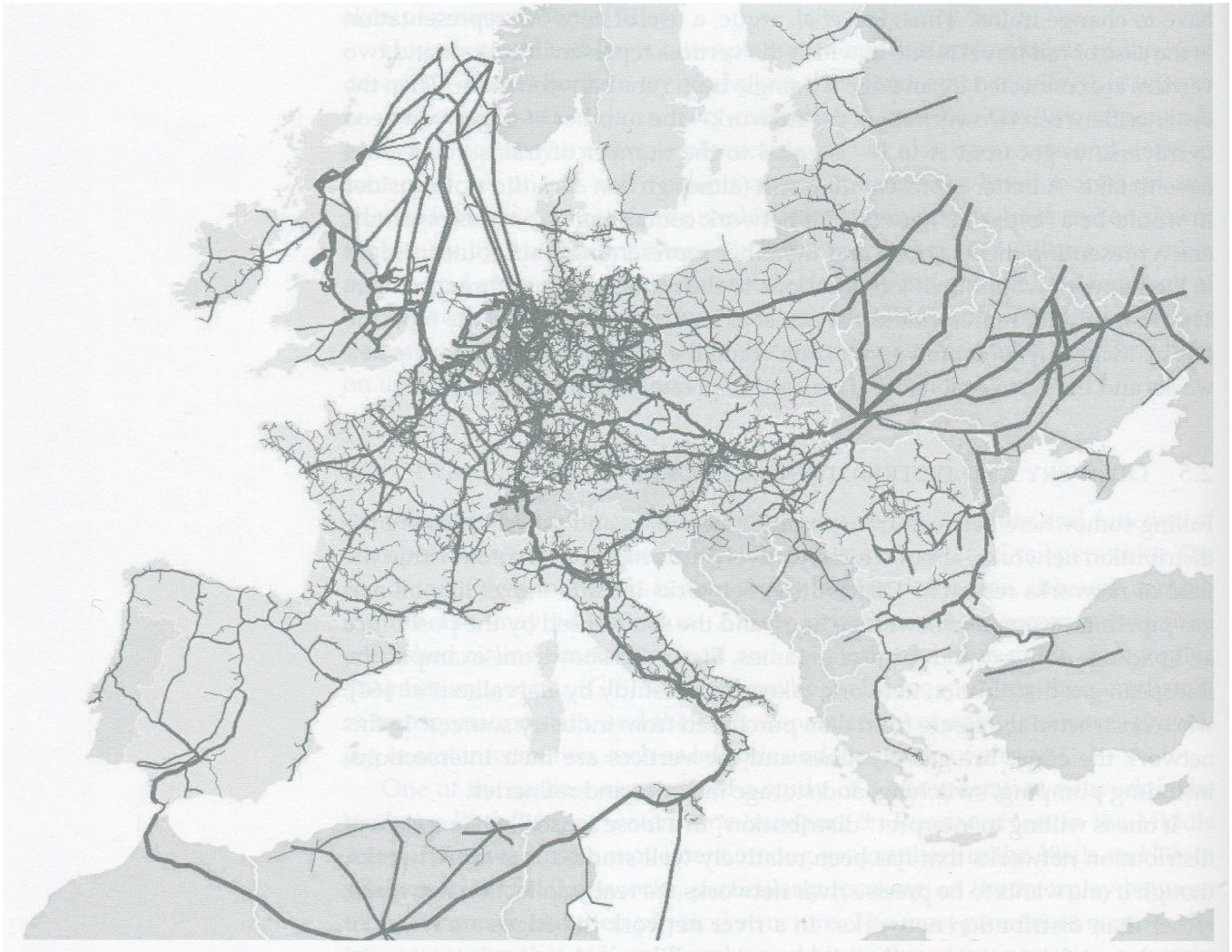


Figure 2.5: The network of natural gas pipelines in Europe. Thickness of lines indicates the sizes of the pipes. Figure created by R. Carvalho *et al.* [64]. Copyright 2009 American Physical Society. Reproduced with permission.

Molecular biology

Metabolic or regulation network,...

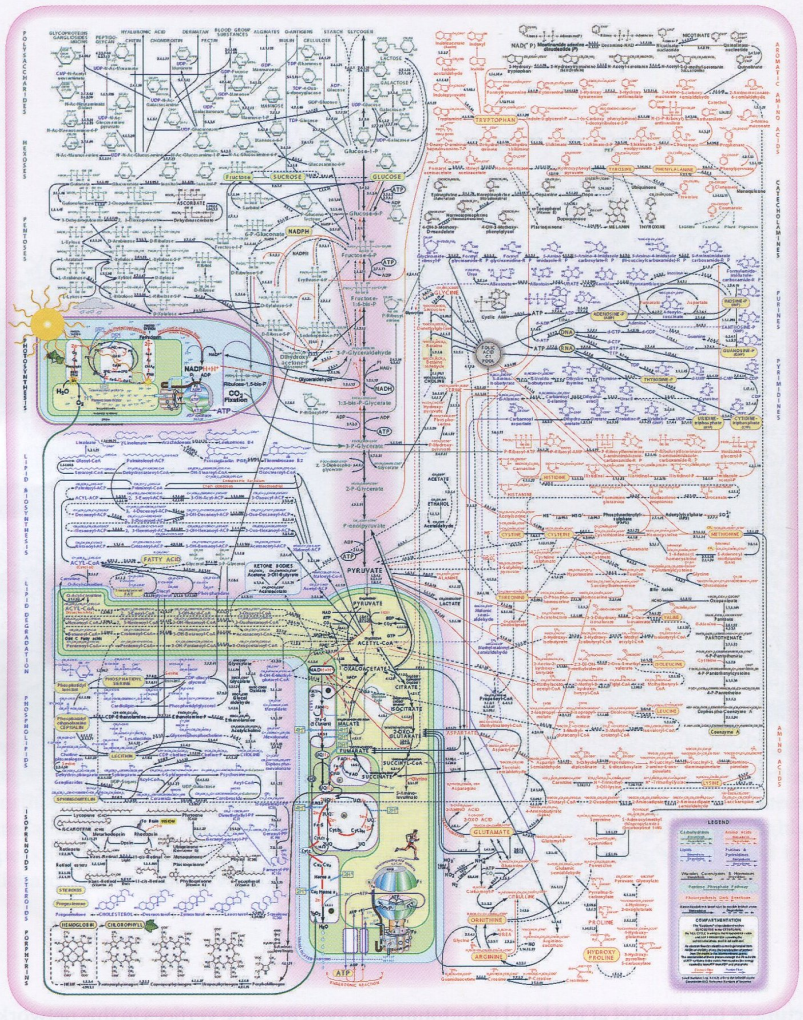


Plate IV: A metabolic network. A wallchart showing the network formed by the major metabolic pathways. Created by Donald Nicholson. Copyright of the International Union of Biochemistry and Molecular Biology. Reproduced with permission.

Social and friendship networks

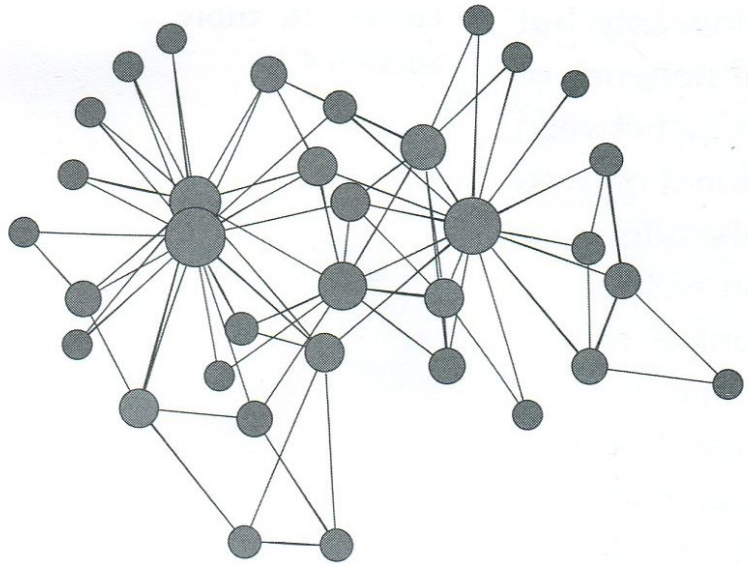
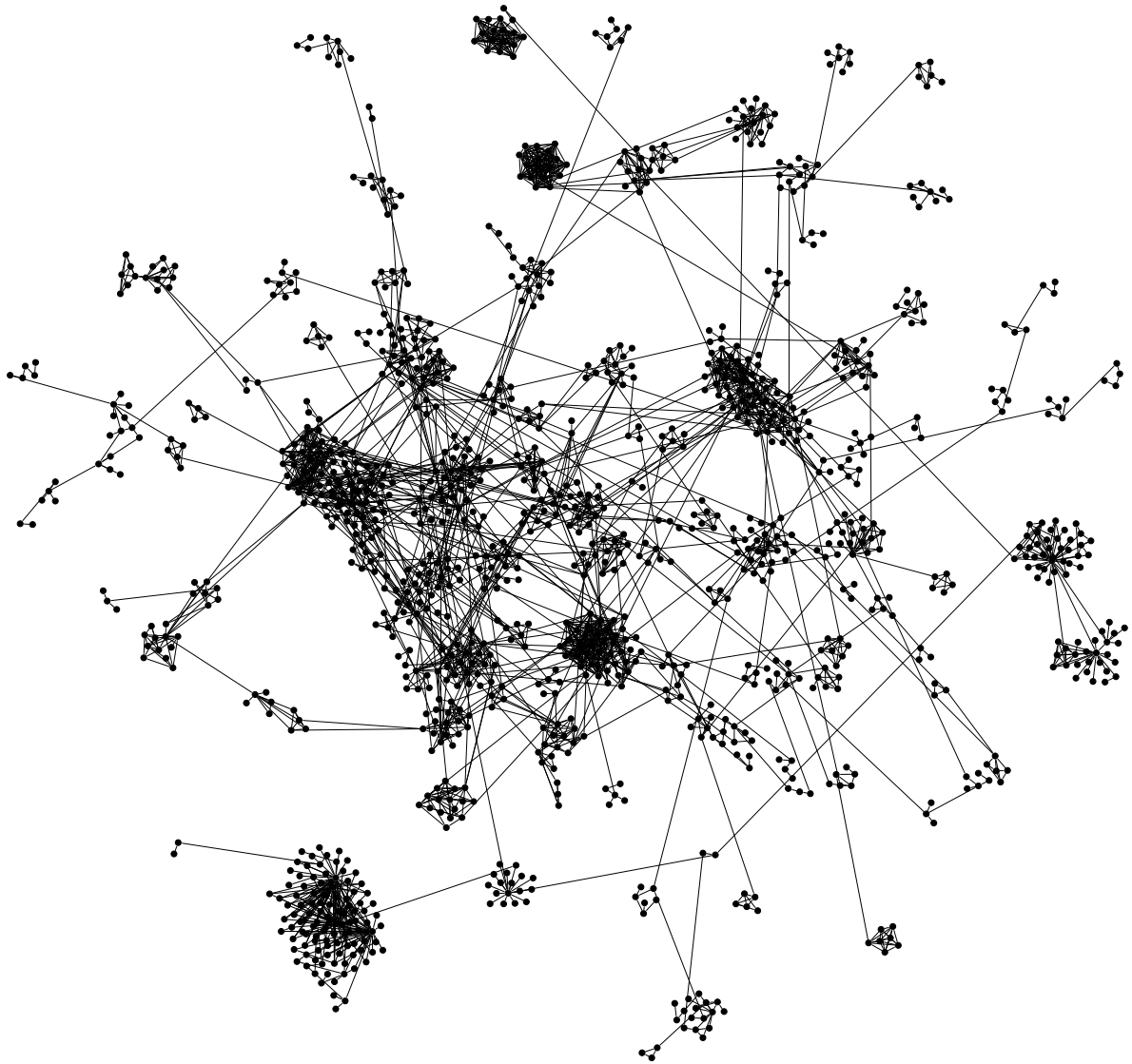


Figure 1.2: Friendship network between members of a club. This social network from a study conducted in the 1970s shows the pattern of friendships between the members of a karate club at an American university. The data were collected and published by Zachary [334].

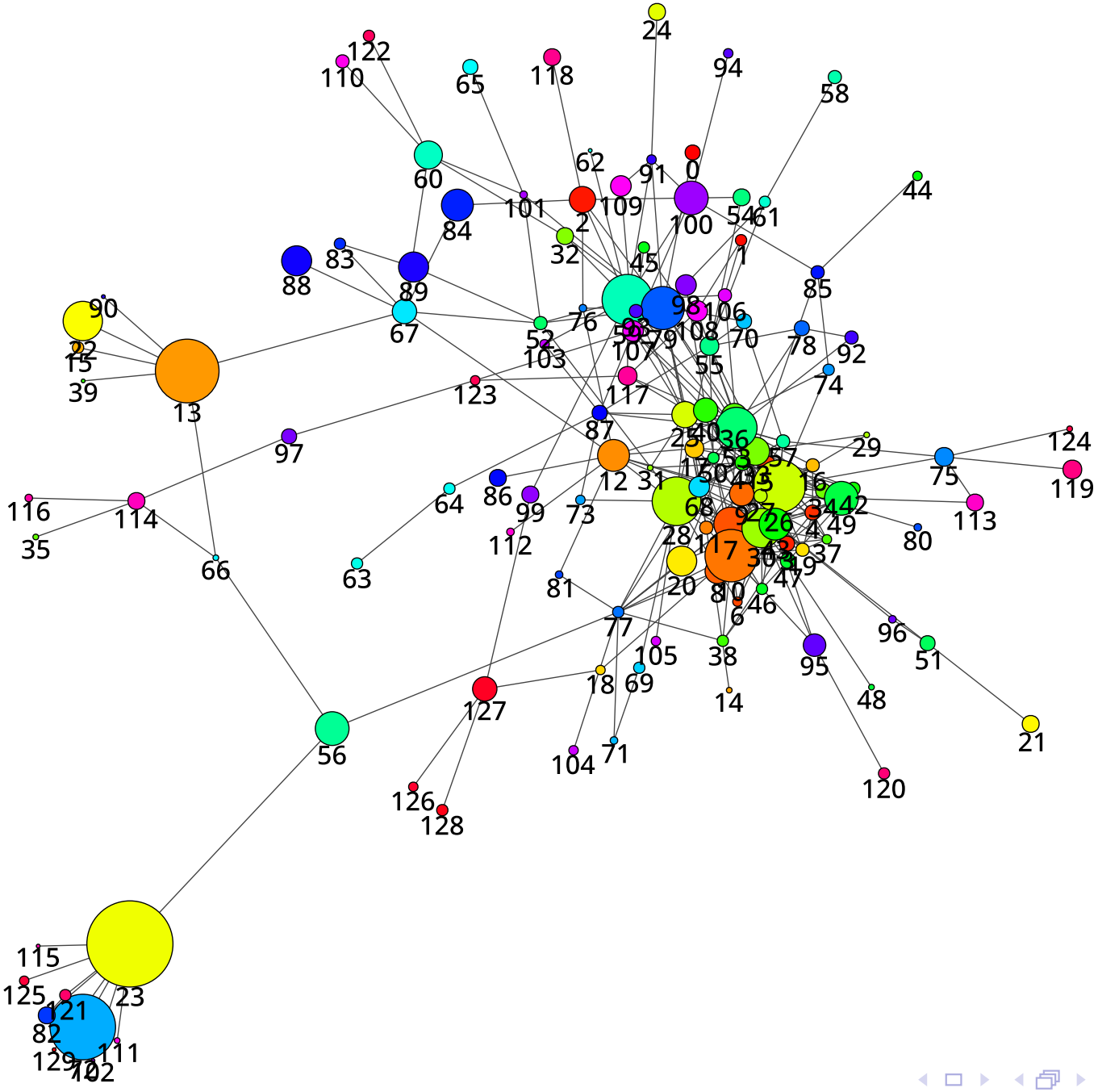
Network of actors. A link is created if two actors played in the same film.

WHO resolutions : network of citations (1948–2021)

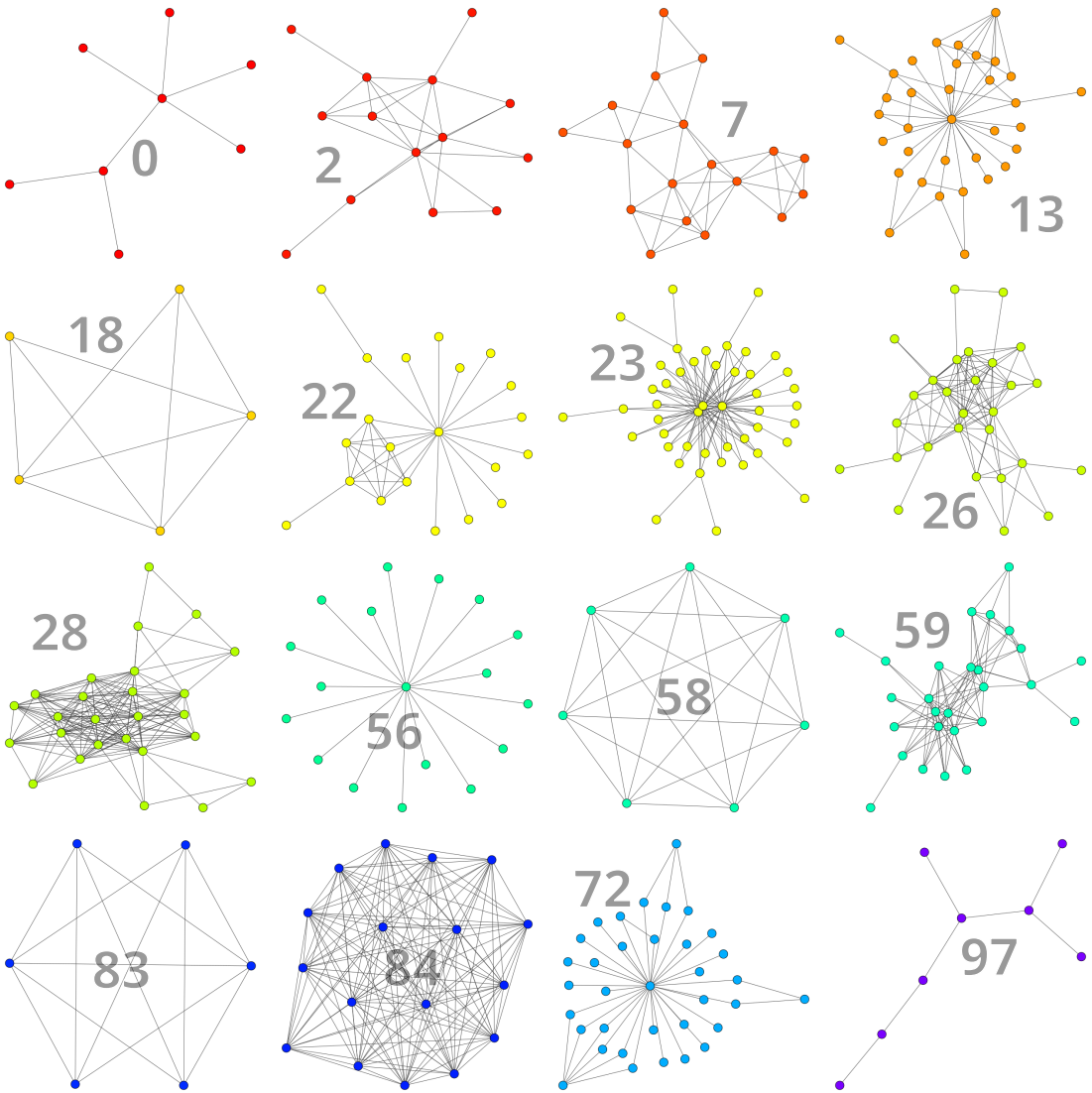


Credit : J.-L. Falcone, D. Wernli

WHO resolutions : network of citations



WHO resolutions : network of citations



- 0 : governing
- 2 : peace, disarmament
- 7 : blindness, rehabilitation
- 13 : invoking article
- 22 : contributions,
- 23 : assessment, members
- 26 : drugs, antimicrobial resistance
- 28 : infant
- 56 : assessed contributions
- 58 : international decade world, world indigenous people
- 59 : developing countries
- 72 : assessment future years
- 83 : lesotho swaziland, struggle southern
- 84 : displaced persons cyprus
- 97 : coordination united nations

Most relevant words in resolution title

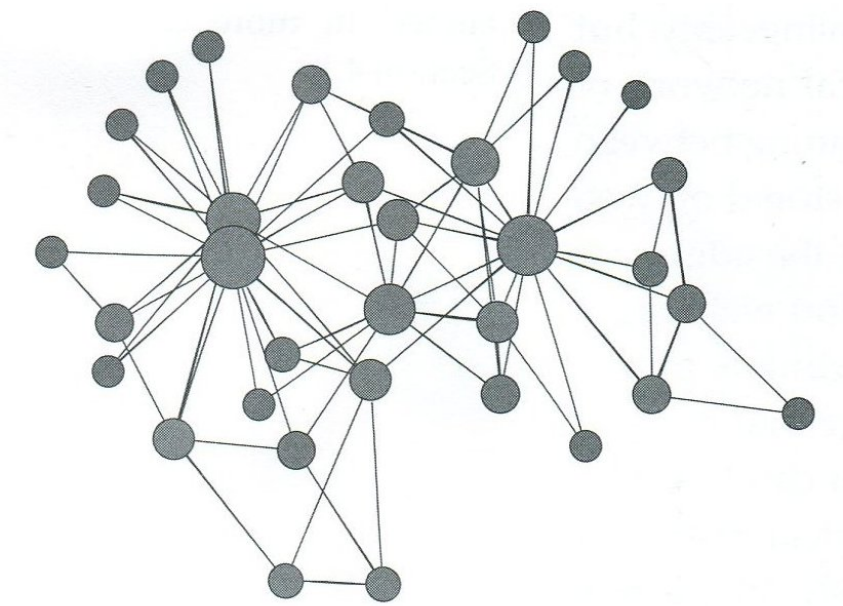
Nature possible des noeuds et des arcs

Network	Vertex	Edge
Internet	Computer or router	Cable or wireless data connection
World Wide Web	Web page	Hyperlink
Citation network	Article, patent, or legal case	Citation
Power grid	Generating station or substation	Transmission line
Friendship network	Person	Friendship
Metabolic network	Metabolite	Metabolic reaction
Neural network	Neuron	Synapse
Food web	Species	Predation

Table 6.1: Vertices and edges in networks. Some examples of vertices and edges in particular networks.

5.2 Definitions and basic concepts

- Vertices or nodes (sommets, noeuds) and edges or links (arcs, liens)
- Order n (number of nodes) and size m (number of edges) of the graph
- Undirected, directed and weighted graphs
- Sub-graphs, cliques,...



Definitions and basic concepts

- Connected nodes and graph components

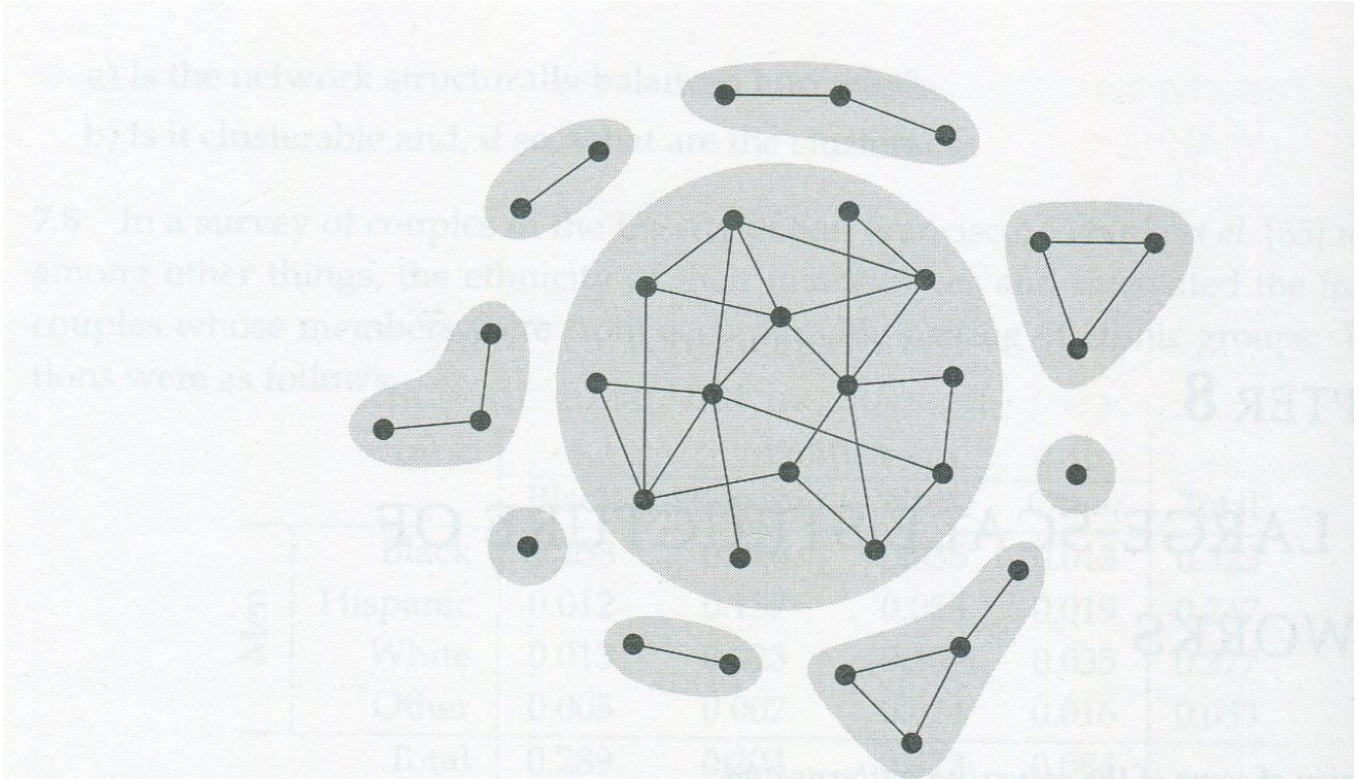


Figure 8.1: Components in an undirected network. In most undirected networ
is a single large component occupying a majority, or at least a significant fra
the network, along with a number of small components, typically consisting o
handful of vertices each.

Adjacency matrix and node degree

- Adjacency matrix A_{ij} : $A_{ij} = 1$ if there is a link from j to i
- Degree k_i, k_i^{in}, k_i^{out}
- $k_i^{in} = \sum_j A_{ij}$, $k_i^{out} = \sum_j A_{ji}$

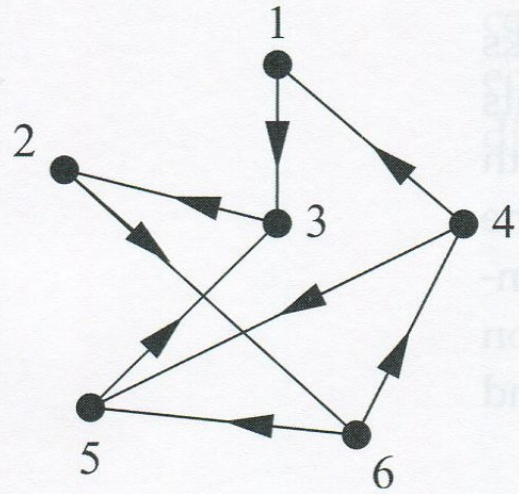
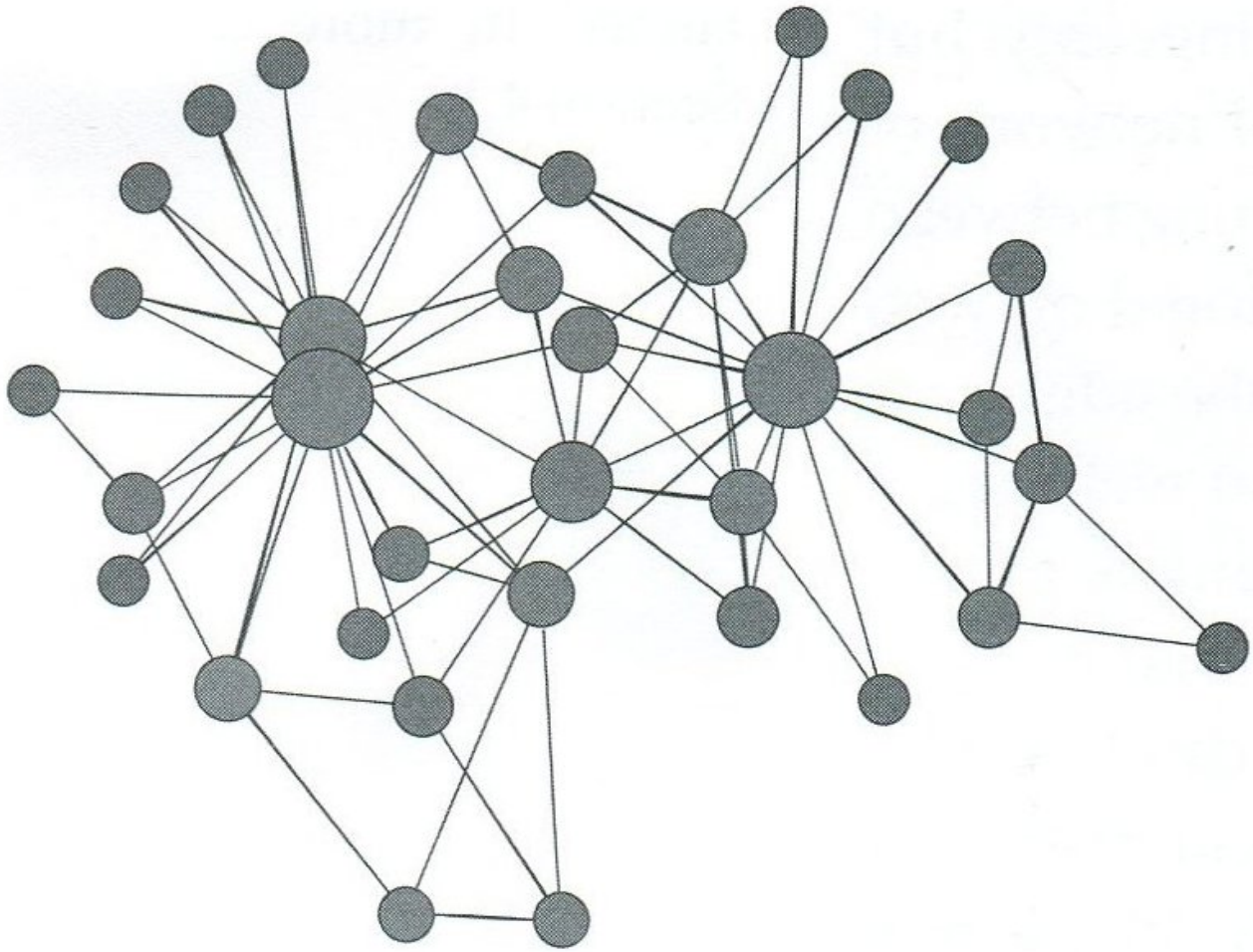


Figure 6.2: A directed network. A small directed network with arrows indicating the directions of the edges.

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

Definitions and basic concepts

- distance d_{ij}, diameter D



d_{ij} is the length of the shortest path between nodes i et j. D is the maximum of these distances over all pairs of nodes.

5.3 Dynamical systems on a graph

- Nodes can have a value that depends on time
- They change their value according to the value of their neighbors

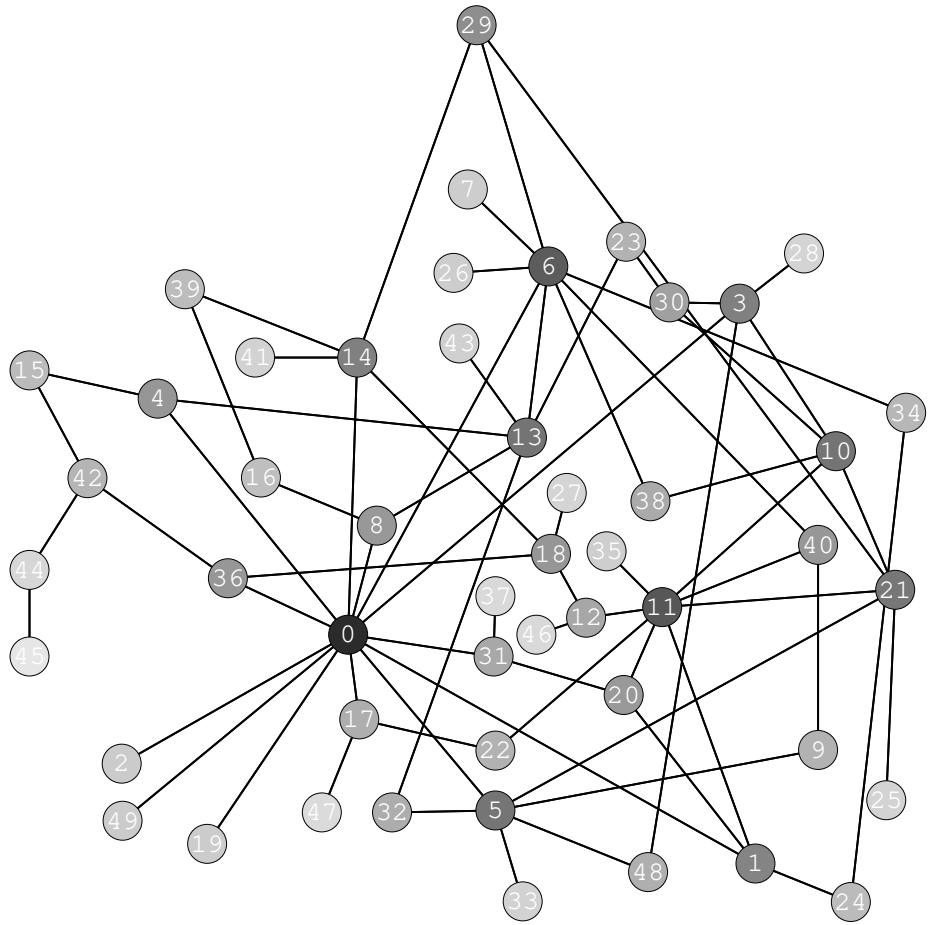
$$s_i(t+1) = F(\{s_j(t)\}$$

where $\{s_j(t)\}$ is the values of nodes j connected to i

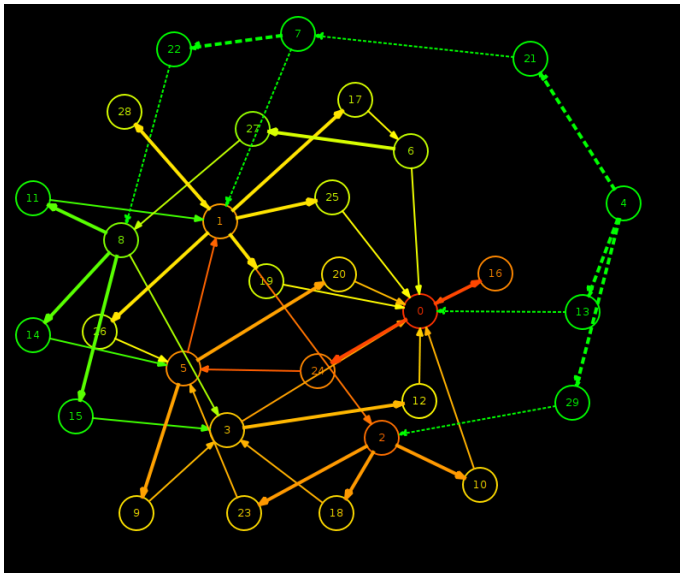
Example :A model of opinion propagation in a social network

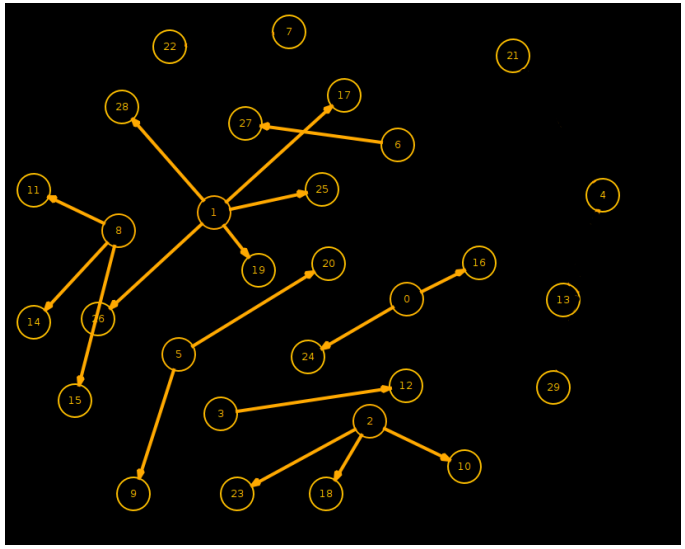
$$s_i(t+1) = \begin{cases} 1 & \text{with probability } p_i(\{s_j(t)\}) \\ 0 & \text{with probability } 1 - p_i(\{s_j(t)\}) \end{cases}$$

How each agent influences the group :



Example :Evolution of relations in a market where goods are exchanged against money





Example : diffusion process on a graph

A diffusion process of a quantity Ψ on a graph means that, at each iteration, each node i sends an amount $D\Psi_i$ to its k_i neighbors and received an amount $D\Psi_j$ from its neighbors j, where D is a given coefficient.

This can be expressed be expressed as

$$\Psi_i(t+1) = \Psi_i(t) + D\left(\sum_{j \in \mathrm{Neighbor}(i)} \Psi_j(t) - k_i\Psi_i(t)\right)$$

Example : diffusion process on a graph

Using the adjacency matrix, this becomes

$$\Psi_i(t+1) = \Psi_i(t) + D \sum_j (A_{ij} - k_i \delta_{ij}) \Psi_j(t)$$

where the Kronecker symbol is

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

Laplacian matrix

we define

$$L = A - d$$

as the Laplacian matrix. Here d is the diagonal matrix such that $d_{ii} = k_i$, which contains the degree of each node.
Thus, the diffusion process can be expressed as

$$\Psi_i(t+1) = \Psi_(t) + D \sum_j L_{ij} \Psi_j(t)$$

or, in a full matrix form

$$\Psi(t+1) - \Psi(t) = L\Psi(t)$$

Laplacian Matrix

For a graph which is a 1D Cartesian grid, we obtain

$$\frac{\partial \Psi_i}{\partial t} = \frac{D}{\Delta x^2} \left(\Psi_{i-1} + \Psi_{i+1,} - 2\Psi_i \right)$$

which is the Taylor expansion in space of

$$\frac{\partial \Psi}{\partial t} = D \nabla^2 \Psi = D \frac{\partial^2}{\partial x^2} \Psi$$

5.4 Properties of complex networks

- Real graphs are not random
- Their topology reflect the properties of the system they describe.
- Analyzing the structure of a graph reveals some of these properties
- For instance, pattern of connection may be more frequent than others
- Or, tightly connected nodes indicate a probable common role in the system.

Examples of properties of real graphs

M. Newman, p 237

	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	r	Ref(s).
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.208	16,323
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.276	88,253
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.120	89,146
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.363	234,236
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.127	234,236
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1				9,10
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16		103
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.092	248
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.029	34
	Sexual contacts	Undirected	2 810					3.2				197,198
Information	WWW `nd.edu`	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.067	13,28
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7				56
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–				280
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.157	184
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44		97,116
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.189	66,111
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	−0.003	323
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.033	294
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.016	239
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.119	315
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.154	115
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.366	6,282
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.240	166
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.156	164
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.263	160
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.326	209
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.226	323,328

Table 8.1: Basic statistics for a number of networks. The properties measured are: type of network, directed or undirected; total number of vertices n; total number of edges m; mean degree c; fraction of vertices in the largest component S (or the largest weakly connected component in the case of a directed network); mean geodesic distance between connected vertex pairs ℓ; exponent α of the degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient C from Eq. (7.41); clustering coefficient C_{WS} from the alternative definition of Eq. (7.44); and the degree correlation coefficient r from Eq. (7.82). The last column gives the citation(s) for each network in the bibliography. Blank entries indicate unavailable data.

Also, see Caldarelli, table 1.2

Small world

- Often, in real complex networks, one observes that the average distance between nodes, as well as the diameter, are small with respect to n, the number of nodes.
- One is close to everyone (see Milgram's experiment)

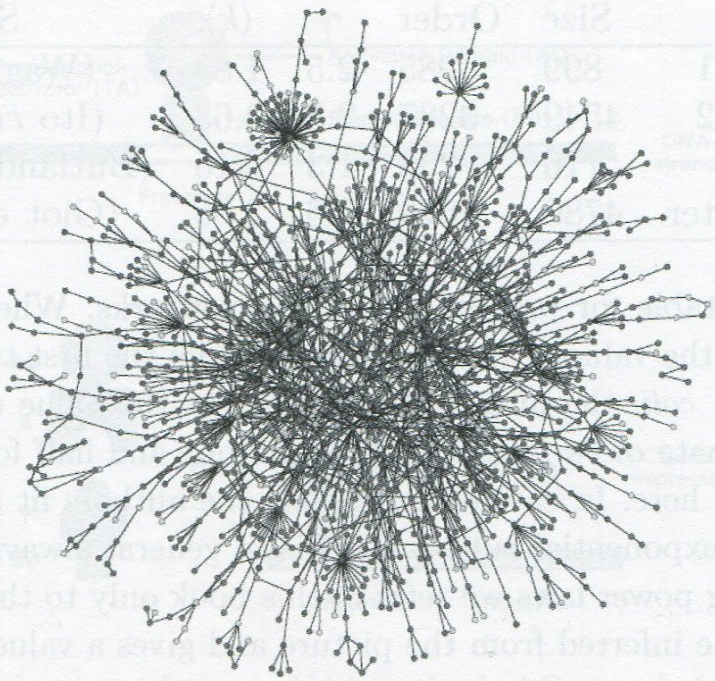


Fig. 6.6 The graph of the protein interaction network of *Saccharomyces cerevisiae*. Courtesy of H. Jeong (Jeong *et al.*, 2001).

Small world

- Small typically means that $D = \mathcal{O}(\log n)$
- Classical networks, such as rings of grids do not have such a property.

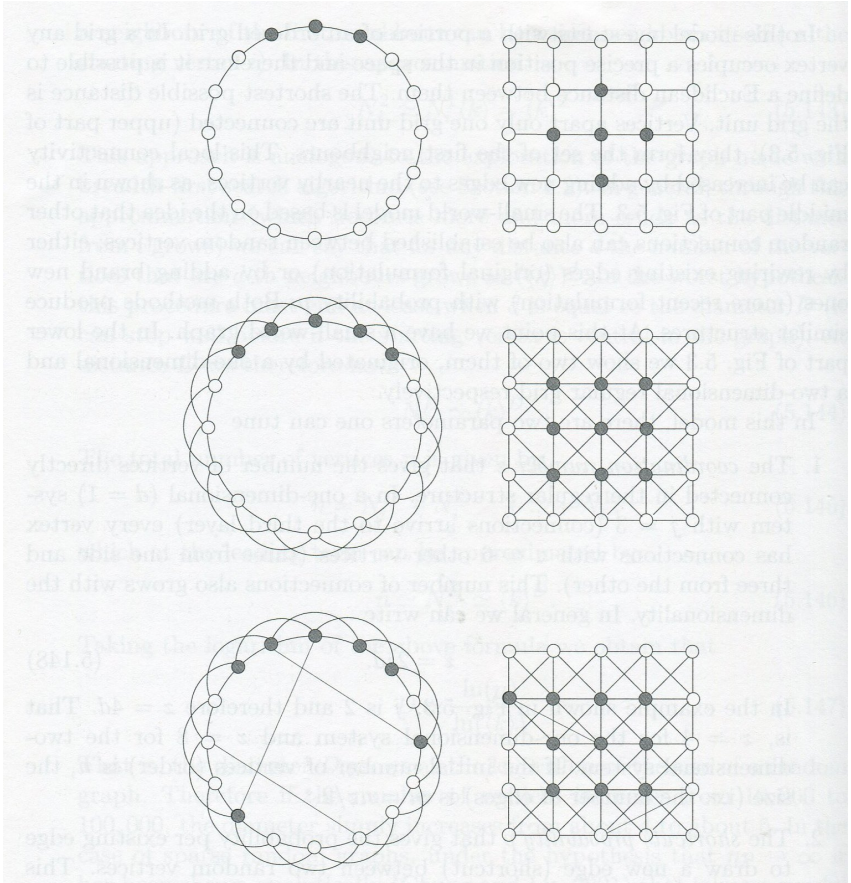


Fig. 5.3 On the first row a regular one-dimensional lattice on the left and a regular two-dimensional lattice on the right. In the second row the same lattices with extra edges increasing the local connectivity. On the last line we have the small-world lattices with shortcuts.

Degree distribution

- In a graph, each node is characterized by its degree k
- An important property of graphs is the degree distribution
- It means that for each possible value of $k \in \{0, 1, 2, 3, \ldots\}$ one gives the number $n(k)$ of nodes having this degree
- The fraction of nodes with degree k

$$p_k = \frac{n(k)}{n}$$

gives the probability that a node chosen at random has degree k.

Degree distribution

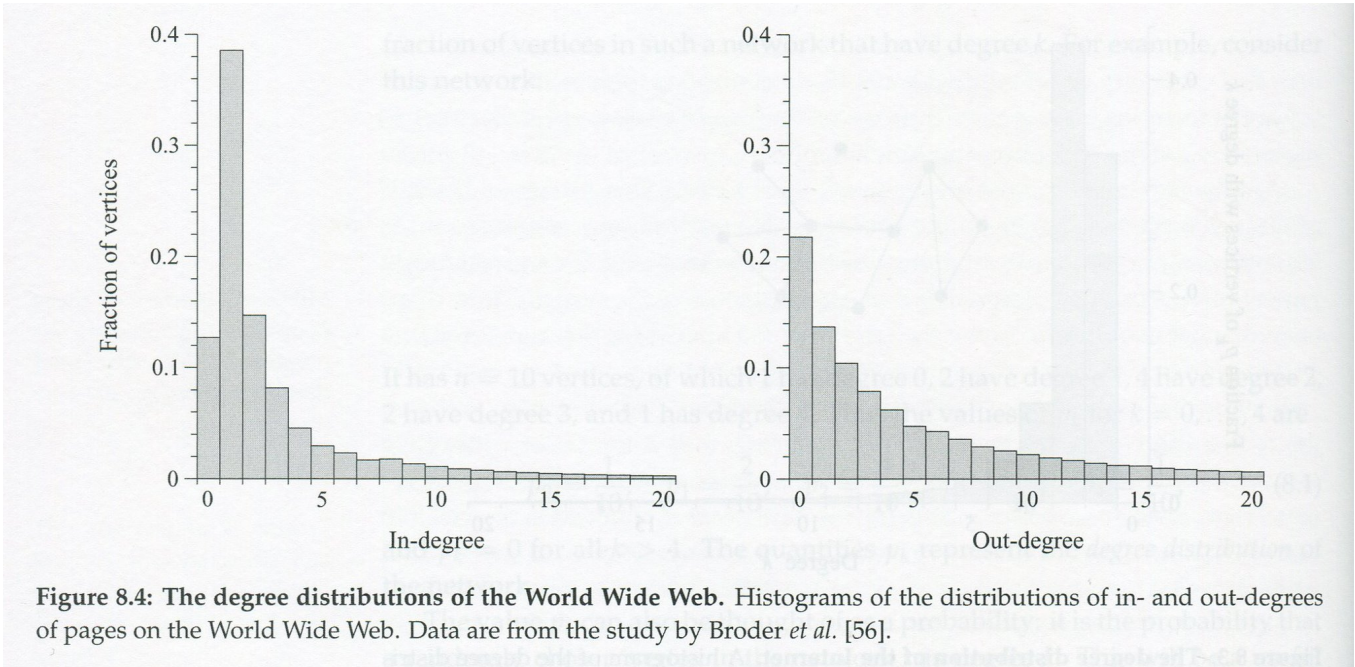


Figure 8.4: The degree distributions of the World Wide Web. Histograms of the distributions of in- and out-degrees of pages on the World Wide Web. Data are from the study by Broder *et al.* [56].

Image Newman p. 246

«Scale-free» Networks

- For many real network, one observes that the degree distribution obeys a **power law**.

$$p_k = Ck^{-\alpha}$$

  at least for an interval of values of k
- α is the **exponent** of the distribution law. Often, one observes that $2 \leq \alpha \leq 3$
- Such networks are called **scale-free** because no characteristic scales exist for k : its average and variance are infinite
- Such distributions are often called **fat-tailed** distributions

«Scale-free» Networks

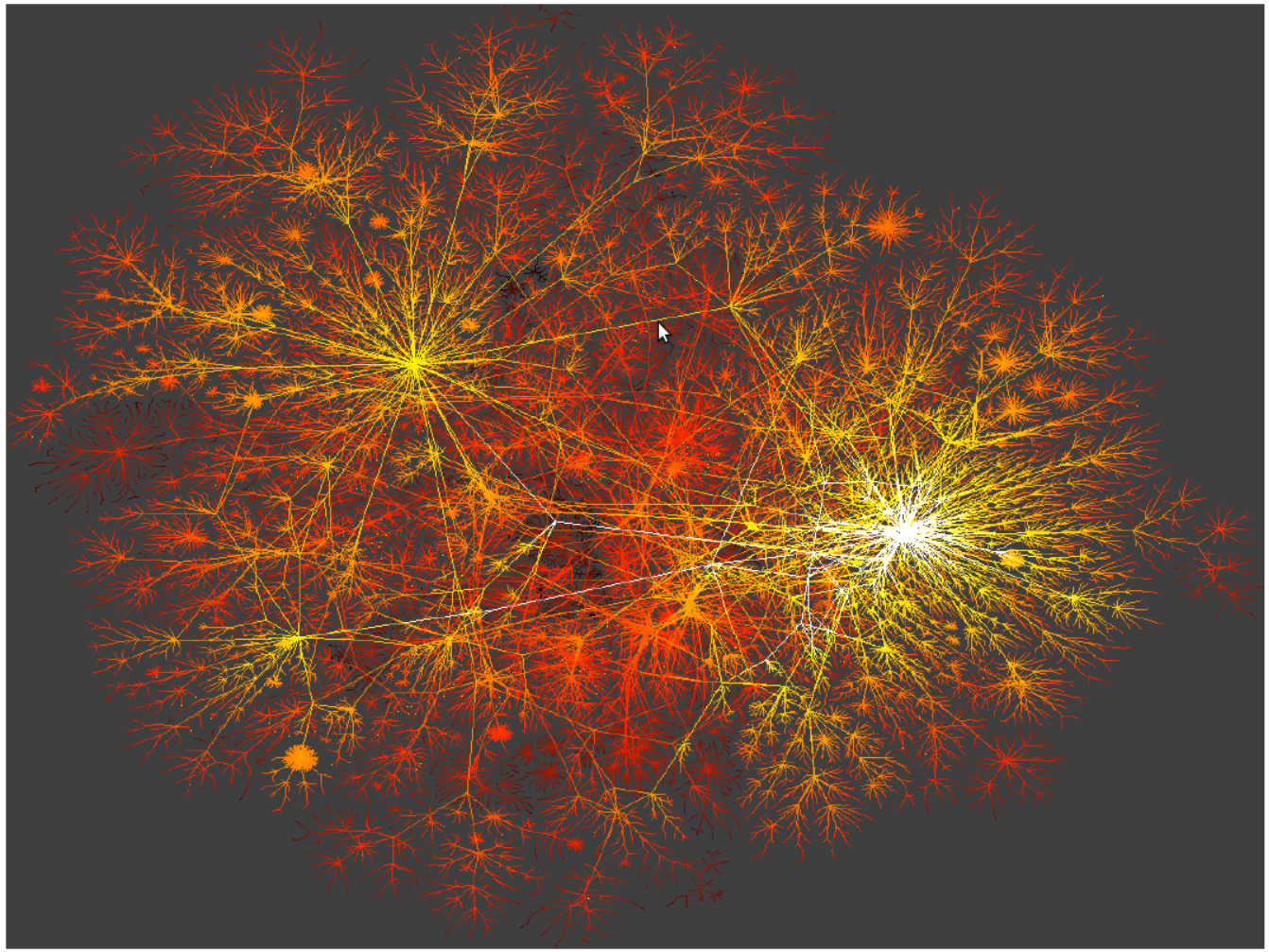


Image G.Caldarelli

Clustering Coefficient

It measures the number of neighbors that are themselves connected
The clustering coefficient C_i for node i is defined as

$$C_i = \frac{\text{number of pairs of neighbors of } i \text{ that are connected}}{\text{number of pairs of neighbors}}$$

In terms of graph quantities

$$C_i = \frac{1}{k_i(k_i-1)/2} \sum_{j,k} A_{ij} A_{ik} A_{jk}$$

This actually measures the presence of *triangles* in the graph, that is the cases where i is connected to j and to k, and j and k are also connected.
So, there is a **clustering** of 3 individuals.

Coefficient de clustering

For a random graph [1], the average clustering coefficient is

$$C_{\mathrm{random-graph}} = C_{rg} = \frac{1}{n} \frac{\left(< k^2 > - < k > \right)^2}{< k >^3}$$

- In real graphs, C usually differs from C_{rg}.
- For instance, in a collaboration network of physicists, one observes

$$C = 0.45 \qquad \text{instead of} \qquad C_{rg} = 0.023$$

  for comparable average degrees.
- This shows the social trend to create new collaborations among two collaborators of a same scientists.

1. See section 5.8

5.5 Centrality Measure

- **Centrality** is a measure of the importance of a node with respect to the others in the graph.
- This suggests that high centrality nodes play an important role in the system described by the graph.
- There are several criteria to define such an importance.
- The **degree centrality** is simply the degree of the node.
- In a social network, this reflects the fact that highly connected persons are probably important.
- But there are several other centrality metrics : **eigenvector centrality**, **page rank**, **closeness** and **betweenness** centrality.

Eigenvector centrality

- The idea of this centrality is that the importance (or score) x_i of a node i is determined by the importance of the nodes pointing to it.
- The score of a node is proportional to the **sum** of the scores of its incoming neighbors.

Eigenvector centrality

- Let $x = (x_1, x_2, ..., x_n)$ be such that x_i is the eigenvector centrality of node i
- Mathematically, the above definition can be expressed as

$$x_i = \mu \sum_j A_{ij} x_j \qquad \text{or, else} \qquad x = \mu A x$$

- with A the adjacency matrix and μ a yet unknown coefficient of proportionality.

Eigenvector centrality

- A solution to this equation is to take x as and eigenvector (vecteur propre) of A

$$Ax = \lambda x$$

with λ, the associated eigenvalue (valeur propre), and

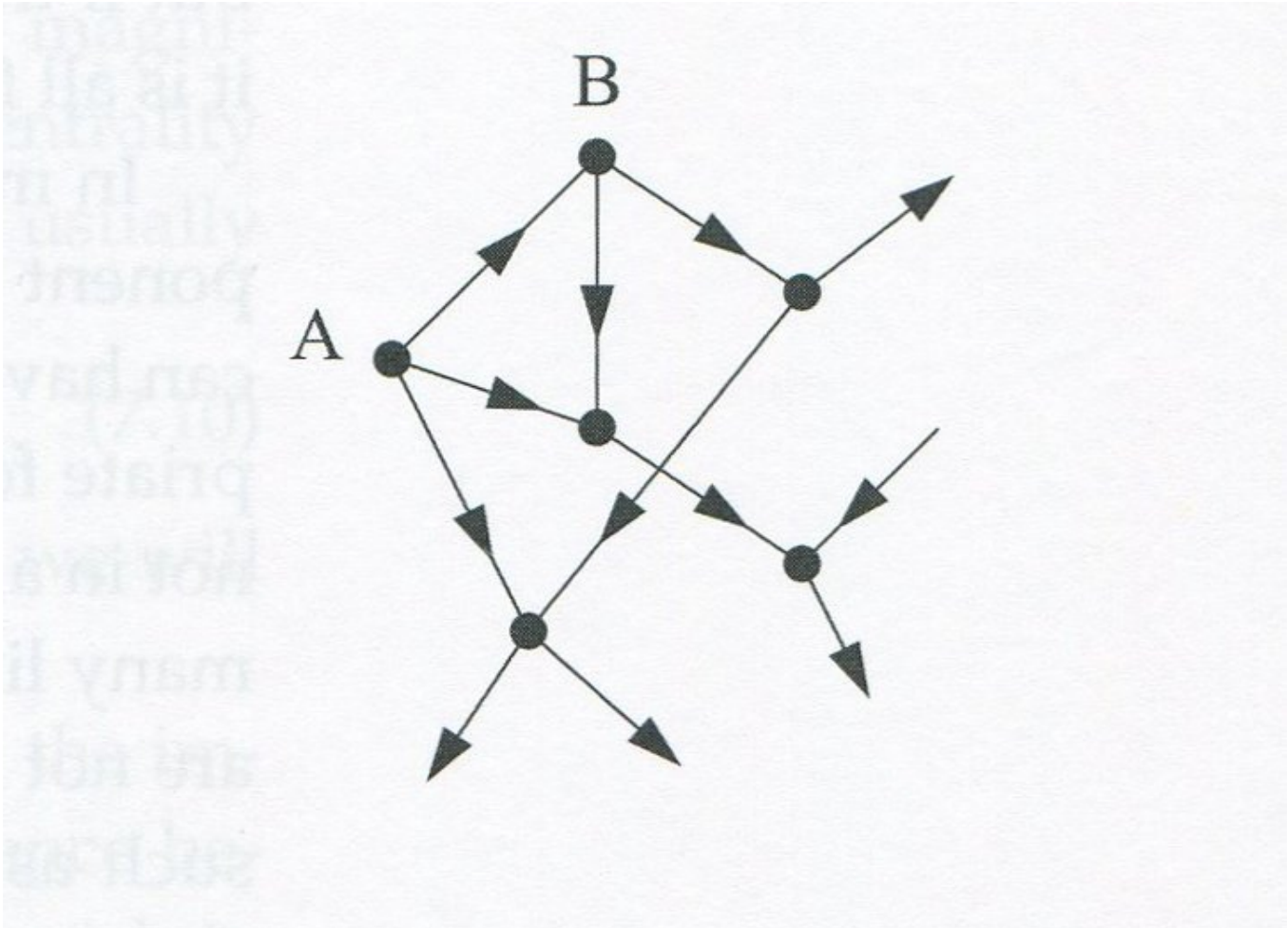
$$\mu = \lambda^{-1}$$

.

- Among the possible eigenvector x, one chooses the one with the largest eignevalue λ.

Eigenvector centrality

- This definition is however not fully satisfactory, in particular for directed graphs.
- The importance of a node depends on the importance of the nodes pointing to it, and not on the importance of the node it points to (otherwise, it would be easy to be important)



Eigenvector centrality

- Thus, a node A having only outgoing links will have centrality 0.
- This could be OK, except if it points to a node B which receives only links from such nodes of null centrality.
- B would also get centrality 0, which is not fair as it is referred to by other nodes.
- Only the nodes of a strongly connected component acquire a non-zero eigenvector centrality.

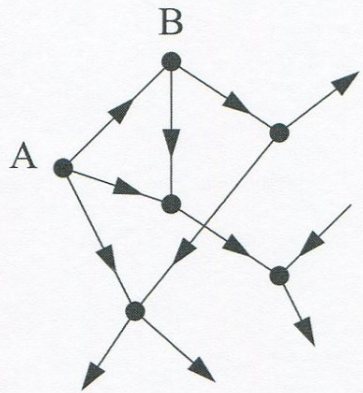


Figure 7.1: A portion of a directed network. Vertex A in this network has only outgoing edges and hence will have eigenvector centrality zero. Vertex B has outgoing edges and one in-

Katz Centrality

A solution is to give a baseline centrality β to all the nodes
Katz centrality x is then defined as

$$x = \alpha A x + \beta$$

or

$$x = (\mathrm{I} - \alpha A)^{-1}\beta$$

where α is a free parameter.
If $\alpha \to 0$, the centrality becomes $x = \beta$.
The parameter α is taken small enough, otherwise, this equation won't have a solution. This happens when

$$\det(A - \alpha^{-1}\mathrm{I}) = 0$$

and, in particular, the first time for $\alpha^{-1} = \lambda$, the largest eigenvalue of A.

Page rank

- Katz centrality has also a weakness. A node of large importance pointing to many other nodes will transfer its importance to all of them.
- That would be the case of a webpage holding a large list of information. As such, it is an important page, However the pointed paged won't have the same importance.
- It is therefore important to normalized the transmitted centrality by the number of outgoing links.
- This is the **page rank** centrality proposed by Google

$$x = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

Closeness centrality (proximité)

A completely different definition of centrality is the **closeness centrality** (centralité de proximité) C_i of node i.
It is defined as the closeness of node i with respect to all the others

$$C_i = \frac{1}{\frac{1}{n}\sum_j d_{ij}} = \frac{n}{\sum_j d_{ij}}$$

where d_{ij} is the distance (shortest path) between node i and j.
So C_i is the inverse of the average distance from i to all the other nodes.

Closeness centrality

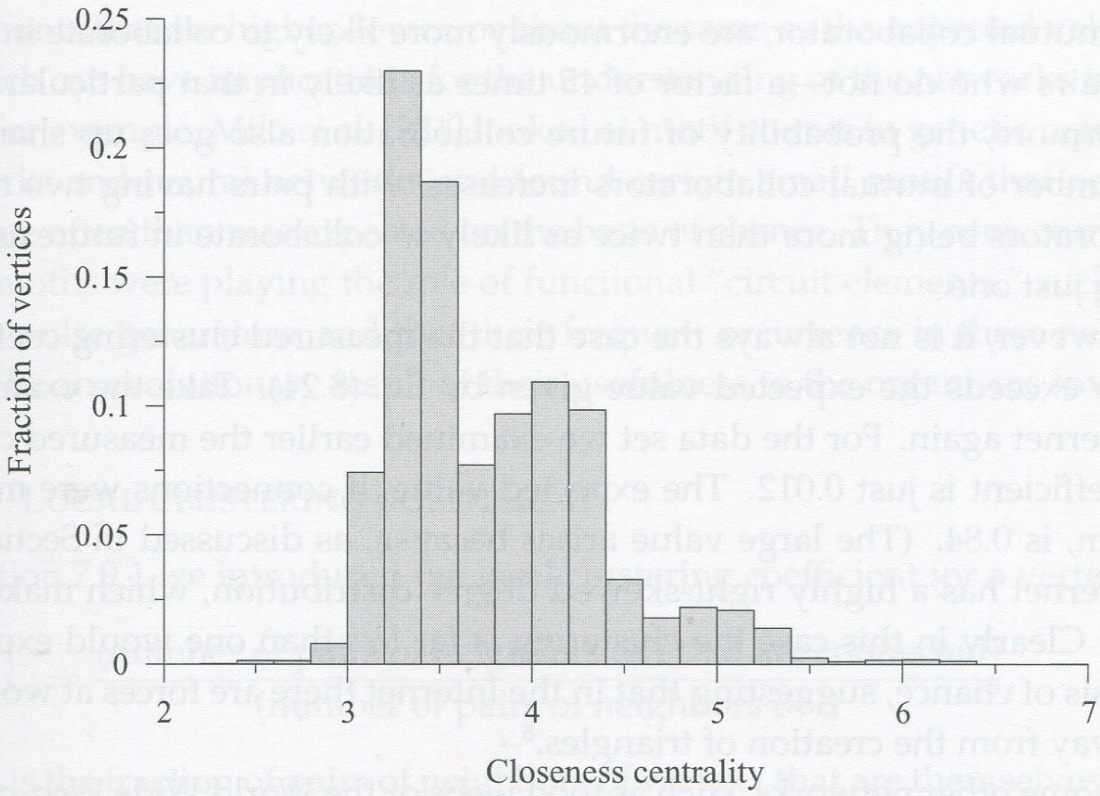


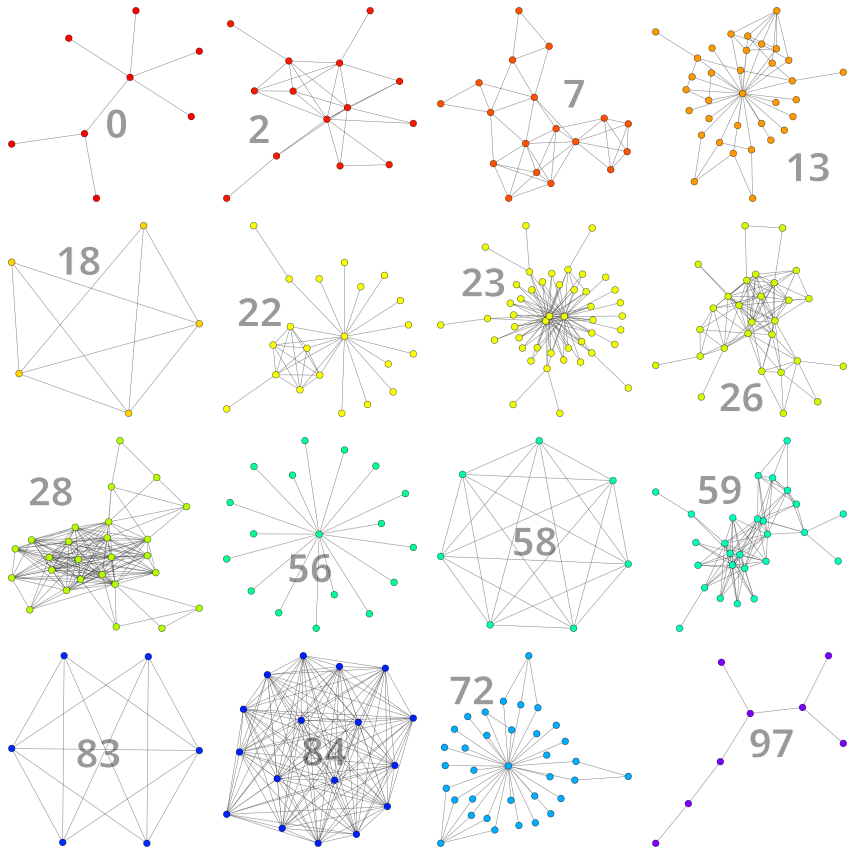
Figure 8.11: Histogram of closeness centralities of vertices on the Internet. Unlike Fig. 8.10 this is a normal non-cumulative histogram showing the actual distribution of closeness centralities. This distribution does not follow a power law.

Betweenness centrality : importance of edges

- The **betweenness** centrality measures the importance of the **edges** of the graphs.
- Let ℓ be an edge and let n_{ij}^{ℓ} the number of shortest paths between nodes i et j which use edge ℓ.
- Let g_{ij} be the number of shortest paths between i and j.
- The betweenness centrality C_{ℓ} of edge ℓ is then defined as

$$C_{\ell} = \sum_{ij} \frac{n_{ij}^{\ell}}{g_{ij}}$$

5.6 Similarity between nodes and assortativity



(Examples : see also Caldarelli, p. 28)

- Similarity in terms of a common property :
- Same number of neighbors, or same neighbors ;
- Same structure of connections
- In a company, people having identical functions will have similar relation with their surrounding

Assortativité (homophilie)

- The concept of **assortative mixing** describes the tendency of a node to link with the nodes that share the same property.
- In a social network, links preferably connect similar persons.
- More rarely, one observes a *diassortative mixing*, the association of persons with difference : for instance, a marriage is a link between persons of different gender.
- A network is said to be **assortative** if a significant fraction of its edges connects nodes sharing a common feature.

assortativité (homophily)

Friendship network between students in a school in USA

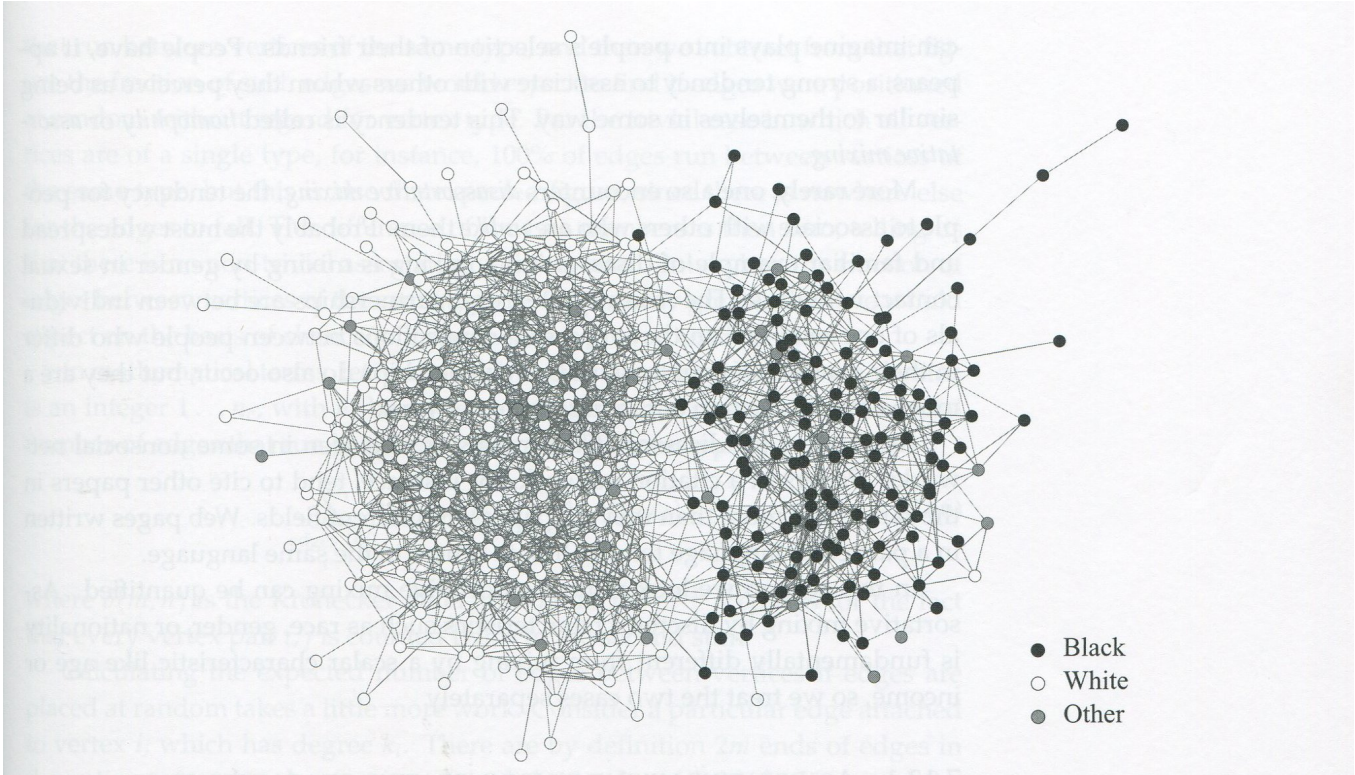


Figure 7.10: Friendship network at a US high school. The vertices in this network represent 470 students at a US high school (ages 14 to 18 years). The vertices are color coded by race as indicated in the key. Data from the National Longitudinal Study of Adolescent Health [34,314].

One observes a separation of the groups according to the ethnic origin of the students. Student with same ethnic origin are more likely to be friends.

Degree Assortativity / desassortativity

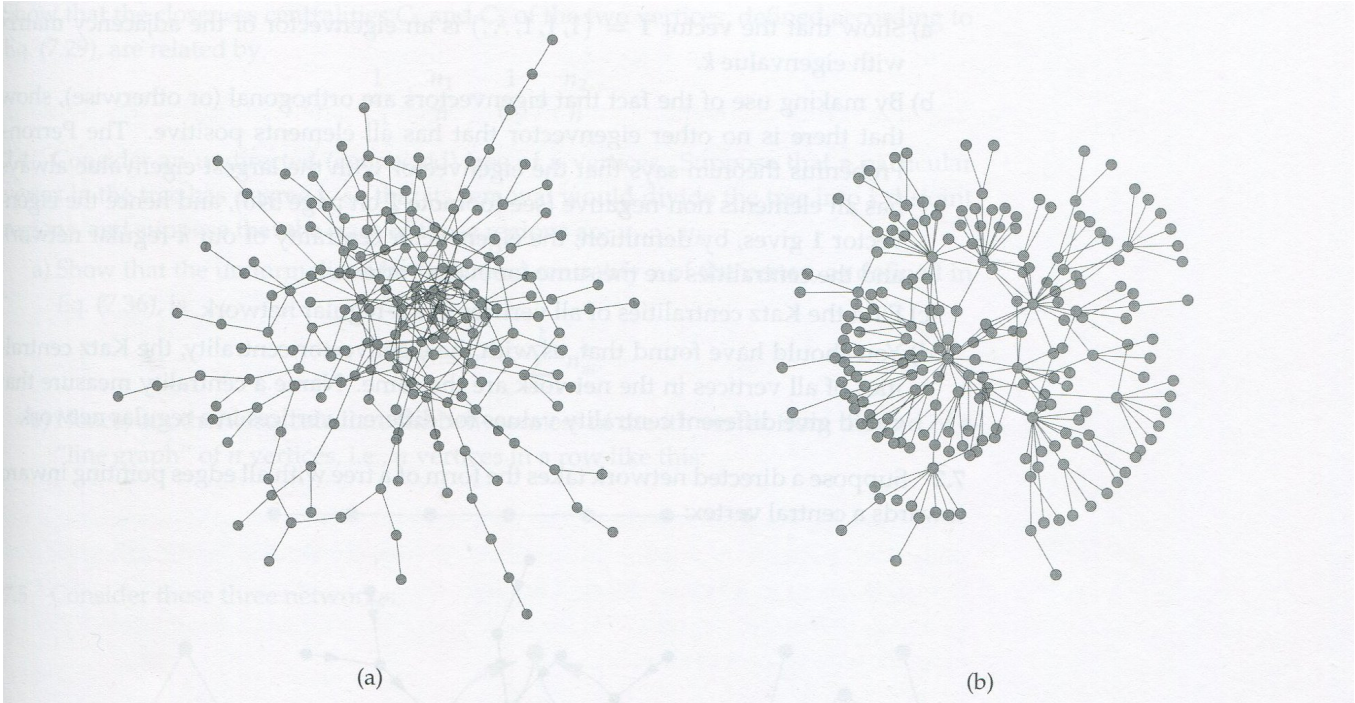


Figure 7.12: Assortative and disassortative networks. These two small networks are not real networks—they were computer generated to display the phenomenon of assortativity by degree. (a) A network that is assortative by degree, displaying the characteristic dense core of high-degree vertices surrounded by a periphery of lower-degree ones. (b) A disassortative network, displaying the star-like structures characteristic of this case. Figure from Newman and Girvan [249]. Copyright 2003 Springer-Verlag Berlin Heidelberg. Reproduced with kind permission of Springer Science and Business Media.

Modularity

- The **modularity** is a measure of the **assortativity** of a graph.
- Let c_i be the attribute of node i in the graph.
- Let us assume that there are ℓ possible attributes :

$$c_i \in \{1, 2, \ldots \ell\}$$

- N_m is defined as the number of links that connect similar nodes in the sense of the value c_i

$$N_m = \sum_{\text{edges } (i,j)} \delta(c_i, c_j) = \frac{1}{2} \sum_{ij} A_{ij} \delta(c_i, c_j)$$

- where $\delta(c_i, c_j)$ the Kronecker function $\delta(x, y) = 1$ if $x = y$, and 0 otherwise.
- The factor $1/2$ in front of the sum is due to the fact that each pair is counted twice.

Modularity

- This number N_m should be compared to a reference value to see if it indicates assortativity or not.
- For this reason, one computes the expected number N_a of links one would have if similar nodes they were connected at random.
- For this, we impose to preserve the degree of each node.

Modularity

- One has $2m$ edge extremities to be connected at random in the graph
- If node i has degree k_i, the probability that an edge taken at random is attached to i is $k_i/(2m)$.
- Let us consider the k_j edges attached to node j. Each of them has a probability $k_i/(2m)$ to reach i.
- The expected number of links between i and j is then

$$\frac{k_i k_j}{2m}$$

if nodes are connected at random.

Modularity

- Therefore, the expected number N_a of links between similar nodes in a random graph is

$$N_a = \frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(c_i, c_j)$$

- $N_m - N_a$ gives the difference between the measured and expected number of edges connecting similar nodes
-

$$N_m - N_a = \frac{1}{2} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

Modularity

The quantity

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

is called **modularité**.

- By construction, $Q < 1$.
- If $Q > 0$, there is assortativity.
- If $Q < 0$ there is deassortativity

The matrix

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

is caleed the **modularity matrix**. It is common in the analysis of complex network

Modularity

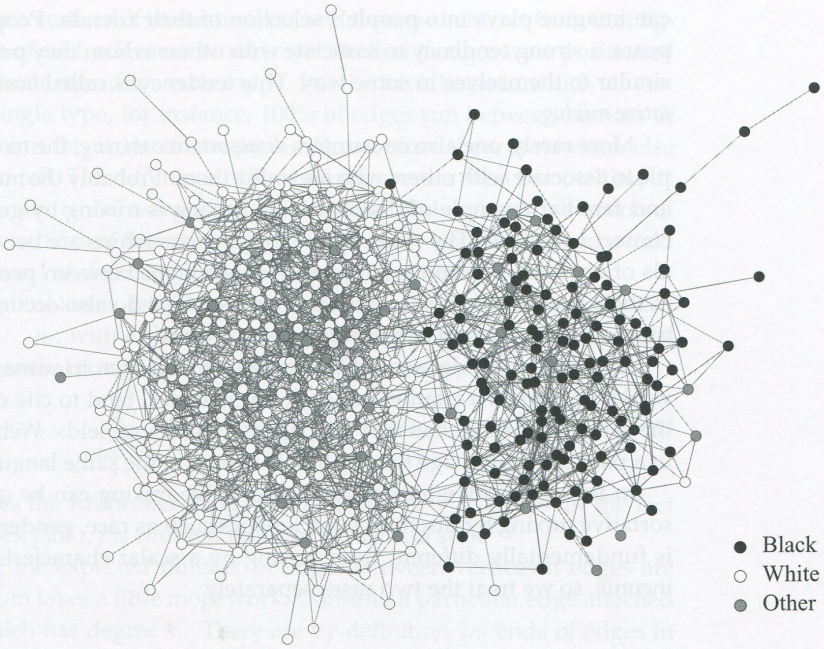


Figure 7.10: Friendship network at a US high school. The vertices in this network represent 470 students at a US high school (ages 14 to 18 years). The vertices are color coded by race as indicated in the key. Data from the National Longitudinal Study of Adolescent Health [34,314].

For the cases of the friendship network in this American school, one gets

$$Q = 0.305$$

which clearly indicates an assortative mixing with respect to the ethnic origin.

5.7 Clustering and Community search

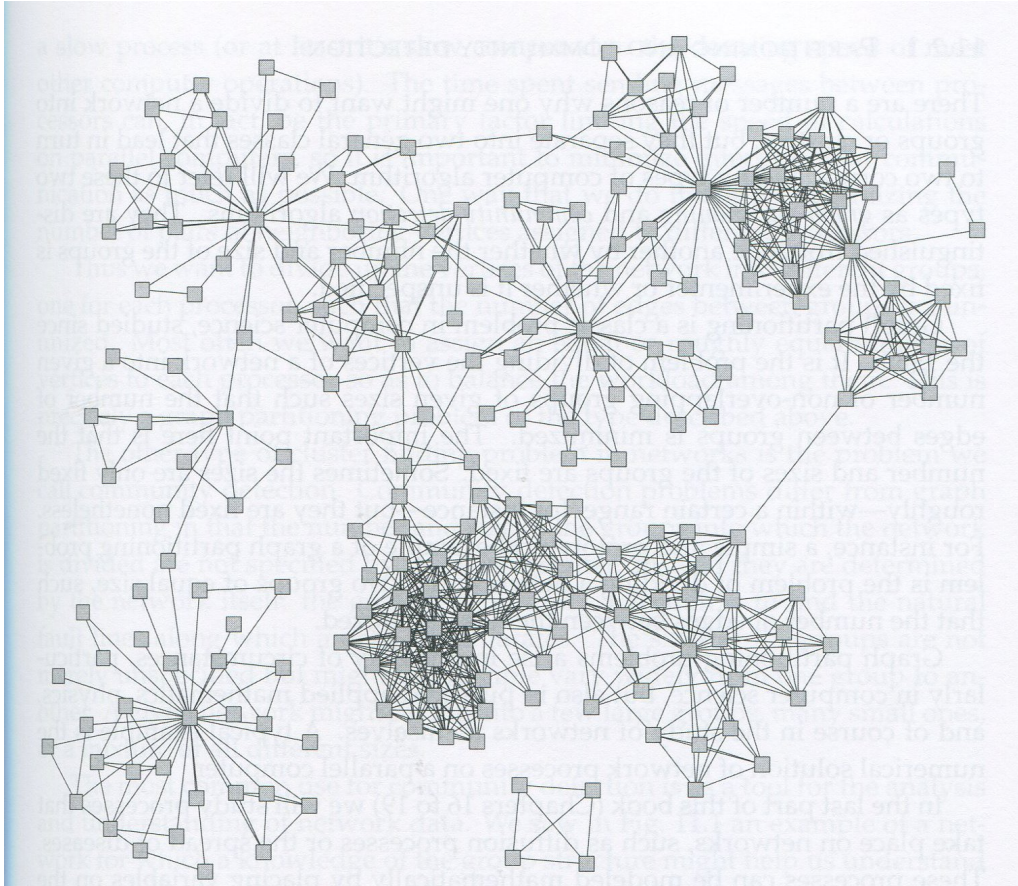


Figure 11.1: Network of coauthorships in a university department. The vertices in this network represent scientists in a university department, and edges links pairs of scientists who have coauthored scientific papers. The network has clear clusters or "community structure," presumably reflecting divisions of interests and research groups within the department.

The graph of publications of a department reveals its structure in laboratories.

Community detection

- This is the division of a graph in sub-graphs, or **communities**.
- Informally, a community is a a group of nodes that have more in common together than with nodes in another community.
- The number of communities and their size is unknown a priori, and should be discovered with an appropriate algorithm.

Community detection algorithms

- There are several algorithms in the literature. They differ by the criteria used to defined a community.
- A common approach consists in labeling the nodes so as to maximize the **modularity** or **assortativity**. Nodes with the same label belong to the same community.
- Another criteria is based on the **betweenness centrality** : by cutting the links of high betweenness will isolate the communities

Communities based on the betweenness centralitya

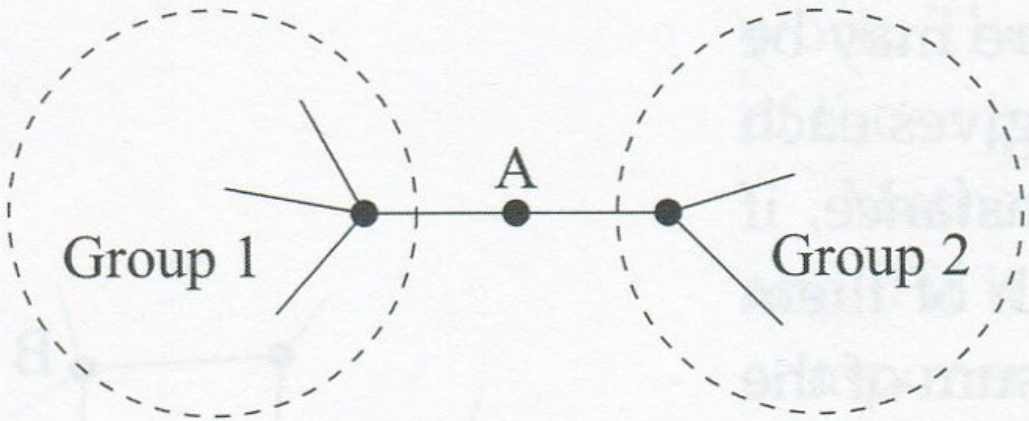


Figure 7.2: A low-degree vertex with high betweenness. In this sketch of a network, vertex A lies on a bridge joining two groups of other vertices. All paths between the groups must pass through A, so it has a high betweenness even though its degree is low.

Detection Algorithms based on the modularity

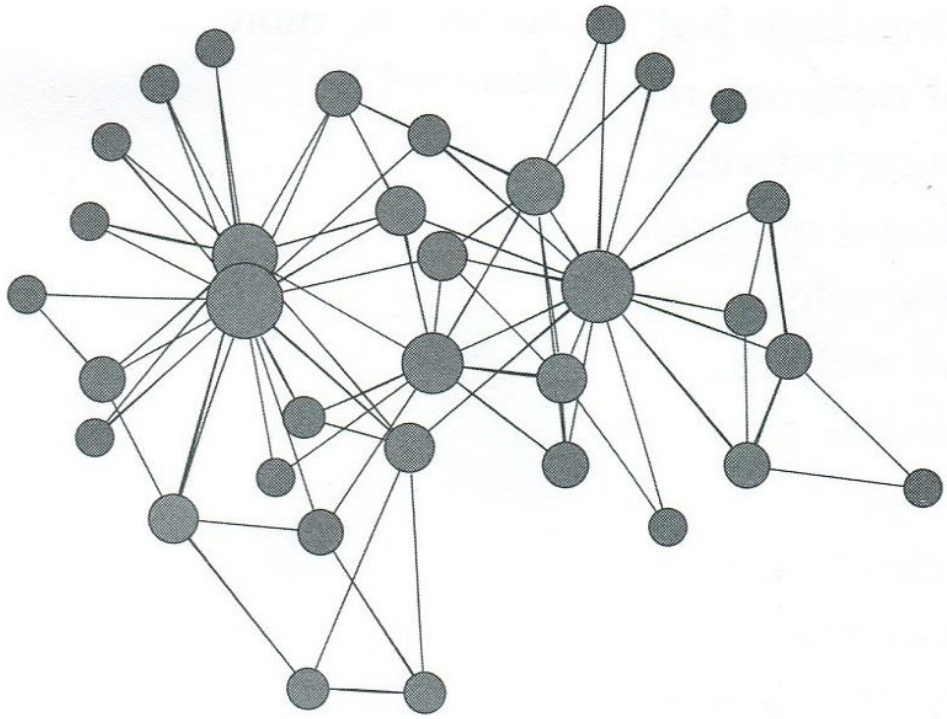
(See M. Newman, section11.7)

- An attribute c_i is associated with each node i, labeling the community to which i belongs.
- The problem is to find the value of c_i which maximizes the modularity Q.
- One can use metaheuristics (such as simulated annealing) to vary the values of c_i until Q reached a maximum.
- A difficulty is that one does not know how many values of c_i should be considered (i.e the number of communities).

Detection Algorithms based on the modularity

- If one is looking for a division in two communities, the algorithm is much easier.
- $c_i = \pm 1$
- One can show that the **eigenvector** u of the modularity matrix B for the largest eigenvalue gives a decomposition in two communities :
- One chooses $c_i = 1$ if $u_i > 0$ and $c_i = -1$ otherwise.

Example : Zachary's Karate club



- This graph was established by the sociologist Zachary, in the 1970s.
- It indicates the friendship between the members of a karate club in a US university.

(See M. Newman, Fig 1.2, p. 6 and Fig 11.4, p. 373)

The Karate club : community detection

- At some stage, the members of the club start disagreeing about the membership fee : increase them or not.
- After some times, the club split and a second club was created.
- If one applies a community detection on the initial friendship graph, one obtains the two communities shown in next the Figure (gray regions).
- The nodes are shown in black or white, according to the actual division of the club members.

The Karate club : community detection

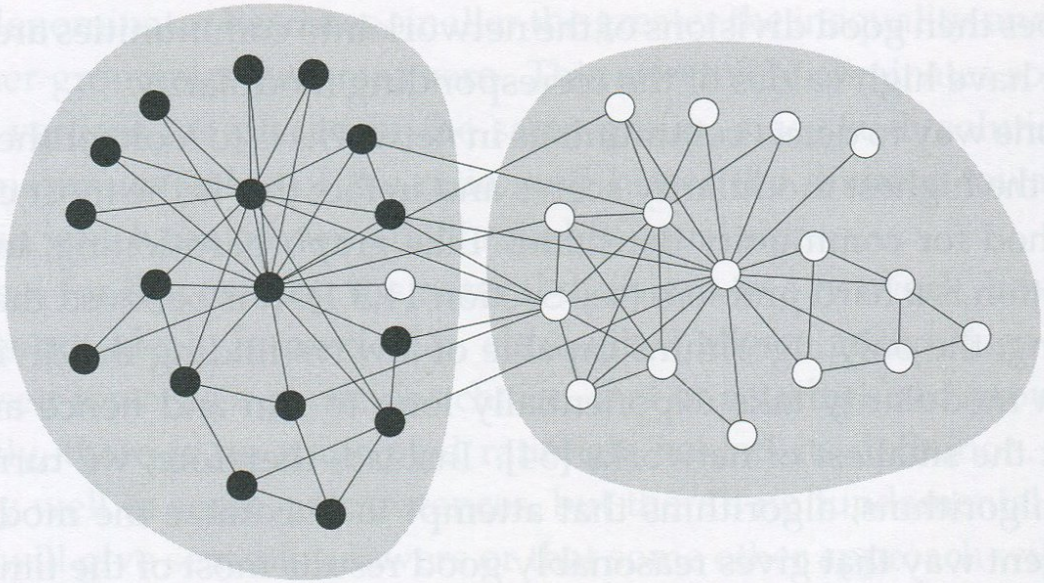


Figure 11.4: Modularity maximization applied to the karate club network. W
we apply our vertex-moving modularity maximization algorithm to the karate c
network, the best division found is the one indicated here by the two shaded regi
which split the network into two groups of 17 vertices each. This division is very nea
the same as the actual split of the network in real life (open and solid circles), follow
the dispute among the club's members. Just one vertex is classified incorrectly.

There is an excellent agreement between the prediction (community detection) and reality (actual split). Only one individual is wrongly classified among the 34 members.

5.8 Models of graphs and synthetic graphs

In practice, large graphs are often generated by a computer model.

- Random graph (Erdös-Rényi, 1960)
- Watts-Strogatz (1998) : regular network with random short cuts (small world graph)
- Barabási-Albert model : preferential attachment (scale-free network)
- ...

Random graphs (Erdös-Rényi)

Each of the $n(n-1)/2$ possible edge is kept with probability p.

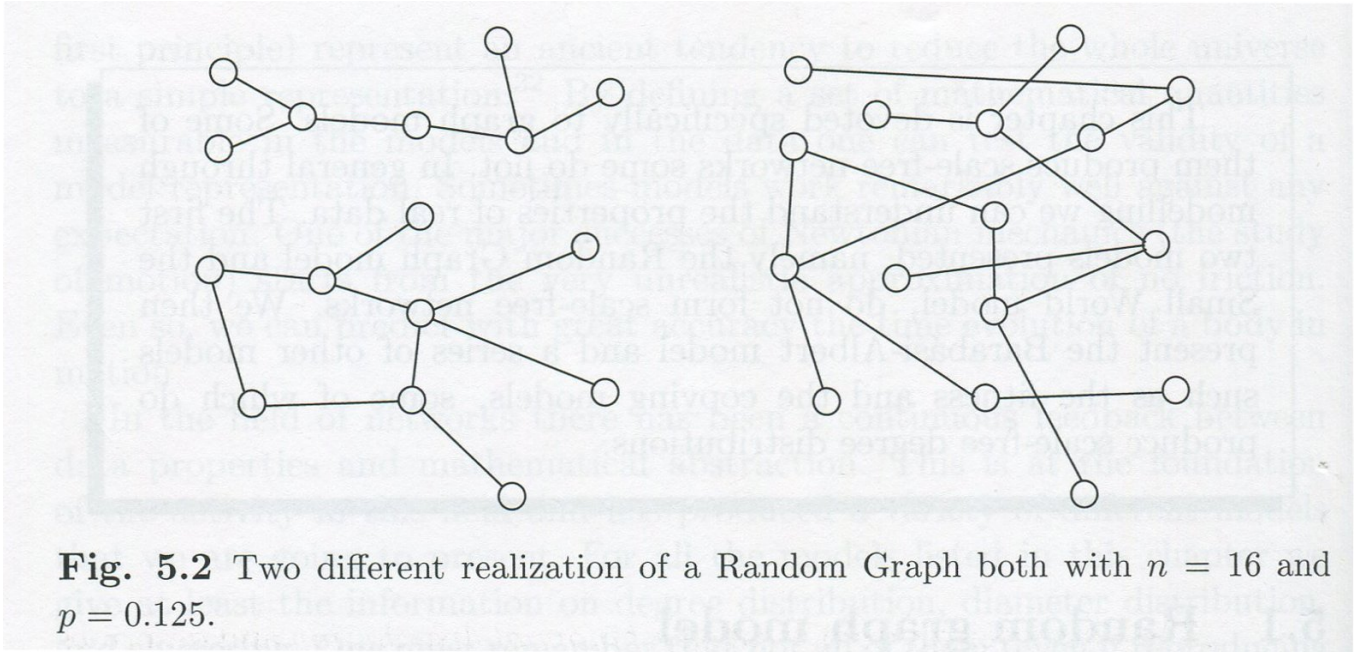


Fig. 5.2 Two different realization of a Random Graph both with $n = 16$ and $p = 0.125$.

The degree distribution follows a binomial law and

$$\langle k \rangle = np$$

The clustering coefficient can be computed analytically.

Watts-Strogatz

Small world networks obtained by adding random edges between nodes of a regular network.

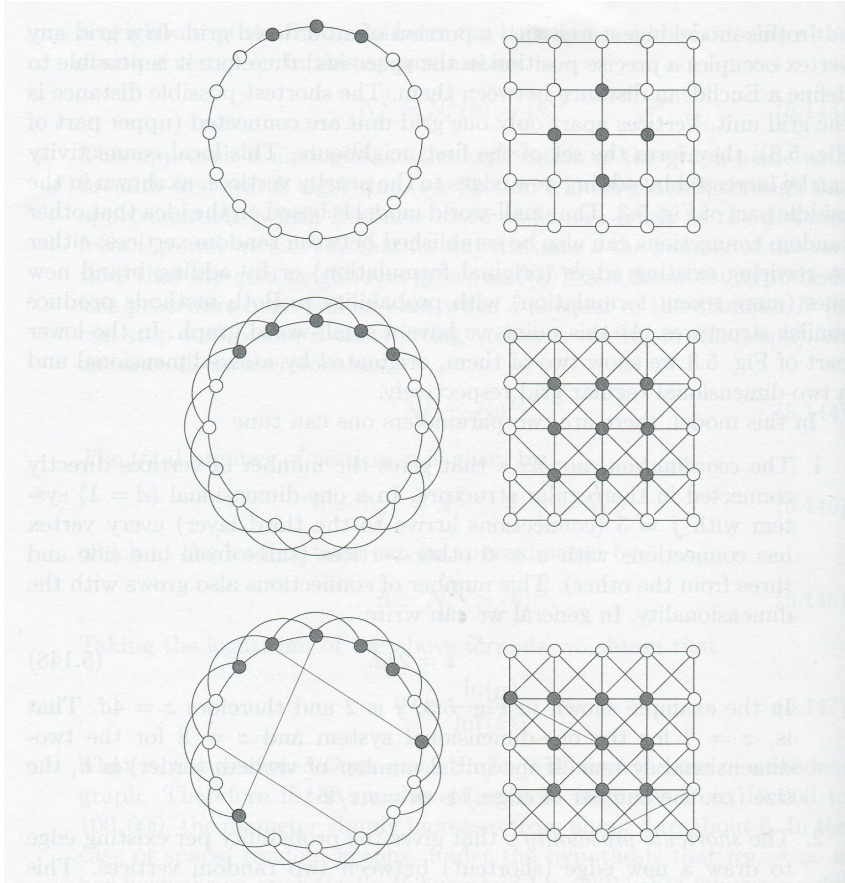


Fig. 5.3 On the first row a regular one-dimensional lattice on the left and a regular two-dimensional lattice on the right. In the second row the same lattices with extra edges increasing the local connectivity. On the last line we have the small-world lattices with shortcuts.

Barabási-Albert Model

- One considers n_0 initial vertices, connected in an arbitrary way.
- One adds the $n - n_0$ remaining nodes, one by one to the graph.
- For each of them, one creates m_0 edges to the already placed nodes.
- These edges are connected to a node i with probability

$$p_i = k_i / \sum_j k_j$$

Barabási-Albert Model

This mechanism is called **preferential attachment** as the probability to connect to a node increases as the degree of this node increases.

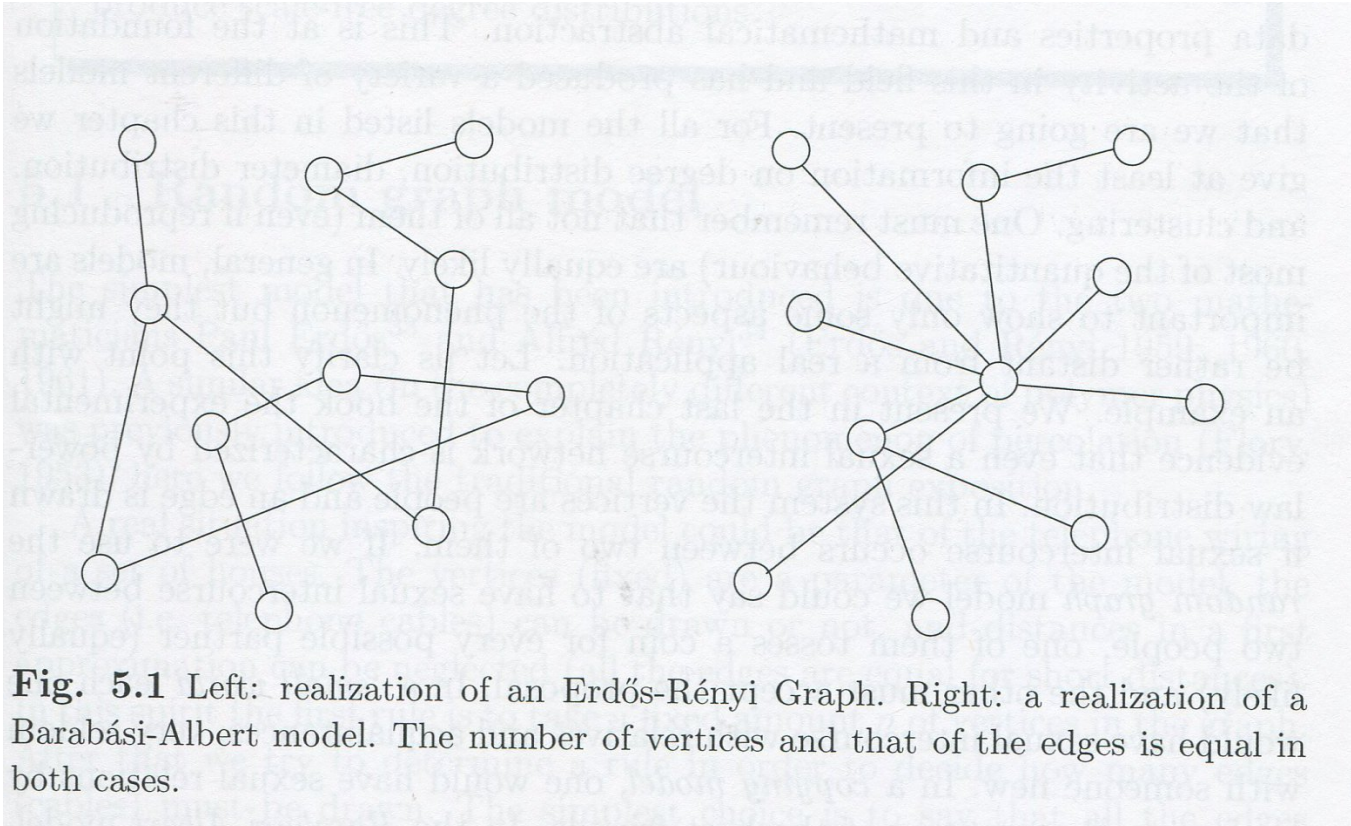


Fig. 5.1 Left: realization of an Erdős-Rényi Graph. Right: a realization of a Barabási-Albert model. The number of vertices and that of the edges is equal in both cases.

Barabási-Albert Model

- Degree distribution is a power law : $p_k \propto k^{-\alpha}$
- Typical exponent $\alpha = 2.26$
- Scale-free network.