Network Science

Class 2: Graph Theory

Ganesh Bagler

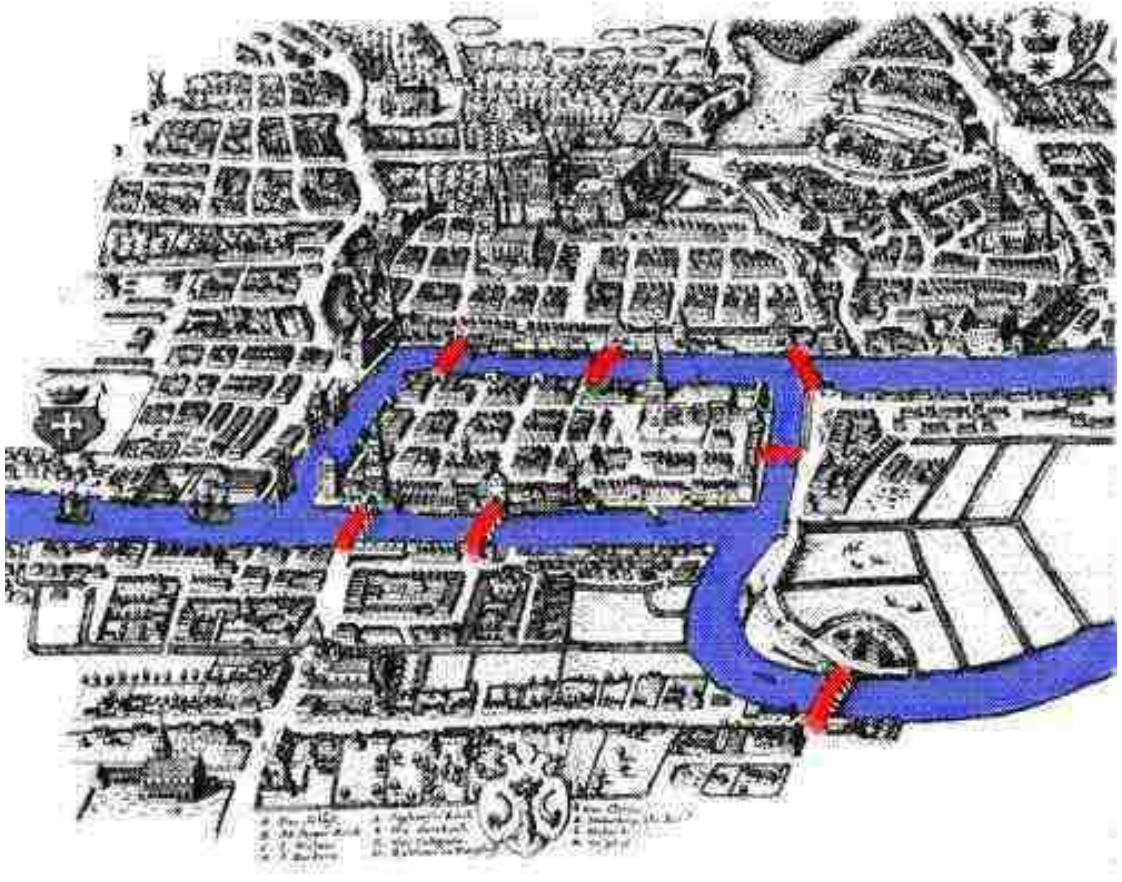
— Adapted from —

Albert-László Barabási

(With Roberta Sinatra)

The Bridges of Konigsberg

Königsberg, 1726



Königsberg Problem: Origin of Graph Theory

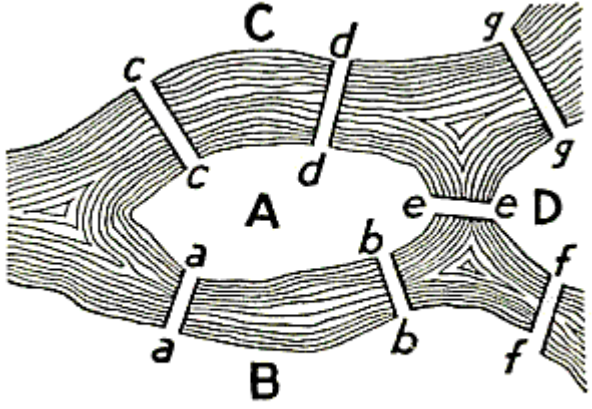
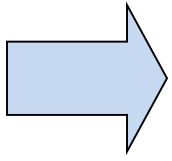
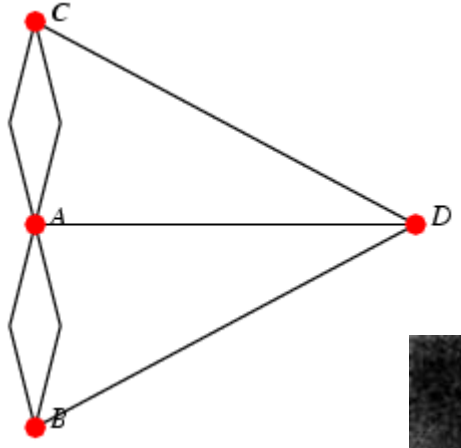
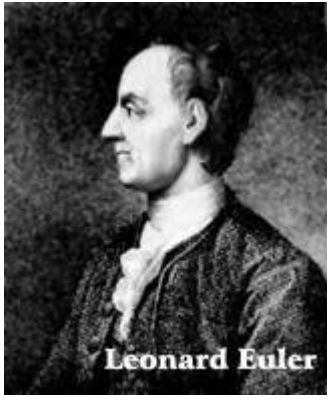


FIGURE 98. *Geographic Map: The Königsberg Bridges.*







Can one walk across the seven bridges and never cross the same one twice?

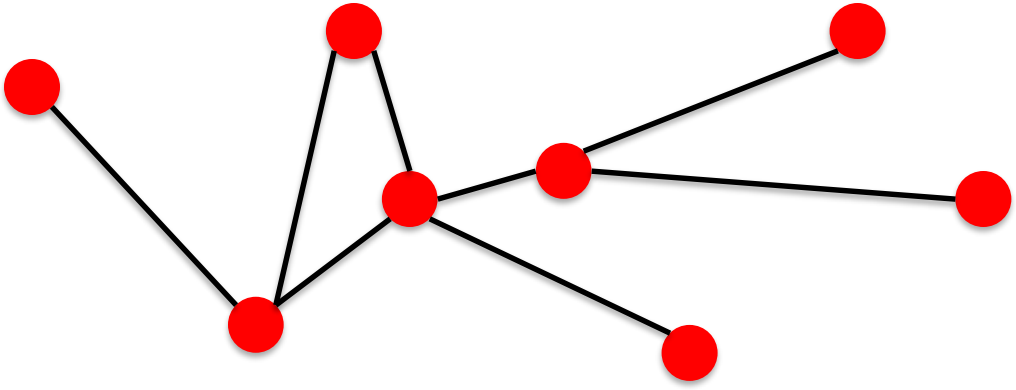
1735: **Euler's theorem:**

(a) If a graph has more than two nodes of odd degree, there is no path.

(b) If a graph is connected and has no odd degree nodes, it has at least one path.

Networks and graphs

COMPONENTS OF A COMPLEX SYSTEM



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

NETWORKS OR GRAPHS?

network often refers to real systems

- www,
- social network
- metabolic network.

Language: (Network, node, link)

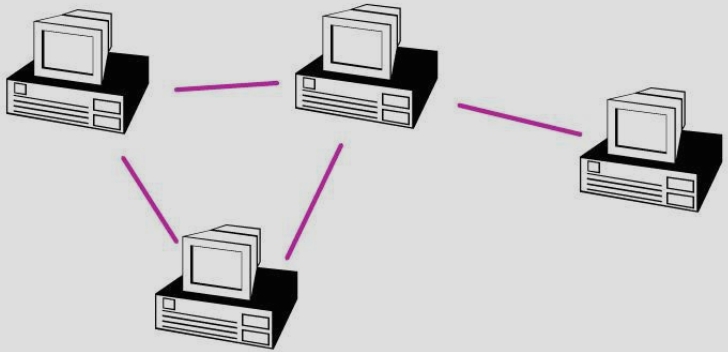
graph: mathematical representation of a network

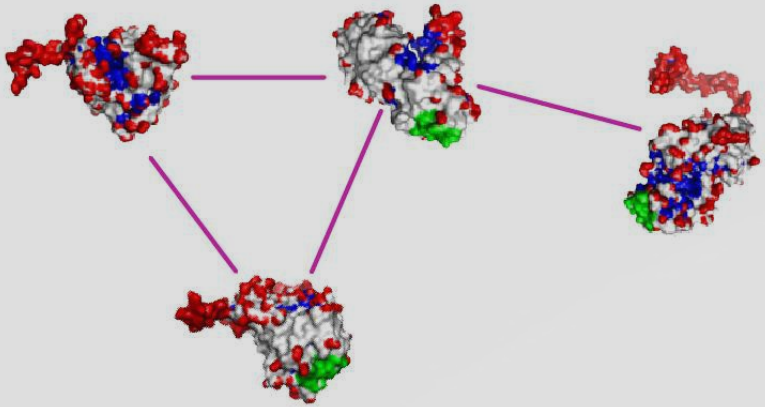
- web graph,
- social graph (a Facebook term)

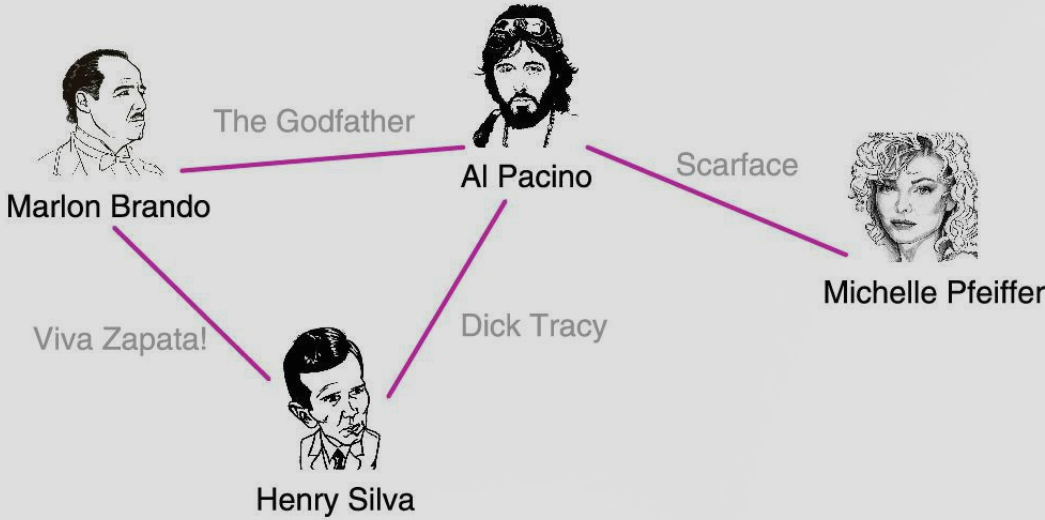
Language: (Graph, vertex, edge)

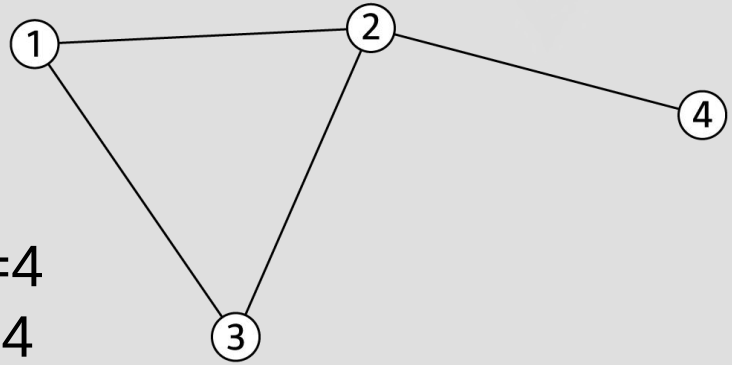
We will try to make this distinction whenever it is appropriate, but in most cases we will use the two terms interchangeably.

A COMMON LANGUAGE









N=4
L=4

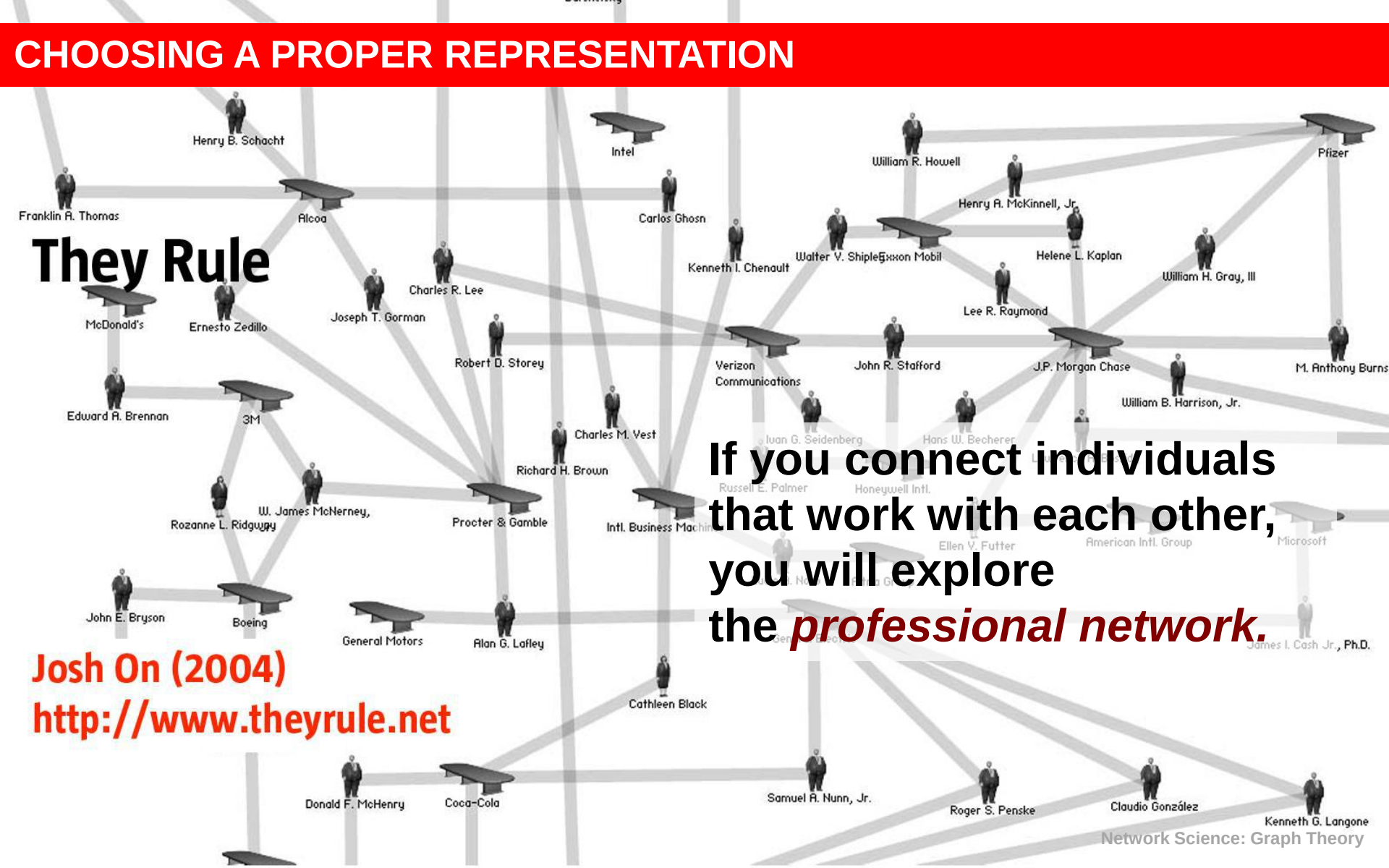
CHOOSING A PROPER REPRESENTATION

The choice of the proper network representation determines our ability to use network theory successfully.

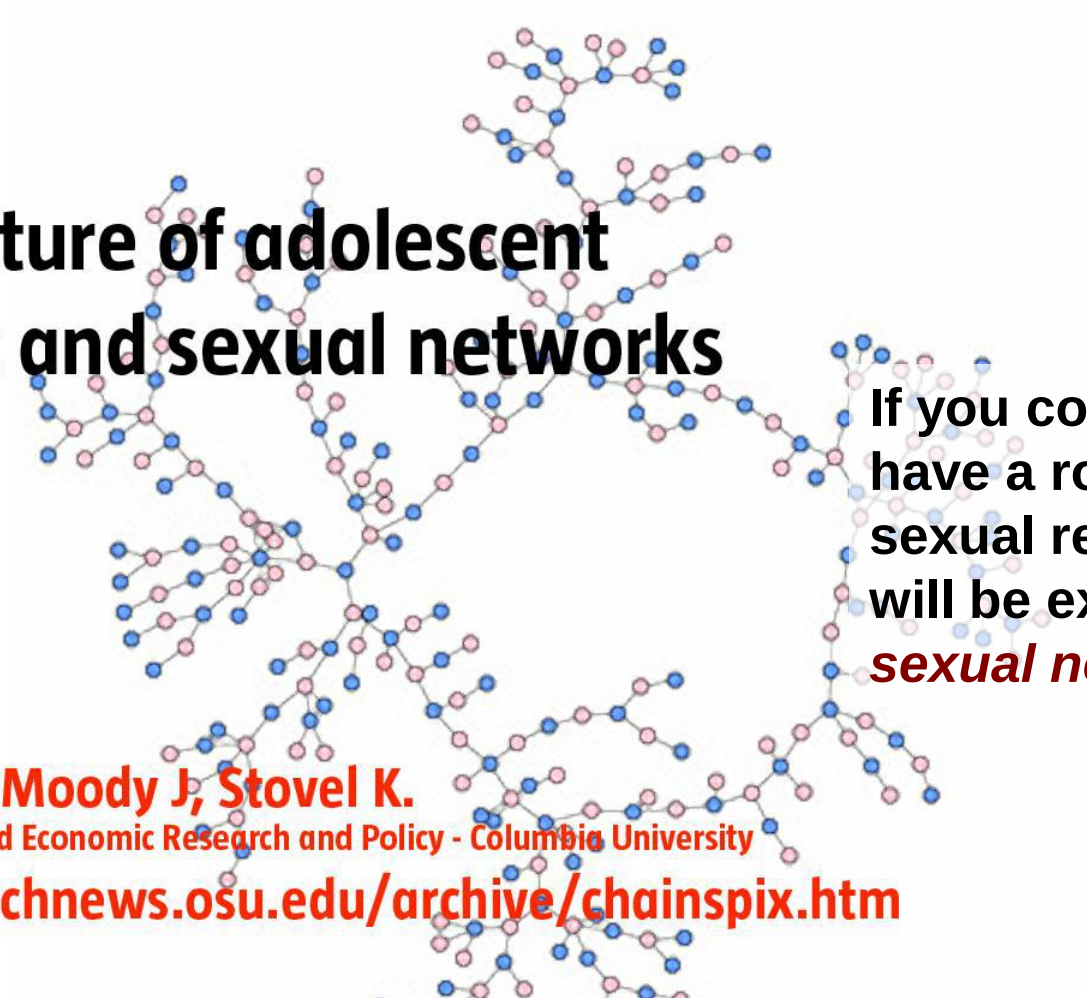
In some cases there is a unique, unambiguous representation.
In other cases, the representation is by no means unique.

For example, the way we assign the links between a group of individuals will determine the nature of the question we can study.

CHOOSING A PROPER REPRESENTATION



CHOOSING A PROPER REPRESENTATION



The structure of adolescent romantic and sexual networks

If you connect those that have a romantic and sexual relationship, you will be exploring the *sexual networks*.

Bearman PS, Moody J, Stovel K.
Institute for Social and Economic Research and Policy - Columbia University
http://researchnews.osu.edu/archive/chainspix.htm

CHOOSING A PROPER REPRESENTATION

If you connect individuals based on their first name (*all Peters connected to each other*), what will you will be exploring?

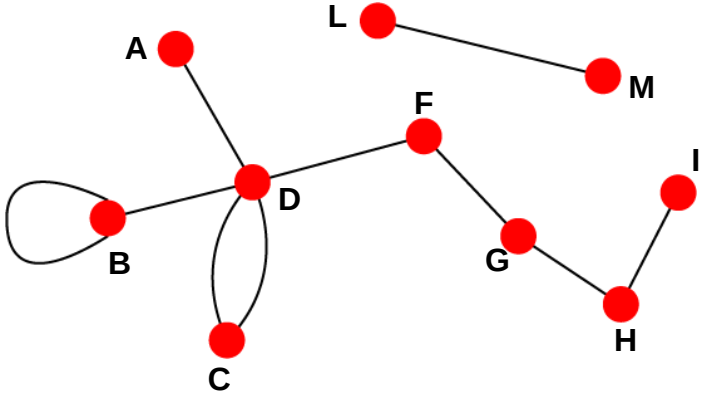
It is a network, nevertheless.

UNDIRECTED VS. DIRECTED NETWORKS

Undirected

Links: undirected (*symmetrical*)

Graph:

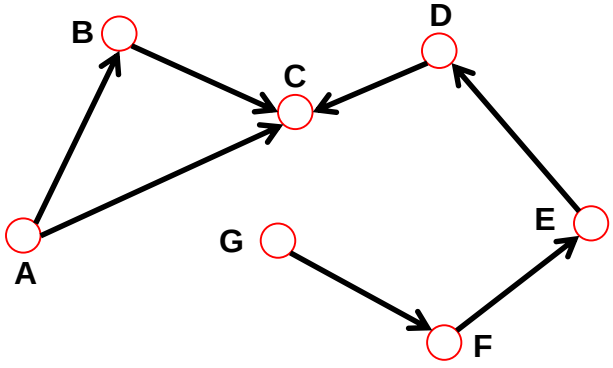


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:



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

Directed links :
URLs on the www
phone calls
metabolic reactions

Section 2.2

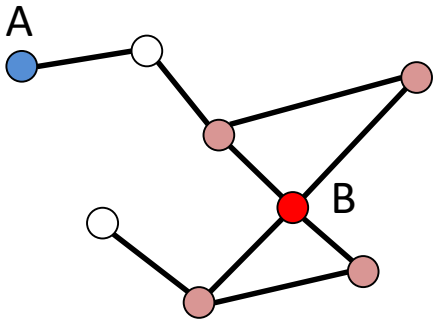
Reference Networks

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

Degree, Average Degree and Degree Distribution

NODE DEGREES

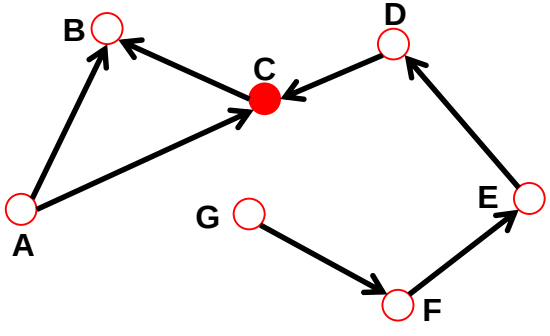
Undirected



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 $k^{out} = 0$.

A BIT OF STATISTICS

BRIEF STATISTICS REVIEW

Four key quantities characterize a sample of N values $x_1, \ldots, x_N$:

Average (mean):

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

The n^{th} moment:

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i^n$$

Standard deviation:

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(x_i - \langle x \rangle \right)^2}$$

Distribution of x:

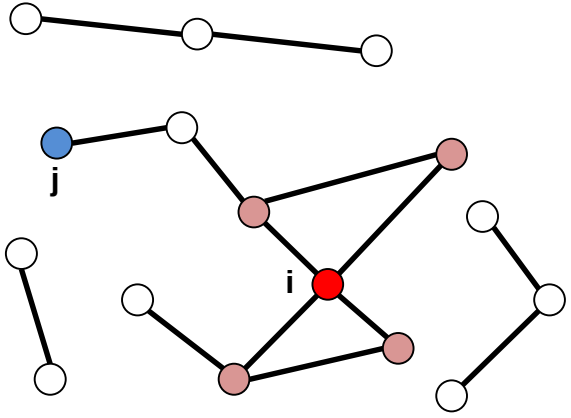
$$p_x = \frac{1}{N} \sum_i \delta_{x,x_i}$$

where p_x follows

$$\sum_i p_x = 1 \quad \left(\int p_x \, dx = 1 \right)$$

AVERAGE DEGREE

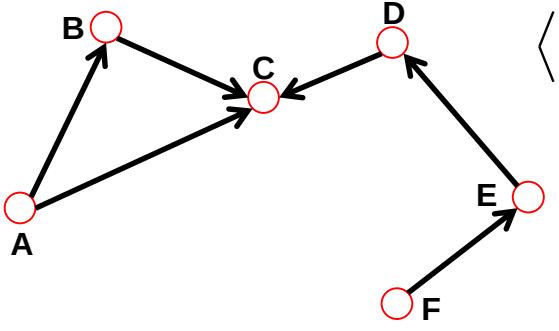
Undirected



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

N – the number of nodes in the graph

Directed



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

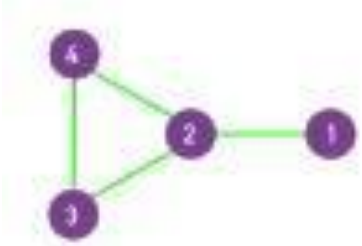
$$\langle k \rangle = \frac{L}{N}$$

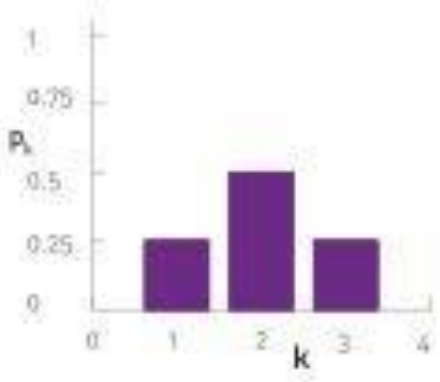
Average Degree

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

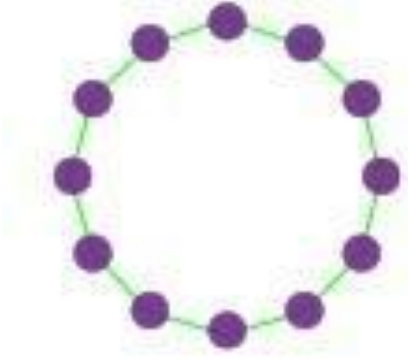
P(k): probability that a randomly chosen node has degree *k*

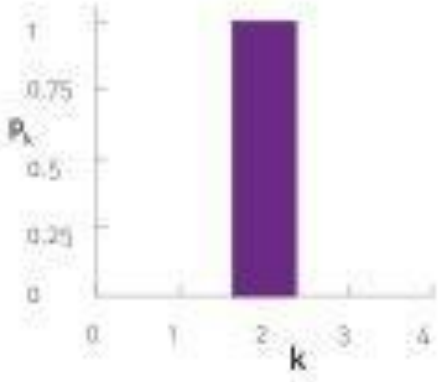




N_k = # nodes with degree k

P(k) = N_k / N ❾ **plot**





DEGREE DISTRIBUTION

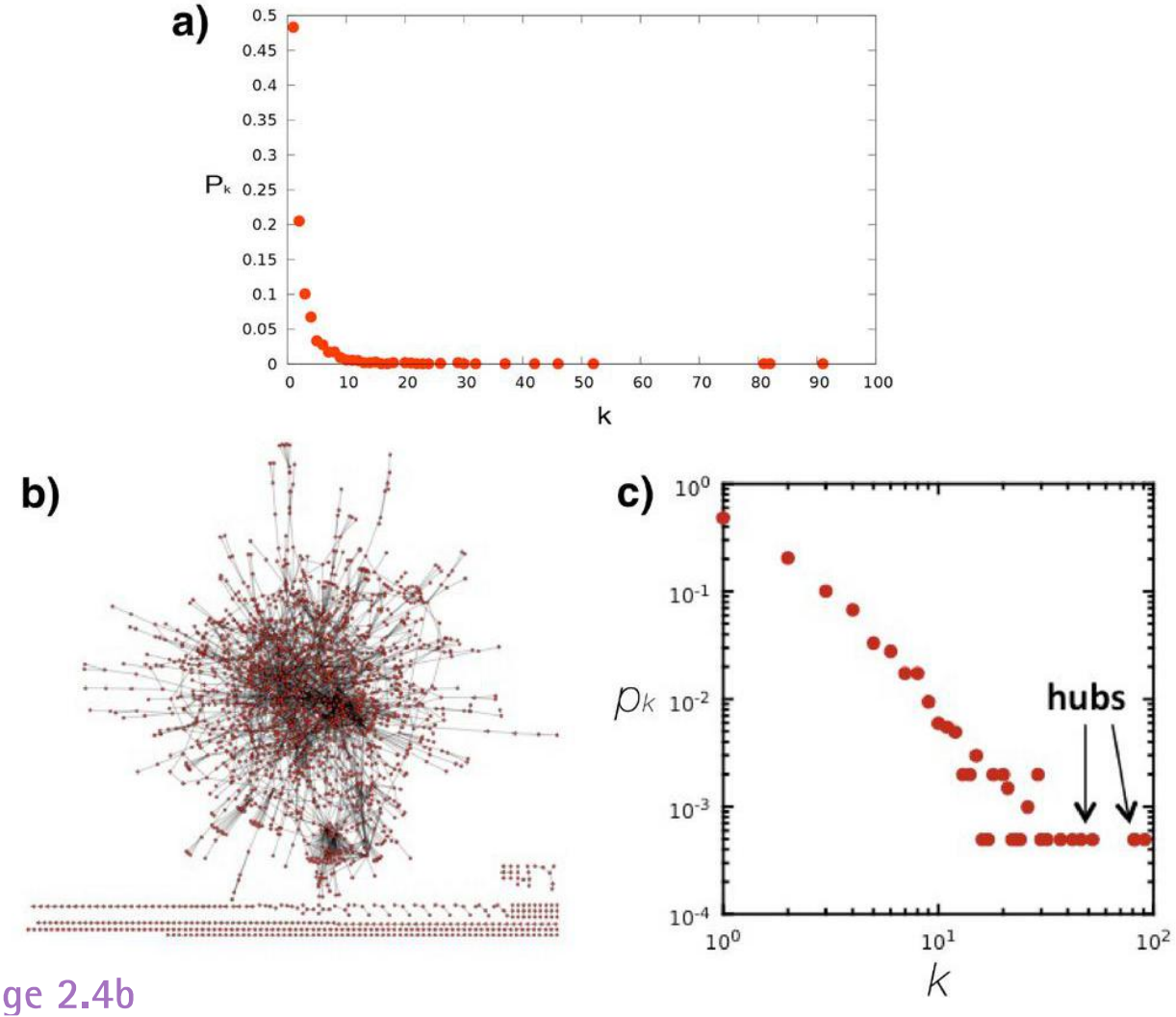


Image 2.4b

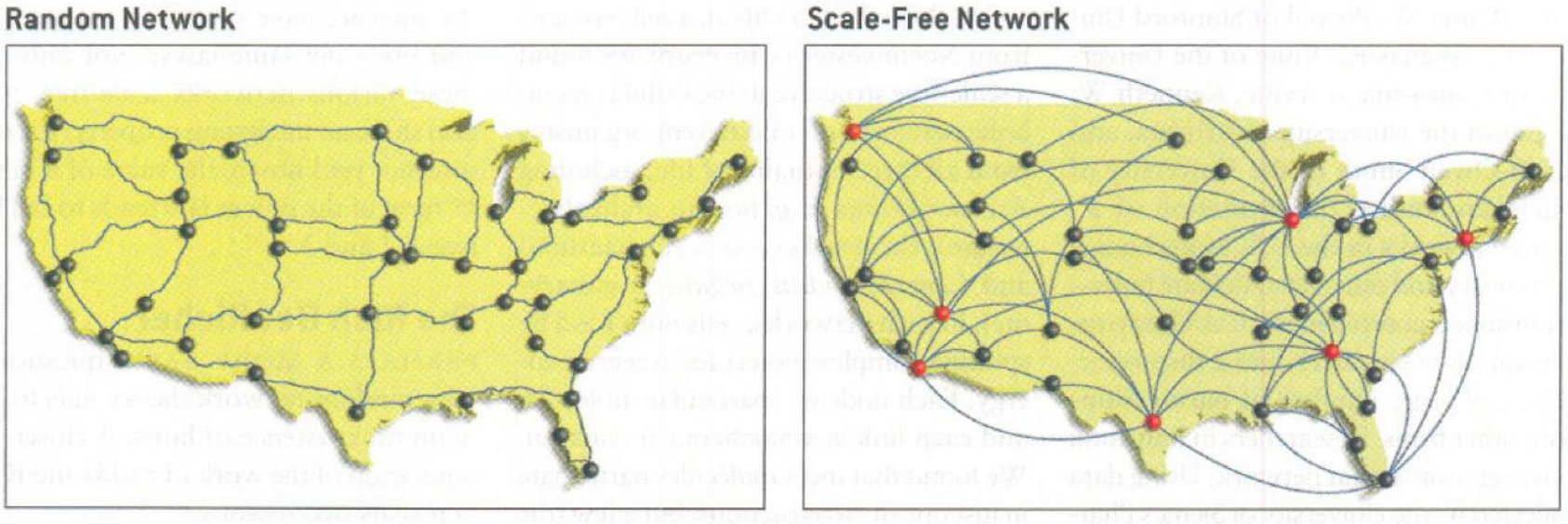
DEGREE DISTRIBUTION

Scale-Free Networks

Scientists have recently discovered that various complex systems have an underlying architecture governed by shared organizing principles. This insight has important implications for a host of applications, from drug development to Internet security

BY ALBERT-LÁSZLÓ BARABÁSI AND ERIC BONABEAU

DEGREE DISTRIBUTION



DEGREE DISTRIBUTION

Overview/*Scale-Free Networks*

- A variety of complex systems share an important property: some nodes have a tremendous number of connections to other nodes, whereas most nodes have just a handful. The popular nodes, called hubs, can have hundreds, thousands or even millions of links. In this sense, the network appears to have no scale.
- Scale-free networks have certain important characteristics. They are, for instance, robust against accidental failures but vulnerable to coordinated attacks.
- Understanding of such characteristics could lead to new applications in many arenas. For example, computer scientists might be able to devise more effective strategies for preventing computer viruses from crippling a network such as the Internet.

DEGREE DISTRIBUTION

The Potential Implications of Scale-Free Networks for . . .

Computing

- Computer networks with scale-free architectures, such as the World Wide Web, are highly resistant to accidental failures. But they are very vulnerable to deliberate attacks and sabotage.
- Eradicating viruses, even known ones, from the Internet will be effectively impossible.

Medicine

- Vaccination campaigns against serious viruses, such as smallpox, might be most effective if they concentrate on treating hubs—people who have many connections to others. But identifying such individuals can be difficult.
- Mapping out the networks within the human cell could aid researchers in uncovering and controlling the side effects of drugs. Furthermore, identifying the hub molecules involved in certain diseases could lead to new drugs that would target those hubs.

Business

- Understanding how companies, industries and economies are interlinked could help researchers monitor and avoid cascading financial failures.
- Studying the spread of a contagion on a scale-free network could offer new ways for marketers to propagate consumer buzz about their products.

Role of Degree Distributions

- Degree distributions dictate the topological and dynamical properties of the network.

- Scale-free networks
  - Heterogeneous degree distributions
  - Importance of 'hubs'
  - Hubs as control elements
  - Structural integrity vs. dynamic integrity

- Scale-free distributions and averages

"Don't cross the river if it on an average only 5 feet deep."

- *Mediocristan vs. Extremistan* (NN Taleb)
  - How to become fat by eating heavily on one day?
  - How to get rich in one day?

DEGREE DISTRIBUTION

Discrete Representation: $\mathbf{p_k}$ is the probability that a node has degree **k**.

Continuum Description: **p(k)** is the pdf of the degrees, where

$$\int_{k_1}^{k_2} p(k)\,dk$$

represents the probability that a node's degree is between $\mathbf{k_1}$ and $\mathbf{k_2}$.

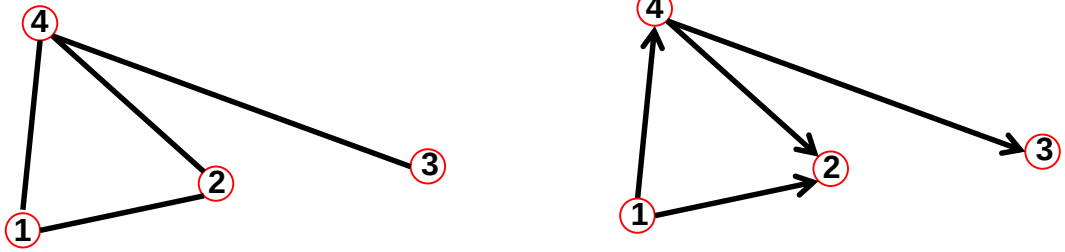
Normalization condition:

$$\sum_{0}^{\infty} p_k = 1 \qquad \int_{K_{\min}}^{\infty} p(k)\,dk = 1$$

where K_{min} is the minimal degree in the network.

Adjacency matrix

ADJACENCY MATRIX



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.

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

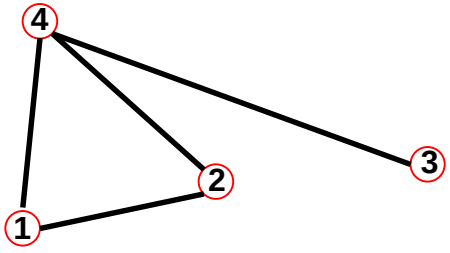
Note that for a directed graph (right) the matrix is not symmetric.

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

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

ADJACENCY MATRIX AND NODE DEGREES

Undirected



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

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

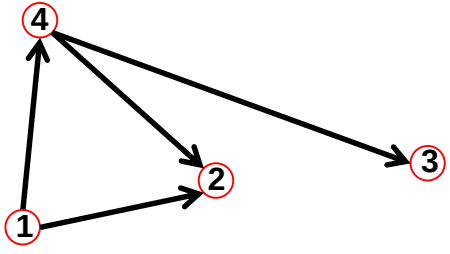
$$k_j = \sum_{i=1}^{N} A_{ij}$$

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

$$A_{ii} = 0$$

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

Directed



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

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

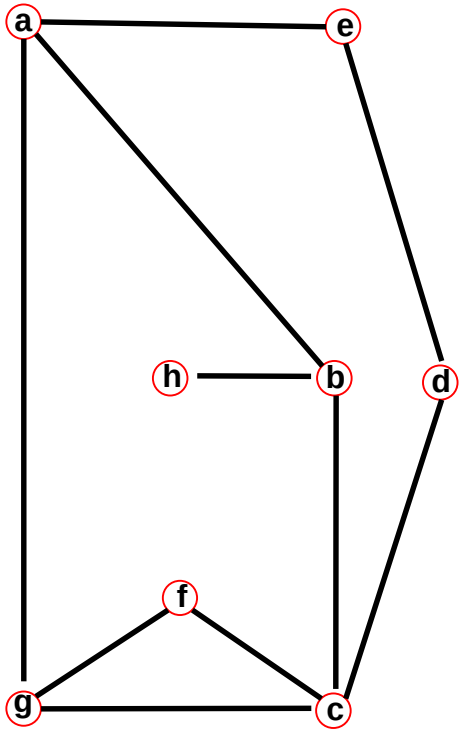
$$k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

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

$$A_{ii} = 0$$

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

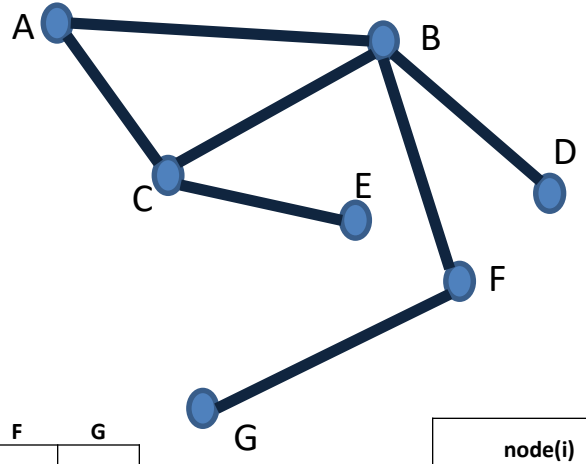
ADJACENCY MATRIX



	a	b	c	d	e	f	g	h
a	0	1	0	0	1	0	1	0
b	1	0	1	0	0	0	0	1
c	0	1	0	1	0	1	1	0
d	0	0	1	0	1	0	0	0
e	1	0	0	1	0	0	0	0
f	0	0	1	0	0	0	1	0
g	1	0	1	0	0	0	0	0
h	0	1	0	0	0	0	0	0

Numerical Representation of a Graph

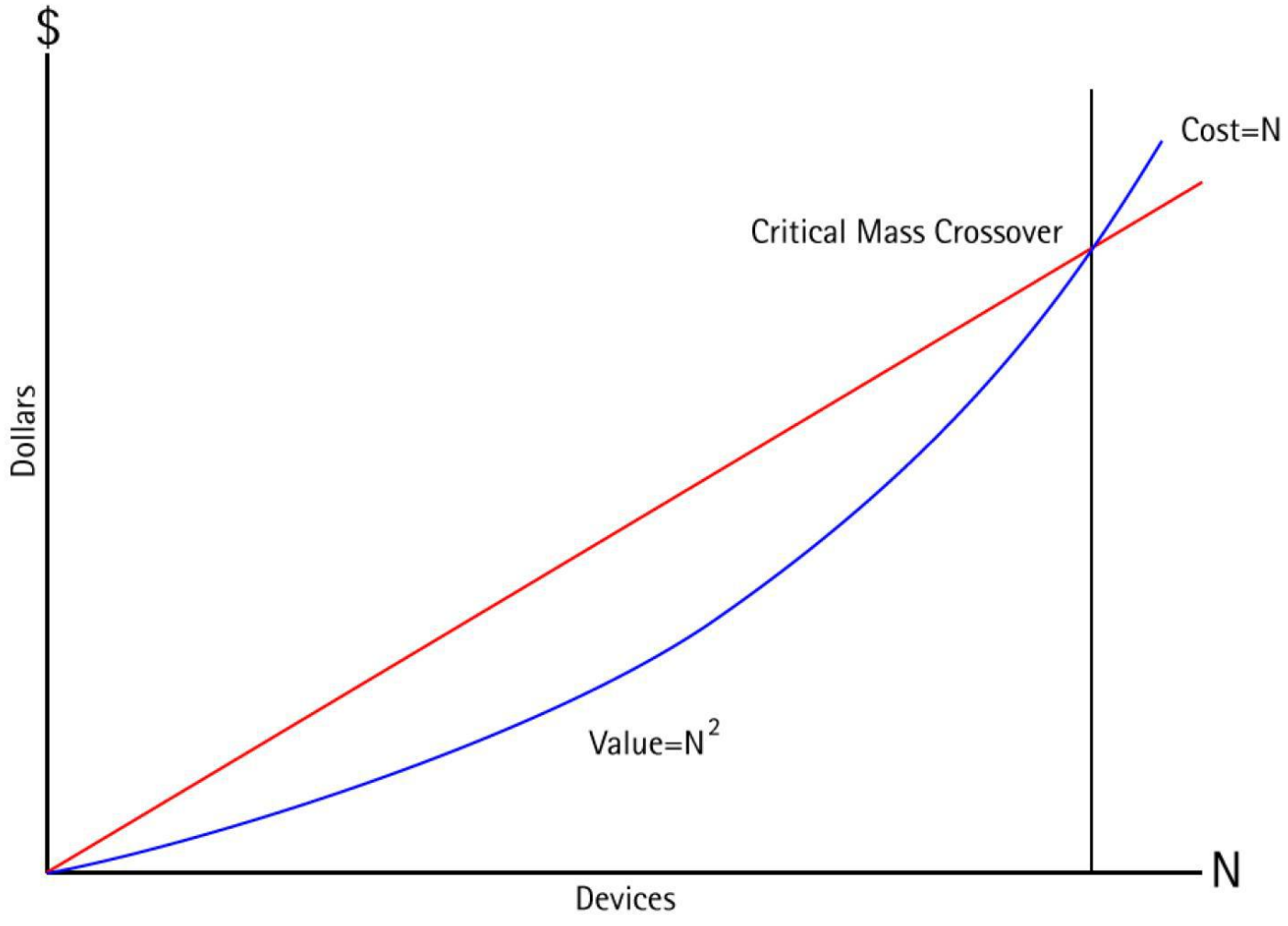
Adjacency matrix and edge list are two of the important numerical (computational) representations of a graph/network.



	A	B	C	D	E	F	G
A		■	■				
B	■		■	■		■	
C	■	■			■		
D		■					
E			■				
F		■					■
G						■	

node(i)	node(j)
A	B
A	C
B	C
B	D
B	F
C	E
F	G

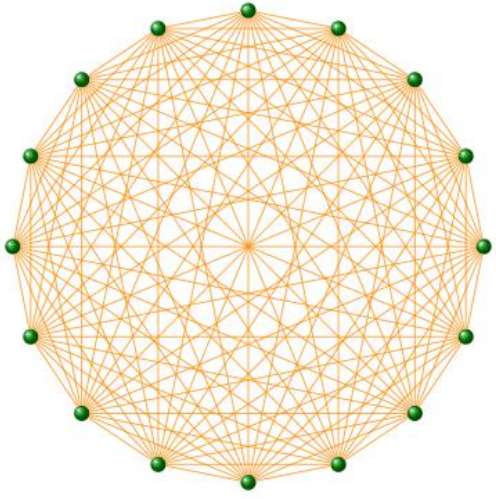
Metcalfe's Law



Real networks are sparse

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 **<k>=N-1**

REAL NETWORKS ARE SPARSE

Most networks observed in real systems are sparse:

$$L \ll L_{max}$$

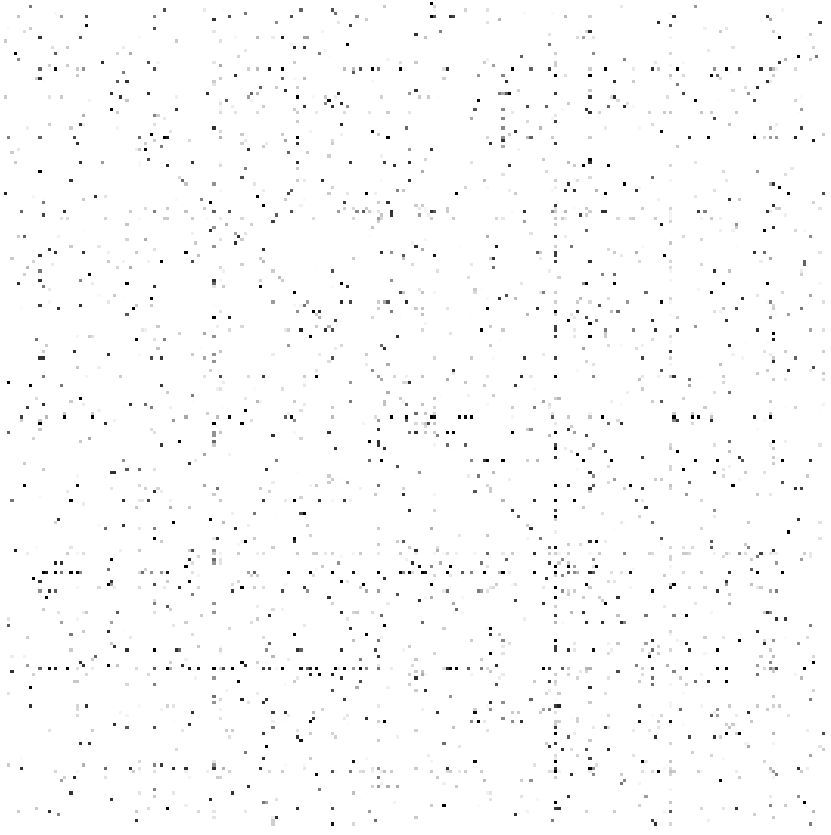
or

$$\langle k \rangle \ll N-1.$$

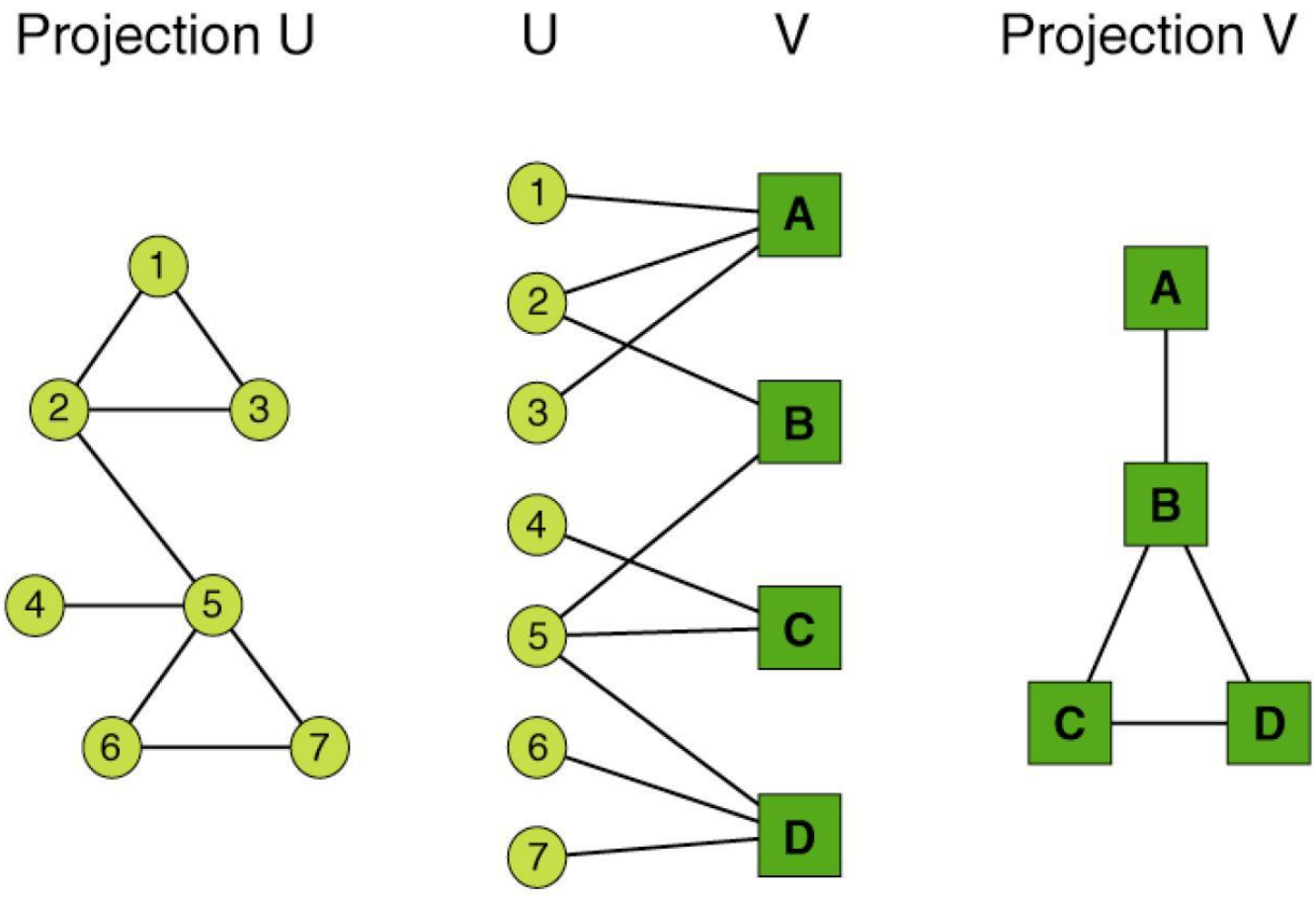
WWW (ND Sample):	N=325,729;	$L=1.4\ 10^6$	$L_{max}=10^{12}$	<k>=4.51
Protein (*S. Cerevisiae*):	N= 1,870;	L=4,470	$L_{max}=10^7$	<k>=2.39
Coauthorship (Math):	N= 70,975;	$L=2\ 10^5$	$L_{max}=3\ 10^{10}$	<k>=3.9
Movie Actors:	N=212,250;	$L=6\ 10^6$	$L_{max}=1.8\ 10^{13}$	<k>=28.78

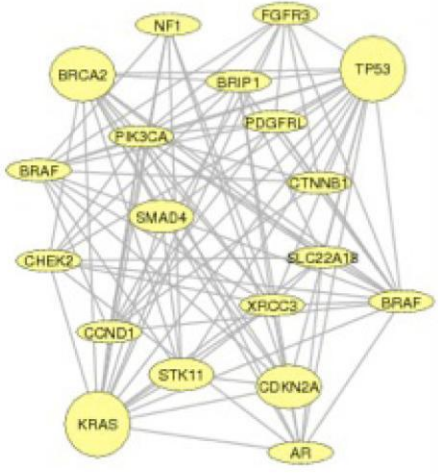
(Source: Albert, Barabasi, RMP2002)

ADJACENCY MATRICES ARE SPARSE

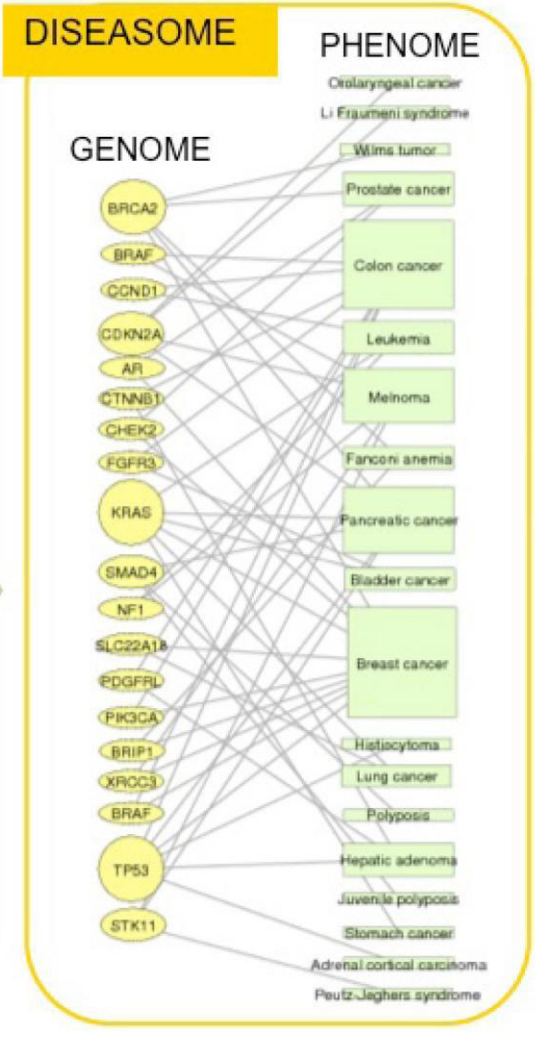


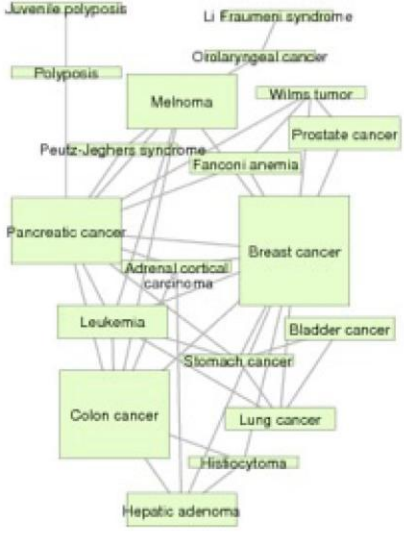
Bipartite Graphs





Gene network





Disease network

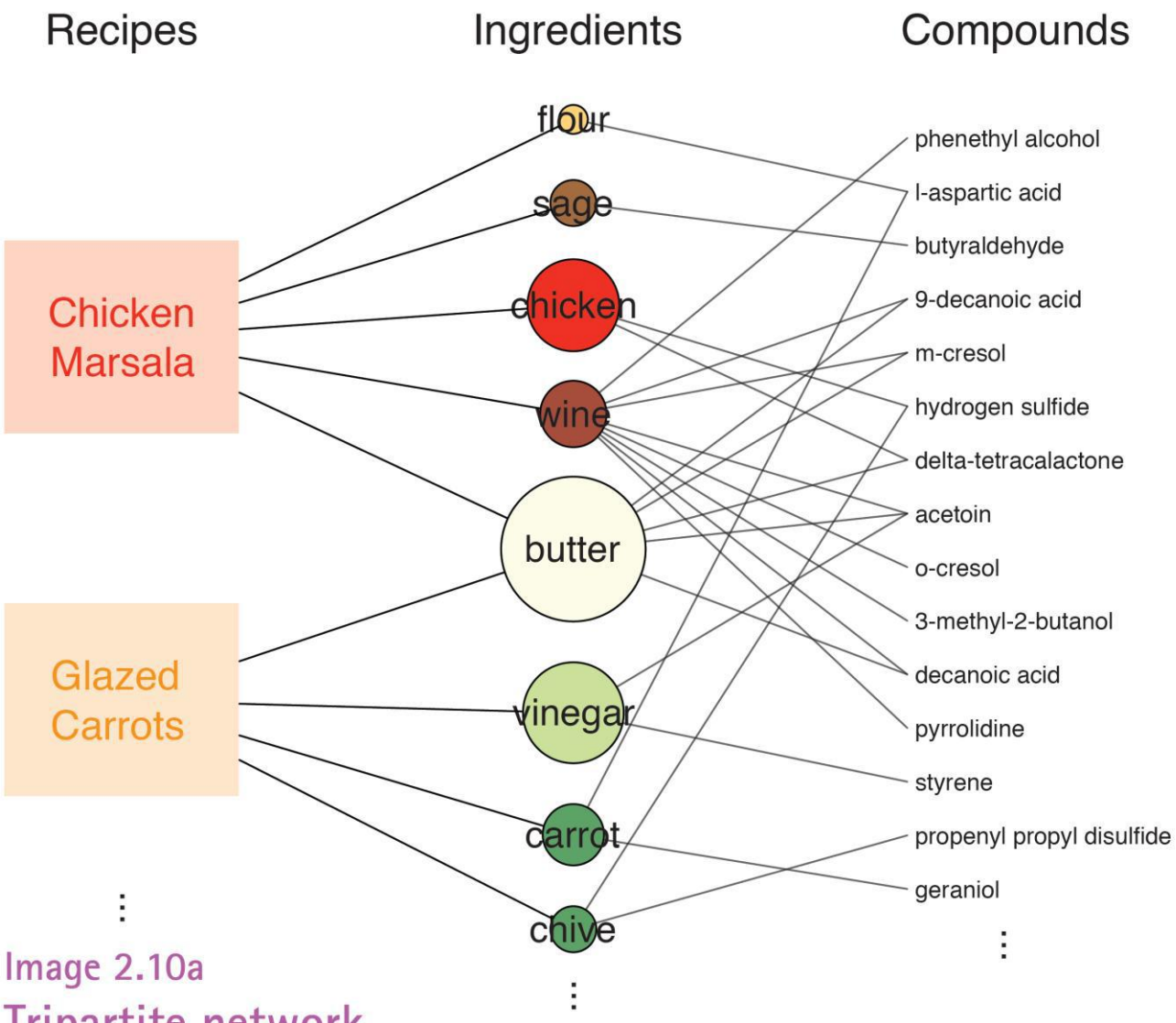


Image 2.10a

Tripartite network.

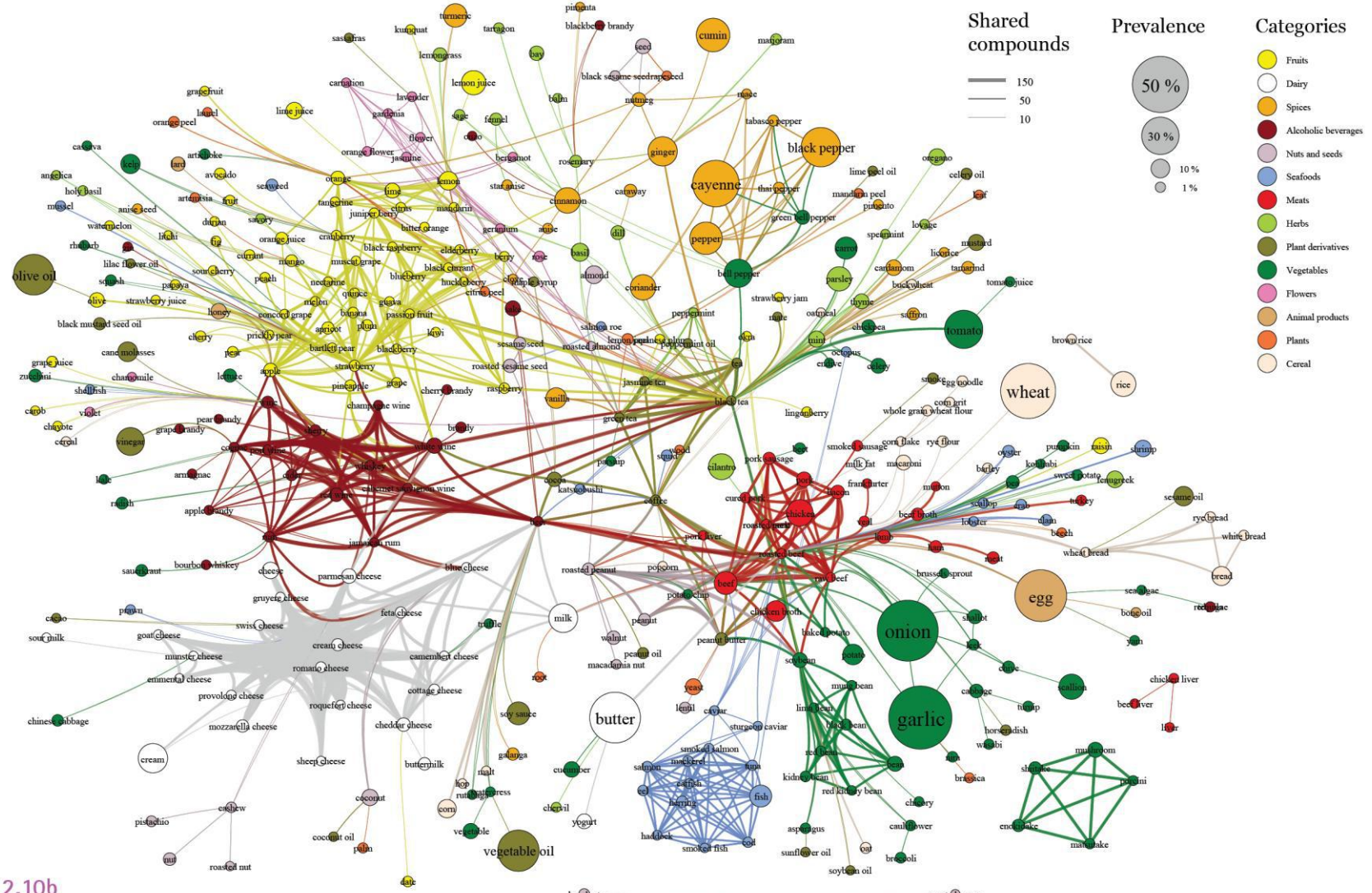
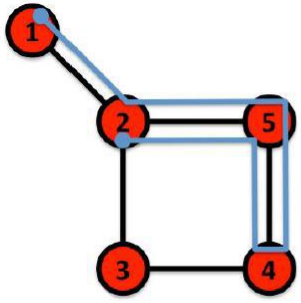


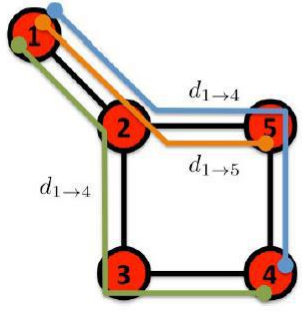
Image 2.10b

Tripartite network.



Path

PATH: A sequence of nodes such that each node is connected to the next node along the path by a link. A path always consists of *n* nodes and *n - 1* links. The length of a path is defined as the number of its links, counting multiple edges multiple times.

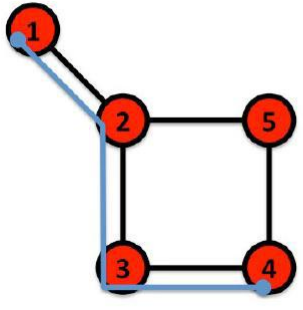


Shortest Path

$d_{1\to4} = 3$

$d_{1\to5} = 2$

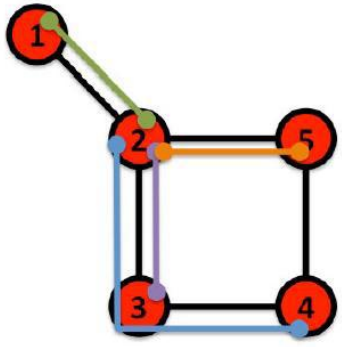
SHORTEST PATH (geodesic path, *d*): the path with the shortest distance *d* between two nodes. We will call it the distance between two nodes.



Diameter

$d_{1\to4} = 3$

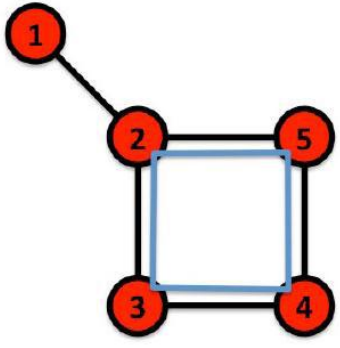
DIAMETER (d_{max}): the longest shortest path in a graph, or the distance between the two furthest away nodes.



Average Path Length

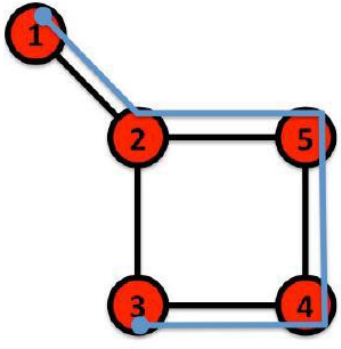
$$(d_{1\to2} + d_{1\to3} + d_{1\to4} + \\ + d_{1\to5} + d_{2\to3} + d_{2\to4} + \\ + d_{2\to5} + d_{3\to4} + d_{3\to5} + \\ + d_{4\to5})/10 = 1.6$$

AVERAGE PATH LENGTH (‹*d*›): the average of the shortest paths between all pairs of nodes.



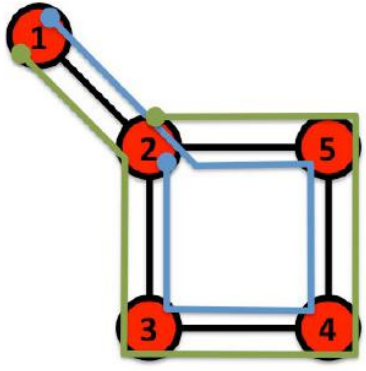
Cycle

CYCLE: a path with the same start and end node.



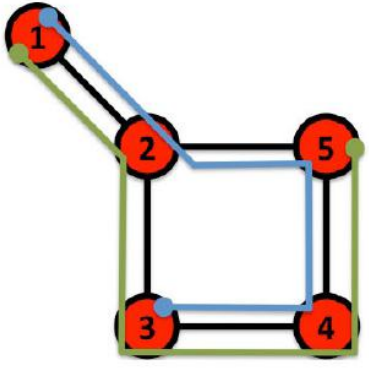
Self-avoiding Path

SELF-AVOIDING PATH: a path that does not intersect itself, i.e. the same node or link does not occur twice along the path.



Eulerian Path

EULERIAN PATH: a path that traverses each link exactly once.



Hamiltonian Path

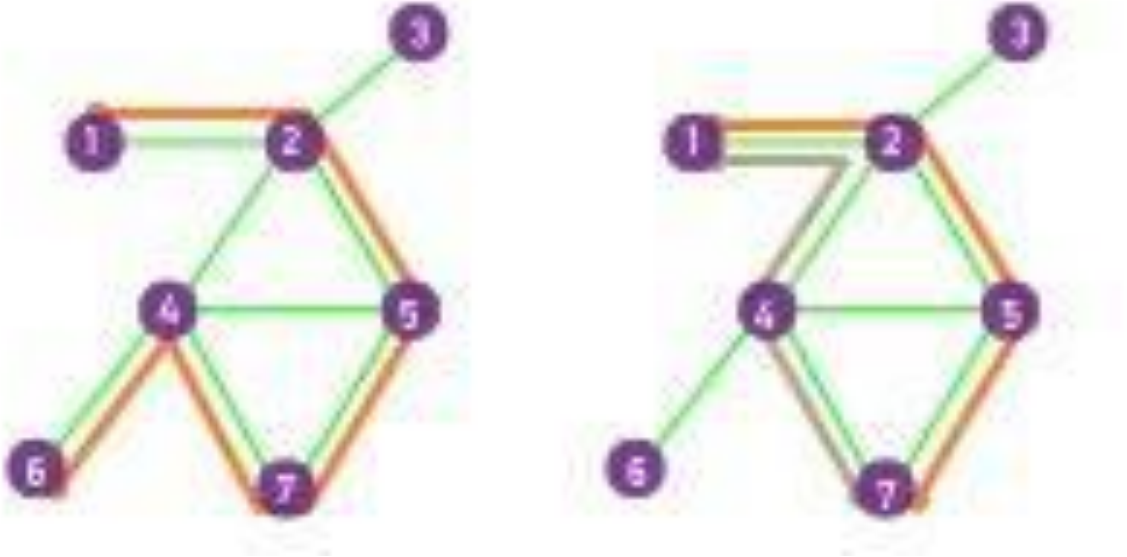
HAMILTONIAN PATH: a path that visits each node exactly once.

PATHS

A *path is* a sequence of nodes in which each node is adjacent to the next one

$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

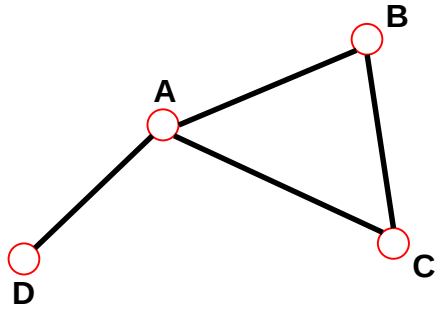
$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$



- In a directed network, the path can follow only the direction of an arrow.

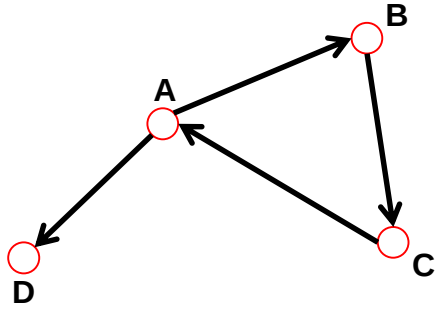
DISTANCE IN A GRAPH

Shortest Path, Geodesic Path



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).

Number of shortest paths between two nodes.

The number of shortest paths, N_{ij}, between nodes *i* and *j* and the distance d_{ij} between them can be determined directly from the adjacency matrix, A_{ij}.

- $d_{ij} = 1$: If there is a link between *i* and *j*, then $A_{ij} = 1$ ($A_{ij} = 0$ otherwise).

- $d_{ij} = 2$: If there is a path of length two between *i* and *j*, then the product of d elements $A_{ik} A_{kj} = 1$ ($A_{ik} A_{kj} = 0$ otherwise). The number of $d_{ij} = 2$ paths between *i* and *j* is

$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik} A_{kj} = \left[A^2 \right]_{ij} \qquad (16)$$

  where [...]*ij* denotes the $(ij)^{th}$ element of a matrix.

- $d_{ij} = d$: If there is a path of length d between *i* and *j*, then $A_{ik} \ldots A_{lj} = 1$ ($A_{ik} \ldots A_{lj} = 0$ otherwise). The number of paths of length d between i and j is

$$N_{ij}^{(d)} = \left[A^d \right]_{ij} \quad . \qquad (17)$$

Equation (17) holds for both directed and undirected networks and can be generalized to multigraphs as well. The distance between nodes i and j is the path with the smallest d for which $N_{ij}^{(d)} > 0$. Despite the mathematical elegancy of Eq. (17), faced with a large network, it is more efficient to use the breadth-first-search algorithm described in Box 2.6.

Box 2.5

N_{ij},number of paths between any two nodes *i* and *j*:

***Length n=1*:** If there is a link between *i* and *j*, then $A_{ij}=1$ and $A_{ij}=0$ otherwise.

Length n=2: If there is a path of length two between *i* and *j*, then $A_{ik}A_{kj}=1$, and $A_{ik}A_{kj}=0$ otherwise.

The number of paths of length 2:

$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik}A_{kj} = [A^2]_{ij}$$

Length n: In general, if there is a path of length *n* between *i* and *j*, then $A_{ik}\ldots A_{lj}=1$ and $A_{ik}\ldots A_{lj}=0$ otherwise.

The number of paths of length *n* between *i* and *j* is*

$$N_{ij}^{(n)} = [A^n]_{ij}$$

**holds for both directed and undirected networks.*

Finding the shortest path: breath first search.

BFS is one of the most frequently used algorithms in network science. Similar to throwing a pebble in a pond and watching the ripples spread from the center, we start from a node and label its neighbors, then the neighbors' neighbors, until we encounter the target node. The number of "*ripples*" needed to reach the target provides the distance. To be specific, the identification of the shortest path between node i and j follows the following steps (Gallery 2.5):

1. Start at node *i*.

2. Find the nodes directly linked to *i*. Label them distance "1" and put them in a queue.

3. 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.

4. Repeat step 3 until you find the target node j or there are no more nodes in the queue.

5. The distance between *i* and *j* is the label of *j*. If j does not have a label, then $d_{ij} = \infty$.

The time complexity of the BFS algorithm, representing the approximate number of steps the computer needs to find dij on a network of *N* nodes and *L* links, is O (*N + L*). It is linear in *N* and *L* as each node needs to be entered and removed from the queue at most once, and each link has to be tested only once.

Box 2.6

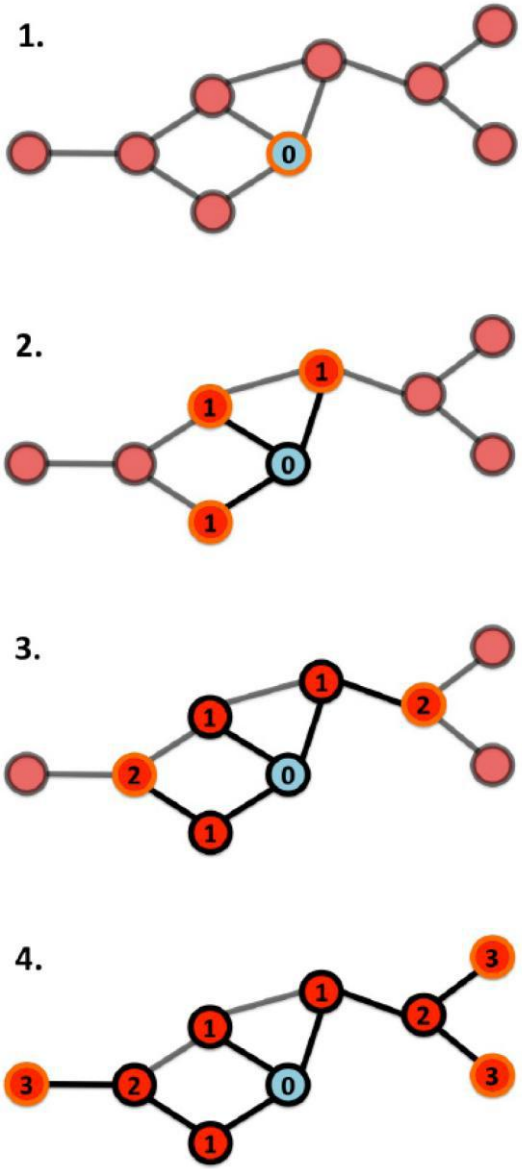


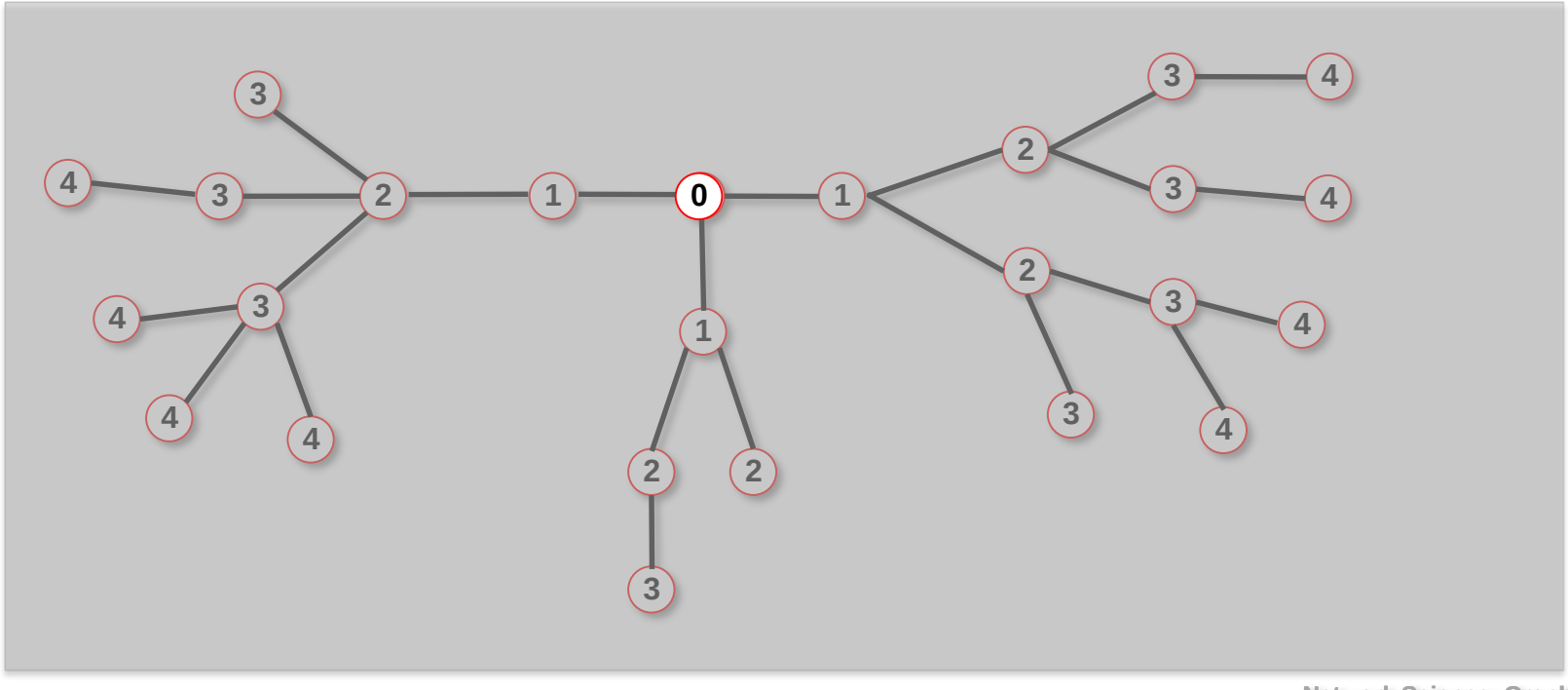
Image 2.13
The BFS algorithm applied to a small network.

Starting from the orange node, labeled "0", we identify all its neighbors, labeling them "1". Then we label "2" the unlabeled neighbors of all nodes labeled "1", and so on, in each iteration increasing the labels, until no node is left unlabeled. The length of the shortest path or the distance d_{0i} between node 0 and some other node i in the network is given by the label on node i. For example, the distance between node 0 and the leftmost node is $d_{03} = 3$.

FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

1.Start at 0.

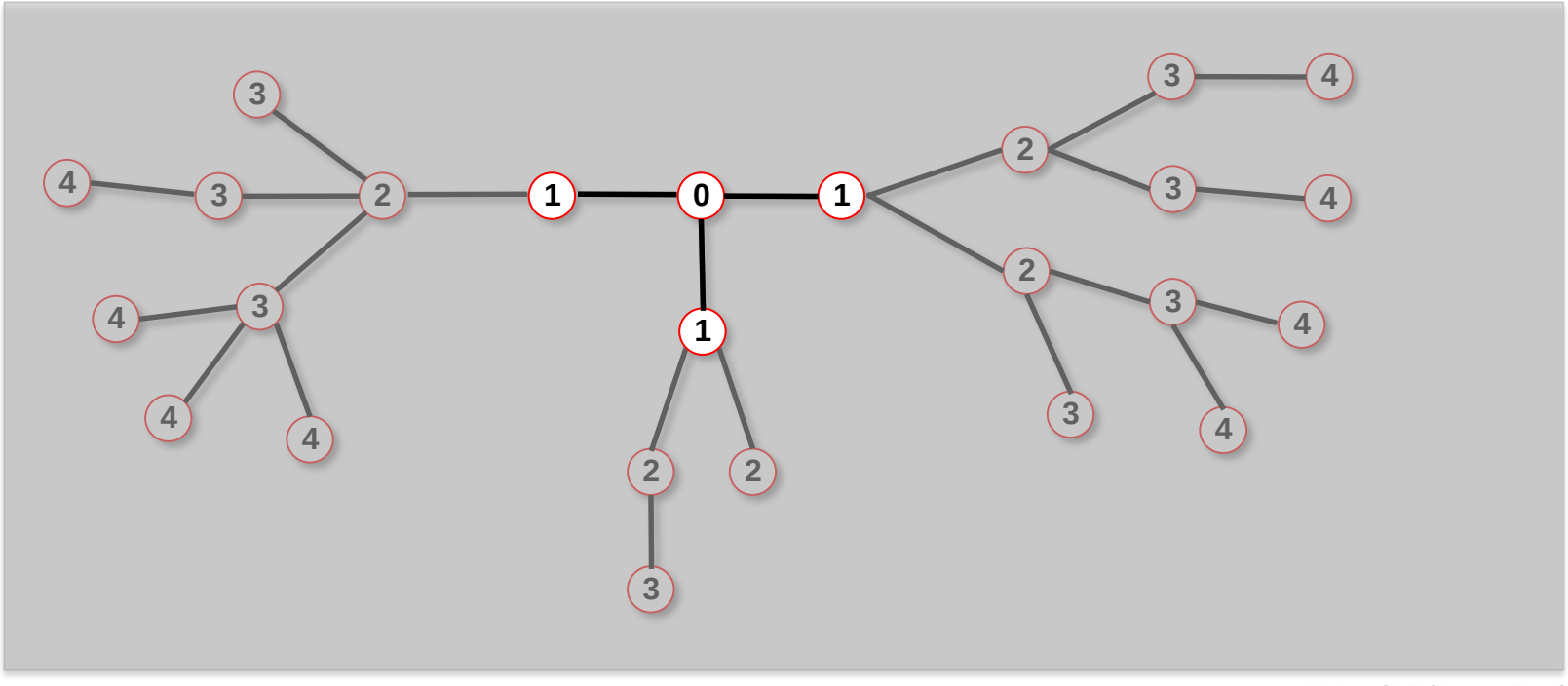


FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

1.Start at 0.

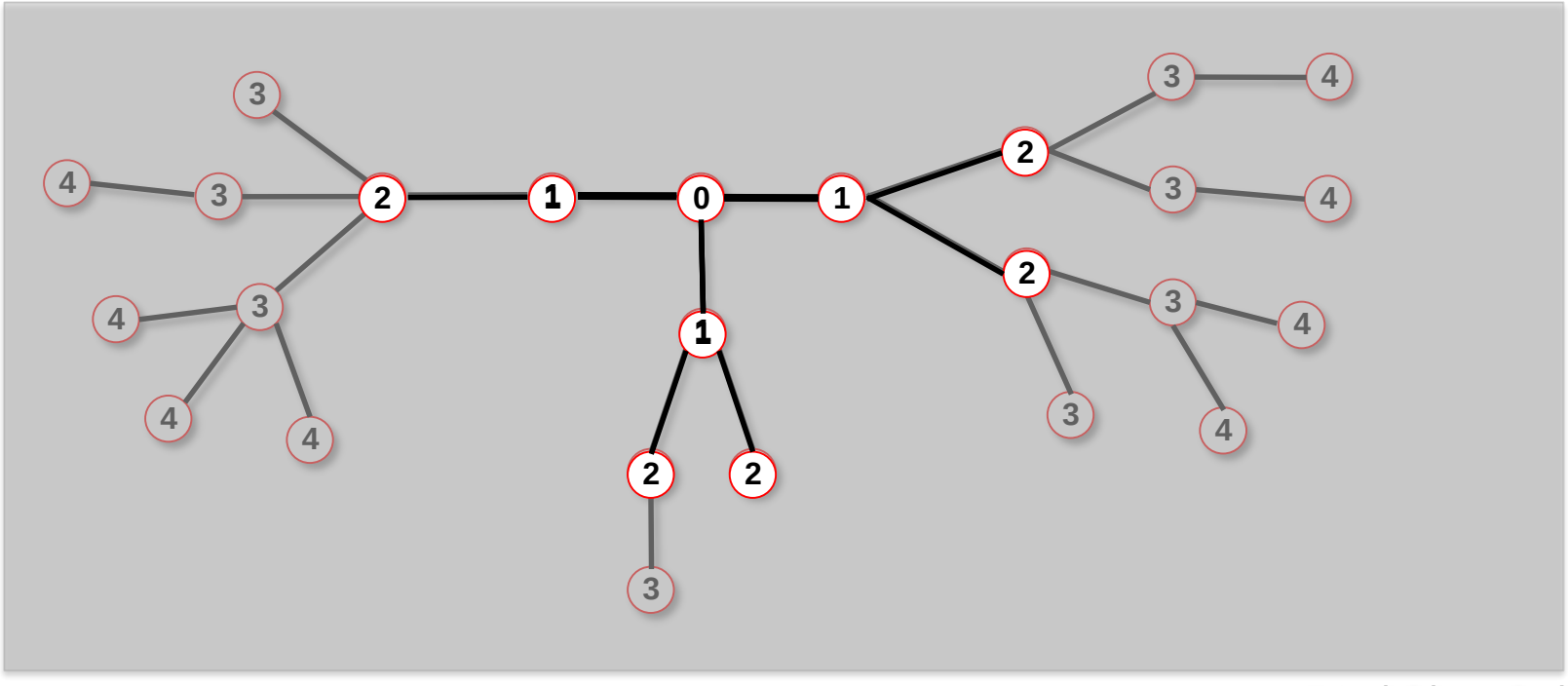
2.Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.



FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

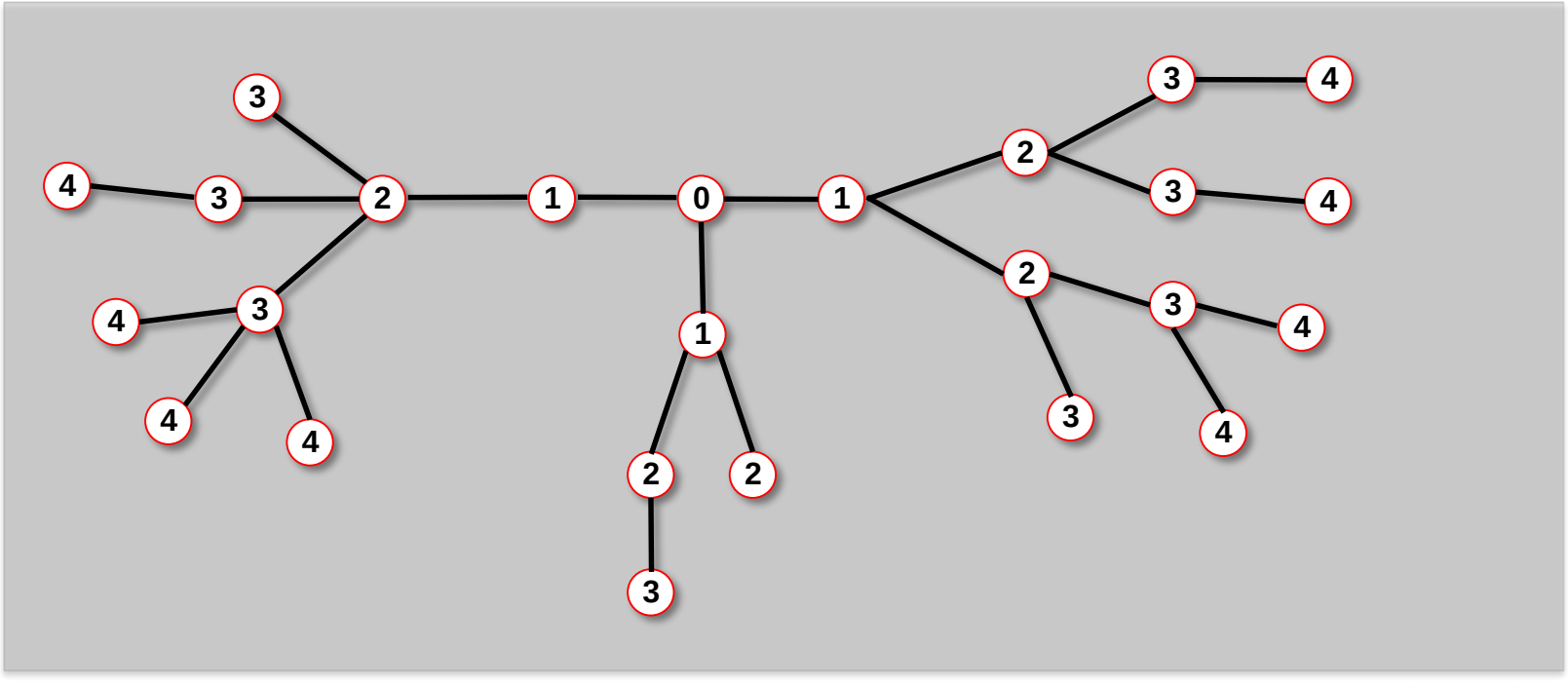
1.Start at 0.
2.Find the nodes adjacent to 0. Mark them as at distance 1. Put them in a queue.
3.Take the first node out of the queue. Find the unmarked nodes adjacent to it in the graph. Mark them with the label of 2. Put them in the queue.



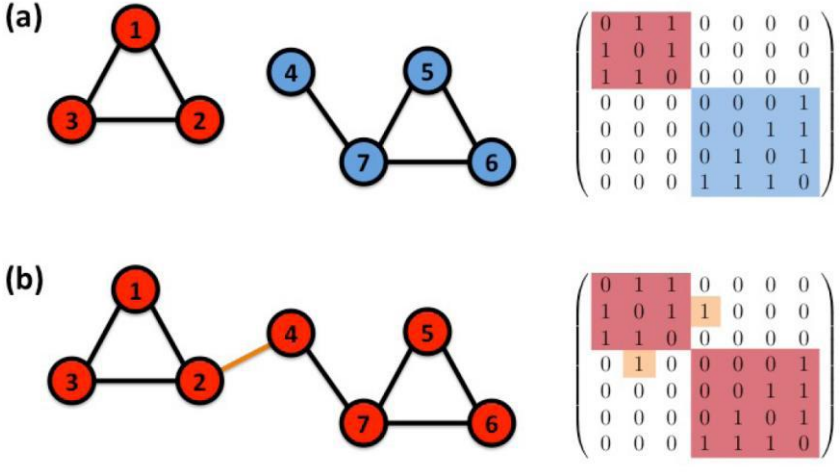
FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

1.Repeat until you find node 4 or there are no more nodes in the queue.
2.The distance between 0 and 4 is the label of 4 or, if 4 does not have a label, infinity.



Connectedness & Components



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

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

While for a small network visual inspection can help us decide if it is connected or disconnected, for a network consisting of millions of nodes connectedness is a challenging question. Several mathematical tools help us identify the connected components of a graph:

- For a disconnected network the adjacency matrix can be rearranged into a block diagonal form, such that all nonzero elements in the matrix are contained in square blocks along the matrix' diagonal and all other elements are zero (Image 2.14a). Each square block will correspond to a component. We can use the tools of linear algebra to decide if the adjacency matrix is block diagonal, helping us to identify the connected components.

- In practice, for large networks the components are more efficient identified using the breadth first search algorithm (Box 2.7).

Connectedness & Components

Finding the connected components of a graph.

- 1. Start from a randomly chosen node i and perform a BFS from this node (Box 2.6). Label all nodes reached this way with $n = 1$. By linking friends to each other, we obtain the *friendship network*, that plays an important role in the spread of ideas, products and habits and is of major interest to sociology, marketing and health sciences.

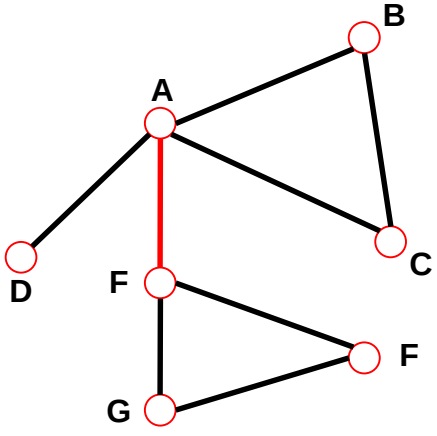
- 2. If the total number of labeled nodes equals N, then the network is connected. If the number of labeled nodes is smaller than N, the network consists of several components. To identify them, proceed to step 3.

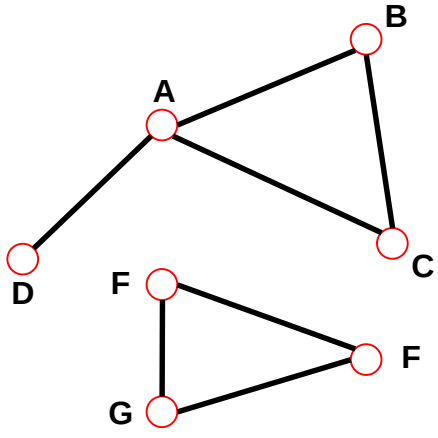
- 3. Increase the label $n \rightarrow n + 1$. Choose an unmarked node j, label it with n. Use BFS to find all nodes reachable from j, label them with n. Return to step 2.

Box 2.7

CONNECTIVITY OF UNDIRECTED GRAPHS

Connected (undirected) graph: any two vertices can be joined by a path.
A disconnected graph is made up by two or more connected components.





Largest Component:
Giant Component

The rest: **Isolates**

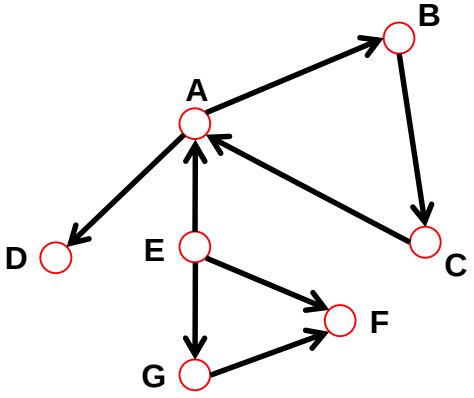
Bridge: if we erase it, the graph becomes disconnected.

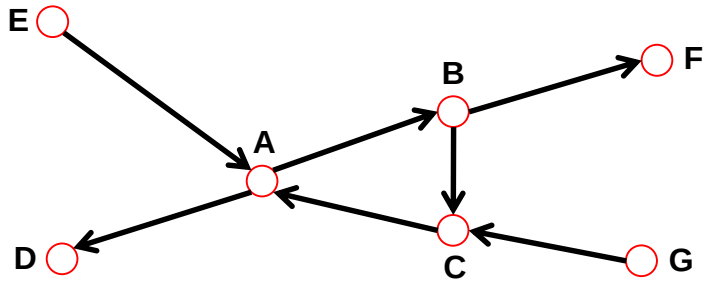
CONNECTIVITY OF DIRECTED GRAPHS

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.



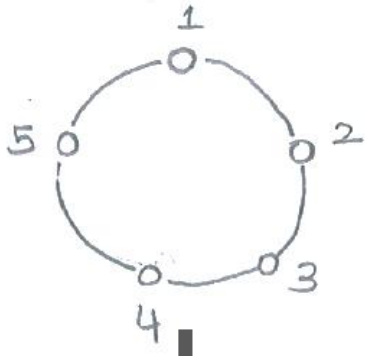


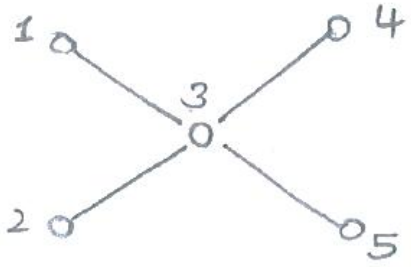
In-component: nodes that can reach the scc,

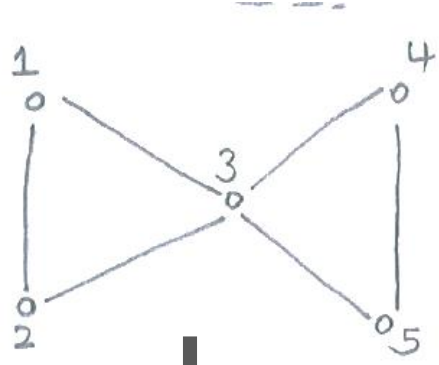
Out-component: nodes that can be reached from the scc.

Graph Theoretical Parameters

COMPUTATION OF NETWORK PARAMETERS: Adjacency Matrix







> Adjacency Matrix (A)

	1	2	3	4	5
1	0	1	0	0	1
2	1	0	1	0	0
3	0	1	0	1	0
4	0	0	1	0	1
5	1	0	0	1	0

	1	2	3	4	5
1	0	0	1	0	0
2	0	0	1	0	0
3	1	1	0	1	1
4	0	0	1	0	0
5	0	0	1	0	0

	1	2	3	4	5
1	0	1	1	0	0
2	1	0	1	0	0
3	1	1	0	1	1
4	0	0	1	0	1
5	0	0	1	1	0

Box 1 | **Network measures**

Network biology offers a quantifiable description of the networks that characterize various biological systems. Here we define the most basic network measures that allow us to compare and characterize different complex networks.

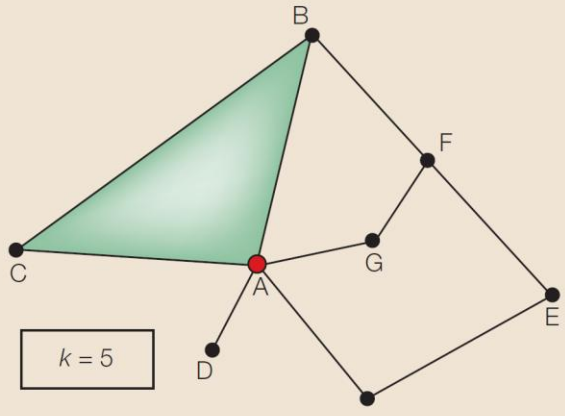
Degree

The most elementary characteristic of a node is its degree (or connectivity), k, which tells us how many links the node has to other nodes. For example, in the undirected network shown in part **a** of the figure, node A has degree $k = 5$. In networks in which each link has a selected direction (see figure, part **b**) there is an incoming degree, k_{in}, which denotes the number of links that point to a node, and an outgoing degree, k_{out}, which denotes the number of links that start from it. For example, node A in part **b** of the figure has $k_{in} = 4$ and $k_{out} = 1$. An undirected network with N nodes and L links is characterized by an average degree $\langle k \rangle = 2L/N$ (where $\langle\rangle$ denotes the average).

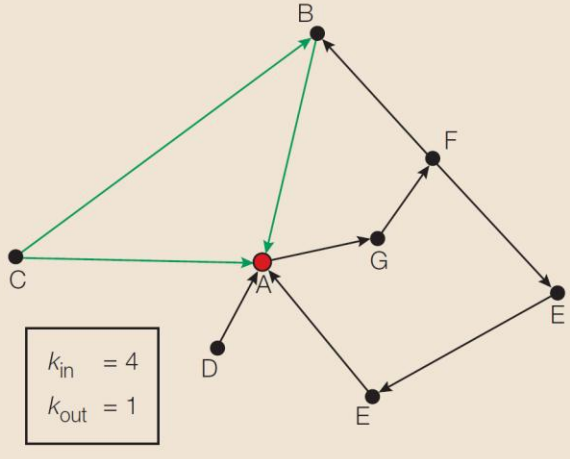
Degree distribution

The degree distribution, $P(k)$, gives the probability that a selected node has exactly k links. $P(k)$ is obtained by counting the number of nodes $N(k)$ with $k = 1, 2 \ldots$ links and dividing by the total number of nodes N. The degree distribution allows us to distinguish between different classes of networks. For example, a peaked degree distribution, as seen in a random network (BOX 2), indicates that the system has a characteristic degree and that there are no highly connected nodes (which are also known as hubs). By contrast, a power-law degree distribution indicates that a few hubs hold together numerous small nodes (BOX 2).

a Undirected network

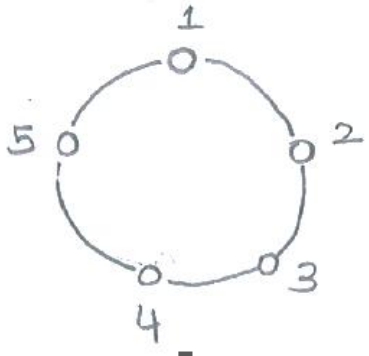


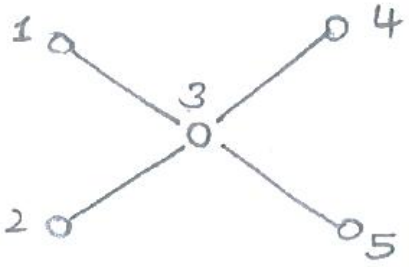
b Directed network

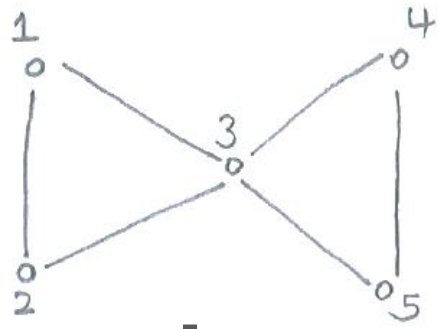


"Network Biology, Understanding the Cell's Functional Organization", Barabasi and Oltvai, Nature Reviews Genetics (2004).

COMPUTATION OF NETWORK PARAMETERS: <k>







$N = 5, \; E = 5$

$$\langle K \rangle = \frac{\sum_{ij} A_{ij}}{N} = \frac{2E}{N}$$

$$= \frac{10}{5} = 2$$

$N = 5, \; E = 4$

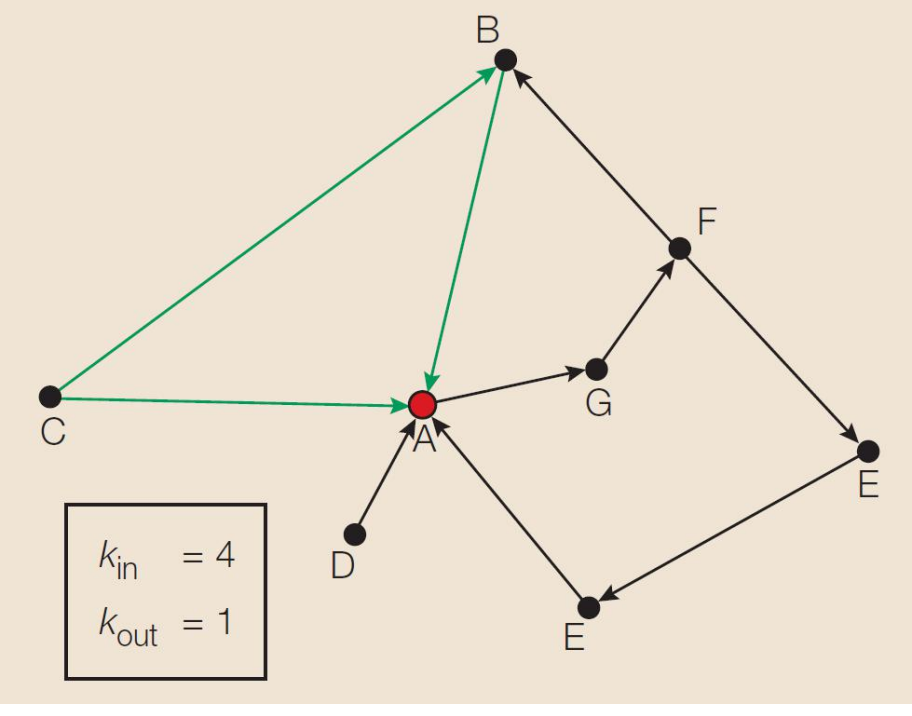
$$\langle k \rangle = \frac{8}{5} = 1.6$$

$N = 5, \; E = 6.$

$$\langle k \rangle = \frac{12}{5} = 2.4$$

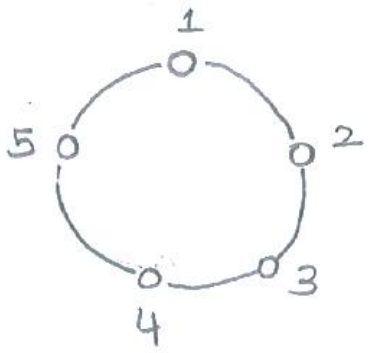
Shortest path and mean path length

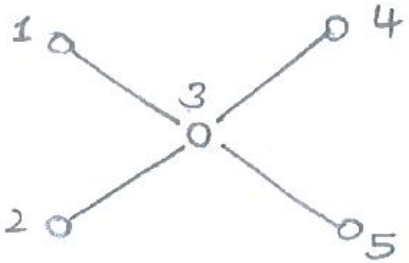
Distance in networks is measured with the path length, which tells us how many links we need to pass through to travel between two nodes. As there are many alternative paths between two nodes, the shortest path — the path with the smallest number of links between the selected nodes — has a special role. In directed networks, the distance ℓ_{AB} from node A to node B is often different from the distance ℓ_{BA} from B to A. For example, in part **b** of the figure, $\ell_{BA} = 1$, whereas $\ell_{AB} = 3$. Often there is no direct path between two nodes. As shown in part **b** of the figure, although there is a path from C to A, there is no path from A to C. The mean path length, $<\ell>$, represents the average over the shortest paths between all pairs of nodes and offers a measure of a network's overall navigability.

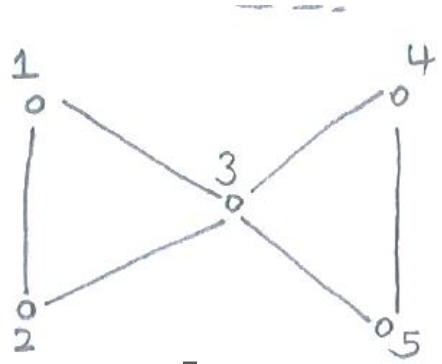


"Network Biology, Understanding the Cell's Functional Organization", Barabasi and Oltvai, Nature Reviews Genetics (2004).

COMPUTATION OF NETWORK PARAMETERS: *Shortest Path Lengths*







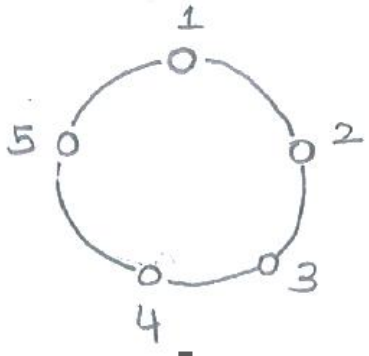
⇒ Shortest Path Lengths (L_{ij}) matrix

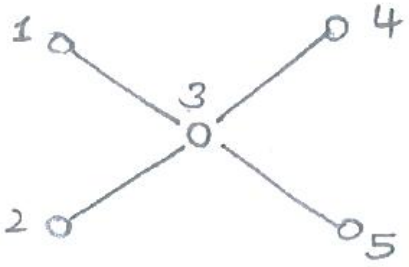
	1	2	3	4	5
1	0	1	2	2	1
2	1	0	1	2	2
3	2	1	0	1	2
4	2	2	1	0	1
5	1	2	2	1	0

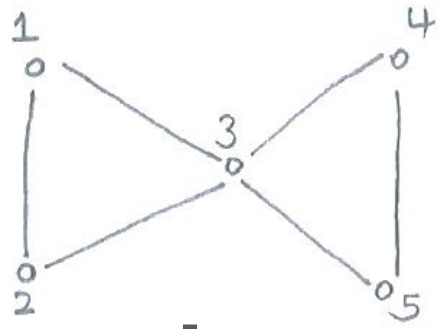
	1	2	3	4	5
1	0	2	1	2	2
2	2	0	1	2	2
3	1	1	0	1	1
4	2	2	1	0	2
5	2	2	1	2	0

	1	2	3	4	5
1	0	1	1	2	2
2	1	0	1	2	2
3	1	1	0	1	1
4	2	2	1	0	1
5	2	2	1	2	0

COMPUTATION OF NETWORK PARAMETERS: *Diameter and L*







Diameter (d) = $\max_{\forall ij}(L_{ij})$

= 2

$L = \frac{\sum L_{ij}\ \forall ij}{N(N-1)} = \frac{15 \times 2}{20}$

= 1.5

d = 2

$L = \frac{32}{20} = 1.6$

d = 2

$L = \frac{28}{20} = 1.4$

Clustering coefficient

In many networks, if node A is connected to B, and B is connected to C, then it is highly probable that A also has a direct link to C. This phenomenon can be quantified using the clustering coefficient[33] $C_I = 2n_I/k(k-1)$, where n_I is the number of links connecting the k_I neighbours of node I to each other. In other words, C_I gives the number of 'triangles' (see BOX 3) that go through node I, whereas $k_I(k_I-1)/2$ is the total number of triangles that could pass through node I, should all of node I's neighbours be connected to each other. For example, only one pair of node A's five neighbours in part **a** of the figure are linked together (B and C), which gives $n_A = 1$ and $C_A = 2/20$. By contrast, none of node F's neighbours link to each other, giving $C_F = 0$. The average clustering coefficient, $\langle C \rangle$, characterizes the overall tendency of nodes to form clusters or groups. An important measure of the network's structure is the function $C(k)$, which is defined as the average clustering coefficient of all nodes with k links. For many real networks $C(k) \sim k^{-1}$, which is an indication of a network's hierarchical character[47,53] (see BOX 2).

The average degree $\langle k \rangle$, average path length $\langle \ell \rangle$ and average clustering coefficient $\langle C \rangle$ depend on the number of nodes and links (N and L) in the network. By contrast, the $P(k)$ and $C(k)$ functions are independent of the network's size and they therefore capture a network's generic features, which allows them to be used to classify various networks.

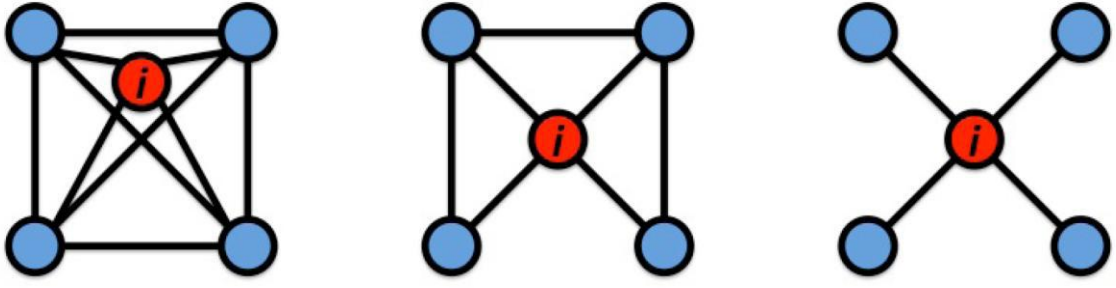
"Network Biology, Understanding the Cell's Functional Organization", Barabasi and Oltvai, Nature Reviews Genetics (2004).

Clustering Coefficient

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

where L_i represents the number of links between the k_i neighbors of node i.

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



$C_i = 1$ $C_i = 1/2$ $C_i = 0$

$C = 1$ $C = 9/14$ $C = 0$

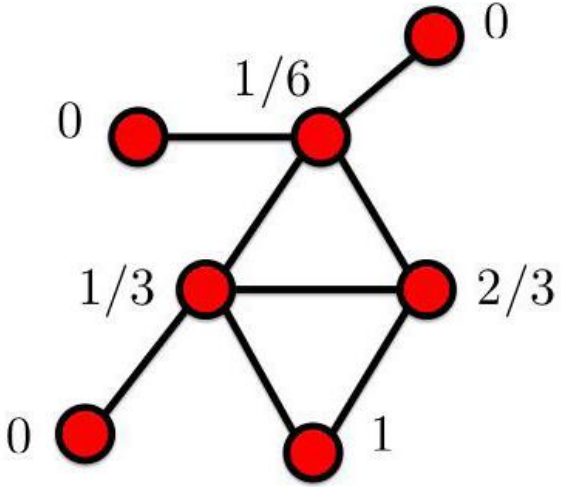
CLUSTERING COEFFICIENT

Clustering coefficient:

what fraction of your neighbors are connected?

- Node i with degree k_i
- C_i in [0,1]

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

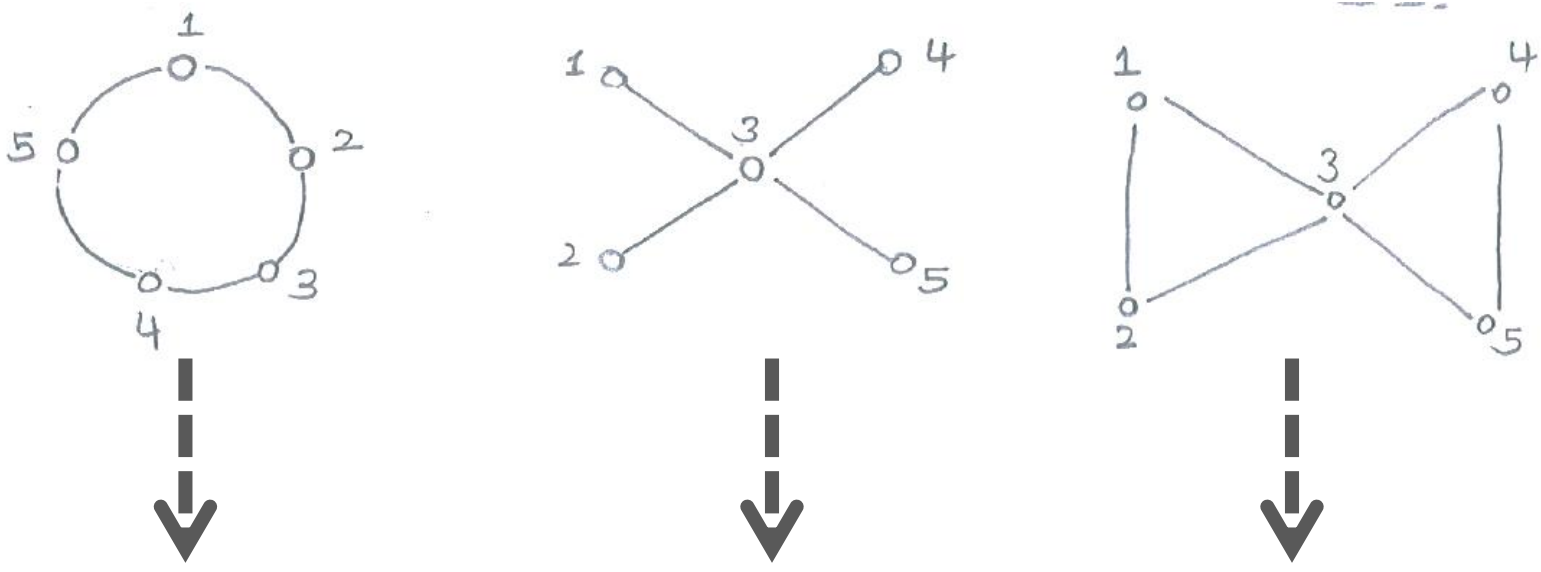


$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

$$C = \frac{3}{8} = 0.375$$

Watts & Strogatz, Nature 1998.

COMPUTATION OF NETWORK PARAMETERS: *Clustering Coefficient*



Compute clustering coefficient of each of the node.

EXERCISE: Study the following network concepts

- Adjacency and incidence (directed and undirected)
- Isomorphism
- Subgraphs
- Cliques
- Walk
- Path
- Cycle
- Connected components
- Connectivity
- Cutsets
- Strongly connected component
- Tree
- Spanning Tree
- Complement of a network
- Regular network
- Empty network
- Cycle networks
- Network coloring
- Bipartite network
- *k*-partite network
- Planar networks

EXERCISE

Let A be the $N \times N$ adjacency matrix of an undirected unweighted network, without self-loops. Let 1 be a column vector of N elements, all equal to 1. In other words $1 = (1, 1, ..., 1)^T$, where the superscript T indicates the *transpose* operation. Use the matrix formalism (multiplicative constants, multiplication row by column, matrix operations like transpose and trace, etc., but avoid the sum symbol Σ) to write expressions for:

a. The vector k whose elements are the degrees k_i of all nodes $i = 1, 2,..., N$.

b. The total number of links, L, in the network.

c. The number of triangles T present in the network, where a triangle means three nodes, each connected by links to the other two (Hint: you can use the trace of a matrix).

d. The vector k_{nn} whose element i is the sum of the degrees of node i's neighbors.

e. The vector k_{nnn} whose element i is the sum of the degrees of node i's second neighbors.

NetworkX Tutorial

- NetworkX Tutorial

https://networkx.github.io/documentation/networkx-1.10/tutorial/index.html