Course of Bioinformatics
Master in Cellular and Molecular Biology
University of Torino

Networks and models of gene regulation

University of Torino

GECO - Computational Epidemiology Group

Mario Giacobini

GECO - Computational Epidemiology Group,
Department of Veterinary Sciences

CBU - Computational Biology Unit, Molecular Biotechnology Center

ARCS - Applied Research on Computational Complex Systems Group,
Department of Computer Science

University of Torino, Italy

Some words you could hear

Kevin Bacon

A/H1N1

Immanuel Kant

Internet

Saddam Hussein

rivers

pandemic

peer-to-peer

Will Smith

La Divina Commedia

Donald Sutherland

sexual contacts

Facebook

World Wide Web

trains

airports

telephones

to get to

gene regulation

epidemiology

Marc Vidal

bacteria

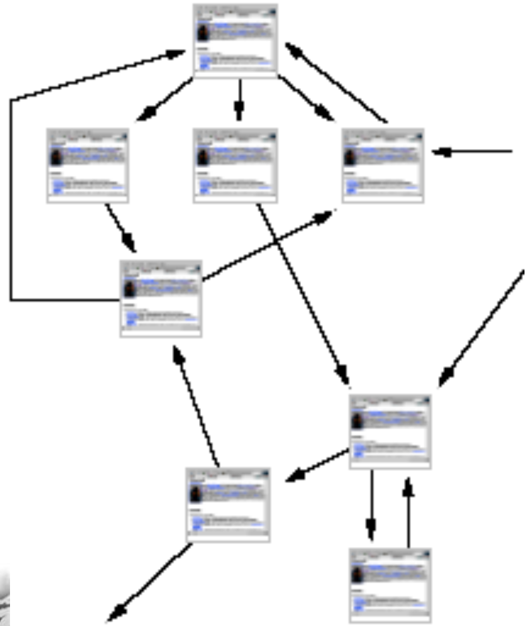
interactions

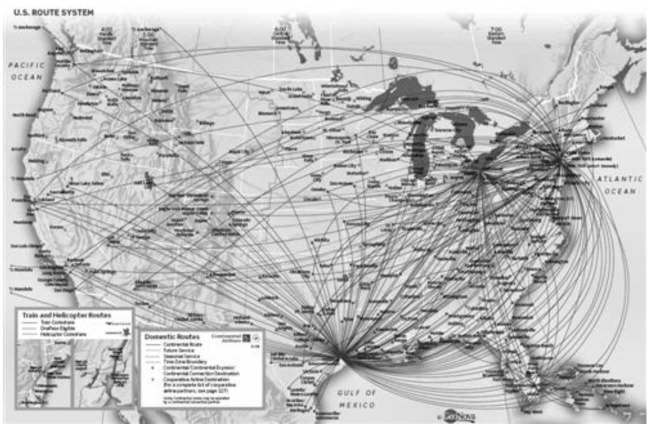
viruses

parasites

metabolic pathways

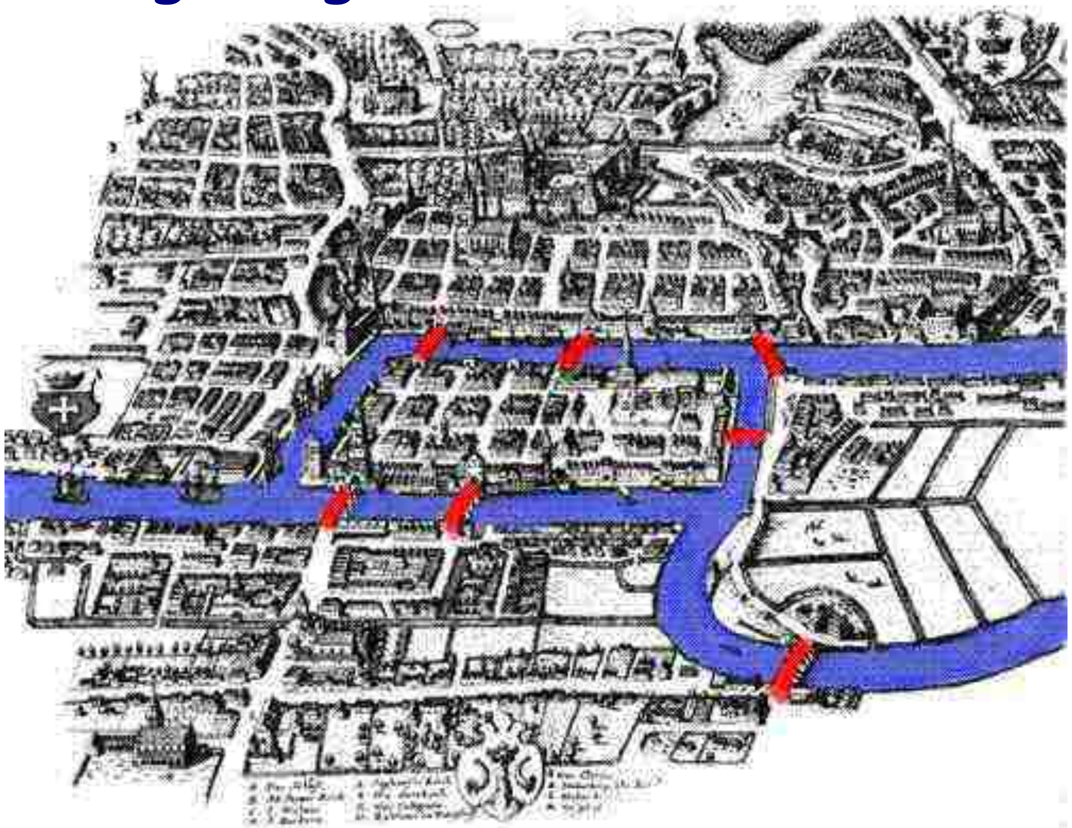
But what do these phenomena have in common?





Representing reality

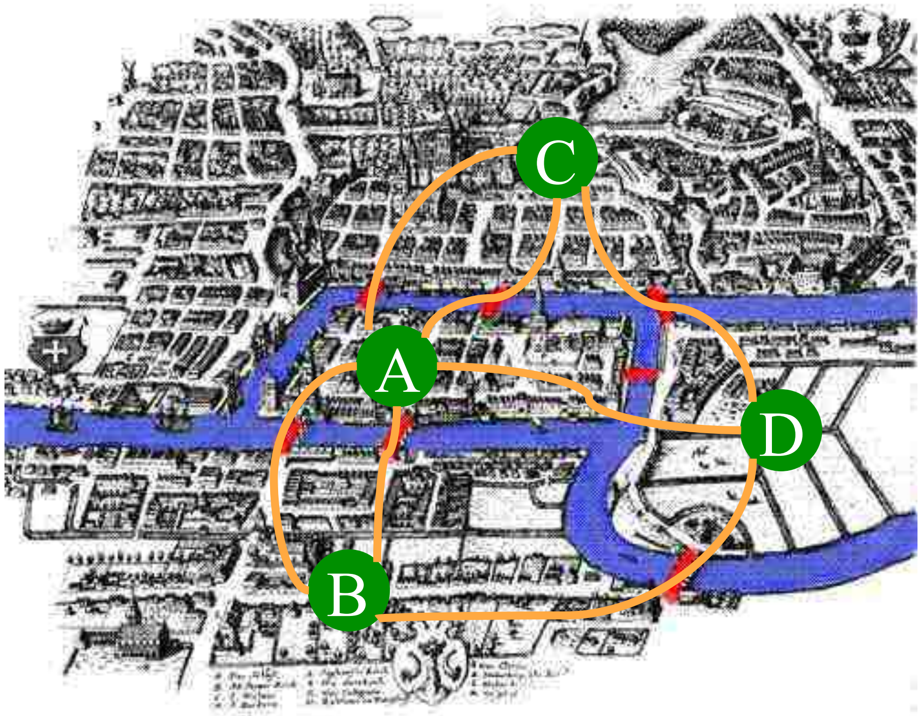
the seven bridges of Königsberg



Leonard Euler [1735]:

"The question is whether a person can plan a walk in such a way that he will cross these bridges once but not more than once. [...] I formulated the following very general problem for myself: given any configuration of the river and the branches into which it may divide, as well as any number of bridges, to determine whether or not it is possible to cross each bridge exactly once."

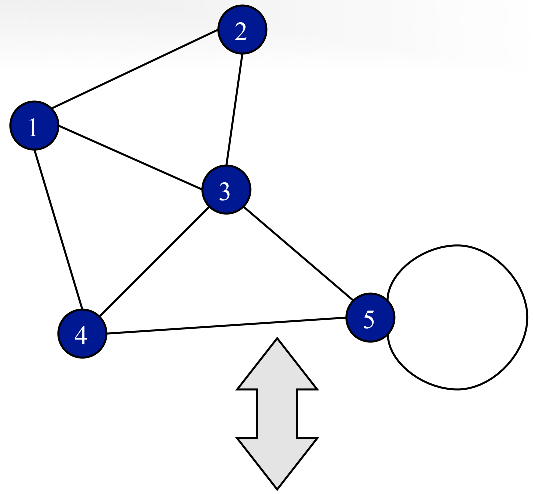
Euler realized that all problems of this form could be represented by replacing area of land by points (called vertices or nodes), and the bridges to and from them by arcs (called edges)



every vertex with an odd number of arcs attached to it has to be either at the beginning or at the end of the path:

a graph has an *euler walk* if at most two vertices have an odd degree

Graphs



a graph (also called **network**) is formed by:

1. a set of nodes also called vertices
2. a set of edges that connect couples of nodes

a graph *G* is defined by the couple ⟨*V*,*E*⟩:

- the number of vertices is the order of *G*
- the number of edges is the size of *G*

A	**1**	**2**	**3**	**4**	**5**
1	0	1	1	1	0
2	1	0	1	0	0
3	1	1	0	1	1
4	1	0	1	0	1
5	0	0	1	1	1

two nodes are said **neighbors** if they are linked by an edge

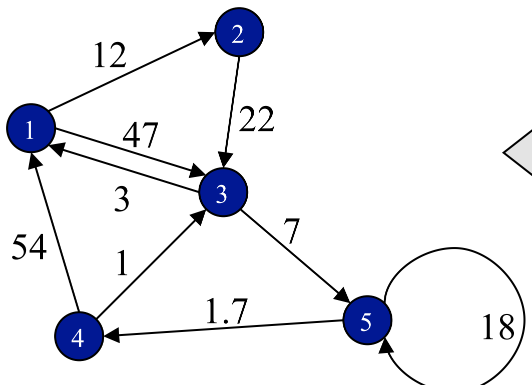
the number of neighbors of a node is called the **degree** of the node

a sequence of nodes $\langle n_1,n_2,\ldots,n_k\rangle$ for which every two nodes ni and ni+1 are adjacent is called walk (or **path**) of length k

- edges may have a direction:
  - these are called **directed or oriented graphs, digraphs**
  - the adjacency matrix is no longer symmetric
  - nodes' in-degrees and out-degrees are defined

- finally, edges may have a weight:
  - these are called **weighted graphs**
  - the values in the adjacency matrix won't be binary
  - nodes have weighted degrees

- example: a directed and weighted



A	1	2	3	4	5
1	0	12	47	0	0
2	0	0	22	0	0
3	3	0	0	0	7
4	54	0	1	0	0
5	0	0	0	1.7	18

Some characteristics of graphs

- the **diameter** D_G of a graph $G(V,E)$ is defined as the maximal distance that can be find between any two of its nodes
- the clustering coefficient C_i of a node i $\in V$ in $G(V,E)$ is the probability that two neighbors of node *i* are connected by an edge
- the **clustering coefficient** CC_G of $G(V,E)$ is the average among the clustering coefficient of its nodes
  - ✓ CC_G can be interpreted as the probability that any two nodes in the graph sharing a common neighbor are connected

- the **degree distribution** $P_G(k)$ in a graph G is the law that describes the probability that a node has degree k

- the **average degree** $\langle k \rangle$ in a graph is:
  - ✓ $\langle k \rangle = 2 \times |E| \,/\, |V|$ if the graph is undirected
  - ✓ $\langle k \rangle = |E| \,/\, |V|$ if the graph is directed (i.e. a digraph)

- the **distance distribution** $P_G(d)$ in a graph G is the law that describes the probability that describes the probability that a couple of nodes has distance d in G

- the **average path length** APL_G in a graph G is the average distance between each couple of nodes

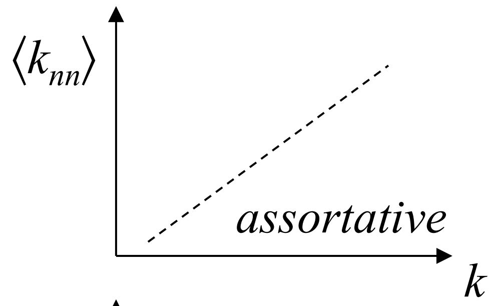
- ↳ it is sometimes observed that in some cases nodes with high degree tend to be connected to nodes with nodes with high degree, while sometimes the opposite happens

we talk of **nodes correlation**:

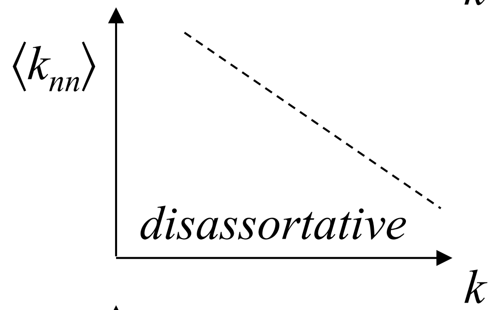
- ✓ in the first case the graph exhibits an assortative mixing or simply an **assortativity**
- ✓ in the second case we talk of **disassortative mixing**
- ✓ it is possible to measure the assortativity of a graph with the **assortative coefficient** r, a particular case of the Pearson correlation coefficient applied to nodes' degrees:
  - when $r > 0$ we have assortative mixing
  - when $r < 0$ we have disassortative mixing

✓ it is also possible to measure the degree of assortativity by plotting on a bidimensional graph the values k of the degrees of the nodes coupled with the relative average degrees of their neighbors $\langle k_{nn} \rangle$

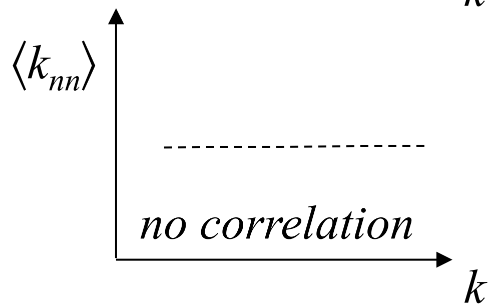
- if the values of $\langle k_{nn} \rangle$ grow when the values of k grow, then we talk of assortative mixing



- if the values of $\langle k_{nn} \rangle$ decrease when the values of k decrease, then we talk of disassortative mixing



- if the values of $\langle k_{nn} \rangle$ remain constant the values of k grow, then we do not have any assortativity



- to define the larger importance of a node with respect to the others in a graph, several measures have been proposed

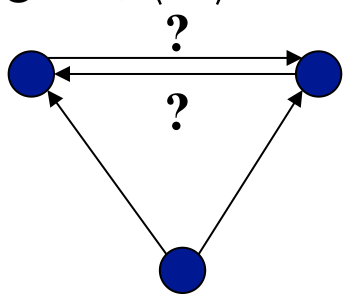
- the most central node can be:
  - ✓ the node with the lowest average distance to all other nodes of the graph (**node centrality**)
  - ✓ the node through which pass most of the shortest path connecting all nodes in the graph, that determine the distances between all nodes (**node betweenness**)

- in a similar way **edge betweenness** can be defined

- a **cluster** is a subset of nodes with more edges connecting each other than those connecting them to the rest of the graph

… in directed and wigthed graphs

- for directed graphs, the definitions of diameter, average path length, assortative e disassortative mixing, node and edge betweenness, clustering and community do not change
- the degree distribution is substituted by the **in- and out-degree distributions**
- the average degree $\langle k \rangle$ by the average in- and out-degrees, $\langle k^{in} \rangle$ and $\langle k^{out} \rangle$
- the clustering coefficient results difficult to define

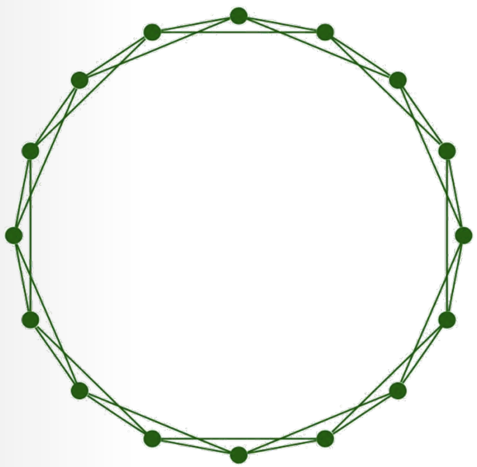


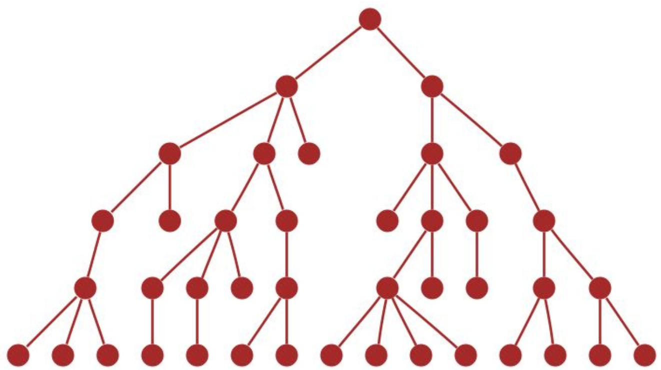
- in weigthed graphs, the different characteristics are redefined taking into account the weigths of the edges

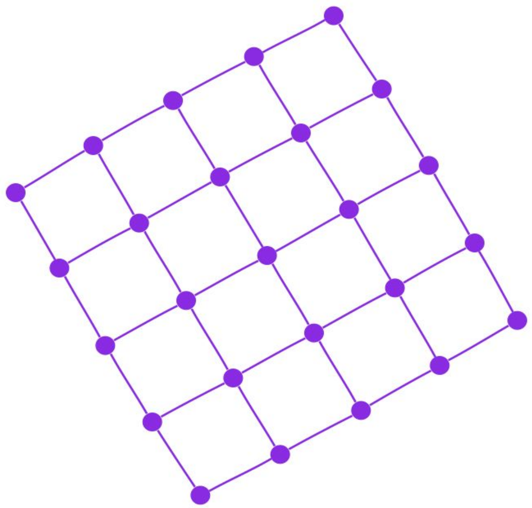
Regular structures

until the second half of the XXth century, many physical, social and biological phenomena were modeled and analyzed using tools of graph theory:

most models used regular structures like lattices and trees



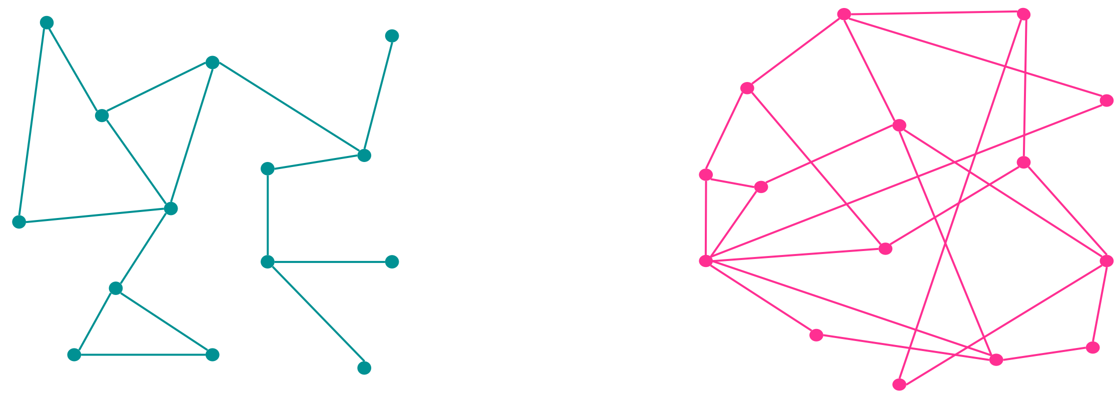


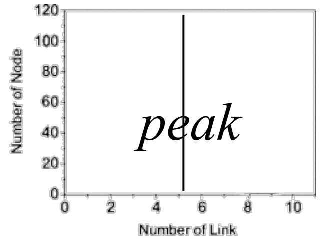
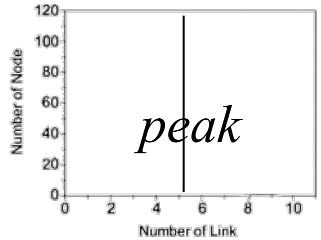
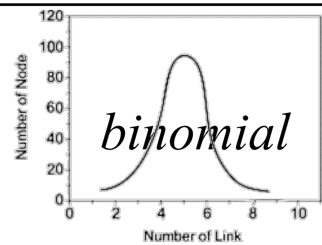


Random structures

the theory of random graphs was introduced by Pàl Erdös and Alfréd Rényi in 1959-61 after Erdös discovery that probabilistic methods were often useful in tackling problems in graph theory

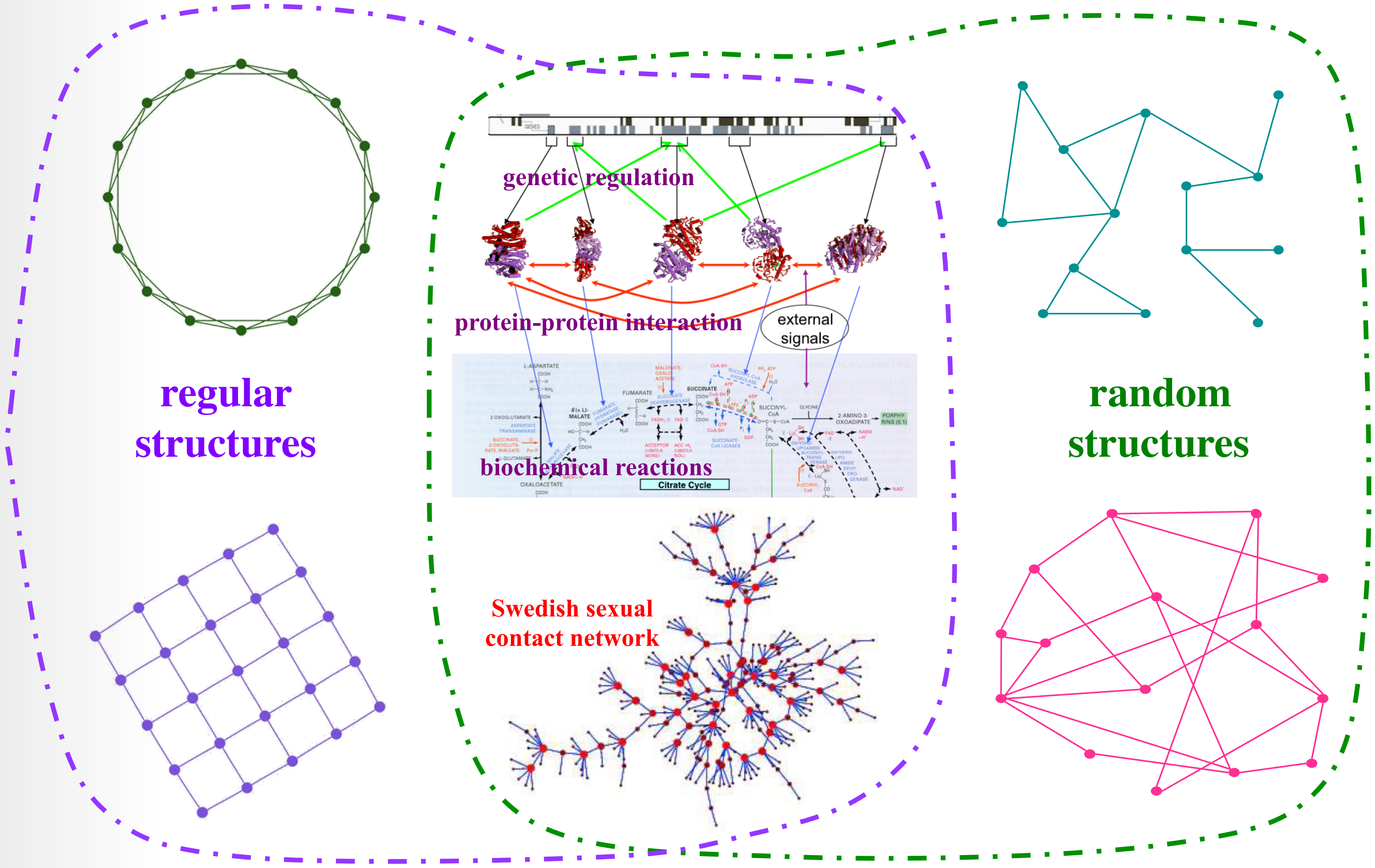
an Erdös-Rényi random graph $G_{N,p}$ is constructed connecting each pair of the N vertices of the graph with probability $p \in [0,1]$ (and non connecting them with probability $1-p$)



graph \ property	**APL**	**CC**	**DD**
[*r*]-ring	$\frac{N}{2r}$ $\uparrow$	$\frac{3(r-1)}{2(2r-1)}$ $\uparrow$	
[*1*$_M$]-torus	$\frac{\sqrt{N}}{2r}$ $\uparrow$	$\frac{3}{7}$ $\uparrow$	
ER-G$_{N,p}$	$\frac{ln(N)}{ln(pN)}$ $\downarrow$	p $\downarrow$	

where *N* is the number of nodes in the graph

Networks: order vs randomness



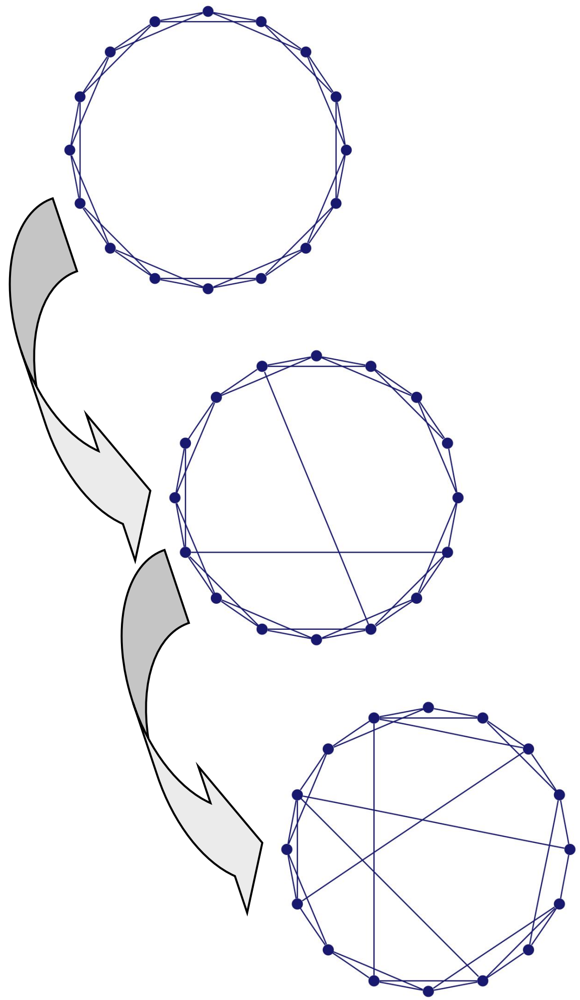
Small-world phenomenon

- theory: *anyone on the planet can be connected to any other person on the planet through a chain of acquaintances that has no more than five intermediaries*
- proposed by F. Karinthy, Hungarian writer, in 1929
- proven mathematically by S. Pool and M. Kochen in 1950's
- life-size test by S. Milgram, American sociologist, in 1967
- popularized by Guare's play 'Six degrees of separation' in 1990
- in 1998, Watts and Strogatz found that real-world networks tend to be highly clustered (like lattices), but have small average path lengths (like random graphs)

Watts-Strogatz small-world graphs

in 1998, Watts and Strogatz found that real-world networks tend to be highly clustered (like lattices), but have small average path lengths (like random graphs)

they raised the possibility of constructing random graphs that have some of the important properties of real-world networks

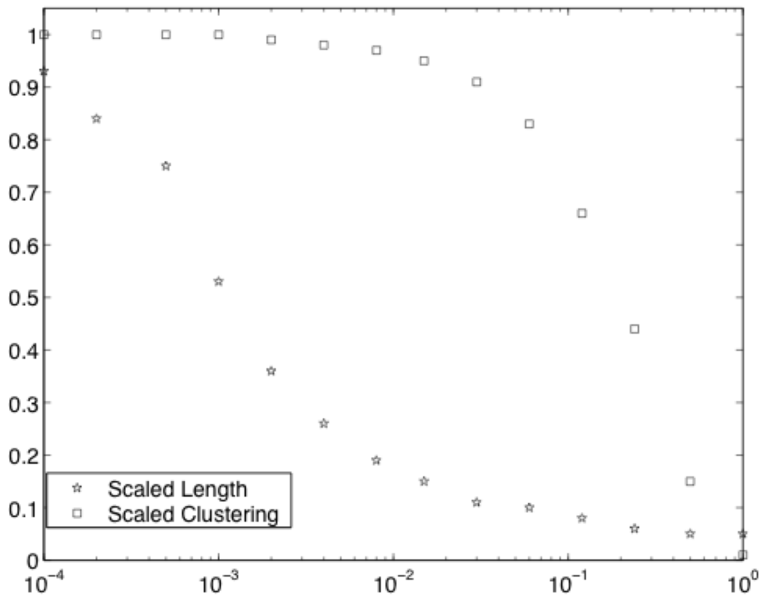


PROPERTIES OF SMALL-WORLD GRAPHS:

$$APL = \frac{N}{k} f(Nk\beta), \text{ where } f(x) \sim \begin{cases} x & \text{for } x >> 1 \\ log(x) & \text{for } x << 1 \end{cases}$$

$$CC = \frac{3(k-1)}{2(2k-1)} (1-\beta)^3$$

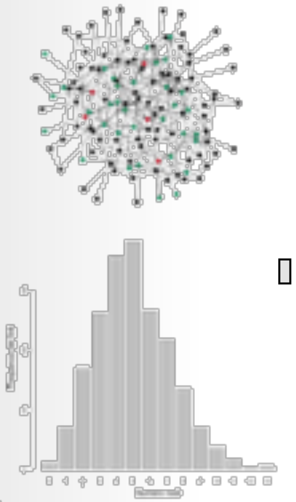
DD ~ binomial



Real-world networks

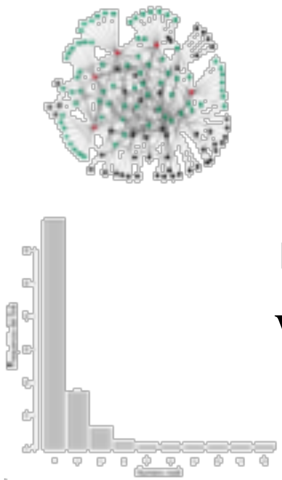
real-world networks are mostly found to be very unlike random graphs and most small worlds in their degree distributions:

far from having a binomial distribution, the degrees of the nodes in most networks are highly right-skewed: their distribution has a long tail of values that are far above the mean



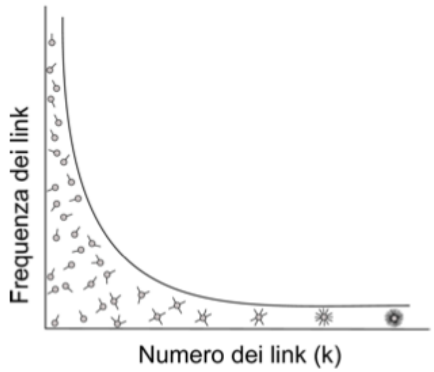
from random

to real graphs



approximation

with a power law

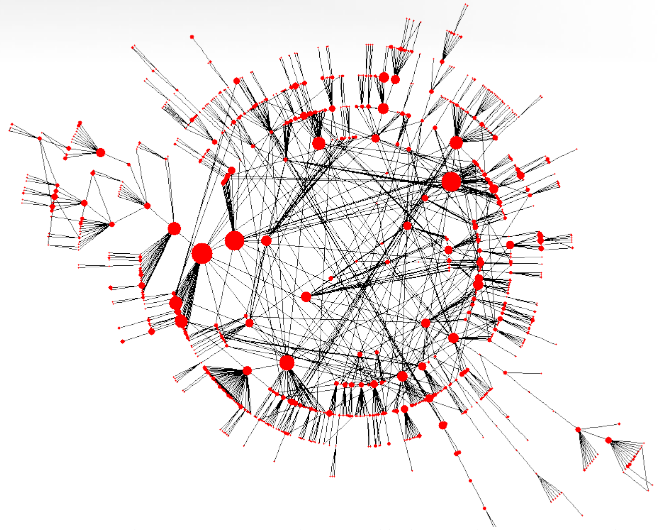


$P(k) \sim k^{-\alpha}$

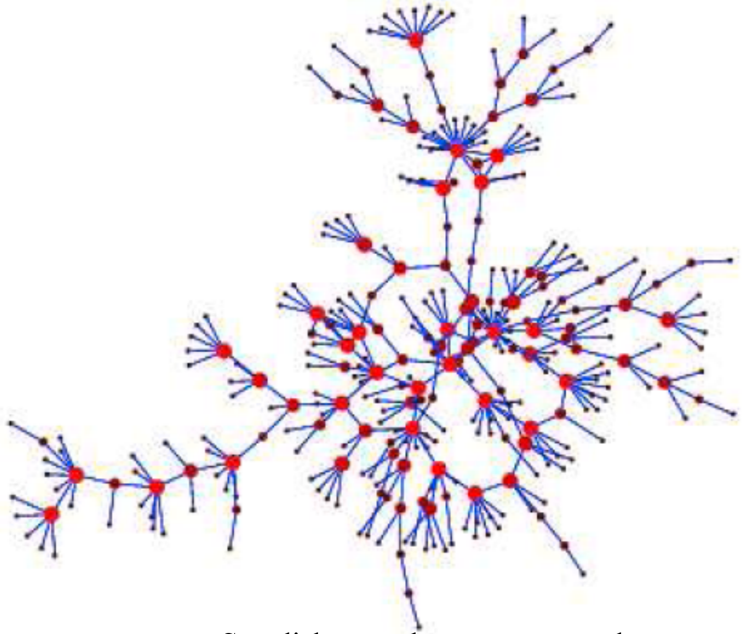
they are usually referred to as ***scale-free networks***

SOCIAL NETWORKS:

- business relationship between companies
- collaboration between film actors
- email communications
- co-authorship between academics
- the patterns of friendship between individuals
- intermarriages between families
- telephone calls
- patterns of sexual contacts
- business communities



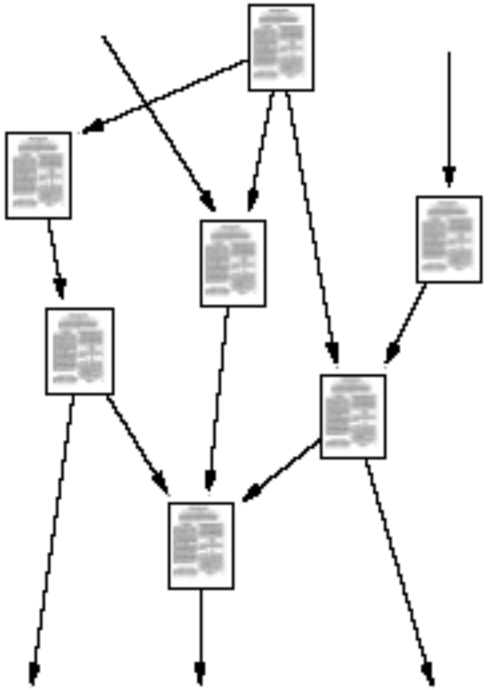
Genetic Programming collaboration network



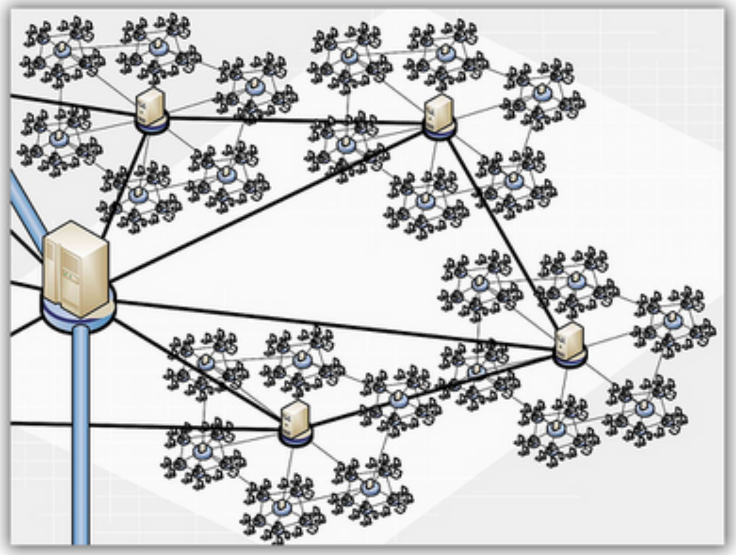
Swedish sexual contact network

INFORMATION NETWORKS:

- citations between US patents
- scientific articles' citations
- the World Wide Web
- peer-to-peer
- linguistics



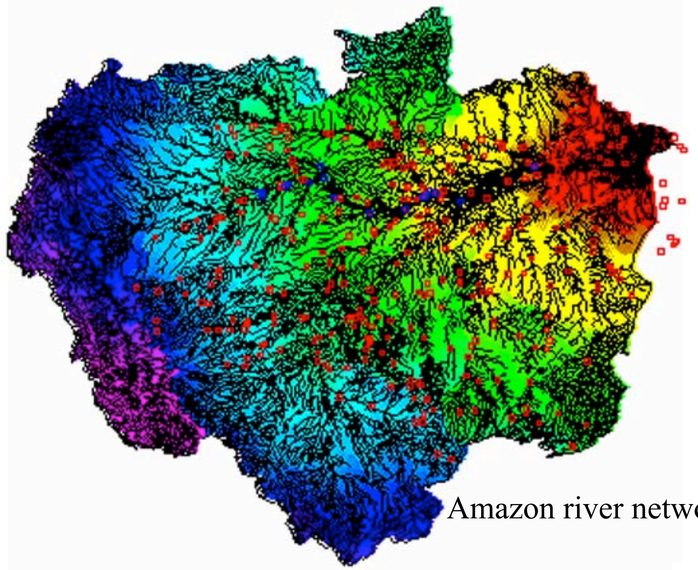
articles citation network



peer to peer network

TECHNOLOGICAL NETWORKS:

- Internet:
  - autonomous systems level
  - routers level
- network of airline routes
- networks of railways
- networks of roads
- river networks
- telephone networks
- mail delivery networks



Amazon river network

USA flight network

BIOLOGICAL NETWORKS:

- food webs of predator-prey interactions between species
- neural and blood vessels
- intra-cellular networks:

  genetic regulation

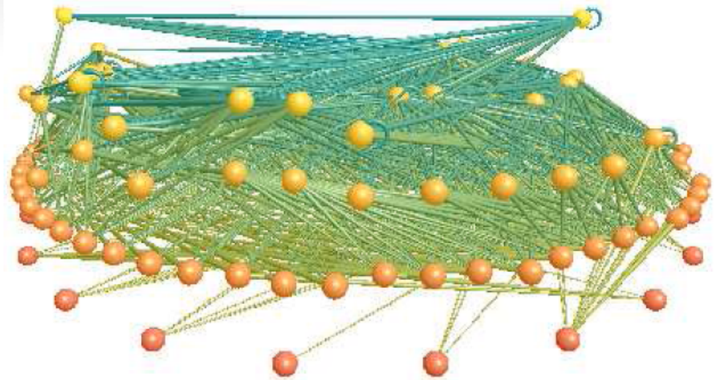
  genome

  protein-protein interaction

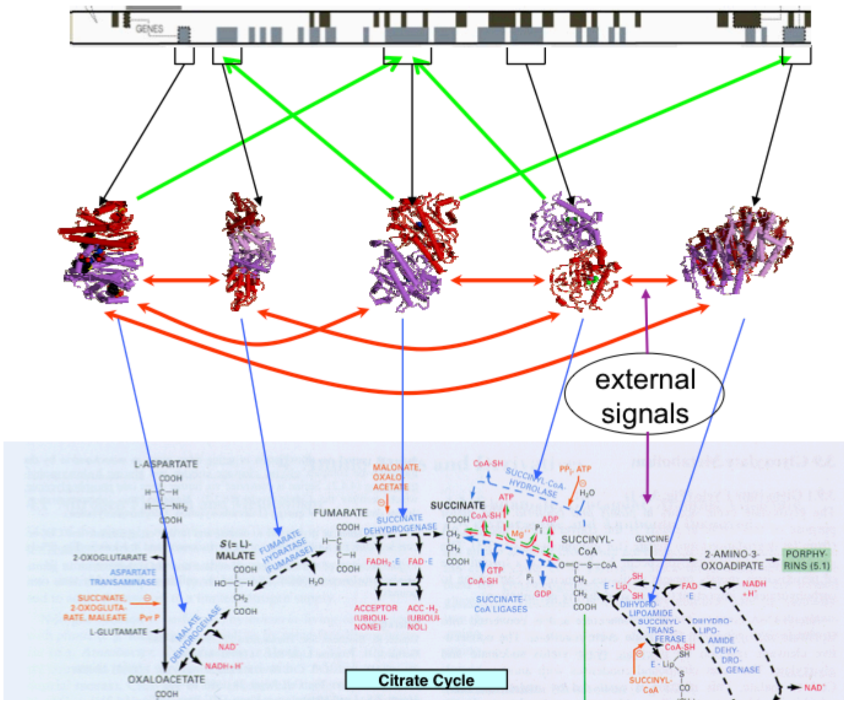
  proteome

  biochemical reactions

  metabolism



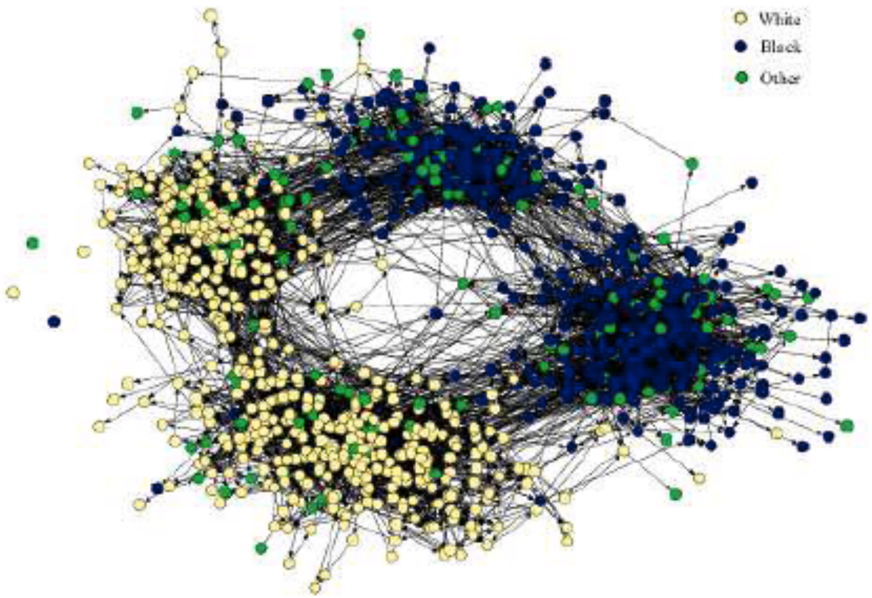
predator-prey network in a freshwater lake



Properties of real-world networks

COMMUNITY STRUCTURES:

most real social networks show community structures, i.e. groups of nodes that have a high density of edges between the group, showing the common experience that people do divide in groups along lines of interest, race, etc



NETWORK RESILIENCE:

networks vary in their level of resilience in vertex removal, mostly resulting in a high robustness with respect to random removals and a low robustness with respect to targeted removals