NETWORK SCIENCE

Graphs and Networks

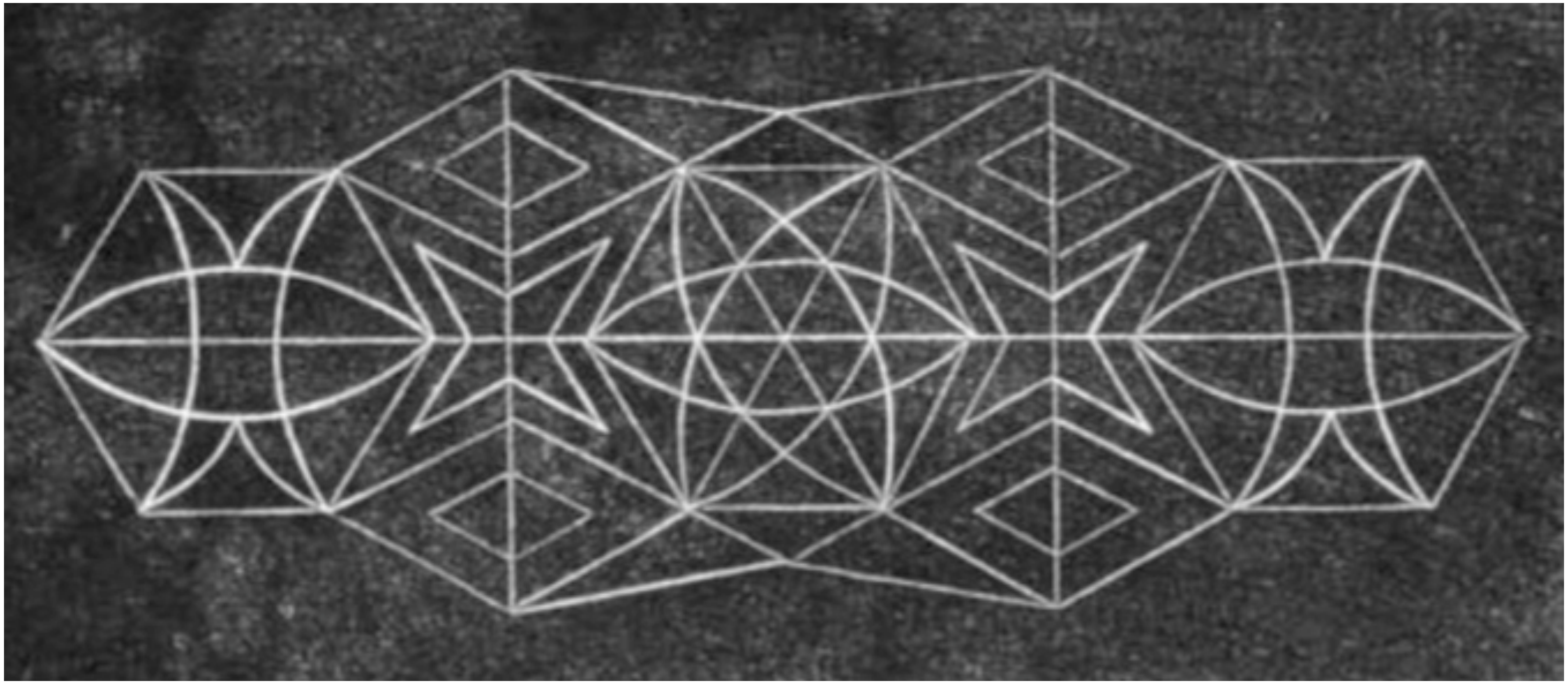
Prof. Marcello Pelillo

Ca' Foscari University of Venice

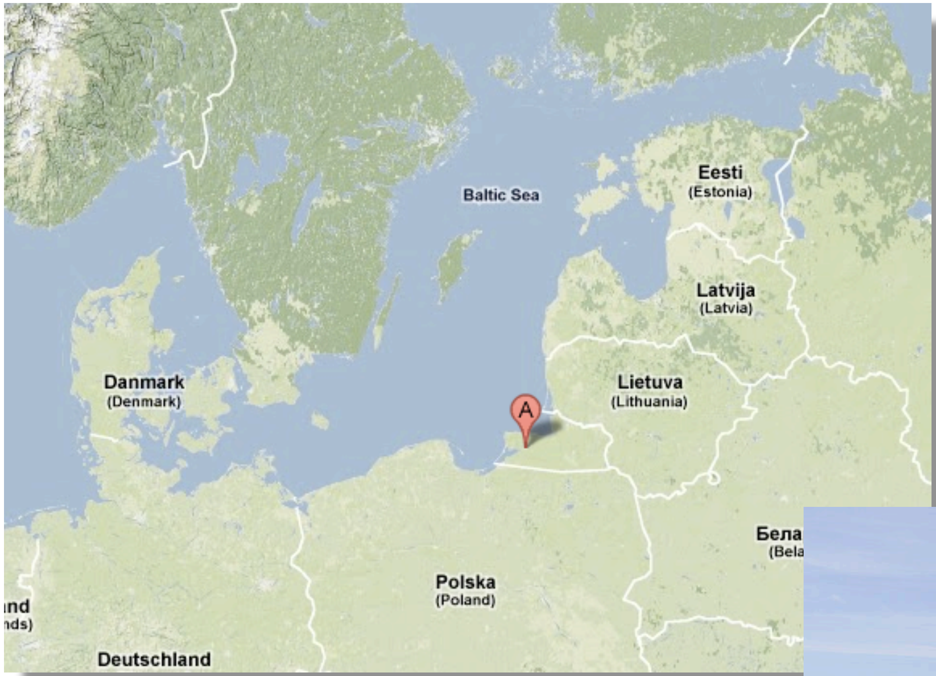
a.y. 2016/17

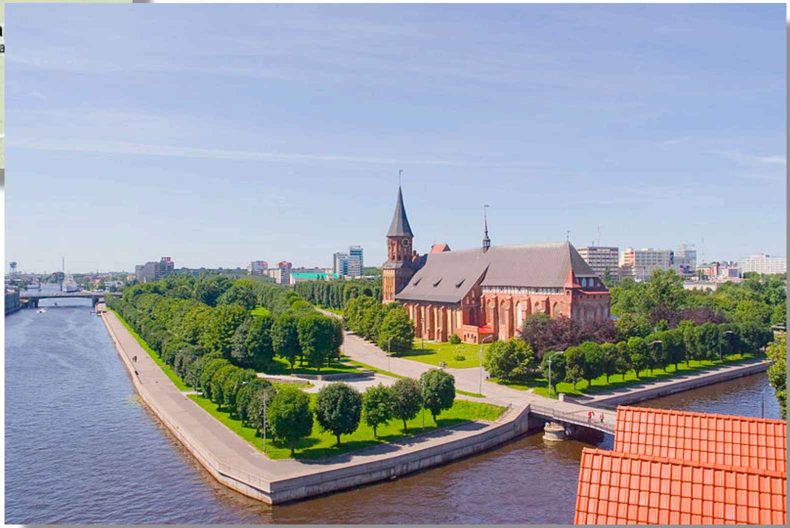
The Bridges of Konigsberg

Drawing Curves with a Single Stroke…



Königsberg
(today's Kaliningrad, Russia)





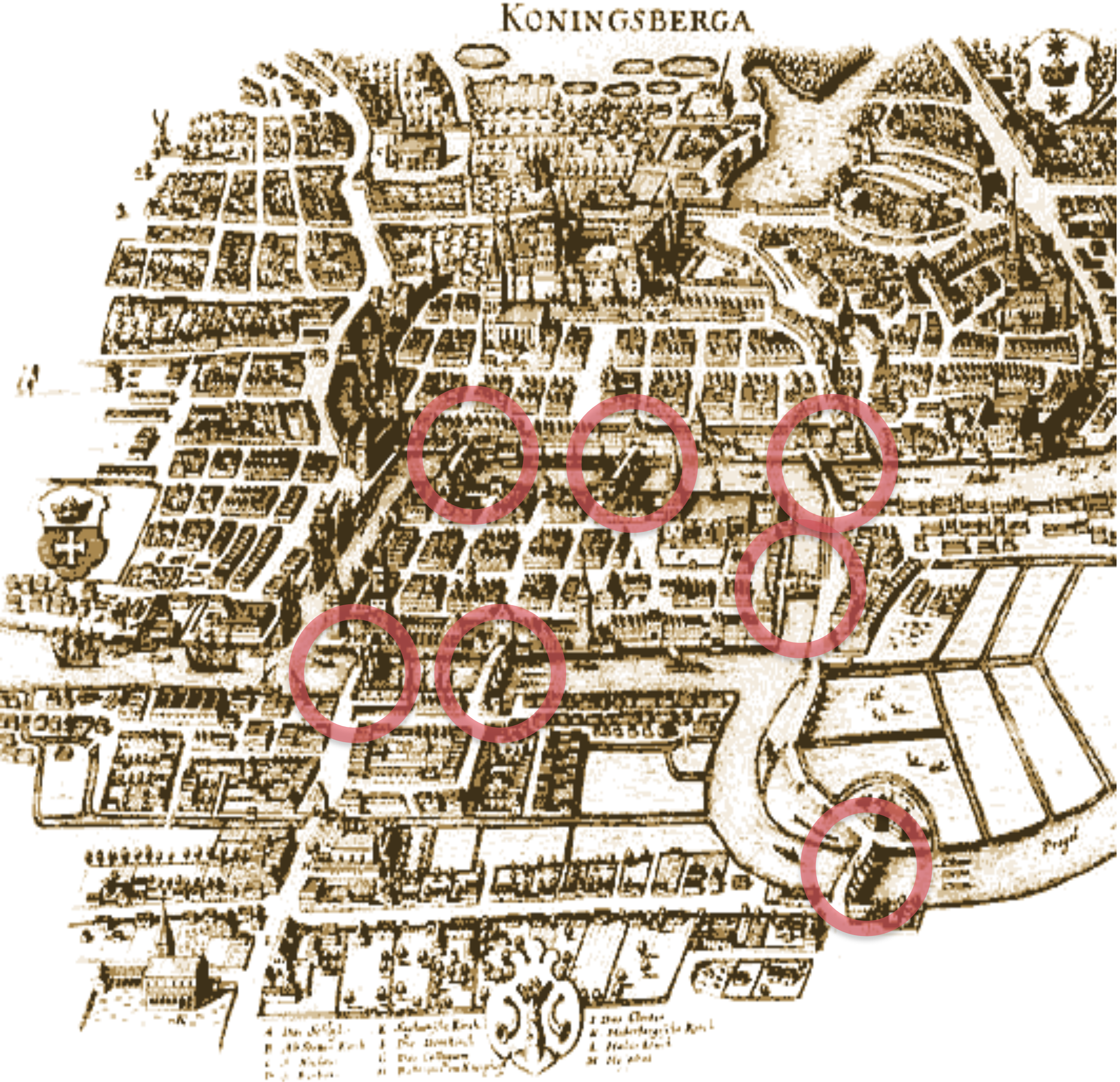
Konigsberg's People

Immanuel Kant
(1724 – 1804)

Gustav Kirchhoff
(1824 – 1887)

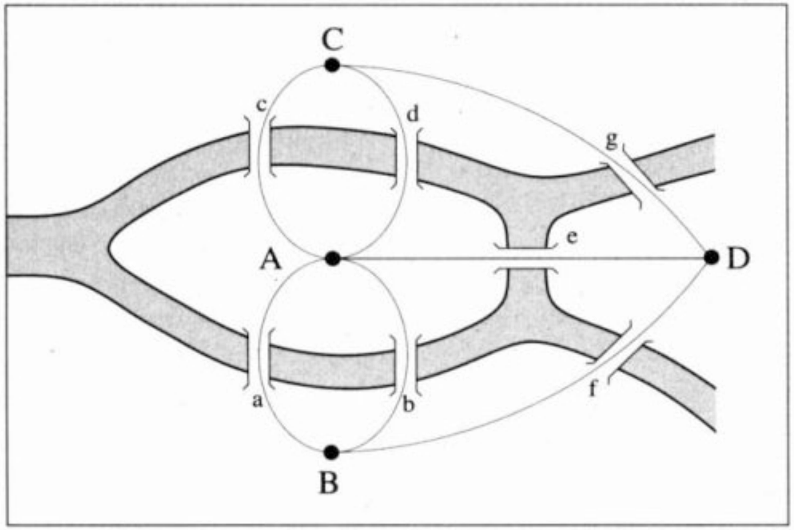
David Hilbert
(1862 – 1943)

THE BRIDGES OF KONIGSBERG



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

THE BRIDGES OF KONIGSBERG



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

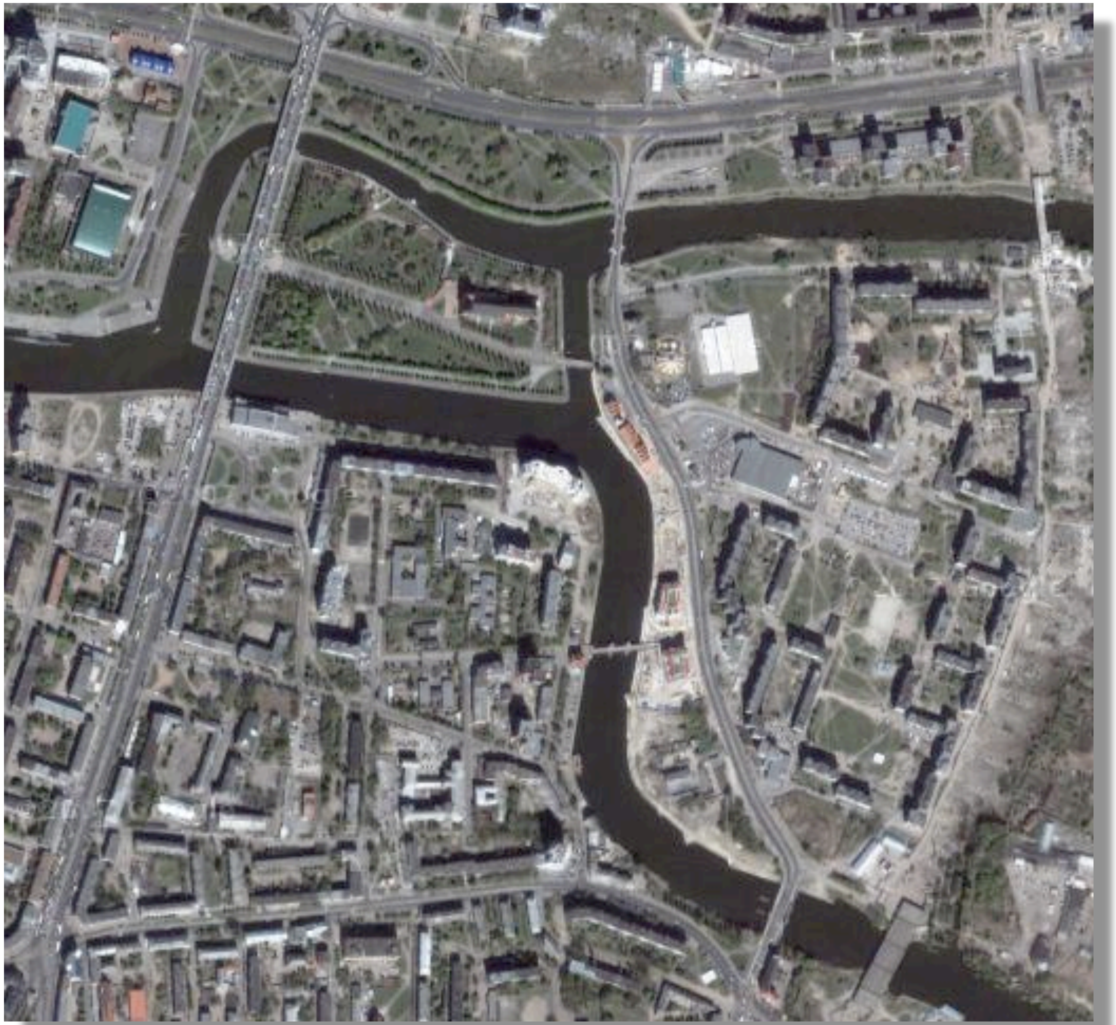
1735: **Euler's theorem:**

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

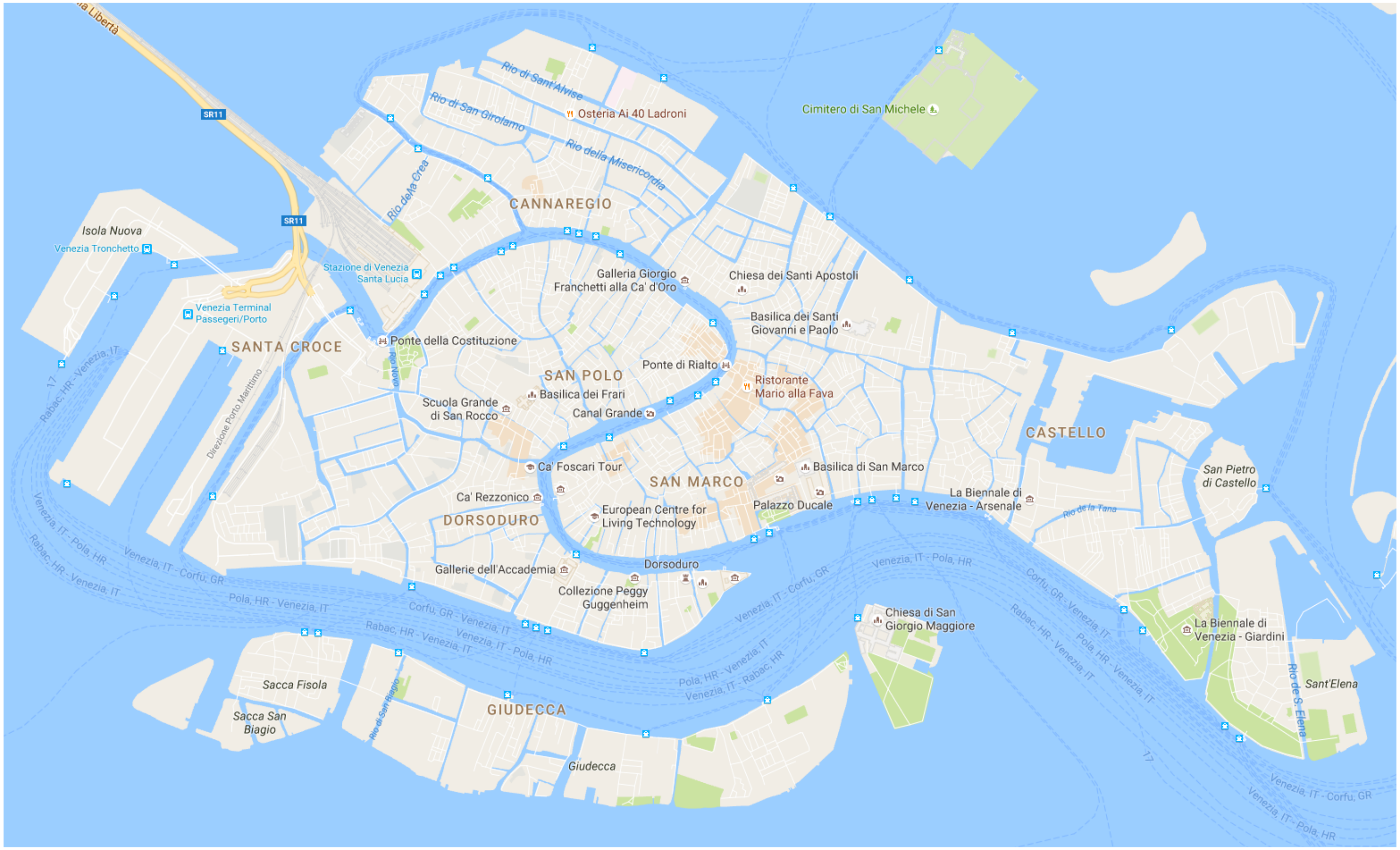
(b) If a graph is connected and has no odd degree nodes, or two such vertices, it has at least one path.

Euler's solution is considered to be the first theorem in graph theory.

The Bridges Today



A “Local” Variation of Euler’s Problem

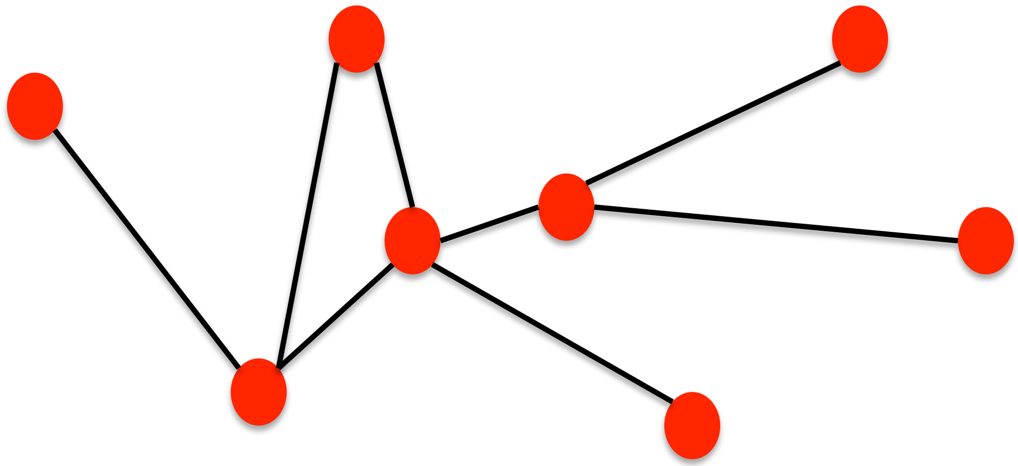


Graphs and networks after the "bridges"

- Laws of electrical circuitry (G. Kirchhoff, 1845)
- Molecular structure in chemistry (A. Cayley, 1874)
- Network representation of social interactions (J. Moreno, 1930)
- Power grids (1910)
- Telecommunications and the Internet (1960)
- Google (1997), Facebook (2004), Twitter (2006), . . .

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)

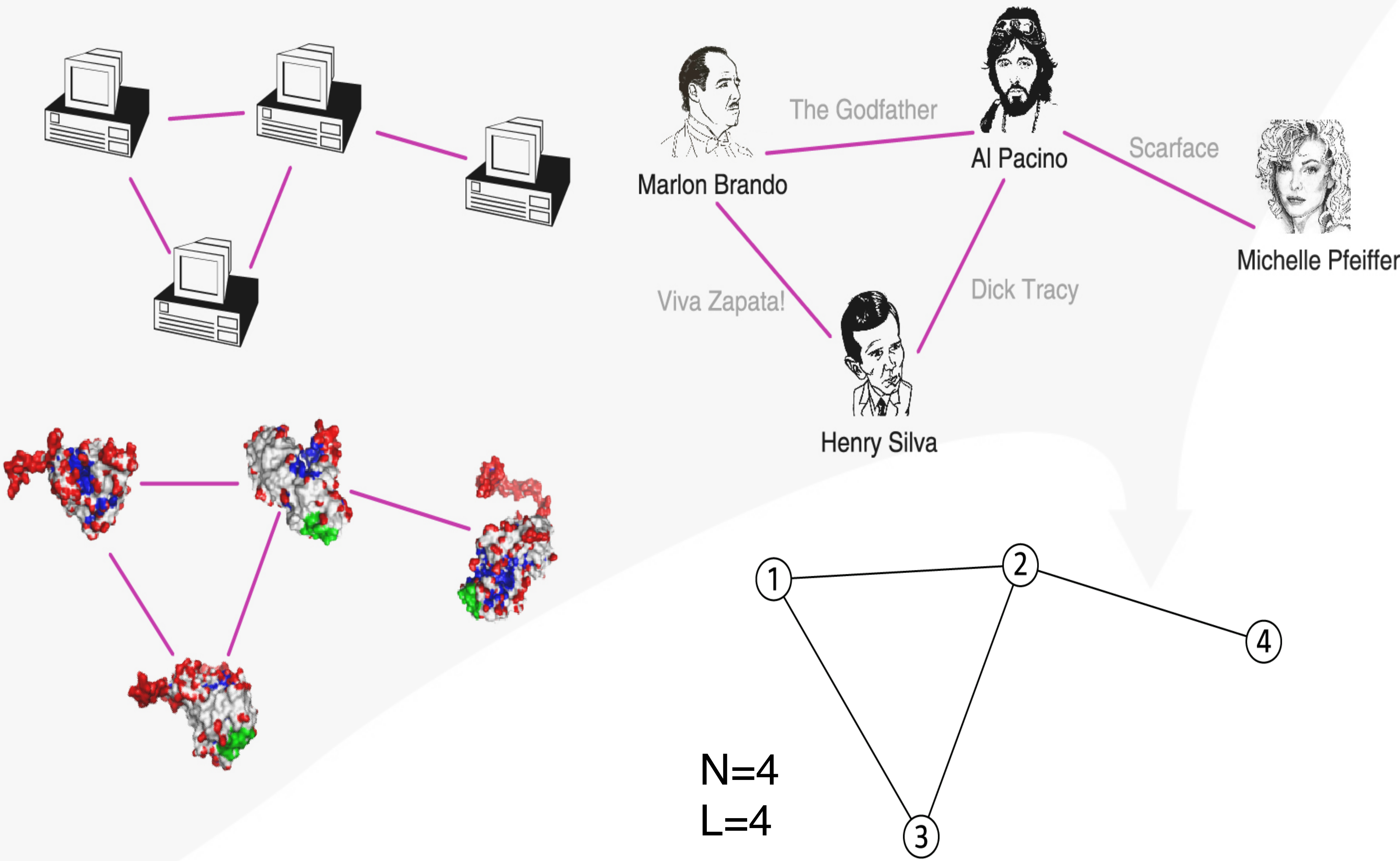
graph: mathematical representation of a network

- 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

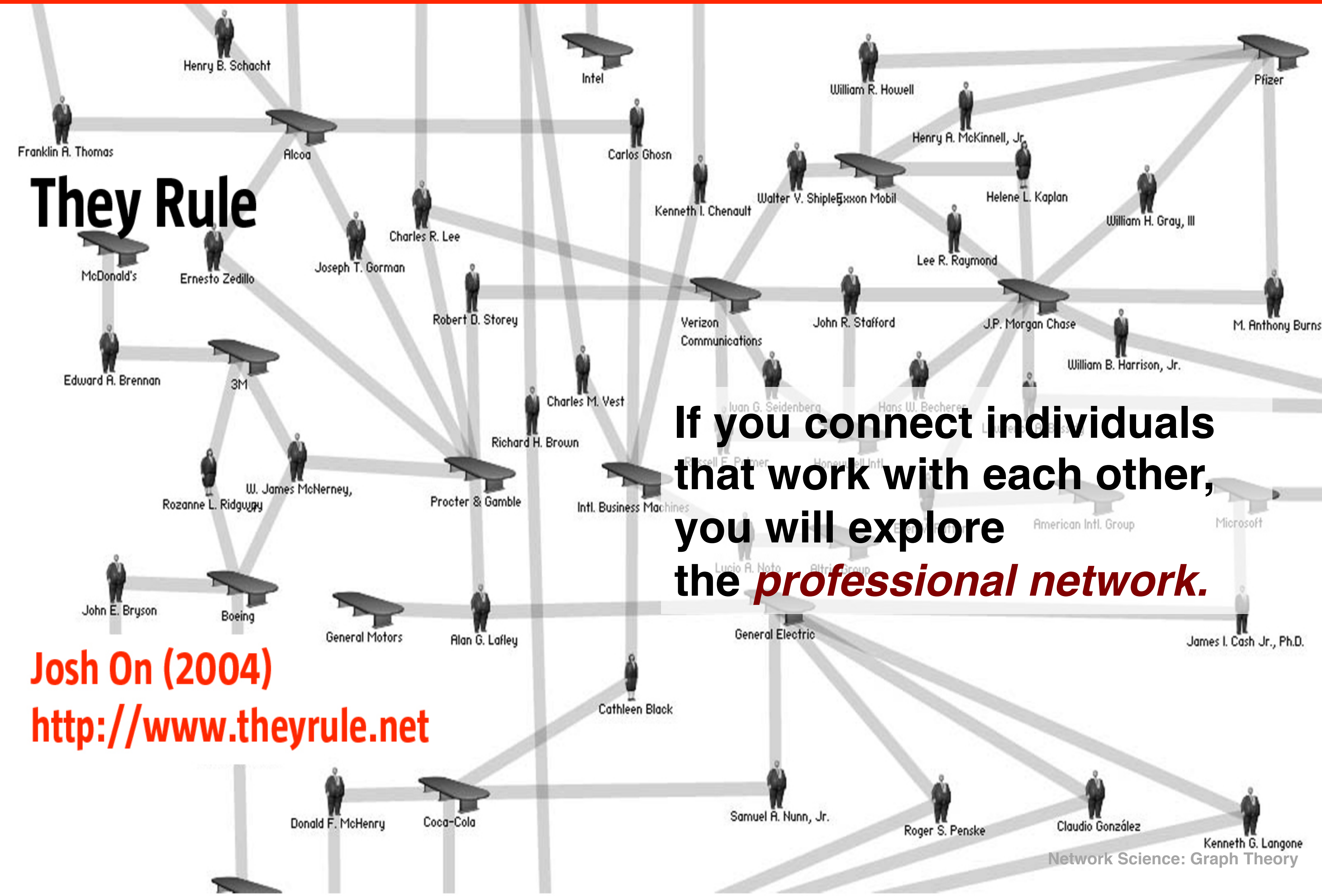


CHOOSING A PROPER REPRESENTATION

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.

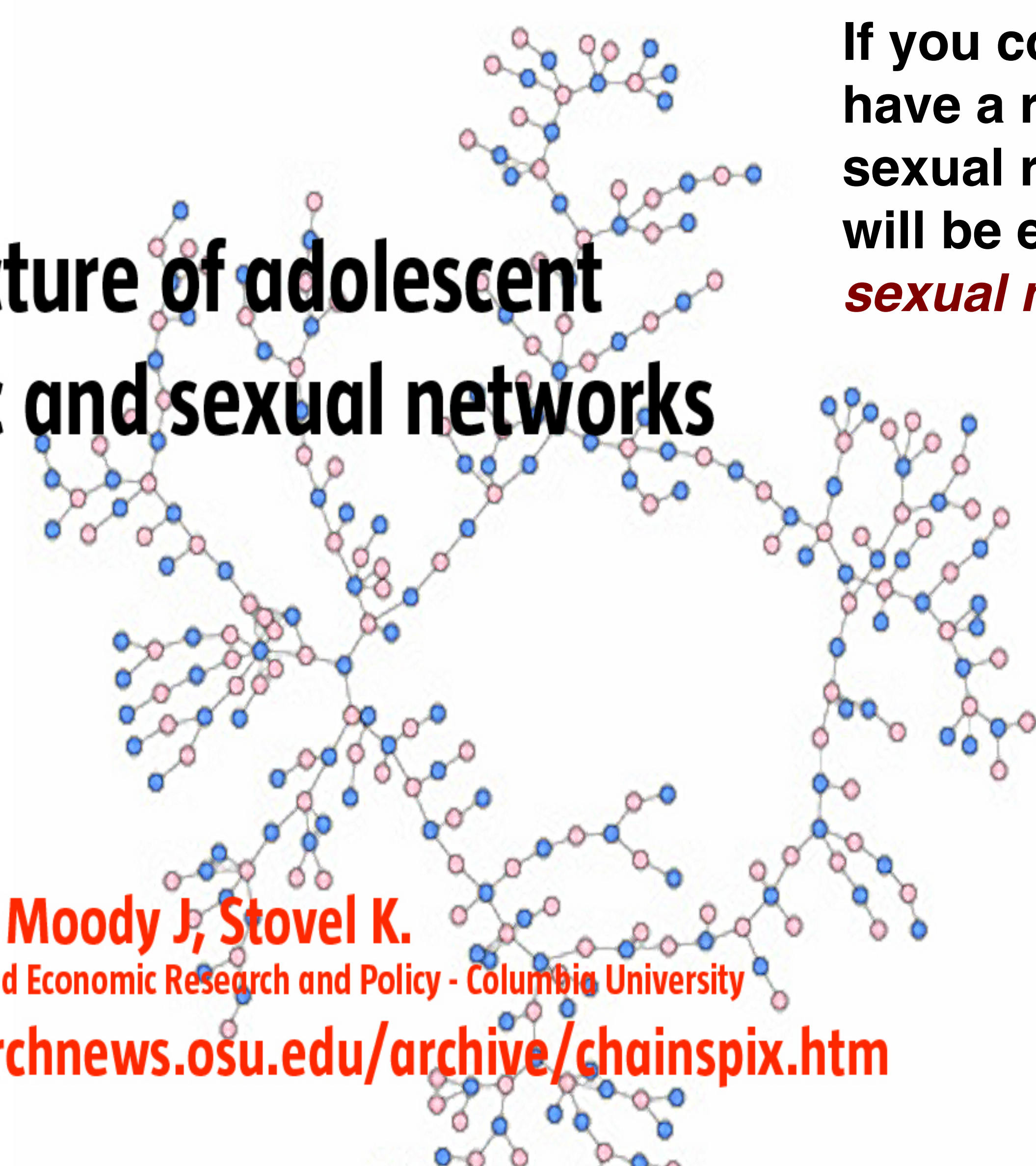
For example, the way we assign the links between a group of individuals will determine the nature of the question we can study.



They Rule

If you connect individuals that work with each other, you will explore the *professional network.*

Josh On (2004)
http://www.theyrule.net



The structure of adolescent romantic and 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

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

CHOOSING A PROPER REPRESENTATION

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

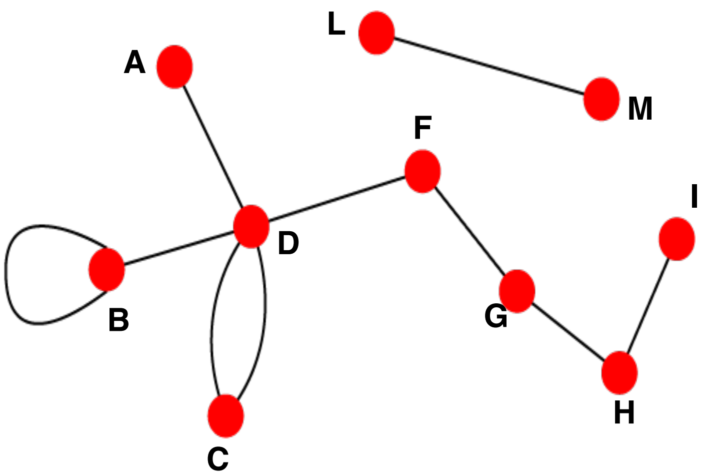
It is a network, nevertheless.

UNDIRECTED VS. DIRECTED NETWORKS

Undirected

Links: undirected (*symmetrical*)

Graph:



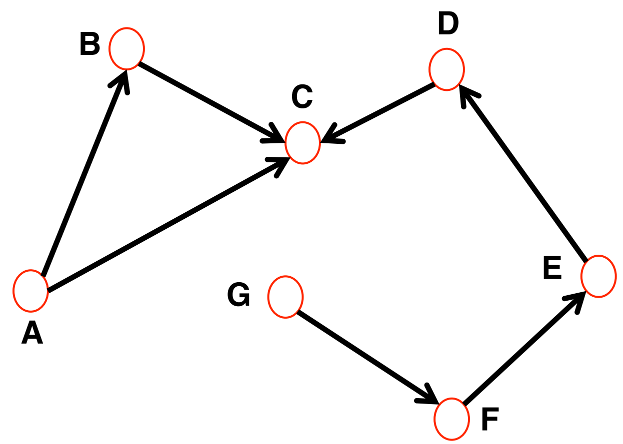
Undirected links :

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

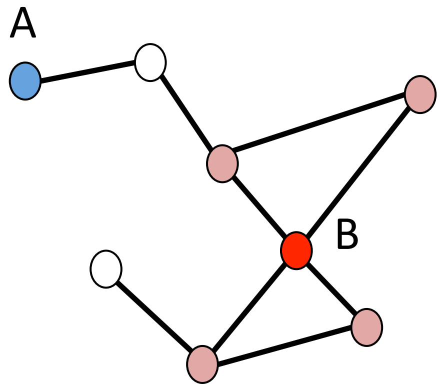
Section 2.2 Reference Networks

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

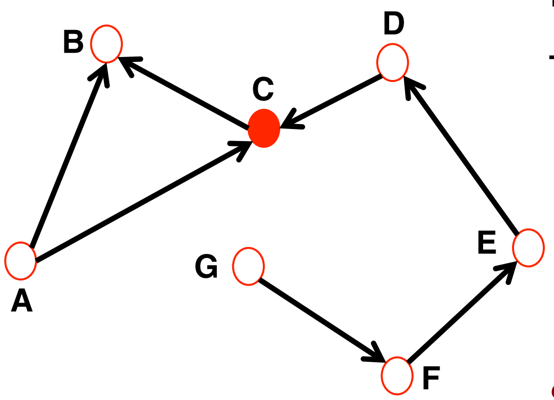
Undirected



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.

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

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 $x_1, \dots, x_N$:

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:

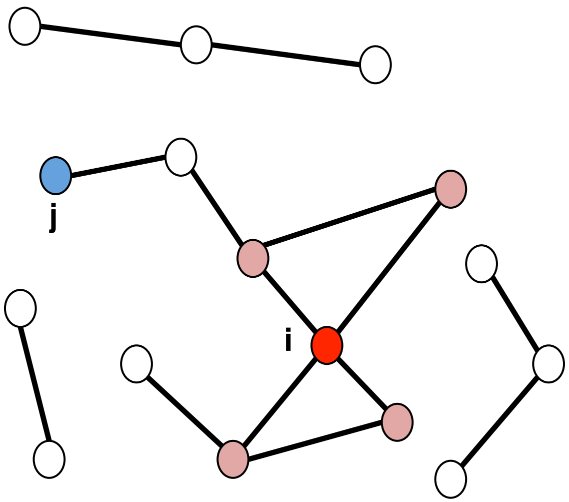
$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

where p_x follows

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

AVERAGE DEGREE

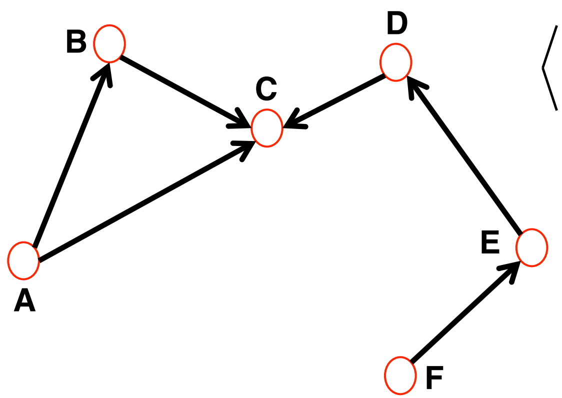
Undirected



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

N – the number of nodes in the graph

Directed



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

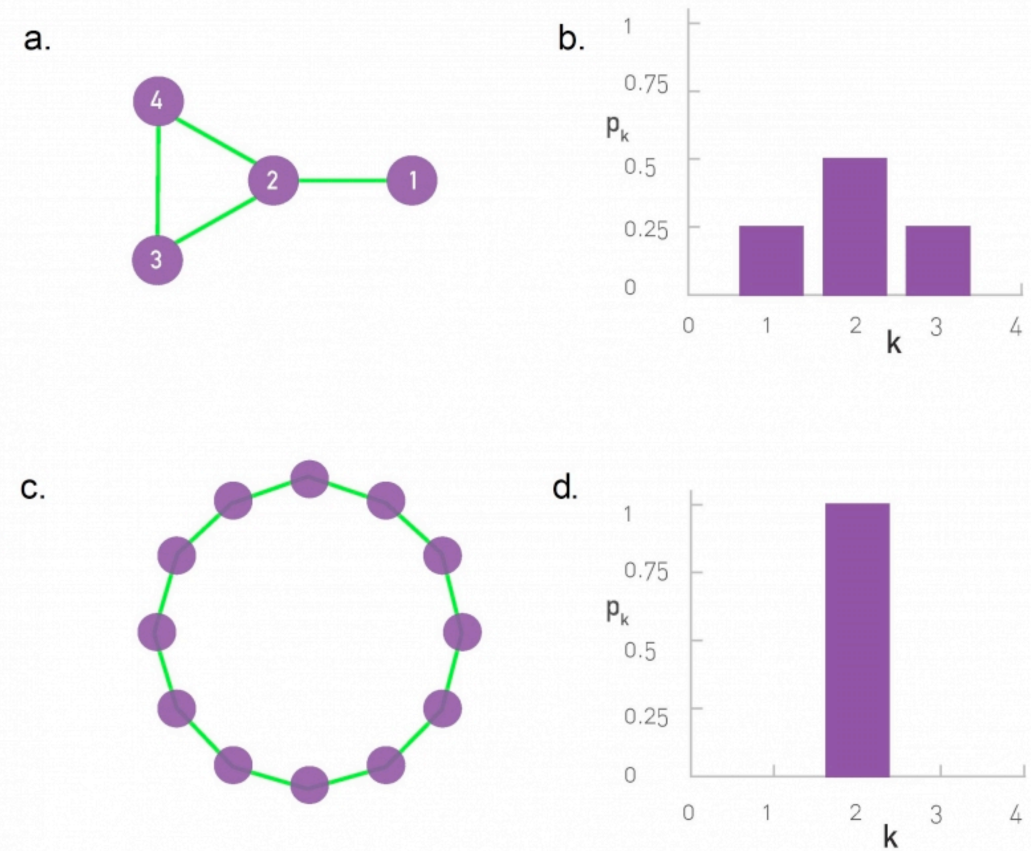
$$\langle k \rangle \equiv \frac{L}{N}$$

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

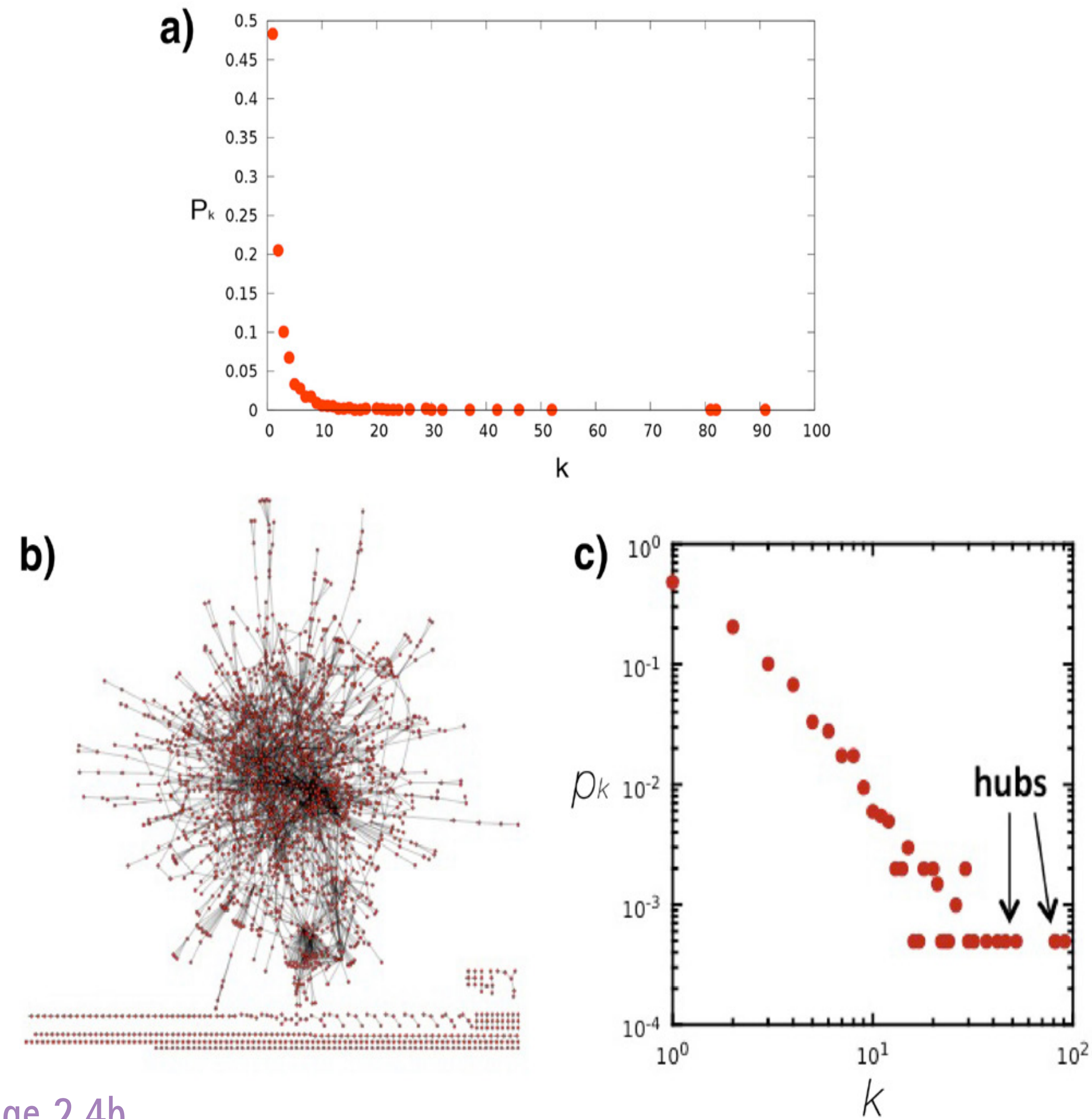
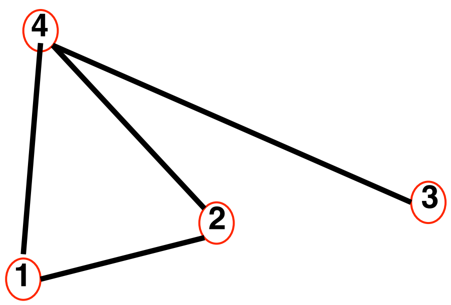
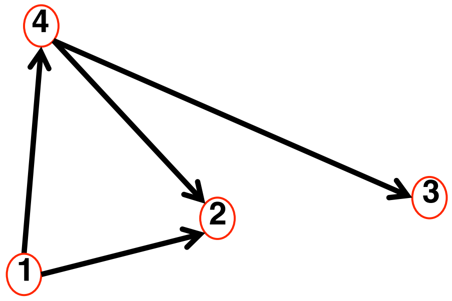


Image 2.4b

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.

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

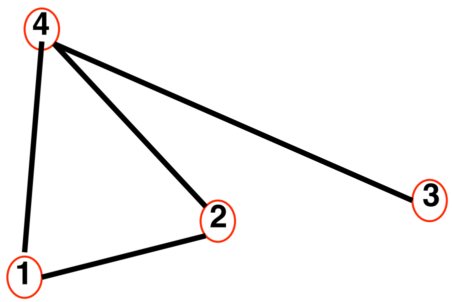
Note that for a directed graph (right) the matrix is not symmetric.

$A_{ij} = 1$ 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



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

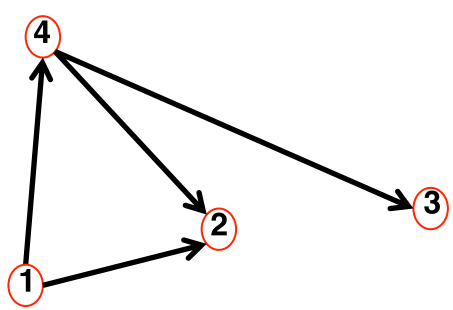
$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

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

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

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

Directed



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

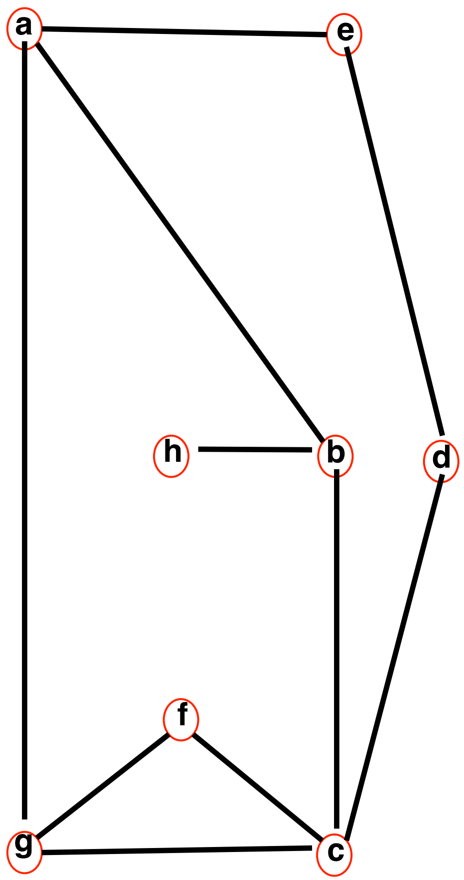
$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

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

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

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

ADJACENCY MATRIX

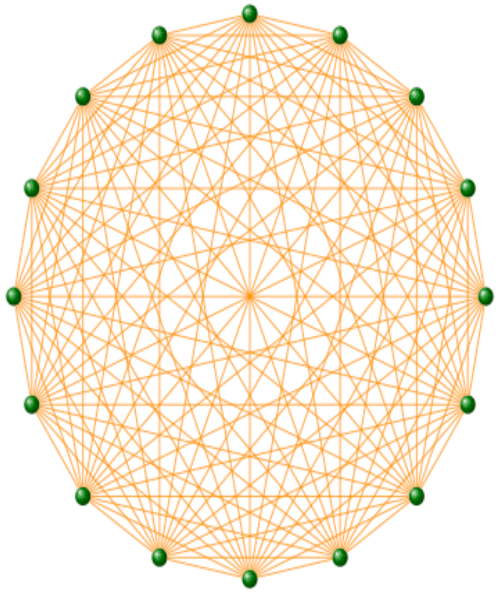


	a	b	c	d	e	f	g	h
a	0	1	0	0	1	0	1	0
b	1	0	1	0	0	0	0	1
c	0	1	0	1	0	1	1	0
d	0	0	1	0	1	0	0	0
e	1	0	0	1	0	0	0	0
f	0	0	1	0	0	0	1	0
g	1	0	1	0	0	0	0	0
h	0	1	0	0	0	0	0	0

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:

$$L \ll L_{max}$$

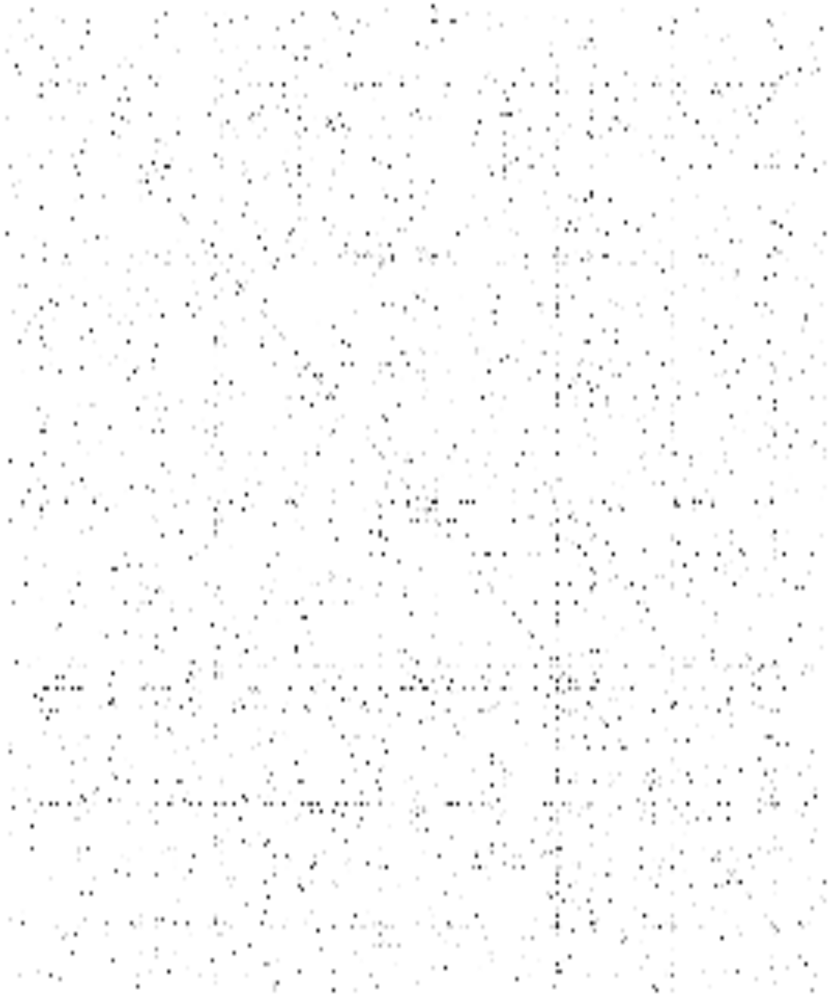
or

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

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

(Source: Albert, Barabasi, RMP2002)

ADJACENCY MATRICES ARE SPARSE



WEIGHTED AND UNWEIGHTED NETWORKS

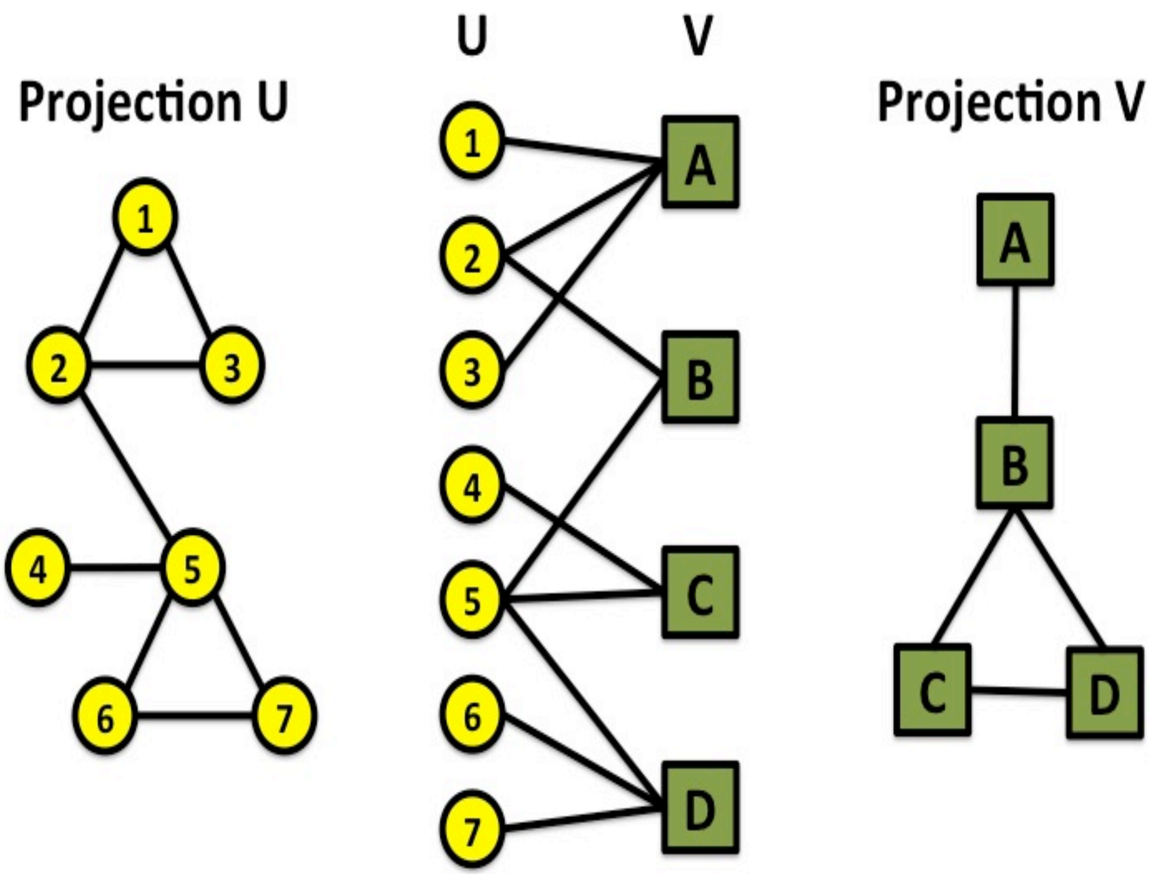
WEIGHTED AND UNWEIGHTED NETWORKS

$$A_{ij} = w_{ij}$$

BIPARTITE NETWORKS

BIPARTITE GRAPHS

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.

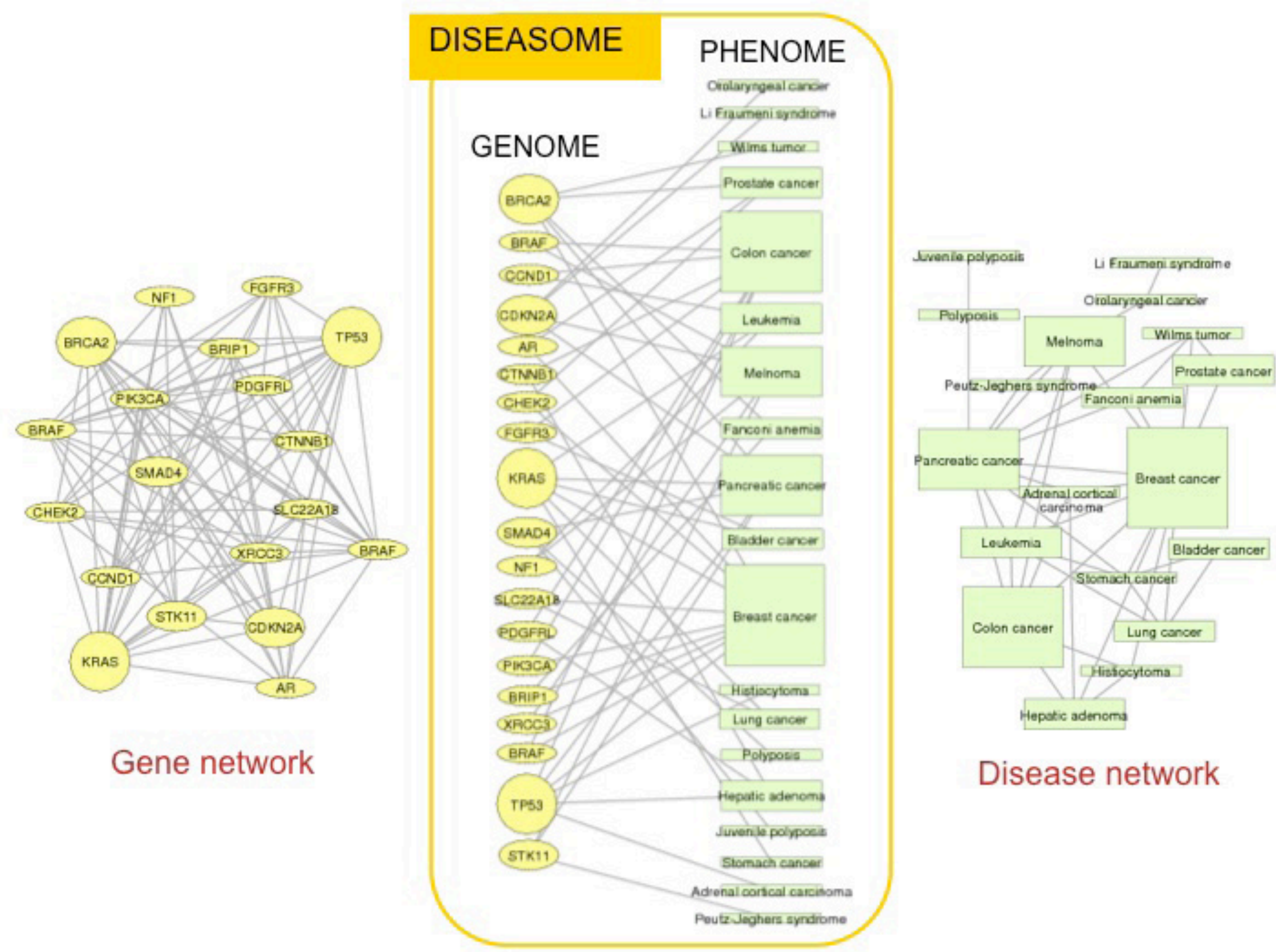


Examples:

Hollywood actor network
Collaboration networks
Disease network (diseasome)

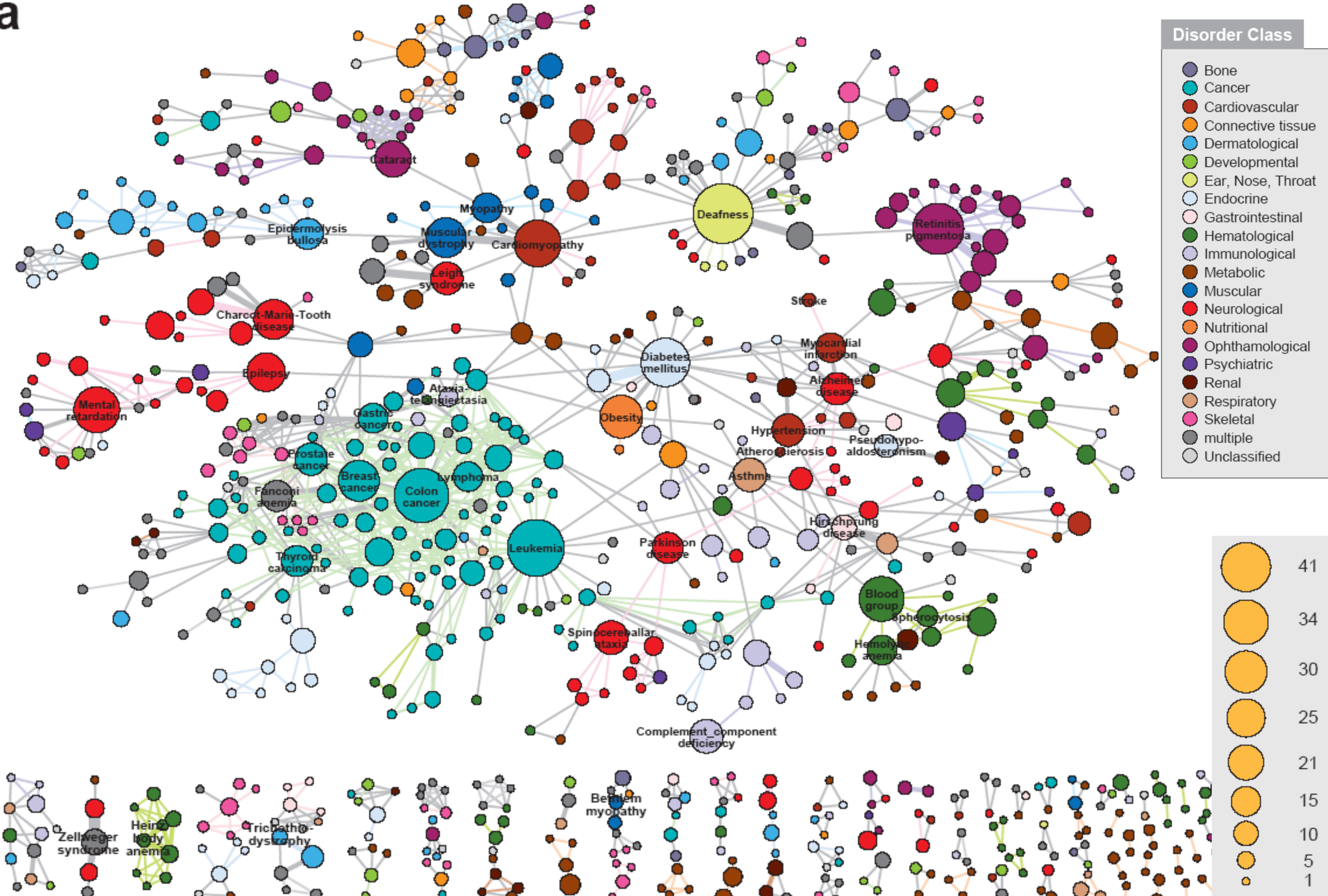
GENE NETWORK – DISEASE NETWORK

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



The *human diseaseome* is a bipartite network, whose nodes are diseases (U) and genes (V), in which a disease is connected to a gene if mutations in that gene are known to affect the particular disease

HUMAN DISEASE NETWORK



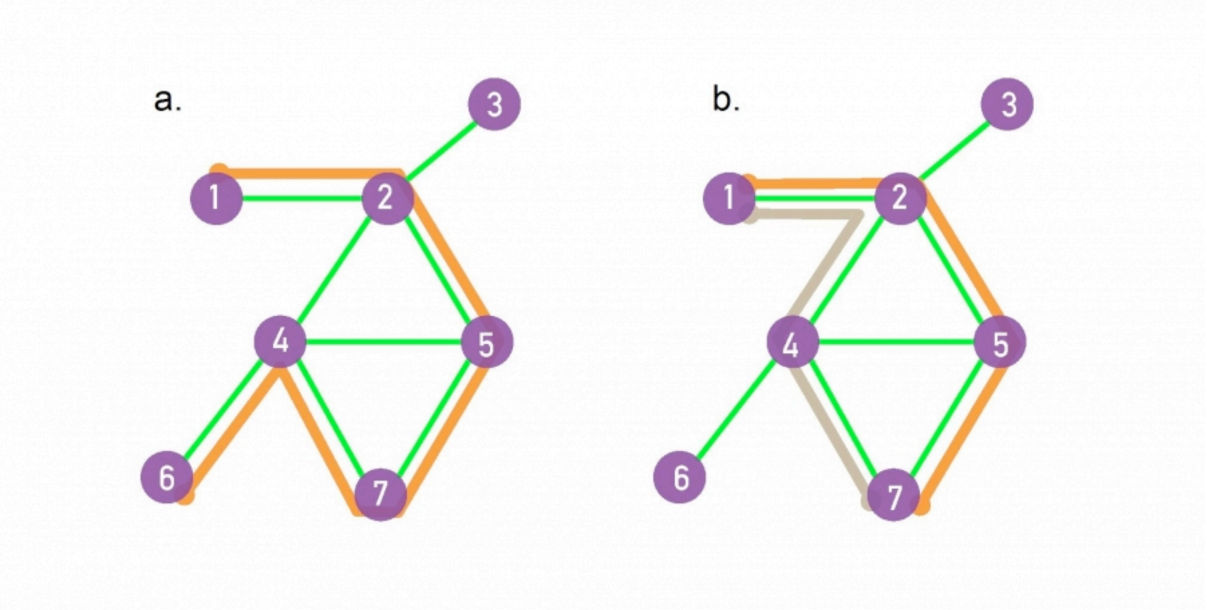
PATHOLOGY

PATHS

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

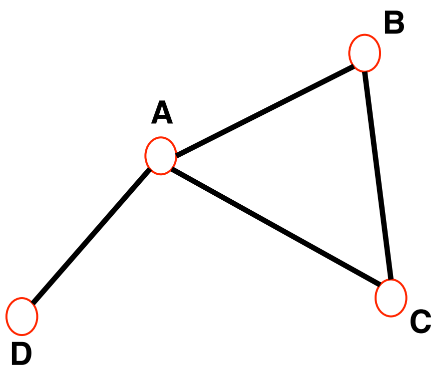
$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

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



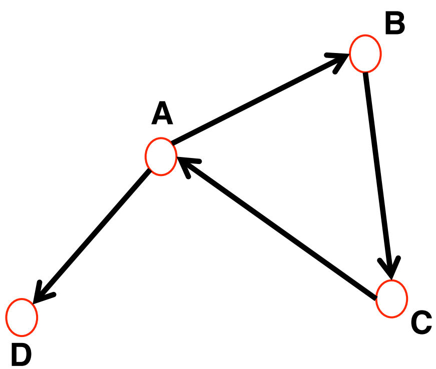
- In a directed network, the path can follow only the direction of an arrow.

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).

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, the number of paths of length *n* between *i* and *j* is*

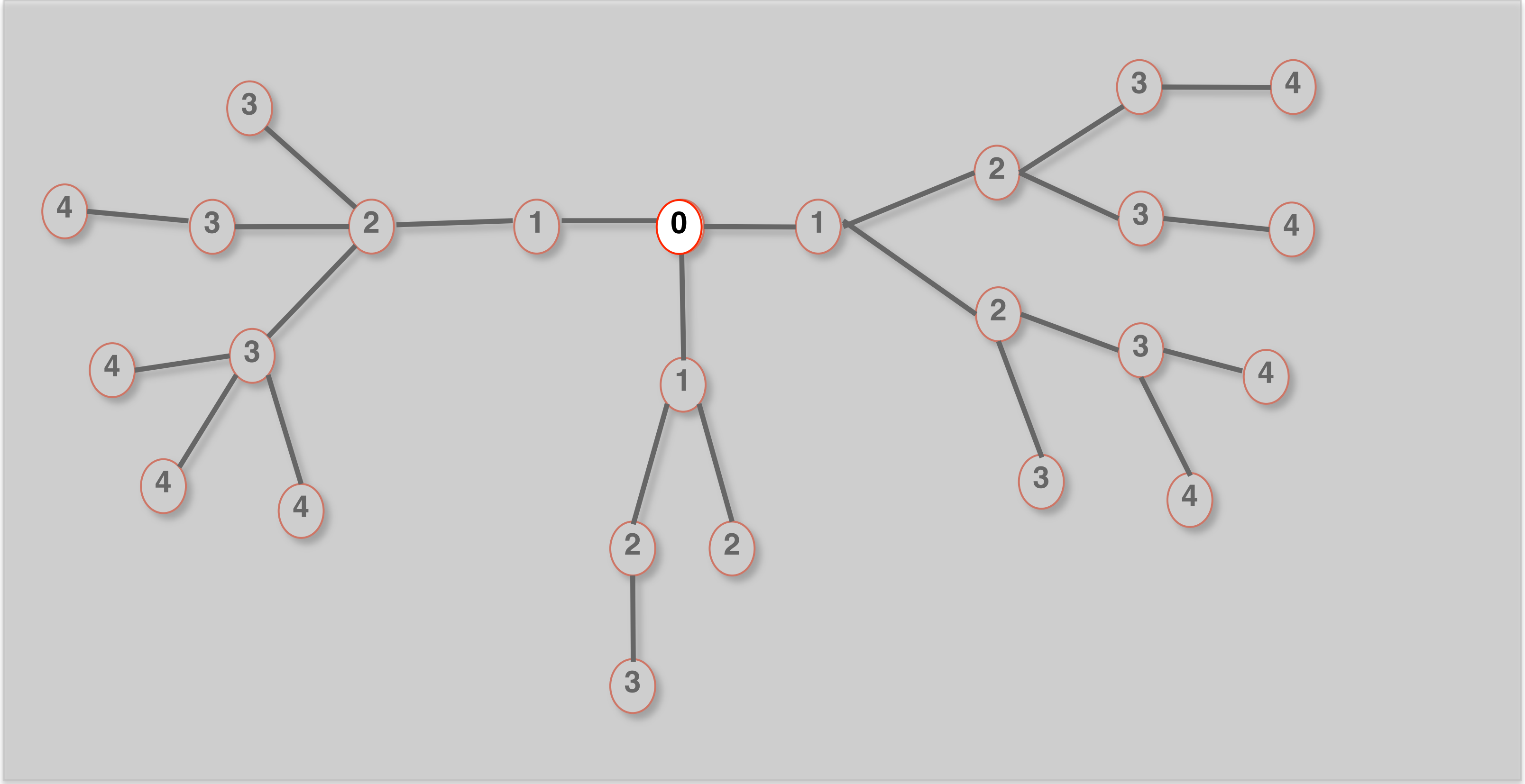
$$N_{ij}^{(n)} = [A^n]_{ij}$$

* *holds for both directed and undirected networks.*

FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

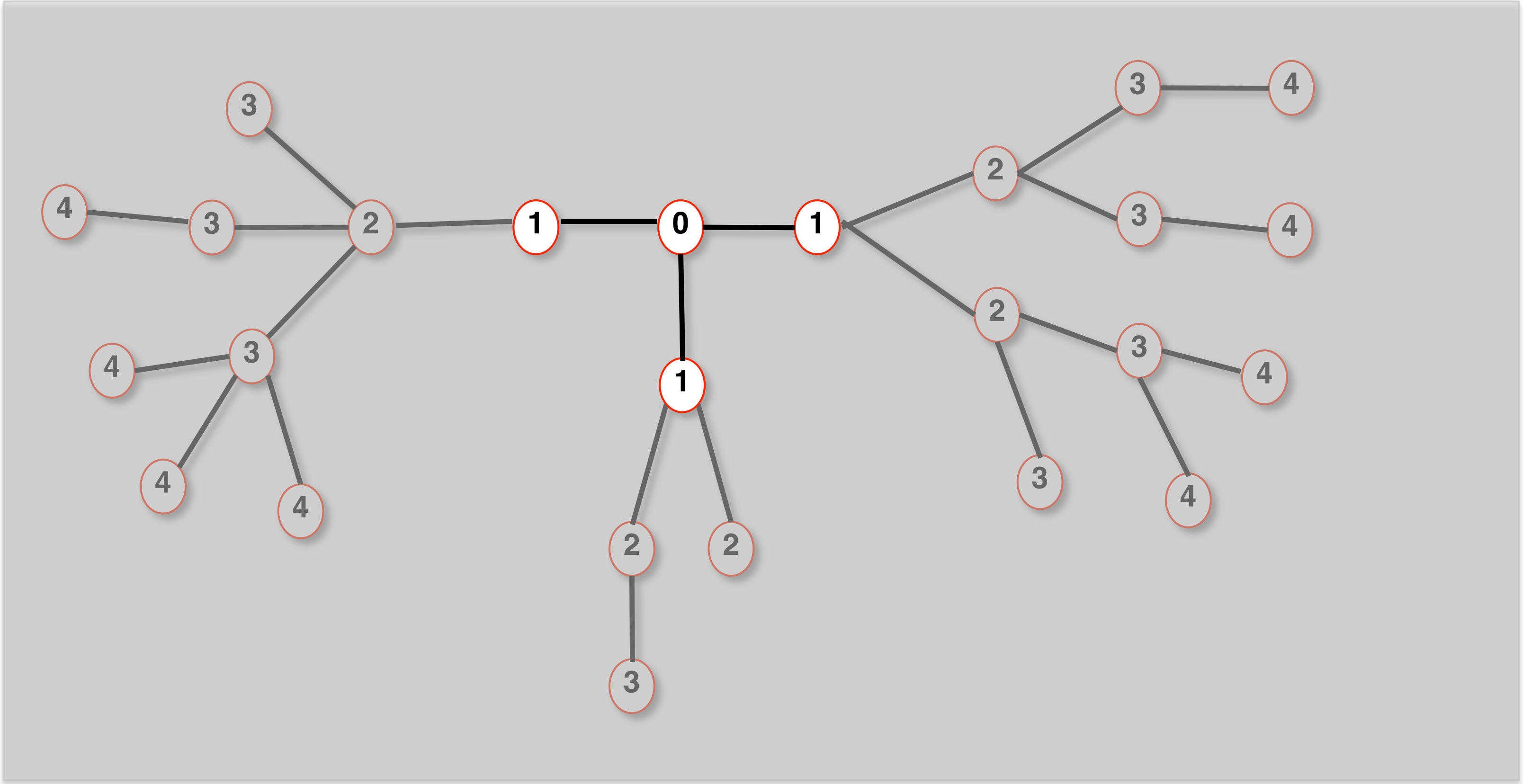
1.Start at 0.



FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

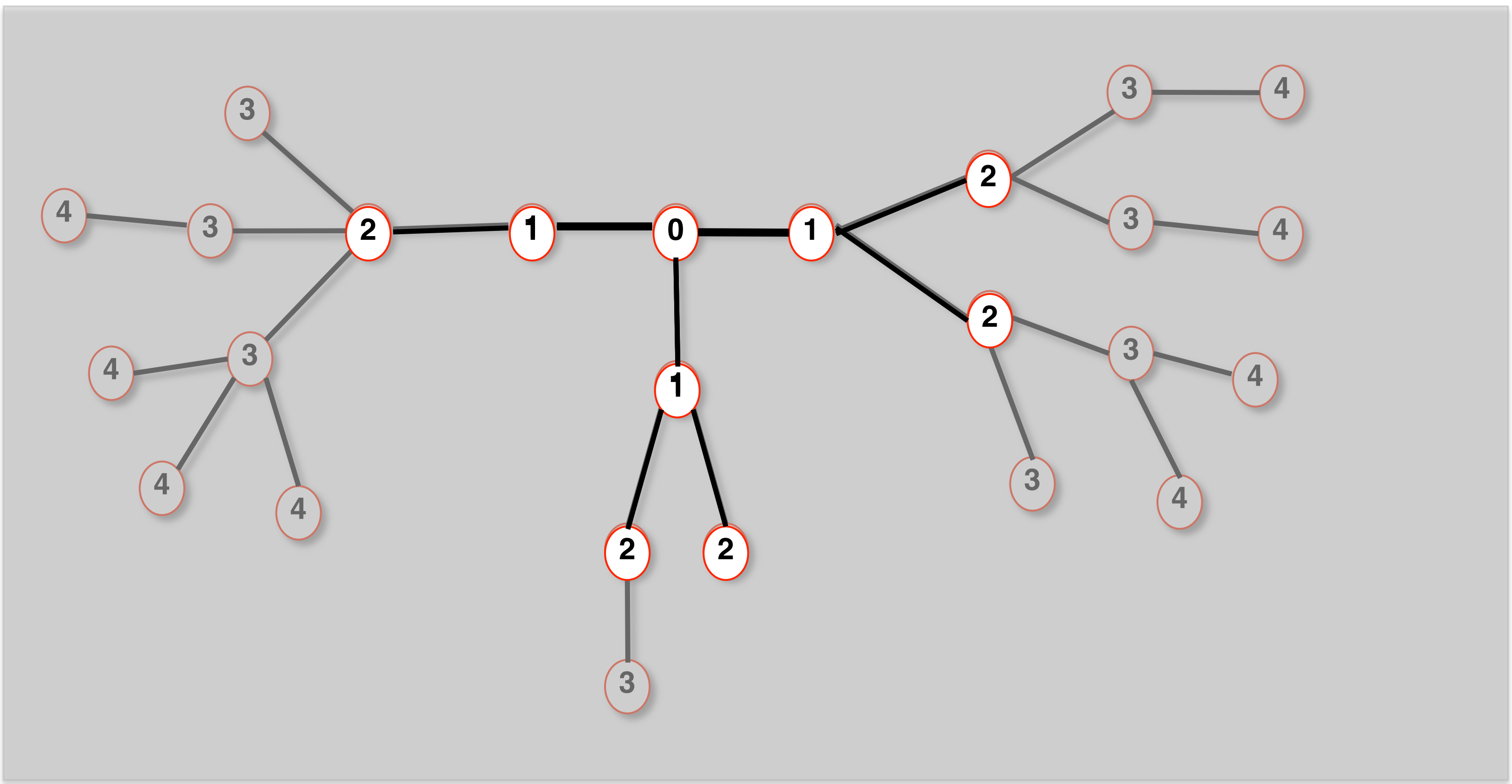
1. Start at 0.
2. Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.



FINDING DISTANCES: BREADTH FIRST SEARCH

Distance between node 0 and node 4:

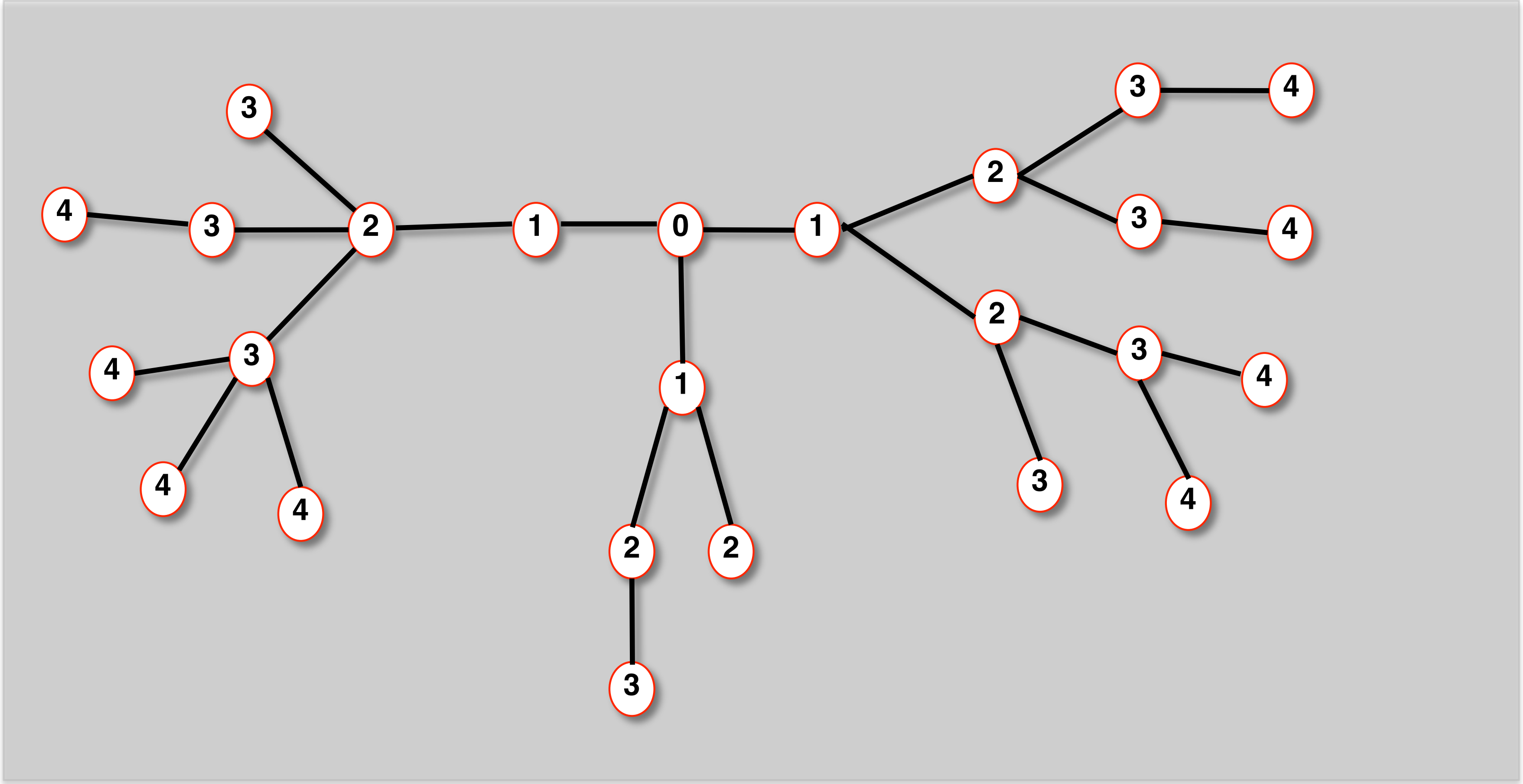
1. Start at 0.
2. Find the nodes adjacent to 0. Mark them as at distance 1. Put them in a queue.
3. Take the first node out of the queue. Find the unmarked nodes adjacent to it in the graph. Mark them with the label of 2. Put them in the queue.



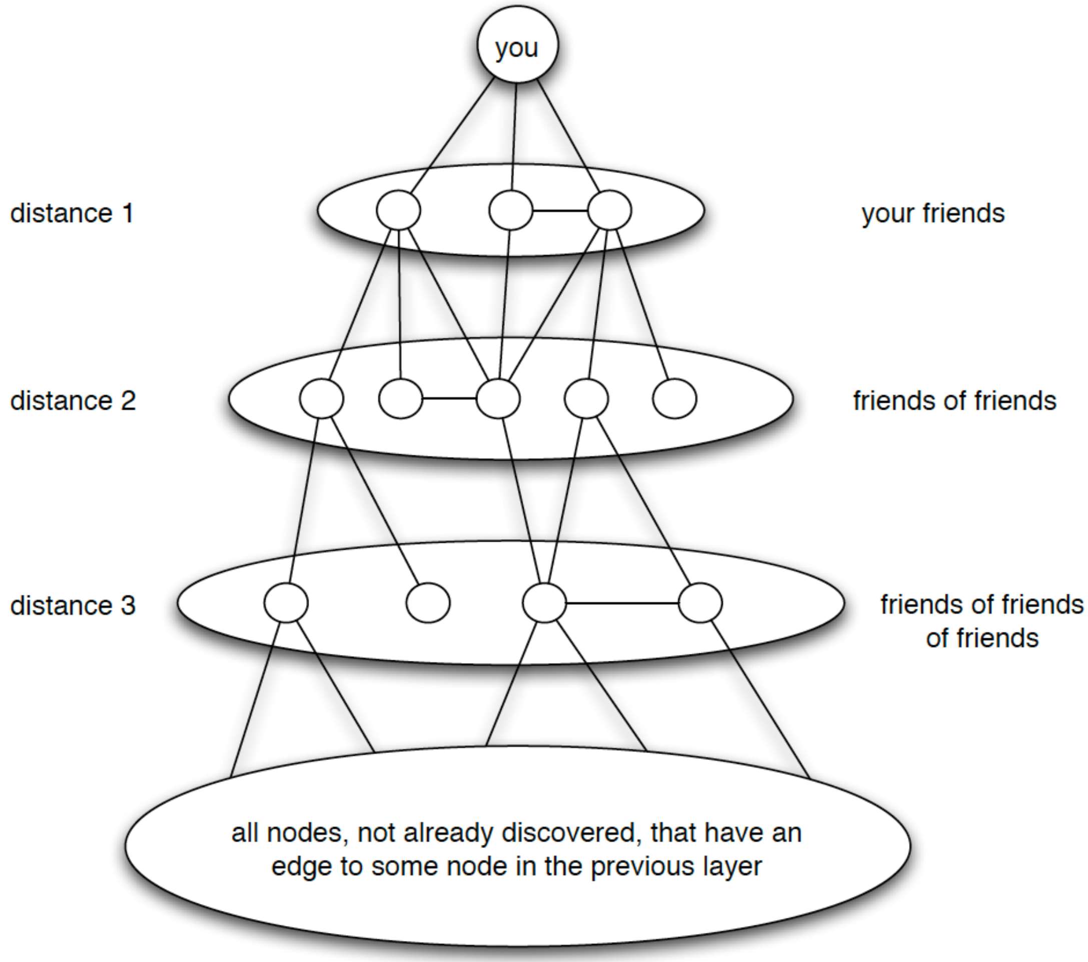
FINDING DISTANCES: BREADTH FIRST SEARCH

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.



FINDING DISTANCES: BREADTH FIRST SEARCH



The computational 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)

TESTING BIPARTITENESS

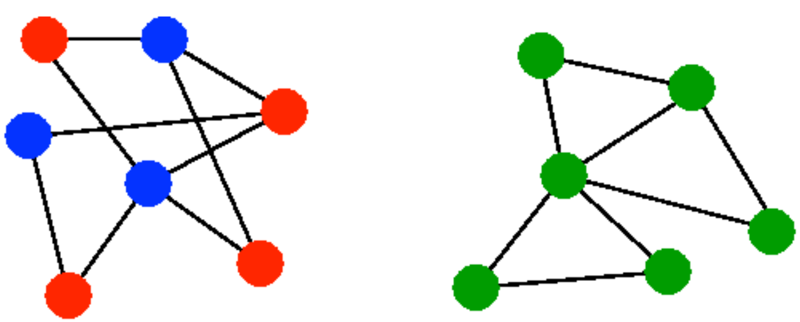
BFS can be used to test bipartiteness, by starting the search at any vertex and giving alternating labels to the vertices visited during the search.

That is, give label 0 to the starting vertex, 1 to all its neighbors, 0 to those neighbors' neighbors, and so on.

If at any step a vertex has (visited) neighbors with the same label as itself, then the graph is not bipartite.

If the search ends without such a situation occurring, then the graph is bipartite.

Try here!



A) A Bipartite Graph   B) A non–Bipartite Graph

Note
A graph is bipartite
iff
it contains no odd cycle.

NETWORK DIAMETER AND AVERAGE DISTANCE

Diameter (d_{max}): the maximum distance between any pair of nodes in the graph.

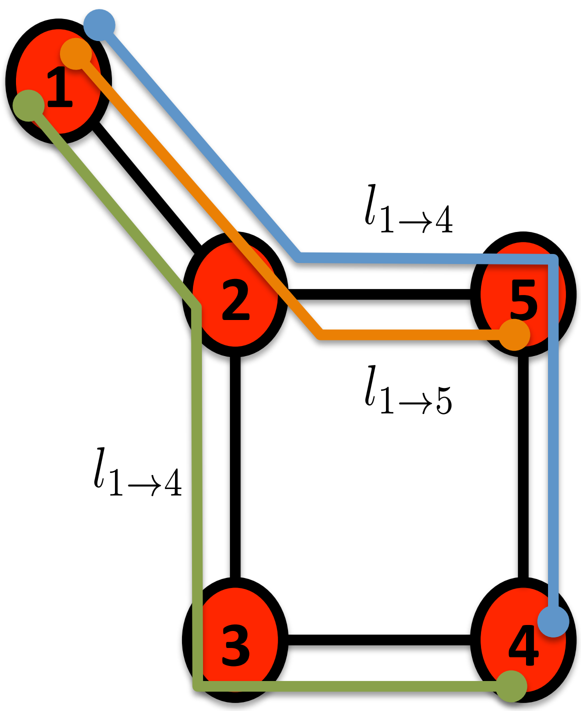
$$d_{\max} \equiv \max_{i \neq j} d_{ij}$$

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

Average distance ($\langle d \rangle$): for a connected graph

$$\langle d \rangle \equiv \frac{1}{N(N-1)} \sum_{i} \sum_{j \neq i} d_{ij}$$

Shortest Path



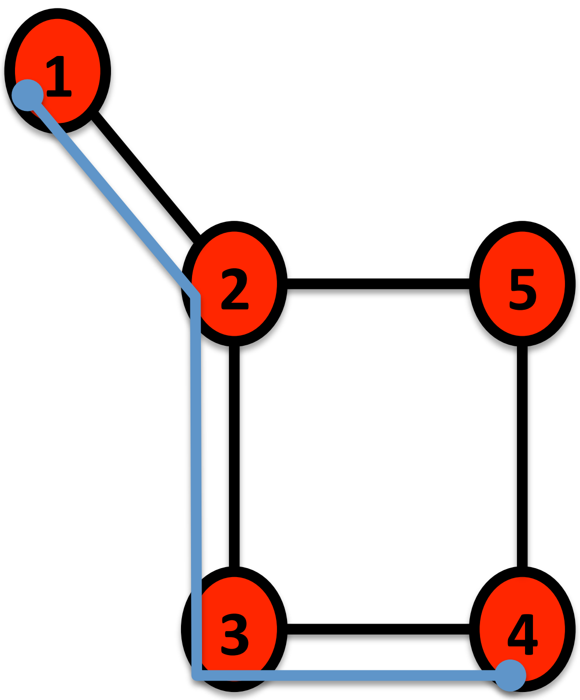
$$l_{1\to4} = 3$$

$$l_{1\to5} = 2$$

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

PATHOLOGY: summary

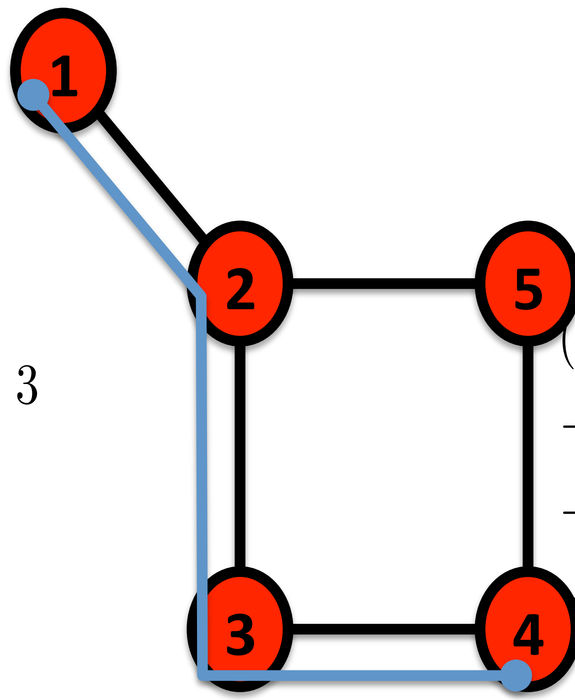
Diameter



$$l_{1\to 4} = 3$$

The longest shortest path in a graph

Average Path Length

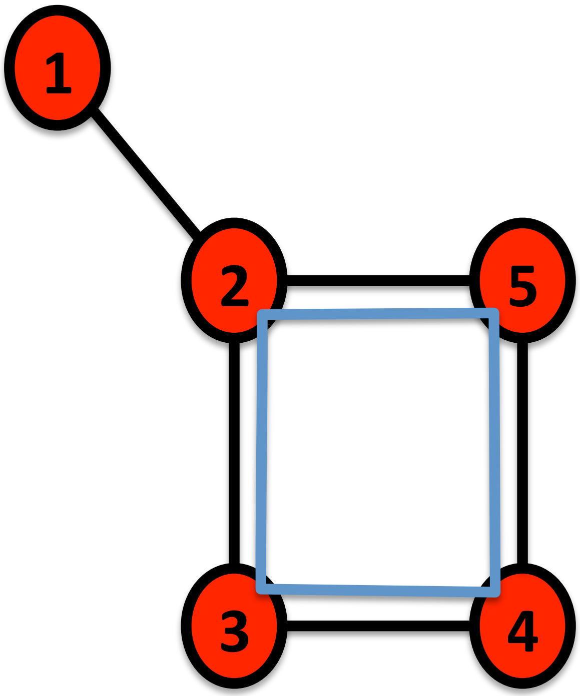


$$(l_{1\to 2} + l_{1\to 3} + l_{1\to 4} + l_{1\to 5} + l_{2\to 3} + l_{2\to 4} + l_{2\to 5} + l_{3\to 4} + l_{3\to 5} + l_{4\to 5})/10 = 1.6$$

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

PATHOLOGY: summary

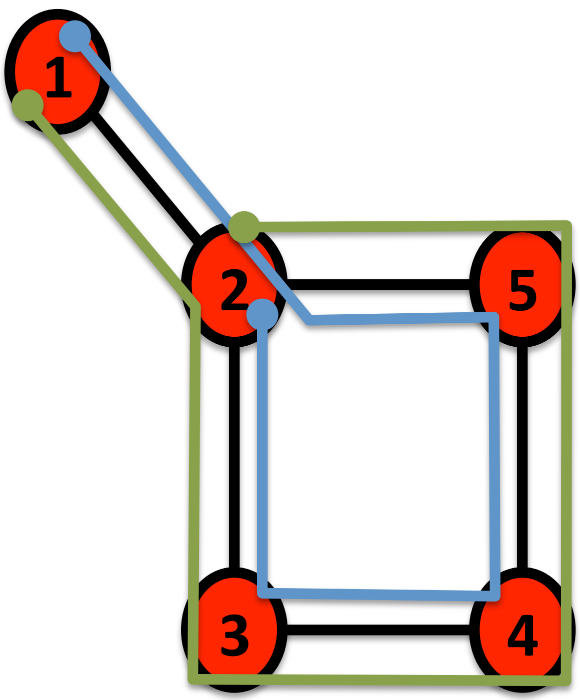
Cycle



A path with the same start and end node.

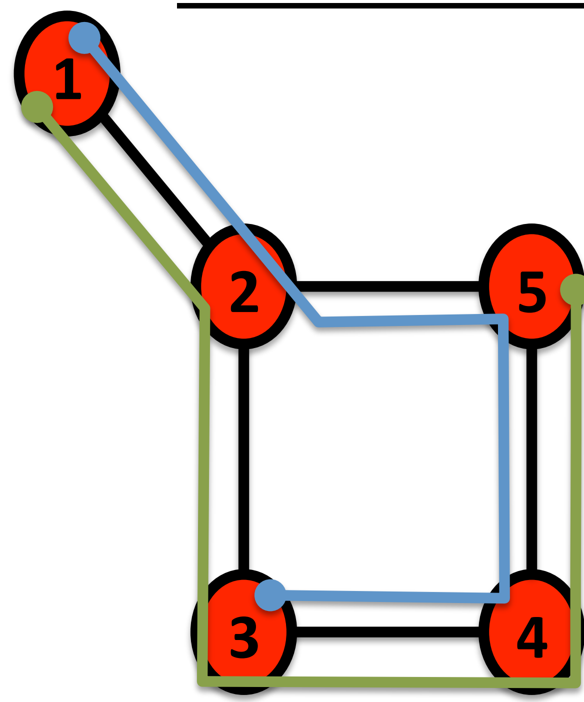
PATHOLOGY: summary

Eulerian Path



A path that traverses each link exactly once.

Hamiltonian Path

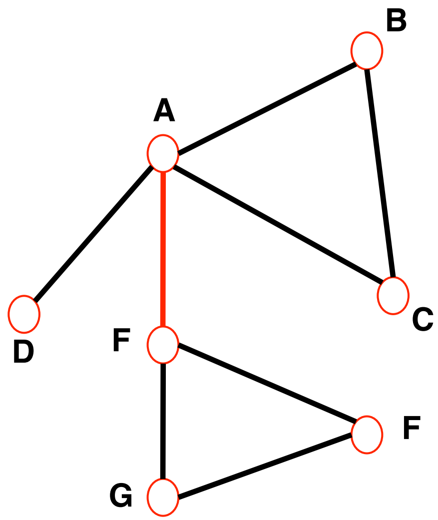


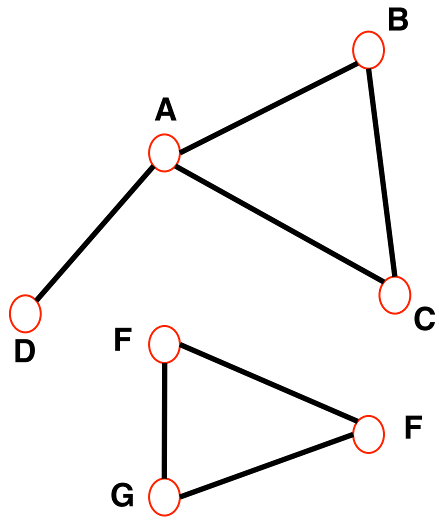
A path that visits each node exactly once.

CONNECTEDNESS

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.



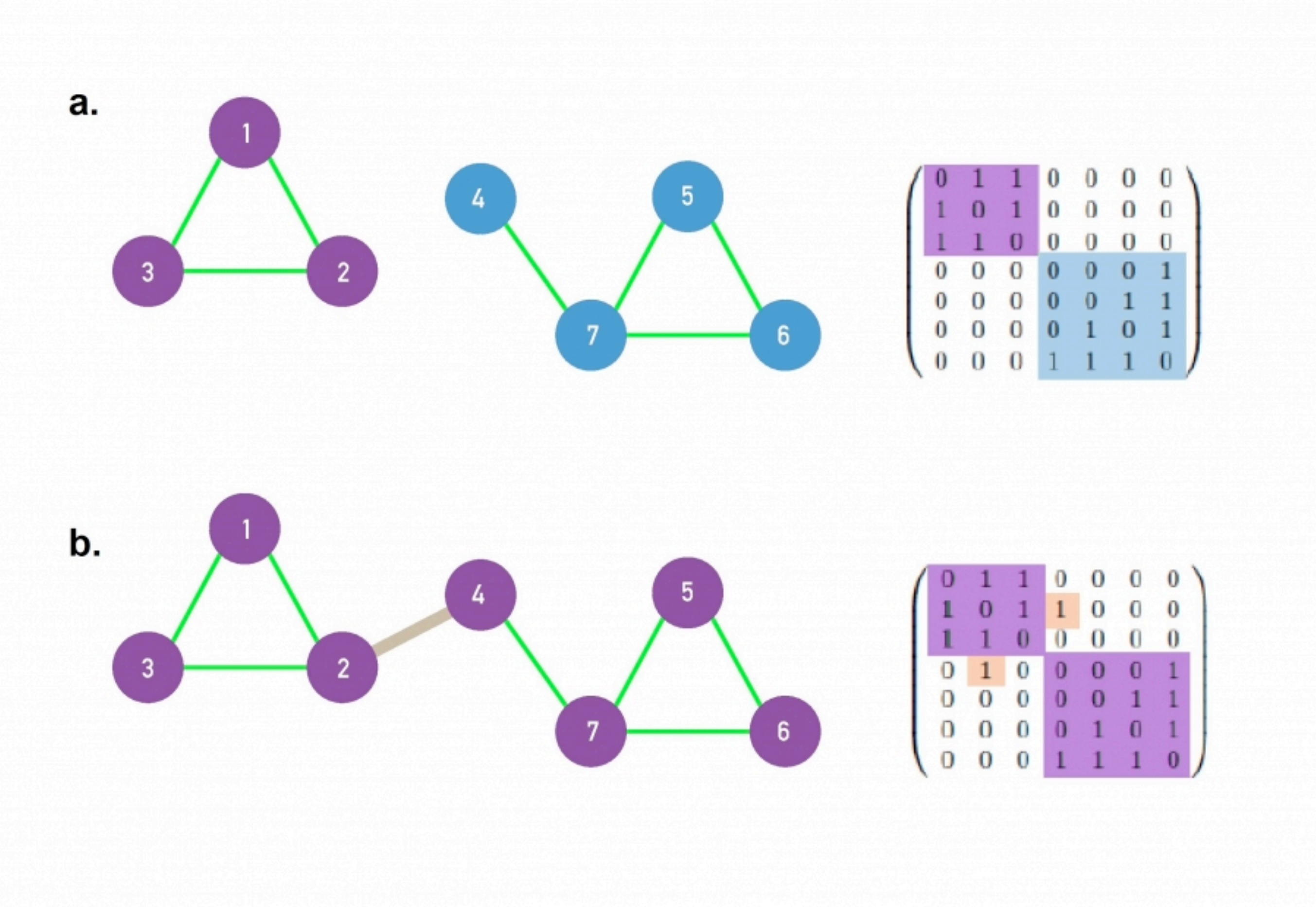


Largest Component:
Giant Component

The rest: **Isolates**

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

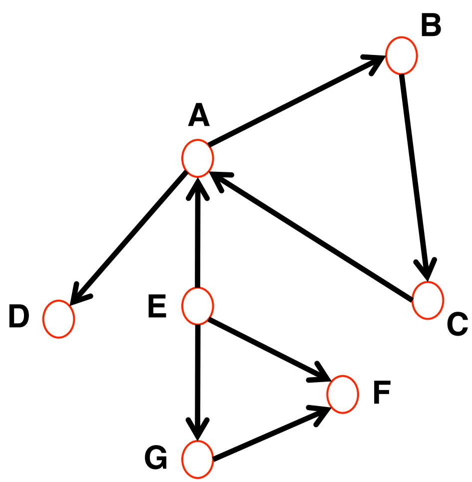
The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:

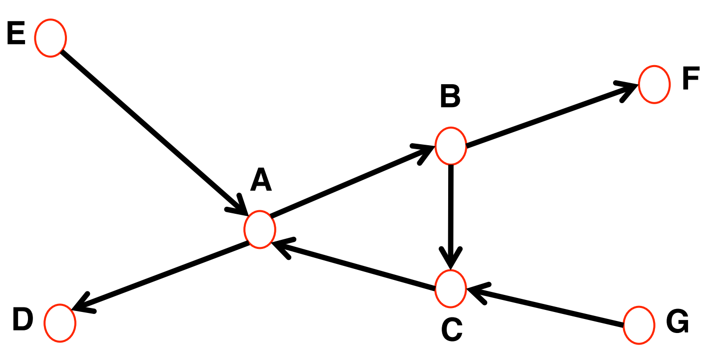


CONNECTIVITY OF DIRECTED GRAPHS

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.





Box 2.6

Finding the Connected Components of a Network

- Start from a randomly chosen node i and perform a BFS (BOX 2.5). Label all nodes reached this way with $n = 1$.
- 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.
- Increase the label $n \to n + 1$. Choose an unmarked node j, label it with n. Use BFS to find all nodes reachable from j, label them all with n. Return to step 2.

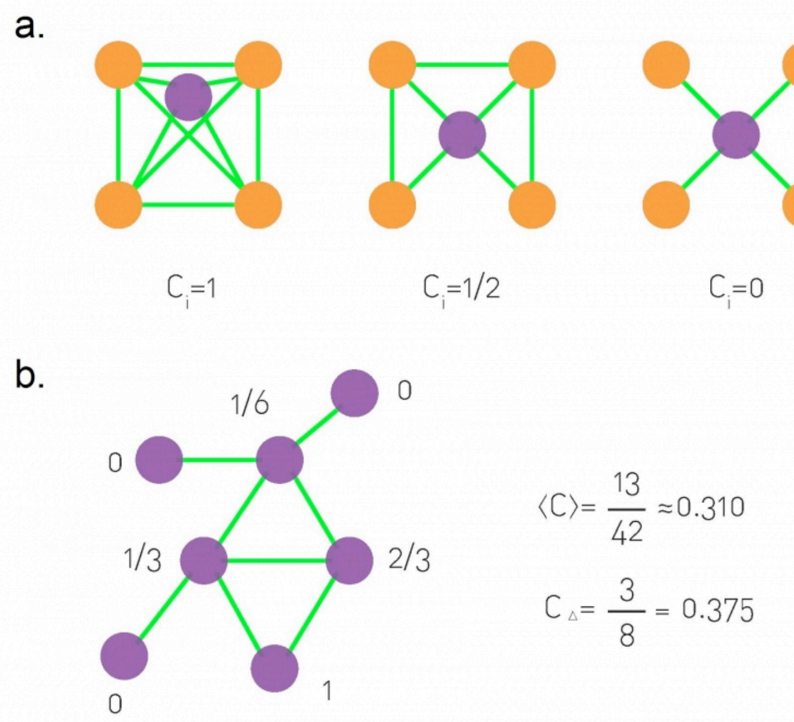
Clustering coefficient and cliques

CLUSTERING COEFFICIENT

What fraction of your neighbors are connected?

- Node i with degree k_i
- e_i = number of links between the k_i neighbors of i

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

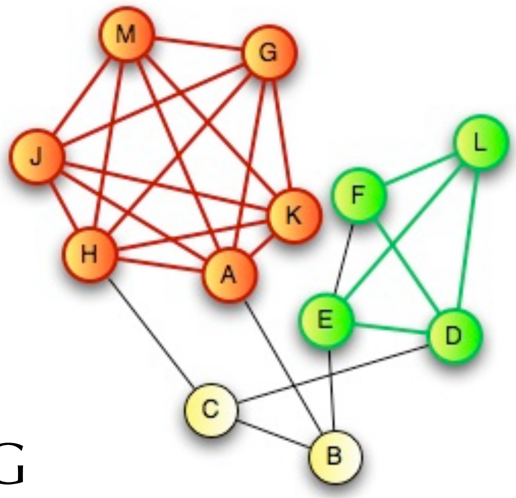


Note: $0 \le C_i \le 1$

Cliques

Given an unweighted undirected graph G=(V,E):

- A clique is a subset of mutually adjacent vertices
- A maximal clique is a clique that is not contained in a larger one
- A maximum clique is a clique having largest cardinality



The clique number, denote $\omega(G)$, is the cardinality of a maximum clique.

Independent set: clique on the complement of G

summary

THREE CENTRAL QUANTITIES IN NETWORK SCIENCE

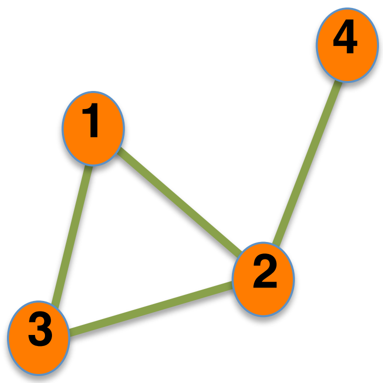
Degree distribution: **P(k)**

Path length: **<*d*>**

Clustering coefficient:

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

Undirected



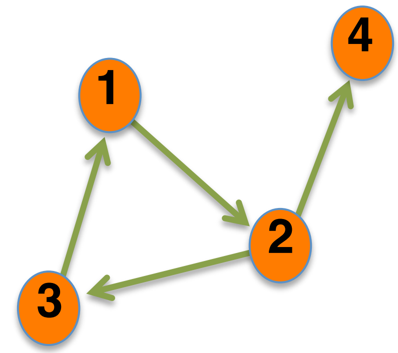
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

Actor network, protein-protein interactions

Directed



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

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

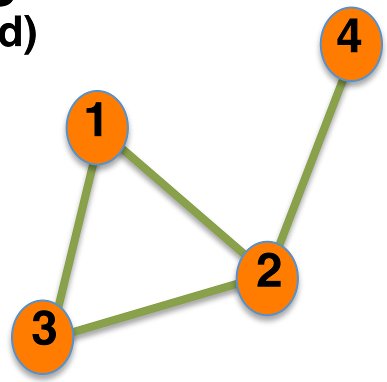
$$L = \sum_{i,j=1}^{N} A_{ij} \qquad <k> = \frac{L}{N}$$

WWW, citation networks

GRAPHOLOGY 2

Unweighted

(undirected)



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

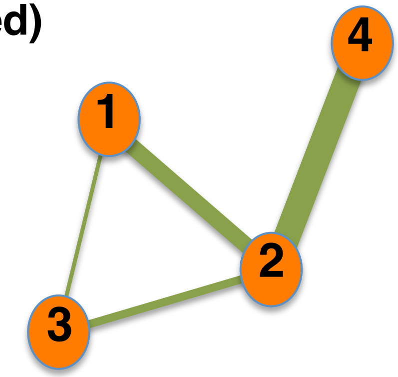
$$A_{ii} = 0 \qquad A_{ij} = A_{ji}$$

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

protein-protein interactions, www

Weighted

(undirected)



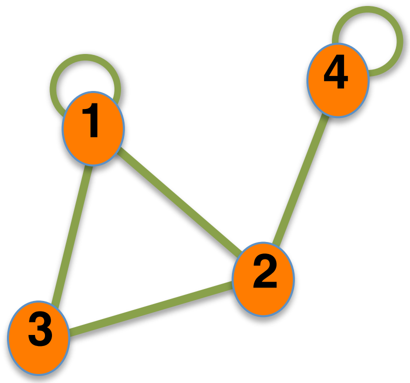
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

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

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

Call Graph, metabolic networks

Self-interactions



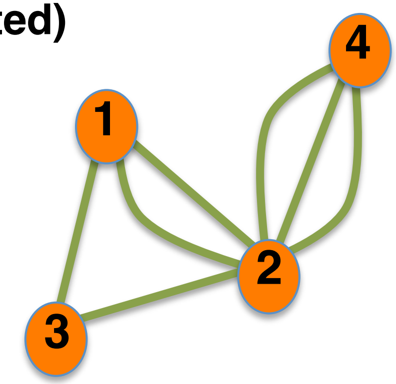
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

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

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

Protein interaction network, www

Multigraph
(undirected)



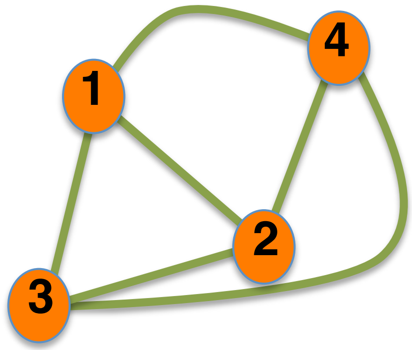
$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

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

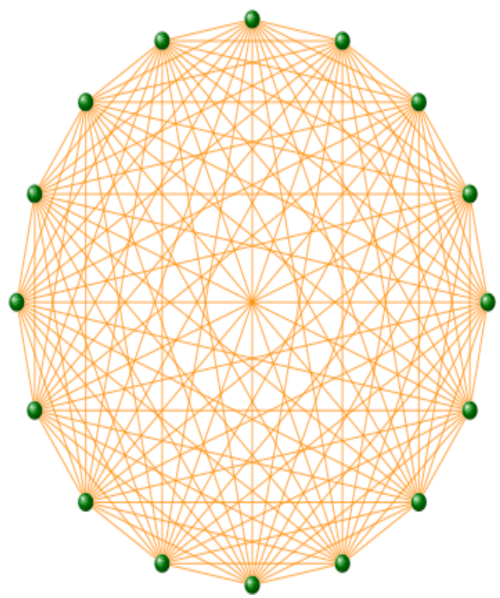
$$L = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad <k>= \frac{2L}{N}$$

Social networks, collaboration networks

Complete Graph (Clique)
(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

GRAPHOLOGY: Real networks can have multiple characteristics

WWW > **d**irected multigraph with self-interactions

Protein Interactions > **u**ndirected unweighted with self-interactions

Collaboration network > **u**ndirected multigraph or weighted.

Mobile phone calls > directed, weighted.

Facebook Friendship links > **u**ndirected, unweighted.