Our ability to reason and comprehend our world requires the coherent activity of billions of neurons in our brain. Our biological existence is rooted in seamless interactions between thousands of genes and metabolites within our cells. These systems are collectively called complex systems, capturing the fact that it is difficult to derive their collective behavior from a knowledge of the system's components.

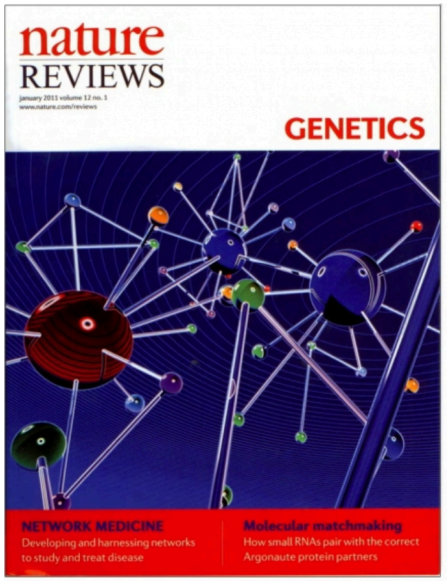
A key discovery of network science is that the architecture of networks emerging in various domains of science, nature, and technology are similar to each other, a consequence of being governed by the same organizing principles. Consequently we can use a common set of mathematical tools to explore these systems.

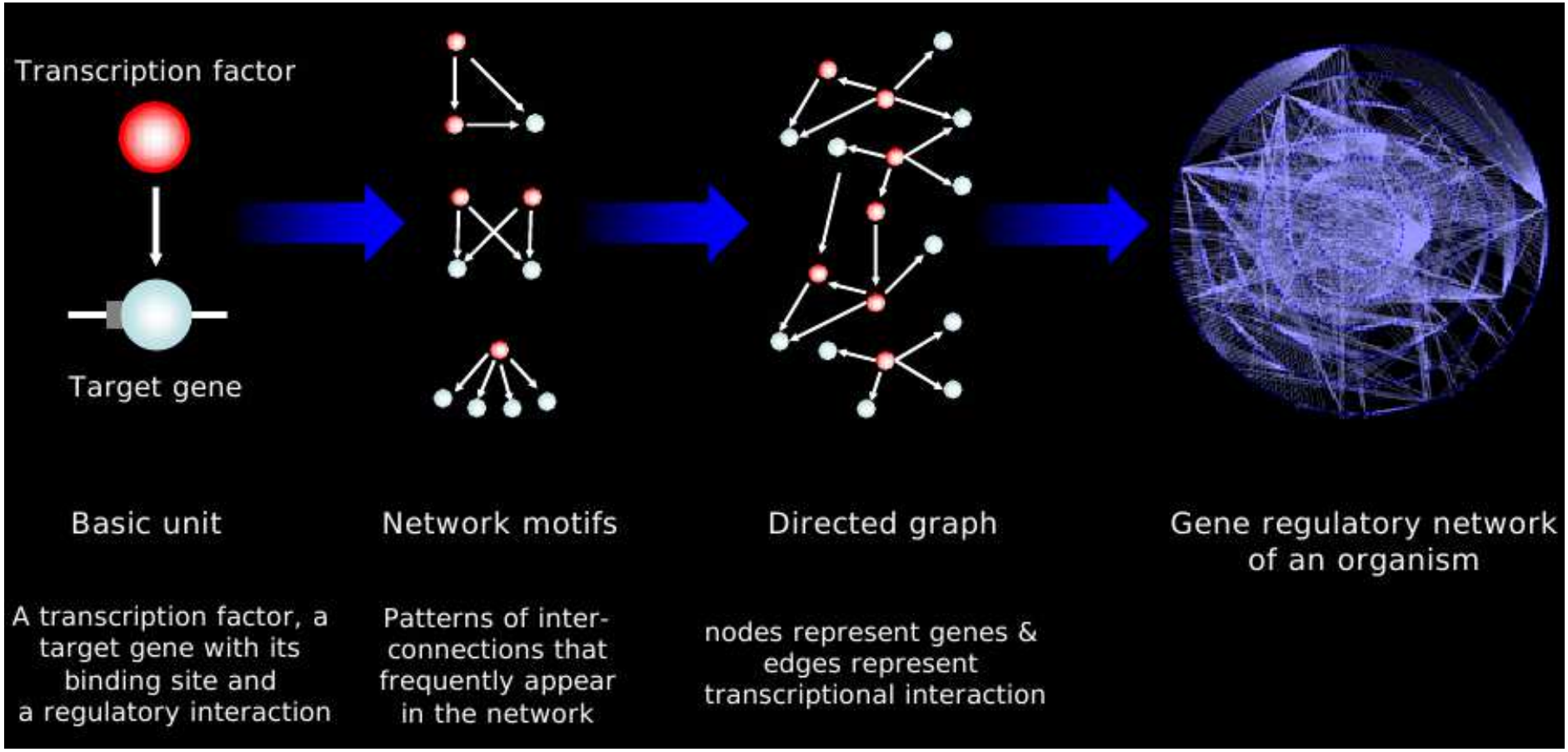
Completed in 2001, the human genome project offered the first comprehensive list of all human genes. Yet, to fully understand how our cells function, and the origin of disease, a f**ull list of genes is not sufficient:** We also **need an accurate map of how genes, proteins, metabolites and other cellular components interact with each other.** Indeed, most cellular processes, from food processing to sensing changes in the environment, rely on molecular networks. The breakdown of these networks is responsible for human diseases.

Network Biology

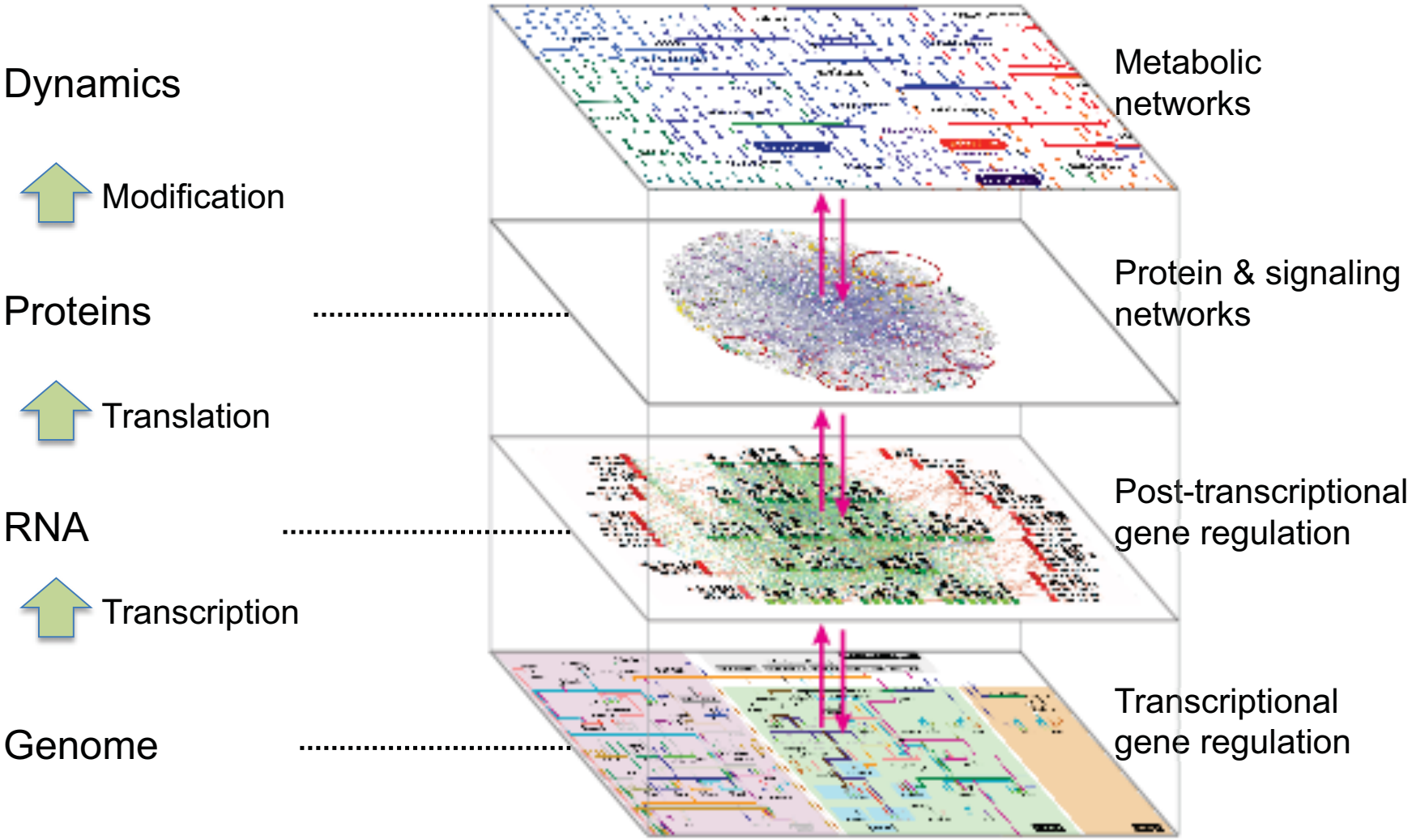
Network Pharmacology

Network Medicine



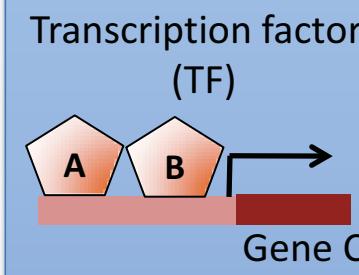

Transcription factor
Target gene
Basic unit
Network motifs
Directed graph
Gene regulatory network of an organism
A transcription factor, a target gene with its binding site and a regulatory interaction
Patterns of inter-connections that frequently appear in the network
nodes represent genes & edges represent transcriptional interaction

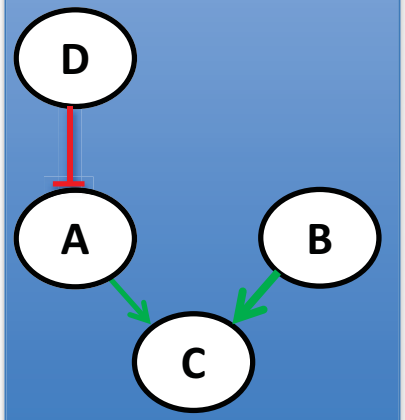
Biological networks at all cellular levels



Five major types of biological networks

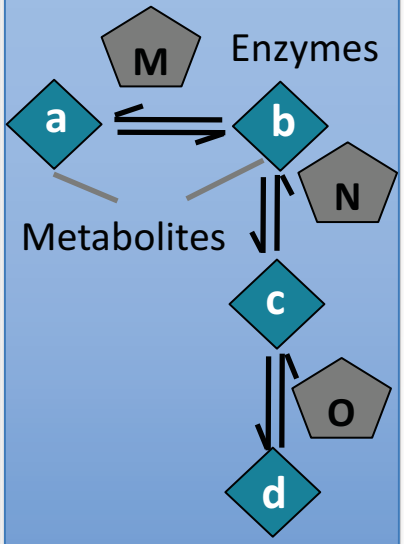
Regulatory network

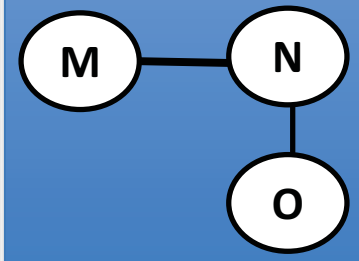




Directed, Signed, weighted

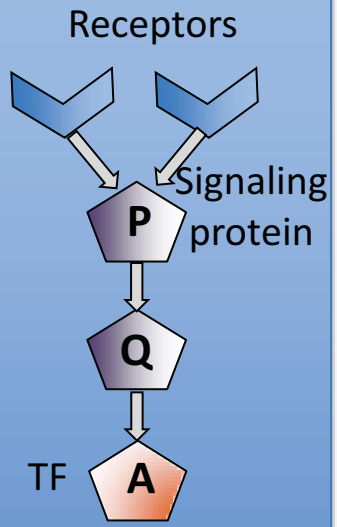
Metabolic network





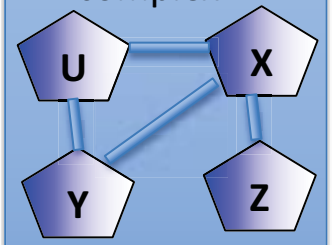
Undirected, weighted

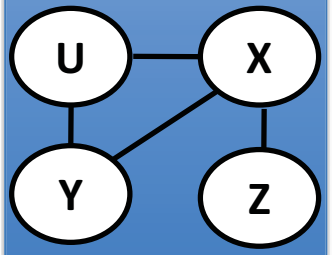
Signaling network



Directed, unweighted

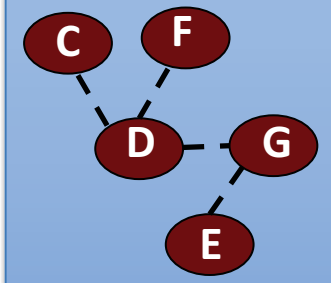
PPI, Protein interaction network

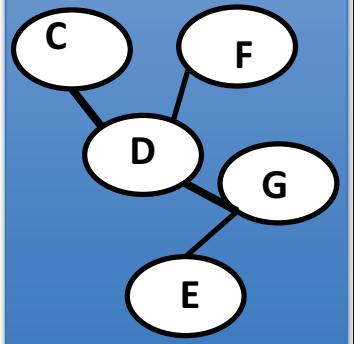




Undirected, unweighted

Functional network (Co-expression)

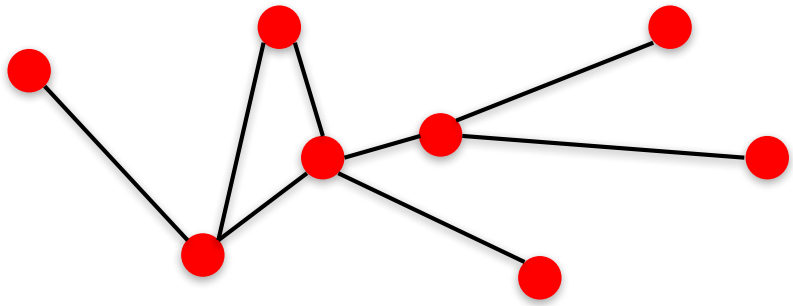




Undirected, weighted

Number of nodes, or N, represents the number of components in the system. We will often call N the size of the network. To distinguish the nodes, we label them with i = 1, 2, ..., N.

Number of links, which we denote with L, represents the total number of interactions between the nodes. Links are rarely labeled, as they can be identified through the nodes they connect.

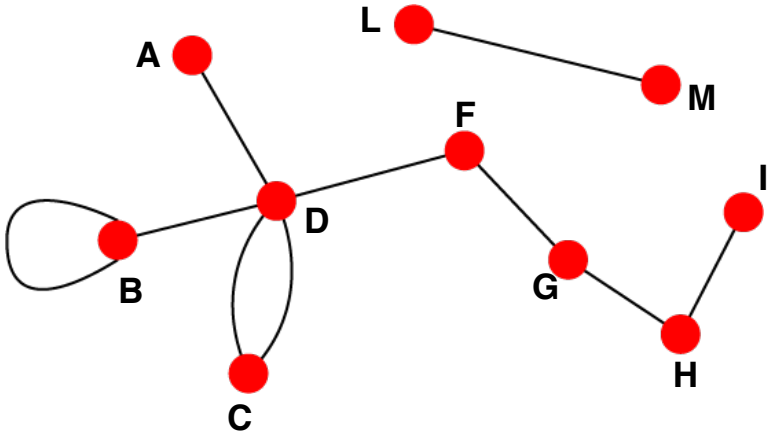


- **components**: nodes, vertices N
- **interactions**: links, edges L
- **system**: network, graph (N,L)

Undirected

Links: undirected (*symmetrical*)

Graph:

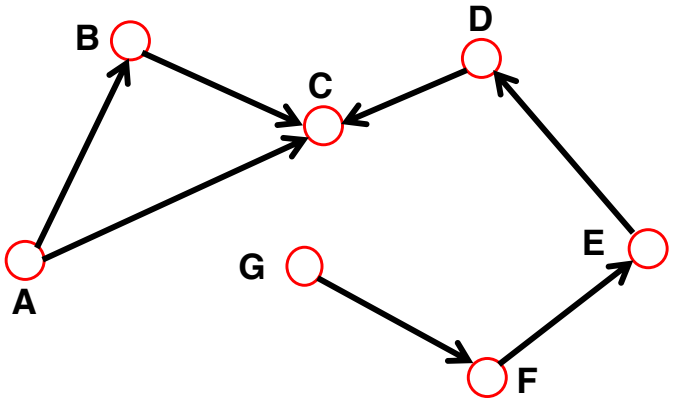


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

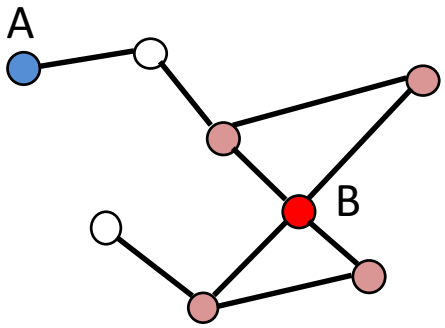
Digraph = directed graph:



An undirected link is the superposition of two opposite directed links.

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L
Internet	Routers	Internet connections	Undirected	192,244	609,066
WWW	Webpages	Links	Directed	325,729	1,497,134
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826
Email	Email addresses	Emails	Directed	57,194	103,731
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908
Citation Network	Paper	Citations	Directed	449,673	4,689,479
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930

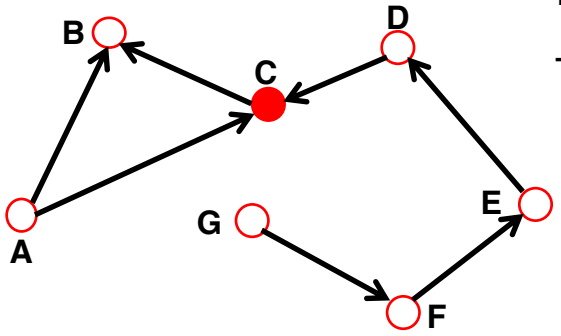
Undirected



Node degree: the number of links connected to the node.

$$k_A = 1 \qquad k_B = 4$$

Directed



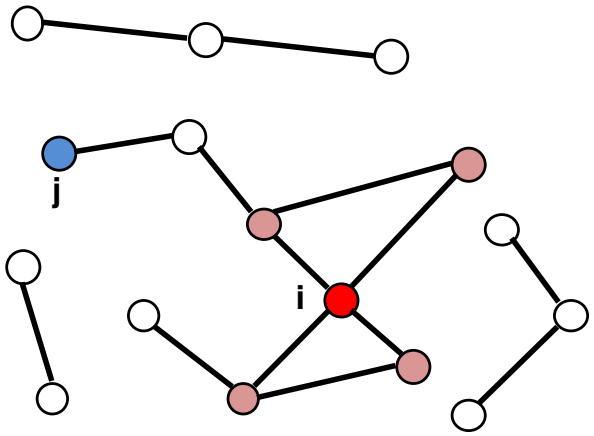
In *directed networks* we can define an in-degree and out-degree.

The (total) degree is the sum of in- and out-degree.

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

Source: a node with $k^{in} = 0$; Sink: a node with $k^{out} = 0$.

Undirected

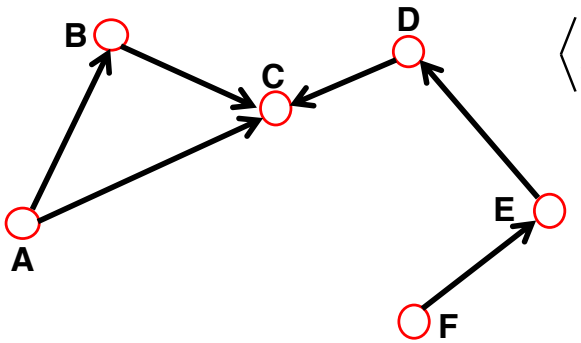


Here the 1/2 factor corrects for the fact that in the sum (2.1) each link is counted twice.

$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$

N – the number of nodes in the graph

Directed



$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

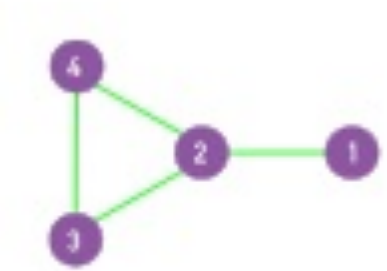
$$\langle k \rangle \equiv \frac{L}{N}$$

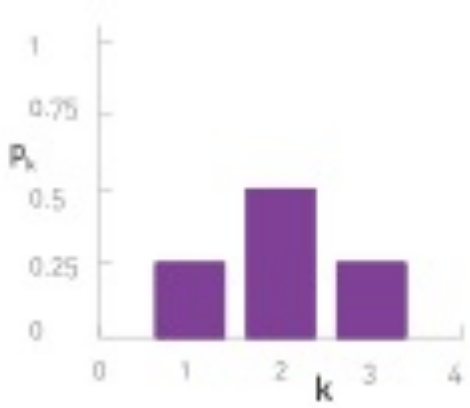
Average degree

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

Degree distribution

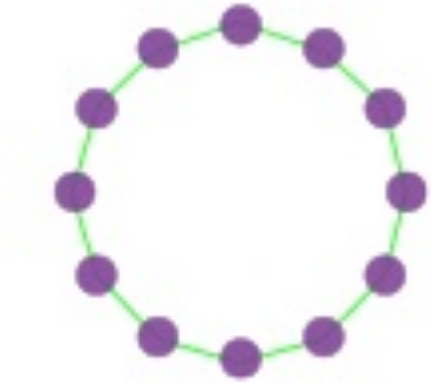
P(k): probability that a randomly chosen node has degree *k*

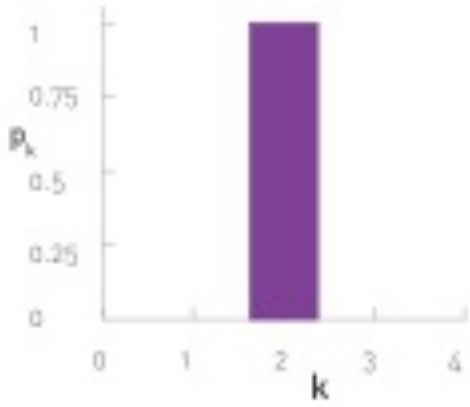




N_k = # nodes with degree k

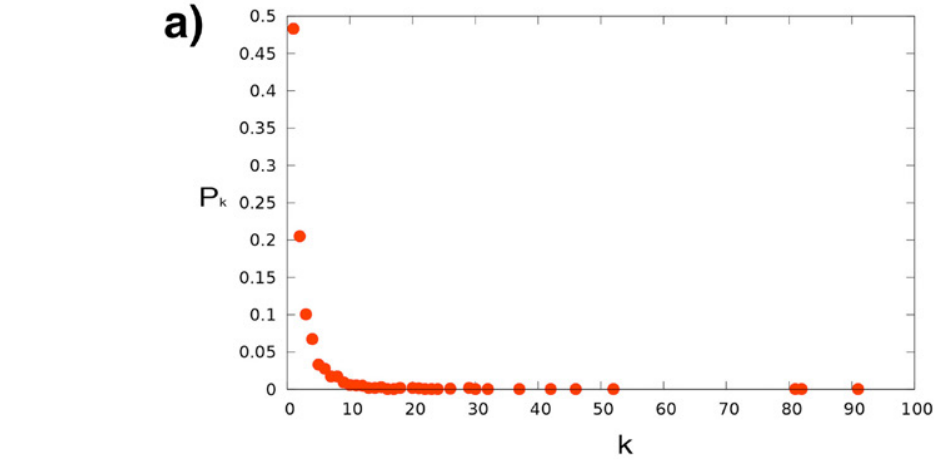
$P(k) = N_k / N$ ➔ plot

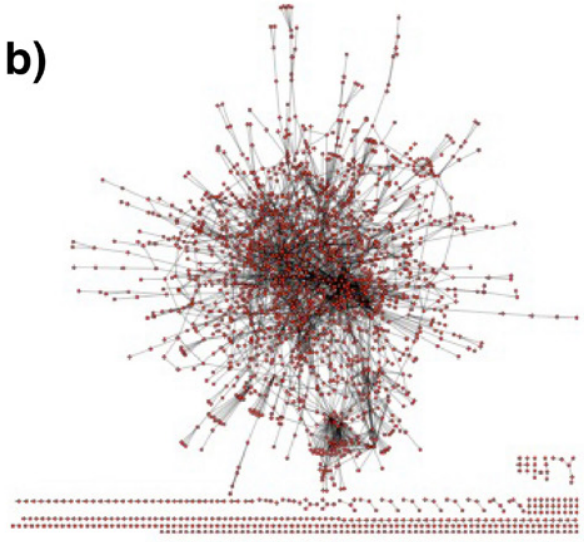


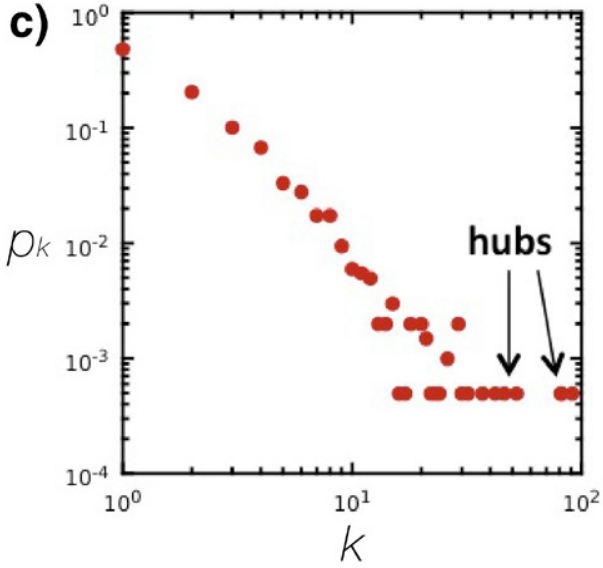


Degree distribution

Degree distribution indicates, the degrees of the proteins in the protein interaction network shown in (b) vary between *k=0* (isolated nodes) and *k=92*, which is the degree of the largest node, called a **hub**.

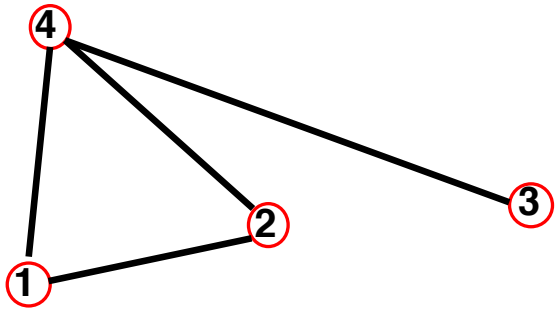


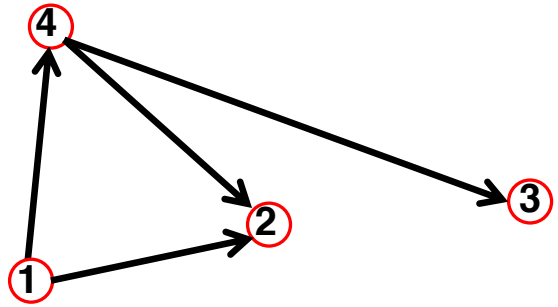




Log-log plot

There are also wide differences in the number of nodes with different degrees: as (a) shows, almost half of the nodes have degree one (i.e. *p1*=0.48), while there is only one copy of the biggest node, hence *p92* = 1/ *N*=0.0005.





Undirect

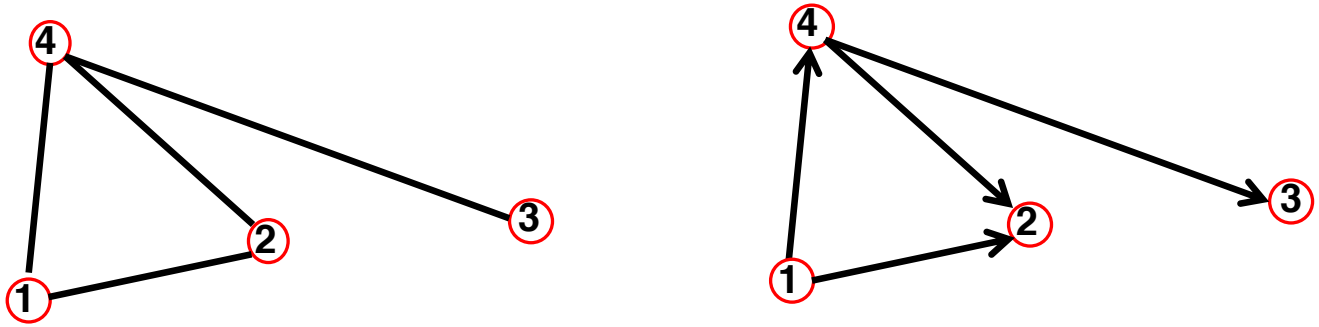
A_{ij}=1 if there is a link between node *i* and *j*

A_{ij}=0 if nodes *i* and *j* are not connected to each other.

Direct

$A_{ij} = 1$ if there is a link pointing from node *j* and *i*

$A_{ij} = 0$ if there is no link pointing from *j* to *i*.



Undirect

A_{ij}=1 if there is a link between node *i* and *j*

A_{ij}=0 if nodes *i* and *j* are not connected to each other.

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

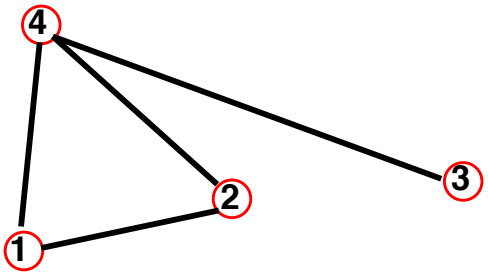
Note that for a directed graph (right) the matrix is not symmetric.

Direct

$A_{ij} = 1$ if there is a link pointing from node *j* and *i*

$A_{ij} = 0$ if there is no link pointing from *j* to *i*.

Undirected



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$k_i = \sum_{j=1}^{N} A_{ij}$$

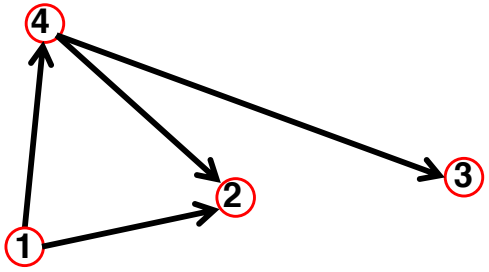
$$k_j = \sum_{i=1}^{N} A_{ij}$$

$$A_{ij} = A_{ji}$$

$$A_{ii} = 0$$

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i = \frac{1}{2}\sum_{ij}^{N} A_{ij}$$

Directed



$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$k_i^{in} = \sum_{j=1}^{N} A_{ij}$$

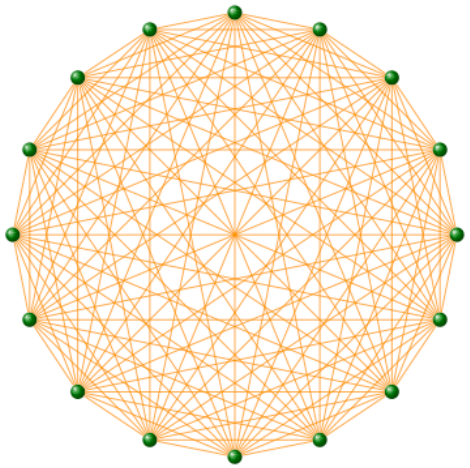
$$k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

$$A_{ij} \neq A_{ji}$$

$$A_{ii} = 0$$

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_j^{out} = \sum_{i,j}^{N} A_{ij}$$

Complete Graph



The maximum number of links a network of N nodes can have is: $L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$

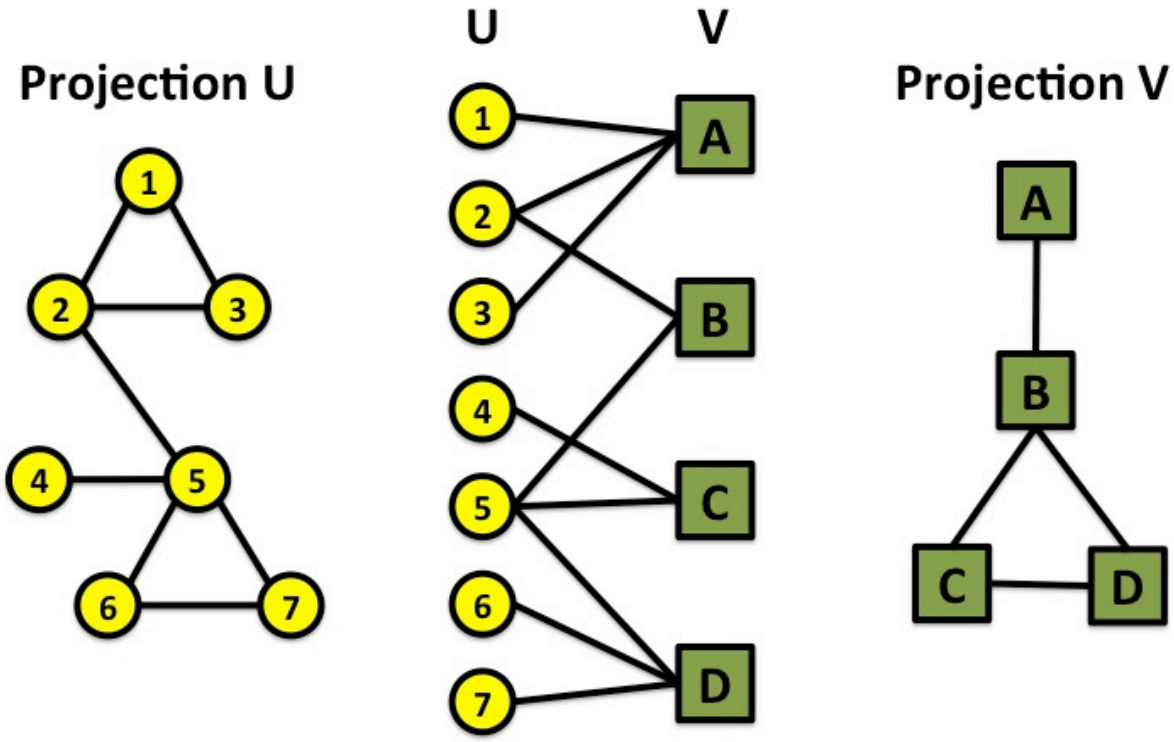
A graph with degree $L=L_{max}$ is called a complete graph, and its average degree is **$\langle k \rangle = N-1$**

Most networks observed in real systems are sparse:

$L \ll L_{max}$
or
$\langle k \rangle \ll N-1$.

Protein (*S. Cerevisiae*): N= 1,870; L=4,470 $L_{max}=10^7$ $\langle k \rangle=2.39$
Coauthorship (Math): N= 70,975; L=2 10^5 $L_{max}=3\ 10^{10}$ $\langle k \rangle=3.9$

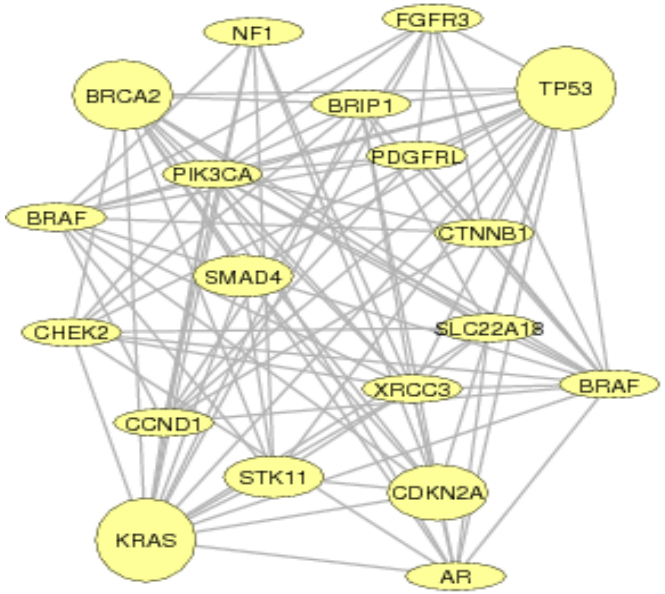
bipartite graph (or **bigraph**) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets.



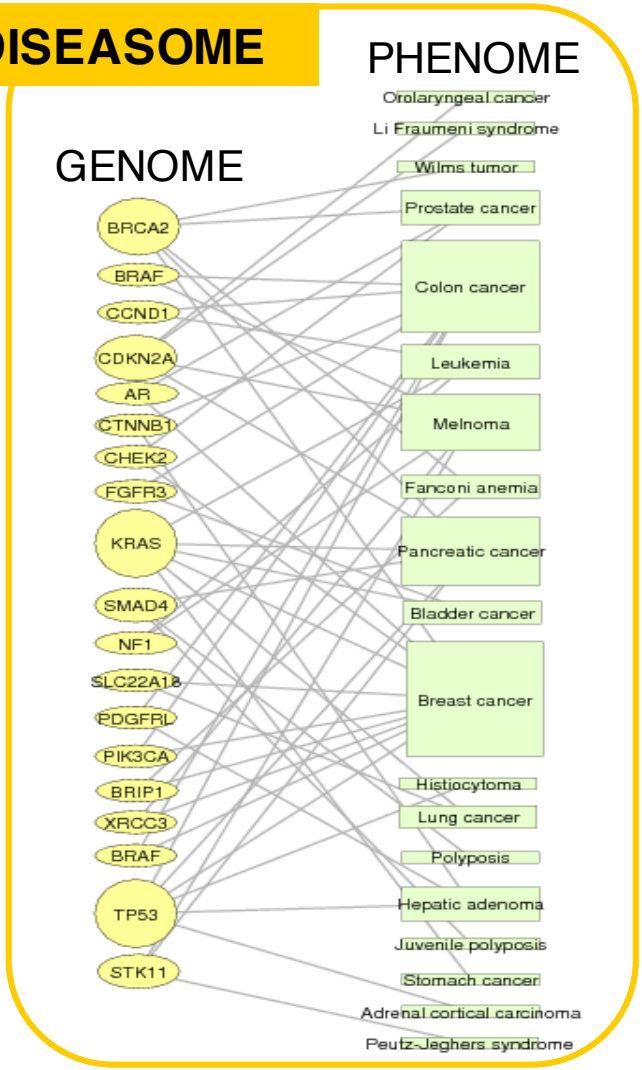
Examples:

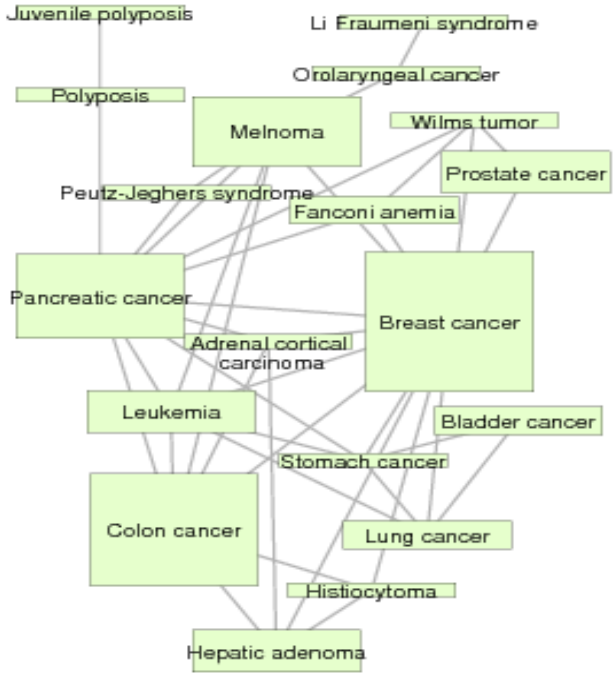
Hollywood actor network
Collaboration networks
Disease network (diseasome)

Bipartite Graph



Gene network

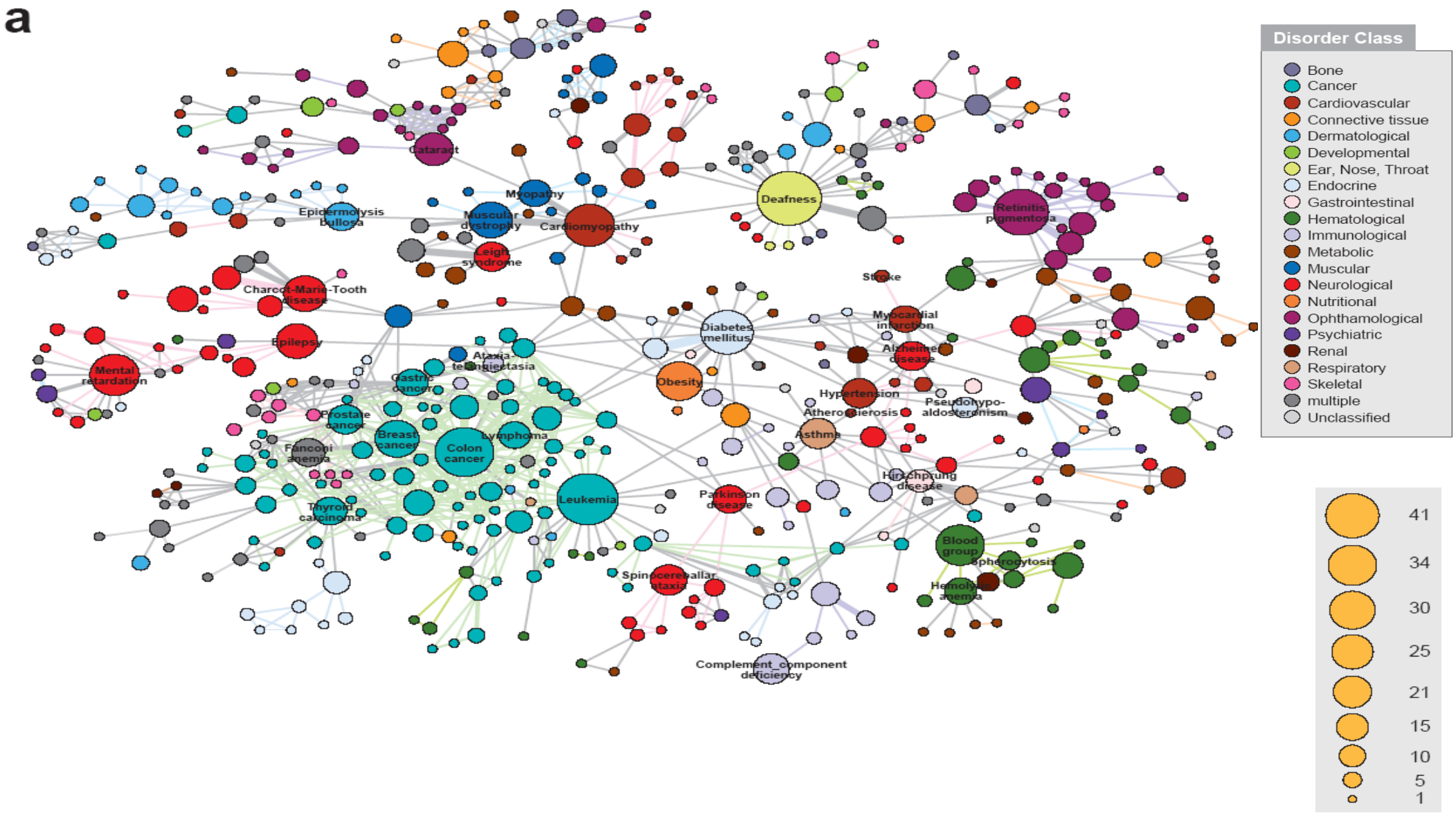




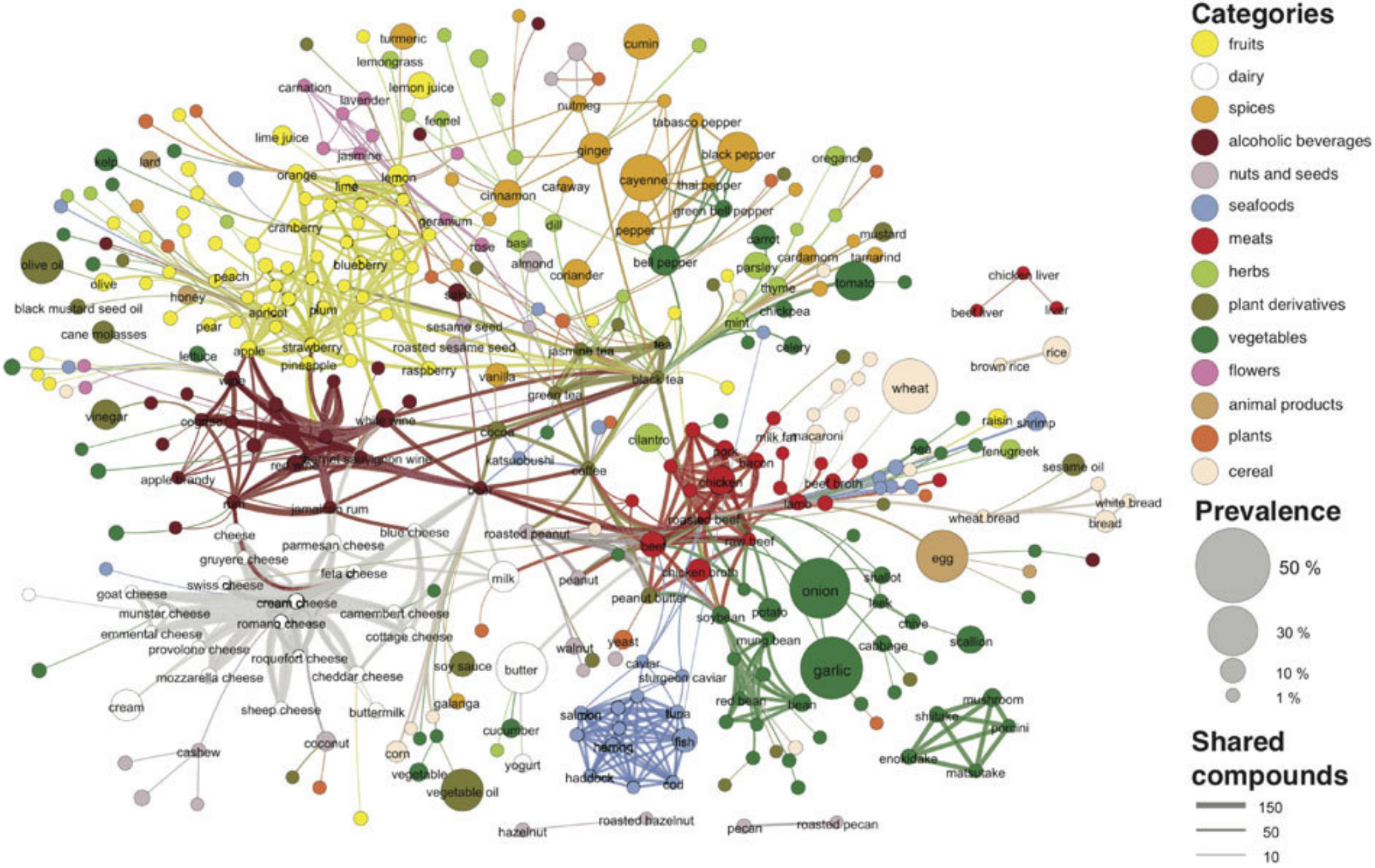
Disease network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

Human Disease Network

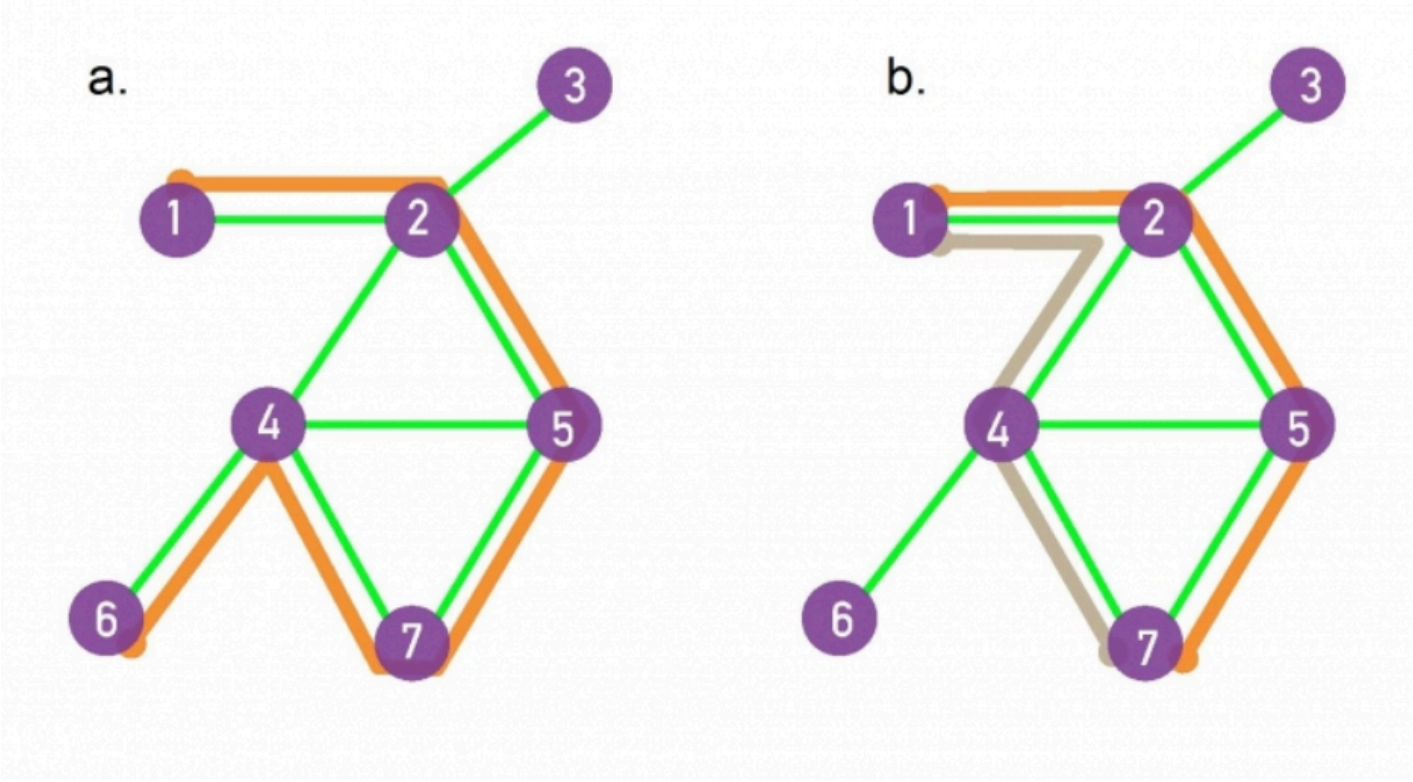


In many applications we need to study weighted networks, where each link (i, j) has a unique weight wij



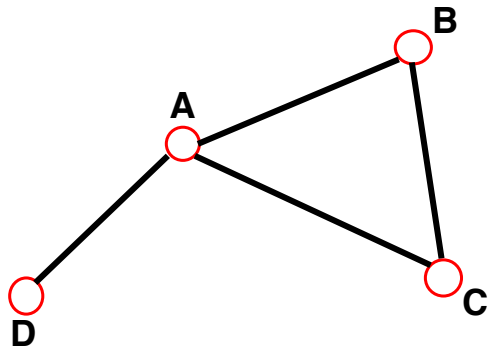
Paths

In networks physical distance is replaced by path length. A path is a route that runs along the links of the network. A path's length represents the number of links the path contains



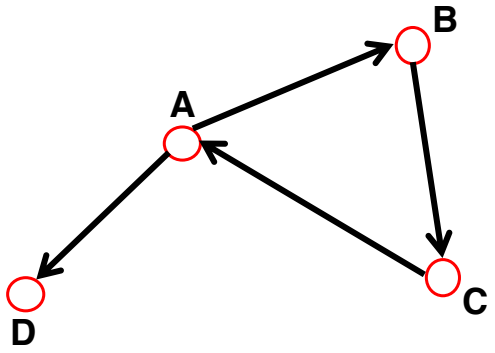
•. A path between nodes i0 and in is an ordered list of n links P = {(i0, i1), (i1, i2), (i2, i3), ... ,(in-1, in)}. The length of this path is n. The path shown in orange in (a) follows the route 1→2→5→7→4→6, hence its length is n = 5.

•. The **shortest paths** between nodes 1 and 7, or the distance d17, correspond to the path with the fewest number of links that connect nodes 1 to 7. There can be multiple paths of the same length, as illustrated by the two paths shown in orange and grey. The network diameter is the largest distance in the network, being dmax = 3 here.



The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them.

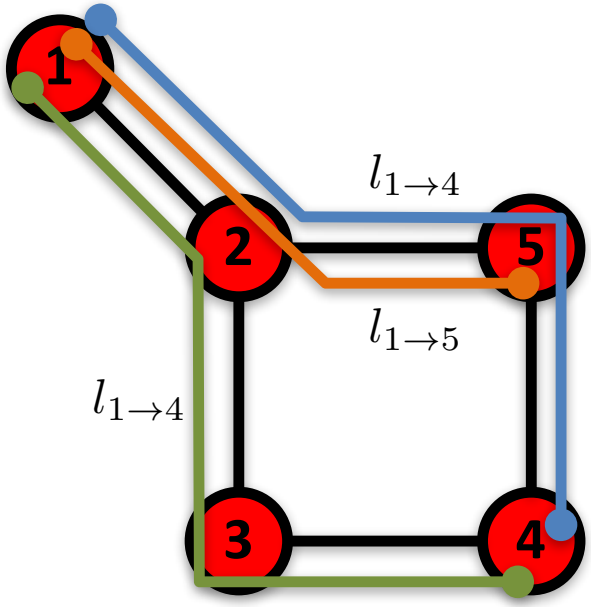
*If the two nodes are disconnected, the distance is infinity.



In directed graphs each path needs to follow the direction of the arrows.

Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

Shortest Path

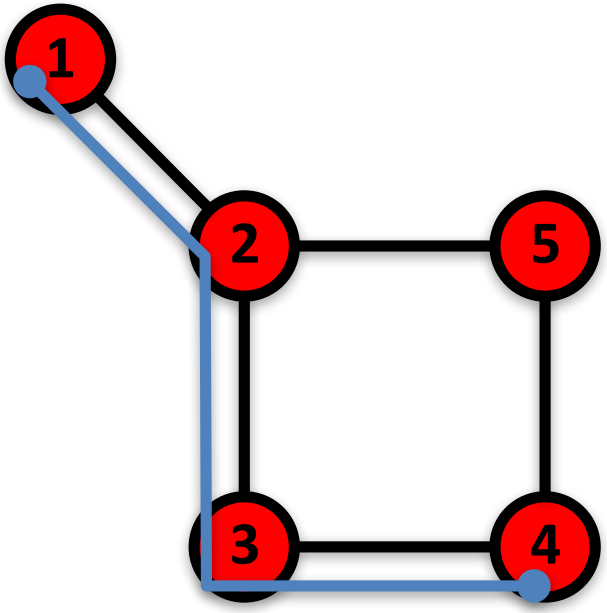


$l_{1\to4} = 3$

$l_{1\to5} = 2$

The path with the shortest length between two nodes (distance).

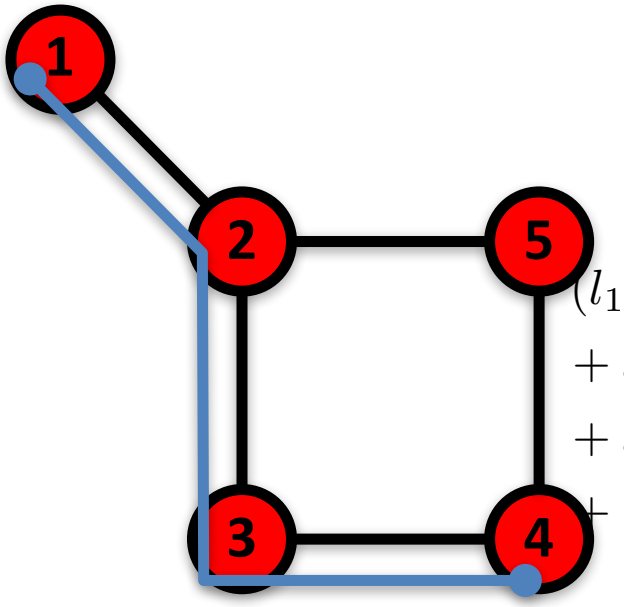
Diameter



$l_{1\to4} = 3$

The longest shortest path in a graph

Average Path Length



$$(l_{1\to2} + l_{1\to3} + l_{1\to4} + l_{1\to5} + l_{2\to3} + l_{2\to4} + l_{2\to5} + l_{3\to4} + l_{3\to5} + l_{4\to5})/10 = 1.6$$

The average of the shortest paths for all pairs of nodes.

Average path length/distance, <d>, for a connected graph:

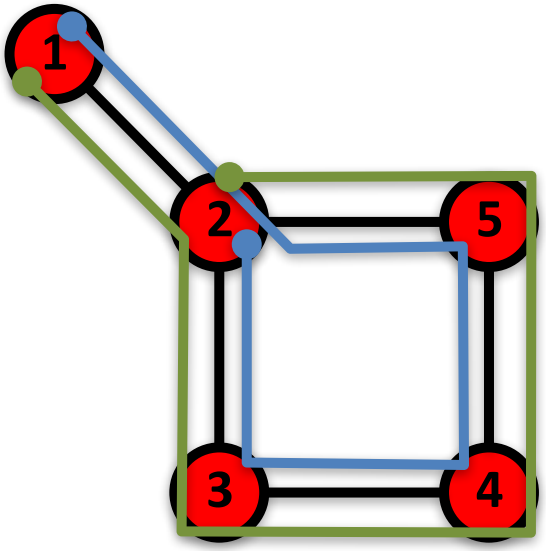
$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j\neq i} d_{ij}$$

where d_{ij} is the distance from node *i* to node j

In an *undirected graph* $d_{ij} = d_{ji}$, *so* we only need to count them once:

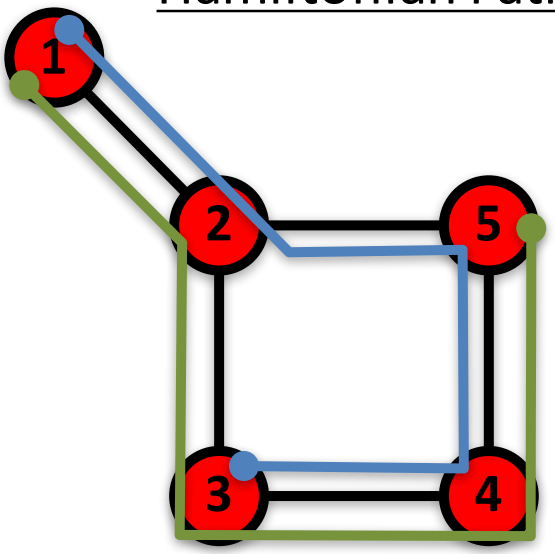
$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i,j>i} d_{ij}$$

Eulerian Path



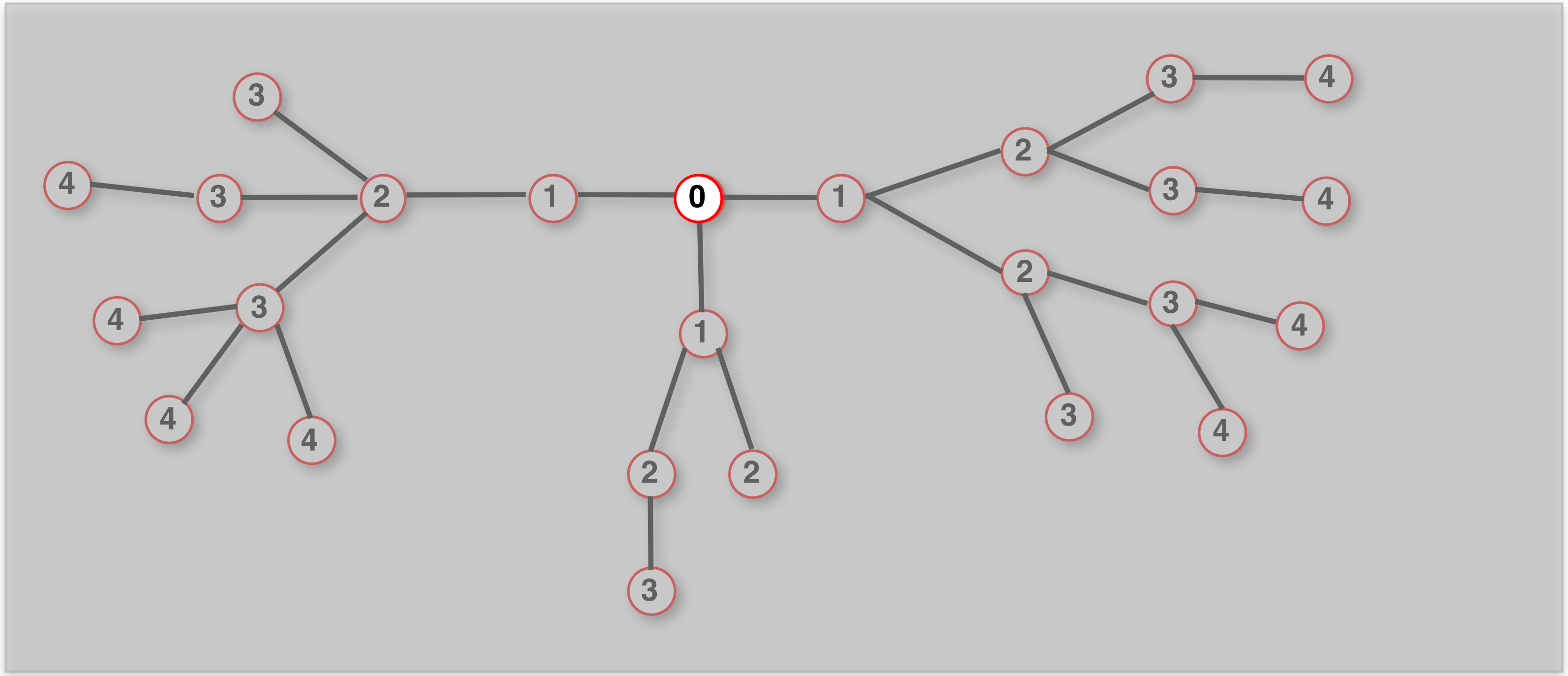
A path that traverses each link exactly once.

Hamiltonian Path

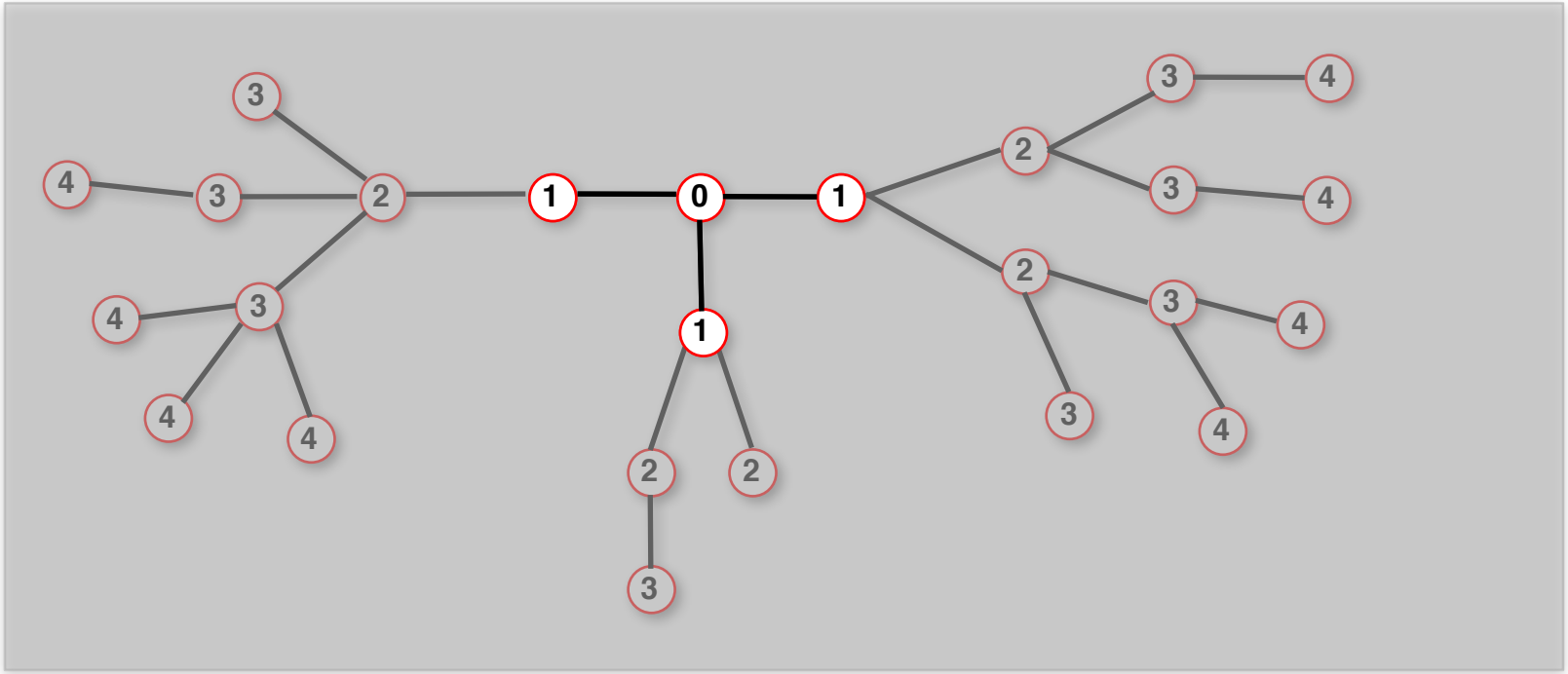


A path that visits each node exactly once.

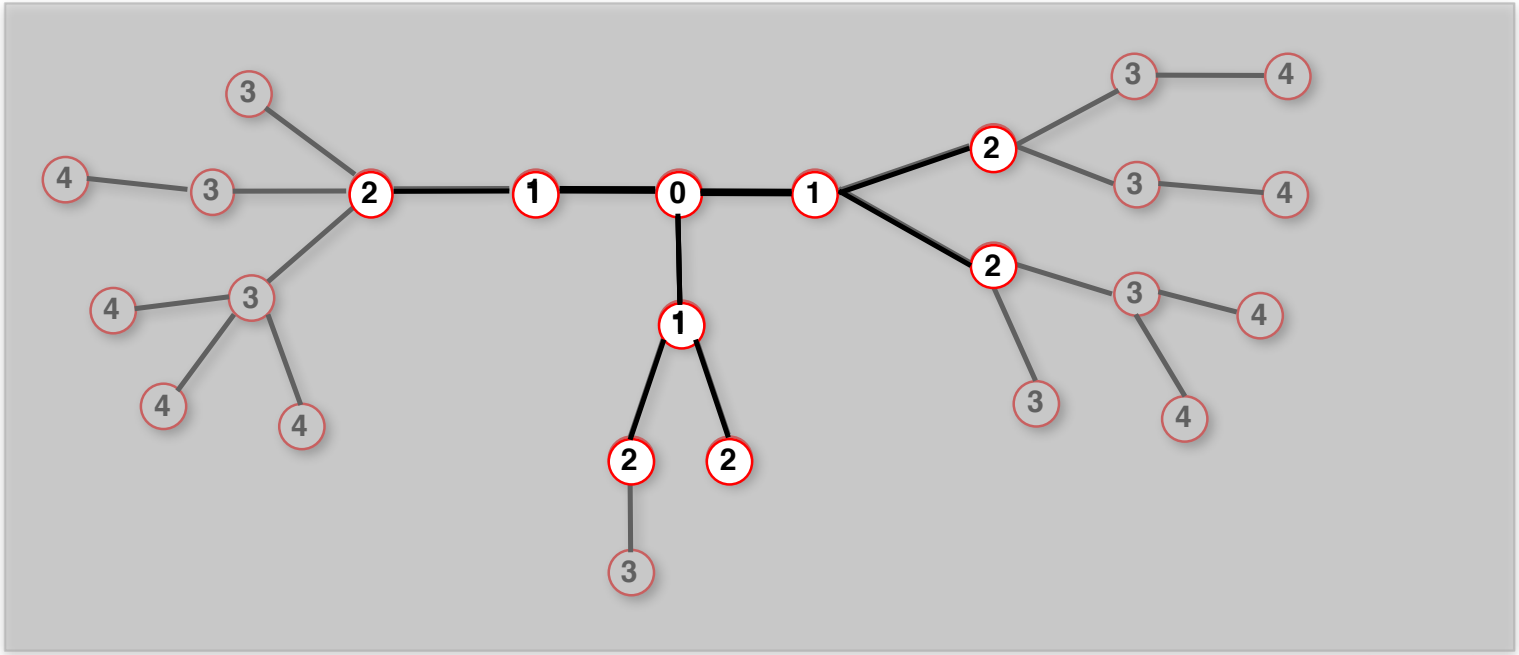
- Start at node i, that we label with "0".
- Find the nodes directly linked to i. Label them distance "1" and put them in a queue.
- Take the first node, labeled n, out of the queue ($n = 1$ in the first step). Find the unlabeled nodes adjacent to it in the graph. Label them with $n + 1$ and put them in the queue.
- Repeat step 3 until you find the target node j or there are no more nodes in the queue.
- The distance between i and j is the label of j. If j does not have a label, then $d_{ij} = \infty$.



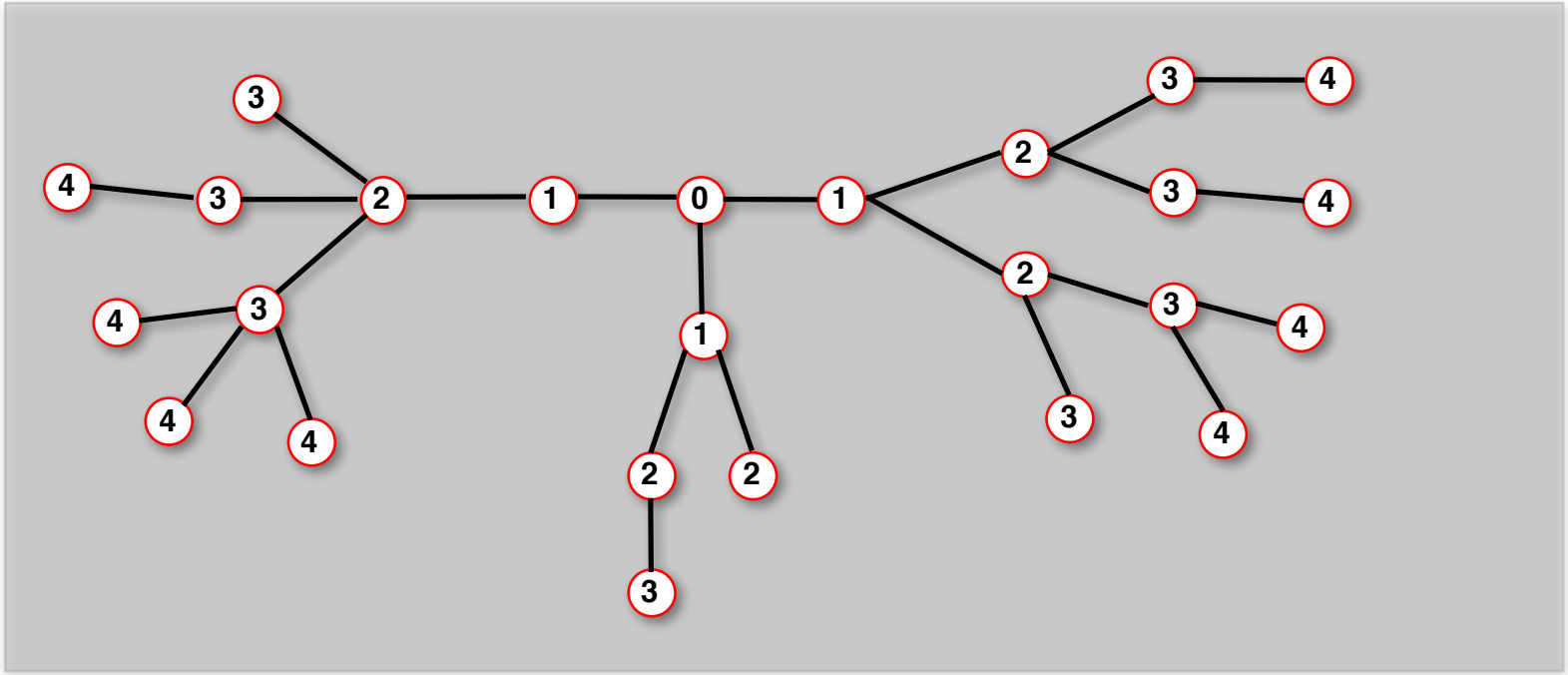
- Start at node i, that we label with "0".
- Find the nodes directly linked to i. Label them distance "1" and put them in a queue.
- Take the first node, labeled n, out of the queue (n = 1 in the first step). Find the unlabeled nodes adjacent to it in the graph. Label them with $n + 1$ and put them in the queue.
- Repeat step 3 until you find the target node j or there are no more nodes in the queue.
- The distance between i and j is the label of j. If j does not have a label, then $d_{ij} = \infty$.



- Start at node i, that we label with "0".
- Find the nodes directly linked to i. Label them distance "1" and put them in a queue.
- Take the first node, labeled n, out of the queue (n = 1 in the first step). Find the unlabeled nodes adjacent to it in the graph. Label them with n + 1 and put them in the queue.
- Repeat step 3 until you find the target node j or there are no more nodes in the queue.
- The distance between i and j is the label of j. If j does not have a label, then $d_{ij} = \infty$.



- Start at node i, that we label with "0".
- Find the nodes directly linked to i. Label them distance "1" and put them in a queue.
- Take the first node, labeled n, out of the queue (n = 1 in the first step). Find the unlabeled nodes adjacent to it in the graph. Label them with n + 1 and put them in the queue.
- Repeat step 3 until you find the target node j or there are no more nodes in the queue.
- The distance between i and j is the label of j. If j does not have a label, then $d_{ij} = \infty$.

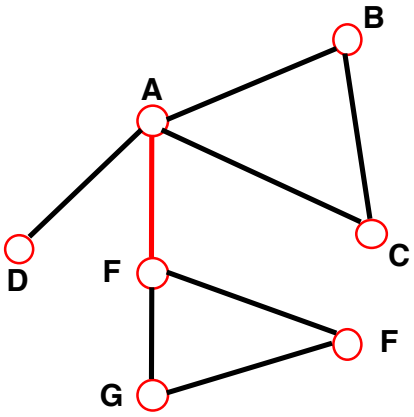


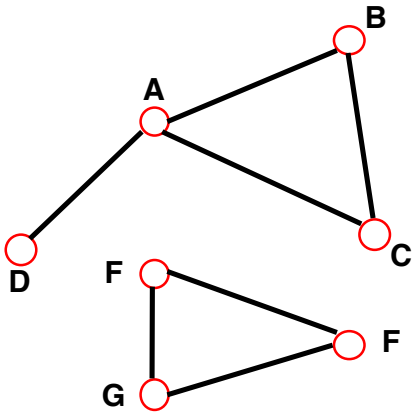
Connectivity in graphs

In an undirected network nodes i and j are connected if there is a path between them. They are disconnected if such a path does not exist, in which case we have dij = ∞.

A network is **connected** if all pairs of nodes in the network are connected.
A network is **disconnected** if there is at least one pair with dij = ∞

A **components** a subset of nodes in a network, so that there is a path between any two nodes that belong to the component, but one cannot add any more nodes to it that would have the same property.





Largest Component:
Giant Component

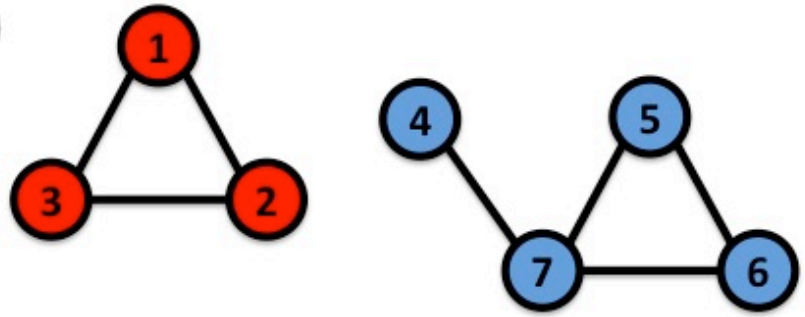
The rest: **Isolates**

Bridge: if we erase it, the graph becomes disconnected.

Disconnected graph in PN?

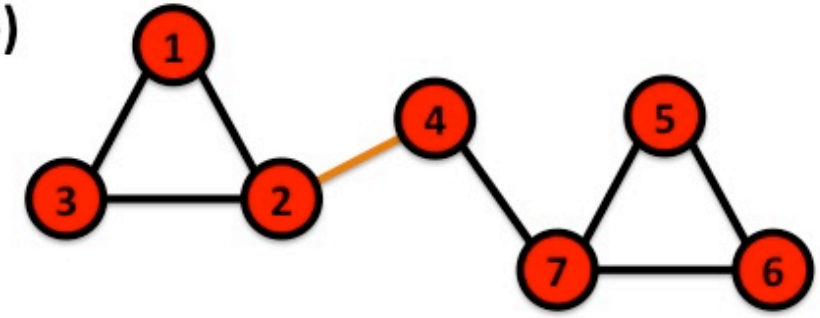
The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

(a)



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

(b)

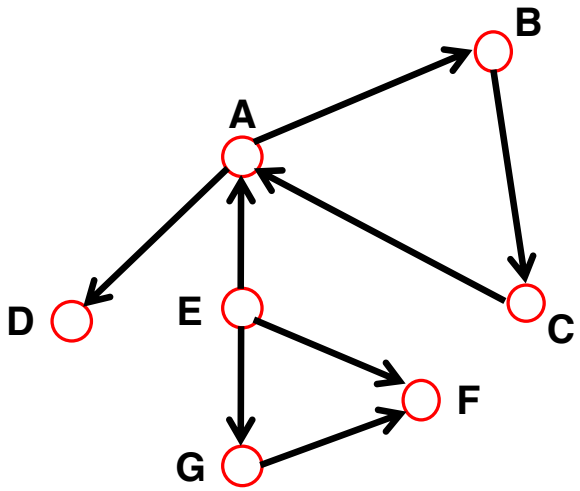


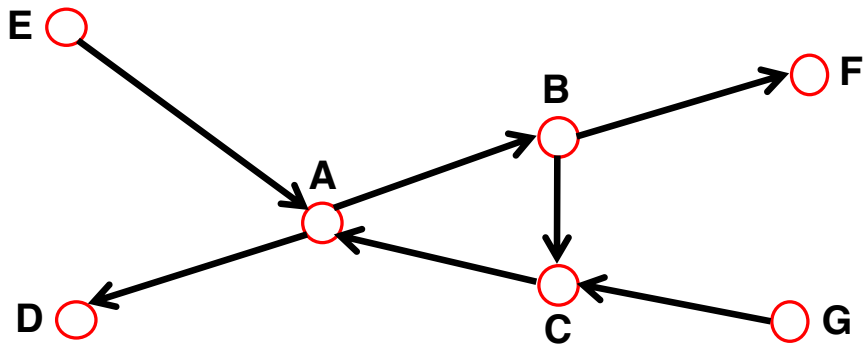
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Connectivity in graphs

Strongly connected directed graph: has a path from each node to every other node and vice versa (e.g. AB path and BA path).
Weakly connected directed graph: it is connected if we disregard the edge directions.

Strongly connected components can be identified, but not every node is part of a nontrivial strongly connected component.





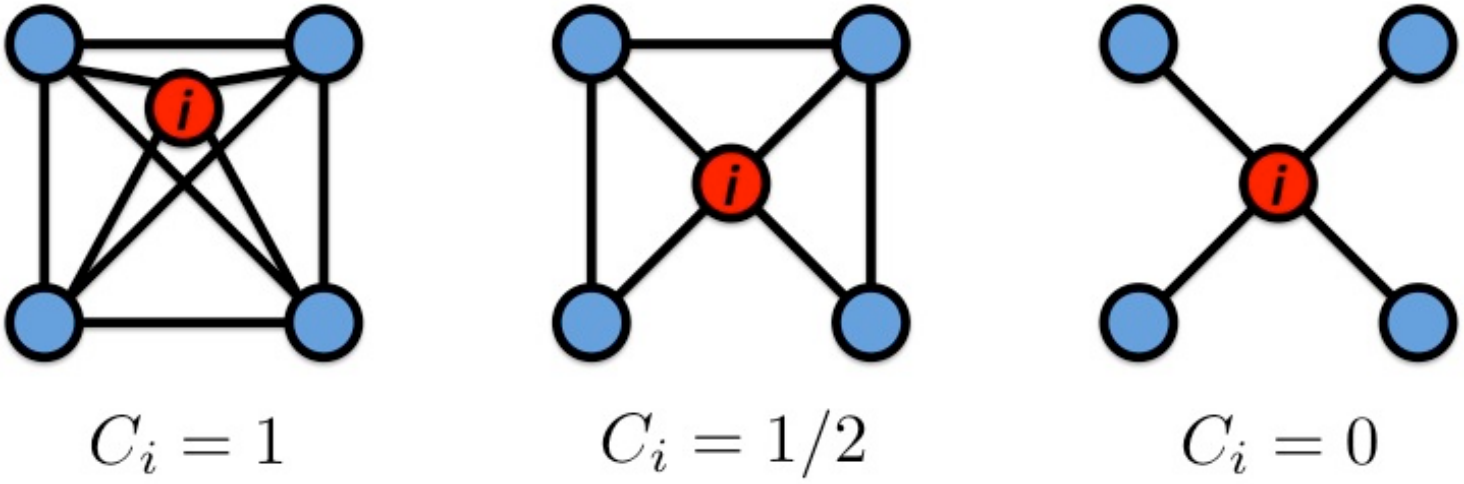
Clustering Coefficient

The clustering coefficient captures the degree to which the neighbours of a given node link to each other. For a node i with degree ki the local clustering coefficient is defined as

$$C_i = \frac{2L_i}{k_i(k_i-1)}$$

where Li represents the number of links between the ki neighbors of node i. Note that Ciis between 0 and 1:

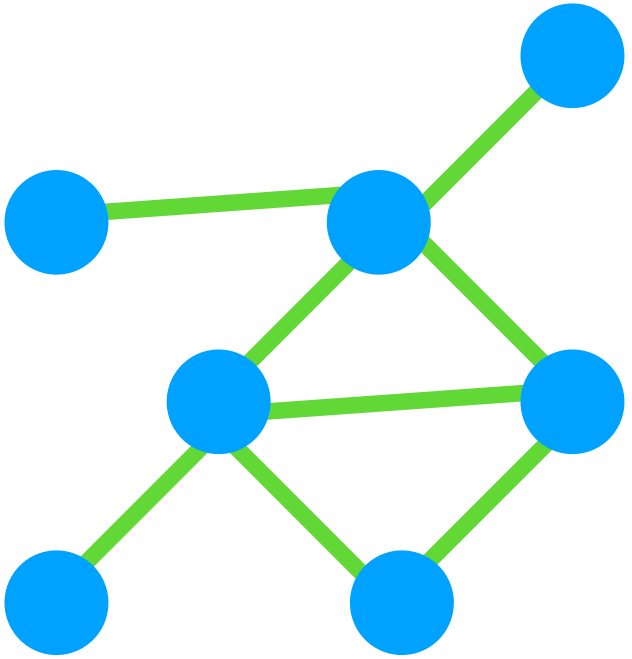
- Ci = 0 if none of the neighbors of node i link to each other.
- Ci = 1 if the neighbors of node i form a complete graph, i.e. they all link to each other.
- Ci is the probability that two neighbors of a node link to each other. Consequently C = 0.5 implies that there is a 50% chance that two neighbors of a node are linked.



Clustering Coefficient

The degree of clustering of a whole network is captured by the average clustering coefficient, 〈C〉, representing the average of Ci over all nodes i = 1, ..., N [12],

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$



In line with the probabilistic interpretation 〈C〉 is the probability that two neighbors of a randomly selected node link to each other.

$$C_{\Delta} = \frac{3 \times NumberOfTriangles}{NumberOfConnectedTriples}$$

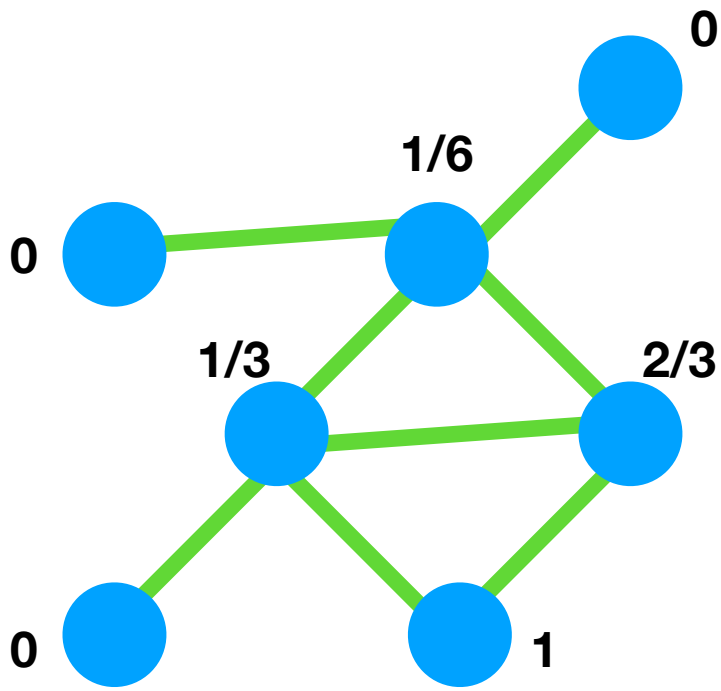
Connected triplet is an ordered set of three nodes ABC such that A connects to B and B connects to C. For example, an A, B, C triangle is made of three triplets, ABC, BCA and CAB. In contrast a chain of connected nodes A, B, C, in which B connects to A and C, but A does not link to C, forms a single open triplet ABC. The factor three in the numerator of (2.17) is due to the fact that each triangle is counted three times in the triplet count.

Clustering Coefficient

The degree of clustering of a whole network is captured by the average clustering coefficient, 〈C〉, representing the average of Ci over all nodes i = 1, ..., N [12],

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$

In line with the probabilistic interpretation 〈C〉 is the probability that two neighbors of a randomly selected node link to each other.



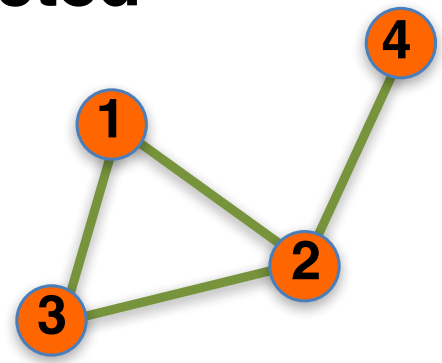
$$C_\Delta = \frac{3 \times NumberOfTriangles}{NumberOfConnectedTriples}$$

$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

$$C_\Delta = \frac{3}{8} = 0.375$$

Connected triplet is an ordered set of three nodes ABC such that A connects to B and B connects to C. For example, an A, B, C triangle is made of three triplets, ABC, BCA and CAB. In contrast a chain of connected nodes A, B, C, in which B connects to A and C, but A does not link to C, forms a single open triplet ABC. The factor three in the numerator is due to the fact that each triangle is counted three times in the triplet count.

Undirected



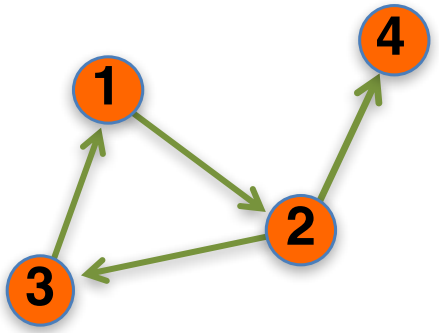
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad < k >= \frac{2L}{N}$$

Actor network, protein-protein interactions

Directed



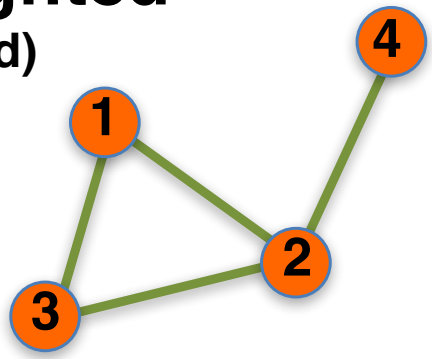
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} \neq A_{ji}$$

$$L = \sum_{i,j=1}^{N} A_{ij} \qquad < k >= \frac{L}{N}$$

WWW, citation networks

Unweighted
(undirected)



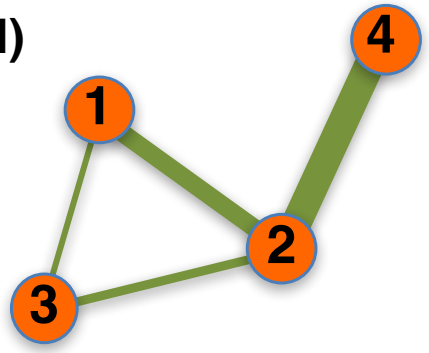
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad < k >= \frac{2L}{N}$$

protein-protein interactions, www

Weighted
(undirected)



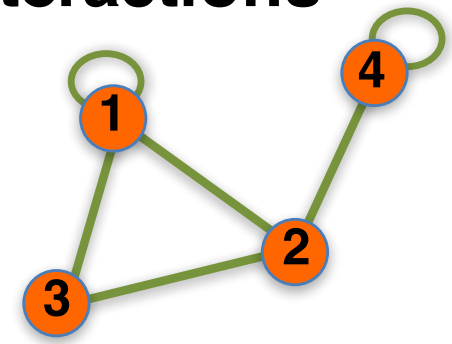
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad < k >= \frac{2L}{N}$$

Call Graph, metabolic networks

Self-interactions



$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

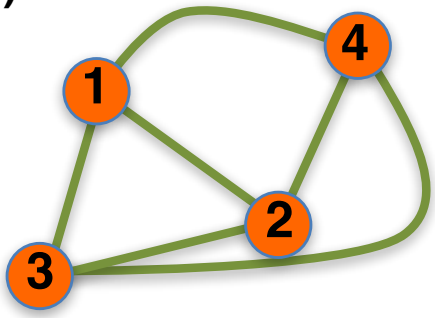
$$A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii} \qquad ?$$

Protein interaction network, www

Complete Graph

(undirected)

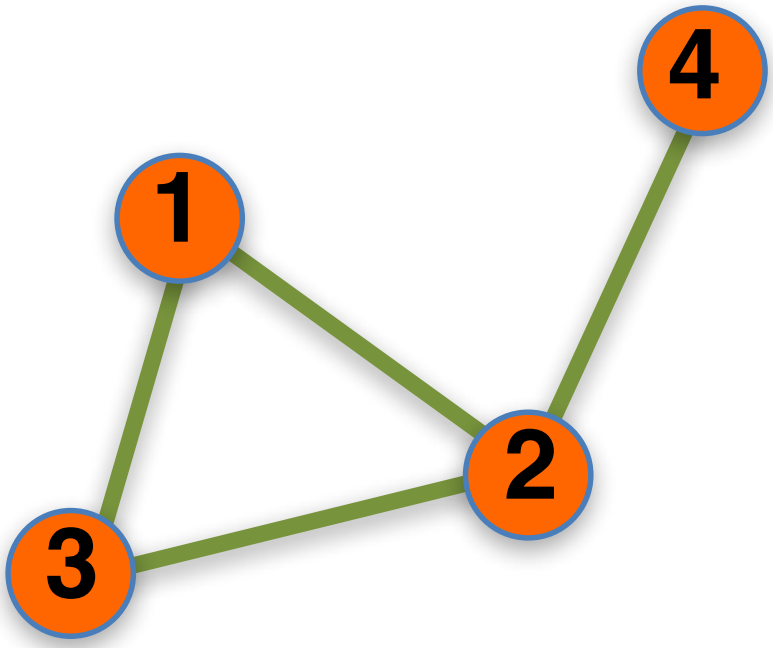


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \qquad A_{i \neq j} = 1$$

$$L = L_{\max} = \frac{N(N-1)}{2} \qquad < k >= N - 1$$

Actor network, protein-protein interactions



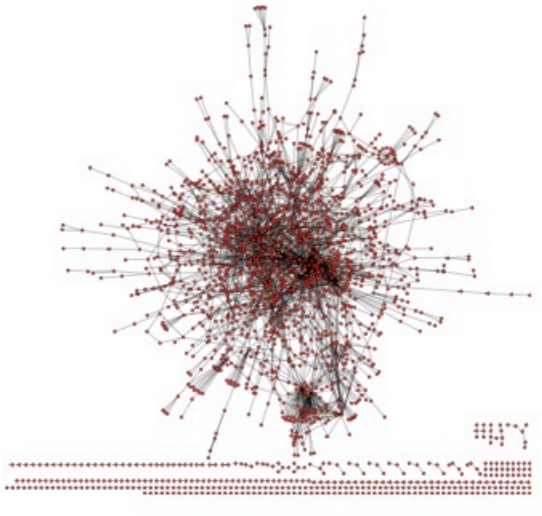
Degree distribution?

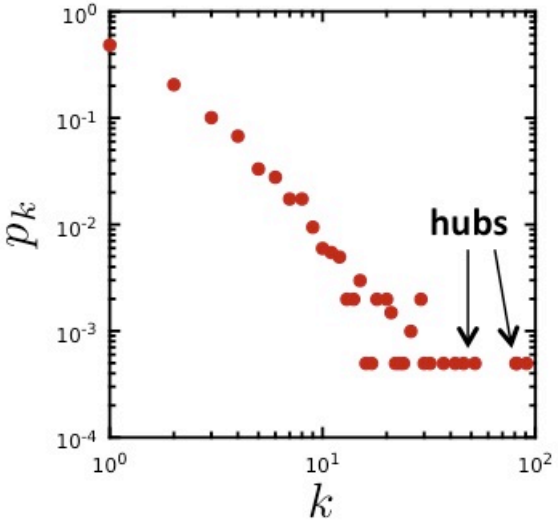
Cluster Coefficient for each node

Average path length and network diameter

Protein Protein Interaction Network

A protein-protein interaction network, where two proteins are connected if there is experimental evidence that they can bind to each other in the cell.



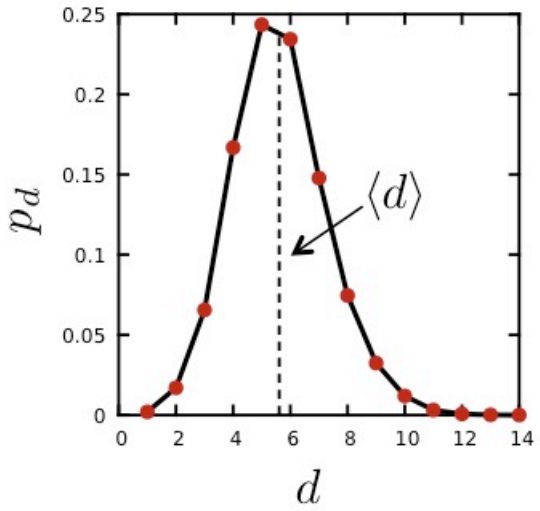


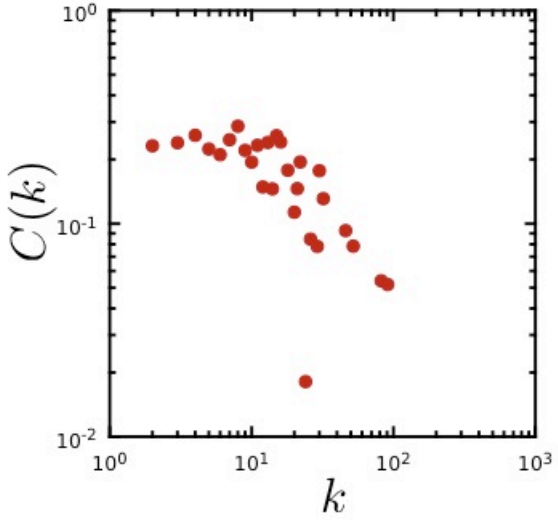
Undirected network

N=2,018 proteins as nodes

L=2,930 binding interactions as links.

Average degree <k>=2.90.



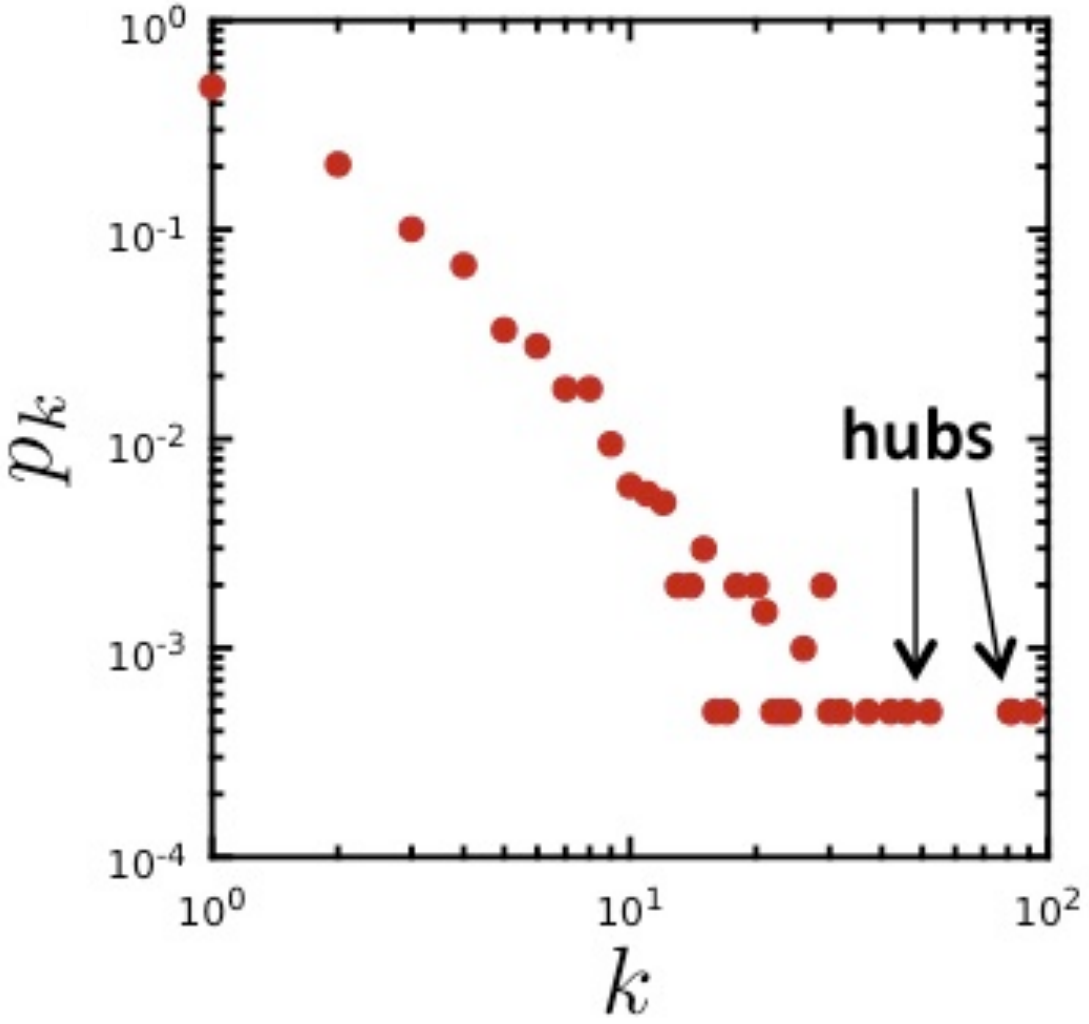


Not connected: 185 components

the largest (giant component) 1,647 nodes

Mean Cluster coefficient = 0.12

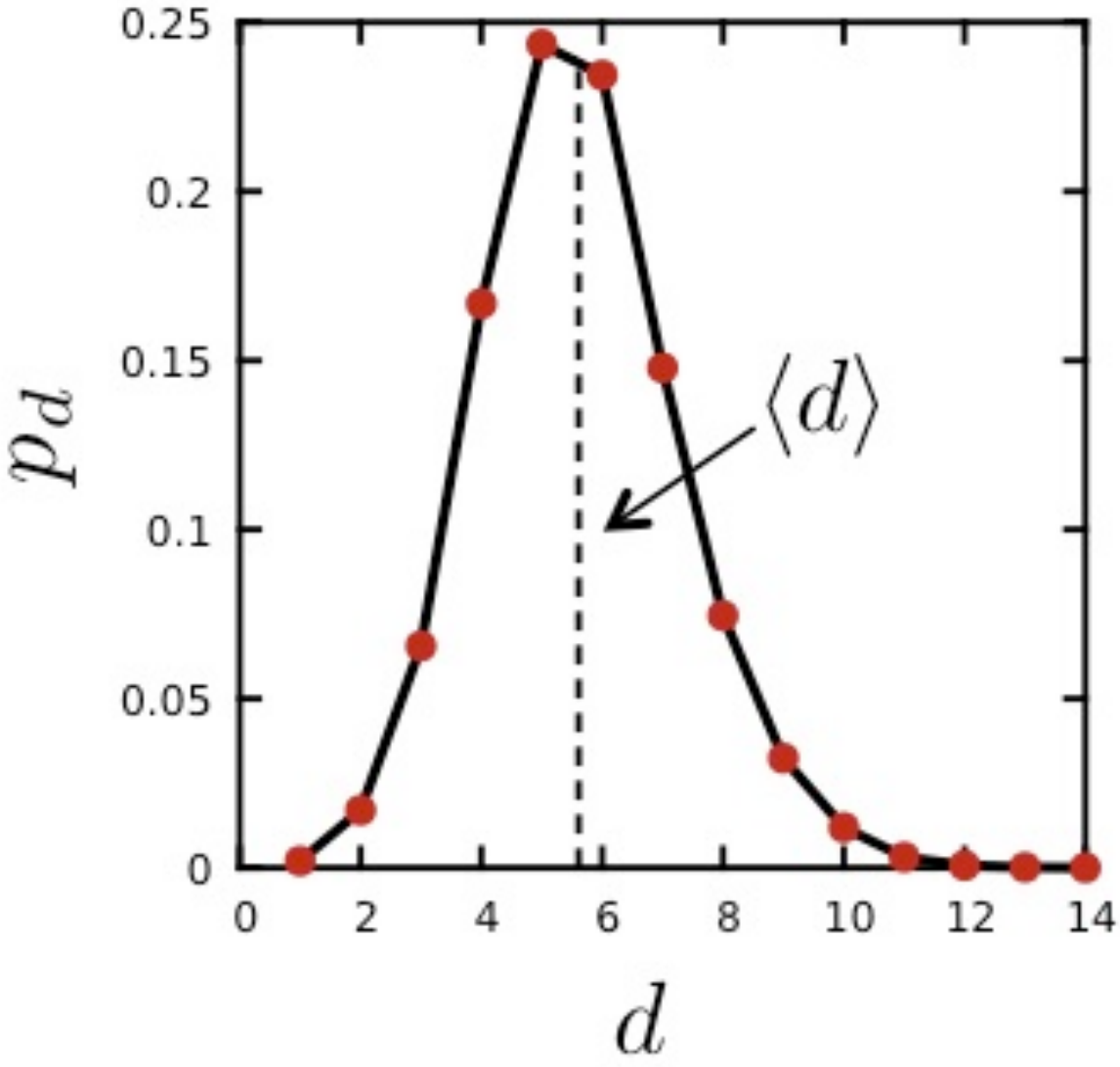
Protein Protein Interaction Network



p_k is the probability that a node has degree **k**.

N_k = # nodes with degree k

$p_k = N_k / N$



d_{max}=14

<d>=5.61

Most networks we encounter do not have the comforting regularity of a crystal lattice or the predictable radial architecture of a spider web, but it is not always constructed by regularity.

Defining Random Networks

There are **two definitions** of a random network:

- **G(N, L) Model:** N labeled nodes are connected with L randomly placed links. Erdős and Rényi used this definition in their string of papers on random networks [2-9]
- **G(N, p) Model:** Each pair of N labeled nodes is connected with probability p, a model introduced by Gilbert.

Hence, the **G(N, p) model fixes the probability p that two nodes are connected** and the **G(N, L)) model fixes the total number of links L**. While in the G(N, L) model the average degree of a node is simply ‹k› = 2L/N, other network characteristics are easier to calculate in the G(N, p) model. The G(N, p) model, is ease to calculate key network characteristics, and in real networks the number of links rarely stays fixed.

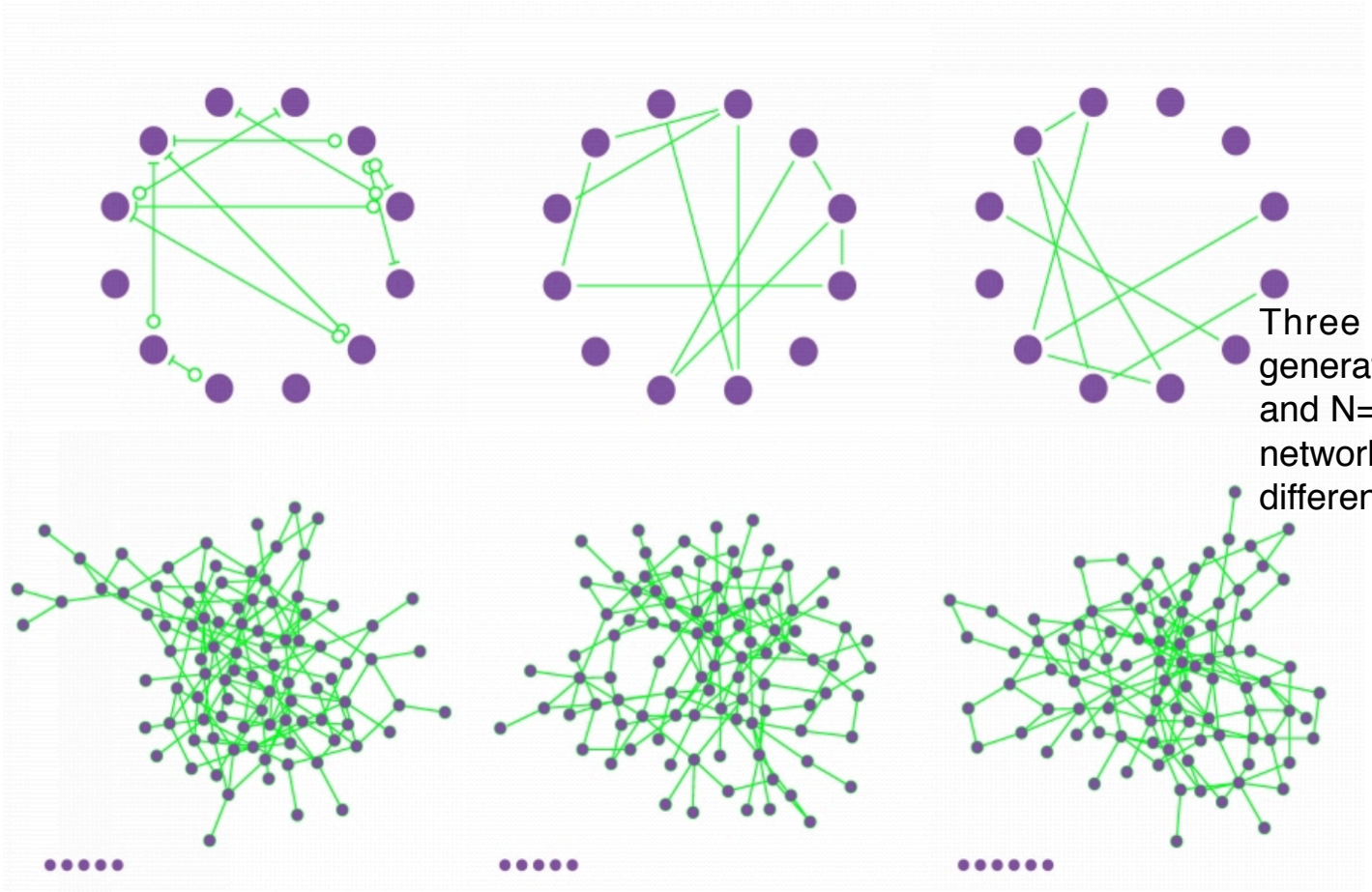
To construct a random network we follow these steps:

- Start with N isolated nodes.
- Select a node pair and generate a random number between 0 and 1. If the number exceeds p, connect the selected node pair with a link, otherwise leave them disconnected.
- Repeat step (2) for each of the N(N-1)/2 node pairs.

In summary the number of links in a random network varies between realizations. Its expected value is determined by N and p. If we increase p a random network becomes denser: The average number of links increase linearly from ‹L› = 0 to Lmax and the average degree of a node increases from ‹k› = 0 to ‹k› = N-1.

Random Network

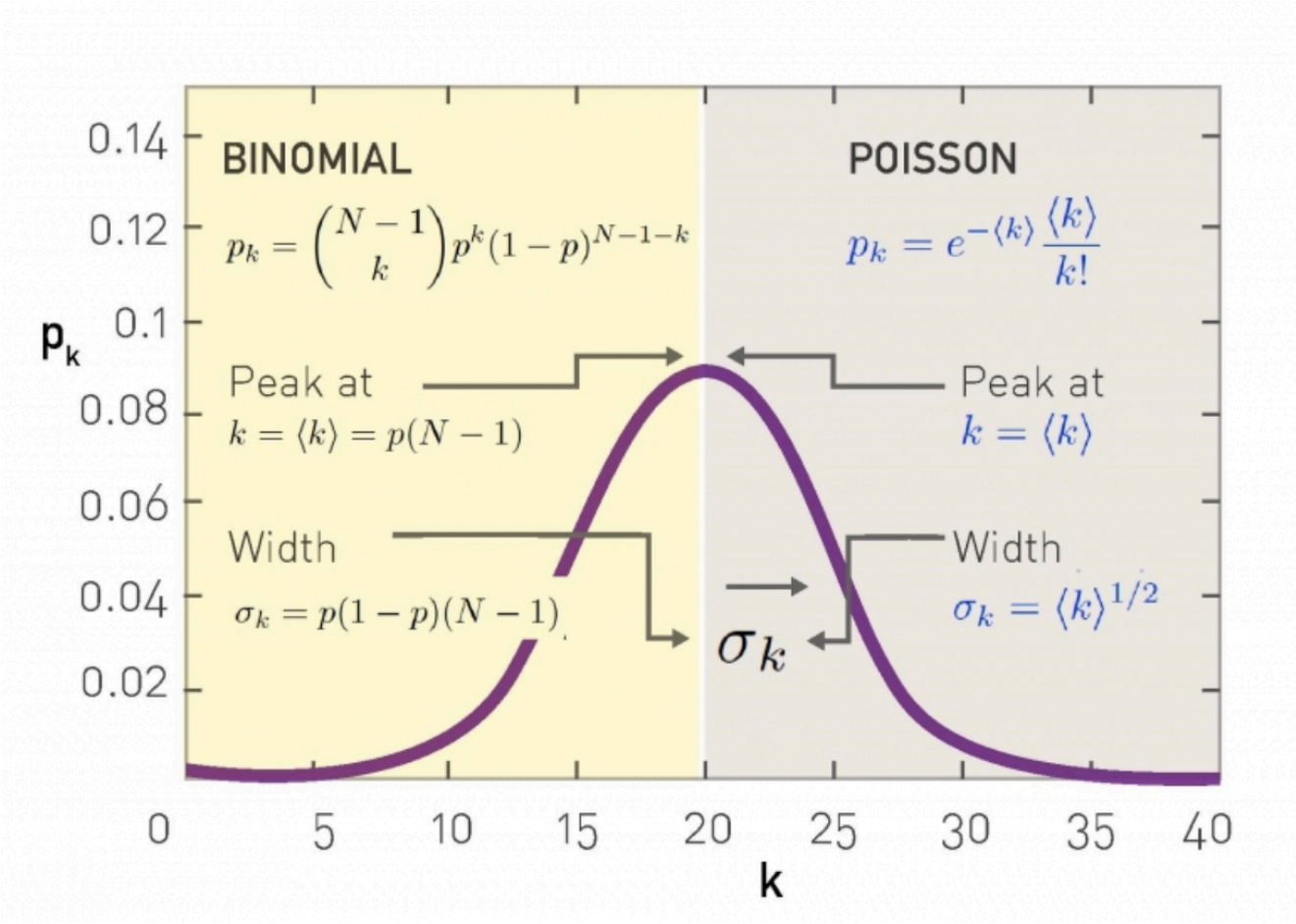
In a given realization of a random network some nodes gain numerous links, while others acquire only a few or no links. These differences are captured by the degree distribution, pk, which is the probability that a randomly chosen node has degree k.



Three realizations of a random network generated with the same parameters p=1/6 and N=12. Despite the identical parameters, the networks not only look different, but they have a different number of links as well (L=10, 10, 8).

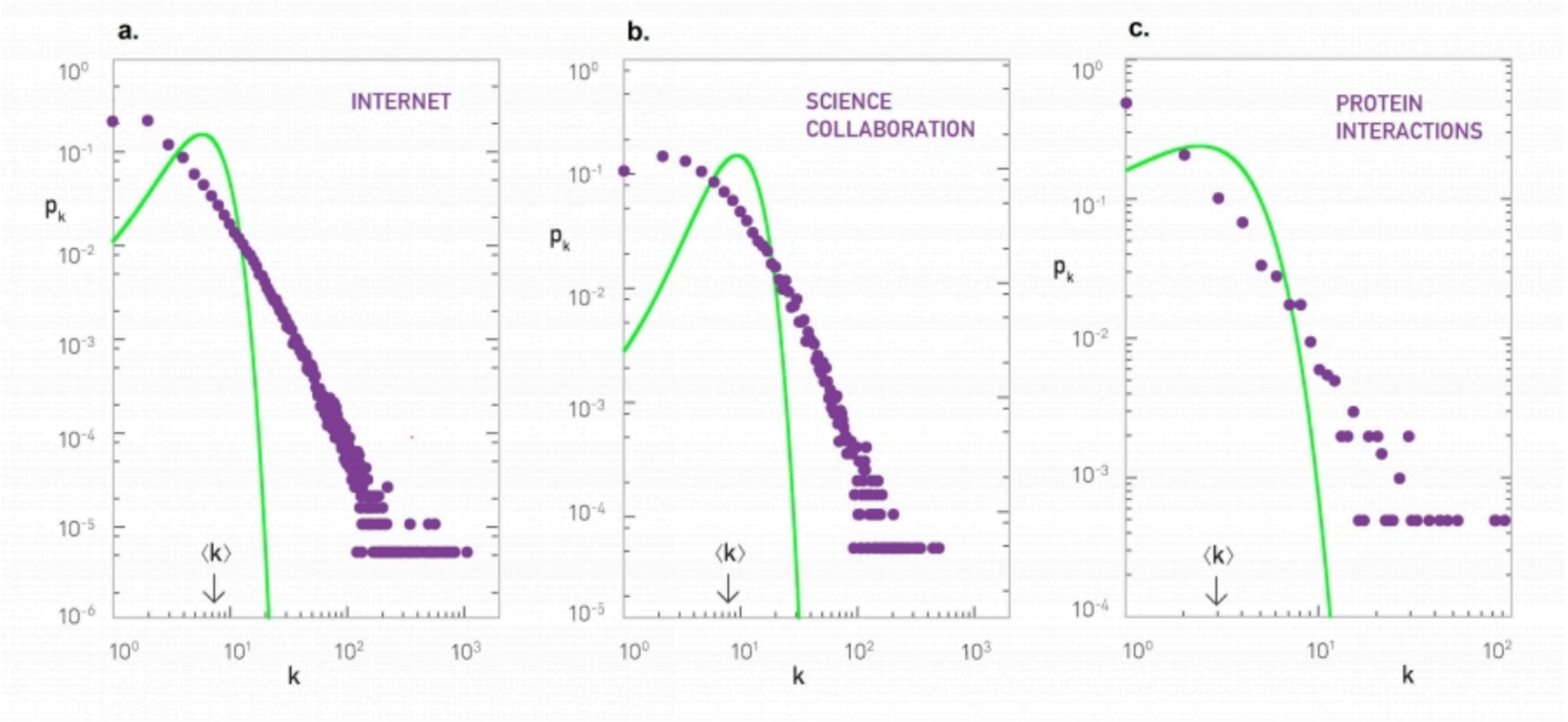
Three realizations of a random network with p=0.03 and N=100. Several nodes have degree k=0, shown as isolated nodes at the bottom.

In a given realization of a random network some nodes gain numerous links, while others acquire only a few or no links. These differences are captured by the degree distribution, pk, which is the probability that a randomly chosen node has degree k.



The exact form of the **degree distribution of a random network is the binomial distribution**(left half). For **N ›› ‹k›** the **binomial is well approximated by a Poisson distribution** (right half). As both formulas describe the same distribution,they have the identical properties, but they are expressed in terms of different parameters: The binomial distribution depends on p and N, while the Poisson distribution has only one parameter, ‹k›. It is this simplicity that makes the Poisson form preferred in calculations.

Random Network: not for real life!

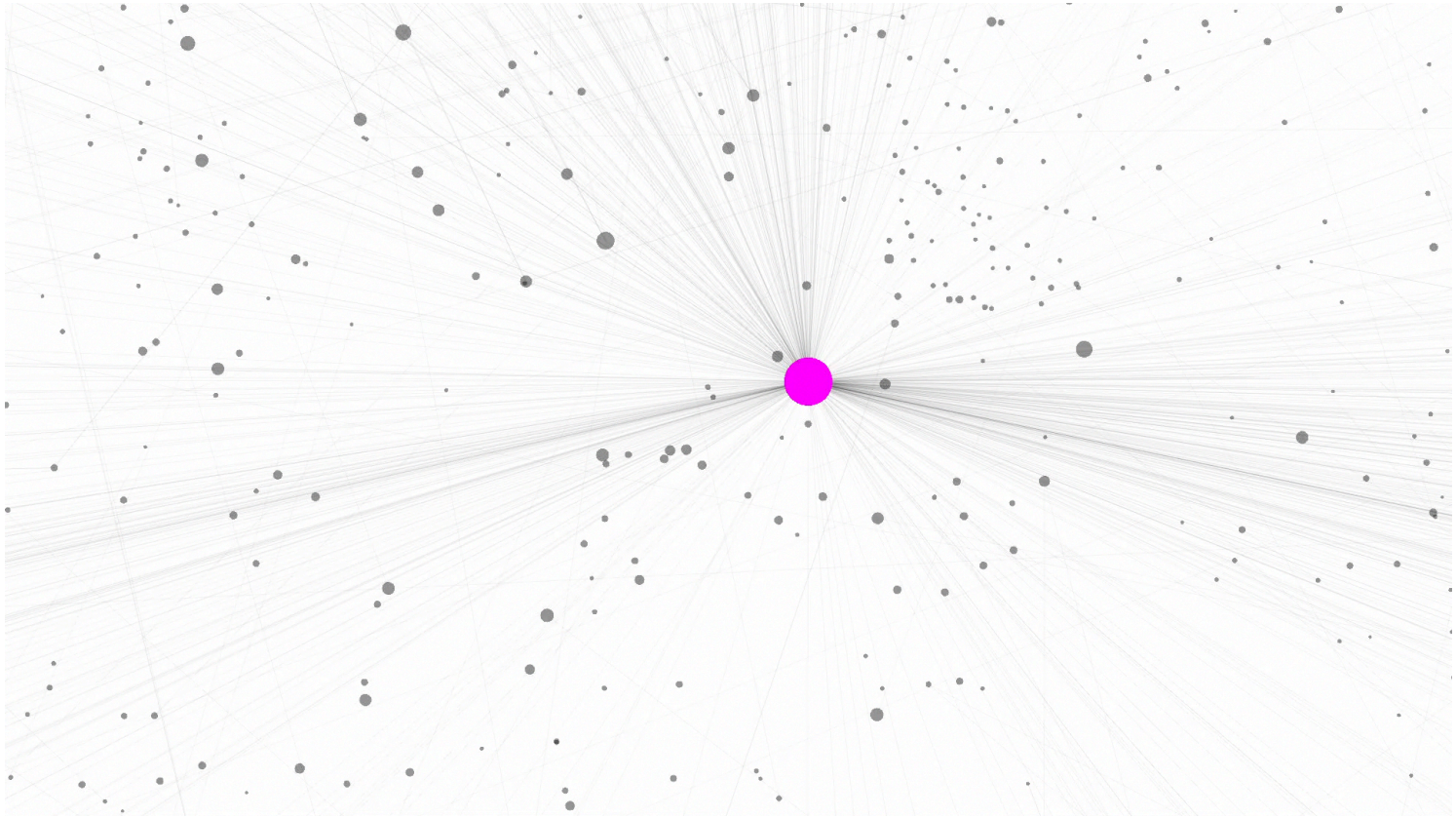


The degree distribution of the (a) Internet, (b) science collaboration network, and (c) protein interaction network.
The **green line corresponds to the Poisson prediction**, obtained by measuring ‹k› for the real network and then plotting. The significant deviation between the data and the Poisson fit indicates that the random network model underestimates the size and the frequency of the high degree nodes, as well as the number of low degree nodes. Instead the random network model predicts a larger number of nodes in the vicinity of ‹k› than seen in real networks.

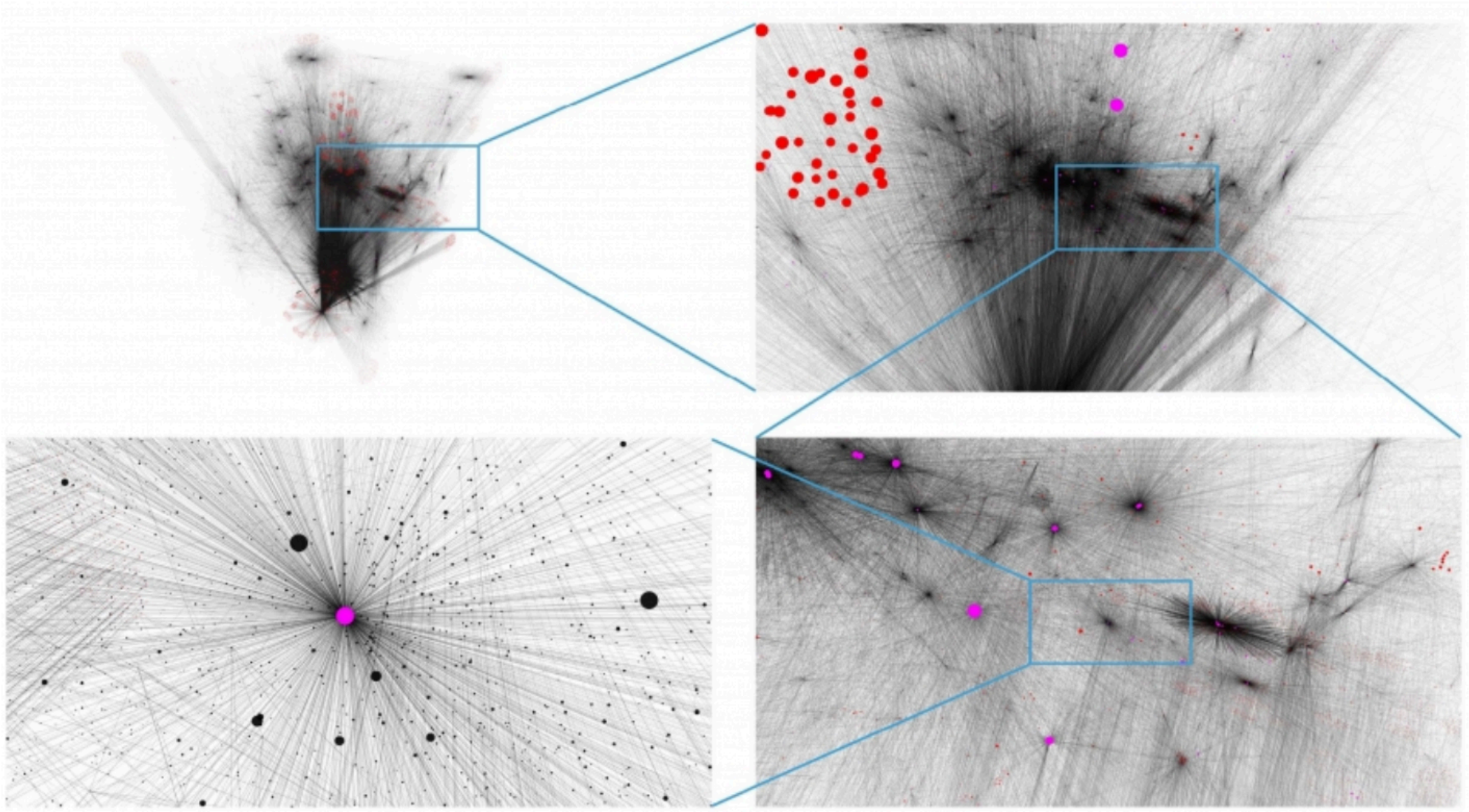
Scale Free Network

The World Wide Web is a network whose nodes are documents and the links are the uniform resource locators (URLs) that allow us to “surf” with a click from one web document to the other. With an estimated size of over one trillion documents (N≈1012), the Web is the largest network humanity has ever built. It exceeds in size even the human brain (N ≈ 1011 neurons).

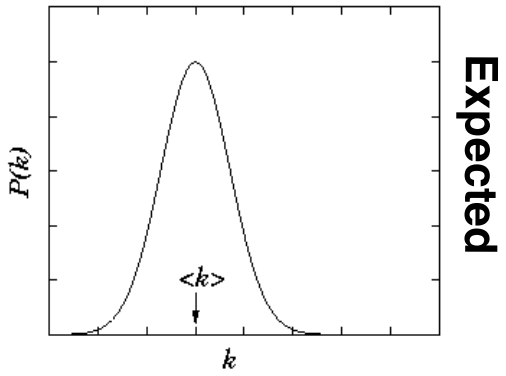
The first map of the WWW obtained with the explicit goal of understanding the structure of the network behind it was generated by Hawoong Jeong at University of Notre Dame. He mapped out the nd.edu domain, consisting of about 300,000 documents and 1.5 million links



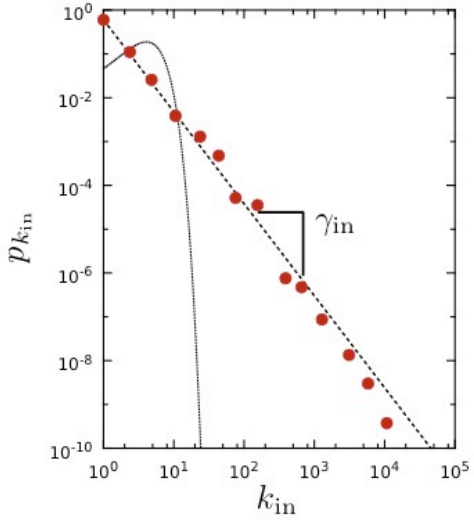
Scale Free Network

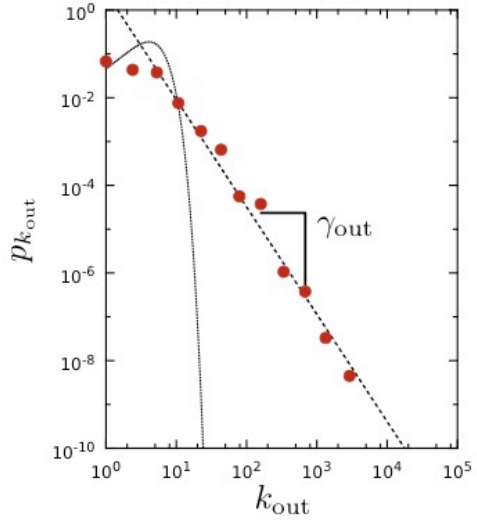


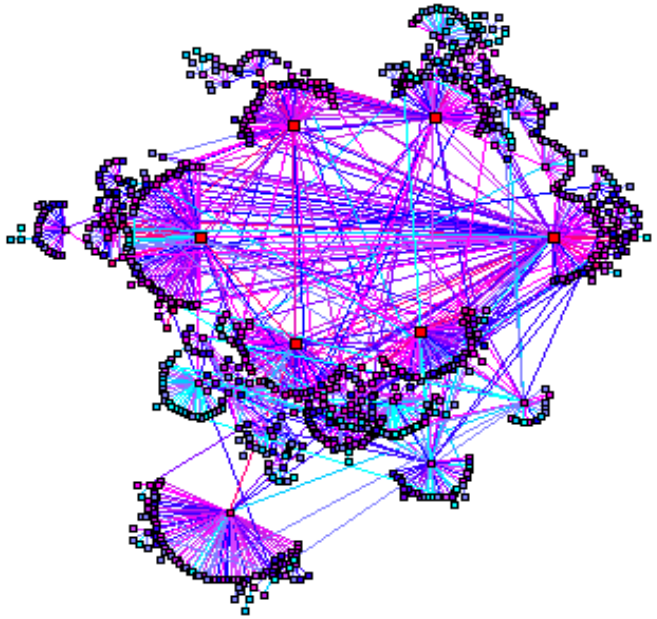
Scale Free Network



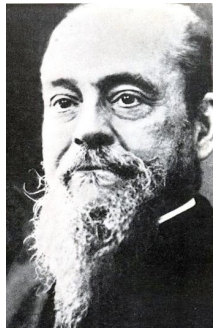
Expected

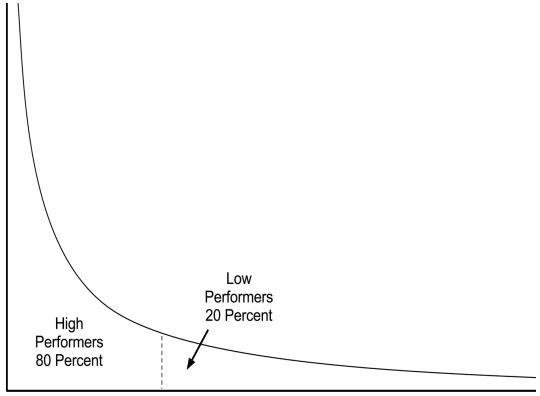




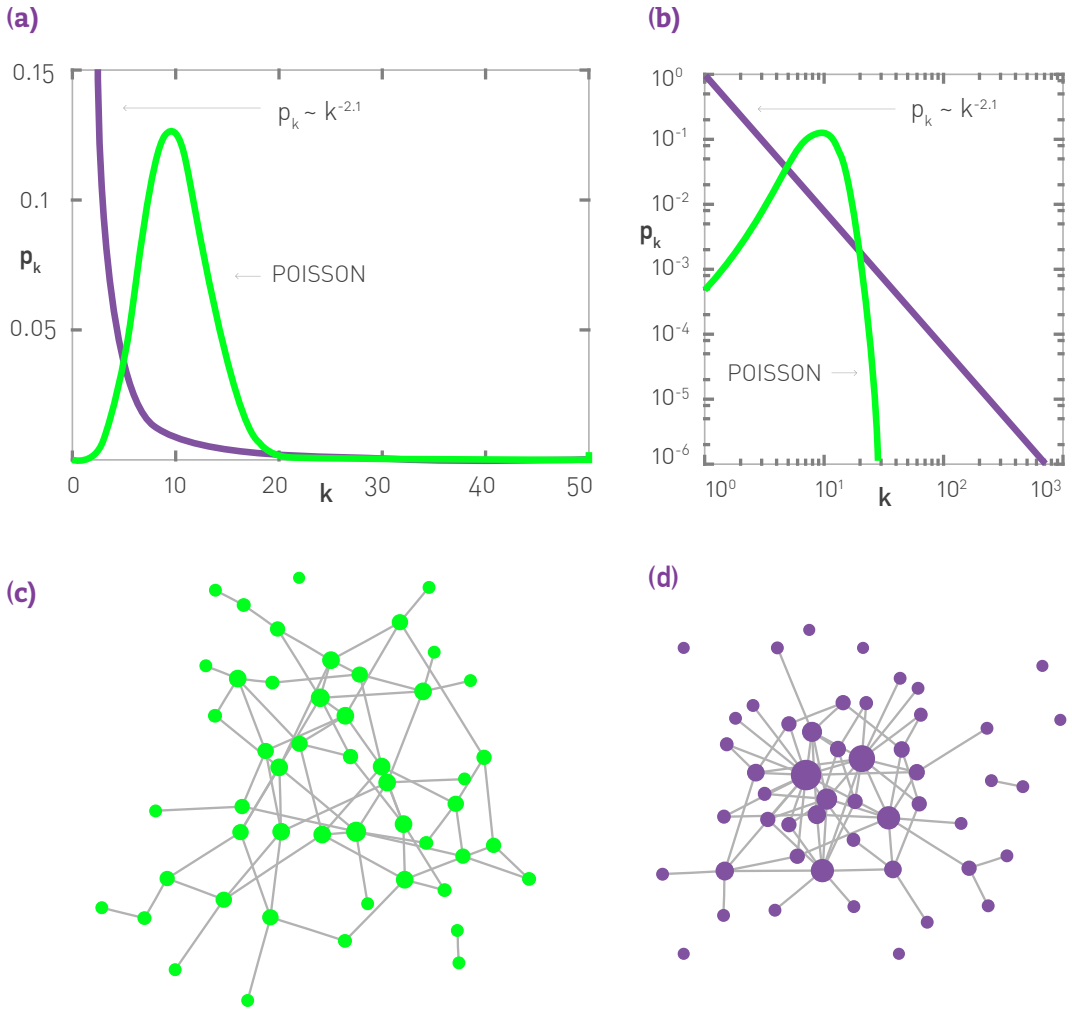


Vilfredo Pareto, a 19th century economist, noticed that in Italy a few wealthy individuals earned most of the money, while the majority of the population earned rather small amounts. He connected this disparity to the observation that incomes follow a power law, representing the first known report of a power law distribution. His finding entered the popular literature as the 80/20 rule: roughly 80 percent of money is earned by only 20 percent of the population. The 80/20 emerges in many areas



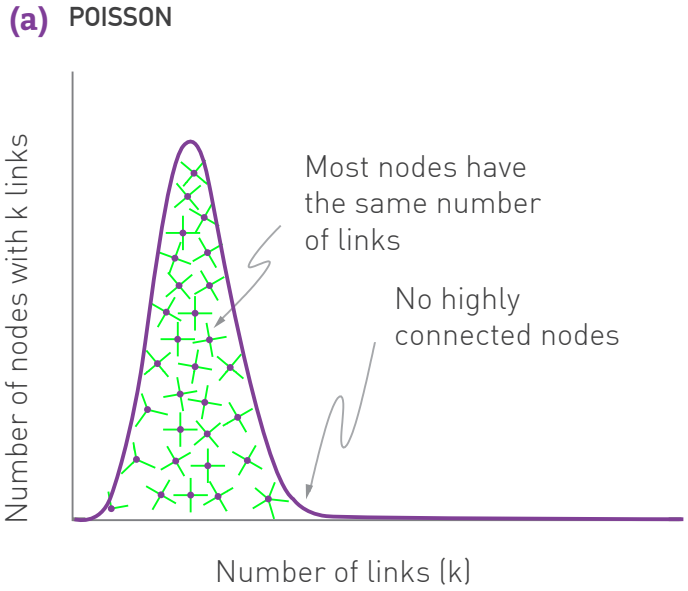


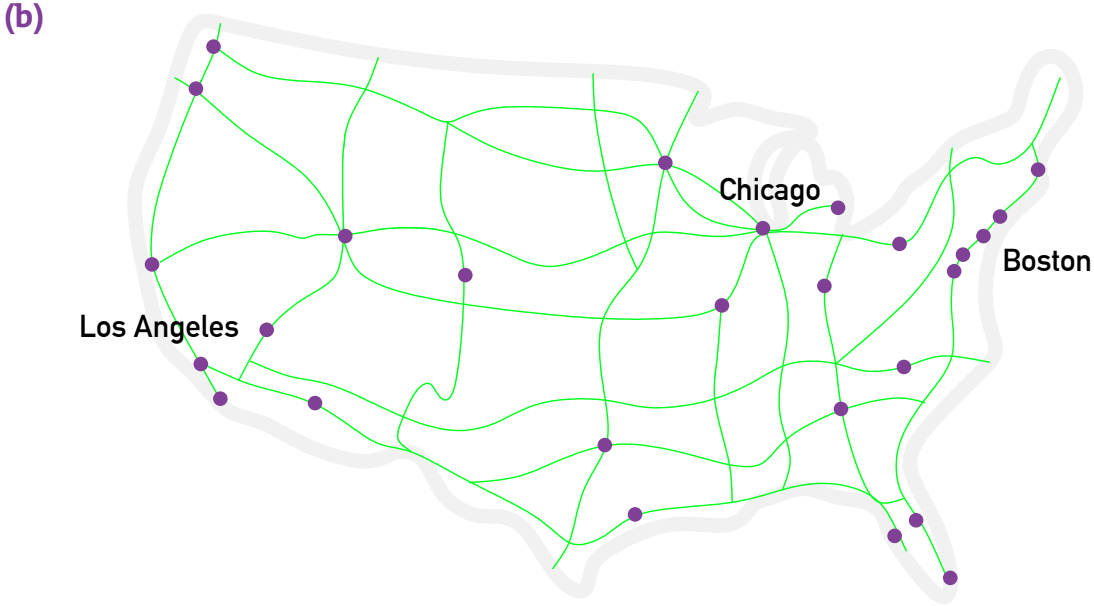
$$P(k) \approx k^{\gamma}$$

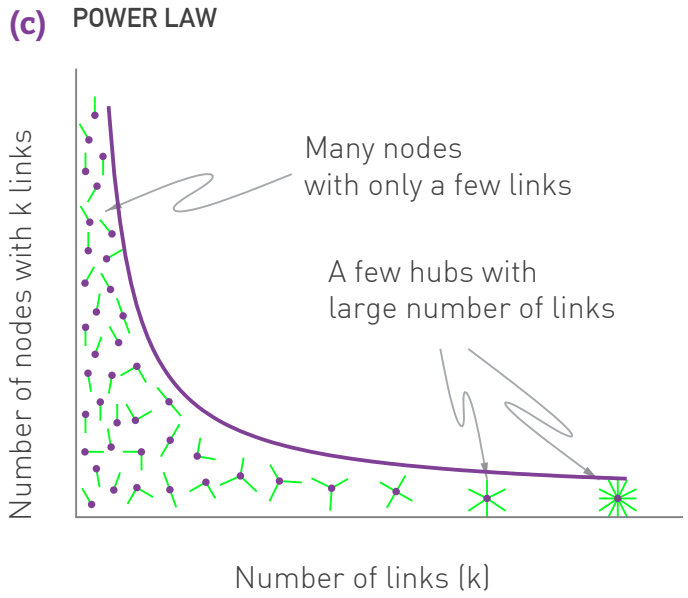


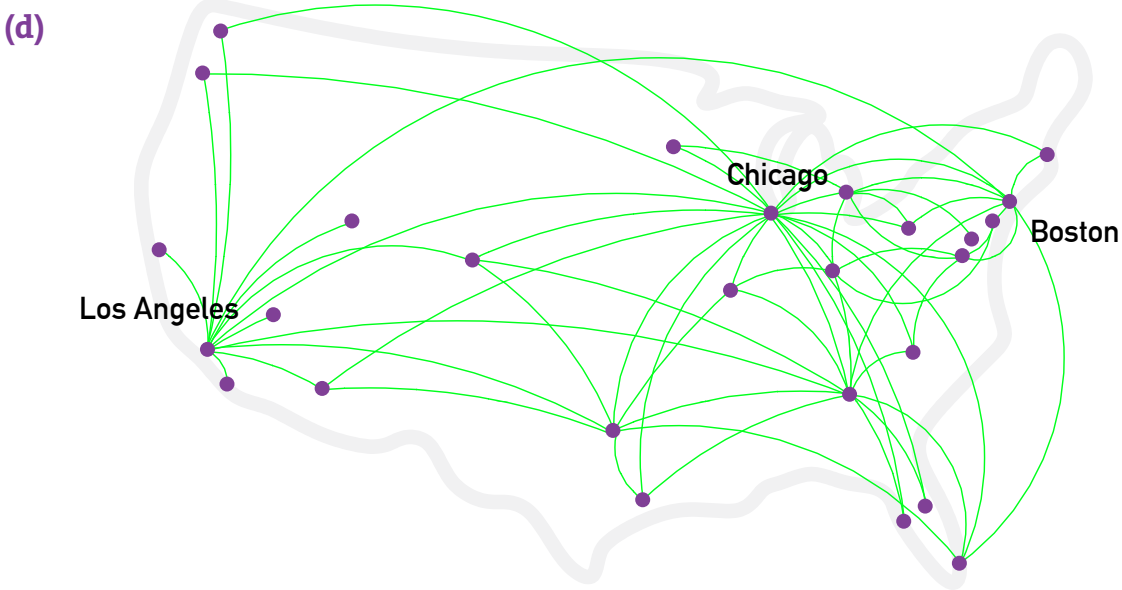
There appears to be little difference between them at the first look. Yet. There is some quite relevant different: the at high degrees the power law curve is always higher than the exponential.

Scale Free Network

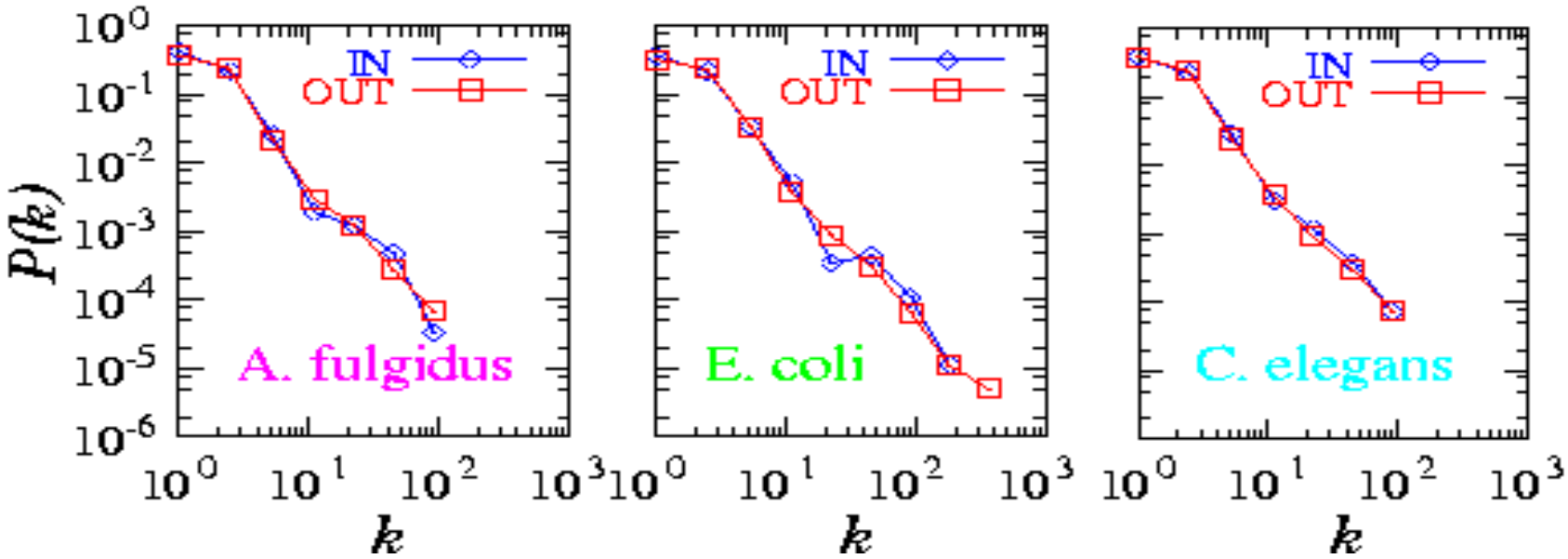








$P(k) \approx k^{\gamma}$



Organisms from all three domains of life are scale-free!

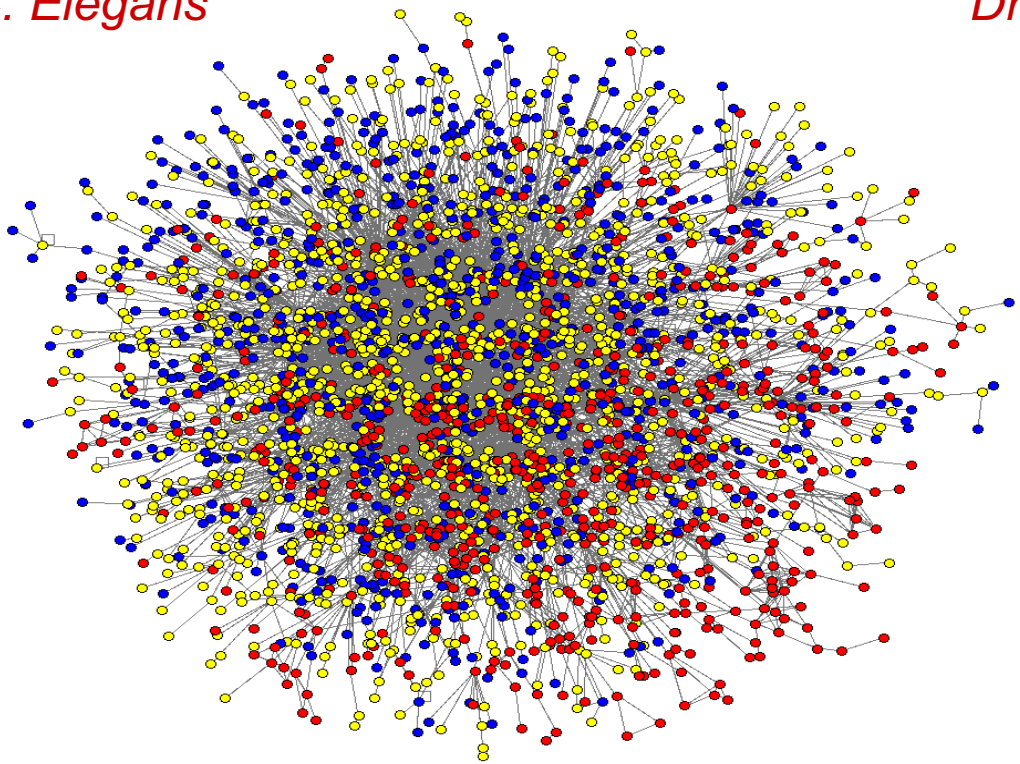
$$P_{in}(k) \approx k^{-2.2}$$

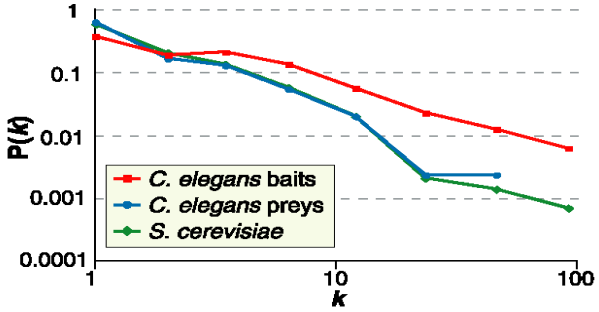
$$P_{out}(k) \approx k^{-2.2}$$

H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi, *Nature*, 407 651 (2000)

Scale Free Network

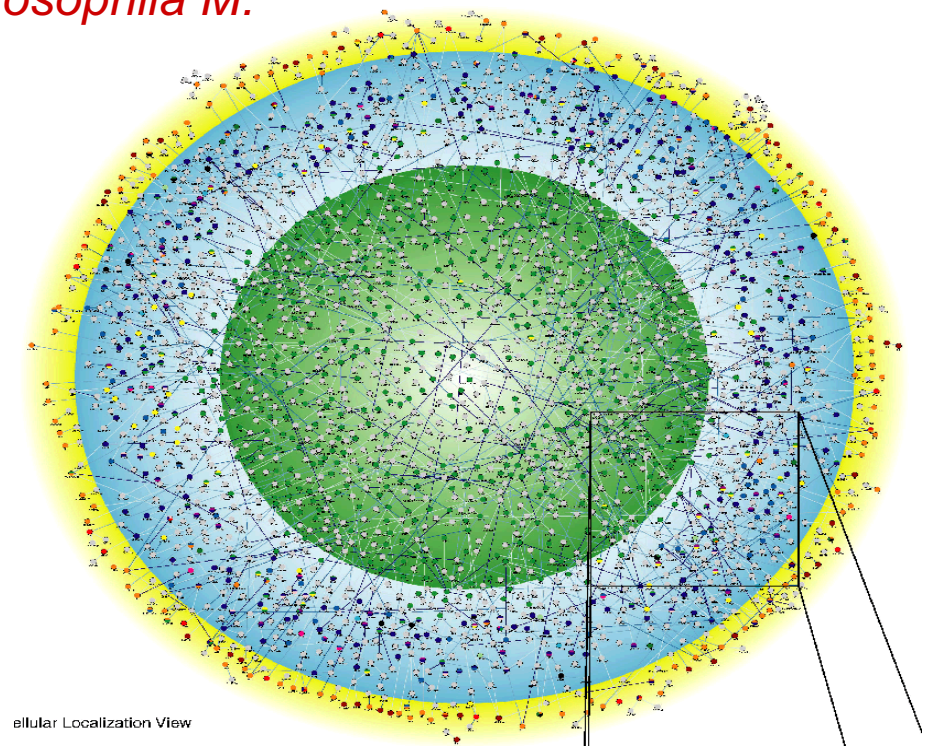
C. Elegans

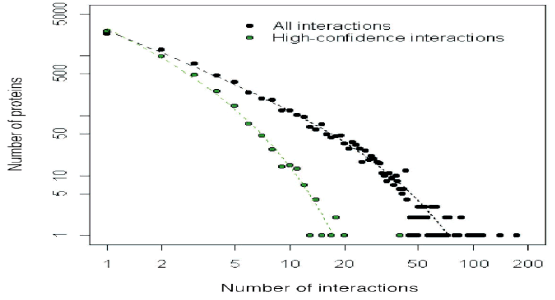




Li *et al.* Science 2004

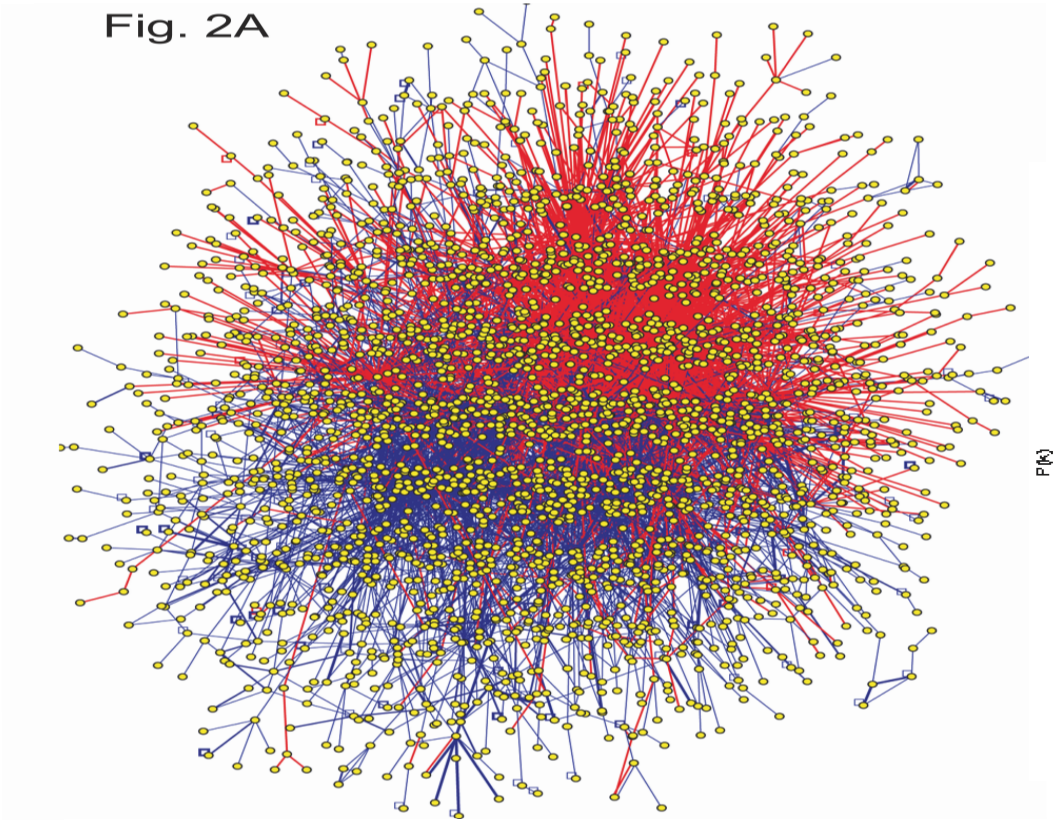
Drosophila M.

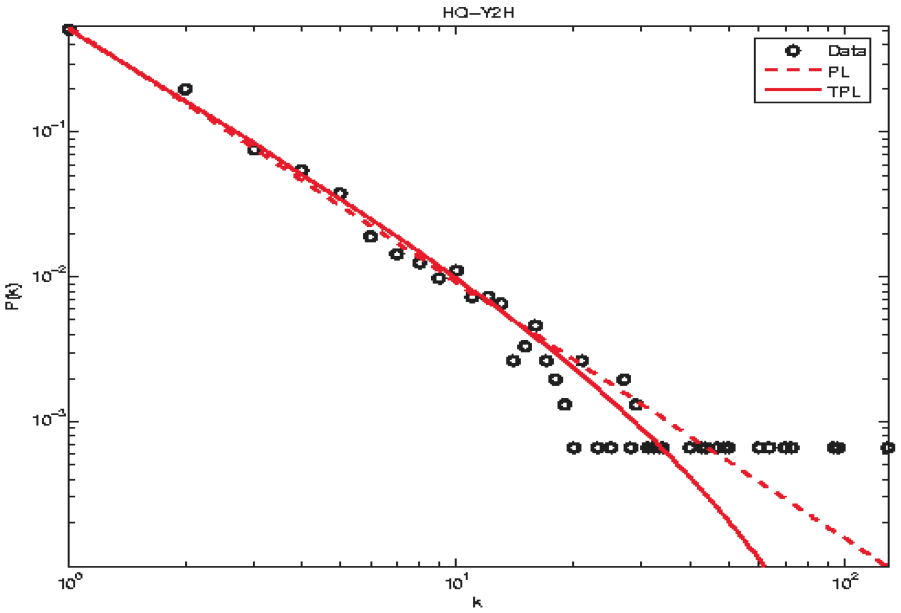




Giot *et al.* Science 2003

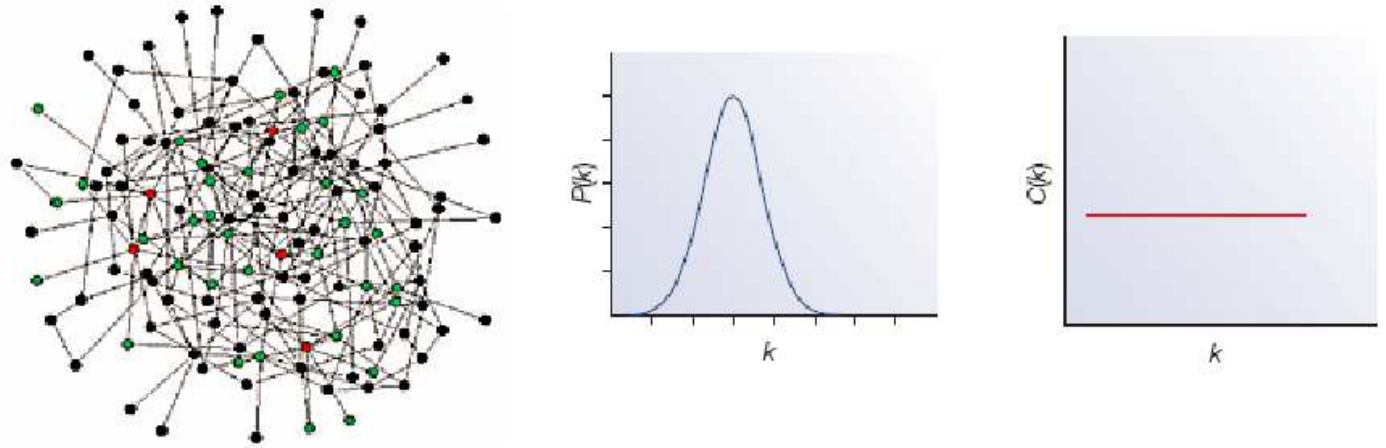
Scale Free Network





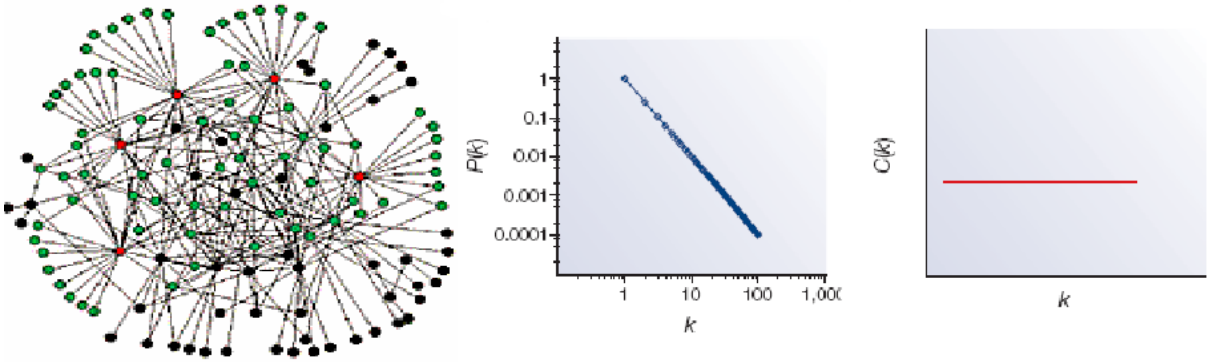
Rual *et al.* Nature 2005; Stelze *et al.* Cell 2005

Random networks The ER model start with N nodes and connects each pair of nodes with probability p. The node degree follow a Poisson distribution, means that most modes have approximately the same number of links. The clustering coefficient is independent of node's degree. The mean path length si proportional to the logarithm of network size $l \approx logN$.



Scale-free networks

Most nodes are poorly while a few are highly connected (Hubs). The degree distribution approximates a power law: $P(k) \approx k^{\gamma}$, where γ is the degree exponent. The smaller the γ, the more important is the role of the Hubs. Most biological networks have $2 < \gamma < 3$. For $\gamma > 3$, Hubs are irrelevant and the network behaves like a random network. The mean shortest path length is proportional to $\approx \ (log(log_N))$ (ie. Much shorter than Small World Property).



$$P(k) \approx k^{\gamma}$$

$$\langle d \rangle \begin{cases} const. & \gamma = 2 \\ \ln\ln N & 2 < \gamma < 3 \\ \dfrac{\ln N}{\ln\ln N} & \gamma = 3 \\ \ln N & \gamma > 3 \end{cases}$$

Ultra Small World

Small World

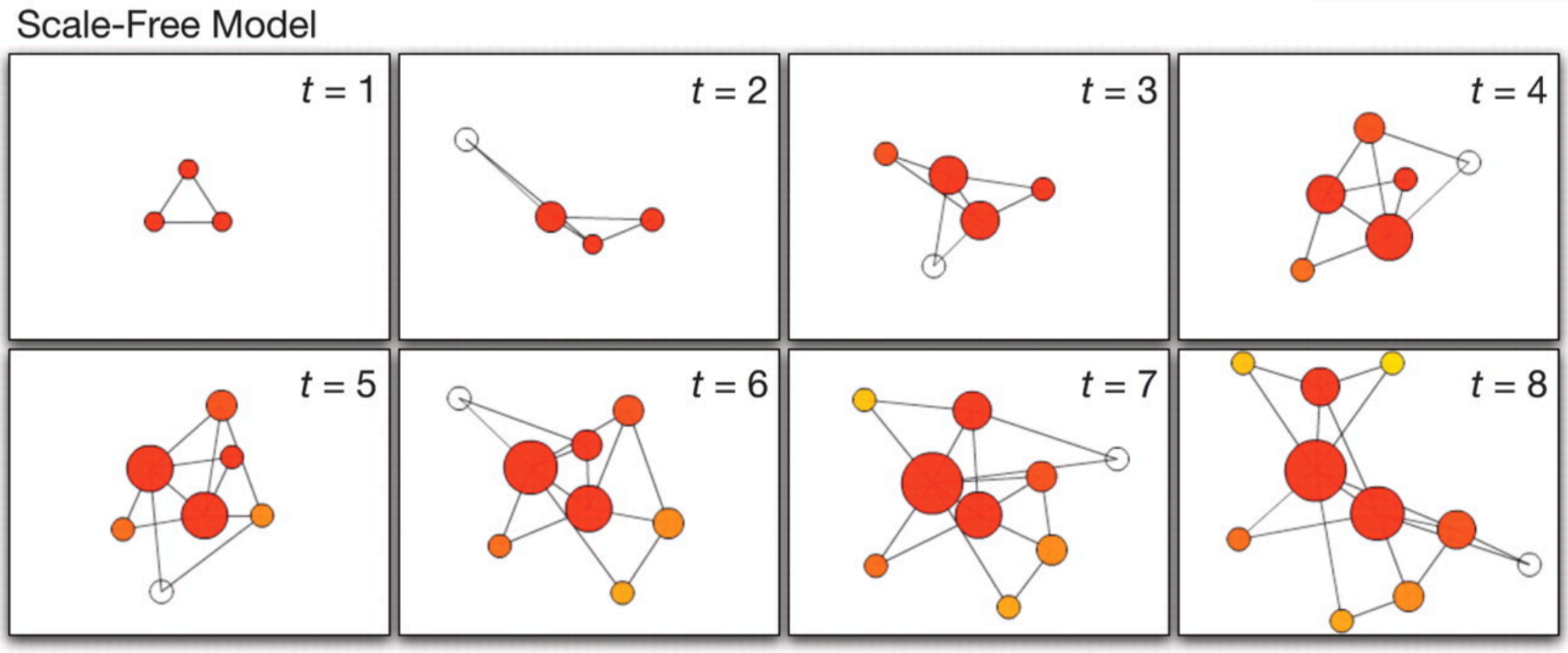
Size of the biggest hub is of order O(N). Most nodes can be connected within two layers of it, thus the average path length will be independent of the system size.

The average path length increases slower than logarithmically. In a random network all nodes have comparable degree, thus most paths will have comparable length. In a scale-free network the vast majority of the path go through the few high degree hubs, reducing the distances between nodes.

Some key models produce γ=3, so the result is of particular importance for them. This was first derived by Bollobas and collaborators for the network diameter in the context of a dynamical model, but it holds for the average path length as well.

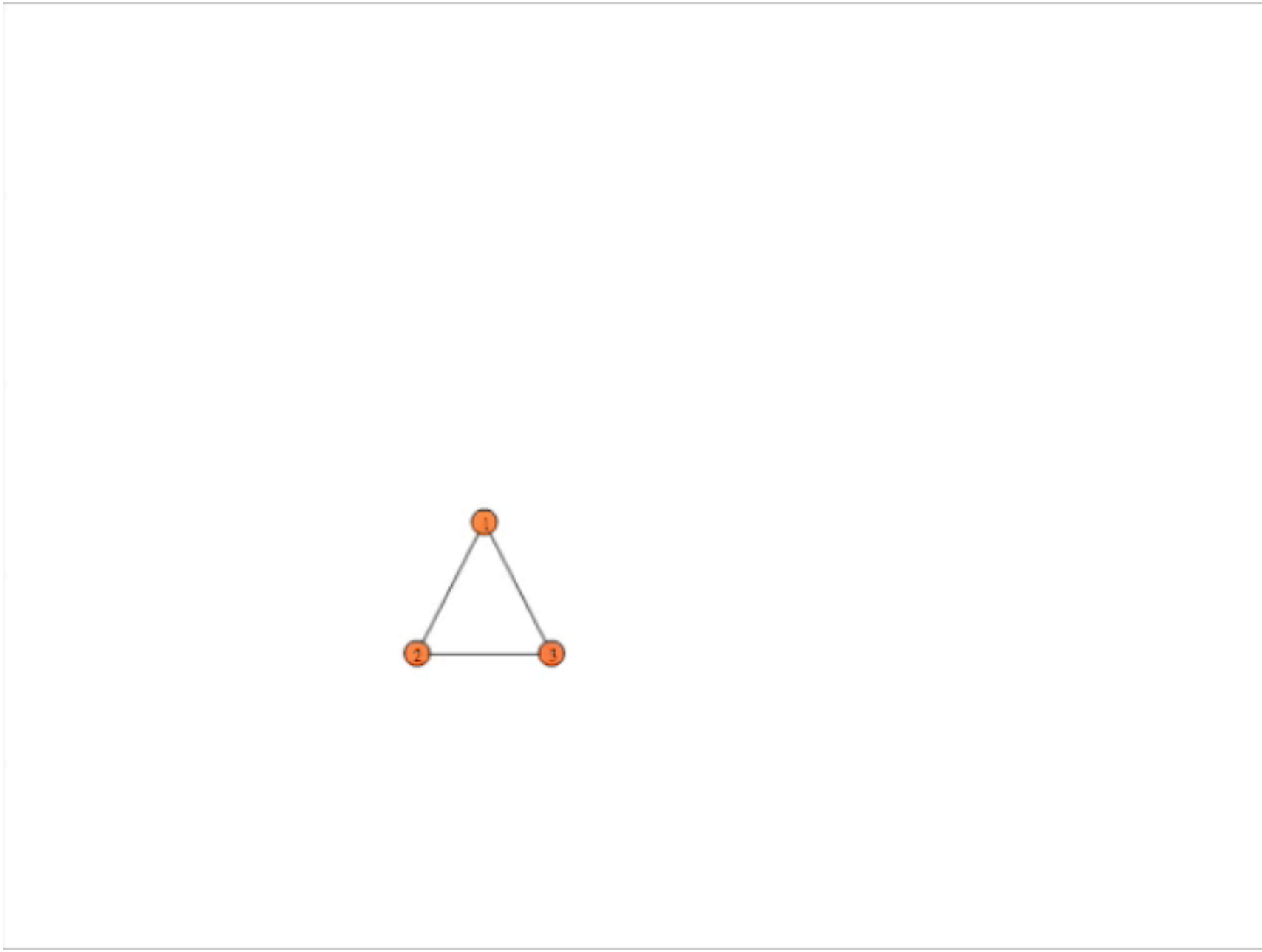
The second moment of the distribution is finite, thus in many ways the network behaves as a random network. Hence the average path length follows the result that we derived for the random network model earlier.

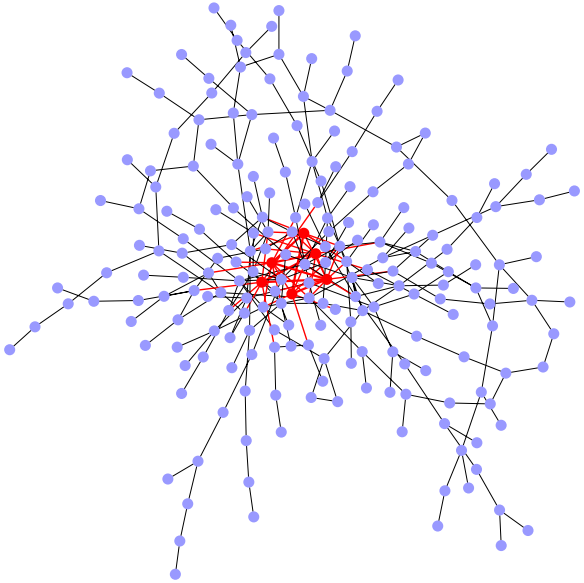
Scale Free Network



Starting from three connected nodes (top left), in each image a new node (shown as an empty circle) is added to the network. When deciding where to link, **new nodes prefer to attach to the more connected nodes**, a process known as preferential attachment. Thanks to growth and preferential attachment, a rich-gets-richer process is observed, which means that the highly connected nodes acquire more links than those that are less connected, leading to the natural emergence of a few highly connected hubs. The node size, which was chosen to be proportional to the node's degree, illustrates the natural emergence of hubs as the largest nodes.

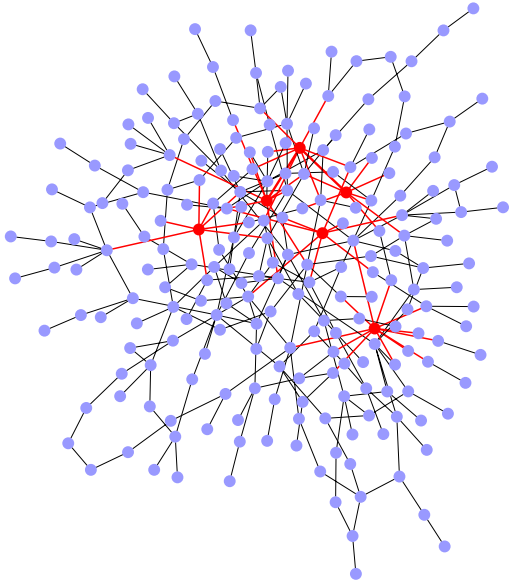
Scale Free Network





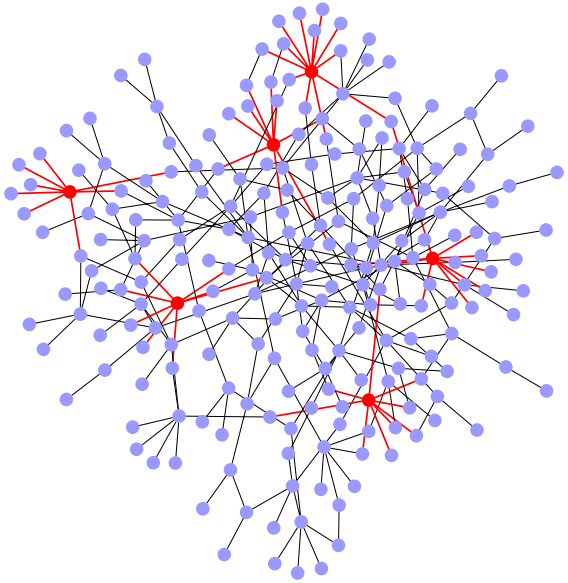
Assortative:

hubs show a tendency to link to each other.



Neutral:

nodes connect to each other with the expected random probabilities.



Disassortative:

Hubs tend to avoid linking to each other.

Quantifying degree correlations (three approaches):

→ full statistical description (Maslov and Sneppen, Science 2001)
→ degree correlation function (Pastor Satorras and Vespignani, PRL 2001)
→ correlation coefficient (Newman, PRL 2002)

Network	n	r
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
Marine food web (i)	134	−0.247
Freshwater food web (j)	92	−0.276
Random graph (u)		0
Callaway *et al.* (v)		$\delta/(1+2\delta)$
Barabási and Albert (w)		0

Social networks are *assortative*

Biological, technological networks are *disassortative*

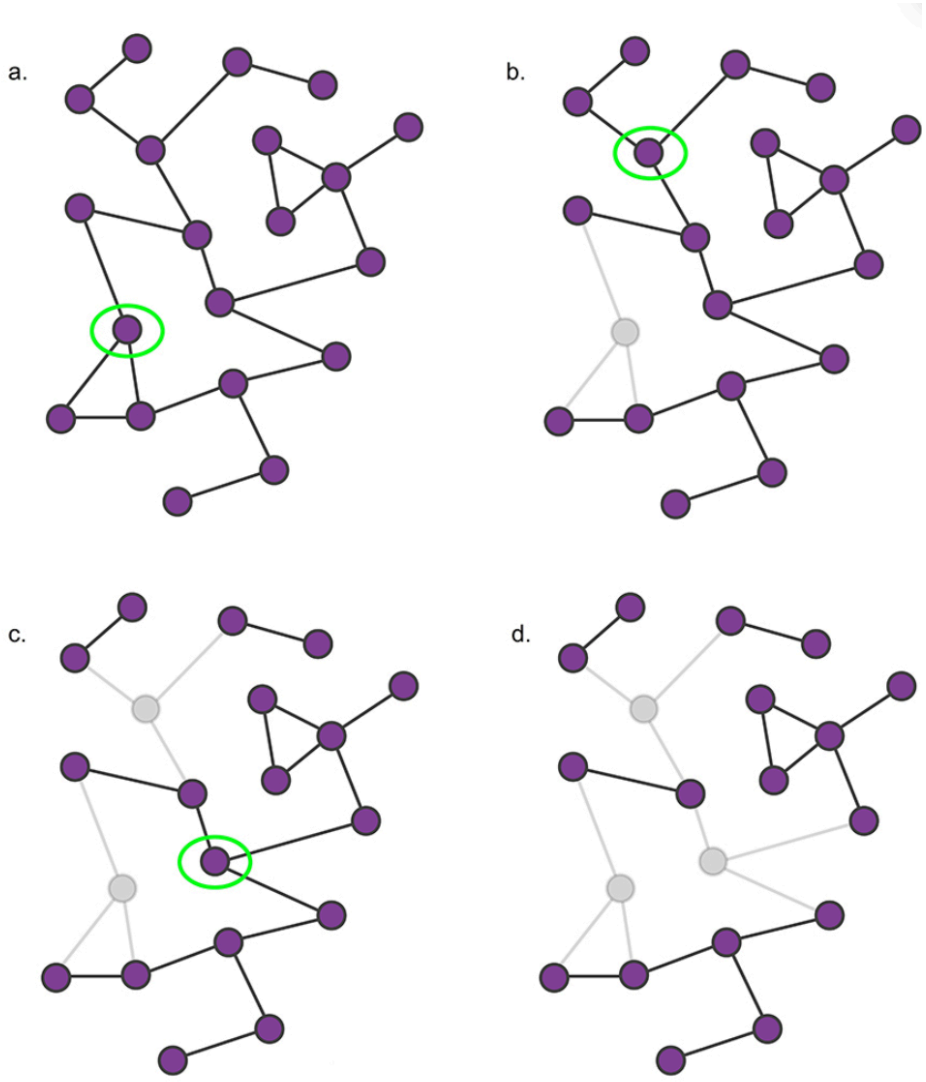
r>0: assortative network:
Hubs tend to connect to other hubs.

r<0: disassortative network:
Hubs tend to connect to small nodes.

Robustness

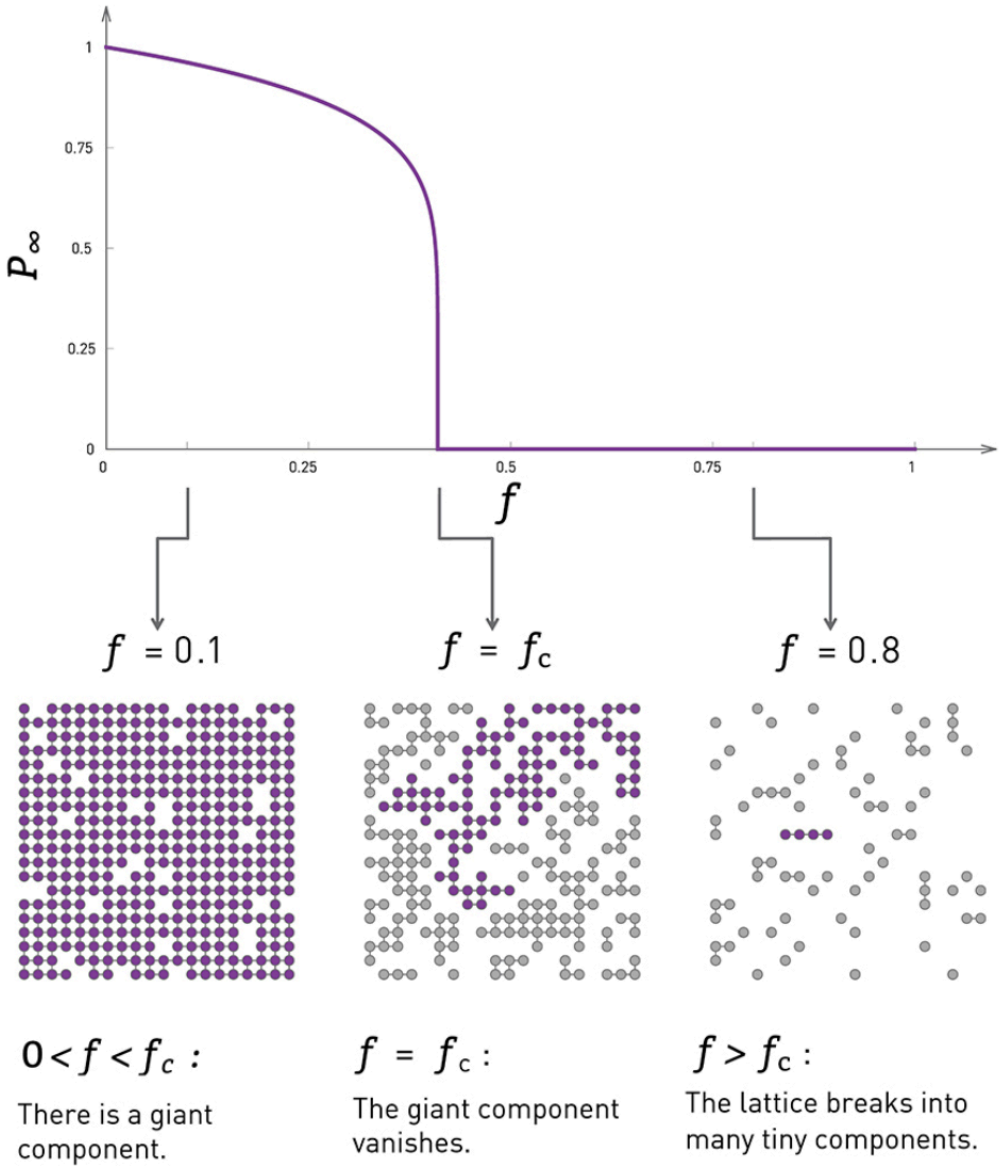
Robustness is a central question in biology and medicine, helping us understand why some mutations lead to diseases and others do not.

Networks play a key role in the robustness of biological, social and technological systems. Indeed, a cell's robustness is encoded in intricate regulatory, signaling and metabolic networks



The removal of a single node has only limited impact on a network's integrity. The removal of several nodes, however, can break a network into several isolated components. Obviously, the more nodes we remove, the higher are the chances that we damage a network, prompting us to ask: **How many nodes do we have to delete to fragment a network into isolated components?**

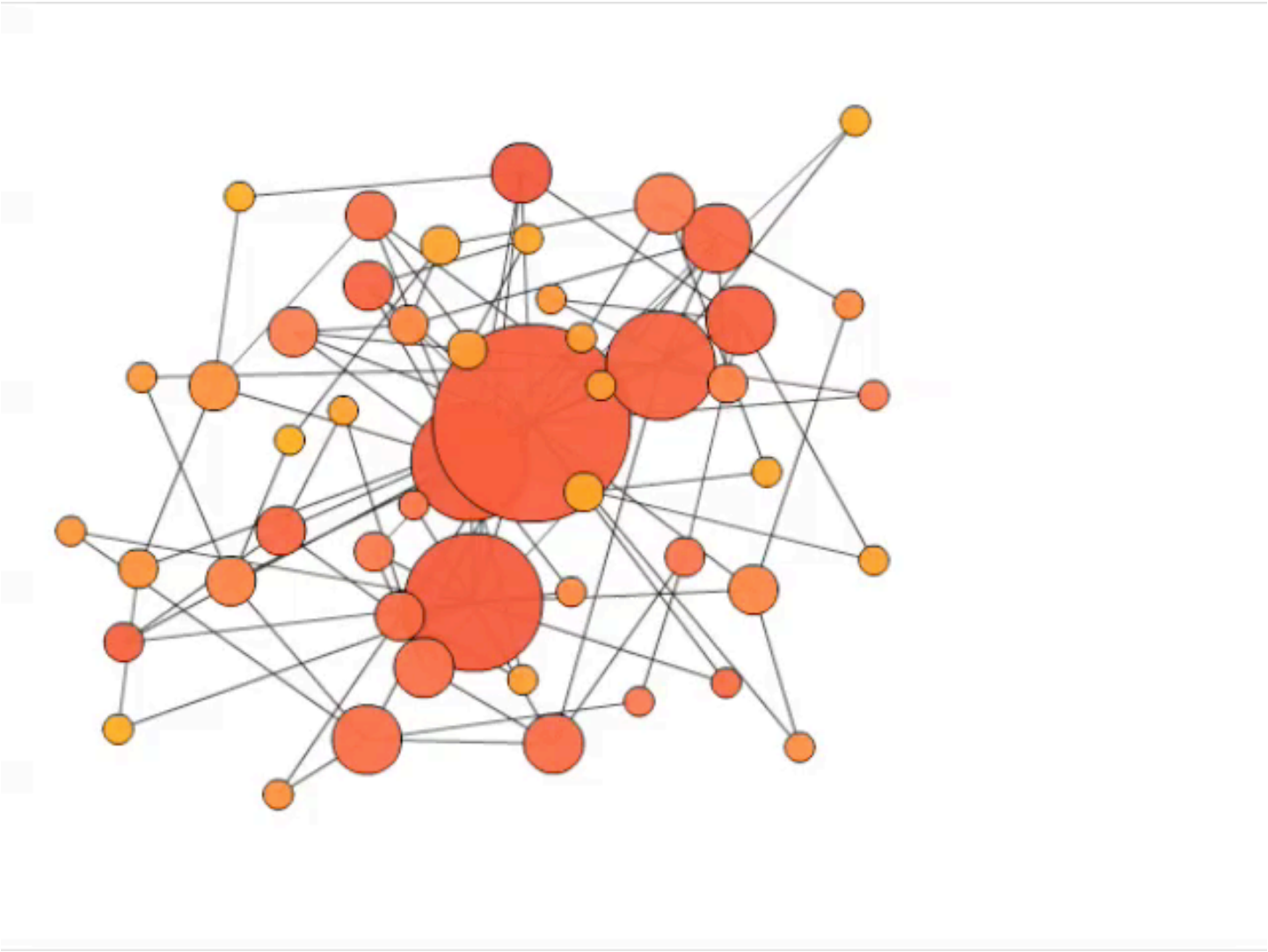
Robustness



If f is small, the missing nodes do little damage to the network. Increasing f, however, can isolate chunks of nodes from the giant component. Finally, for sufficiently large f the giant component breaks into tiny disconnected components

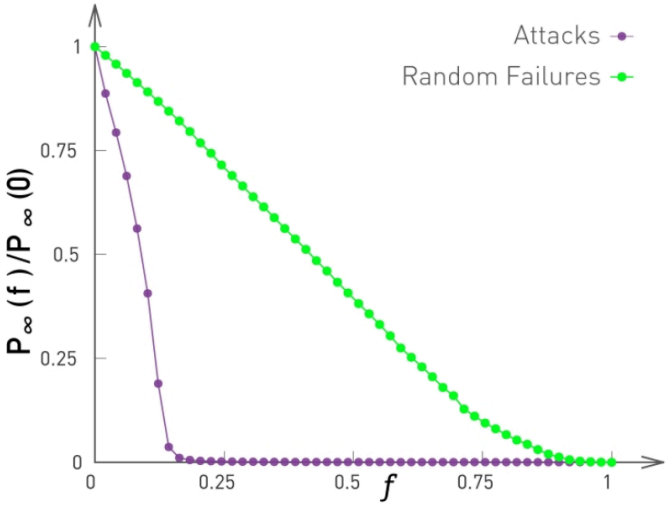
This fragmentation process is not gradual, but it is characterized by a critical threshold fc: For any f< fc we continue to have a giant component. Once f exceeds fc, the giant component vanishes. This is illustrated by the f-dependence of P∞, representing the probability that a node is part of the giant component: P∞ is nonzero under fc, but it drops to zero as we approach fc.

Robustness



To illustrate the robustness of a scale-free network we start from the scale-free network. Next we randomly select and remove nodes one-by-one. As the movie illustrates, despite the fact that we remove a significant fraction of the nodes, the network refuses to break apart.

Robustness

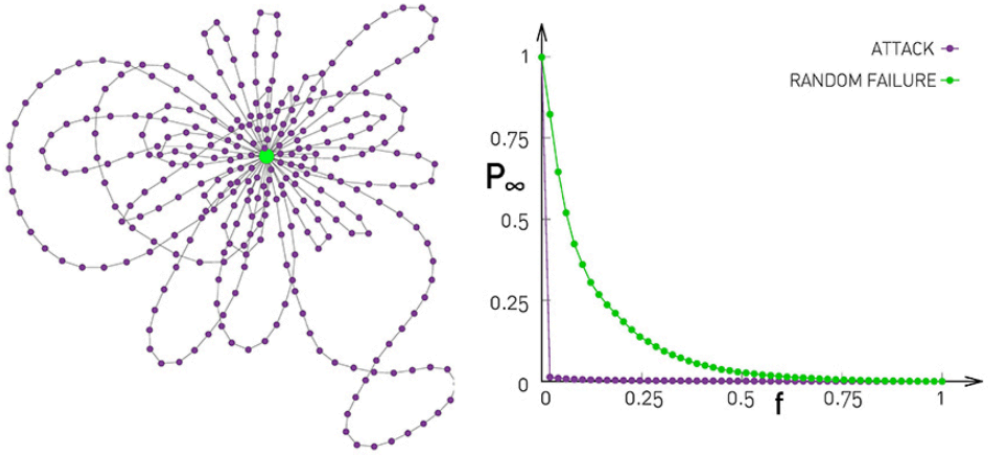


ork	Random Failures (Real Network)	Random Failures (Randomized Network)	Attack (Real Network)
Internet	0.92	0.84	0.16
WWW	0.88	0.85	0.12
Power Grid	0.61	0.63	0.20
Mobile Phone Calls	0.78	0.68	0.20
Email	0.92	0.69	0.04
Science Collaboration	0.92	0.88	0.27
Actor Network	0.98	0.99	0.55
Citation Network	0.96	0.95	0.76
E. Coli Metabolism	0.96	0.90	0.49
Protein Interactions	0.88	0.66	0.06

The estimated fc for random node failures (second column) and attacks (fourth column) for ten reference networks

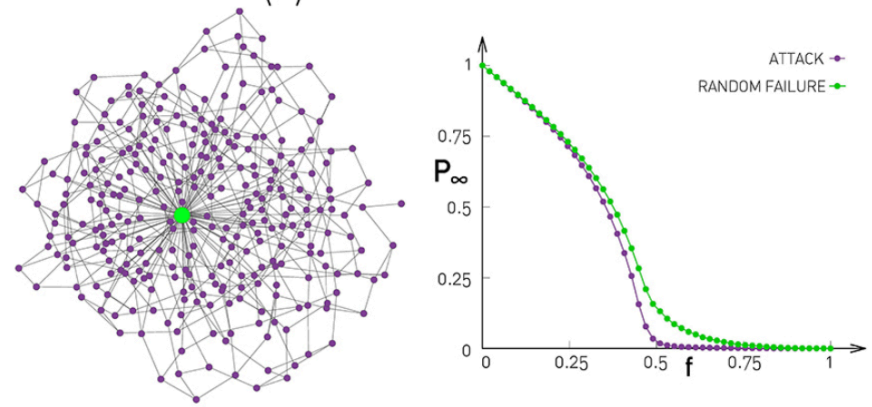
Robustness

a. $\langle k \rangle = 2$



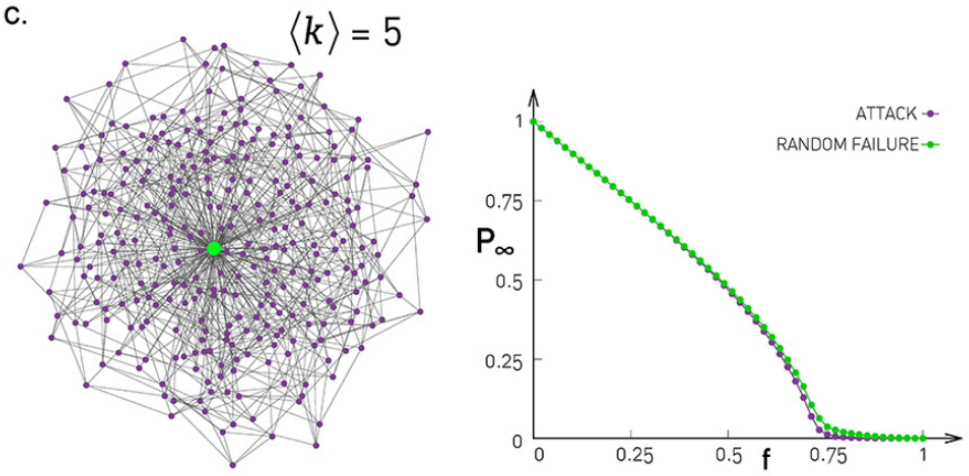
For small 〈k〉 the hub holds the network together. Once we remove this central hub the network breaks apart. Hence the attack and error curves are well separated, indicating that the network is robust to random failures but fragile to attacks.

b. $\langle k \rangle = 3$



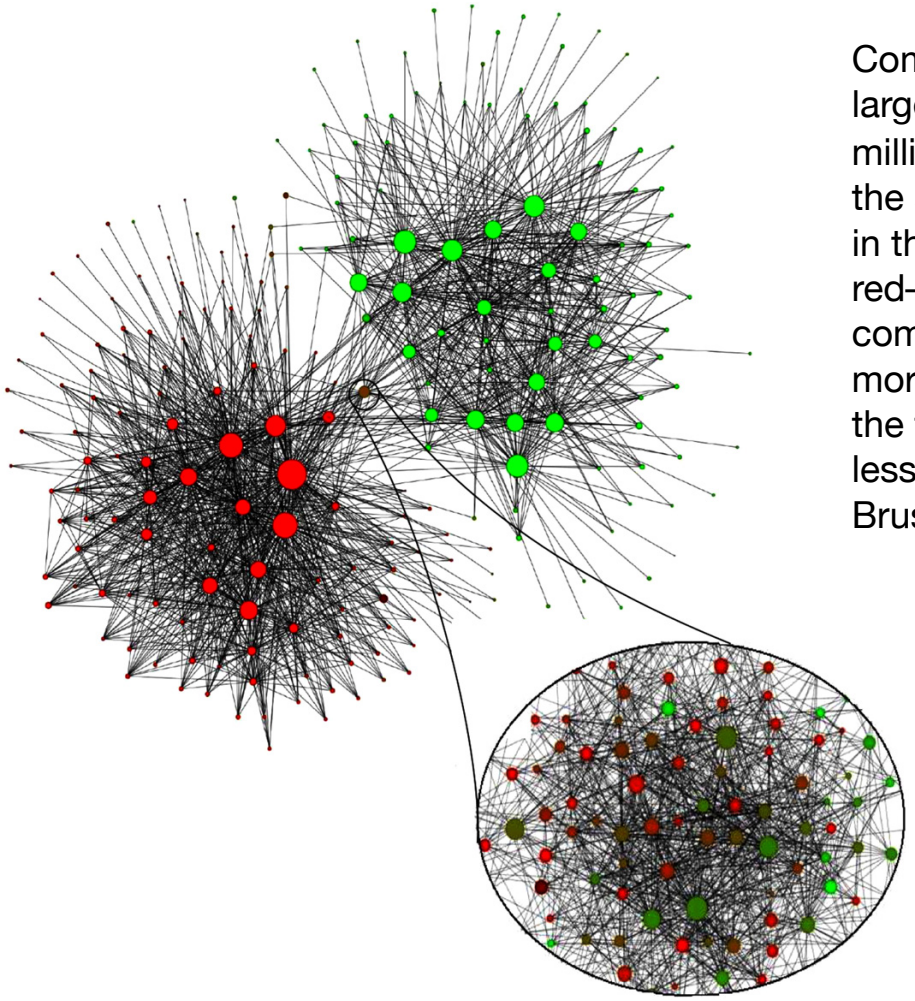
•For larger 〈k〉 a giant component emerges, that exists even without the central hub. Hence while the hub enhances the system's robustness to random failures, it is no longer essential for the network.

c. $\langle k \rangle = 5$



For even larger 〈k〉 the error and the attack curves are indistinguishable, indicating \that the network's response to attacks and random failures is indistinguishable. In this case the network is well connected even without its central hub.

Belgium appears to be the model bicultural society: 59% of its citizens are Flemish, speaking Dutch and 40% are Walloons who speak French.



Communities extracted from the call pattern of the consumers of the largest Belgian mobile phone company. The network has about two million mobile phone users. The nodes correspond to communities, the size of each node being proportional to the number of individuals in the corresponding community. The color of each community on a red–green scale represents the language spoken in the particular community, red for French and green for Dutch. Only communities of more than 100 individuals are shown. The community that connects the two main clusters consists of several smaller communities with less obvious language separation, capturing the culturally mixed Brussels, the country's capital.

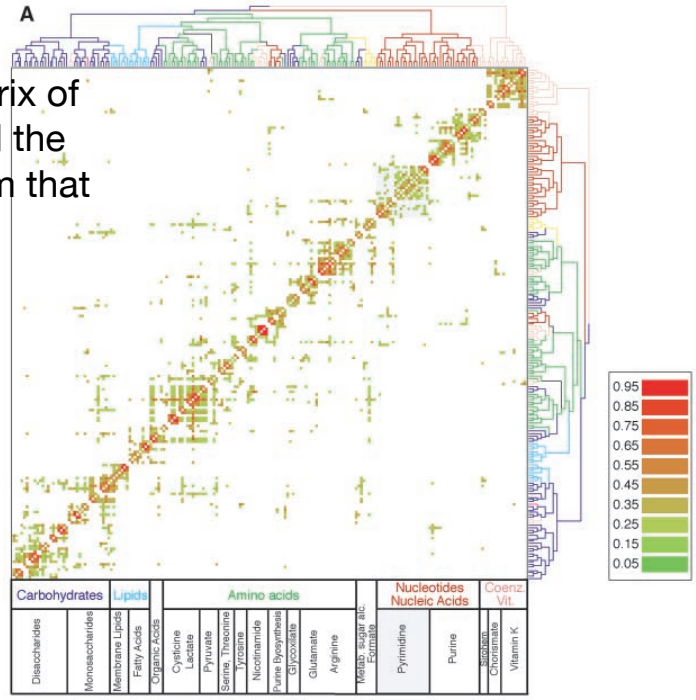
In network science we call a **community** a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities.

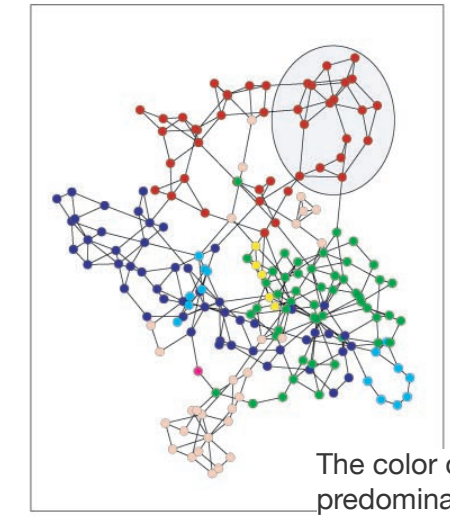
Communities play a particularly important role in our understanding of how specific biological functions are encoded in cellular networks.

Communities play a particularly important role in understanding human diseases. Indeed, proteins that are involved in the same disease tend to interact with each other. This finding inspired the disease module hypothesis, stating that each disease can be linked to a well-defined neighborhood of the cellular network.

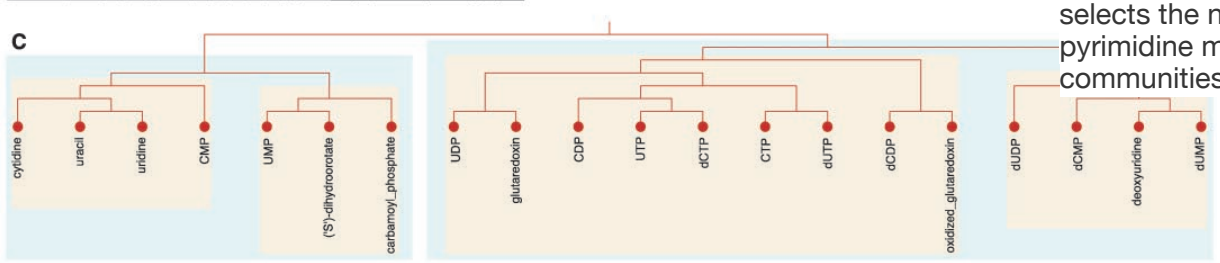
Communities

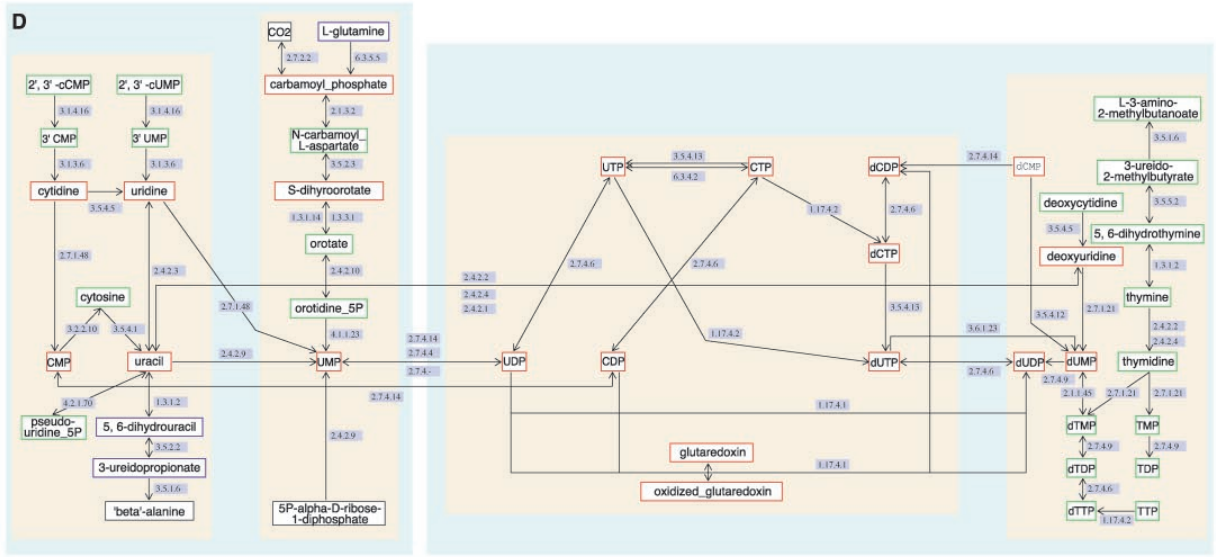
The topologic overlap matrix of the E. coli metabolism and the corresponding dendrogram that allows us to identify the modules





The color of each node, capturing the predominant biochemical class to which it belongs, indicates that different functional classes are segregated in distinct network neighborhoods. The highlighted region selects the nodes that belong to the pyrimidine metabolism, one of the predicted communities.





Maximum Cliques: In graph theoretic terms this means that a community is a complete subgraph, or a clique.

Strong and Weak Communities

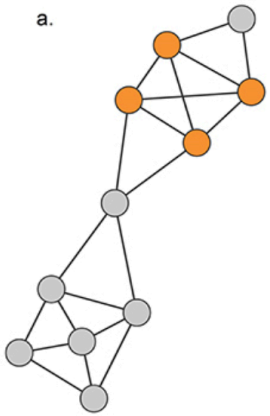
To relax the rigidity of cliques, consider a connected subgraph C of NC nodes in a network. The **internal degree kiint** of node i is the number of links that connect i to other nodes in C. The **external degree kiextis** the number of links that connect i to the rest of the network. If kiext=0, each neighbor of i is within C, hence C is a good community for node i. If kiint=0, then node i should be assigned to a different community. These definitions allow us to distinguish two kinds of communities

strong community:
each node has more links within the community than with the rest of the graph.

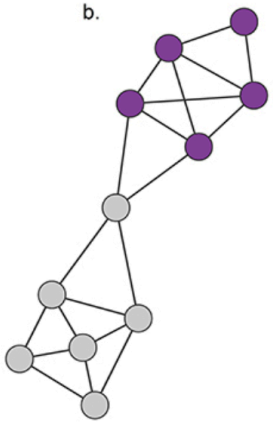
$$k_i^{\text{int}}(C) > k_i^{\text{ext}}(C)$$

weak community:
the total internal degree of the subgraph exceeds its total external degree,

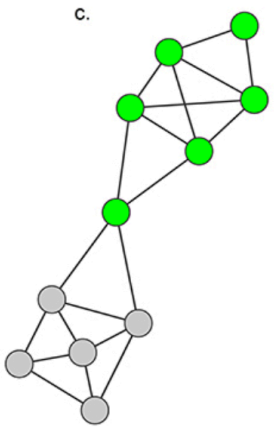
$$\sum_{i \in C} k_i^{\text{in}}(C) > \sum_{i \in C} k_i^{\text{out}}(C)$$



A clique correspond s to a complete subgraph.



A strong community



A weak community

To uncover the community structure of large real networks we need algorithms whose running time grows polynomially with N. **Hierarchical clustering.**

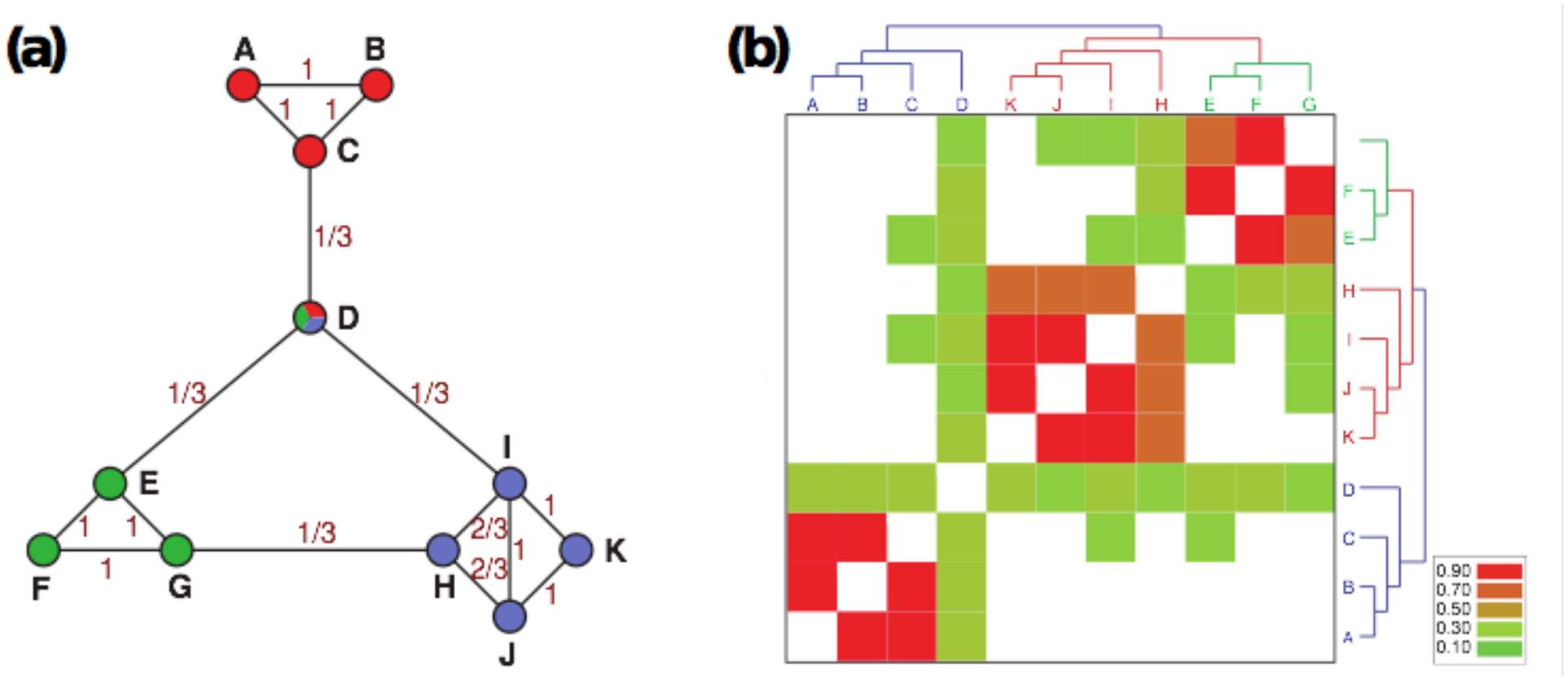
The starting point of hierarchical clustering is a **similarity matrix,** whose elements xij indicate the distance of node i from node j. In community identification the similarity is extracted from the relative position of nodes i and j within the network.

Step 1: Define the Similarity Matrix.

- high for node pairs that likely belong to the same community;
- low for those that likely belong to different communities.
- Nodes that connect directly to each other and/or share multiple neighbors are more likely to belong to the same dense local neighborhood, hence their should be large.

Topological overlap matrix:

$J_N(i,j)$: number of common neighbors of node i and j;
(+1) if there is a direct link between i and j;



J=1 if nodes and have the same neighbors (A&B)

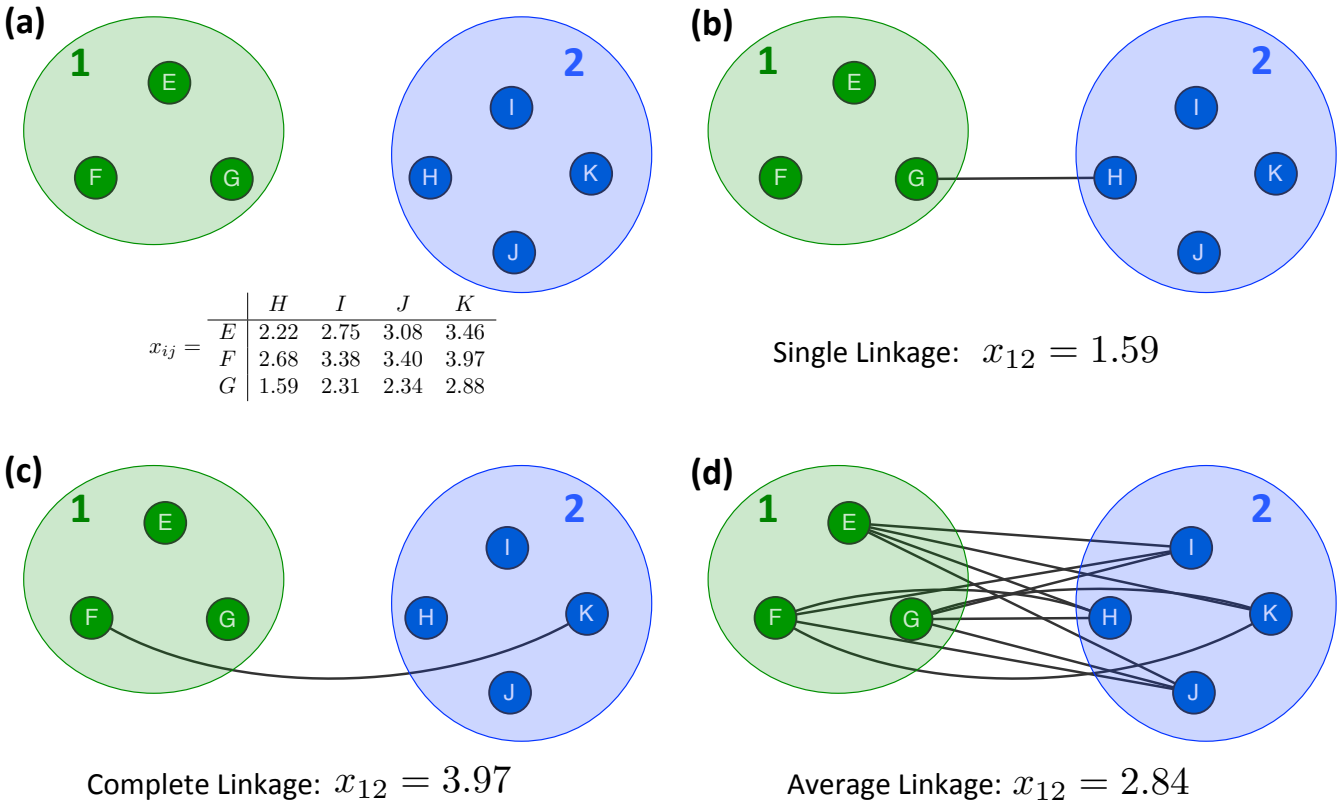
J=0 if do not have common neighbors, nor do they link to each other (A&E)

Members of the same dense local neighborhood have high JN(i,j), like nodes H, I, J, K

Step 2: Decide Group Similarity.

- groups are merged based on their mutual similarity:

single, *complete* and *average cluster similarity*



$x_{ij} =$

	H	I	J	K
E	2.22	2.75	3.08	3.46
F	2.68	3.38	3.40	3.97
G	1.59	2.31	2.34	2.88

Single Linkage: $x_{12} = 1.59$

Complete Linkage: $x_{12} = 3.97$

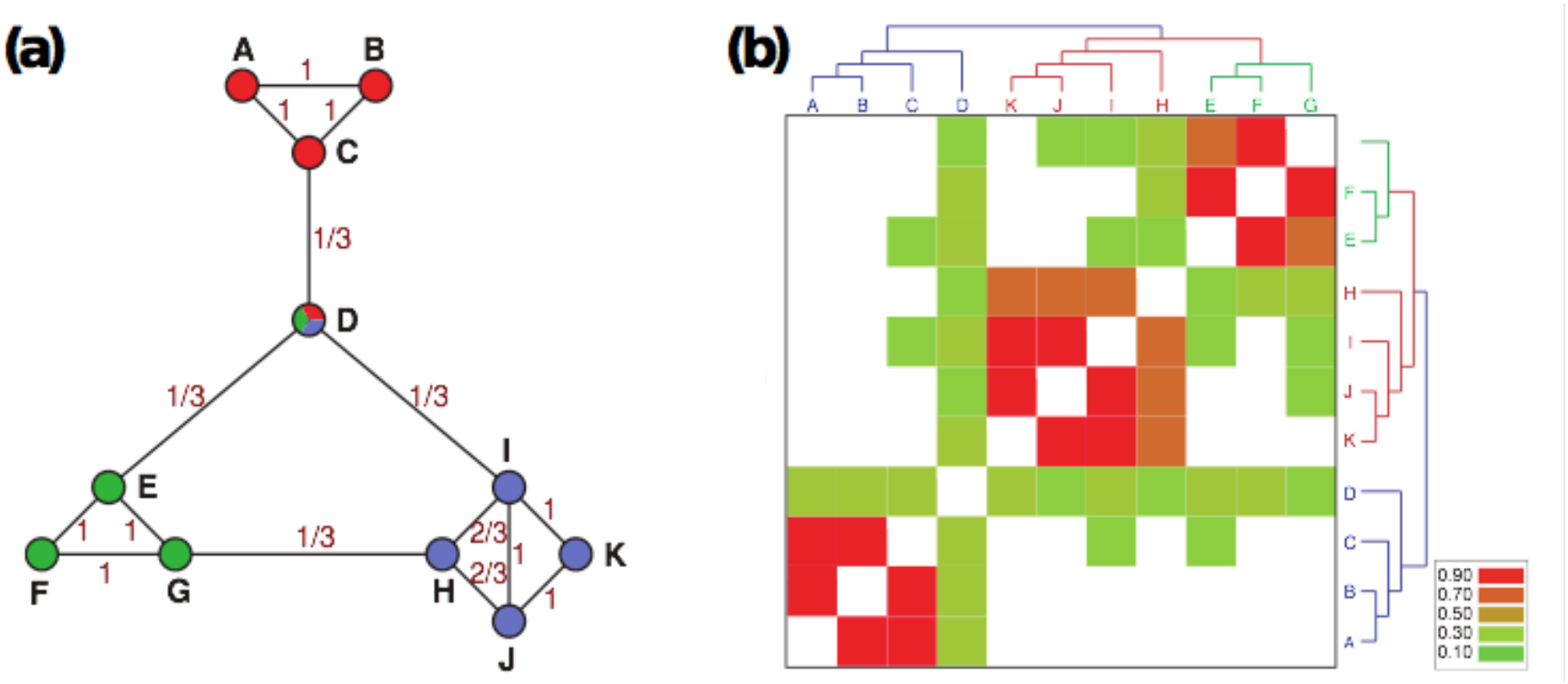
Average Linkage: $x_{12} = 2.84$

Step 3: Apply Hierarchical Clustering

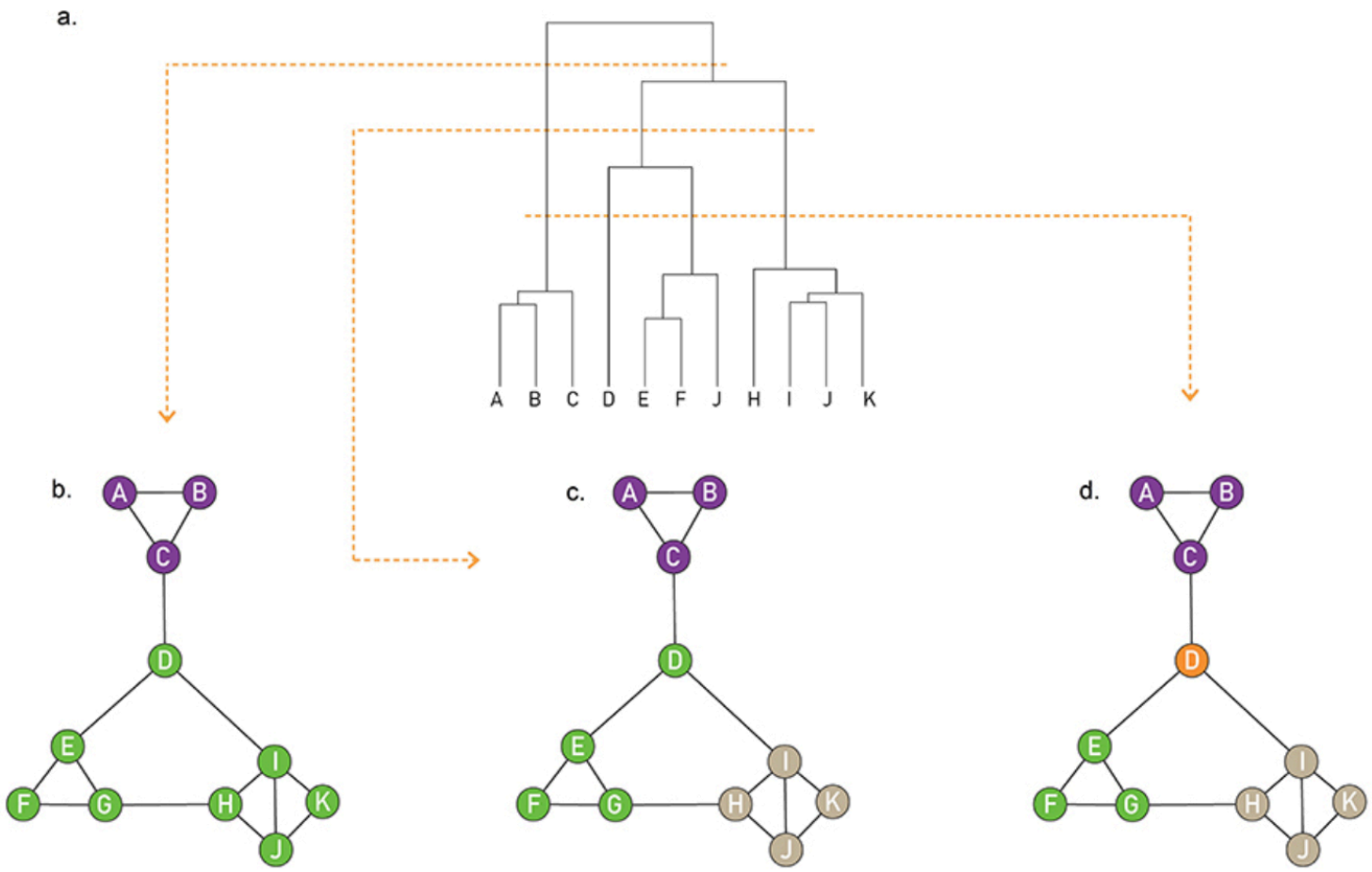
- Assign each node to a community of its own and evaluate for all node pairs. The initial similarities between these “communities” are simply the node similarities.
- Find the community pair with the highest similarity and merge them to form a single community.
- Calculate the similarity between the new community and all other communities.
- Repeat from Step 2 until all nodes are merged into a single community.

Step 4: Build Dendrogram.

- describes the precise order in which the nodes are assigned to communities.



For example, the dendrogram of Figure 9.13B tells us that the algorithm first merged nodes A and B, K and J and E and F, as each pair has . Next node C was added to the (A, B) community; I to (K, J) and G to (E, F). Eventually this procedure correctly identified the three obvious communities (ABC, EFG, and HIJK). The dendrogram also captures the fact that the EFG and the HIJK communities are closer to each other that to the ABC module.



Ambiguity in Hierarchical Clustering

Hierarchical clustering does not tell us where to cut a dendrogram. Indeed, depending on where we make the cut in the dendrogram of Image 9.9a, we obtain (b) two, (c) three or (d) four communities. While for a small network we can visually decide which cut captures best the underlying community structure, it is impossible to do so in larger networks.

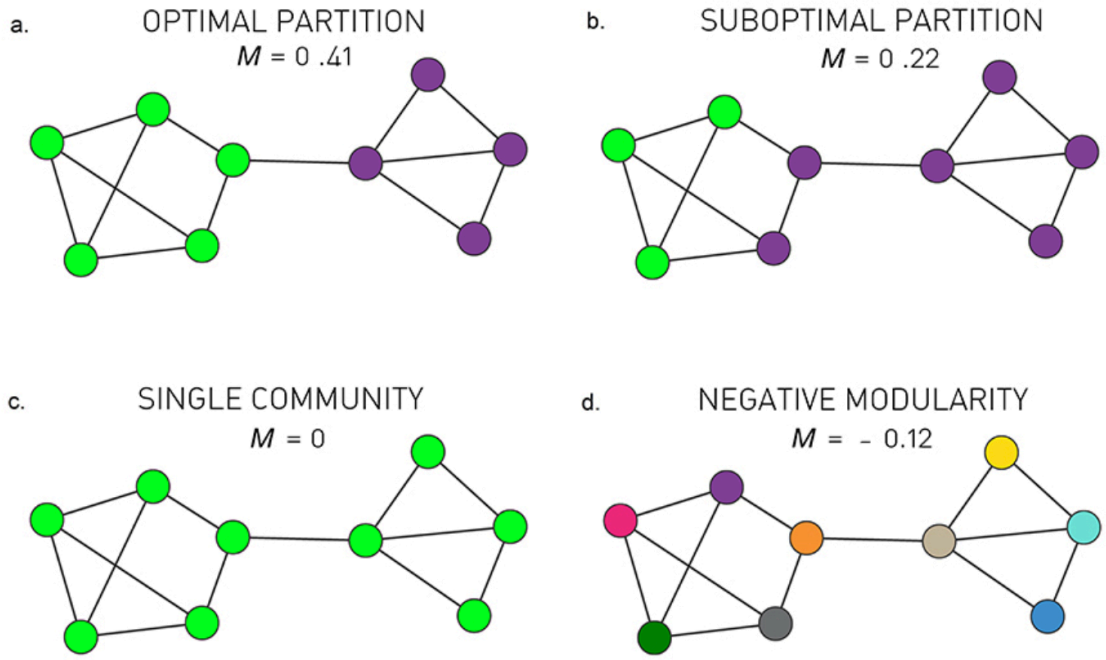
To generalize these ideas to a full network consider the complete partition that breaks the network into nc communities. To see if the local link density of the subgraphs defined by this partition differs from the expected density in a randomly wired network, we define the partition's modularity by summing (9.11) over all nc communities

Higher Modularity Implies Better Partition

The higher is M for a partition, the better is the corresponding community structure. Indeed, in Image 9.16a the partition with the maximum modularity (M=0.41) accurately captures the two obvious communities. A partition with a lower modularity clearly deviates from these communities (Image 9.16b). Note that the modularity of a partition cannot exceed one [31,32].

Zero and Negative Modularity

By taking the whole network as a single community we obtain M=0, as in this case the two terms in the parenthesis of (9.12) are equal (Image 9.16c). If each node belongs to a separate community, we have Lc=0 and the sum (9.12) has nc negative terms, hence M is negative (Image 9.16d).



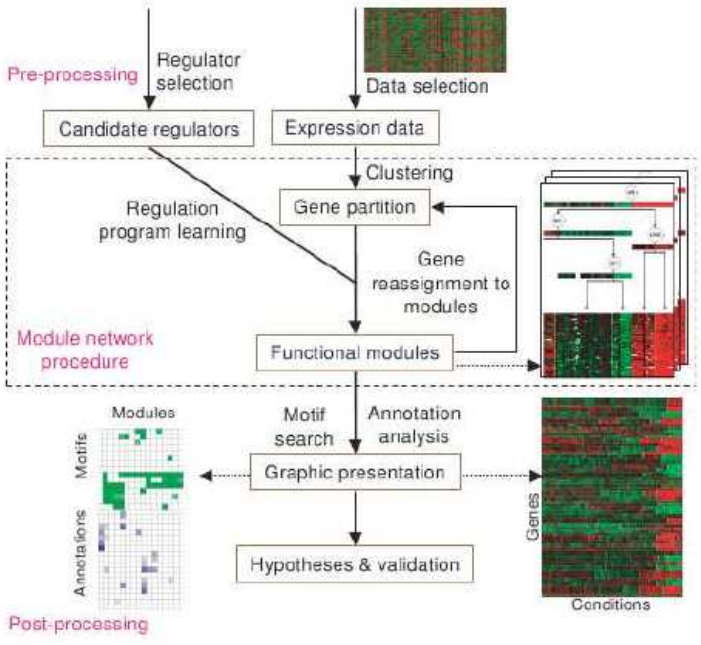
Module networks: identifying regulatory modules and their condition-specific regulators form gene expression data

- The complex functions of a living cell are carried out through the concerted activity of many genes and gene products. This activity is often coordinated by the organization of the genome into regulatory modules, or sets of coregulated genes that share a common function.
- Identifying this organization is crucial for understanding cellular responses to internal and external signals.
- Genome-wide expression profiles

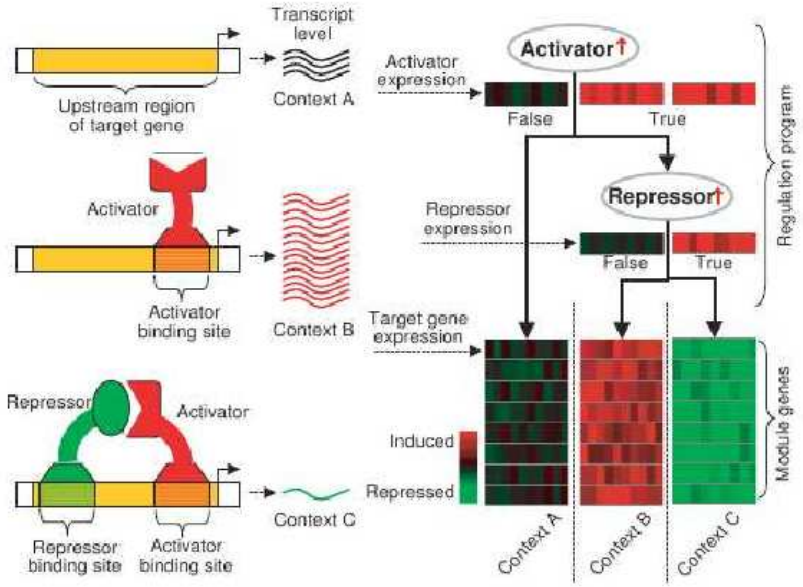
Goal Module networks procedure; a method based on probabilistic graphical models for inferring regulatory modules from gene expression data.

Assumption The regulators are themselves transcriptionally regulated, so that their expression profiles provide information about their activity level.

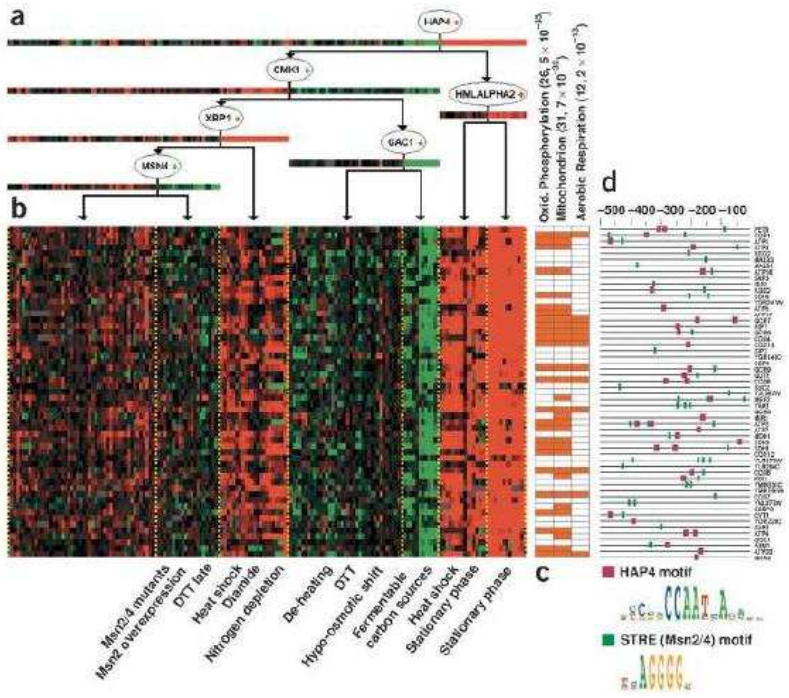
Procedure



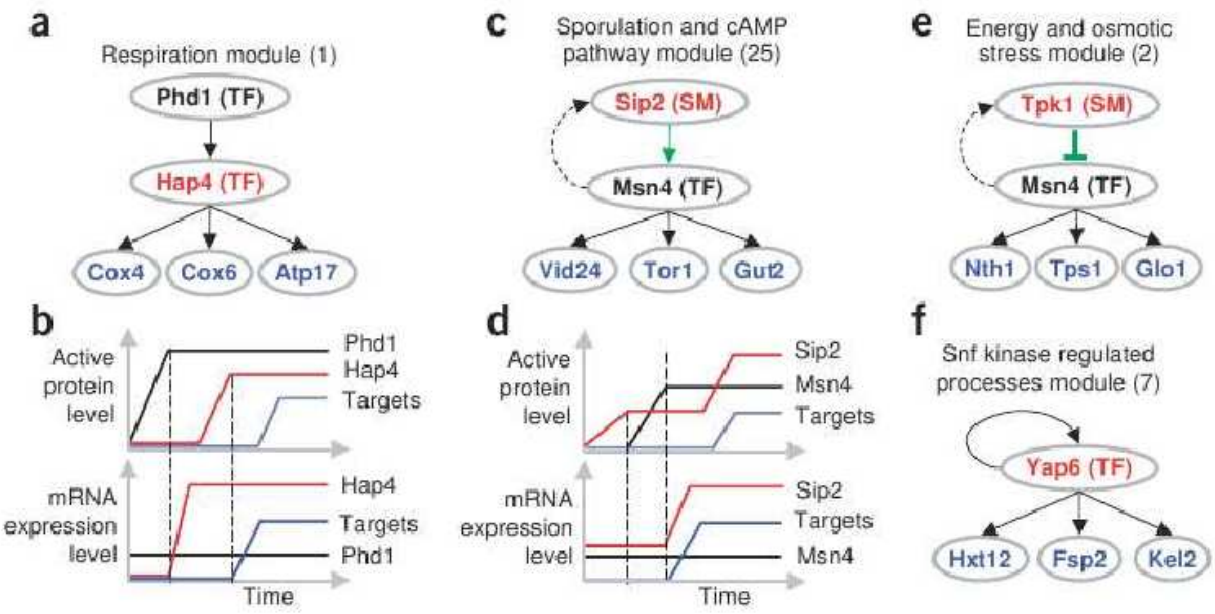
The algorithm searches simultaneously for a partition of genes into modules and for a regulation program for each module that explains the expression behavior of genes in the module. The regulation program of a module.

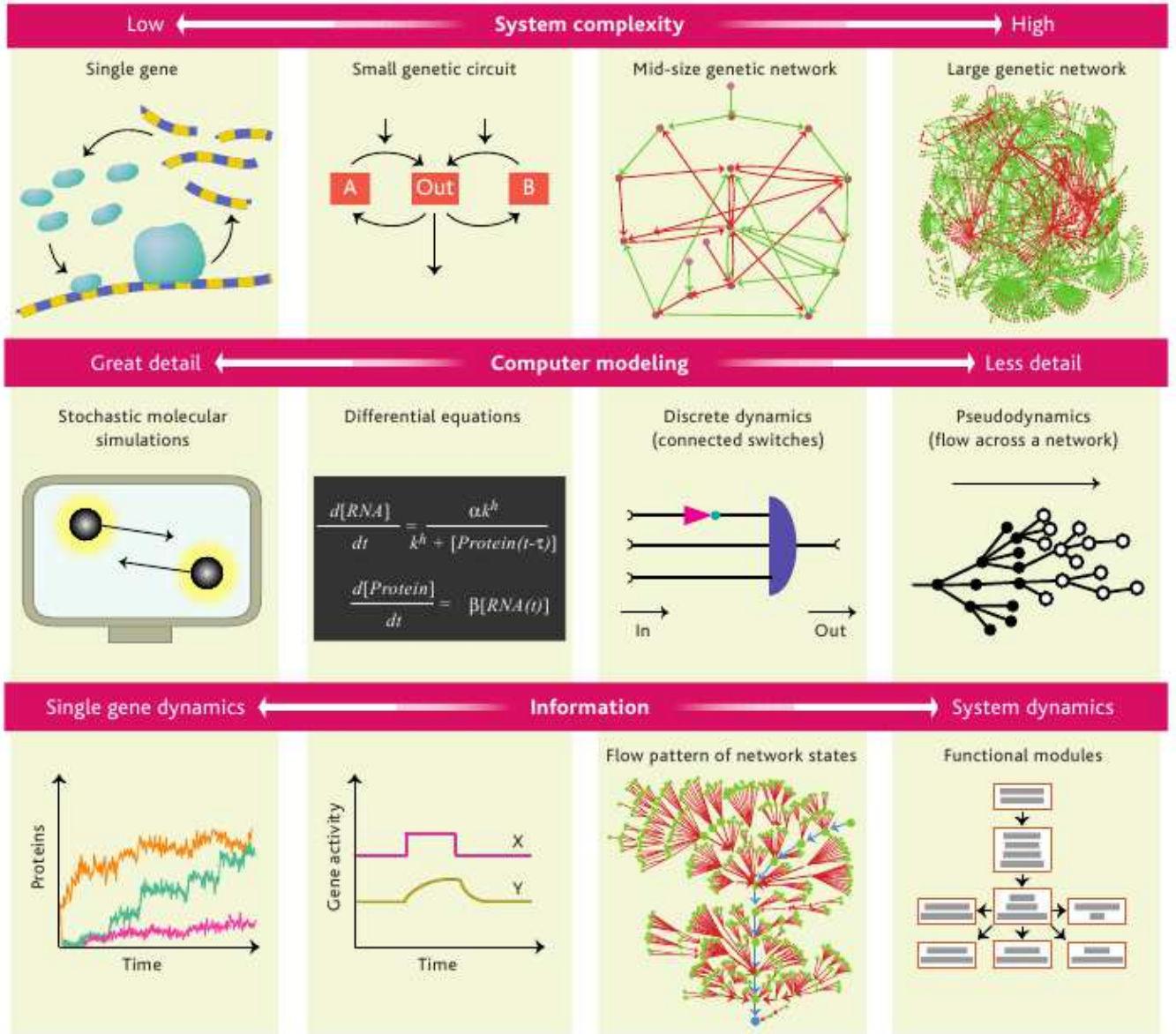


Respiration module Hap4 TF module's top regulator, indeed Hap4-DNA biding sequence motif is present in 29 of 55 genes in the module.



Regulatory components




Low
System complexity
High
Single gene
Small genetic circuit
Mid-size genetic network
Large genetic network
A
Out
B
Great detail
Computer modeling
Less detail
Stochastic molecular simulations
Differential equations
Discrete dynamics (connected switches)
Pseudodynamics (flow across a network)
$\frac{d[RNA]}{dt} = \frac{\alpha k^h}{k^h + [Protein(t-\tau)]}$
$\frac{d[Protein]}{dt} = \beta[RNA(t)]$
In
Out
Single gene dynamics
Information
System dynamics
Proteins
Time
Gene activity
X
Y
Time
Flow pattern of network states
Functional modules