Social Network Analysis

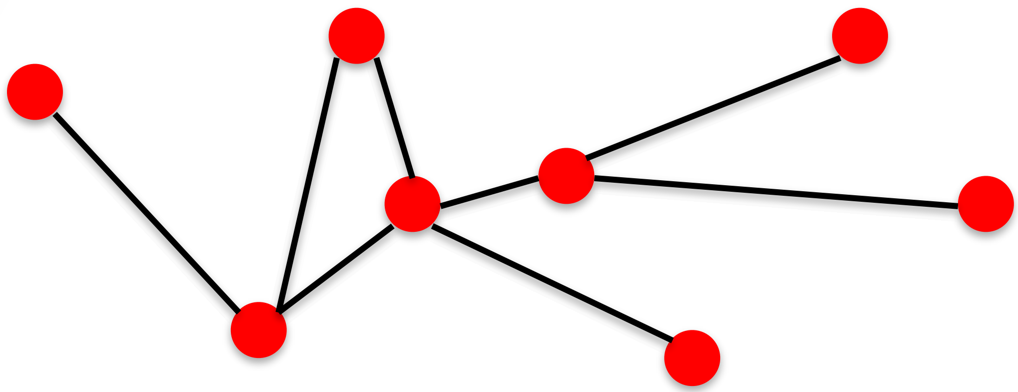
Preliminaries

Course Outline

- **Graph Theory and Social Networks**
- Visualizing Social Networks
- Game Theory
- Information Networks and the World Wide Web
- Network Dynamics
- Applications of SNA in various domains

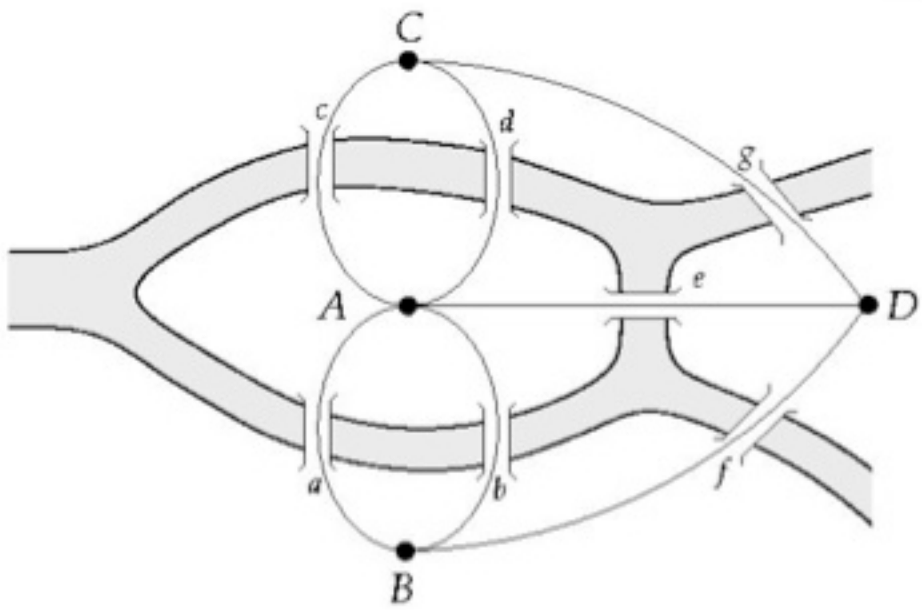
Graph Theory

A Graph (Network)



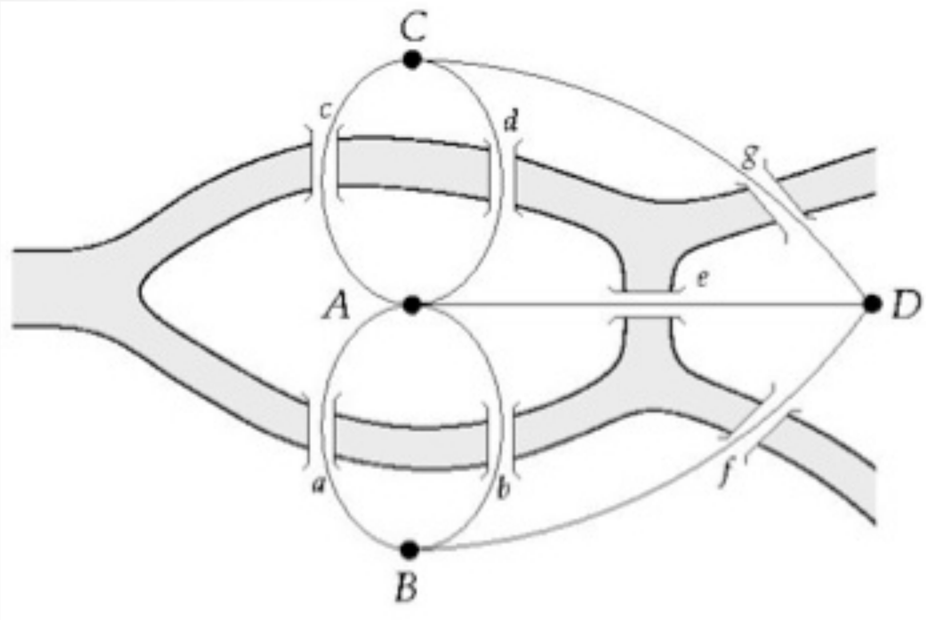
- **components**: nodes, vertices N/V
- **interactions**: links, edges L/E
- **system**: network, graph (N,L)/(V,E)

The Bridges of Königsberg



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

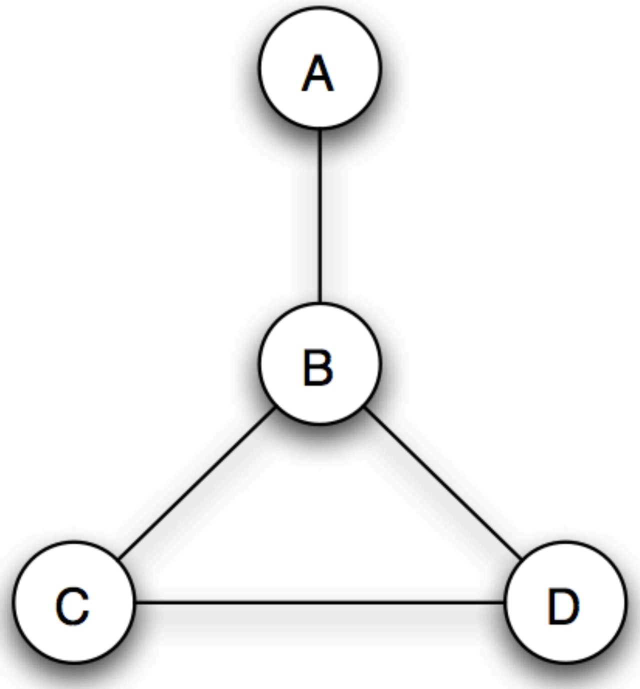
The Bridges of Königsberg



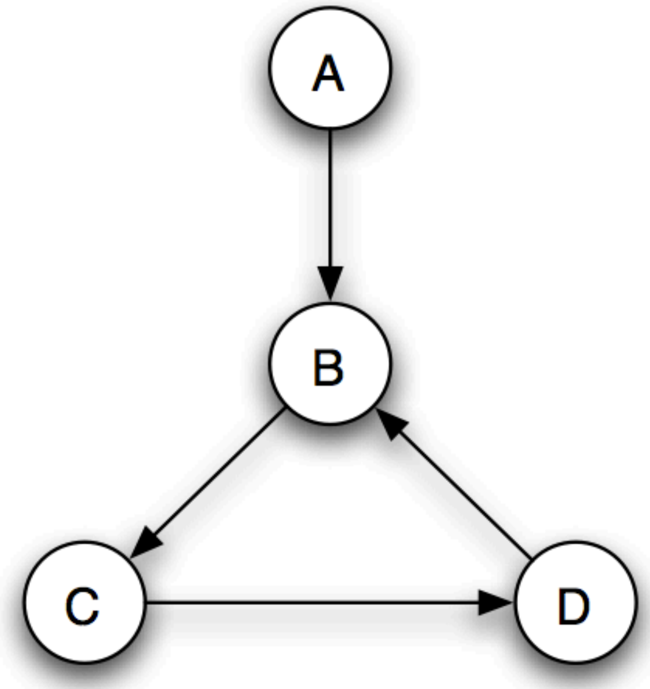
Can one walk across the seven bridges and never cross the same bridge twice?

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, it has at least one path.



(a) *A graph on 4 nodes.*



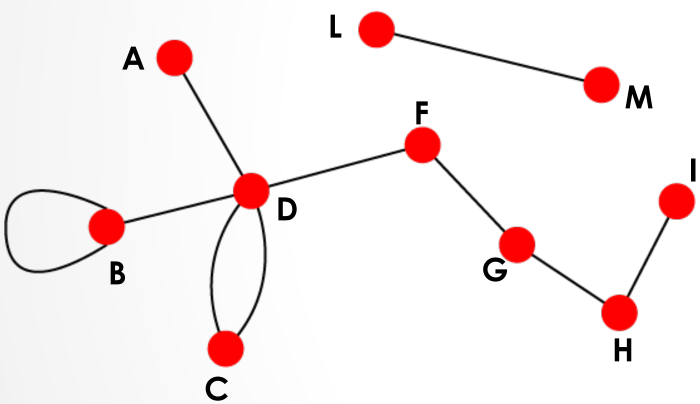
(b) *A directed graph on 4 nodes.*

Directed & Undirected Graphs

Undirected

Links: undirected (*symmetrical*)

Graph:

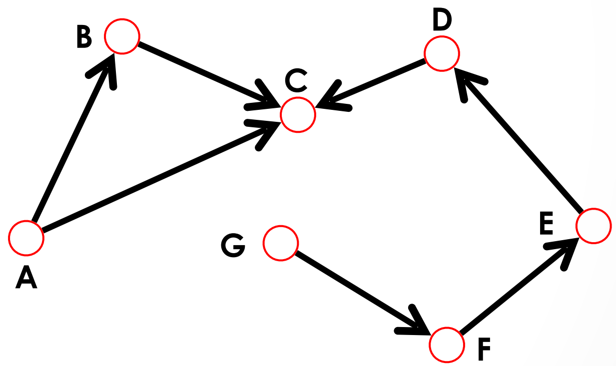


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:



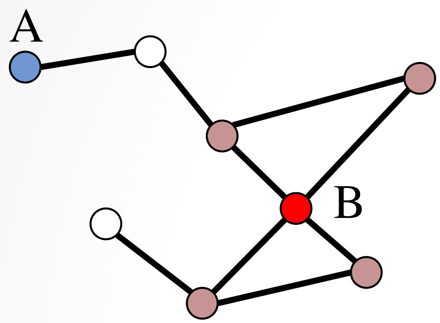
An undirected link is the superposition of two opposite directed links.

Example 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

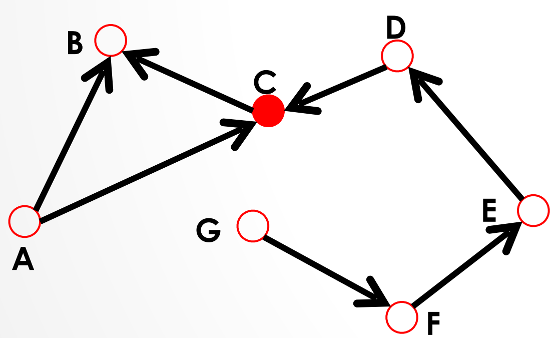
Undirected



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

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

Directed



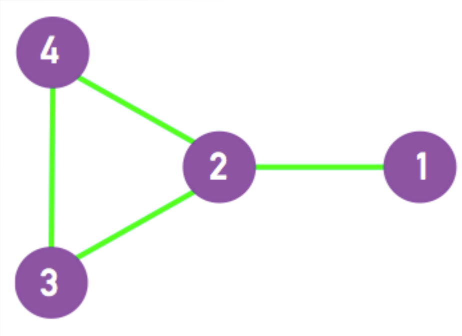
In *directed networks* we can define an in-degree and out-degree. The (total) degree is the sum of in- and out-degree.

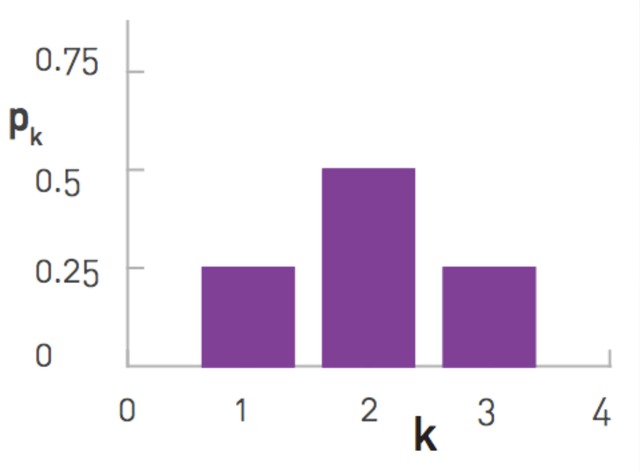
$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

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

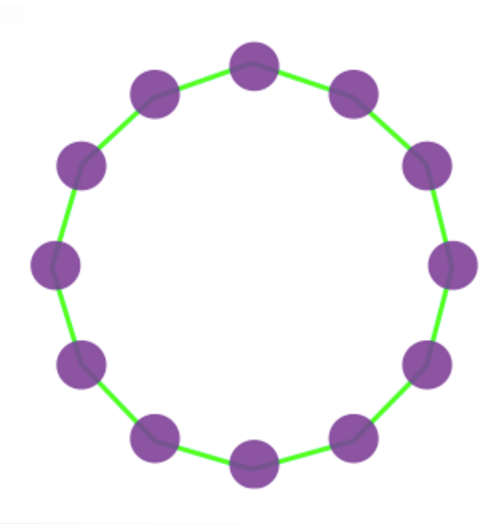
Degree is a "local" property – belongs to a single vertex.

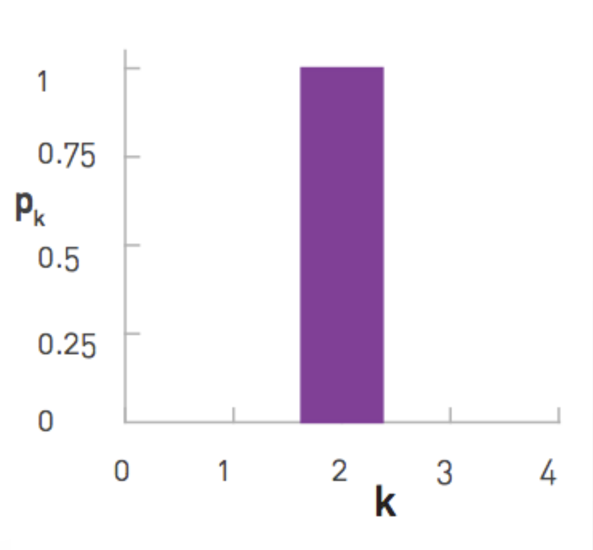
Degree Distribution



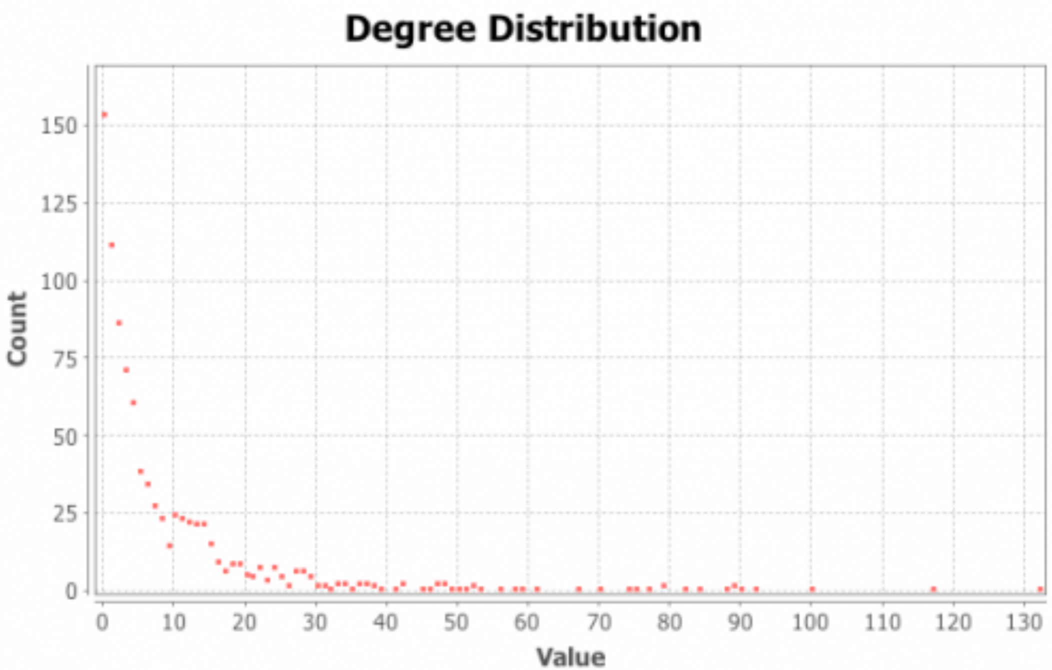


$p_k = \frac{N_k}{N}$ denotes the probability of a vertex having degree k



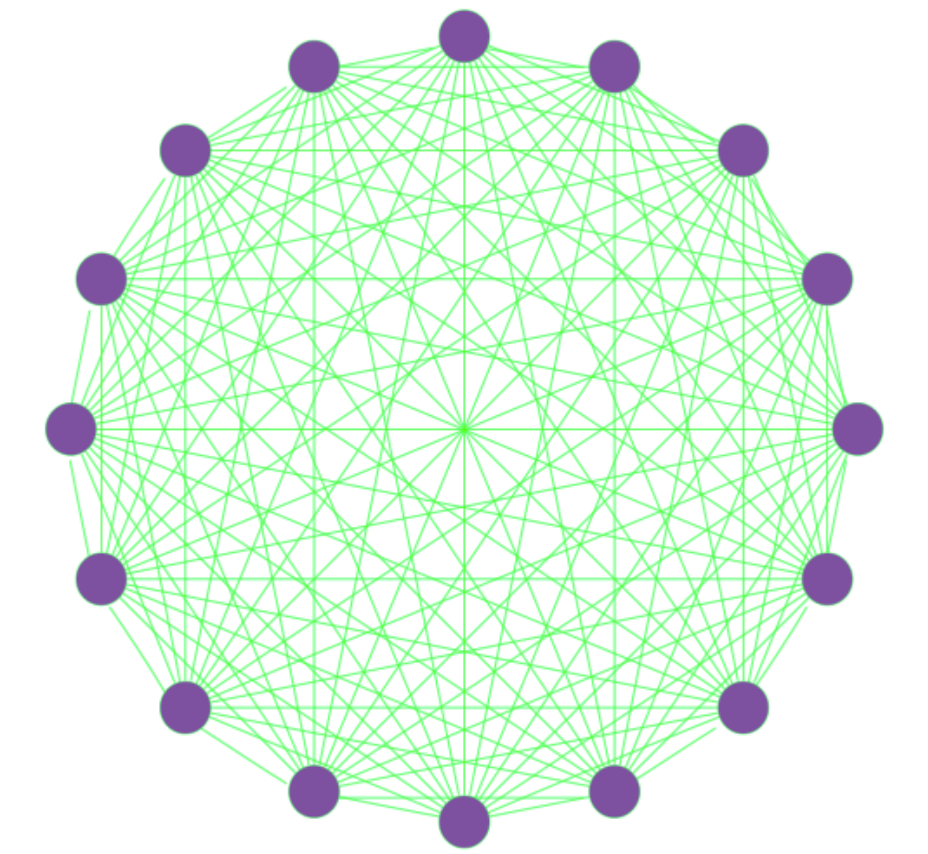


Degree Distribution in Real Networks



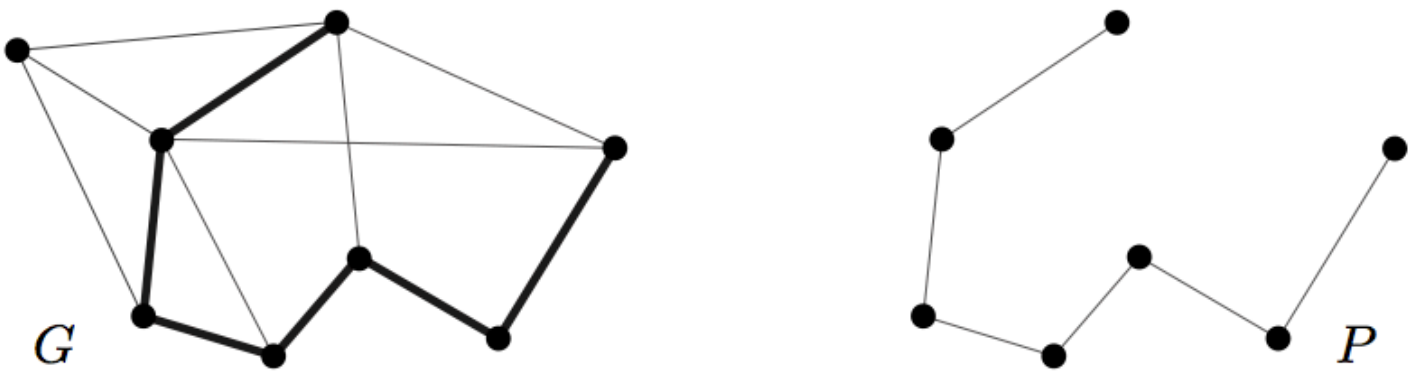
..looks often like this…

Complete Graph



has all pairs of vertices connected with each other:
$|E| = N(N-1)/2$

Paths

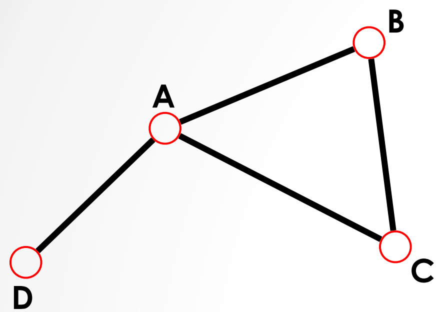


A *path* is a non-empty graph $P = (V, E)$ of the form

$$V = \{ x_0, x_1, \ldots, x_k \} \qquad E = \{ x_0x_1, x_1x_2, \ldots, x_{k-1}x_k \},$$

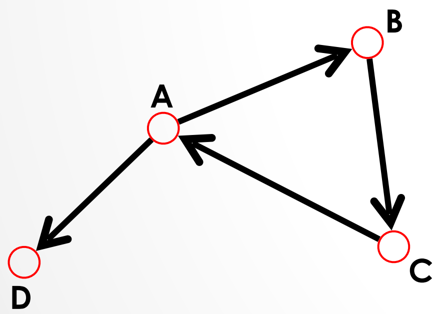
where the x_i are all distinct.

Distance



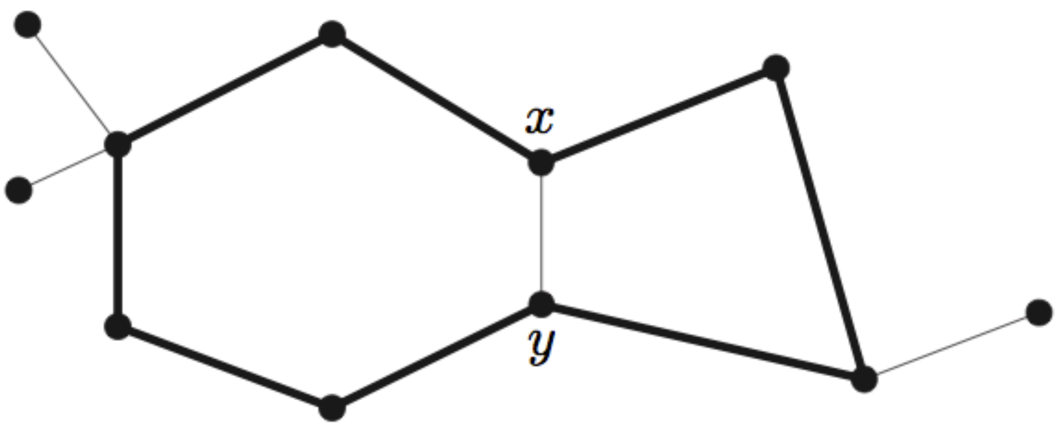
The *distance (shortest path, geodesic path)* between two nodes is defined as the number of edges along the shortest path connecting them.

*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).

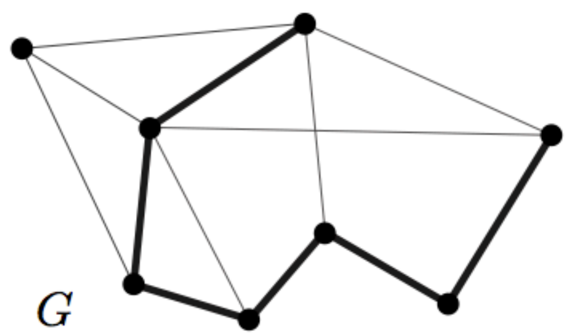
Cycles

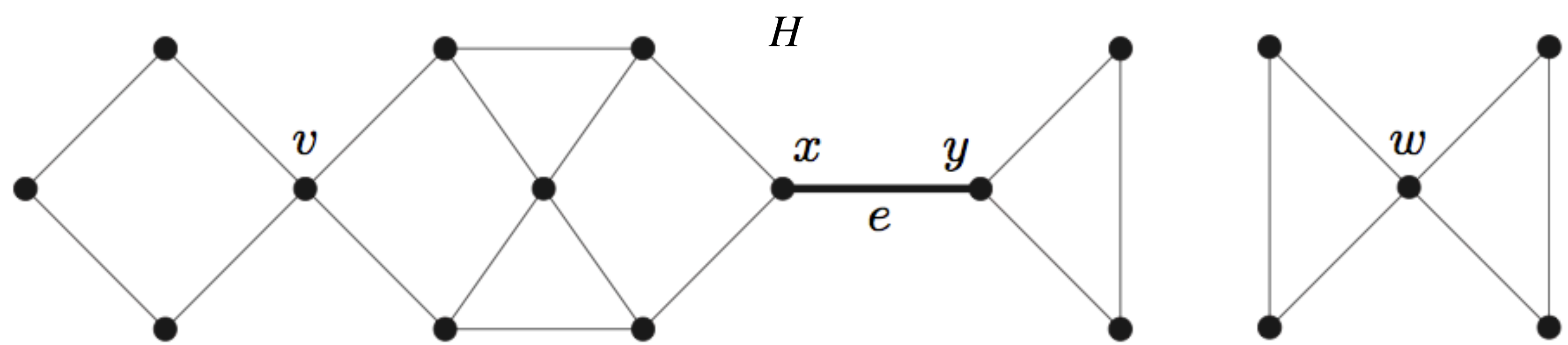


If $P = x_0 \ldots x_{k-1}$ is a path and $k \geqslant 3$,

then the graph $C := P + x_{k-1}x_0$ is called a *cycle*.

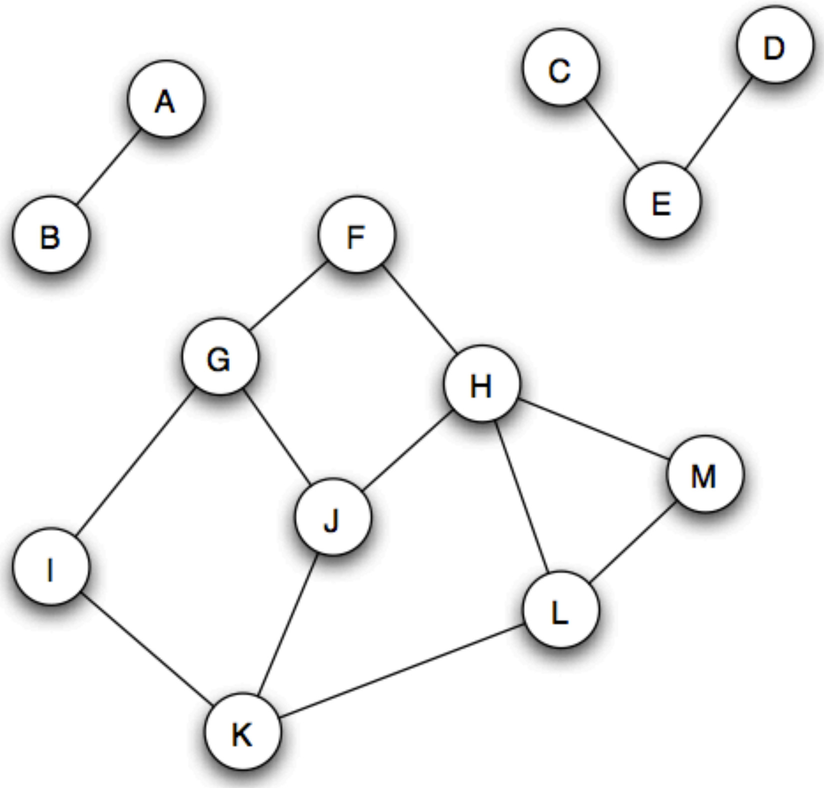
Connectivity





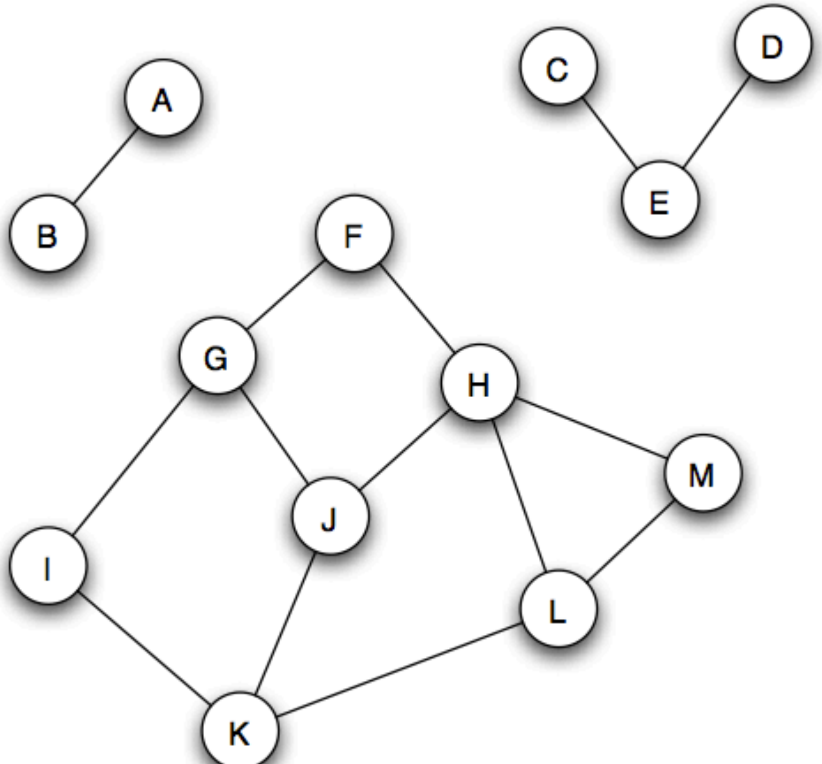
A non-empty graph G is called *connected* if any two of its vertices are linked by a path in G. If $U \subseteq V(G)$ and $G[U]$ is connected, we also call U itself connected (in G).

Connected Components



Can you please try to define a connected component?

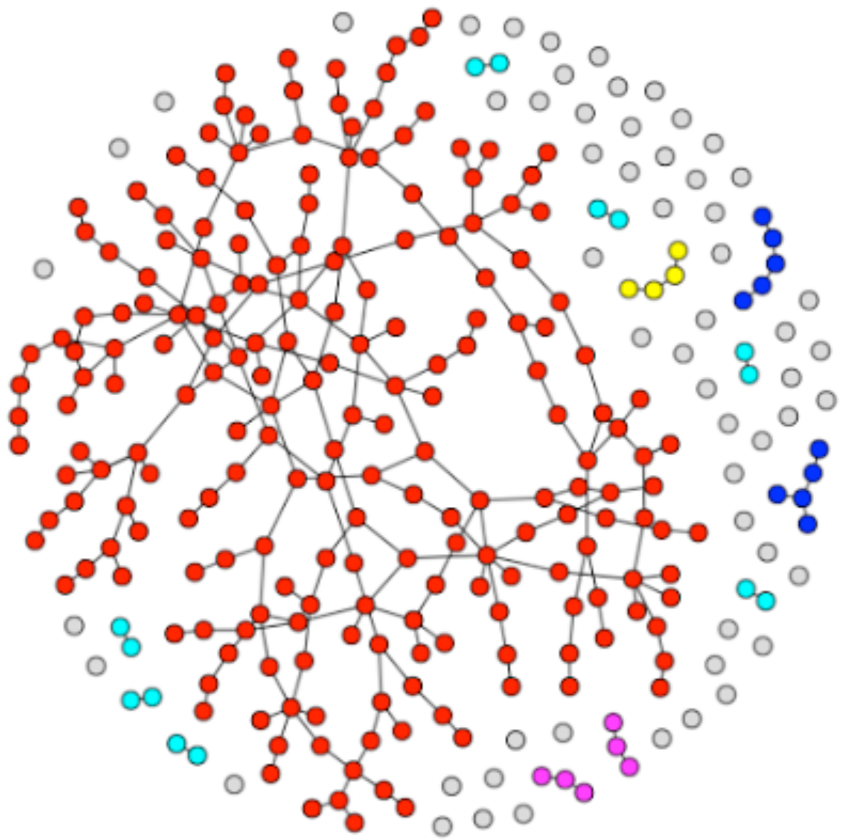
Connected Components



A (maximal) subgraph where there is an undirected path between every pair of vertices. "Maximal": We cannot leave out anyone who is connected. So, F-M is a connected component.

Components is a "global" property – a graph has components.

Giant Component

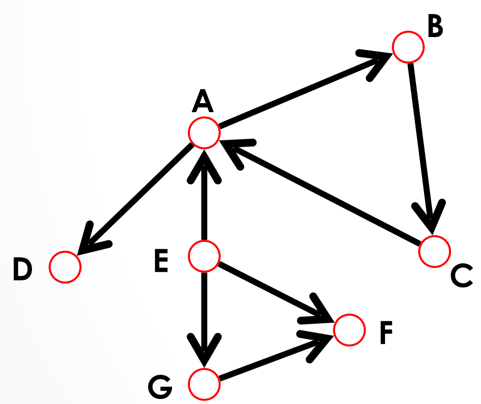


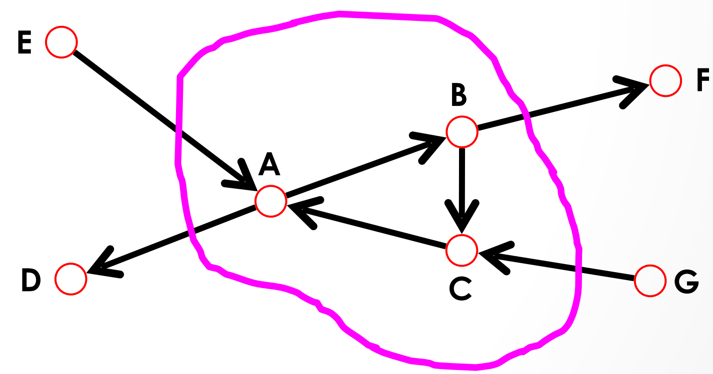
A connected component that contains a **significant** fraction of all the nodes. Many real-life networks possess a giant component.

Strongly-Connected Component

Strongly connected directed graph: has a path from each node to every other node and vice versa (e.g. AB path and BA path).

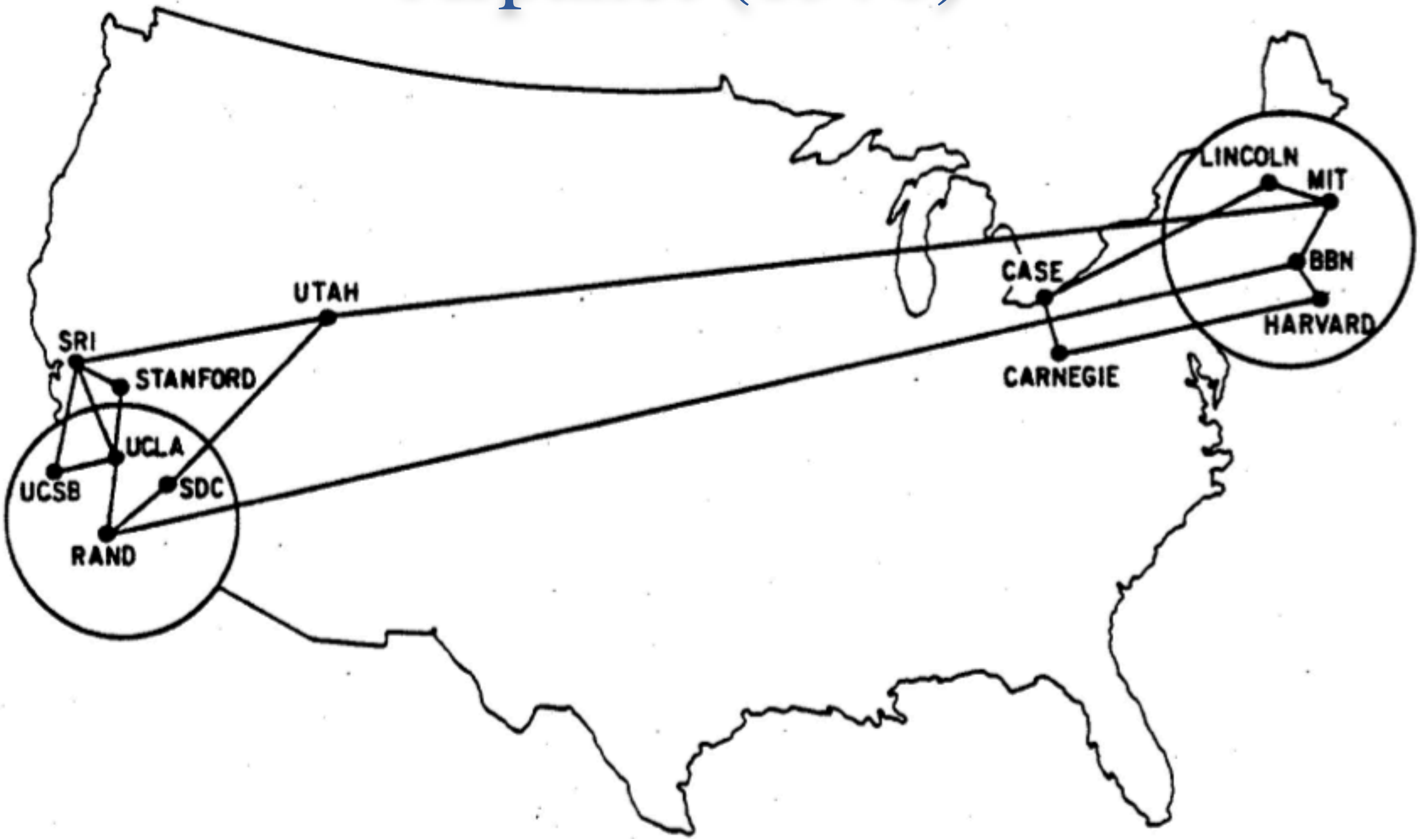
Weakly connected directed graph: it is connected if we disregard the edge directions.





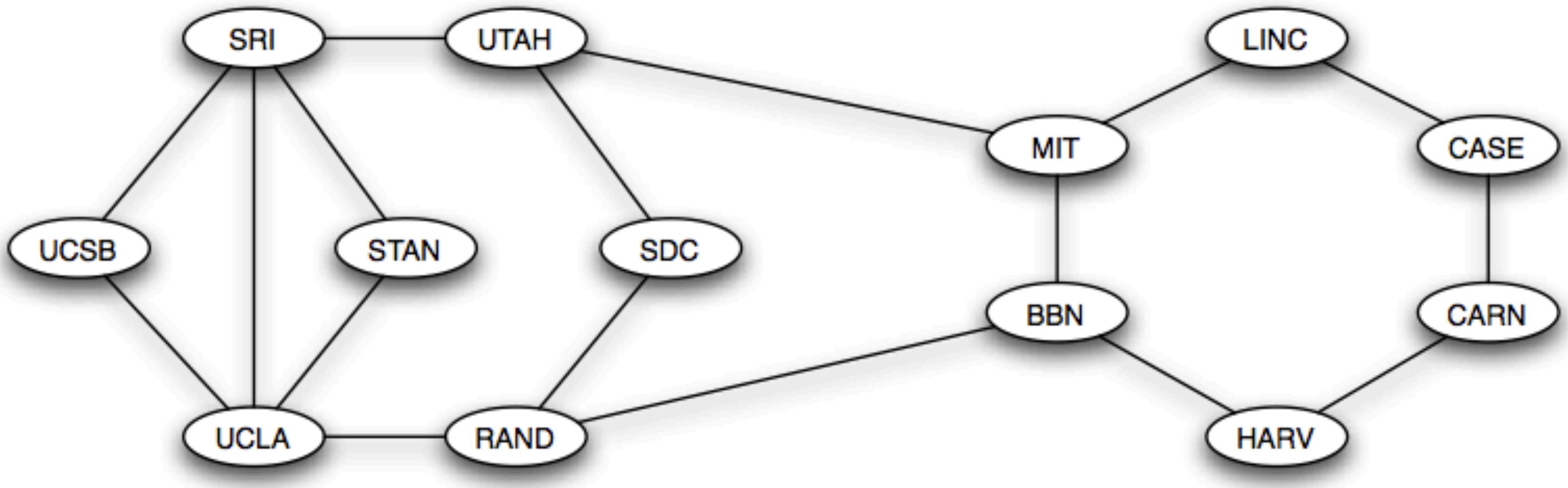
Strongly connected components can be identified, but not every node is part of a nontrivial strongly connected component.

Arpanet (1970)

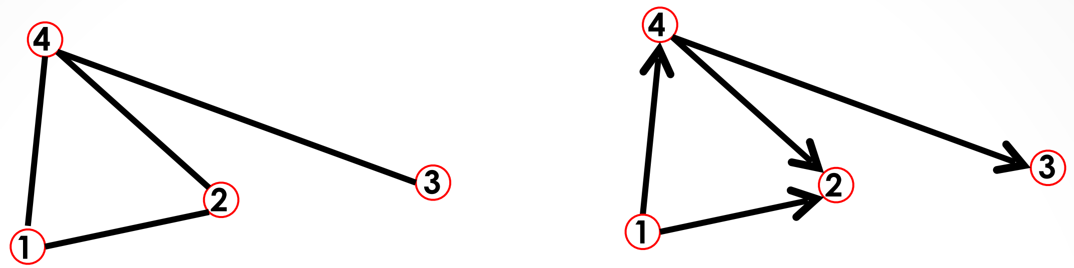


A network depicting the sites on the Internet, then known as the Arpanet, in December 1970. (Image from F. Heart, A. McKenzie, J. McQuillian, and D. Walden [7]; on-line at http://som.csudh.edu/cis/lpress/history/arpamaps/.)

Arpanet as a graph



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.

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

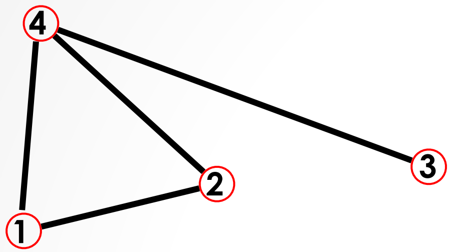
Note that for a directed graph (right) the matrix is not symmetric.

$A_{ij} = 1$ 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 & Node Degrees

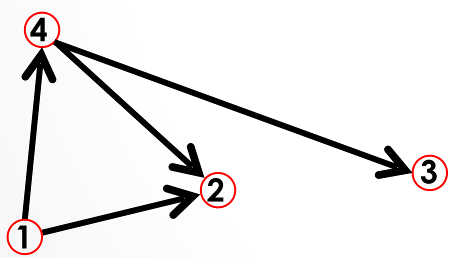
Undirected



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

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

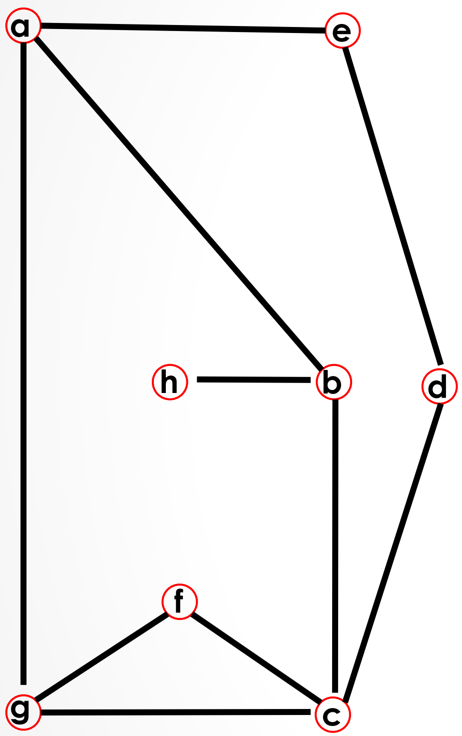
Directed



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

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

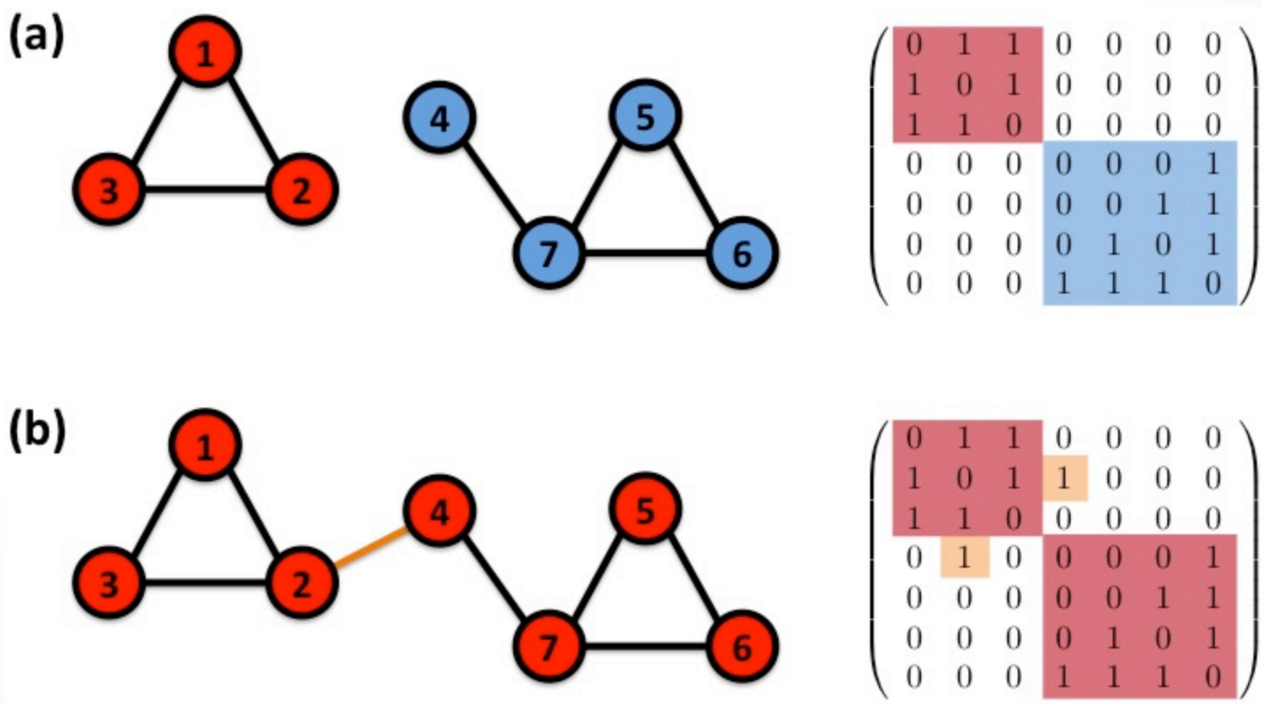
Adjacency Matrix



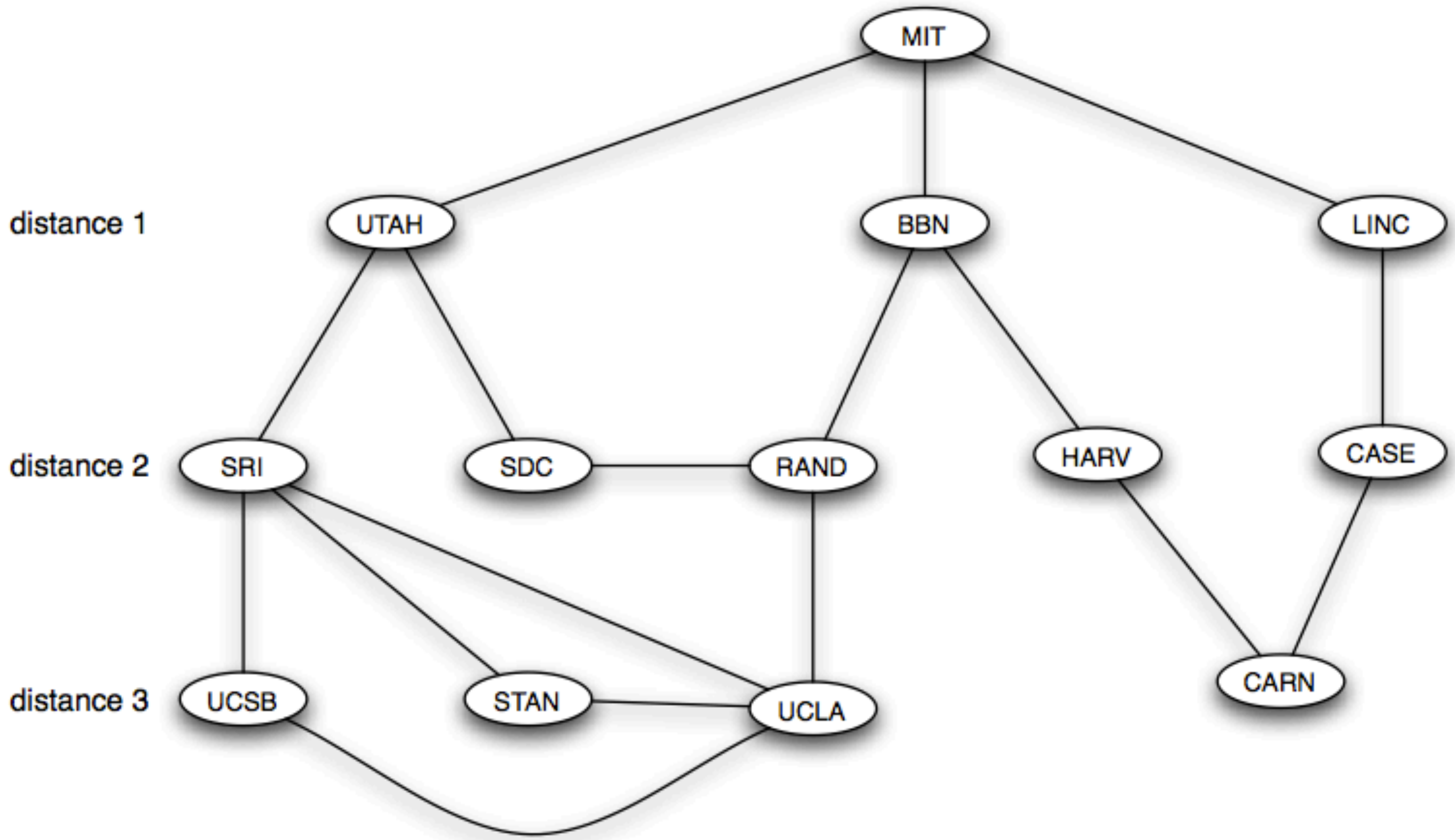
	a	b	c	d	e	f	g	h
a	0	1	0	0	1	0	1	0
b	1	0	1	0	0	0	0	1
c	0	1	0	1	0	1	1	0
d	0	0	1	0	1	0	0	0
e	1	0	0	1	0	0	0	0
f	0	0	1	0	0	0	1	0
g	1	0	1	0	0	0	0	0
h	0	1	0	0	0	0	0	0

Adjacency Matrix – Connected Components

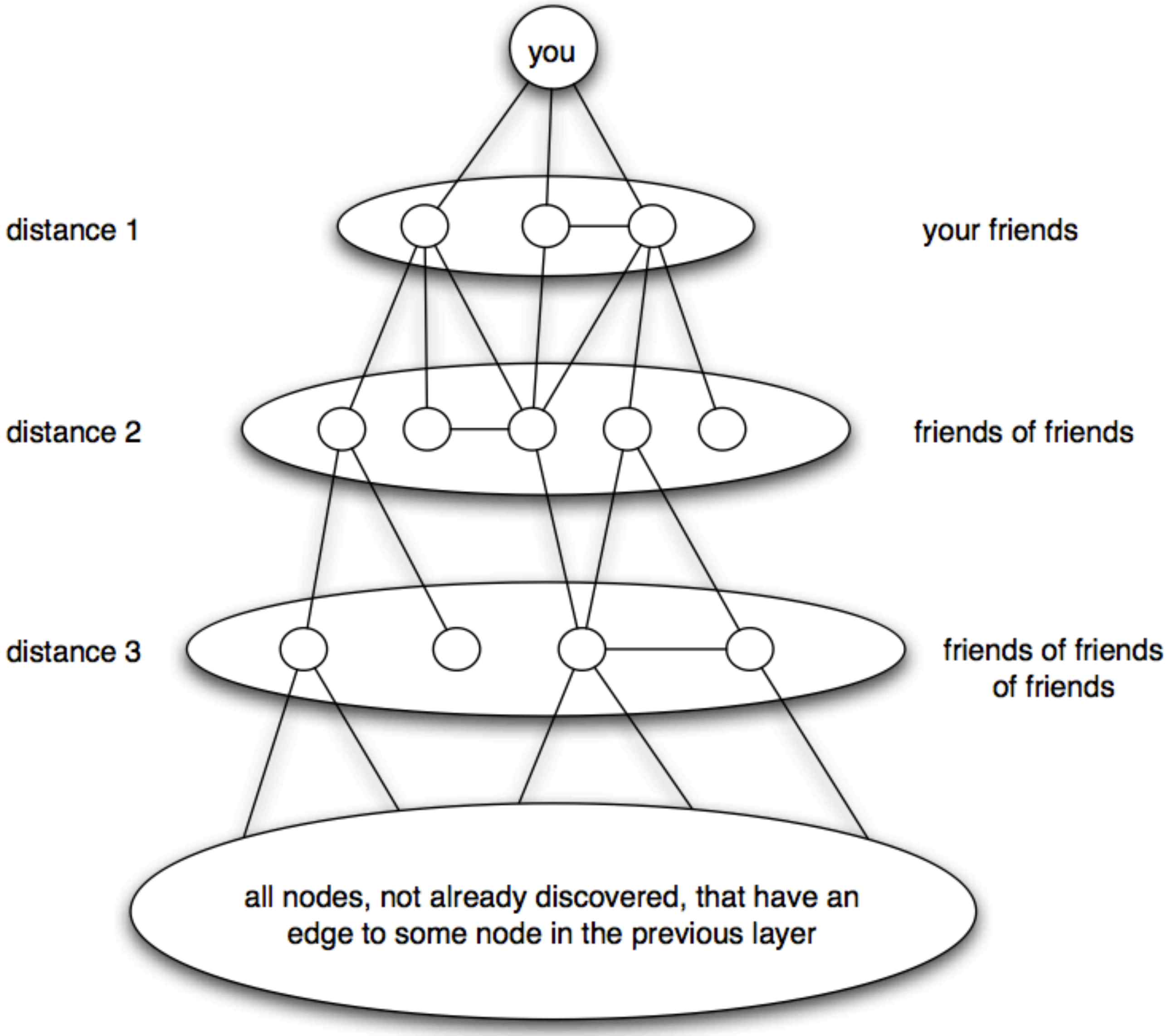
The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero:



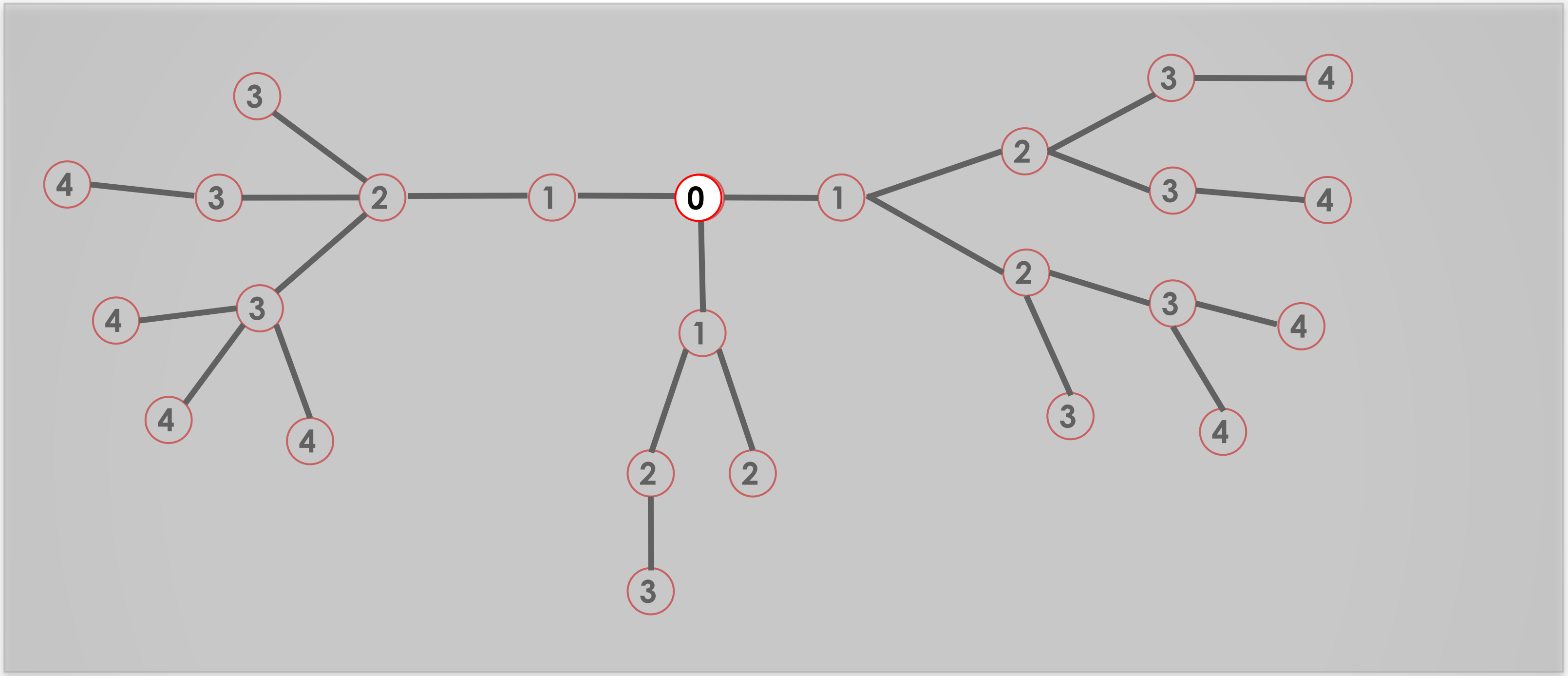
Breadth-first Search



Breadth-first Search



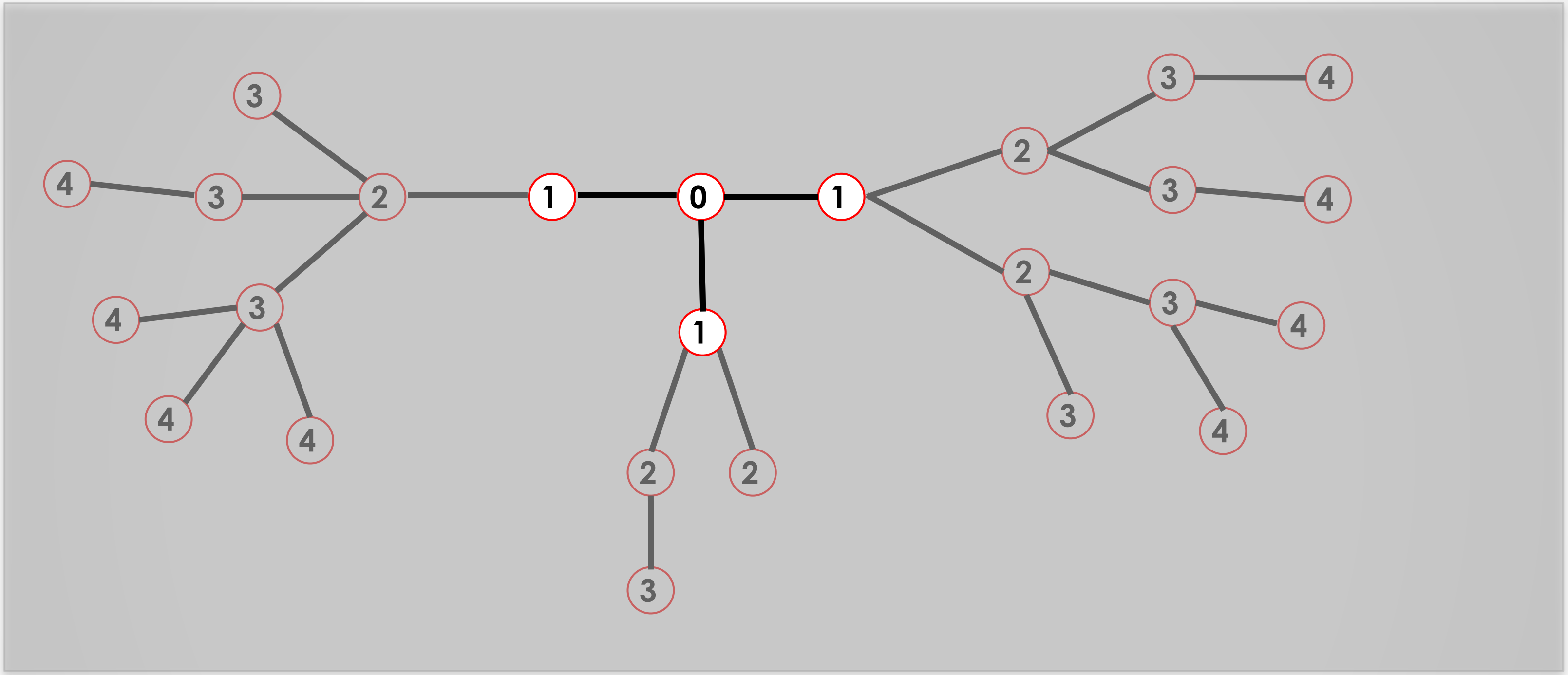
Breadth-first Search (BFS)



Distance between node 0 and node 4:

1. Start at 0.

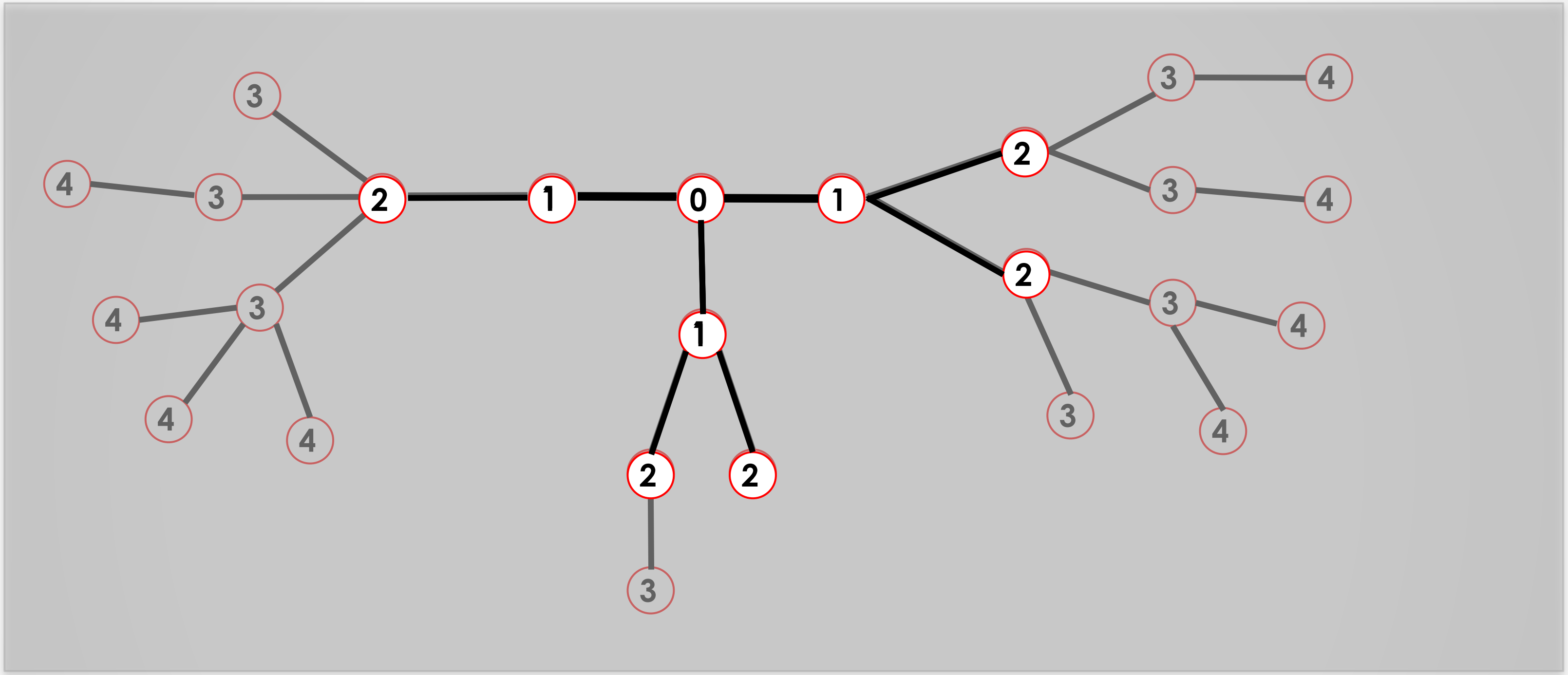
Breadth-first Search (BFS)



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.

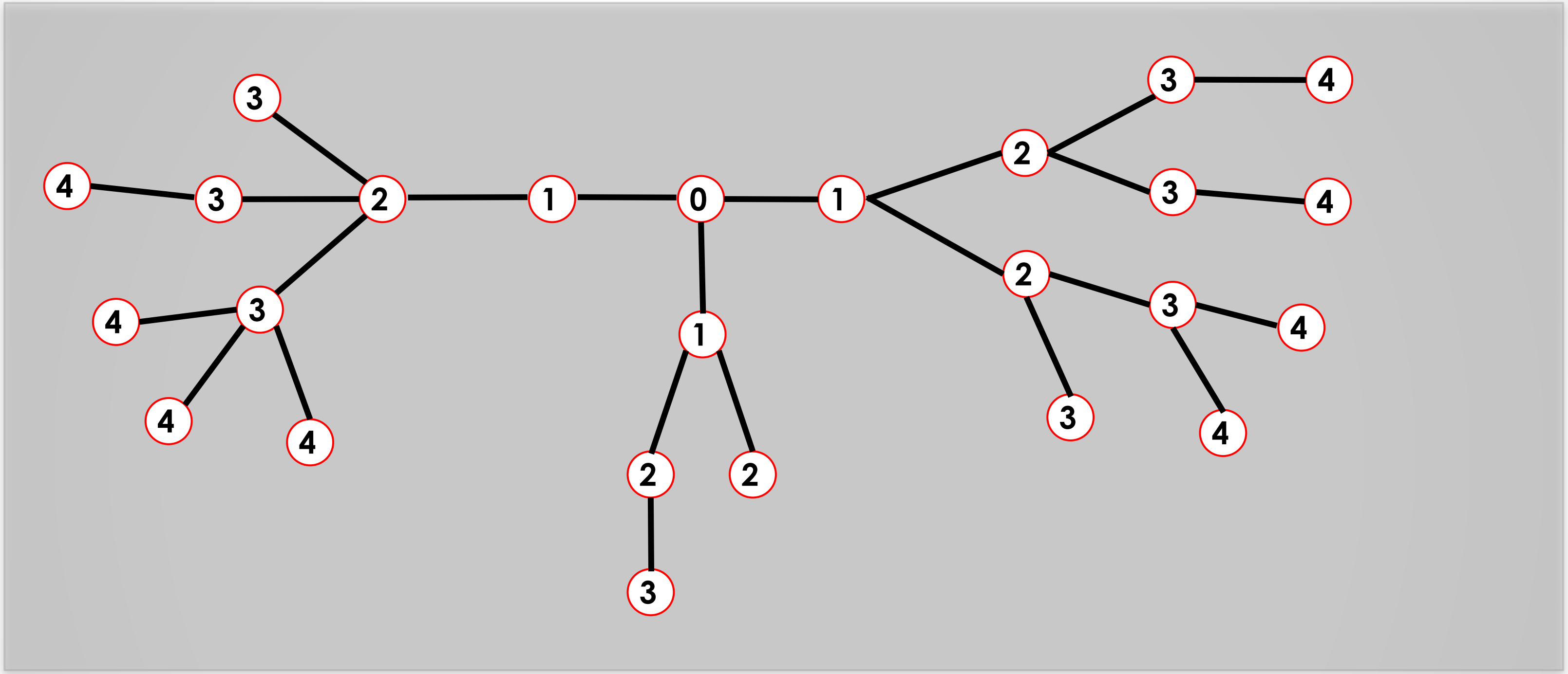
Breadth-first Search (BFS)



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.
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.

Breadth-first Search (BFS)



Distance between node 0 and node 4:

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

Connected Components Using BFS

1. Start from a randomly chosen node i and perform a BFS. Label all nodes reached this way with $n = 1$.

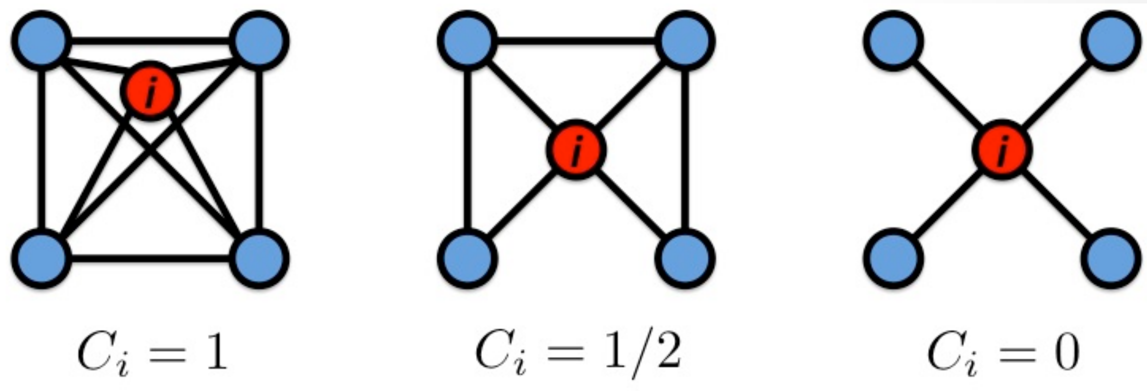
2. If the total number of labeled nodes equals N, then the network is connected. If the number of labeled nodes is smaller than N, the network consists of several components. To identify them, proceed to step 3.

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 all with n. Return to step 2.

Clustering Coefficient

* **What fraction of your neighbours are connected?**
  * Node i with degree k_i
  * Clustering Coefficient C_i for a vertex i is in [0,1]

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



Clustering coefficient is a “local” property – each vertex has one.

Watts & Strogatz, Nature 1998.

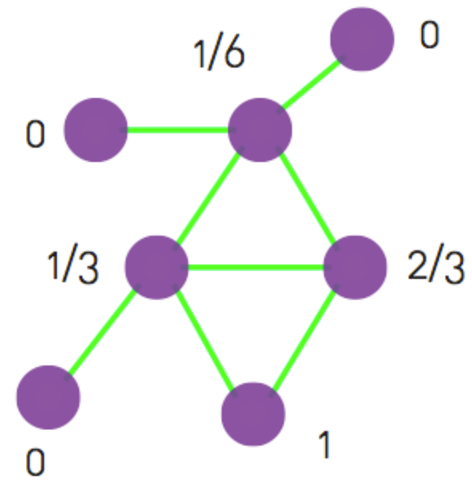
Clustering Coefficient

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

Clustering Coefficient of vertex i

Average Clustering Coefficient of the graph

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



<C> = (0+0+0+1+1/6+1/3+2/3)/7

<C> = 13/42 ≈ 0.310

Network Diameter & Average Distance

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

Average path length (Average distance), <d>, for a directed graph:

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

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

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

$$\langle d \rangle \equiv \frac{2}{N(N-1)} \sum_{i,j>i} d_{ij}$$

Central Quantities in Network Science

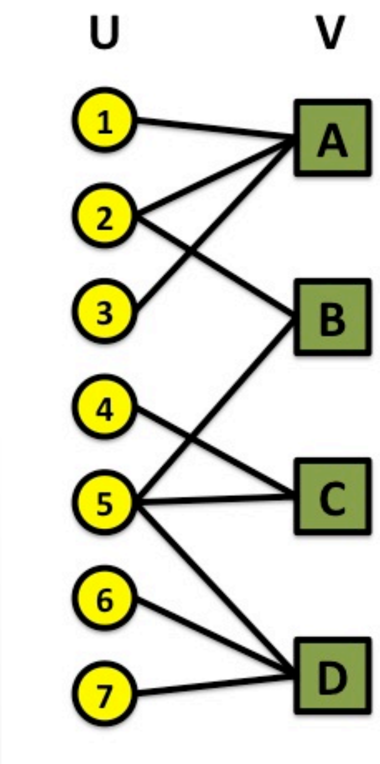
Degree distribution: P(k)

Path length: <d>

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

Bipartite Graph

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:

U – People, V – Hobbies
U – Recipies, V – Ingredients
U – Documents, V – Keywords

Network Data Sources

1. http://www-personal.umich.edu/~mejn/netdata/
2. https://snap.stanford.edu/data/
3. https://networkdata.ics.uci.edu/index.php

References

1. http://barabasi.com/networksciencebook/
2. https://www.cs.cornell.edu/home/kleinber/networks-book/