CptS 591: Elements of Network Science

Basic Network Properties

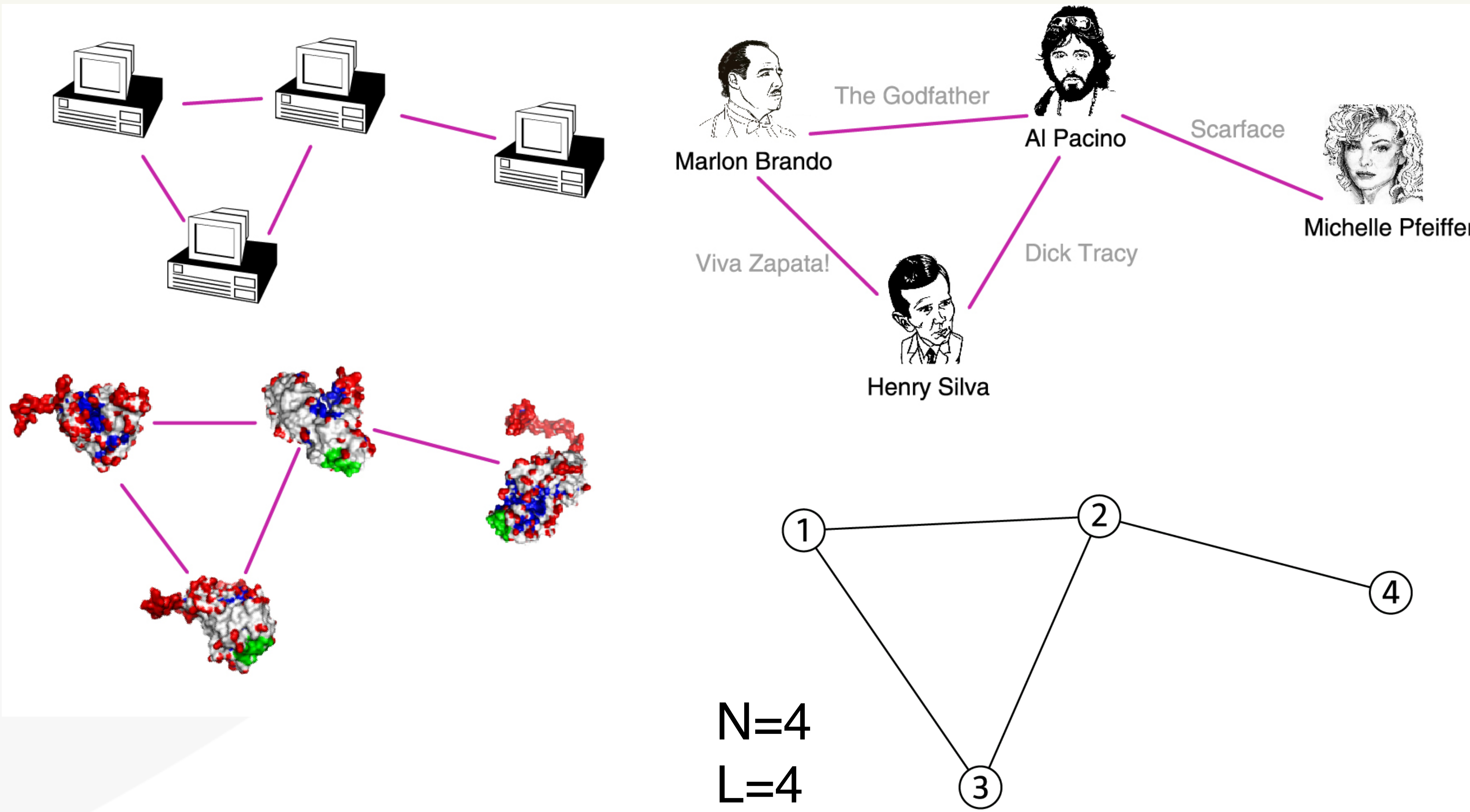
Outline

- Three central quantities:
  - Degree distribution
  - Network diameter and average path length
  - Clustering coefficient

- Associated side discussions:
  - Networks and Adjacency Matrices
  - Types of networks

- Note:
  - This material is adapted from slides of the "Graph Theory" section/chapter in the "Network Science" course/book by Albert-László Barabási.

A Common Language



Choosing a Proper Representation

- The choice of the proper network representation determines our ability to use network science 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.

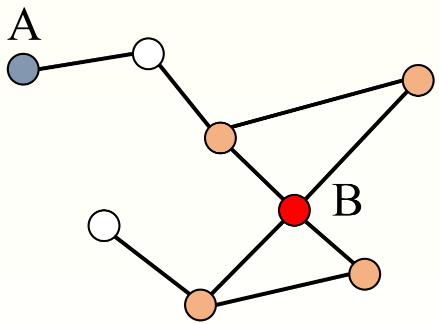
Basic Network Properties

Degree, degree distribution.

Node Degrees

Undirected

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

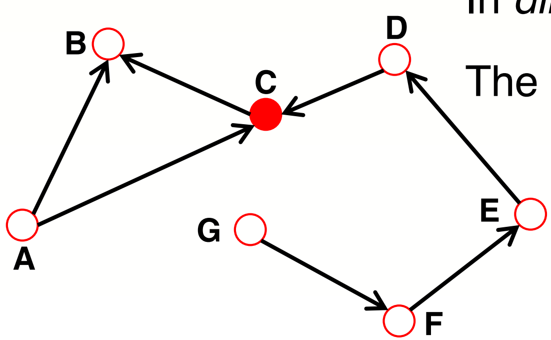


$$k_A = 1 \qquad k_B = 4$$

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 $x_1, \dots, x_N$:

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:

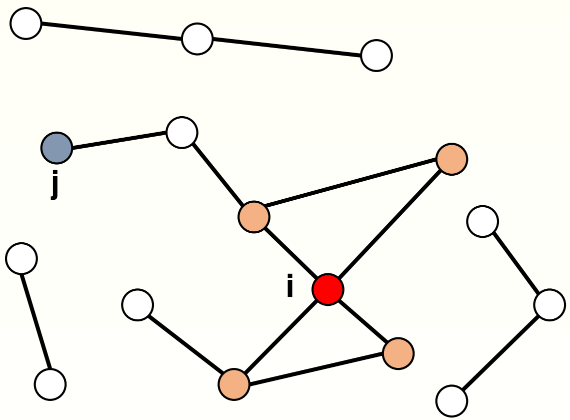
$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

where p_x follows

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

Average Degree

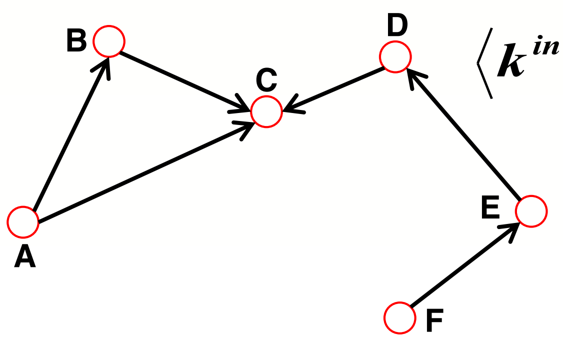
Undirected



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

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

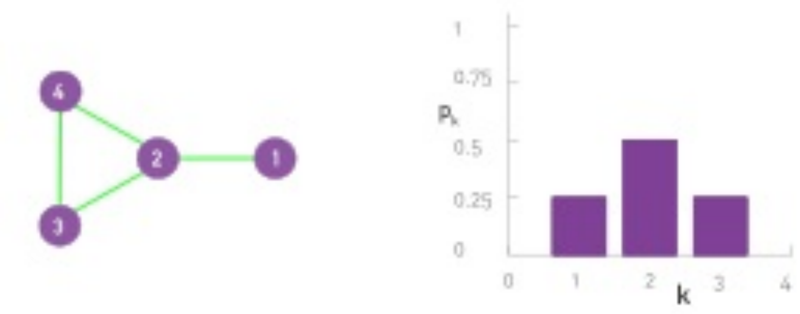
Degree Distribution

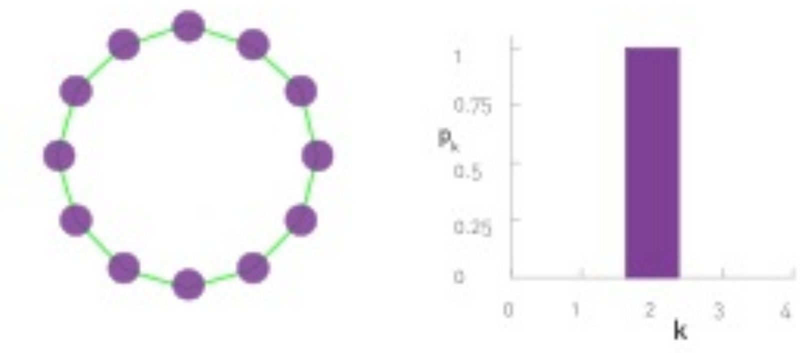
Degree distribution

P(k): probability that a

randomly chosen node

has degree *k*

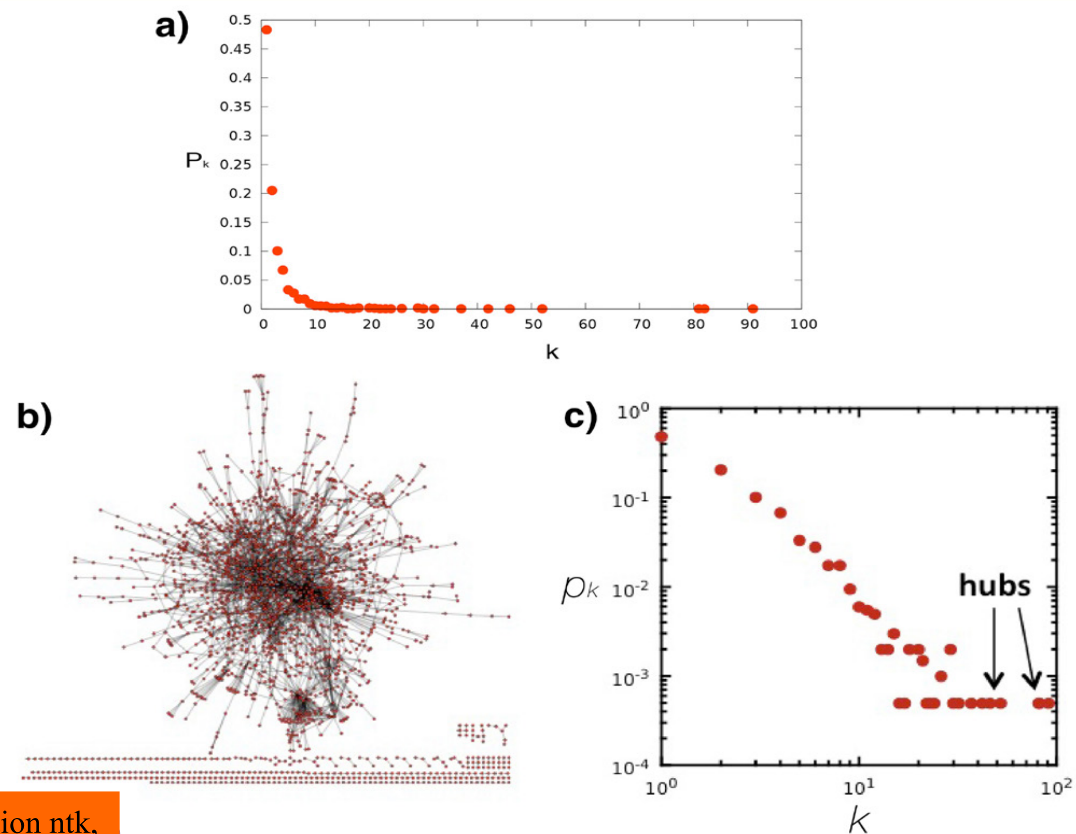




N_k = # nodes with degree k

$P(k) = N_k / N$ ➔ plot

Degree Distribution



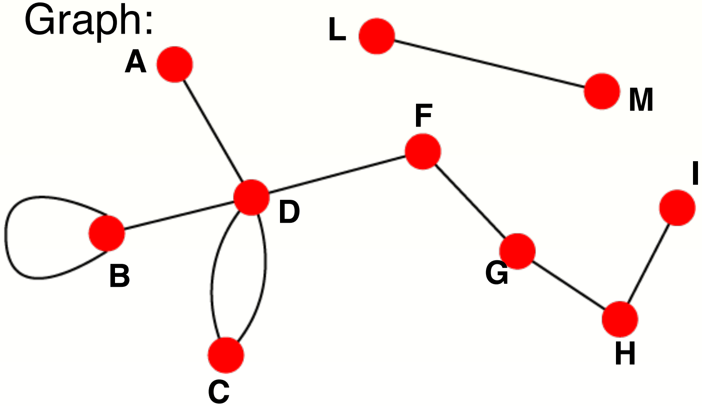
Protein interaction ntk, roughly 2k nodes.

Undirected vs Directed Networks

Undirected

Links: undirected (*symmetrical*)

Graph:



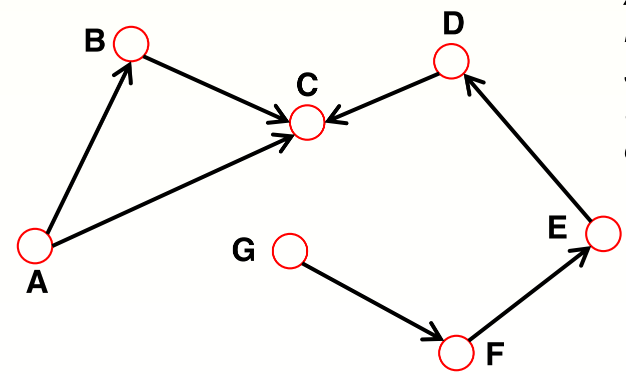
Undirected links :

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

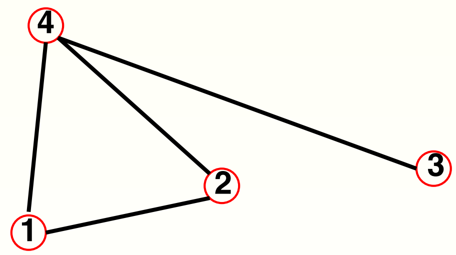
(Reference) Networks

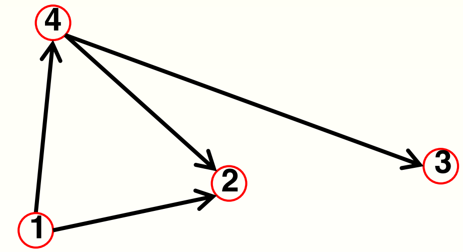
NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L
Internet	Routers	Internet connections	Undirected	192,244	609,066
WWW	Webpages	Links	Directed	325,729	1,497,134
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826
Email	Email addresses	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

Networks and Matrices

Adjacency Matrices

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.

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

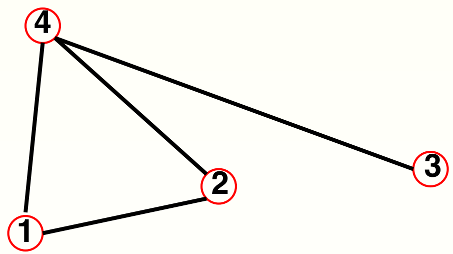
Note that for a directed graph (right) the matrix is not symmetric.

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

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

Adjacency Matrix and Node Degrees

Undirected



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

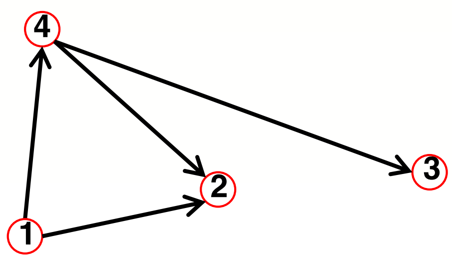
$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

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

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

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

Directed



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

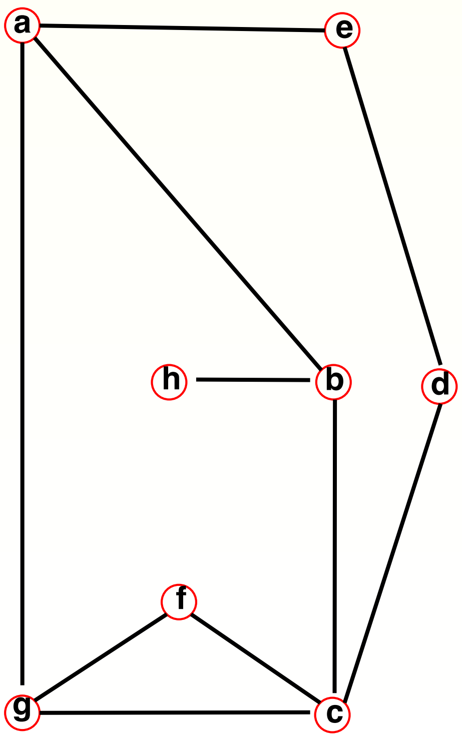
$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

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

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

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

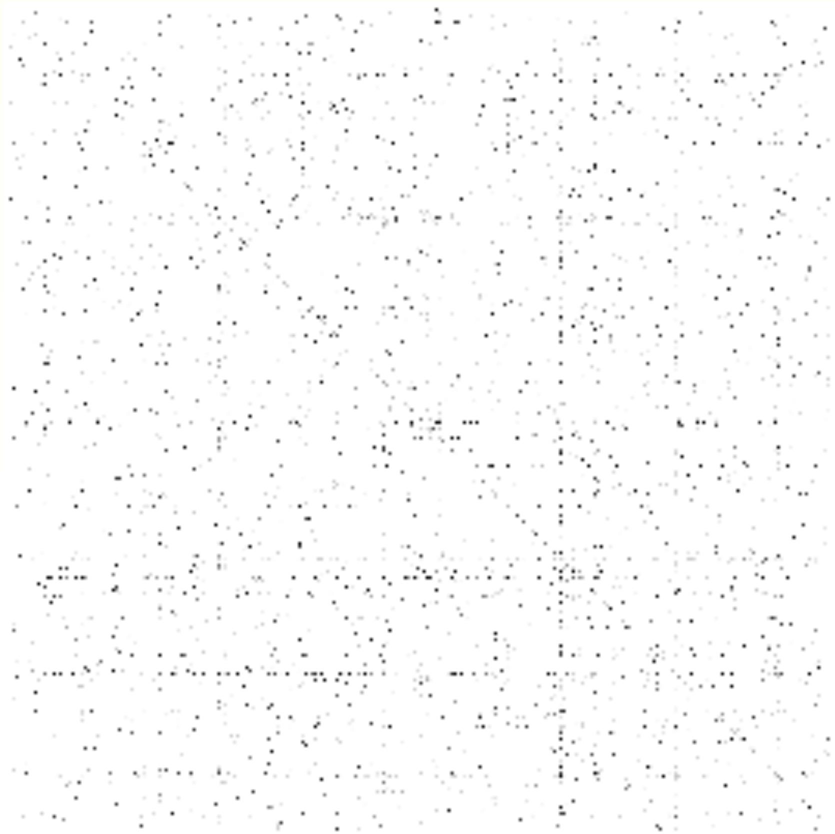
Adjacency Matrix



	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

The adjacency matrix can take far more complicated forms for a larger network

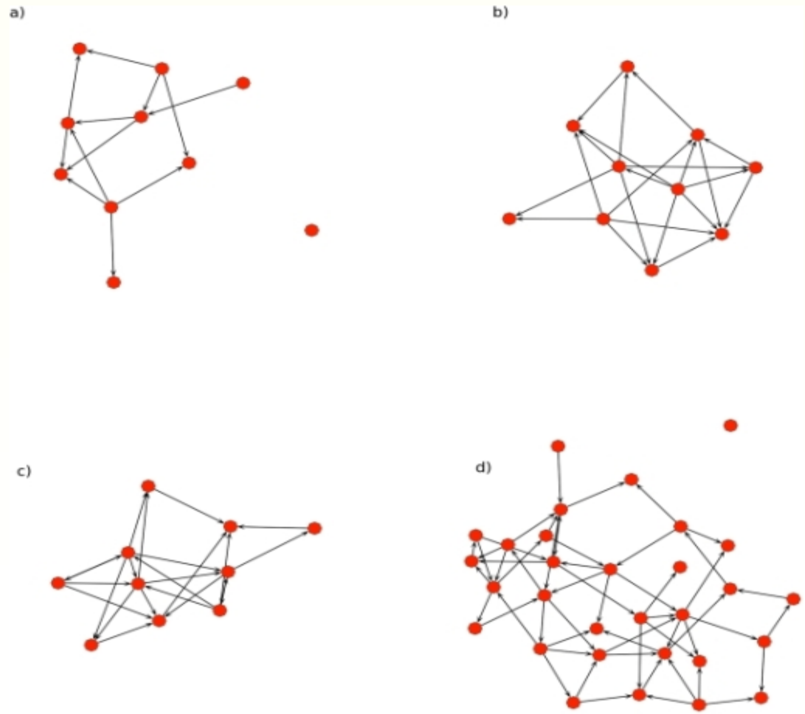
Adjacency Matrices are Sparse



Networks and Matrices

Sparseness

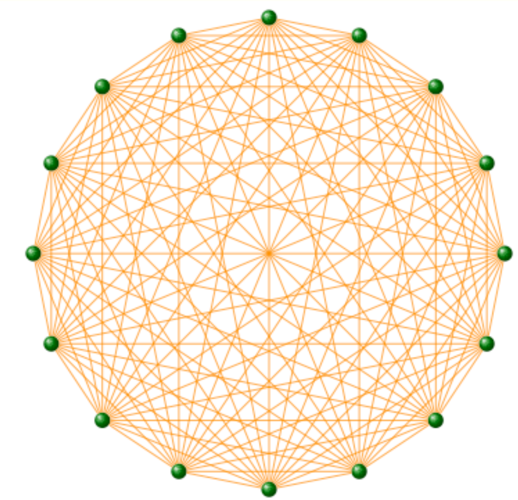
Real networks are sparse



Complete Graph (Clique)

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 d $L=L_{max}$ links is called a complete graph, and its average degree is **<k>=N-1**

Real World 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 x 10^{10}	<k>=3.9
Movie Actors:	N = 212,250;	L=6 10^6	L_{max}=1.8 x10^{13}	<k>=28.78

(Source: Albert, Barabasi, RMP2002)

Weighted and Unweighted Networks

Weighted and Unweighted Networks

A weighted network is represented using an adjacency matrix in which the entries are real numbers, rather than just 0 and 1.

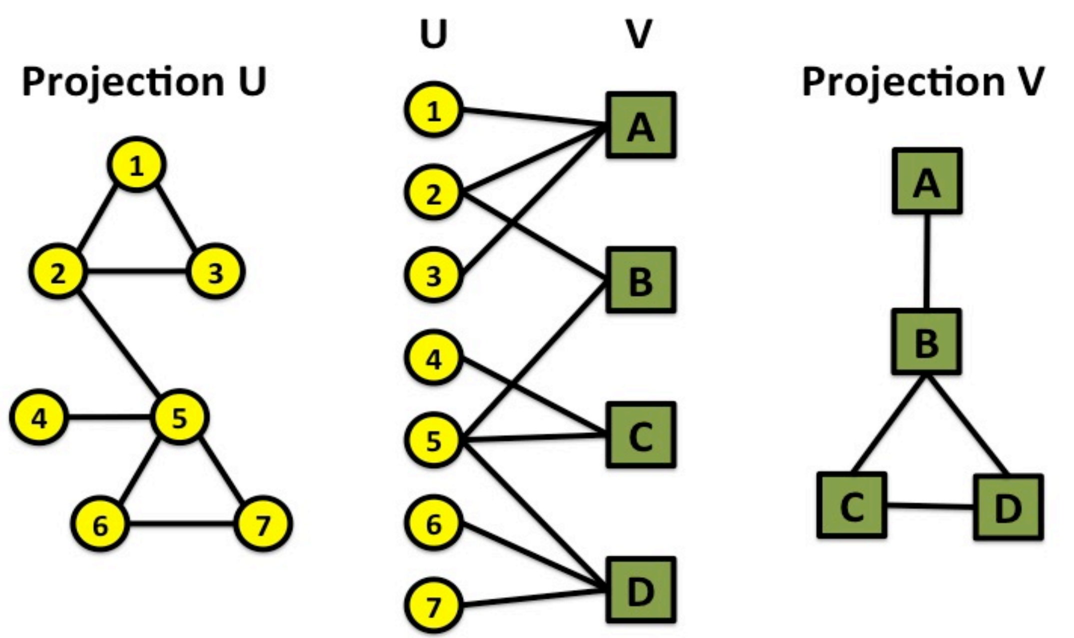
$$A_{ij} = w_{ij}$$

Network types

Bipartite Networks

Bipartite Graphs

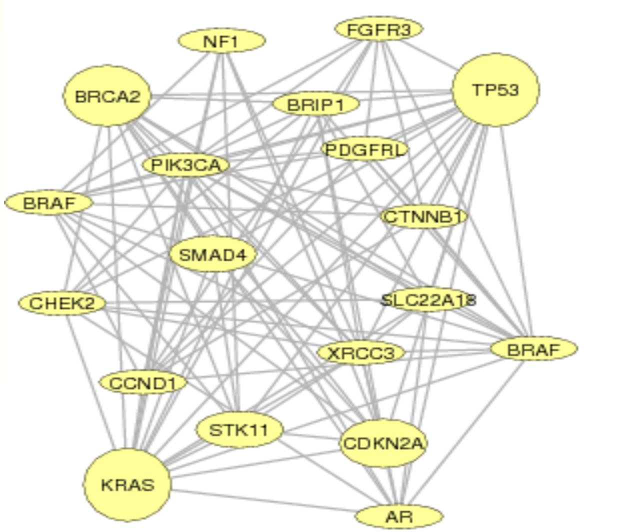
A bipartite graph is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V; that is, U and V are independent sets.



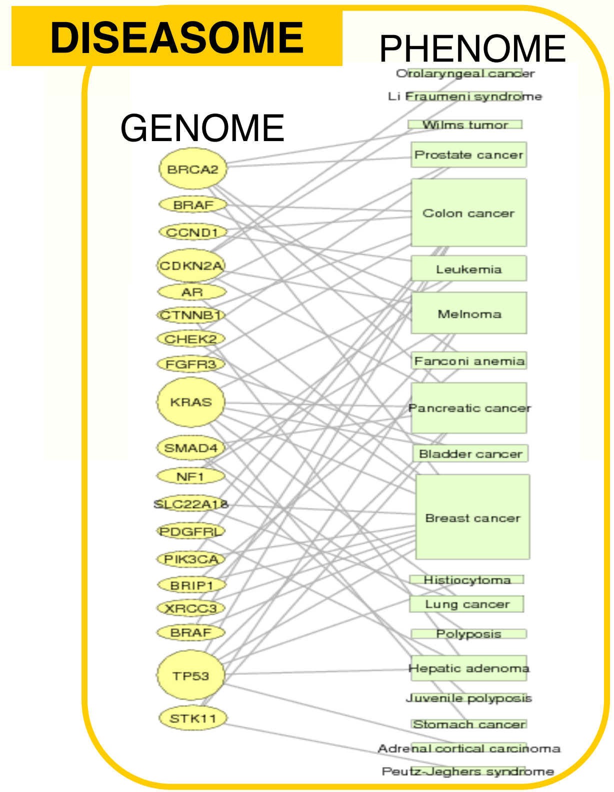
Examples:

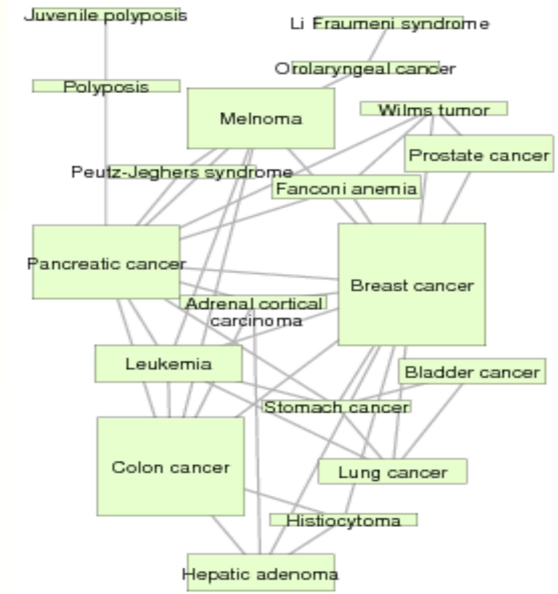
Hollywood actor network
Collaboration networks
Disease network (diseasome)

Gene Network – Disease Network



Gene network

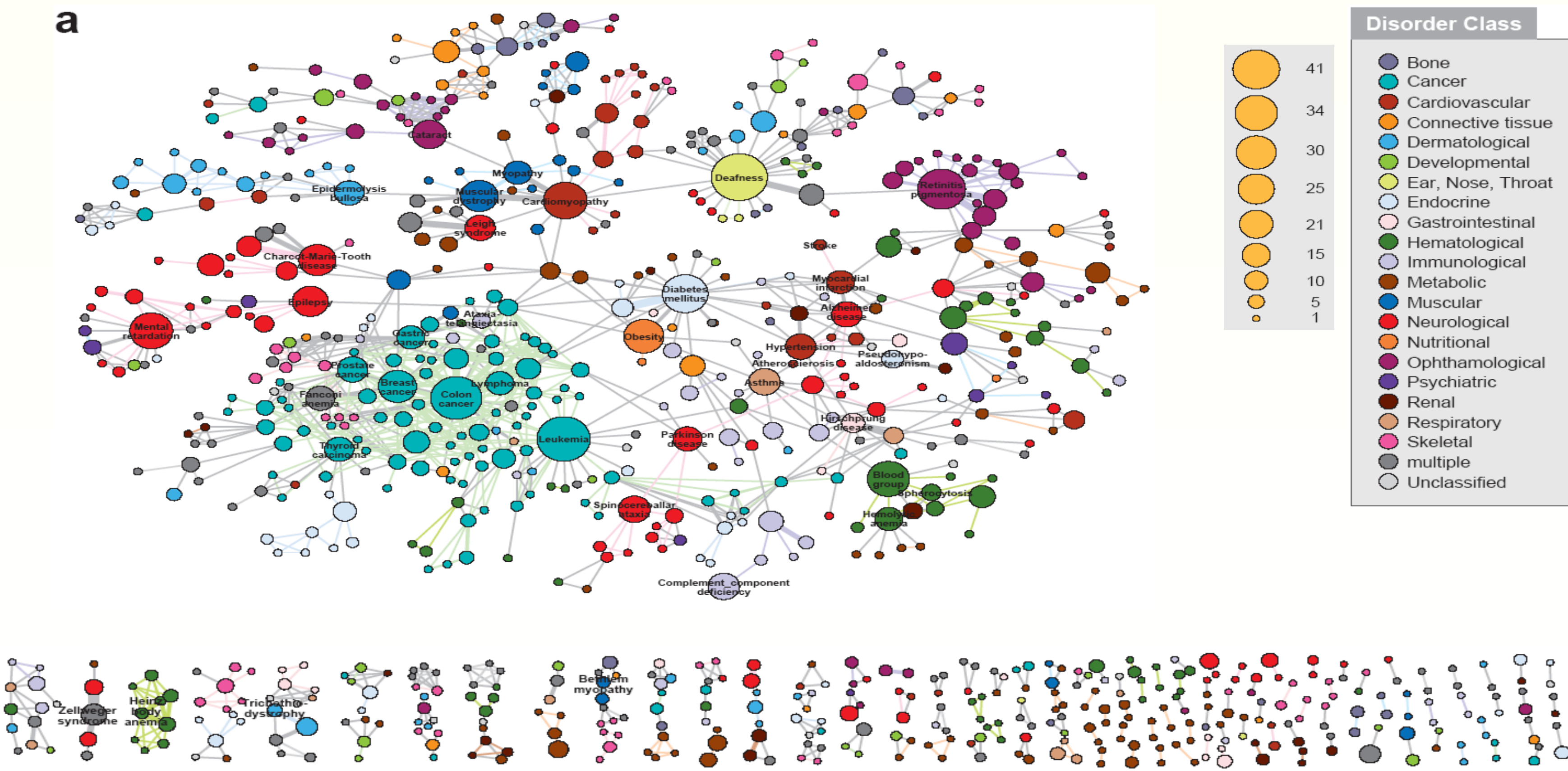




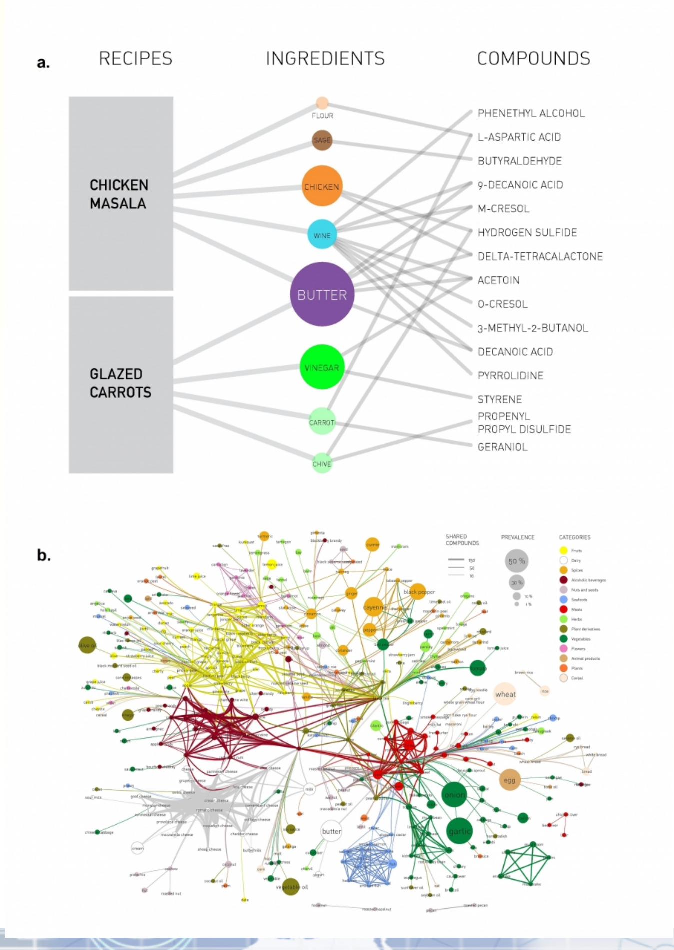
Disease network

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

Human Disease Network



Tripartite network: recipes, ingredients and compounds

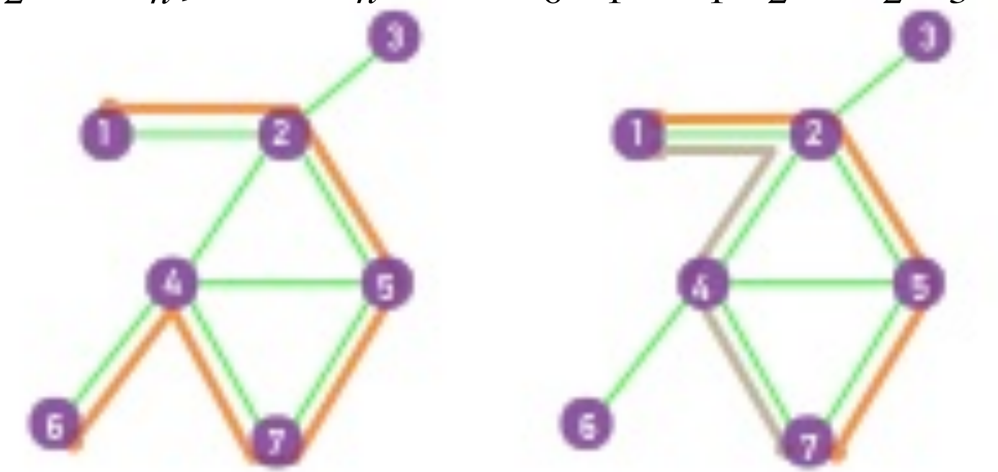


Paths

Paths (reminder)

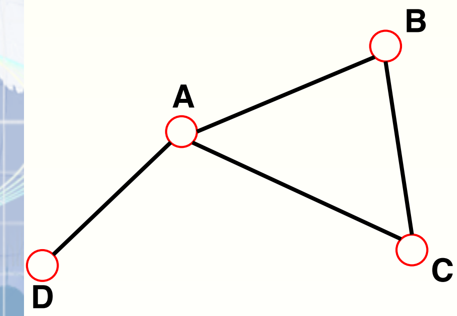
- A *path is* a sequence of nodes in which each node is adjacent to the next one
- P_{i_0,i_n} of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

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



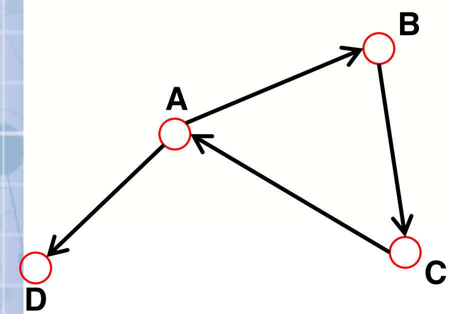
- In a directed network, the path can follow only the direction of an arrow.

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 Paths Between Two Nodes
Adjacency Matrix

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

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

<u>Length n=2:</u> 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:

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

<u>Length n:</u> 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.*

Network Diameter and Average Path

Diameter: $\boldsymbol{d_{max}}$ the maximum distance between any pair of nodes in the graph.

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

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

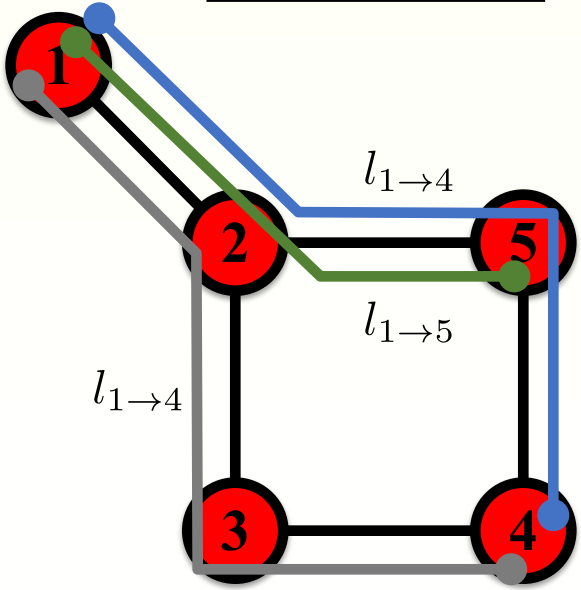
$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$

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

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

Paths: summary/example

Shortest Path



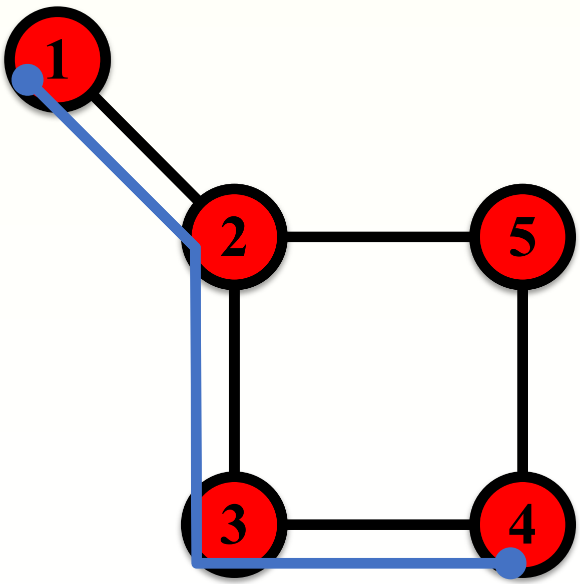
$l_{1\to4} = 3$

$l_{1\to5} = 2$

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

Paths: summary/example

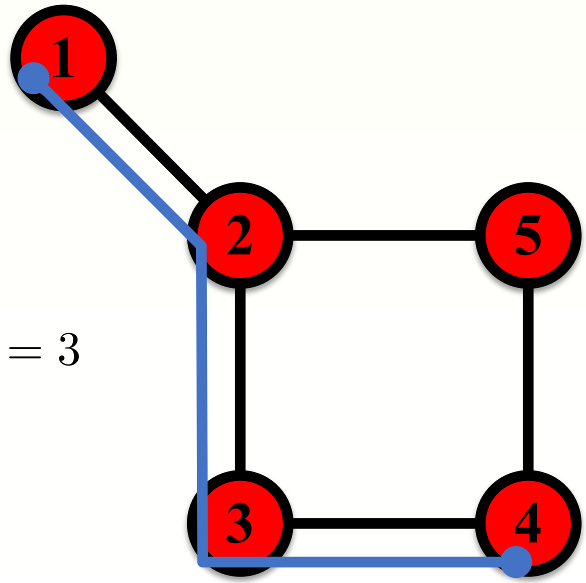
Diameter



$l_{1 \to 4} = 3$

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.