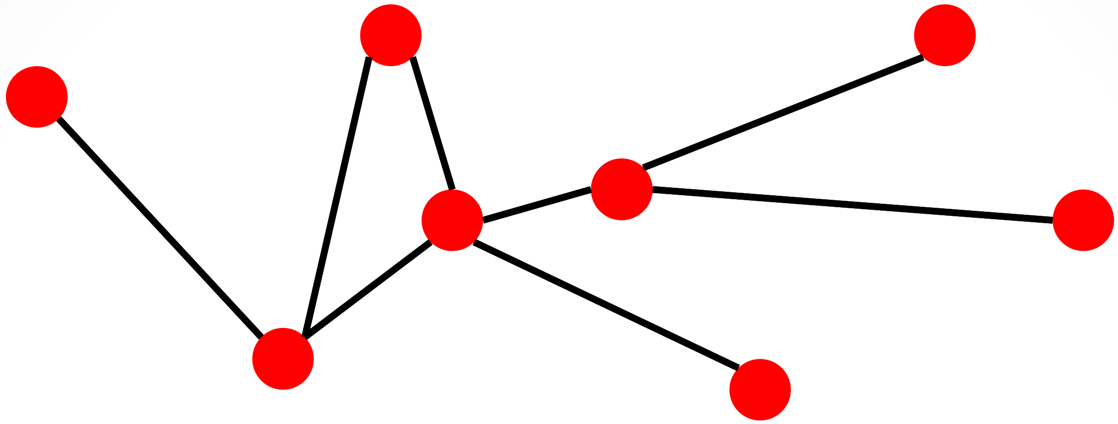
Social Network Analysis

Graphs

Graph



- **Components**: nodes, vertices, actors N, V
- **Interactions**: links, edges, relations L, E
- **System**: network, graph G(V,E)

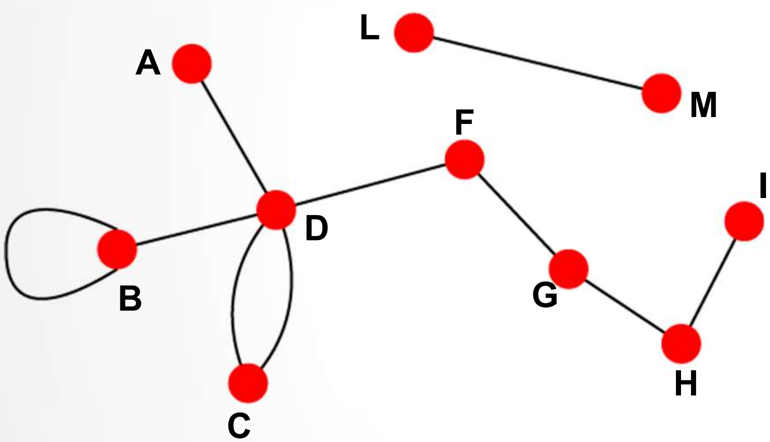
Graphs serve as mathematical models of network structures.

Directed & Undirected Graphs

Undirected

Links: undirected (*symmetrical*)

Graph:

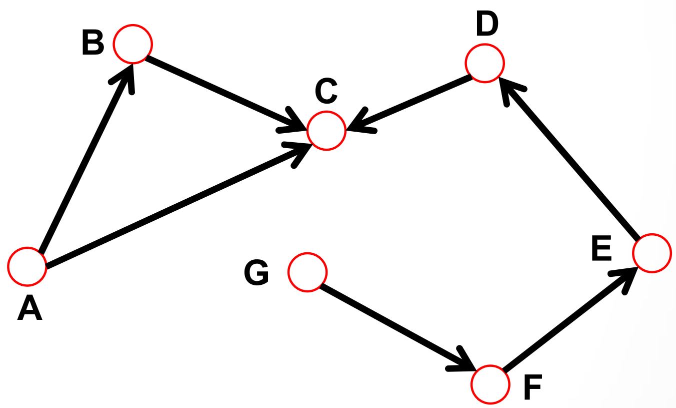


Undirected links :
Co-authorship links
Actor network
Protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:



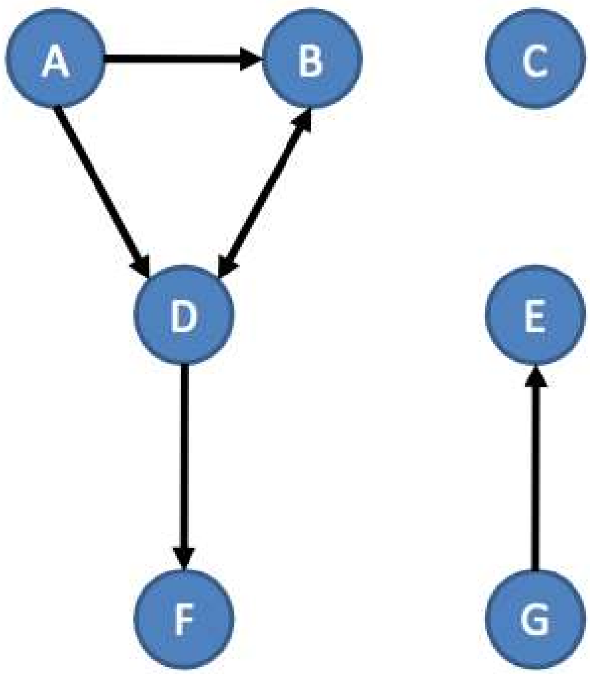
An undirected link is the superposition of two opposite directed links.

Directed links :
Who-calls-whom
Citation network

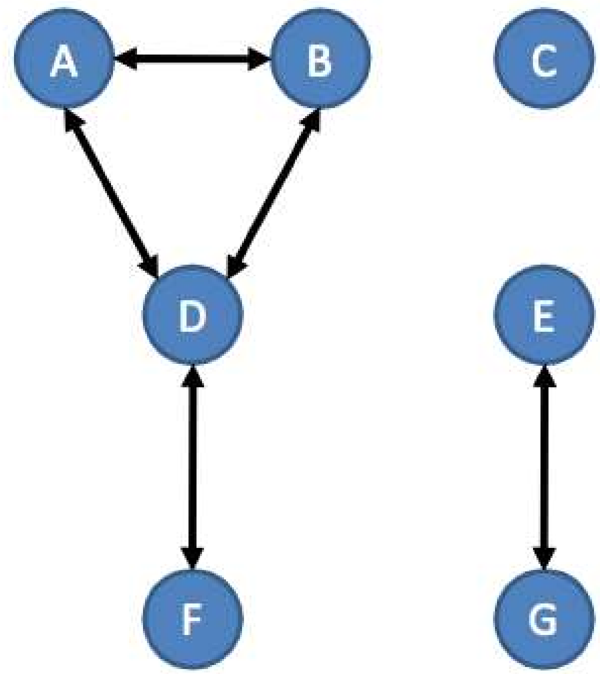
Directed & undirected

- Communication vs. friendship networks

twitter

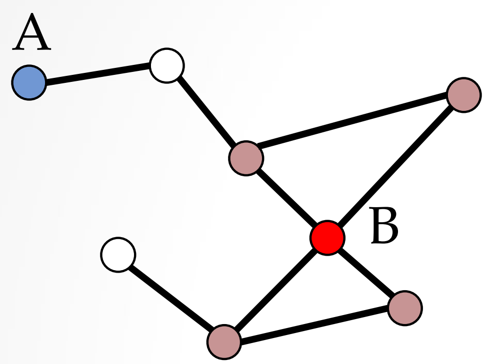


facebook



Degree

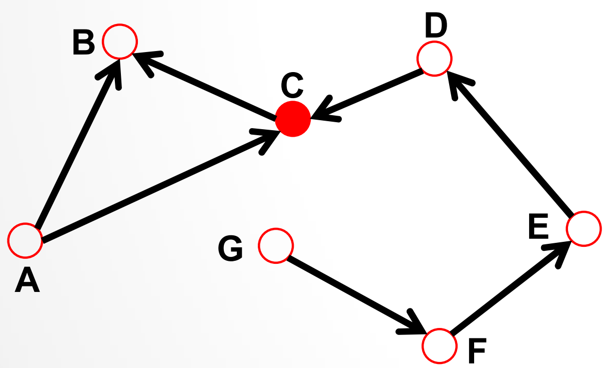
Undirected



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.

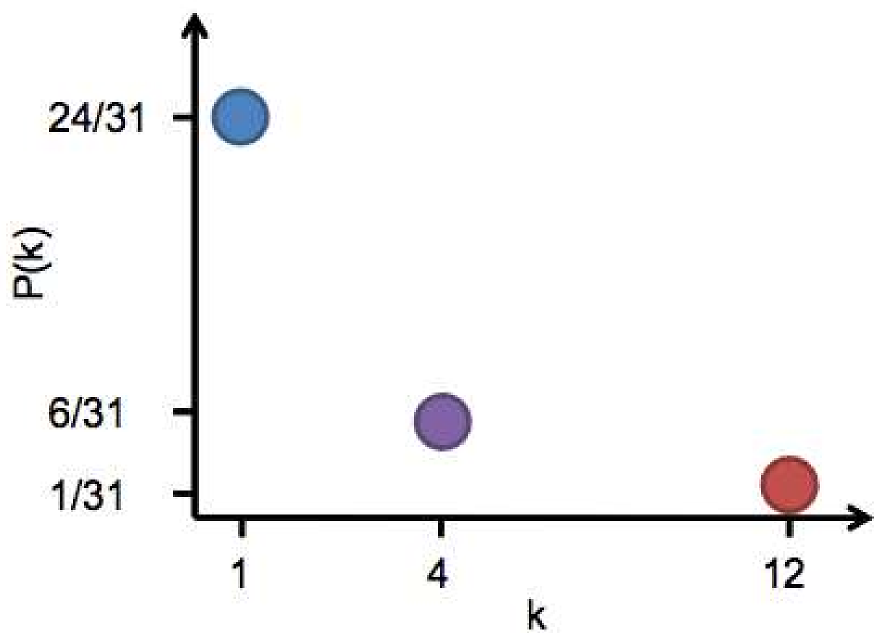
The (total) degree is the sum of in- and out-degree.

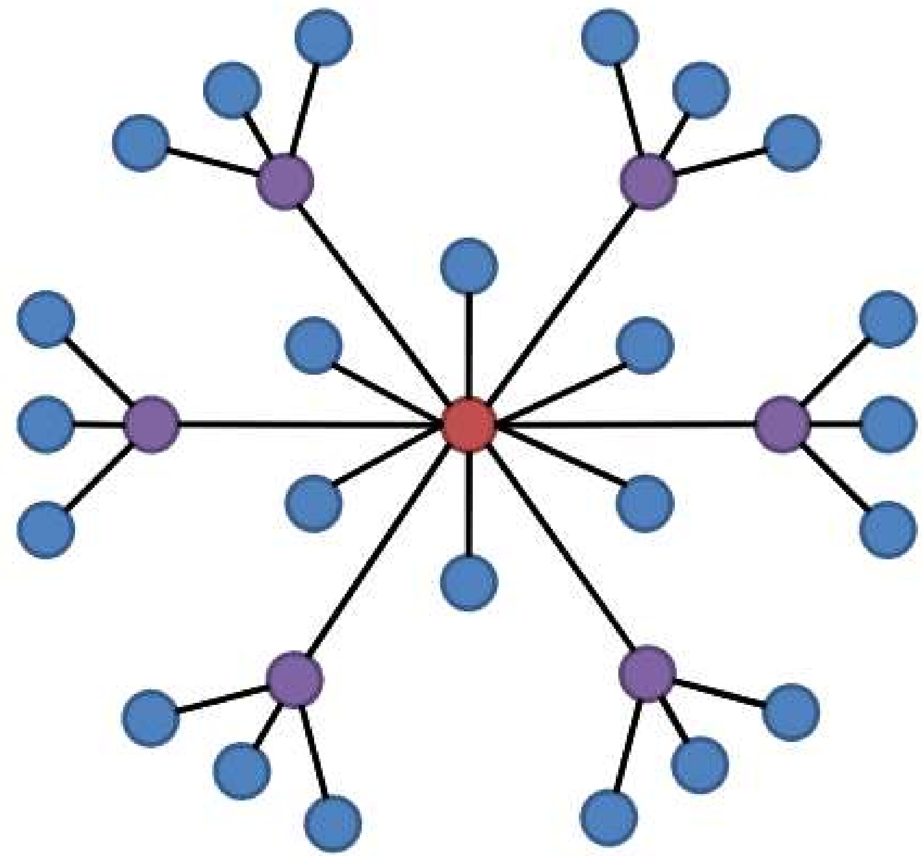
$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

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

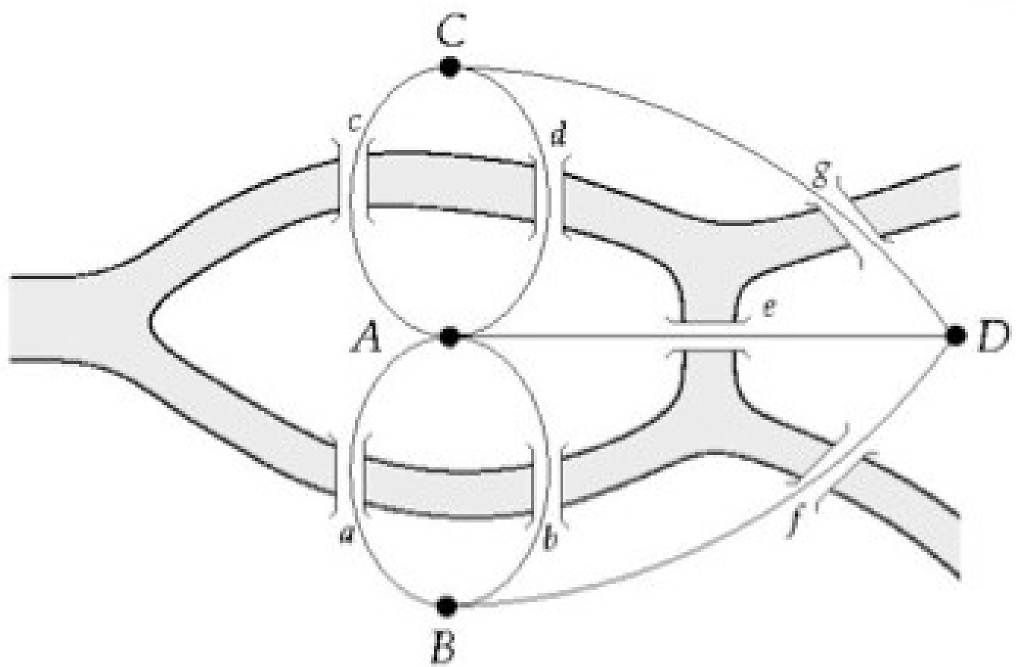
Degree distributions

- What's the probability *P*(*k*) of randomly selecting a node with degree *k* in this network?



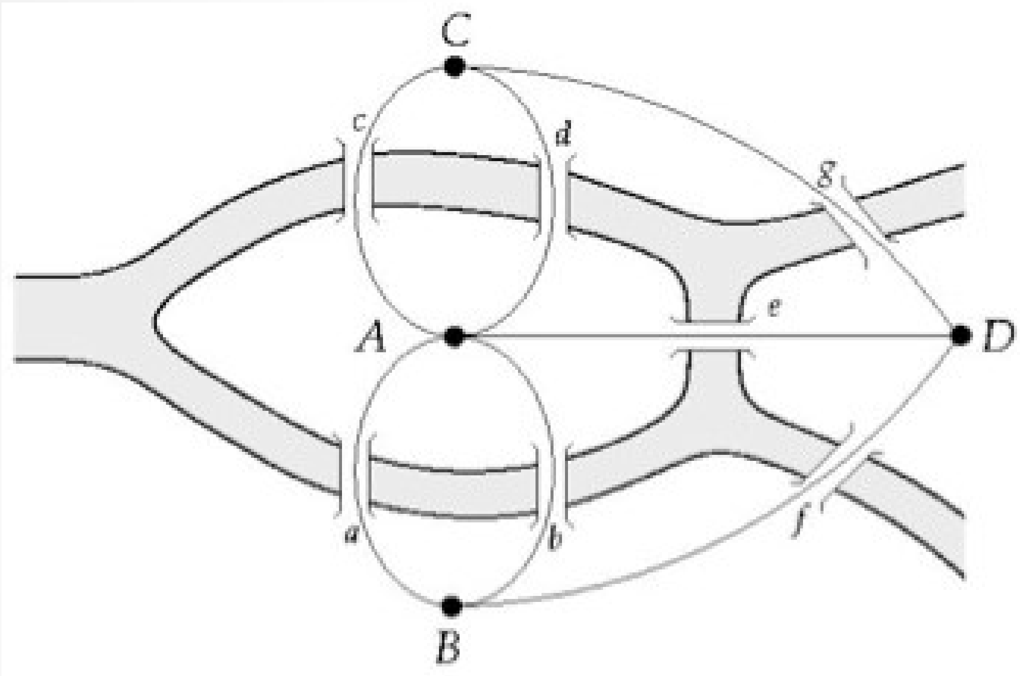


The Bridges of Königsberg



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

The Bridges of Königsberg



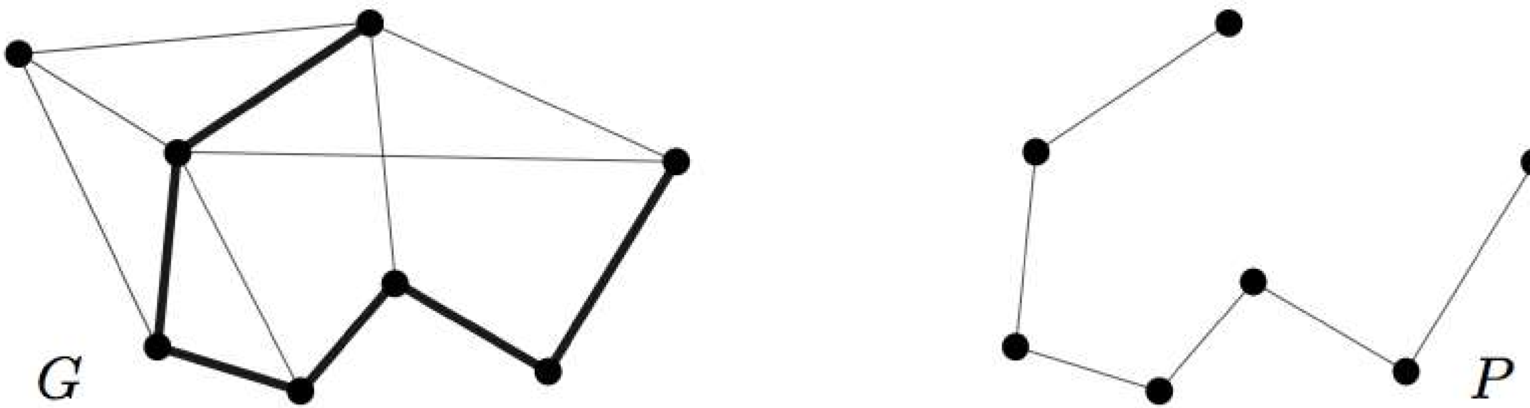
Can one walk across the seven bridges and never cross the same bridge twice?

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.

Led to foundation of Graph Theory

Paths

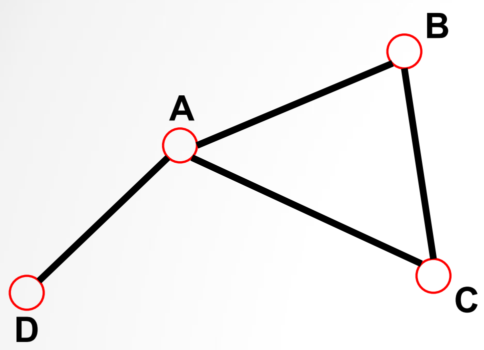


A *path* is a non-empty graph $P = (V, E)$ of the form

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

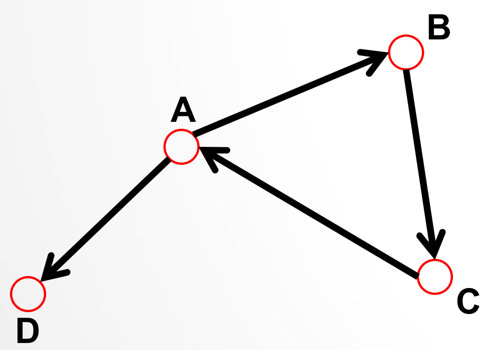
where the x_i are all distinct.

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.

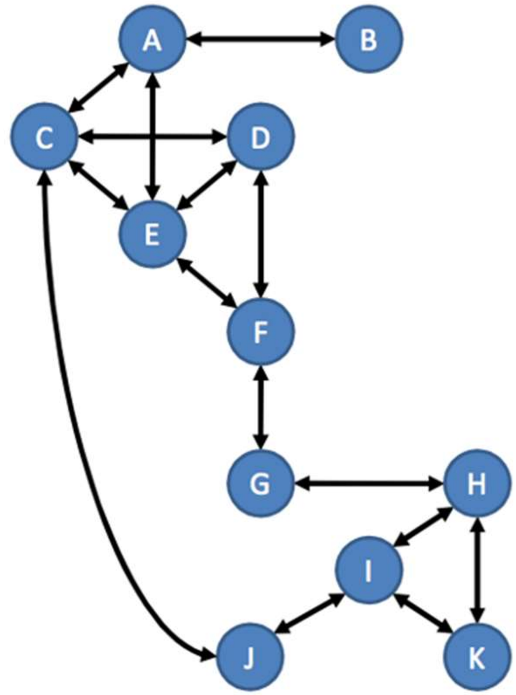


In directed graphs each path needs to follow the direction of the arrows.

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).

Path length

- **Path length**: number of links between two nodes (degrees of separation)
  - BACDE = 4
- **Geodesic**: Shortest path length between two nodes
  - BAE = 2
- **Eccentricity**: Each actor's longest geodesic
- **Diameter**: Network's largest geodesic\eccenctricity
  - BAEFGH\BACJIH



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

$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$ where d_{ij} is the distance from node *i* to node j

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

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

$L_{max} = N(N-1)/2$

Density

- Observed edges in network/ Max possible edges

In particular, for undirected simple graphs, the graph density is defined as

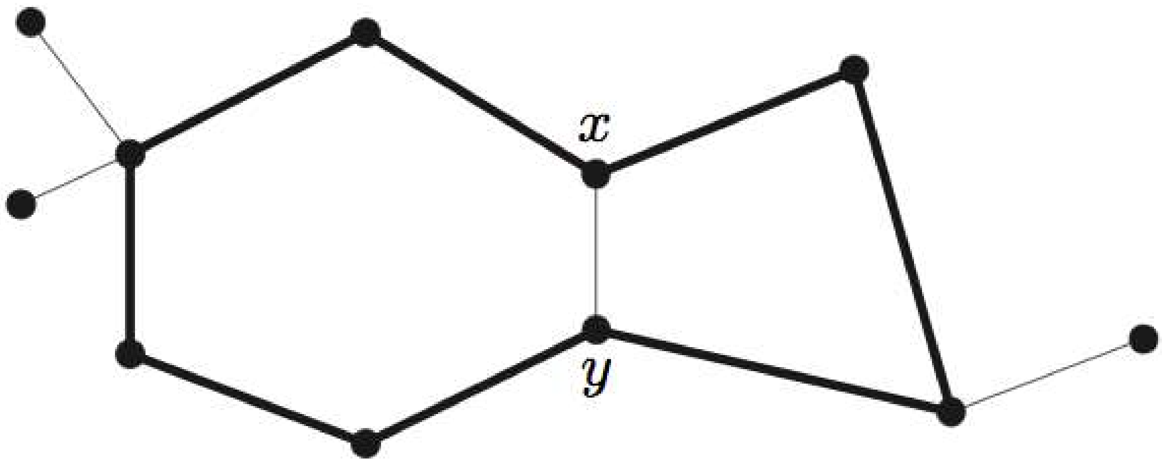
$$D = \frac{2|E|}{|V|(|V|-1)}.$$

While for directed simple graphs, the graph density is defined as

$$D = \frac{|E|}{|V|(|V|-1)},$$

where $|E|$ is the number of edges and $|V|$ is the number of vertices in the graph.

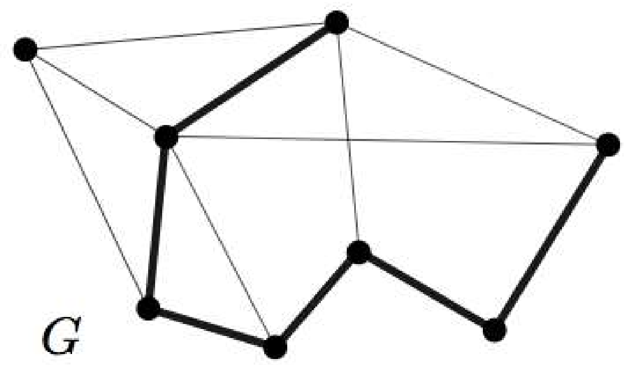
Cycles

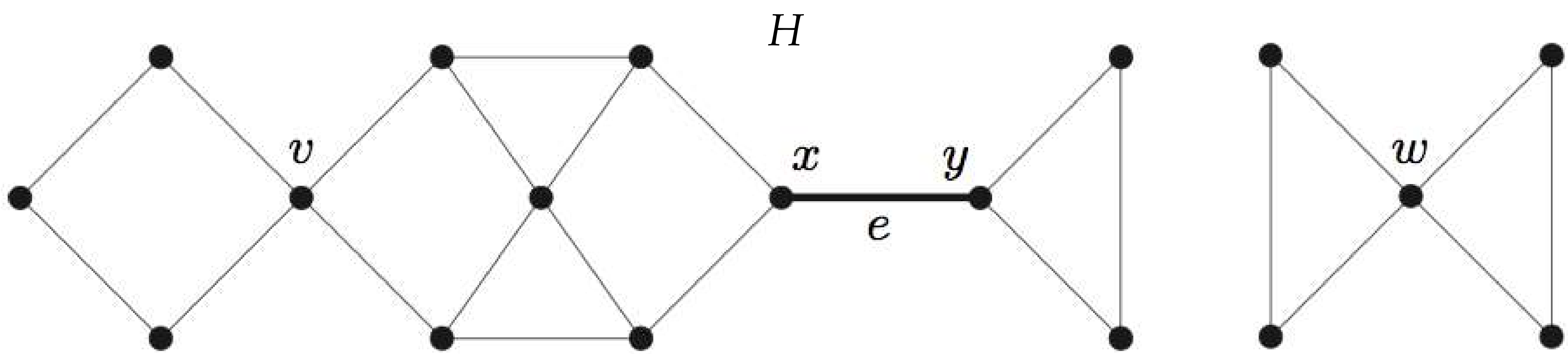


If $P = x_0 \dots x_{k-1}$ is a path and $k \geqslant 3$,

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

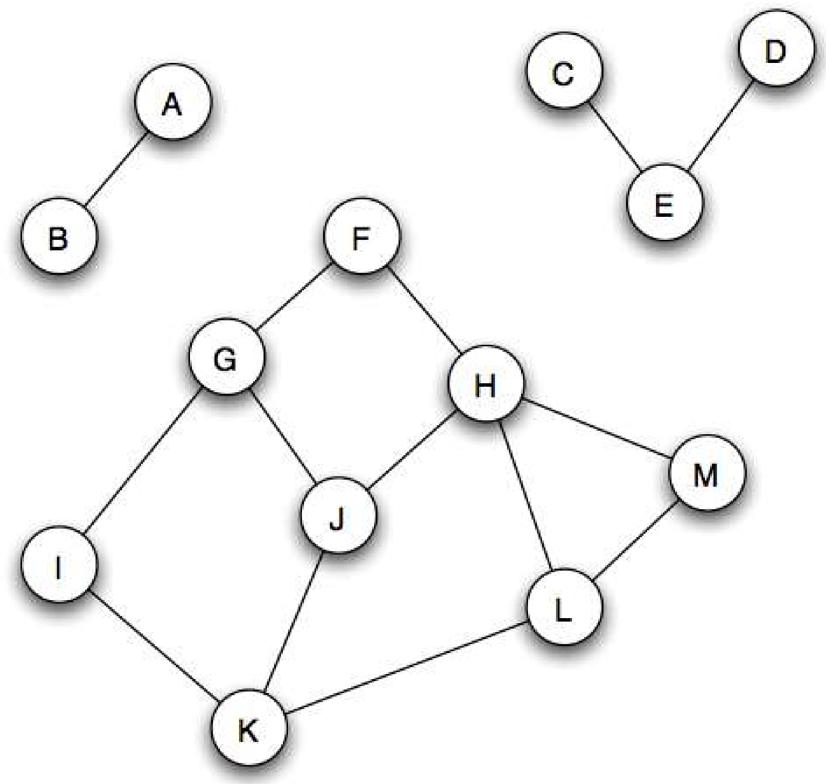
Connectivity





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



A connected component of a graph (often shortened just to the term "component") is a subset of the nodes such that:

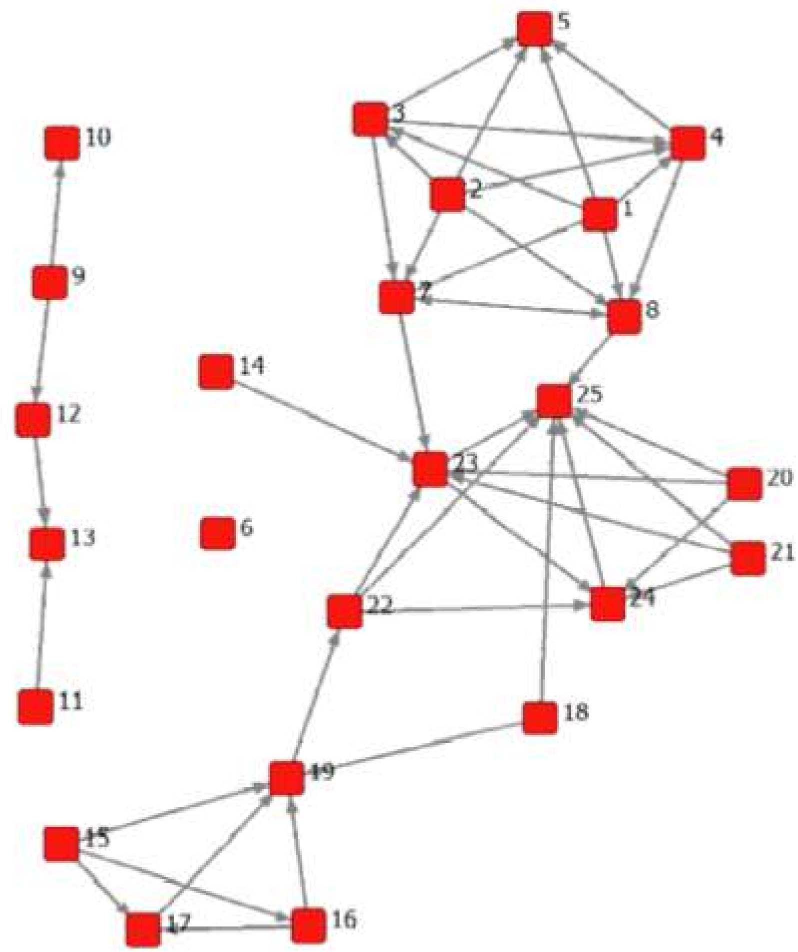
(i) every node in the subset has a path to every other; and

(ii) the subset is not part of some larger set with the property that every node can reach every other.

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

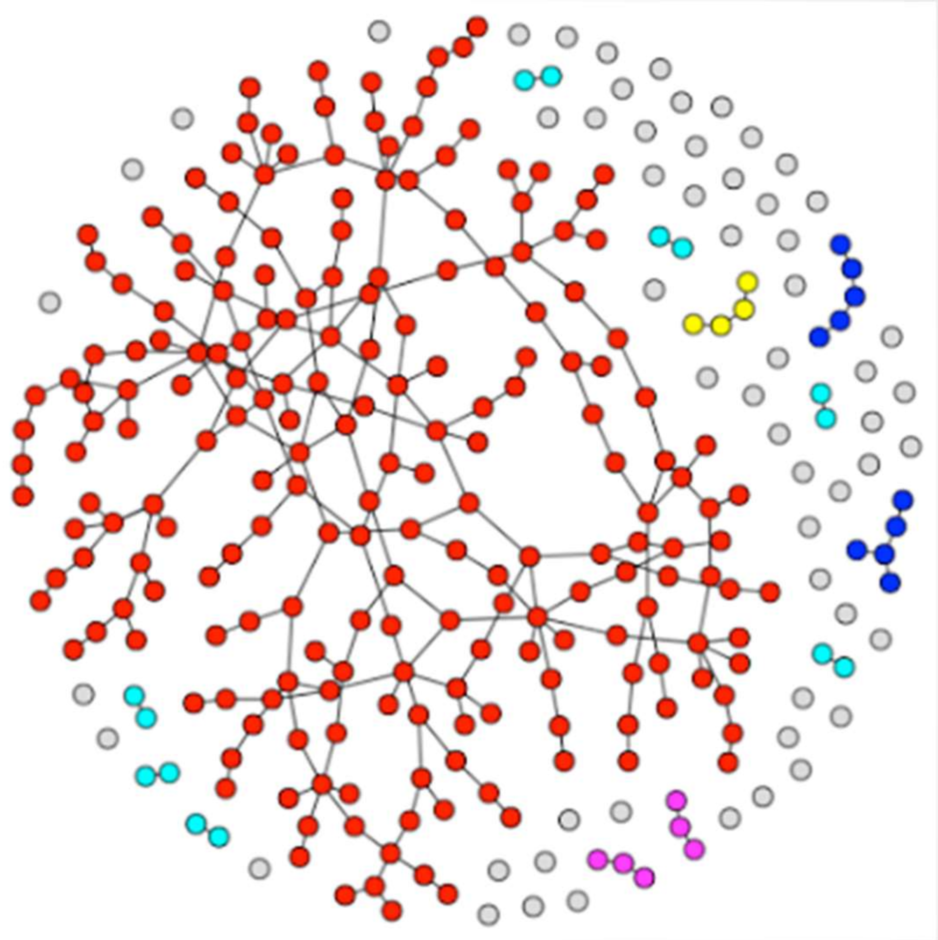
Strong & weak components

- Weak component
  - Set of connected nodes, regardless of direction of connections
- Strong component
  - A directed path must exist between two nodes for them to be in the same component



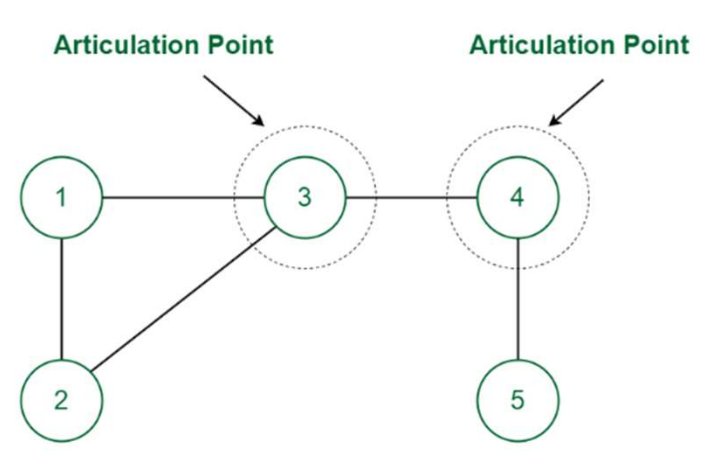
Giant Component

- Across a range of network datasets — large, complex networks often have what is called a giant component, a deliberately informal term for a connected component that contains a significant fraction of all the nodes.
- Moreover, when a network contains a giant component, it almost always contains only one.

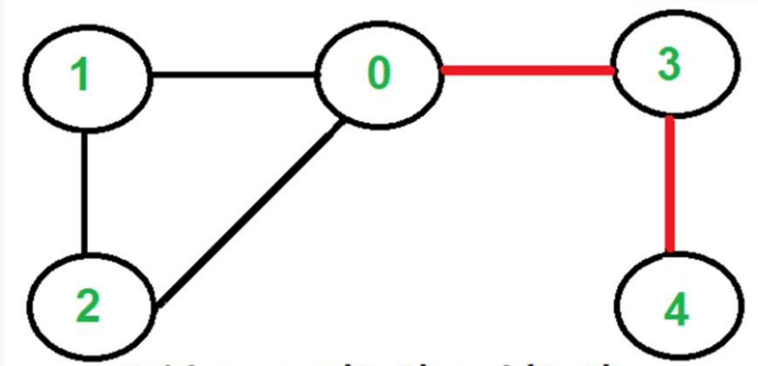


Cutpoints & Bridges

- Cutpoints:
  - A cutpoint, also known as an articulation point c cut vertex, in a graph is a vertex that is shared b two or more blocks and whose removal disconnects the graph.
  - A block is a maximal connected subgraph of a graph that has no cutpoints.

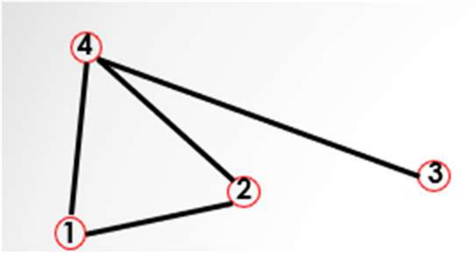


- Bridge:
  - An edge in an undirected connected graph is a bridge if removing it disconnects the graph.
  - For a disconnected undirected graph, the definition is similar, a bridge is an edge removal that increases the number of disconnected components.



Bridges are (0, 3) and (3, 4)

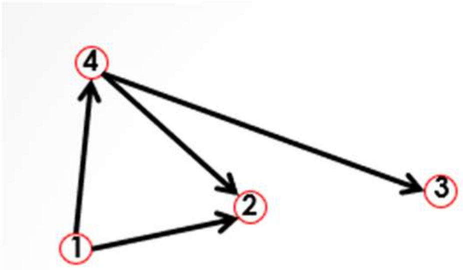
Adjacency Matrix



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

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.



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

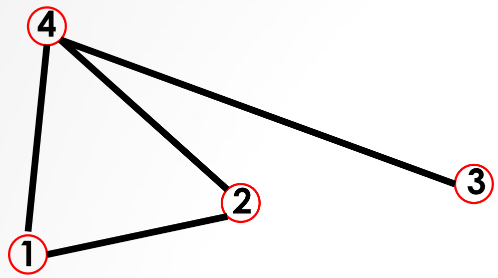
Note that for a directed graph (right) the matrix is not symmetric.

$A_{ij} = 1$ 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 & Node Degrees

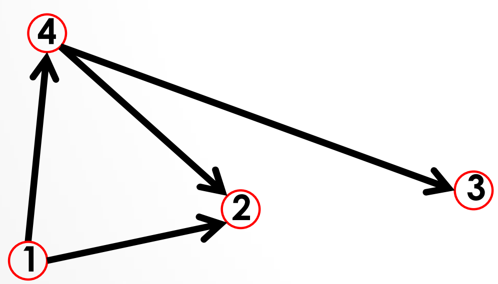
Undirected



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

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

Directed



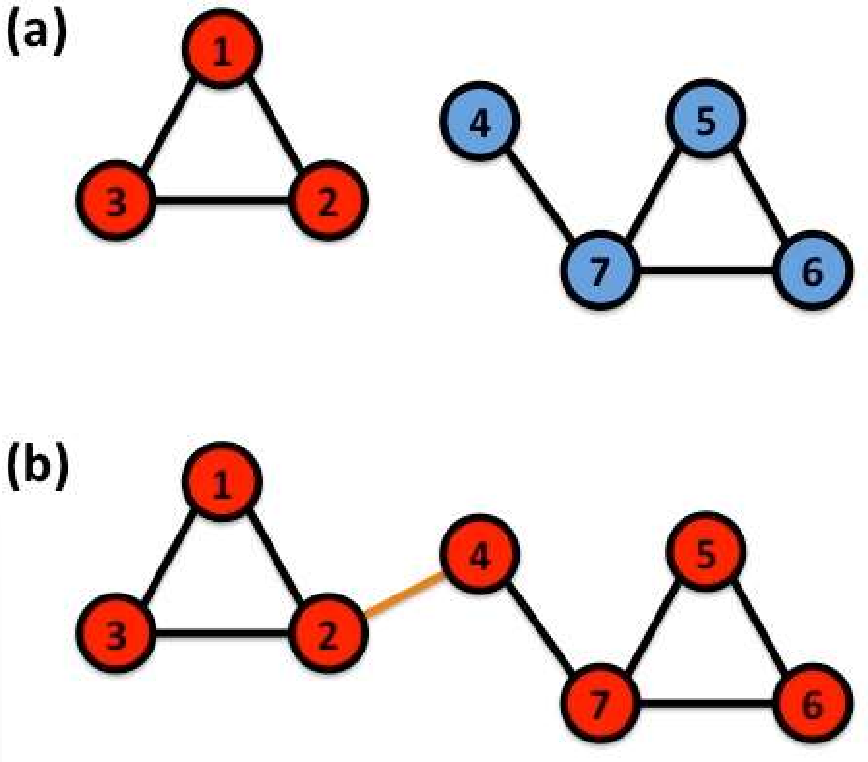
$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

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

$A_{ij} = 1$ can indicate an edge from node j to node i or node i to node j

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:



(a)
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

(b)
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Sociomatrix

	A	B	C	D	E	F	G
A	-	1	0	1	0	0	0
B	1	-	0	1	0	0	0
C	0	0	[illegible]	0	0	0	0
D	1	1	0	-	0	1	0
E	0	0	0	0	-	0	1
F	0	0	0	1	0	-	0
G	0	0	0	0	1	0	-

C has <u>no</u> relationships with A,B,D,E,F

F has a relationship with D

Undirected sociomatrix

	A	B	C	D	E	F	G
A	-	1	0	1	0	0	0
B	1	-	0	1	0	0	0
C	0	0	-	0	0	0	0
D	1	1	0	-	0	1	0
E	0	0	0	0	-	0	1
F	0	0	0	1	0		0
G	0	0	0	0	1	0	-

Symmetric

Symmetric

Symmetric

Directed sociomatrix

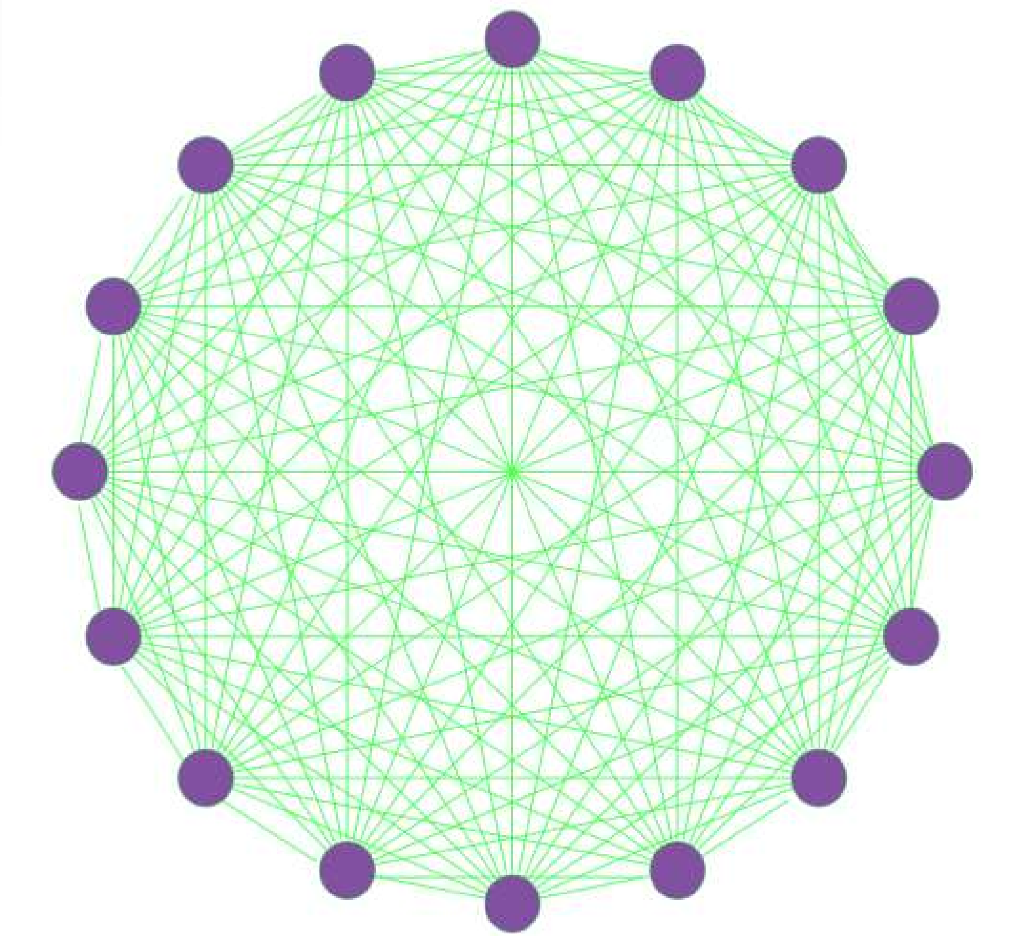
	A	B	C	D	E	F	G
A	-	1	0	1	0	0	0
B	0	-	0	1	0	0	0
C	0	0	-	0	0	0	0
D	0	1	0	-	0	1	0
E	0	0	0	0	-	0	0
F	0	0	0	0	0	-	0
G	0	0	0	0	1	0	-

Unsymmetric

Unsymmetric

Unsymmetric

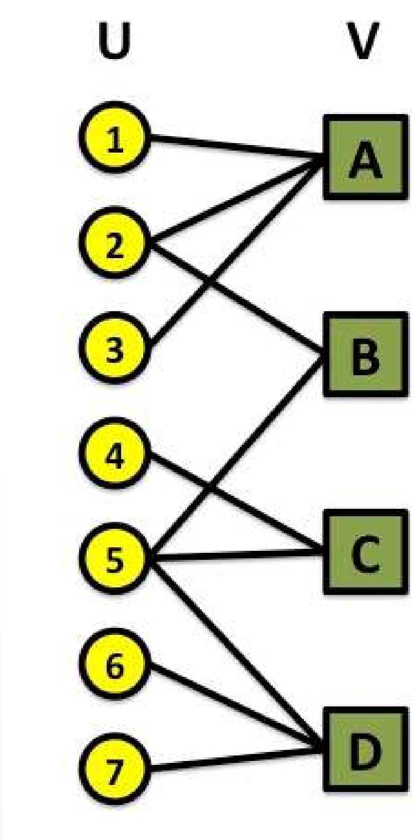
Complete Graph



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

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.

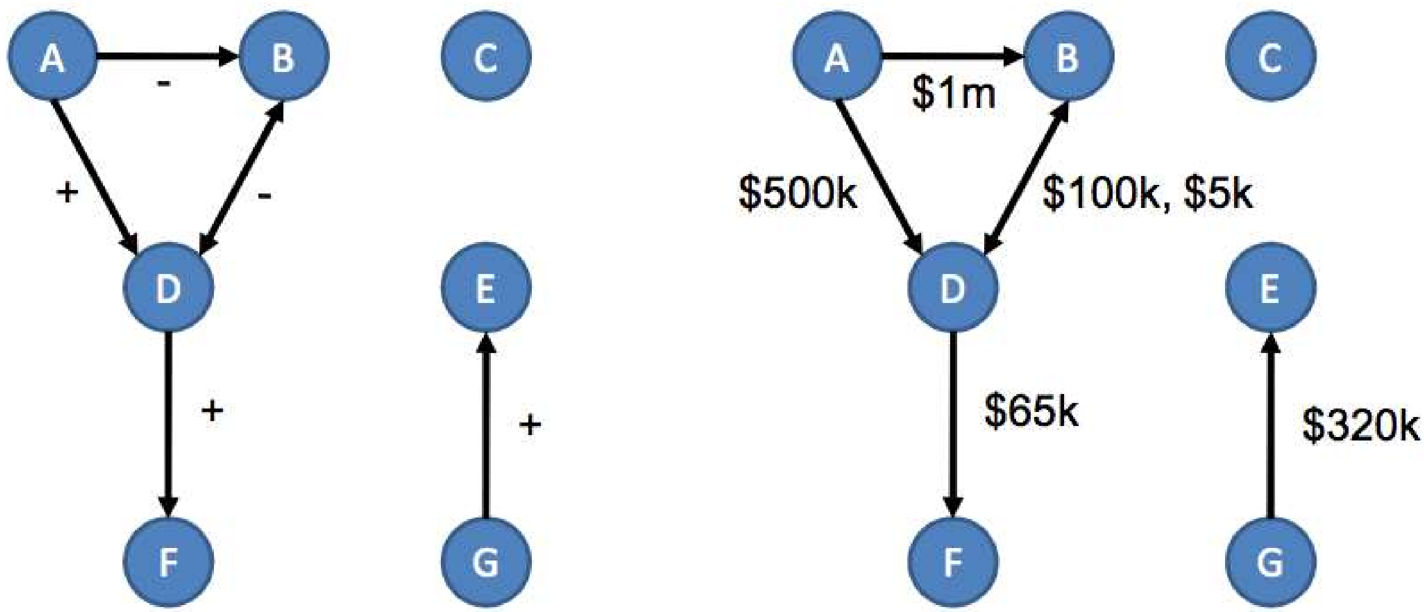


Examples:

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

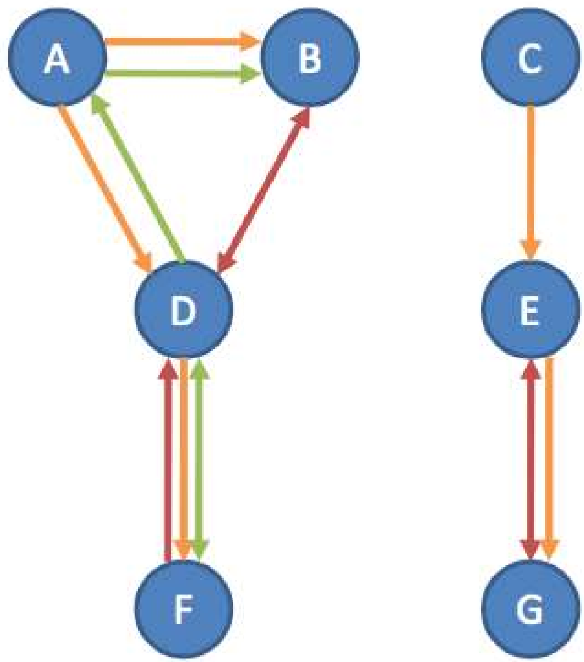
Signed & valued

- Affect in a sorority vs. campaign financing

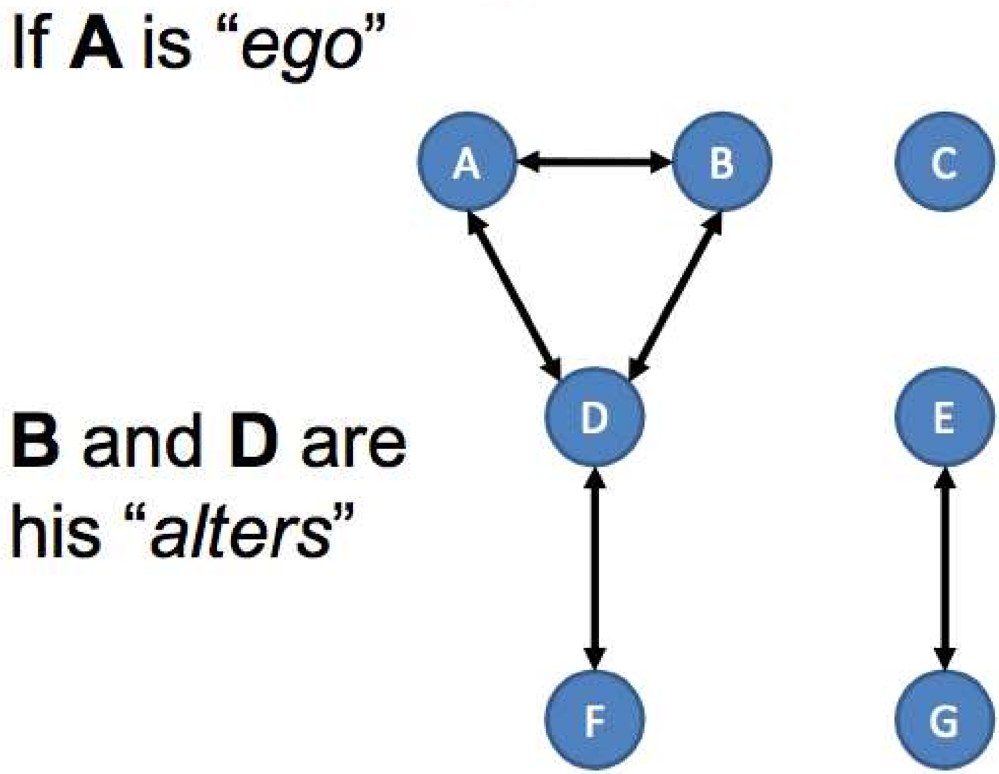


Multi-relational

- Organizations: **authority**, **trust**, & **friendship**



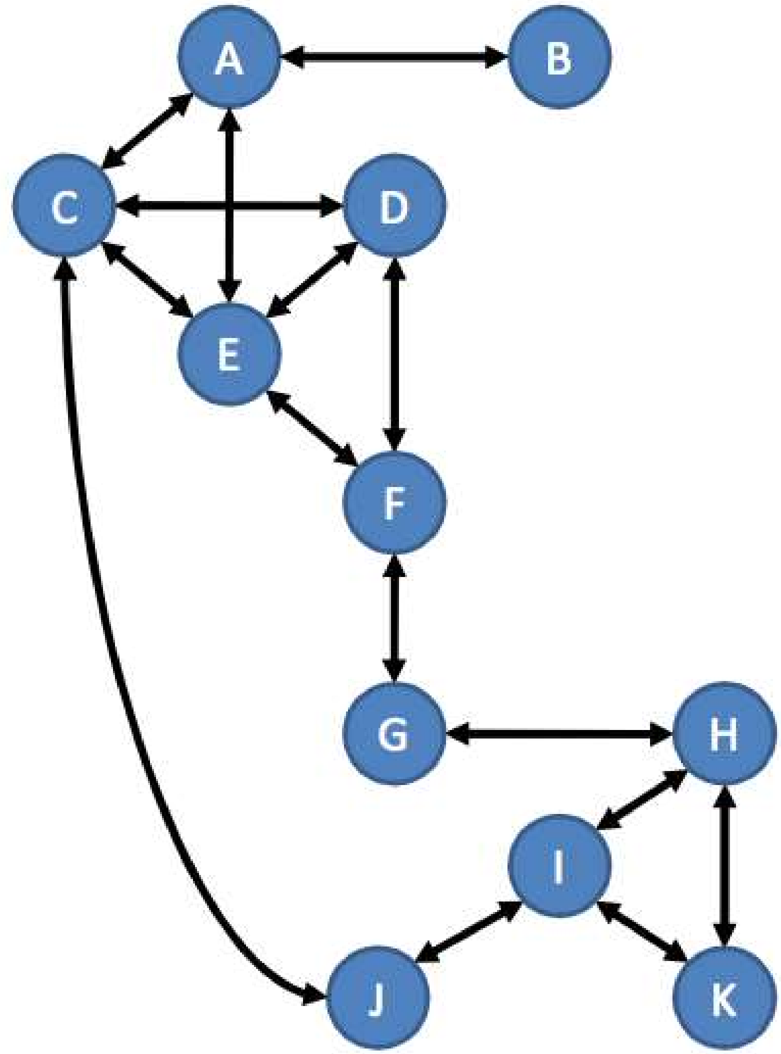
Egos & alters



- Ego-centric networks (or shortened to "ego" networks) are a particular type of network which specifically maps the connections of and from the perspective of a single actor (an "ego").

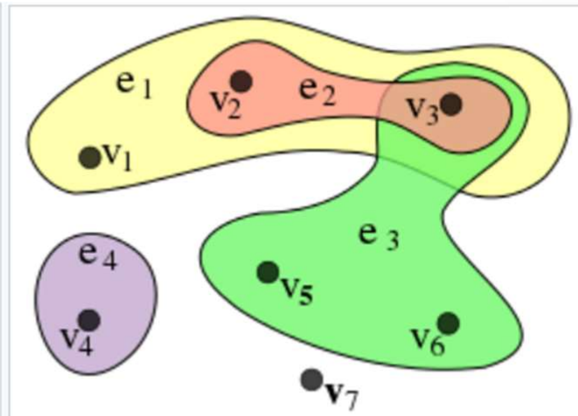
Ego network

- **N-step ego network**: network of all actors and their shared ties, N steps away from ego
  - E's 1-step ego network
  - E's 2-step ego network
  - E's 3-step ego network
  - E's 4-step ego network



Hypergraph

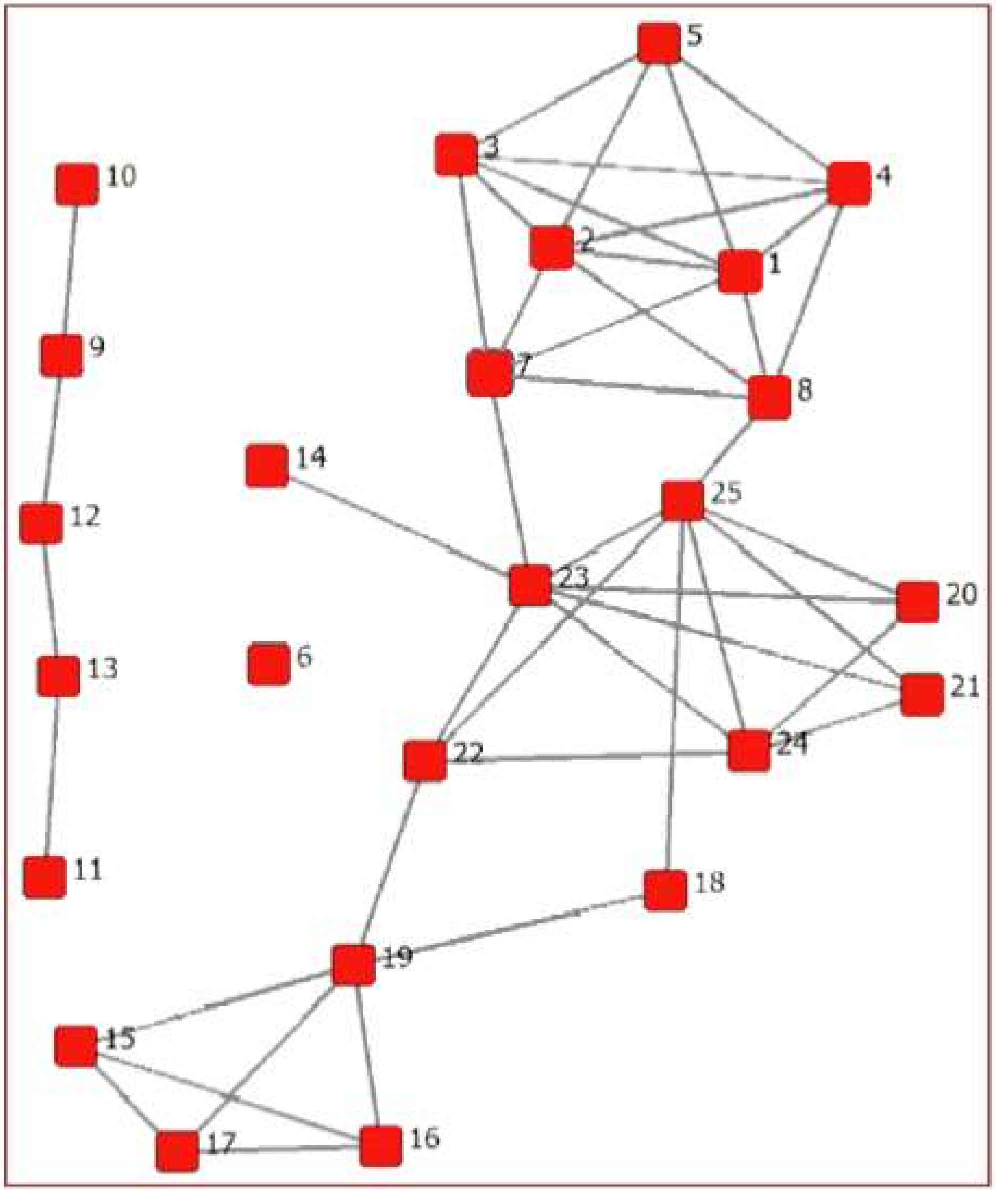
- A **hypergraph** is a generalization of a graph in which an edge can join any number of vertices.
- In contrast, in an ordinary graph, an edge connects exactly two vertices.



An example of an undirected hypergraph, with $X = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ and $E = \{e_1, e_2, e_3, e_4\} = \{\{v_1, v_2, v_3\}, \{v_2, v_3\}, \{v_3, v_5, v_6\}, \{v_4\}\}$. This hypergraph has order 7 and size 4. Here, edges do not just connect two vertices but several, and are represented by colors.

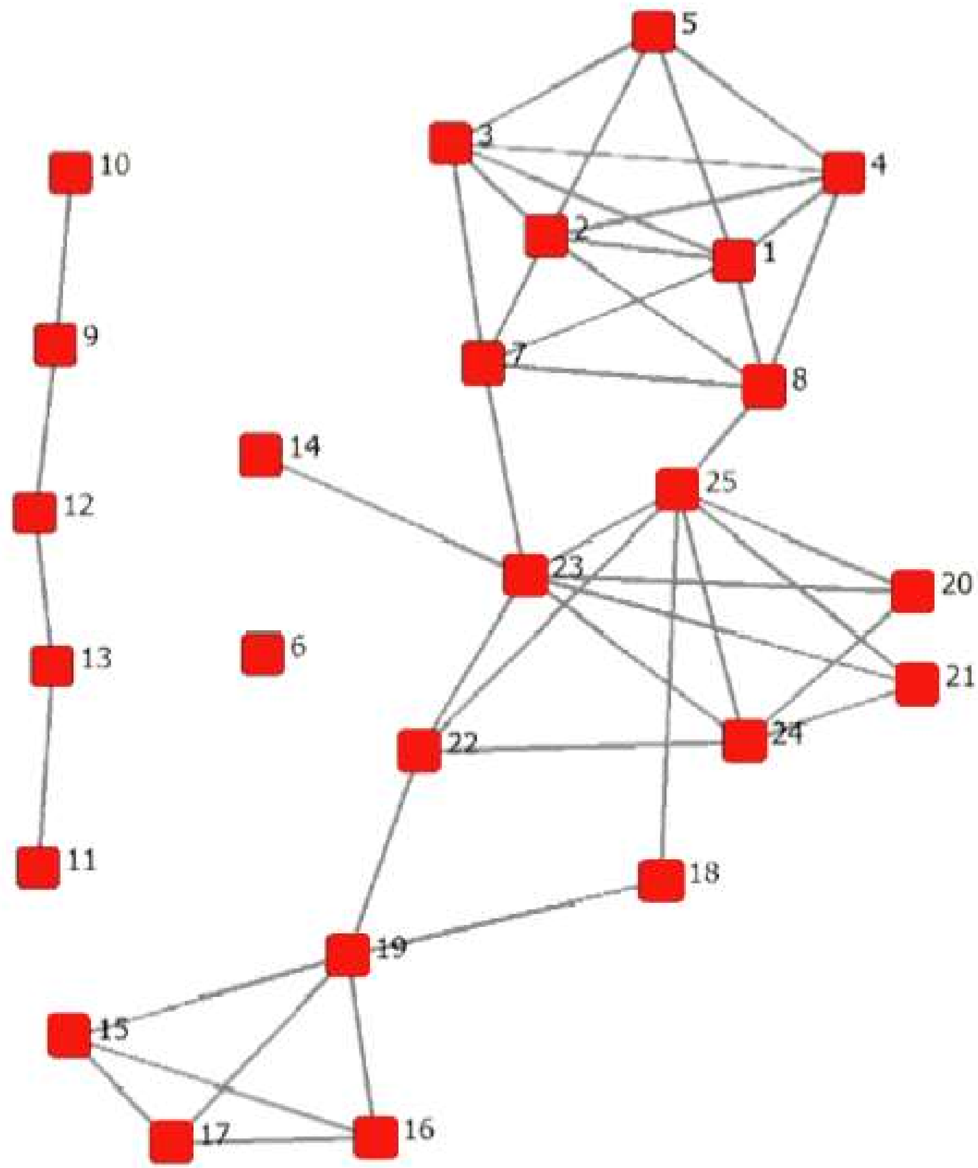
Cliques

- Largest subset of actors that are directly and completely connected to the rest of the set
- "Maximal complete sub-graph"
- 8 is a member of what cliques?



N-Cliques

- Largest subset of actors that are completely-connected with rest of the set ***within N steps***
- N is typically 2
- “Long & stringy”
- Possible for N-clique members to be connected by non-members ☹
- 18 is a member of what 2-cliques?

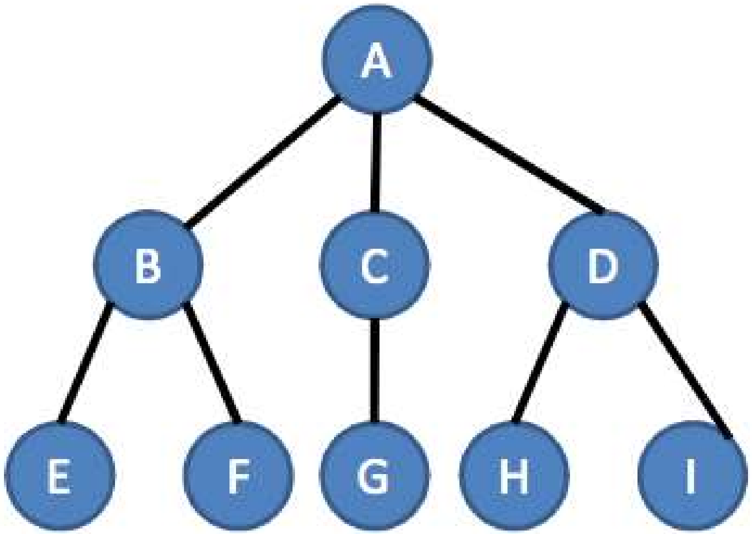


K-Cores

- A k-core of a graph G is a maximal connected subgraph of G in which all vertices have degree at least k.
- Equivalently, it is one of the connected components of the subgraph of G formed by repeatedly deleting all vertices of degree less than k.
- A vertex u has **coreness** c if it belongs to a c-core but not to any (c+1)-core.
- The **degeneracy** of G is the largest k for which G has a k-core.
- The degeneracy of a graph is a measure of how *sparse* it is.

Equivalence

- **Structural equivalence**
  - Sets of actors having exactly the same set of relations as another actor (brothers)
  - {A}, {B}, {C}, {D}, **{E,F}**, {G}, **{H,I}**
- **Automorphic equivalence**
  - Sets of actors having the same patterns of ties and are completely substitutable (cousins)
  - {A}, **{B,D}**, {C}, **{E,F,H,I}**, {G}
- **Regular equivalence**
  - Sets of actors having similar relationships types with other sets (fathers)
  - {A}, **{B,C,D}**, **{E,F,G,H,I}**



Equivalence

- The three types of equivalence (structural, automorphic, and regular) have progressively less strict definitions of what it means for two actors to be "equivalent."
- As we make the definitions less strict (which is not the same as making them less precise!), we are able to understand social networks at increasing levels of abstraction.
- Structural equivalence is the most "concrete" form of equivalence.
  - Two actors are exactly structurally equivalent if they have exactly the same ties to exactly the same other individual actors.
  - Pure structural equivalence can be quite rare in social relations
- Automorphic equivalence is a bit more relaxed.
  - Two actors may not be tied to the same others, but if they are embedded in the same way in the larger structure, they are equivalent.
  - With automorphic equivalence, we are searching for classes of actors who are at the same distance from other sets of actors -- that is, we are trying to find parallel or substitutable sub-structures (rather than substitutable individuals).
- Regular equivalence deserves special attention because it gets at the idea of the "role" that an actor plays with respect to occupants of other "roles" in a structure.
  - The idea of a social role, which is "institutionalized" by normative and sanctioned relationships to other roles is at the very core of the entire sociological perspective.

Node Centrality

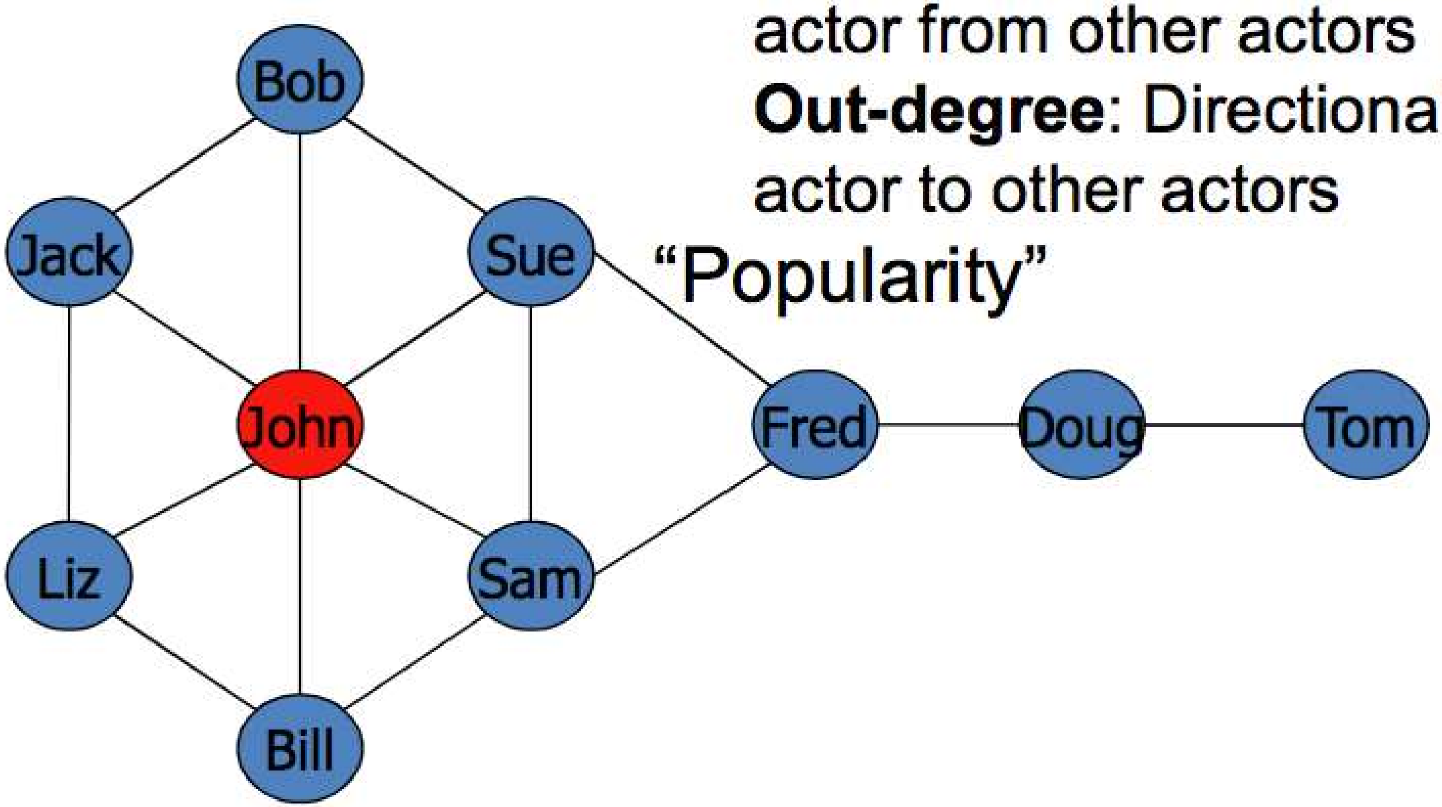
- Node centrality measures how influential a node is in a network.
- Applied to measure how central a node is to the network.
- Centrality is one of the most widely applied measures in Network Science and SNA.
- There are variety of Node centrality measures.

Degree Centrality

Degree: total number of links with other actors

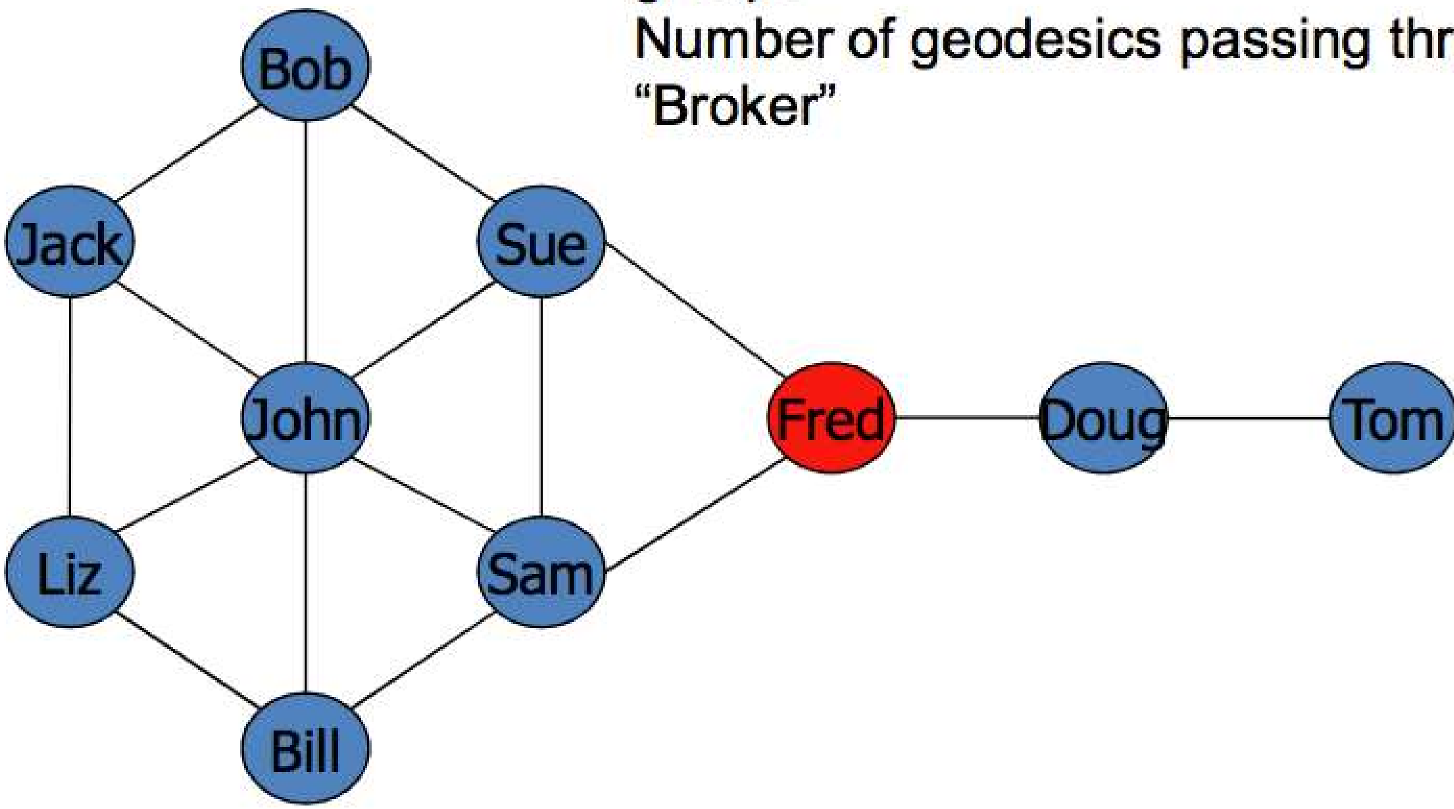
- **In-degree**: Directional links to actor from other actors
- **Out-degree**: Directional links from actor to other actors

"Popularity"



Betweenness Centrality

How much an actor lies between distinct groups
Number of geodesics passing through actor
"Broker"



Betweenness Centrality

- **Betweenness centrality** is a measure of centrality in a graph based on shortest paths.
- For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized.
- The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.

The betweenness centrality of a node v is given by the expression:

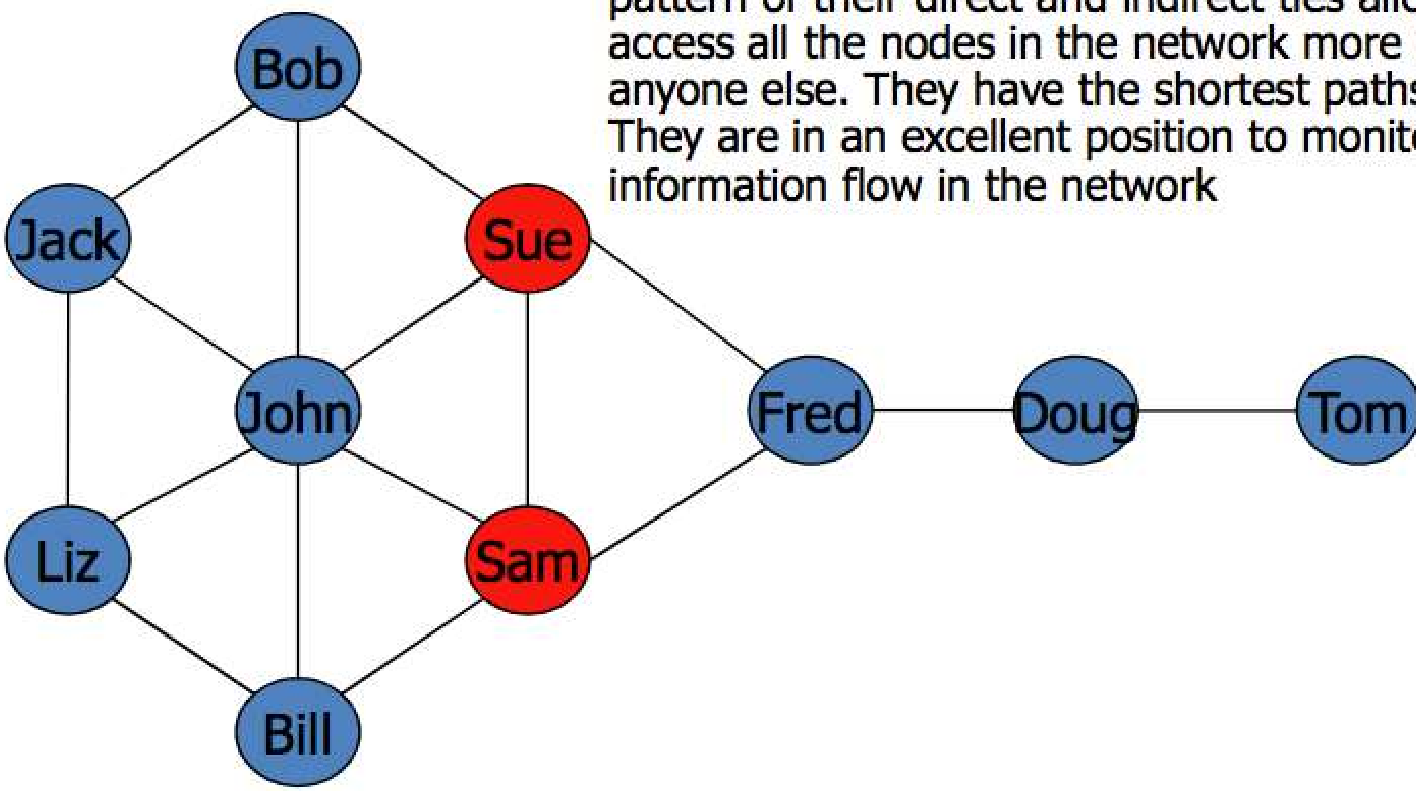
$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v

Closeness Centrality

How easily one actor can reach rest of network
Actor with shortest average path length
"Pulse-taker"

Sue and Sam have fewer connections than John, yet the pattern of their direct and indirect ties allow them to access all the nodes in the network more quickly than anyone else. They have the shortest paths to all others They are in an excellent position to monitor the information flow in the network



Closeness Centrality

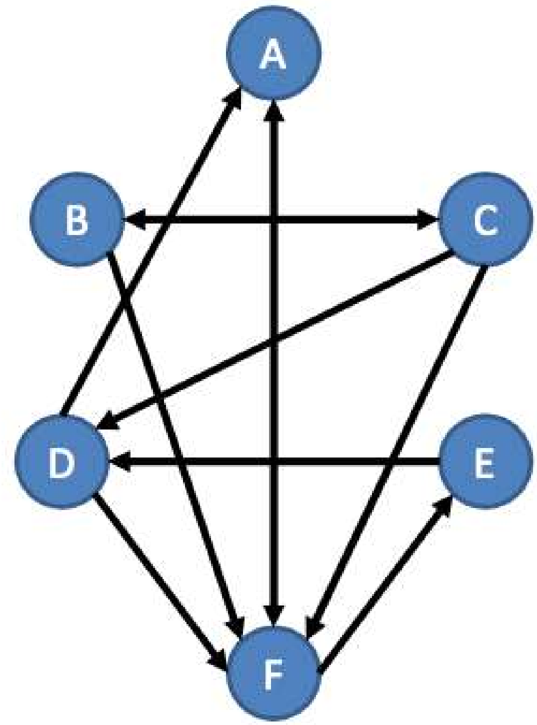
- Closeness centrality of a node is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph.

$$C(x) = \frac{N-1}{\sum_y d(y,x)}$$

- Thus, the more central a node is, the closer it is to all other nodes.

Eigenvector centrality

- A node's centrality is a function of its neighbors' centralities
  - Recompute each node's score as weighted sum of neighbors' centralities
- Highest between and\or degree central actor often have highest eigenvalue central actor, but not always the case with less central actors
  - Nodes B, C, E have equal in-degrees
  - Node E recommended by F, which has high in-degree
  - Nodes B & C only recommended by each other
  - E's eigenvalue > B & C's eigenvalues



- Eigenvector centrality is a measure of the influence of a node in a connected network.
- Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.
- A high eigenvector score means that a node is connected to many nodes who themselves have high scores.

Eigenvector Centrality

Let A be an $n \times n$ matrix and let $X \in \mathbb{C}^n$ be a **nonzero vector** for which

$$AX = \lambda X$$

for some scalar λ. Then λ is called an **eigenvalue** of the matrix A and X is called an **eigenvector** of A associated with λ, or a λ-eigenvector of A.

The set of all eigenvalues of an $n \times n$ matrix A is denoted by $\sigma(A)$ and is referred to as the **spectrum** of A.

Example:

$$\begin{bmatrix} -5 & 2 \\ -7 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 7 \end{bmatrix} = \begin{bmatrix} 4 \\ 14 \end{bmatrix} = 2 \begin{bmatrix} 2 \\ 7 \end{bmatrix}$$

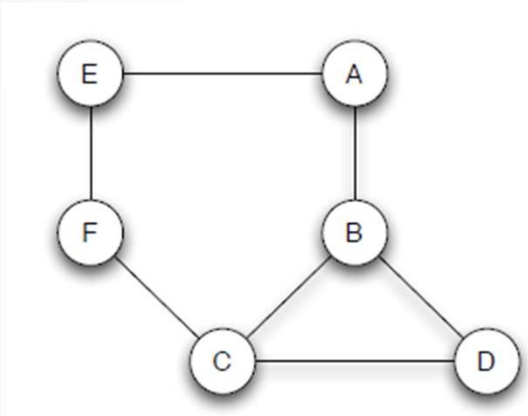
- For a Graph with Adjacency matrix A, the Eigenvector Centrality x_i of node *i* is given by the equation:

$$x_i = \frac{1}{\lambda} \sum_k^N a_{k,i} \, x_k$$

- $\lambda \neq 0$ is the largest eigenvalue
- Google's PageRank is a variant of the Eigenvector Centrality

Exercise

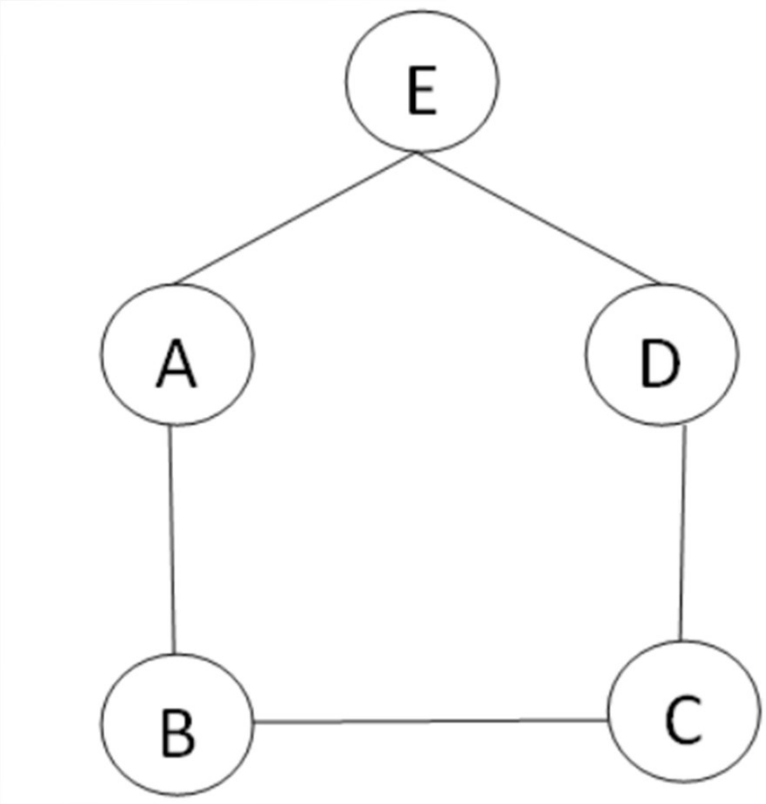
1. We say that a node X is pivotal for a pair of distinct nodes Y and Z if X lies on every shortest path between Y and Z (and X is not equal to either Y or Z).



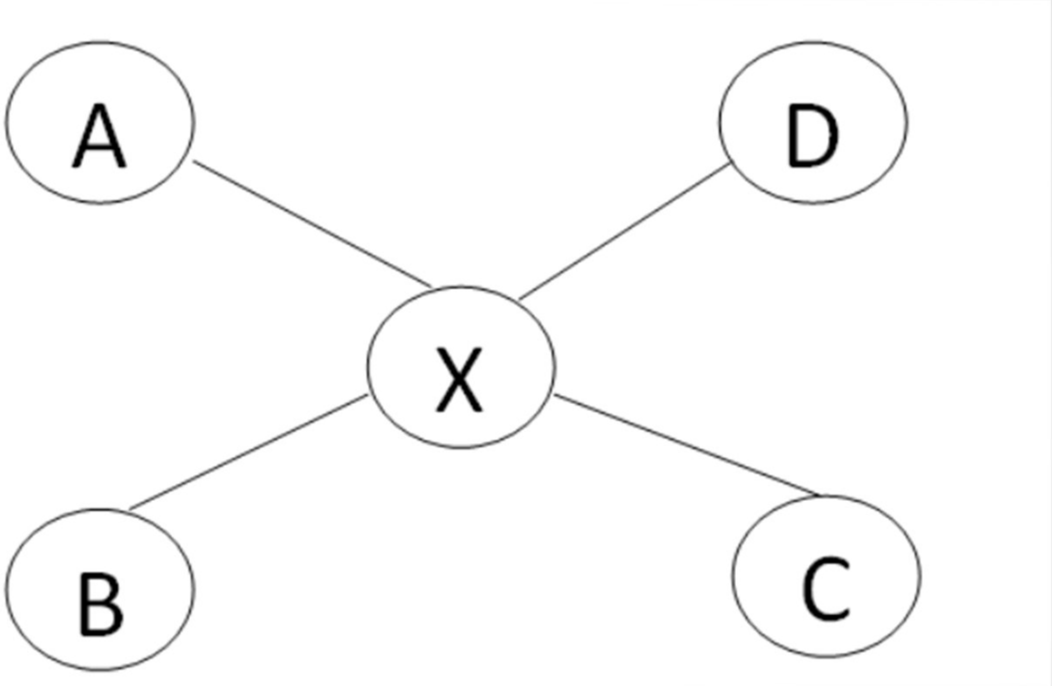
Node B is pivotal for two pairs: the pair consisting of A and C, and the pair consisting of A and D.
On the other hand, node D is not pivotal for any pairs.

a. Give an example of a graph in which every node is pivotal for at least one pair of nodes.
b. Give an example of a graph having at least four nodes in which there is a single node X that is pivotal for every pair of nodes (not counting pairs that include X).

Exercise - Answer



Graph in which every node is pivotal for at least one pair of nodes



Graph having at least four nodes in which there exists a single node X that is pivotal for every pair of nodes

Readings

- Networks, Crowds, and Markets: Reasoning About a Highly Connected World https://www.cs.cornell.edu/home/kleinber/networks-book/
  - Chapter 2.1-2.2
- Social Network Analysis. Tanmoy Chakraborty. Chapter 2