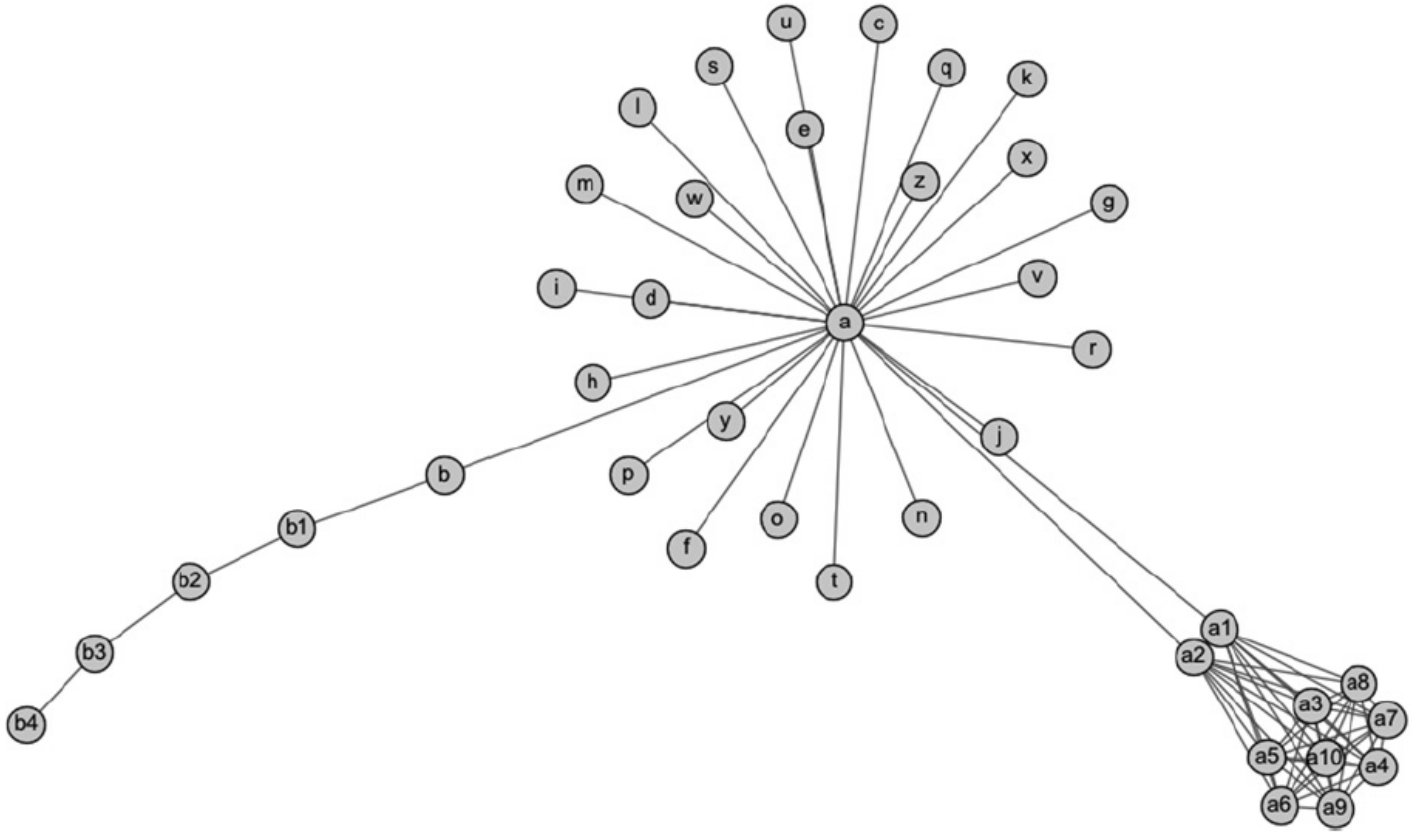
Social Network: Basic Structure and Measures

Pawan Goyal

CSE, IITKGP

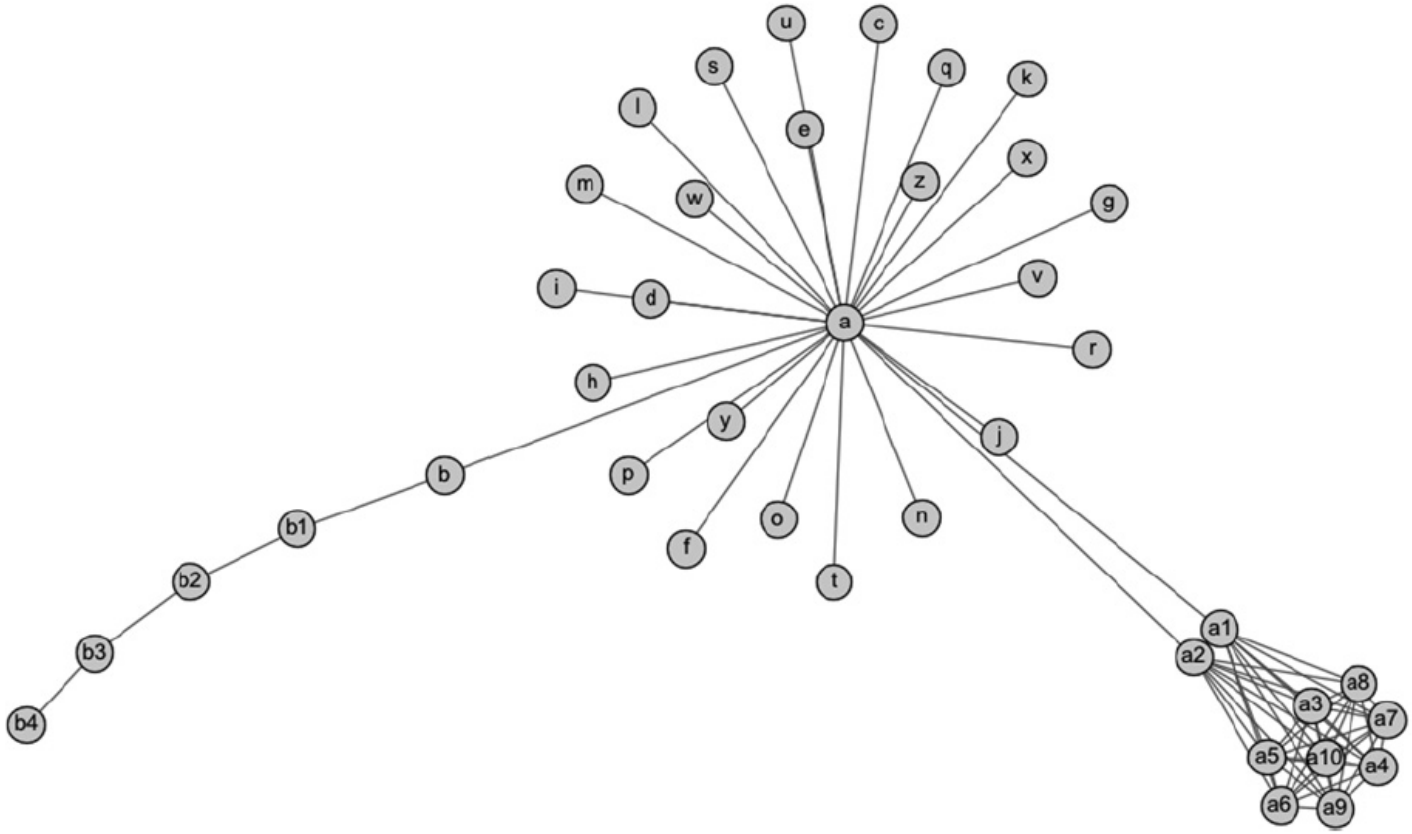
August 19, 2014

A sample social network



Some observations

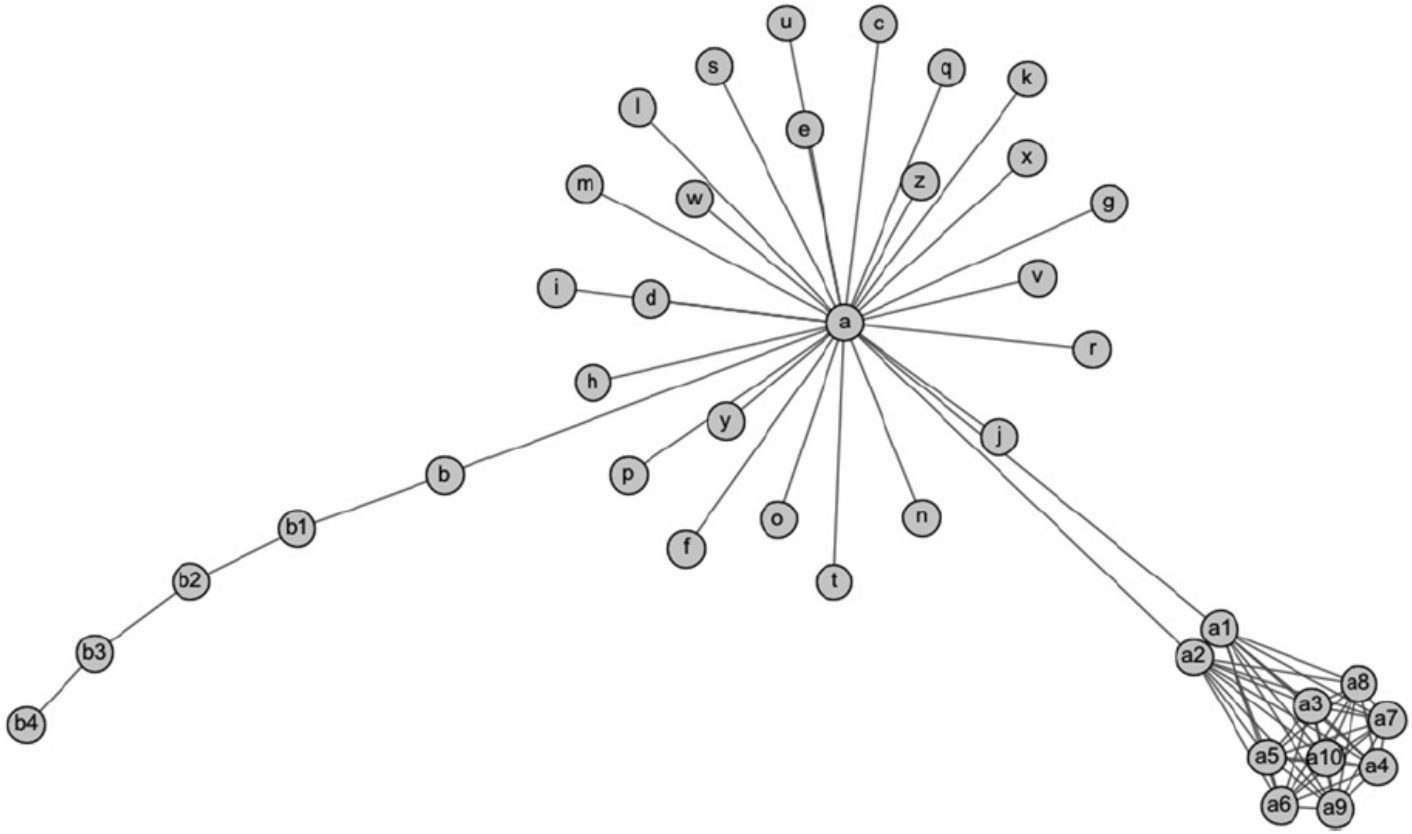
A sample social network



Some observations

Node a has a lot of relationships with other nodes.

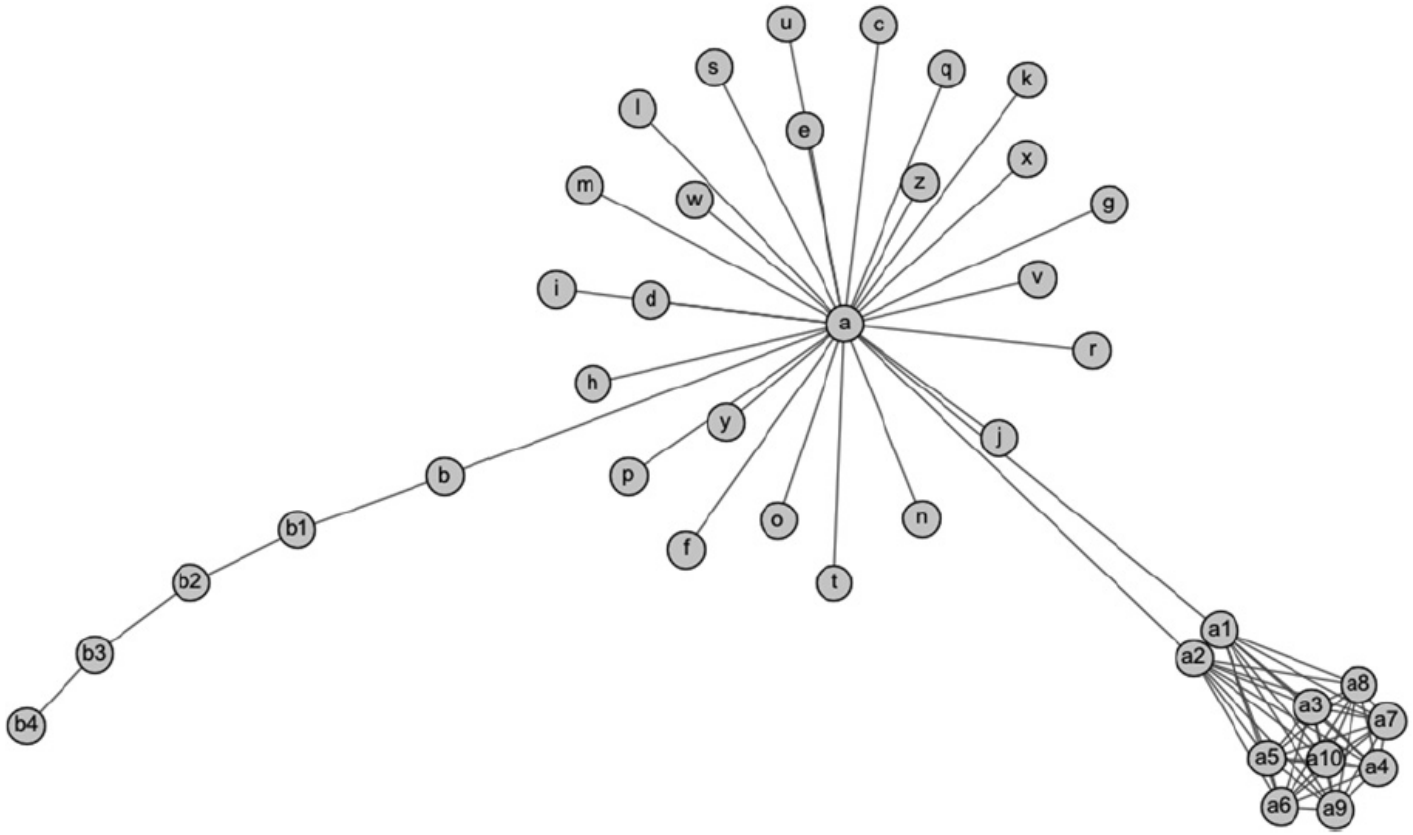
A sample social network



Some observations

There is a long series of relationships from a to b to b_1 to b_2 and so on.

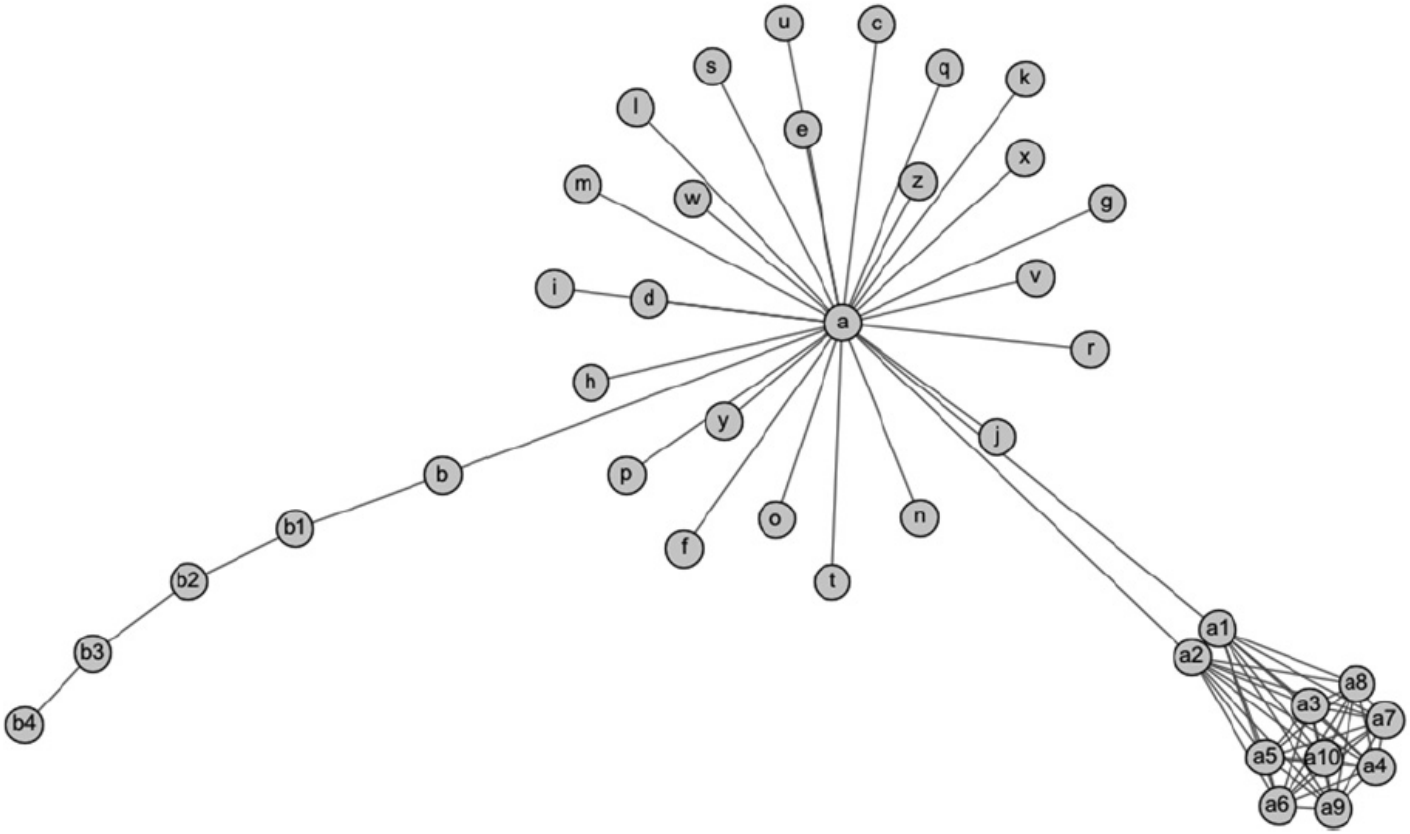
A sample social network



Some observations

There are many relationships among the nodes a_1 through a_{10}, which might be a group of people with very close relationships.

A sample social network



Some observations

Are there formal methods for quantifying these insights: node's importance, connectedness, communities?

Six degrees of separation

- *The idea that people who seem very unlike one another may be connected by a chain of six or fewer acquaintances*

Six degrees of separation

- *The idea that people who seem very unlike one another may be connected by a chain of six or fewer acquaintances*

Six degrees of Kevin Bacon

Connect any actor to Kevin Bacon through co-stars in movies, in as few steps as possible.

Six degrees of separation

- *The idea that people who seem very unlike one another may be connected by a chain of six or fewer acquaintances*

Six degrees of Kevin Bacon

Connect any actor to Kevin Bacon through co-stars in movies, in as few steps as possible.

Erdos Number

More well-established notion among mathematicians and computer scientists, how many co-author relationships separate them from the famous mathematician, Paul Erdos.

Social Network

- *A network or graph is a set of nodes and edges.*

Social Network

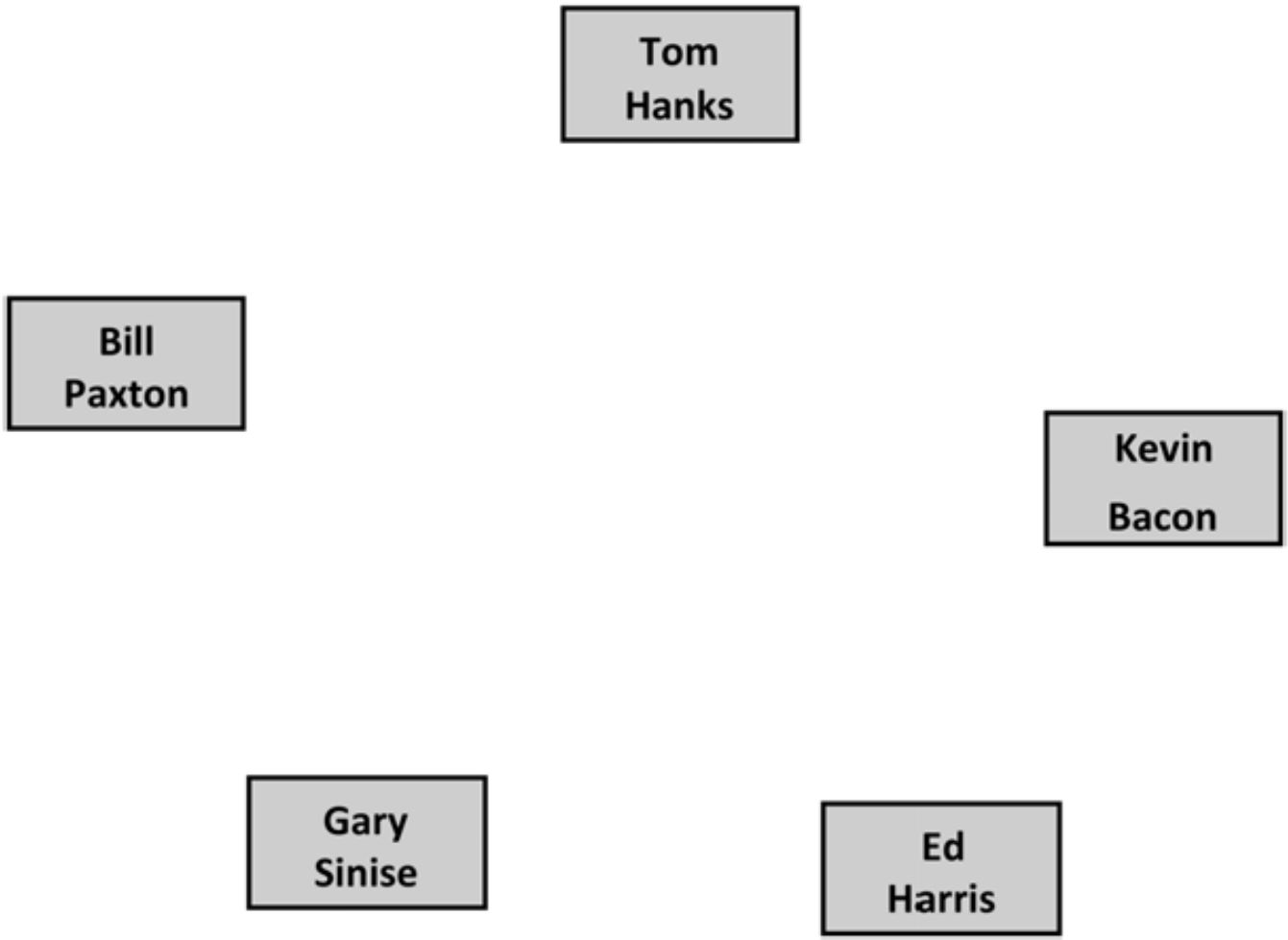
- *A network or graph is a set of nodes and edges.*

Apollo 13

Five stars: Tom Hanks, Gary Sinise, Ed Harris, Bill Paxton and Kevin Bacon.

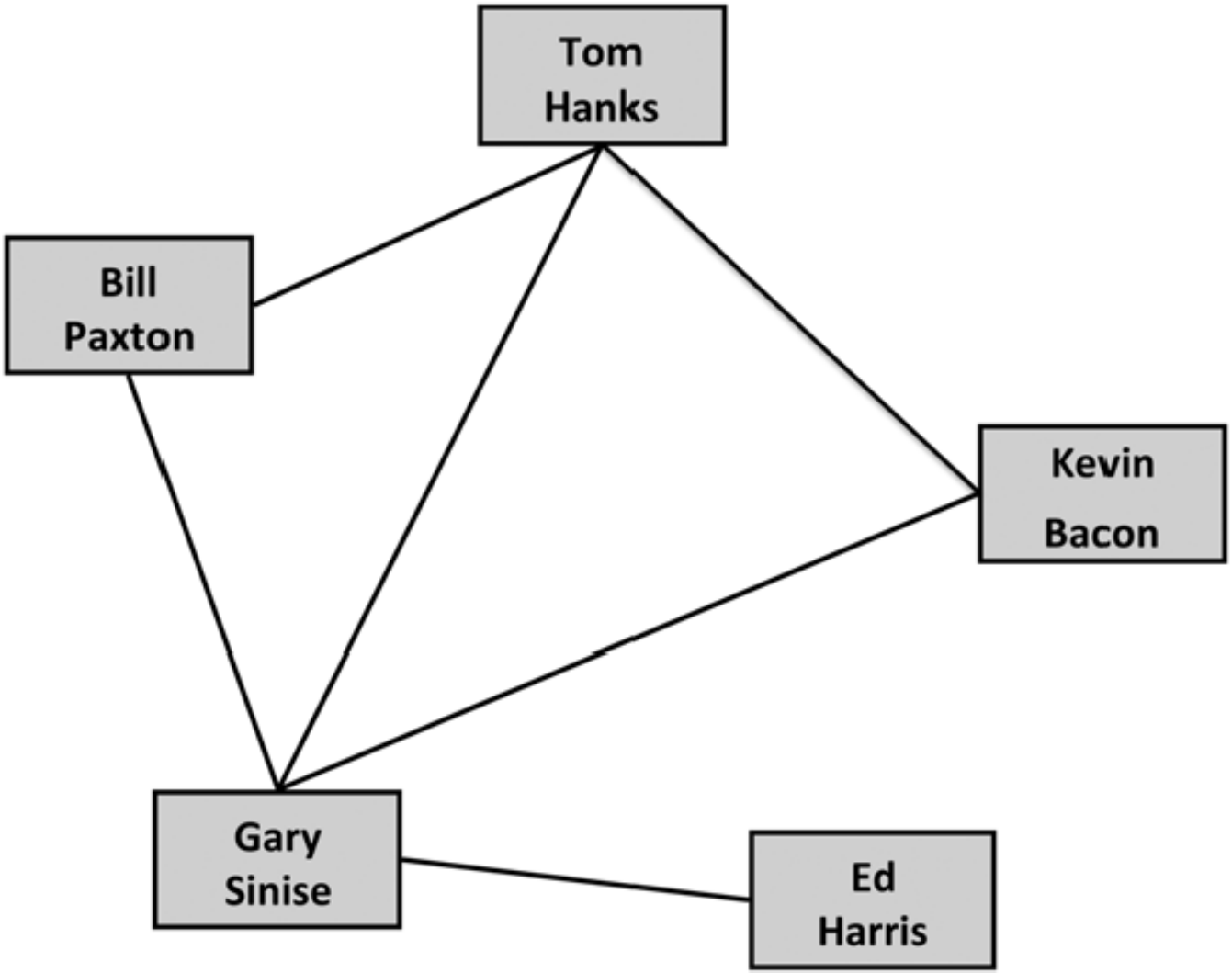
Nodes

Each actor can be represented as a node in the graph.



Edges

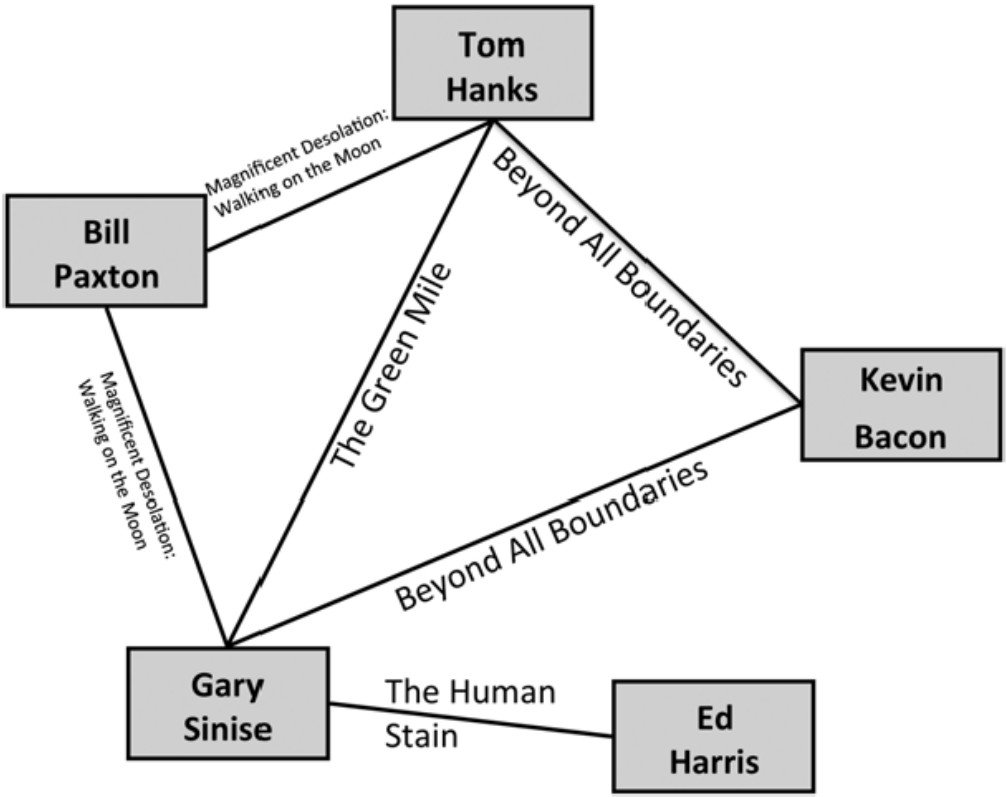
The actors can be linked if they were in another movie together.



Edge Features

Edge Labels

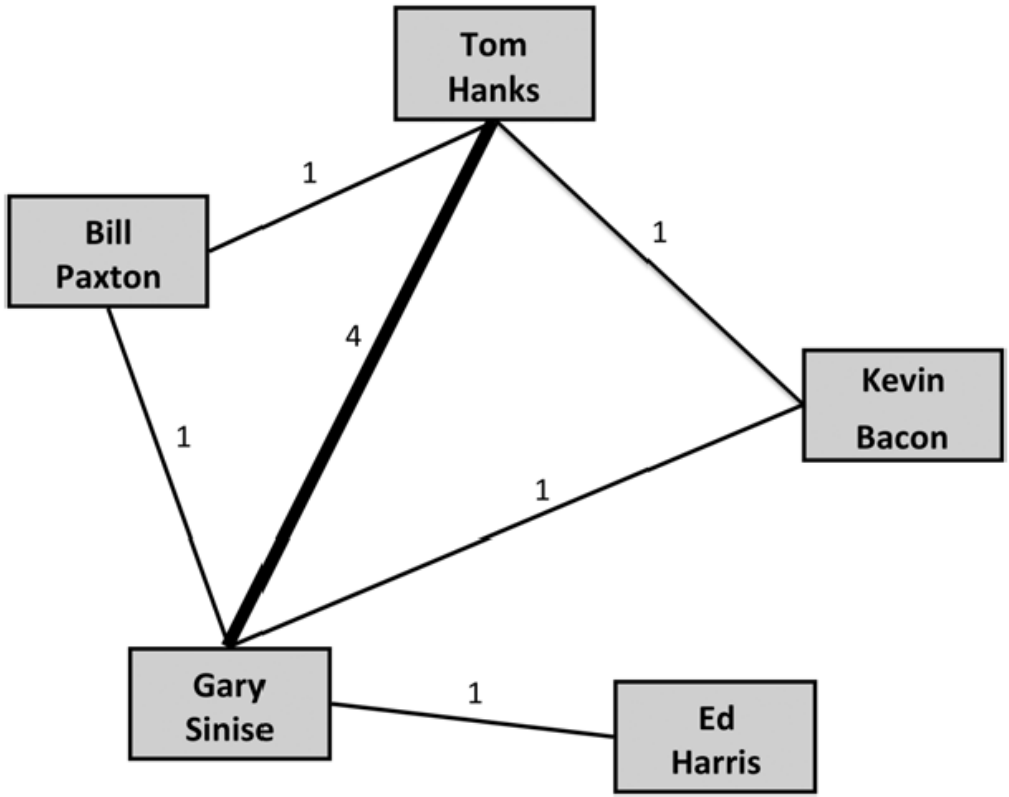
Edge label can give more information about the relationship.



Edge Features

Edge Weights

Edge weight indicates numerical information about a relationship, often the strength of the relationship



Undirected and Directed Network

- *Edges can be either directed or undirected.*
- **Undirected edge** indicates a mutual relationship, whereas
- **Directed edge** indicates a relationship that a node has with the other, not necessarily reciprocated.

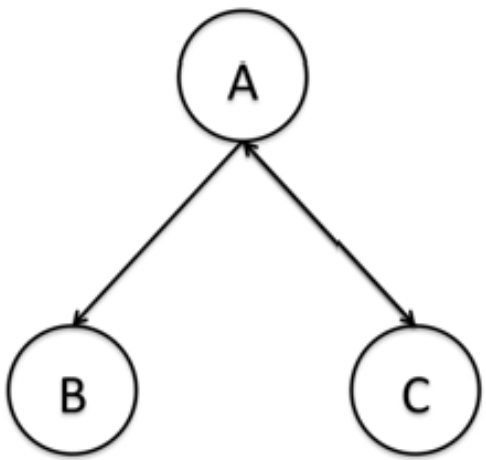
Undirected and Directed Network

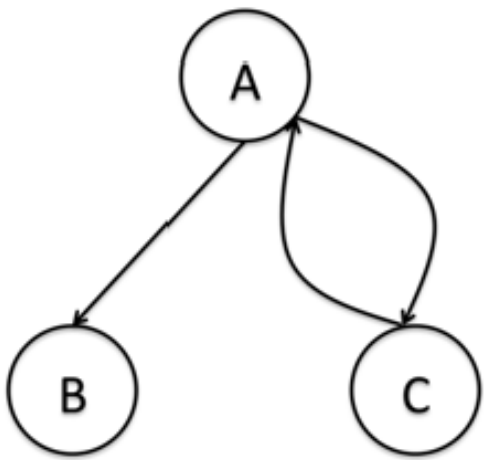
The example under consideration is an undirected network.

Undirected and Directed Network

The example under consideration is an undirected network.

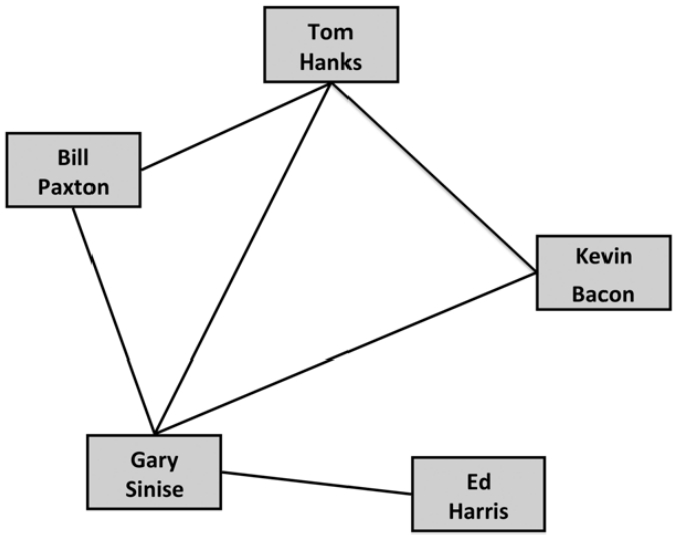
Directed Network Example: Email Communication





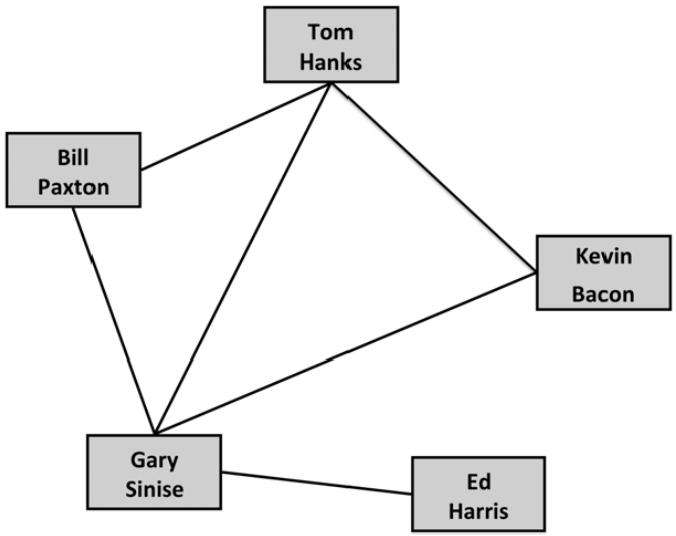
Adjacency Lists

Indicated by listing the pair of nodes that ae connected.



Adjacency Lists

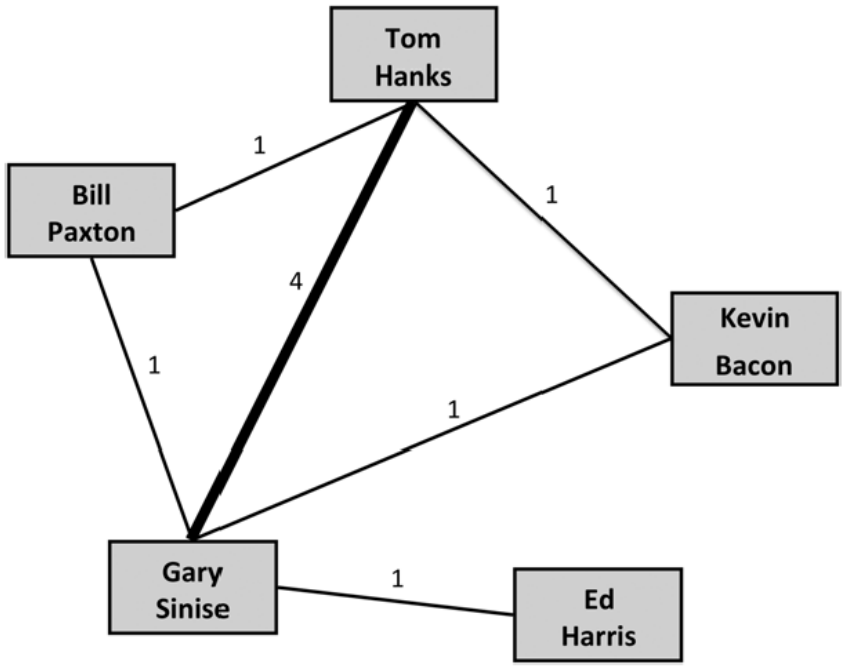
Indicated by listing the pair of nodes that ae connected.



Tom Hanks, Bill Paxton
Tom Hanks, Gary Sinise
Tom Hanks, Kevin Bacon
Bill Paxton, Gary Sinise
Gary Sinise, Kevin Bacon
Gary Sinise, Ed Harris

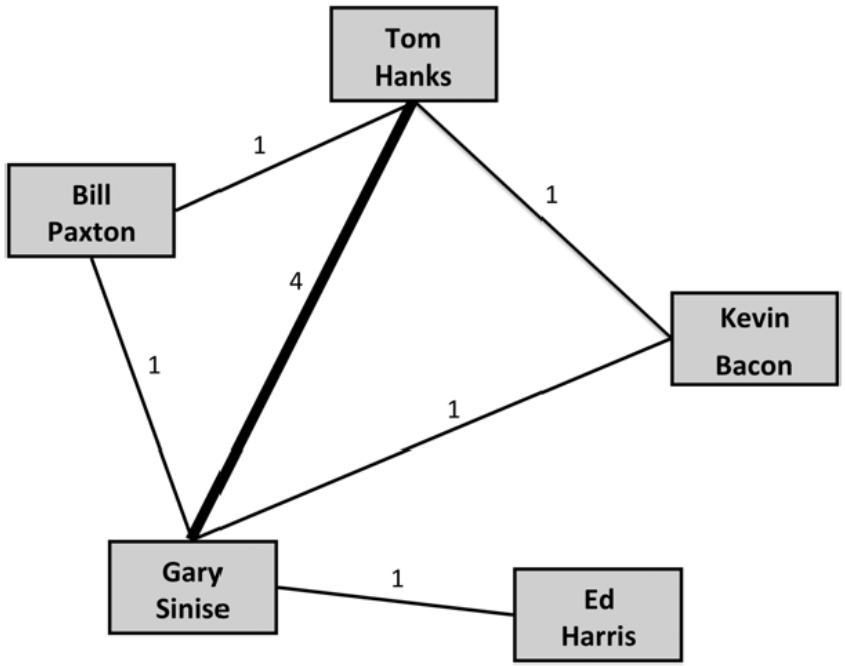
Adjacency Lists

Additional information can be incorporated



Adjacency Lists

Additional information can be incorporated



Tom Hanks, Bill Paxton, 1
Tom Hanks, Gary Sinise, 4
Tom Hanks, Kevin Bacon, 1
Bill Paxton, Gary Sinise, 1
Gary Sinise, Kevin Bacon, 1
Gary Sinise, Ed Harris, 1

Adjacency Matrix

- All the nodes are listed on both the X-axis and Y-axis.
- Values are filled in to the matrix to indicate an edge.

Adjacency Matrix

- All the nodes are listed on both the X-axis and Y-axis.
- Values are filled in to the matrix to indicate an edge.

	Tom Hanks	Bill Paxton	Gary Sinise	Kevin Bacon	Ed Harris
Tom Hanks	0	1	1	1	0
Bill Paxton	1	0	1	0	0
Gary Sinise	1	1	0	1	1
Kevin Bacon	1	0	0	0	0
Ed Harris	0	0	1	0	0

XML formats

```
<Person>
    <name>Tom Hanks</name>
    <connection>Bill Paxton</connection>
    <connection>Gary Sinise</connection>
    <connection>Kevin Bacon</connection>
</Person>
```

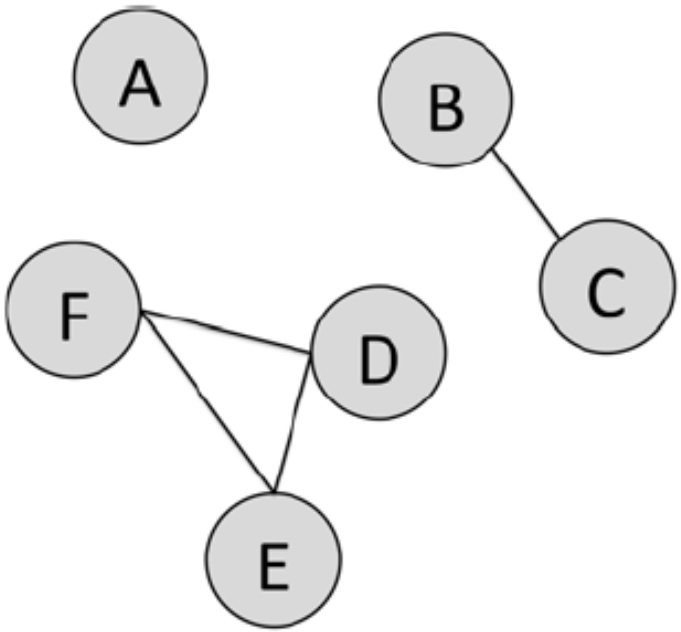
Subnetworks

Subnetwork

When we are considering a subset of the nodes and edges in a graph, it is called a *subnetwork*.

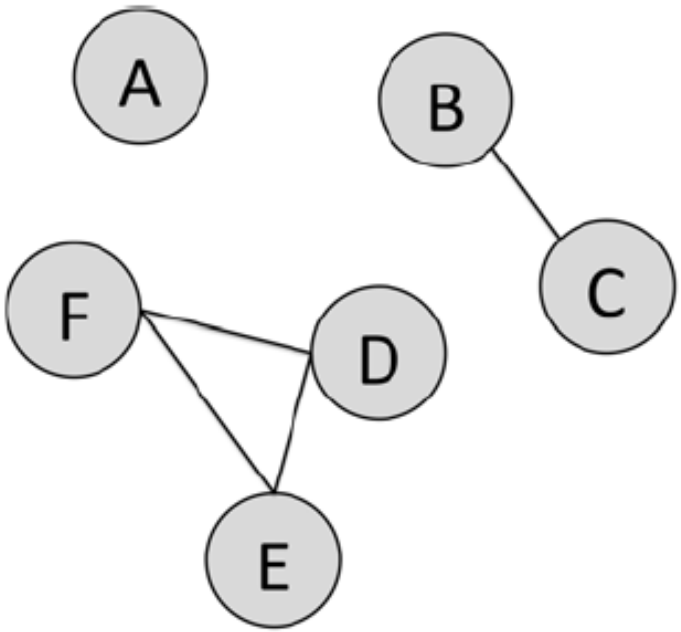
Simplest subnetworks: *Singletons*, nodes that have no edges.

Similarly, *dyad* for two nodes, *triad* for three nodes.



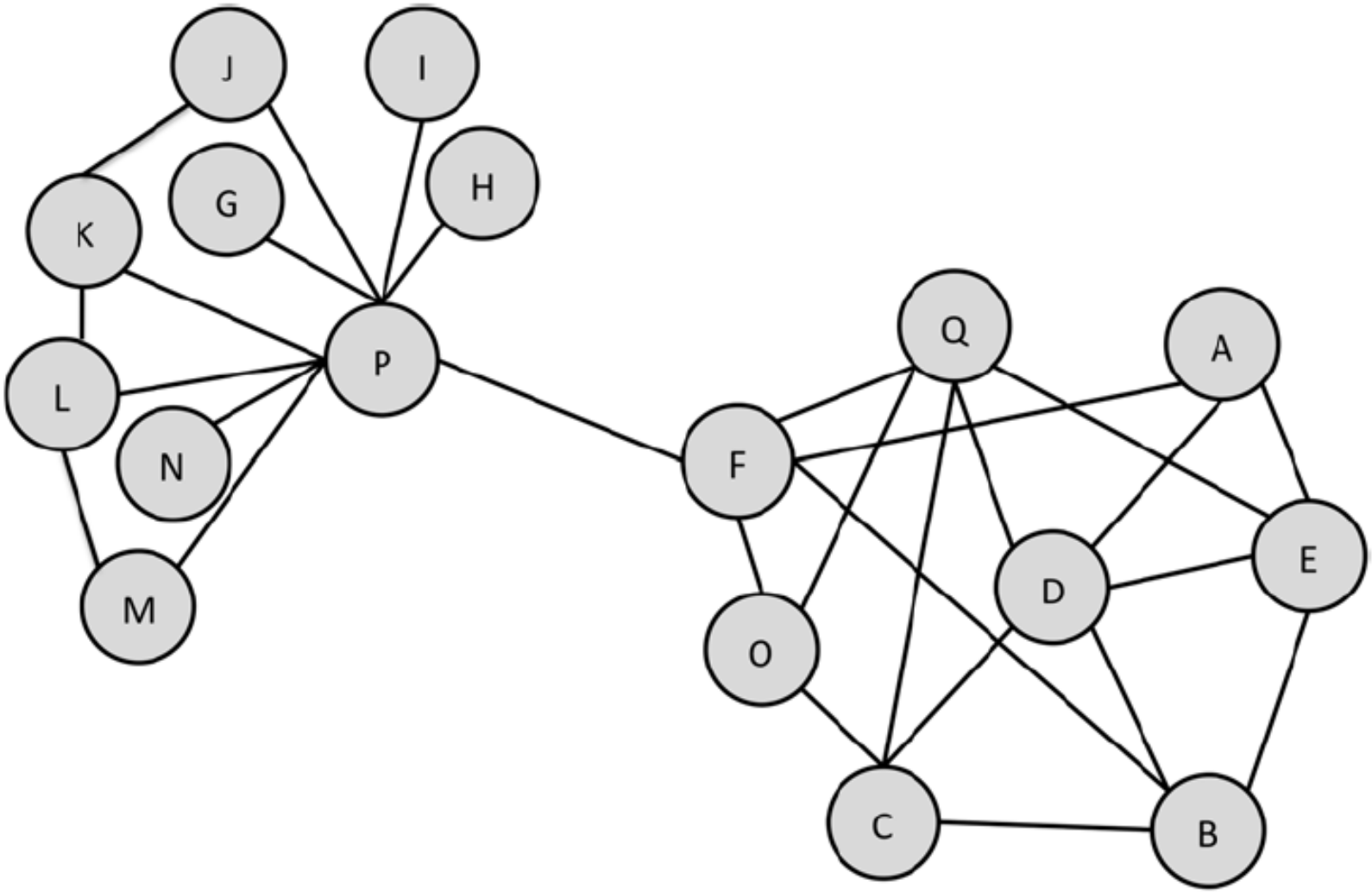
Cliques

- *Property for a groups of nodes*
- If all the nodes in a group are connected to one another, it is called a *clique*.



Clusters

- *No strict definition like there is for a clique*



Egocentric Networks

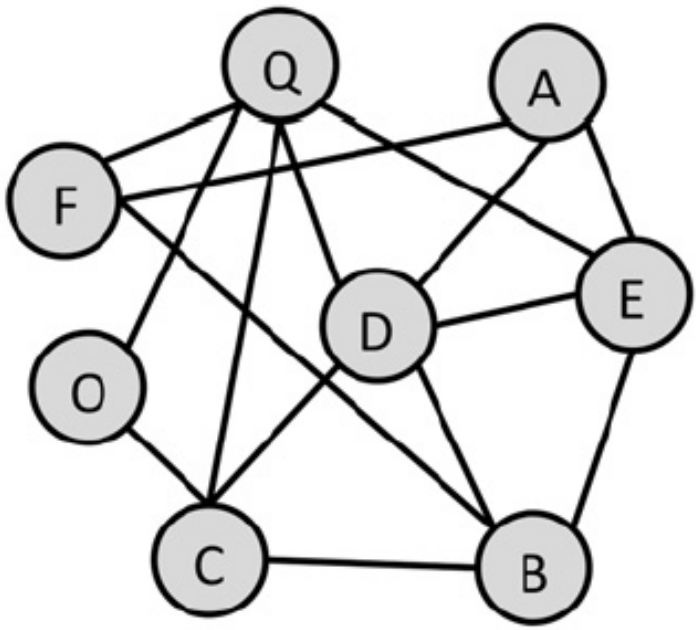
- *This is a network constructed by selecting a node and all of its connections.*

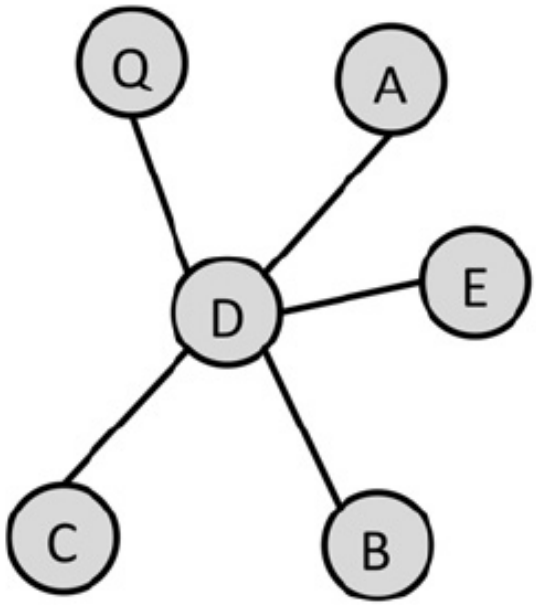
Egocentric Networks

- *This is a network constructed by selecting a node and all of its connections.*

1-degree egocentric network

We are going one step away from D in the network





Egocentric Networks

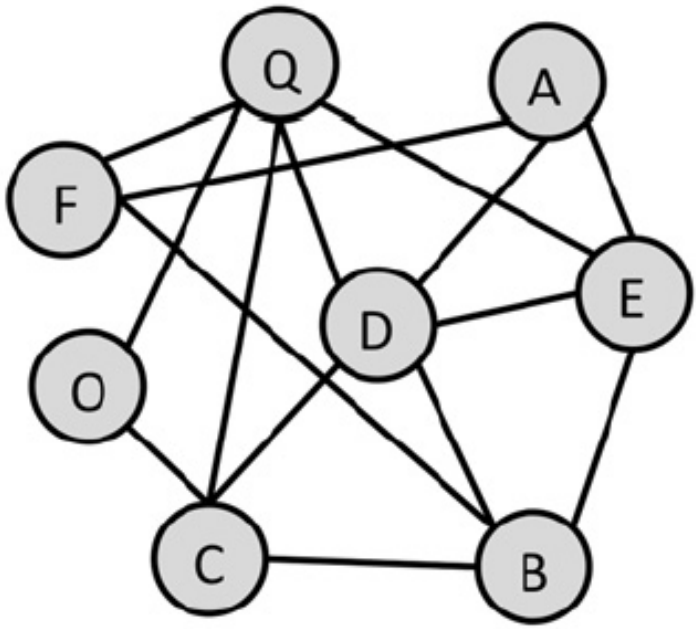
- *This is a network constructed by selecting a node and all of its connections.*

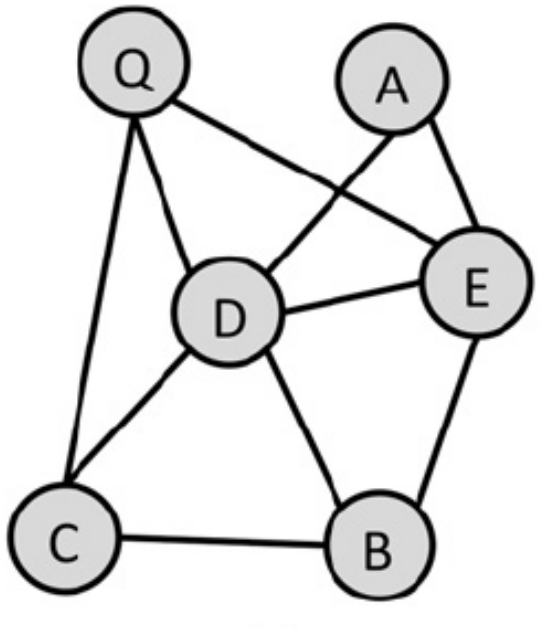
Egocentric Networks

- *This is a network constructed by selecting a node and all of its connections.*

1.5-degree egocentric network

We want to see only D's neighbors and their connections.



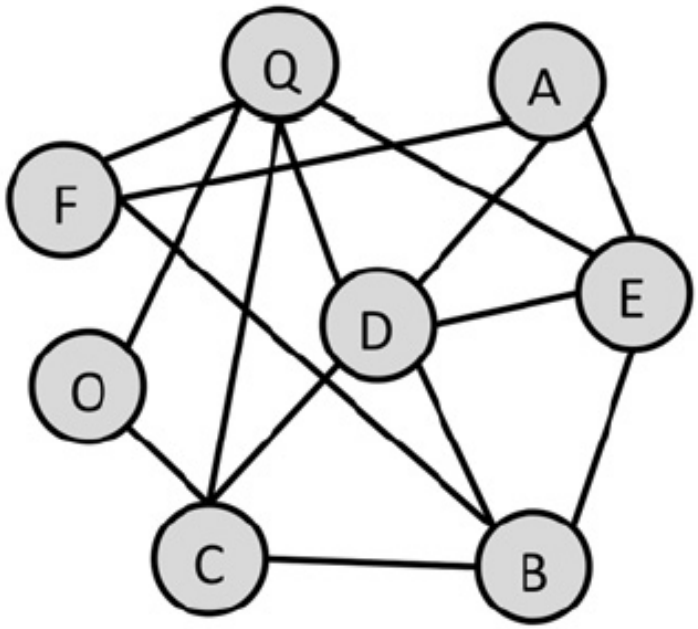


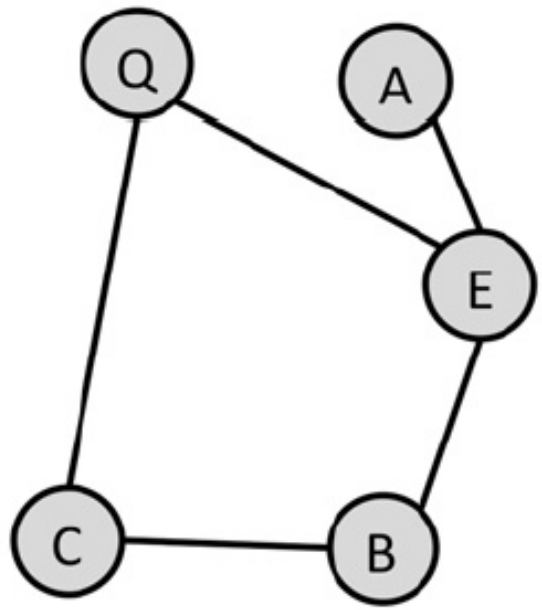
Egocentric Networks

- *This is a network constructed by selecting a node and all of its connections.*

1.5 egocentric network with D excluded

Central node and its edges are excluded.



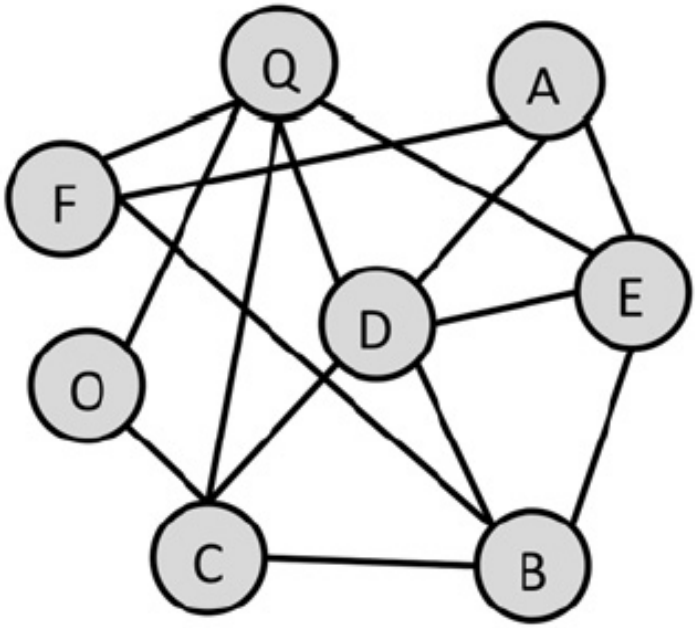


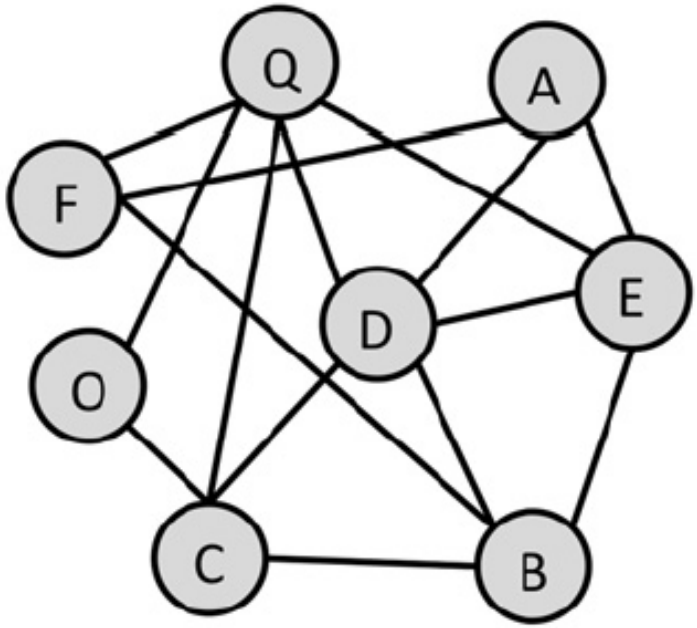
Egocentric Networks

- *This is a network constructed by selecting a node and all of its connections.*

2-degree egocentric network

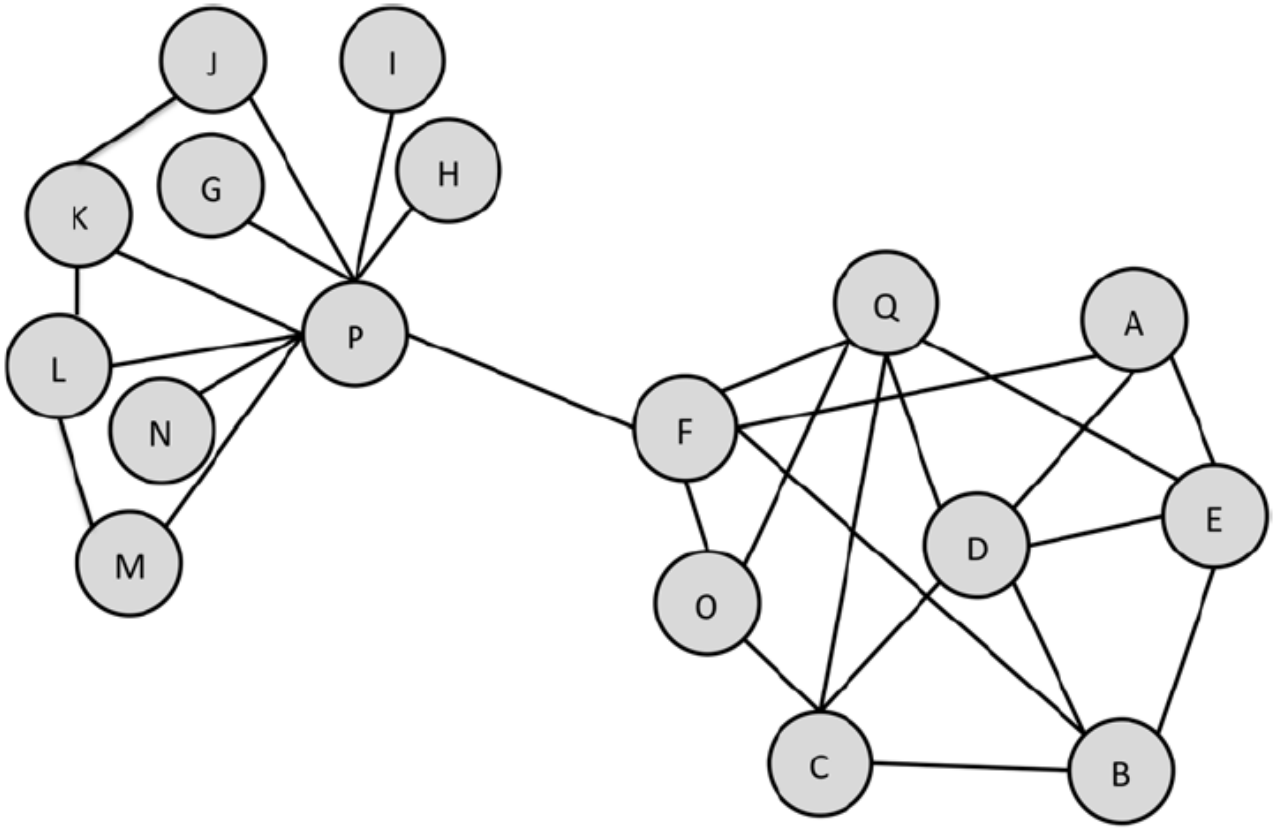
Includes all of D's neighbors, their connections to one another and all of their neighbors.





Paths

- *A path is a series of nodes that can be traversed following edges between them.*



Path connecting node M to node C?

Paths

We are typically interested only in the shortest path from one node to another.

Paths

We are typically interested only in the shortest path from one node to another.

Shortest paths will be an important measure and are sometimes called geodesic distances.

Connectedness

Connectedness

Two nodes in a graph are called connected if there is a path between them in the network.

An entire graph is called connected if all pairs of nodes are connected.

Directed graphs

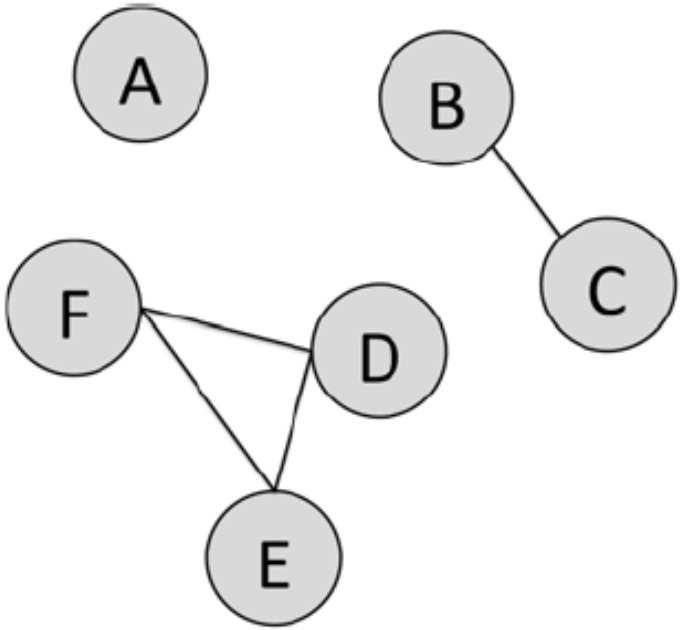
While there may be a set of edges connecting two nodes, they may not all point in the right direction.

Strongly connected: If there are edges that can be followed in the correct direction to find a path between every pair of nodes.

Weakly connected: If a path cannot be found between all pairs of nodes using the direction of the edges, but paths can be found if the directed edges are treated as undirected.

Connected Components

If a graph is not connected, it may have subgraphs that are connected. There are called *connected components*.



This graph contains a three-node connected component, a two-node connected component, and a singleton.

Bridges and Hubs

- *These are two basic concepts that we can use to identify particularly important edges and nodes.*

Bridges and Hubs

- *These are two basic concepts that we can use to identify particularly important edges and nodes.*

Bridge

Intuitively, a bridge is an edge that connects two otherwise separate groups of nodes in the network.

Bridges and Hubs

- *These are two basic concepts that we can use to identify particularly important edges and nodes.*

Bridge

Intuitively, a bridge is an edge that connects two otherwise separate groups of nodes in the network.
Formally, a bridge is an edge that, if removed, will increase the number of connected components in a graph.

Bridges and Hubs

- *These are two basic concepts that we can use to identify particularly important edges and nodes.*

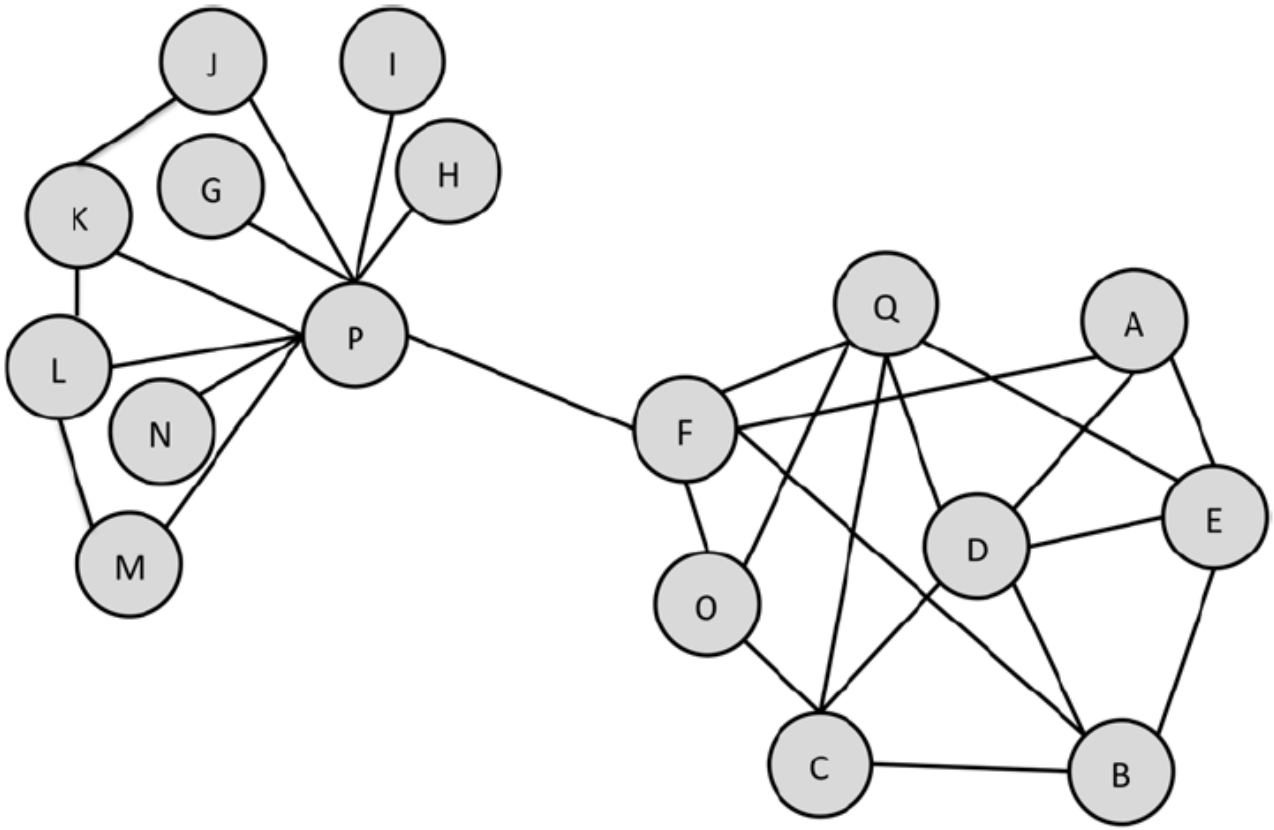
Bridge

Intuitively, a bridge is an edge that connects two otherwise separate groups of nodes in the network.
Formally, a bridge is an edge that, if removed, will increase the number of connected components in a graph.

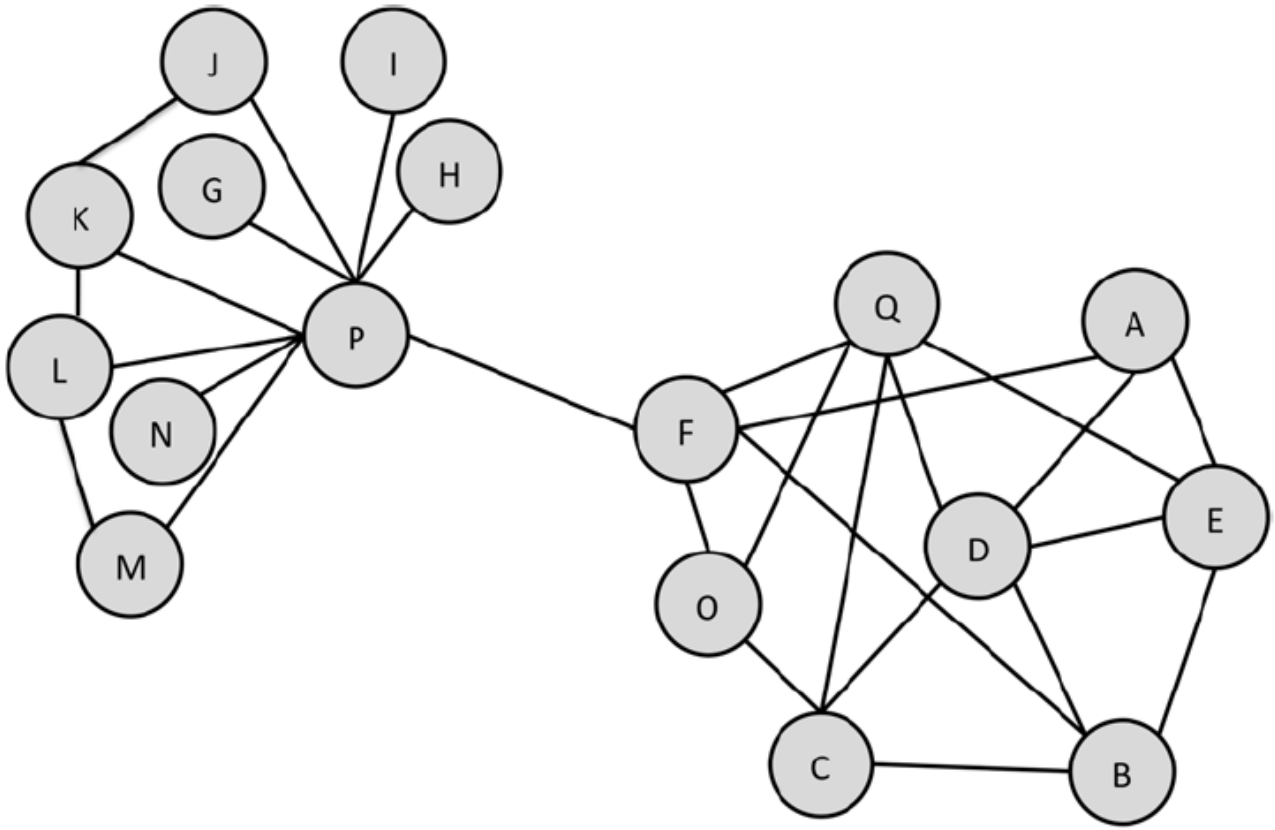
Hubs

The most connected nodes in the network are treated as hubs.

Bridges and Hubs



Bridges and Hubs



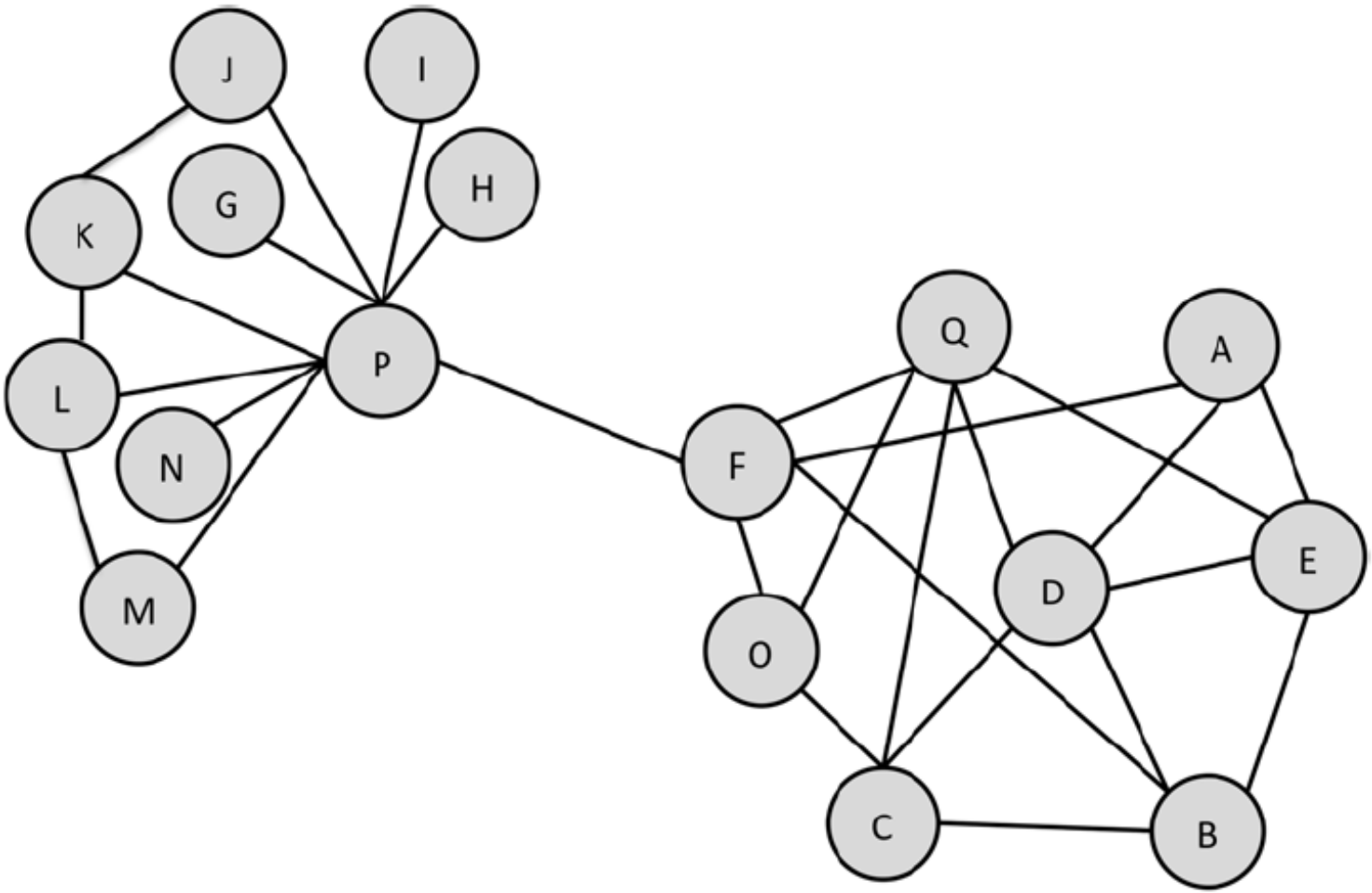
- The edge between nodes P and F is a bridge, because it will give rise to two connected components in the graph
- Node P is a hub because it has many connections to other nodes.

Network properties for nodes and edges

Degree

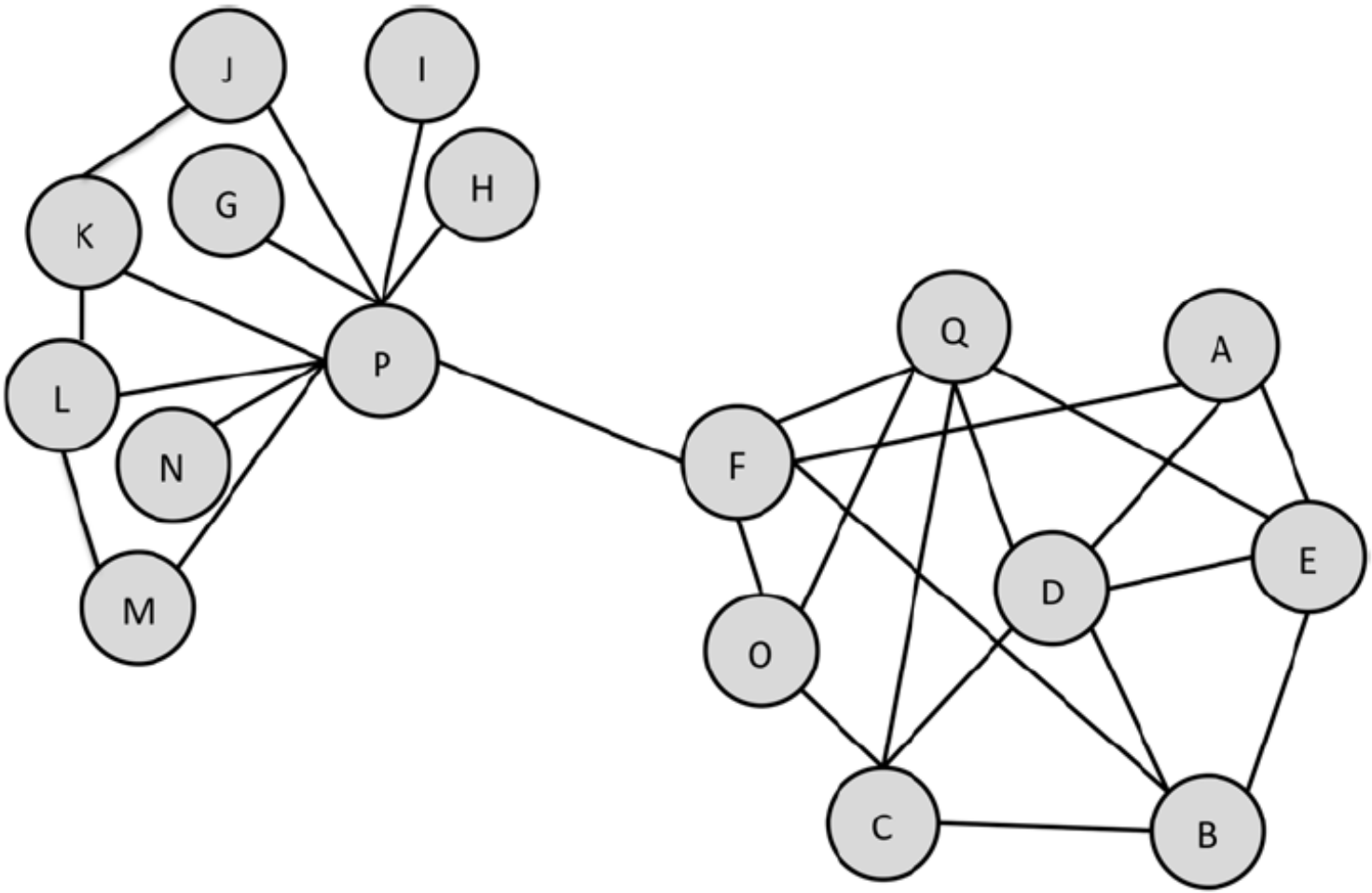
The degree of a node is the number of edges connected to that node. (*undirected*)
Directed graphs: in-degree (number of edges coming into the node) and out-degree (number of edges originating from the node going outward to other nodes)

Degree: Undirected Graph



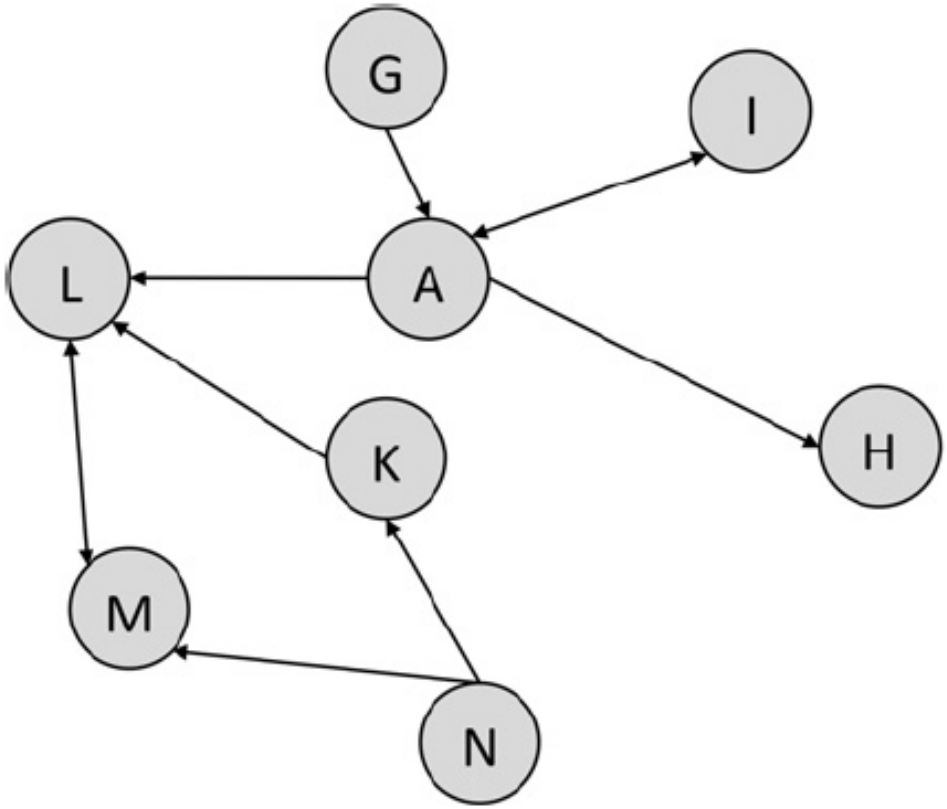
Degrees of nodes P and A?

Degree: Undirected Graph



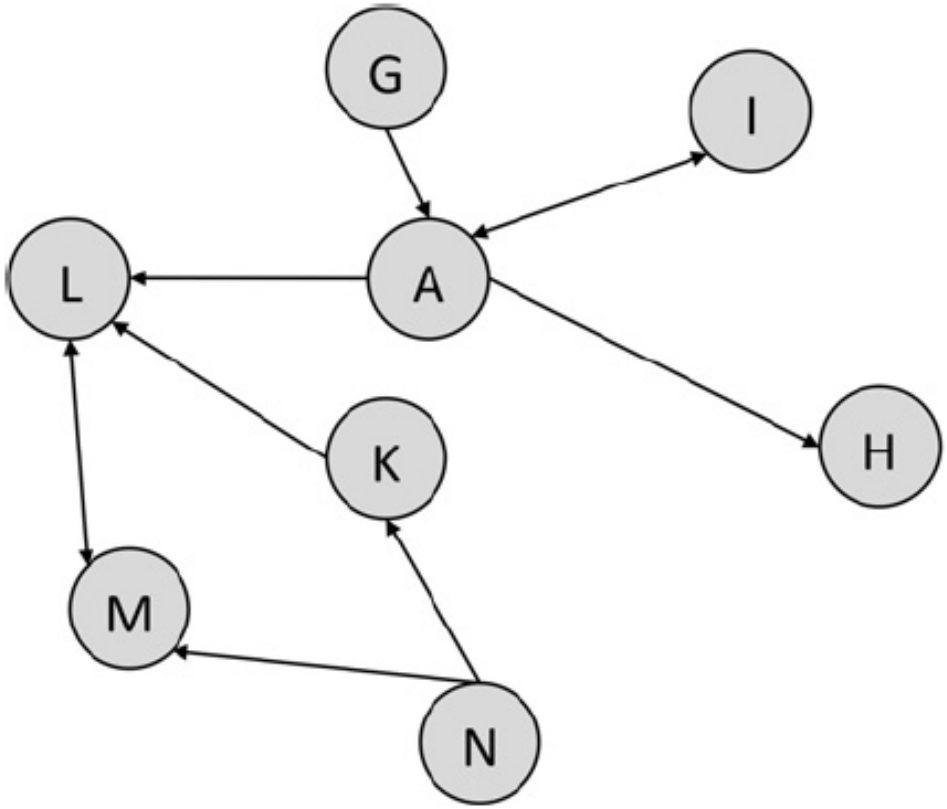
Degrees of nodes P and A?9 and 3

Degree: Directed Graph



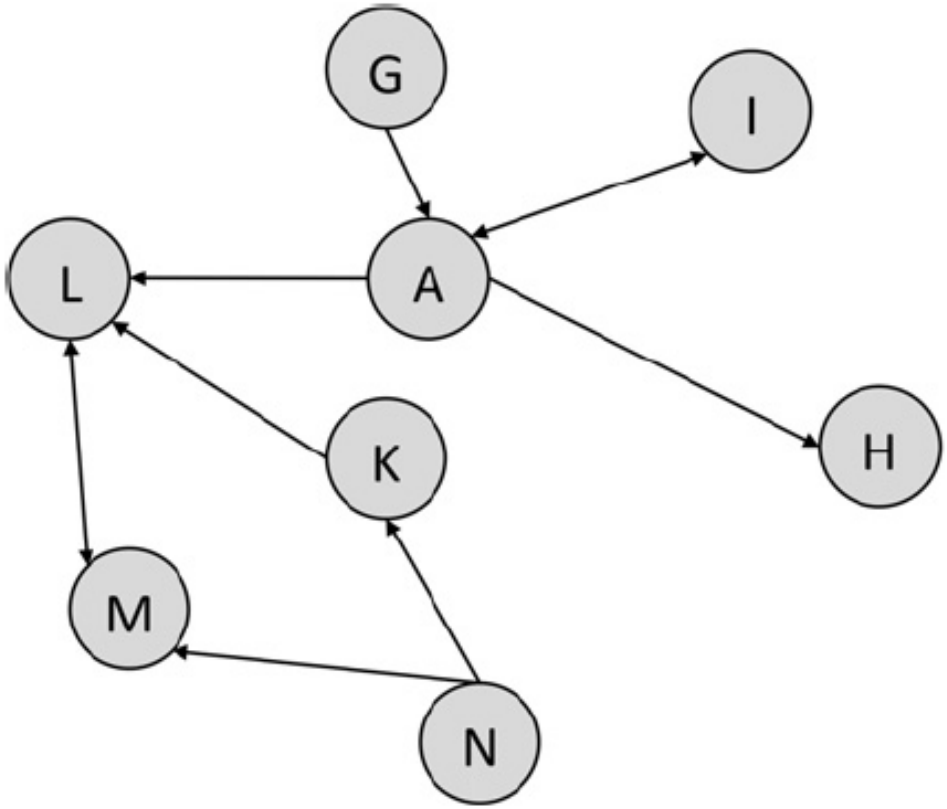
In-degree and out-degree of nodes A:

Degree: Directed Graph



In-degree and out-degree of nodes A: 2 and 3.

Degree: Directed Graph



In-degree and out-degree of nodes A: 2 and 3. Degree of node A is 5.

Centrality

- *Measures how "central" a node is in the network.*

Centrality

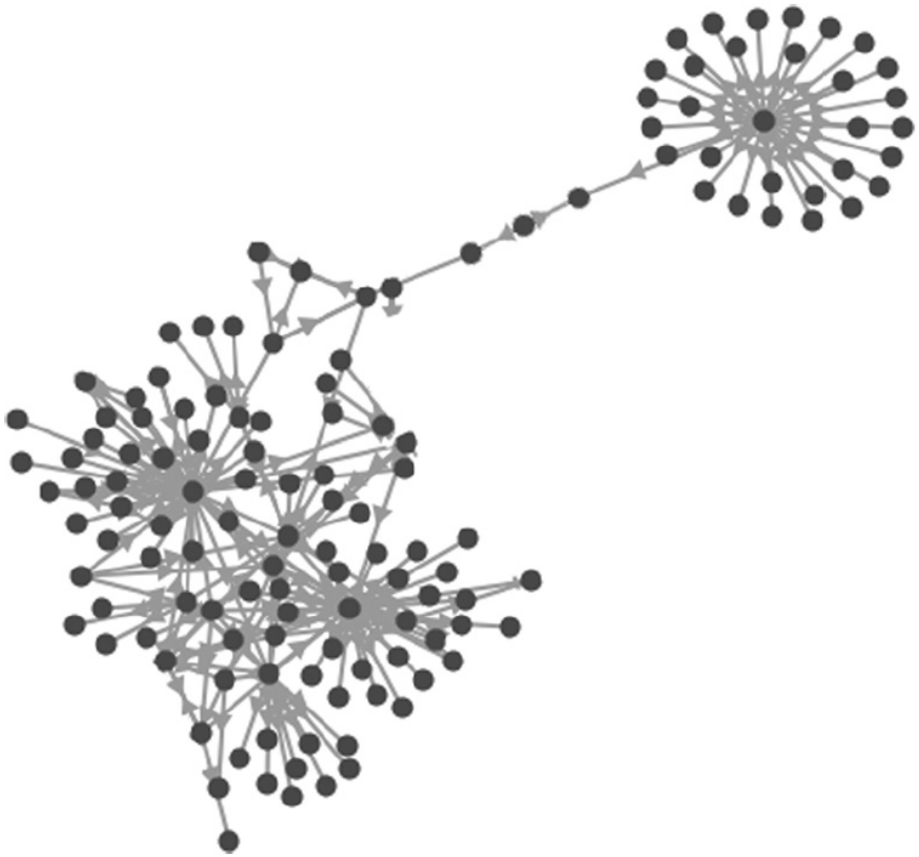
- *Measures how "central" a node is in the network.*
- What counts as "central" may depend on the context.

4 types of centrality (*undirected graphs*)

- Degree centrality
- Closeness centrality
- Betweenness centrality
- Eigenvector centrality

Degree Centrality

- *Degree centrality* of a node is simply its degree - the number of edges it has.

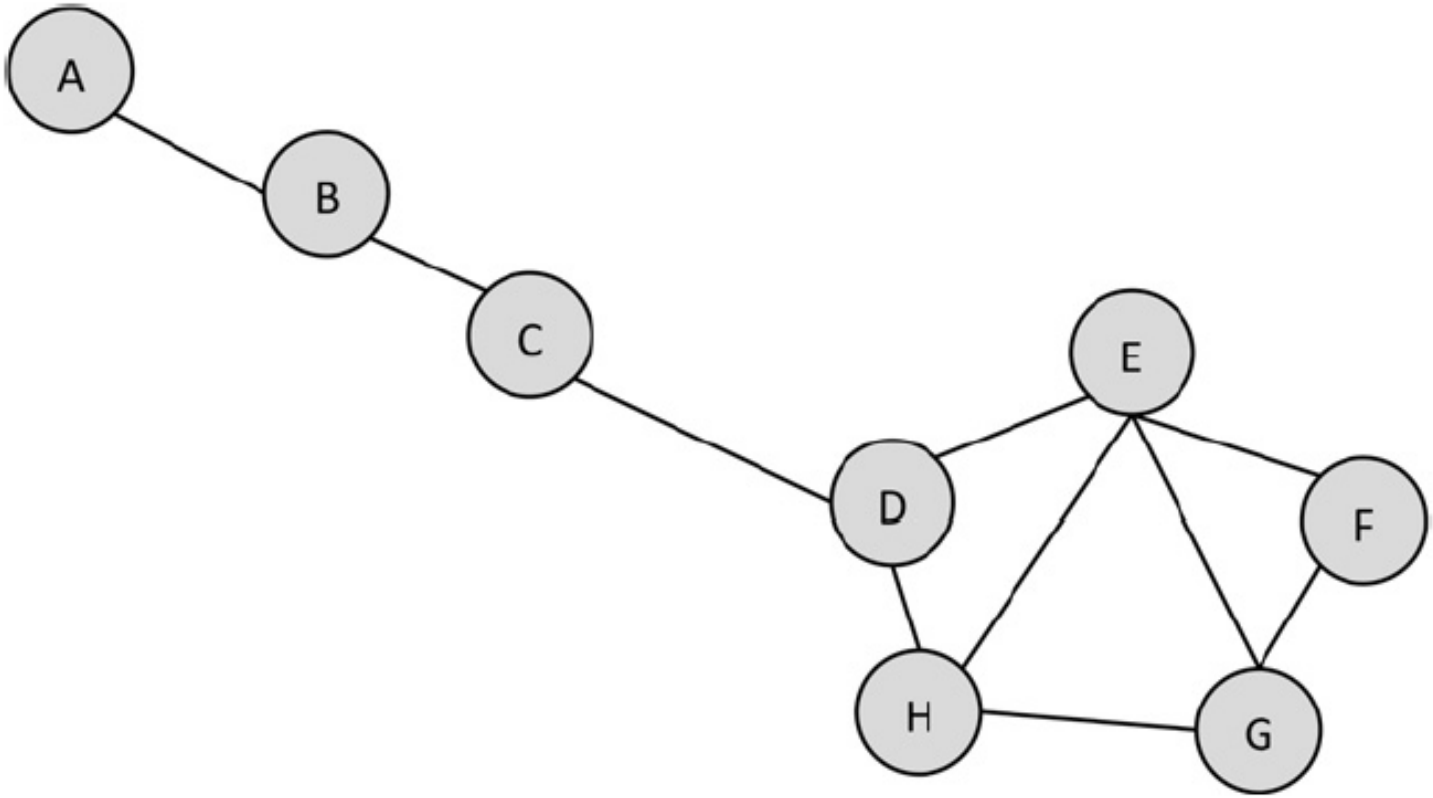


Does not necessarily indicate the importance of a node in connecting others or how central it is to the main group

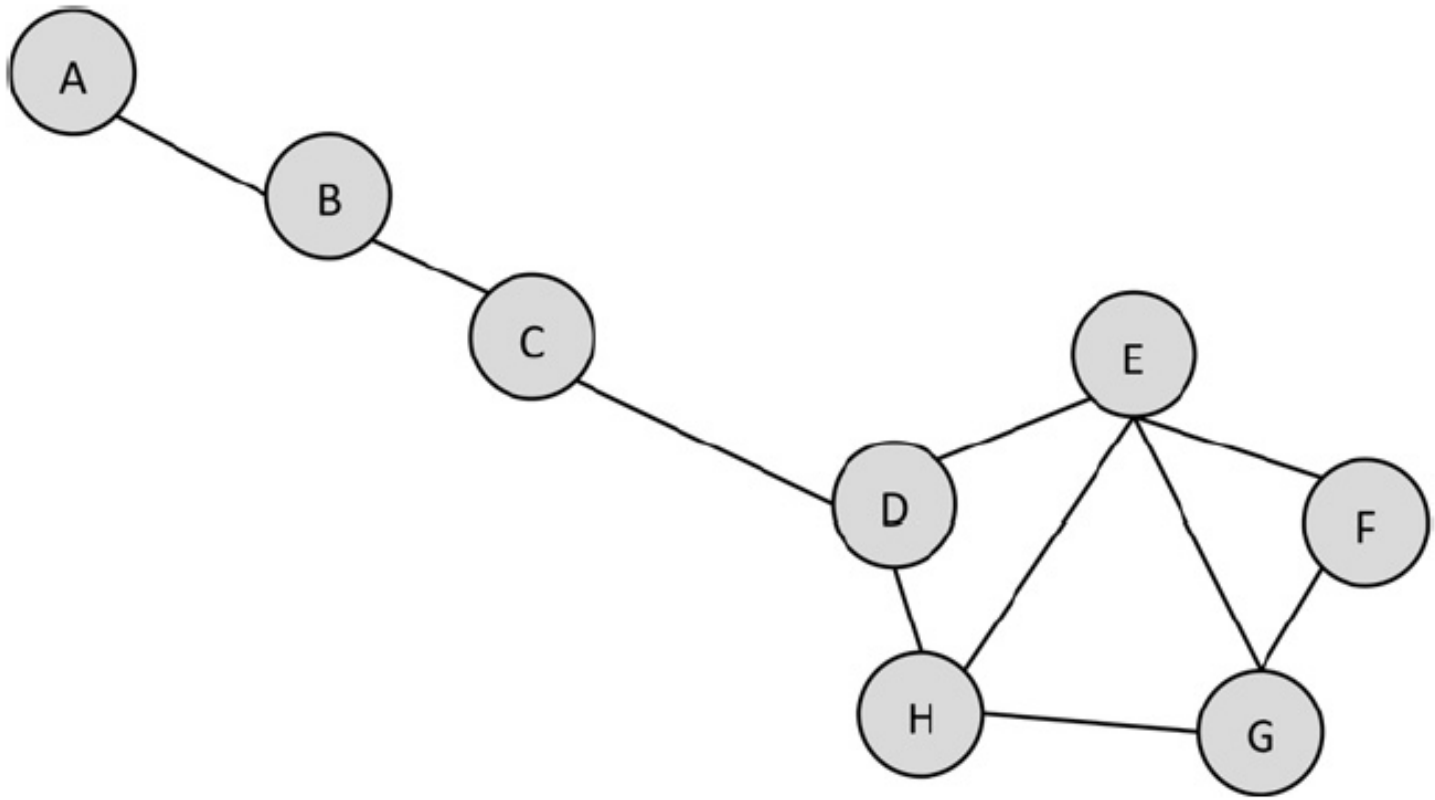
Closeness Centrality

- *Closeness Centrality* indicates how close a node is to all other nodes in the network.
- Calculated as the average of the shortest path length from the node to every other node in the network.

Closeness Centrality

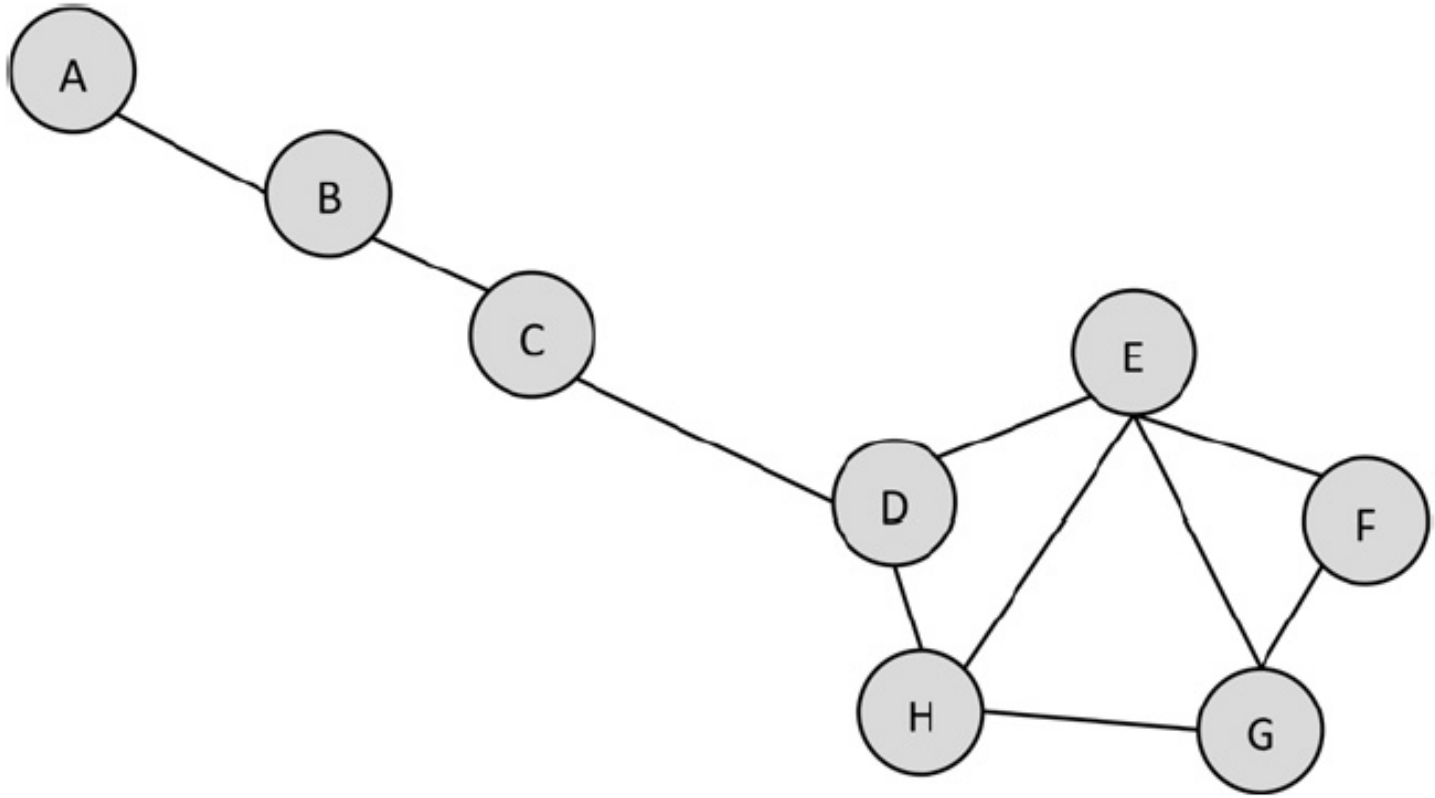


Closeness Centrality



Node D:

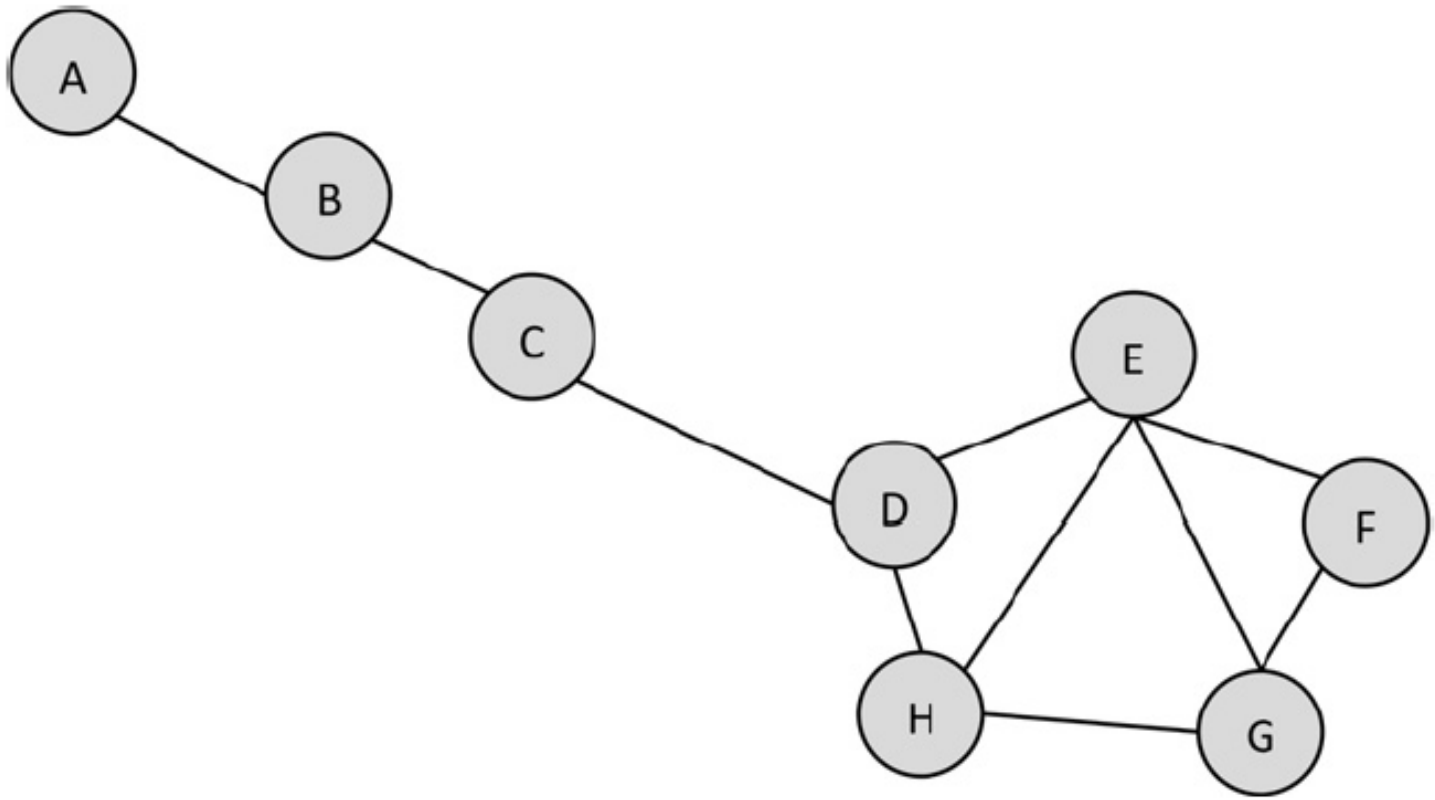
Closeness Centrality



Node D: (3+2+1+1+2+2+1)/7 = 12/7 = 1.71

Node A:

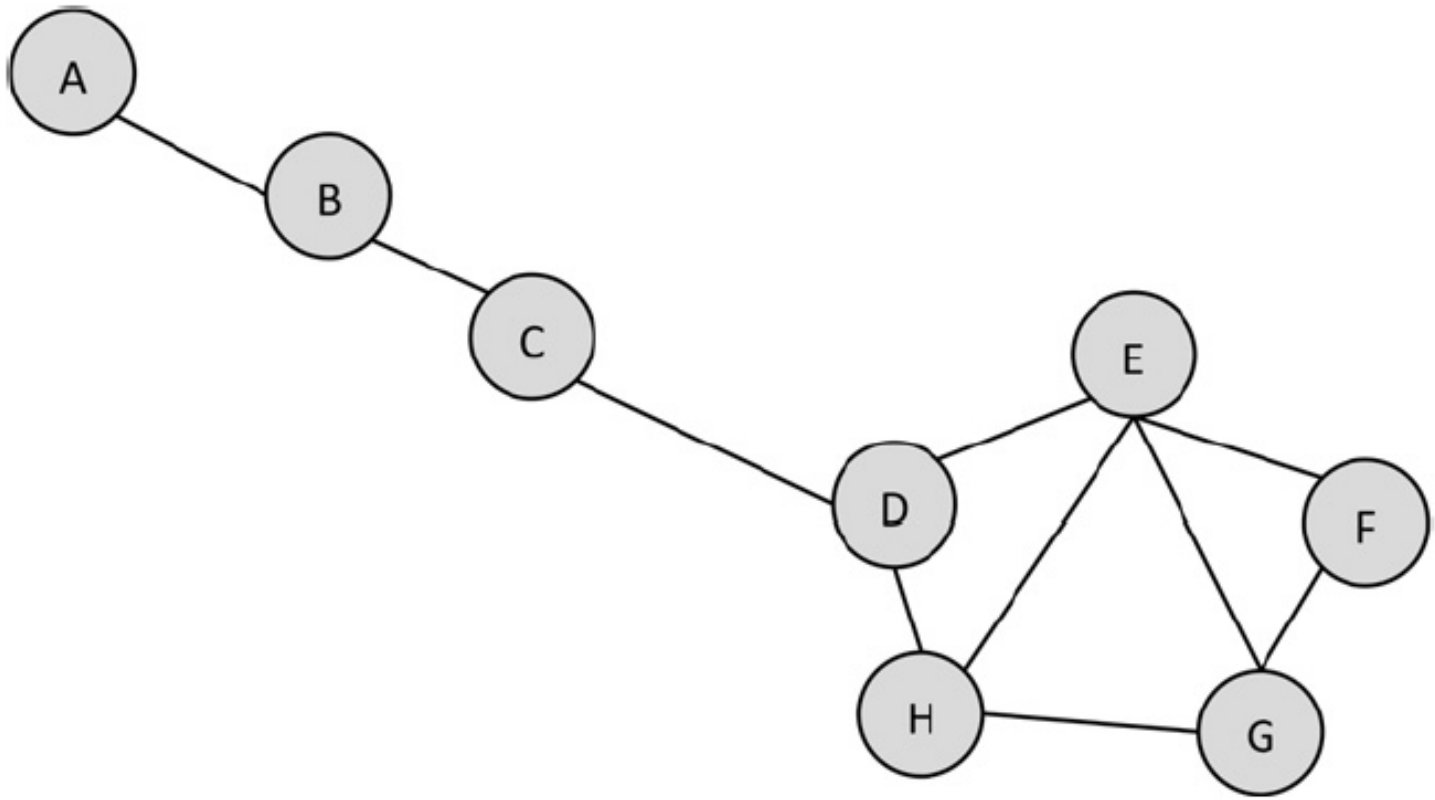
Closeness Centrality



Node D: (3+2+1+1+2+2+1)/7 = 12/7 = 1.71

Node A: 24/7 = 3.43

Closeness Centrality



Node D: (3+2+1+1+2+2+1)/7 = 12/7 = 1.71

Node A: 24/7 = 3.43

Lower values indicate more central nodes.

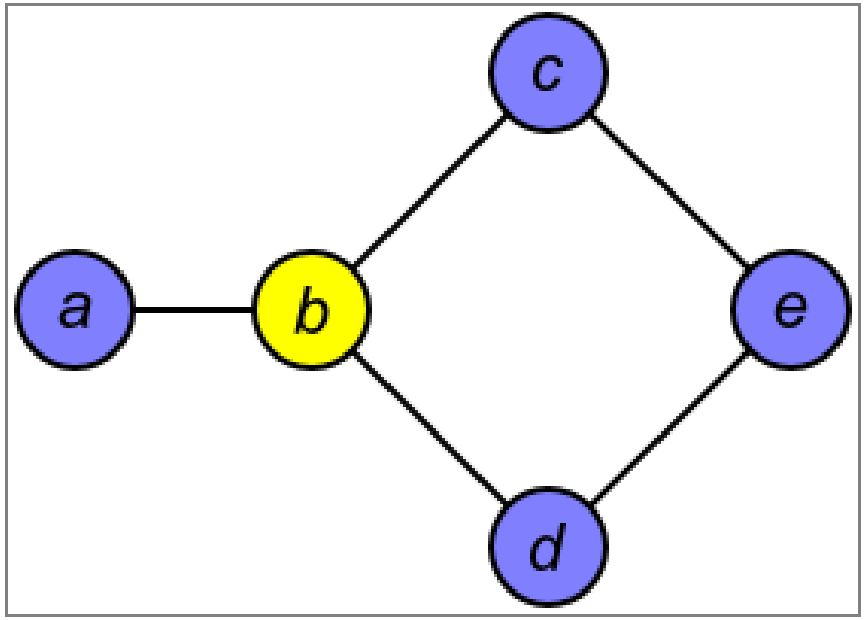
Betweenness centrality

- *Betweenness centrality* measures how important a node is to the shortest paths through the network.

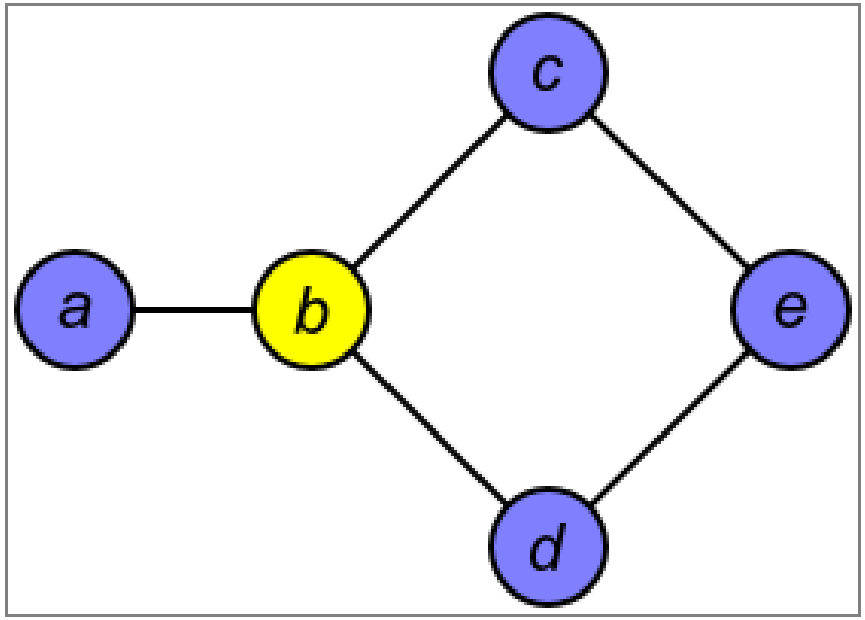
Computing betweenness for a node N

- Select a pair of nodes and find all the shortest paths between the nodes.
- Compute the fraction of those shortest paths that include node N.
- If 5 shortest paths and 3 went through N, fraction: 3/5 = 0.6
- Repeat this for every pair of nodes
- Add up the fractions computed $\rightarrow$ *Betweenness centrality*
- Betweenness *may be normalized* by dividing through the number of pairs

Betweenness centrality

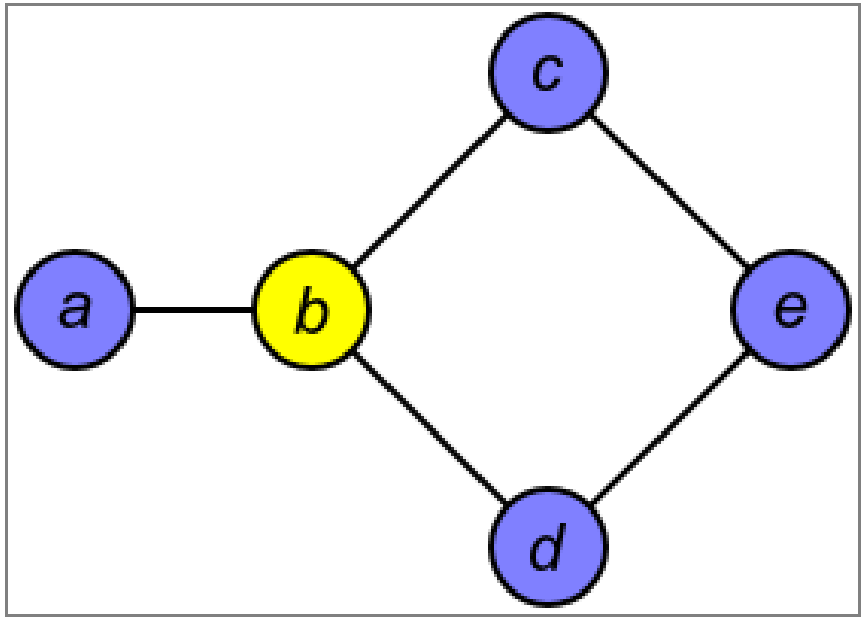


Betweenness centrality



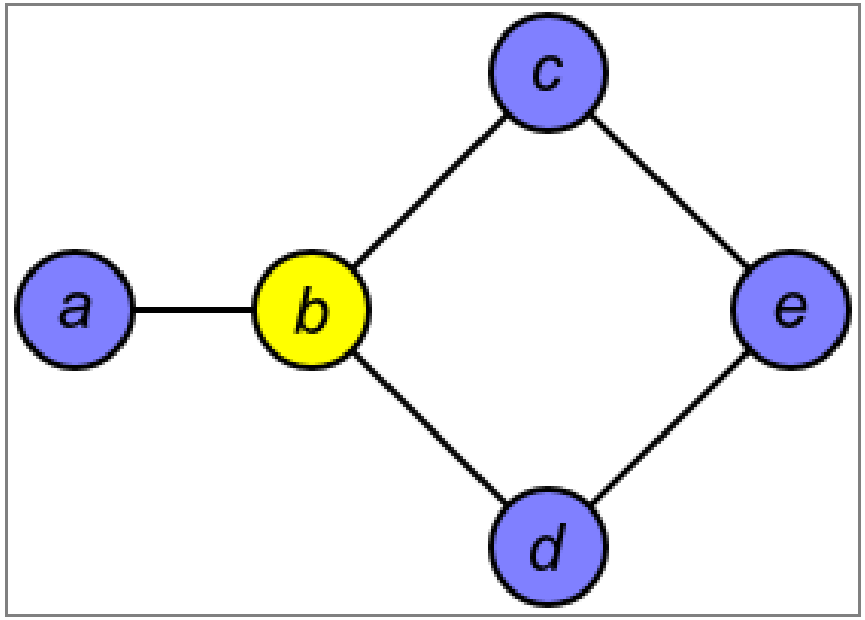
Node b:

Betweenness centrality



Node b: 6 pairs: ac, ad, ae, cd, ce, de

Betweenness centrality



Node b: 6 pairs: ac, ad, ae, cd, ce, de

((1/1) + (1/1) + (2/2) + (1/2) +) + 0) = 3.5/6

Betweenness centrality

- *Captures how important a node is in the flow of information from one part of the network to another.*
- *Is one of the most frequently used centrality measures*.

Betweenness centrality

- *Captures how important a node is in the flow of information from one part of the network to another.*
- *Is one of the most frequently used centrality measures*.

Directed Networks: Several meanings

- A user with high betweenness may be followed by many others who don't follow the same people as the user. (*well-followed*)
- The user may have fewer followers but connect them t many accounts that are otherwise distant. (*reader of many people*)

Eigenvector Centrality

- *Eigenvector Centrality* measures a node's importance while giving consideration to the importance of its neighbors.
- **Ex:** A node with 300 relatively unpopular friends would have lower eigenvector centrality than someone with 300 very popular friends.

Eigenvector Centrality

- *Eigenvector Centrality* measures a node's importance while giving consideration to the importance of its neighbors.
- **Ex:** A node with 300 relatively unpopular friends would have lower eigenvector centrality than someone with 300 very popular friends.
- Sometimes used to measure a node's influence in the network.

Eigenvector Centrality

- PageRank algorithm uses a variant of eigenvector centrality.

Eigenvector Centrality

- PageRank algorithm uses a variant of eigenvector centrality.
- Main principle: links from important nodes are worth more than links from unimportant nodes.

Eigenvector Centrality

- PageRank algorithm uses a variant of eigenvector centrality.
- Main principle: links from important nodes are worth more than links from unimportant nodes.
- All nodes start off equal, but as the computation progresses, nodes with more edges start gaining importance

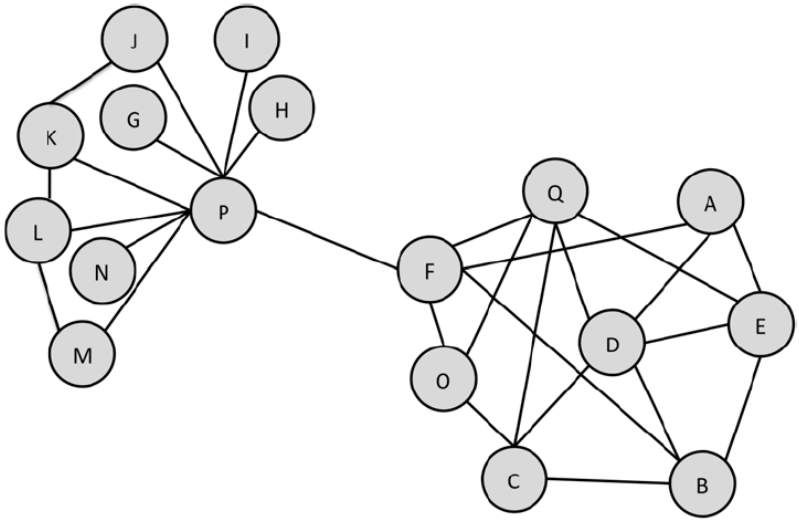
Describing Networks

- *A number of measures can be used to describe the structure of a network as a whole.*
- **Ex:** Density: the number of edges in the graph divided by the number of possible edges, is one of the most common ways

Degree distribution

- *Degree distribution* gives an idea of the degree for all the nodes in the network.
- This shows how many nodes have each possible degree.
- How to compute?

Step 1: Calculate the degree for each node



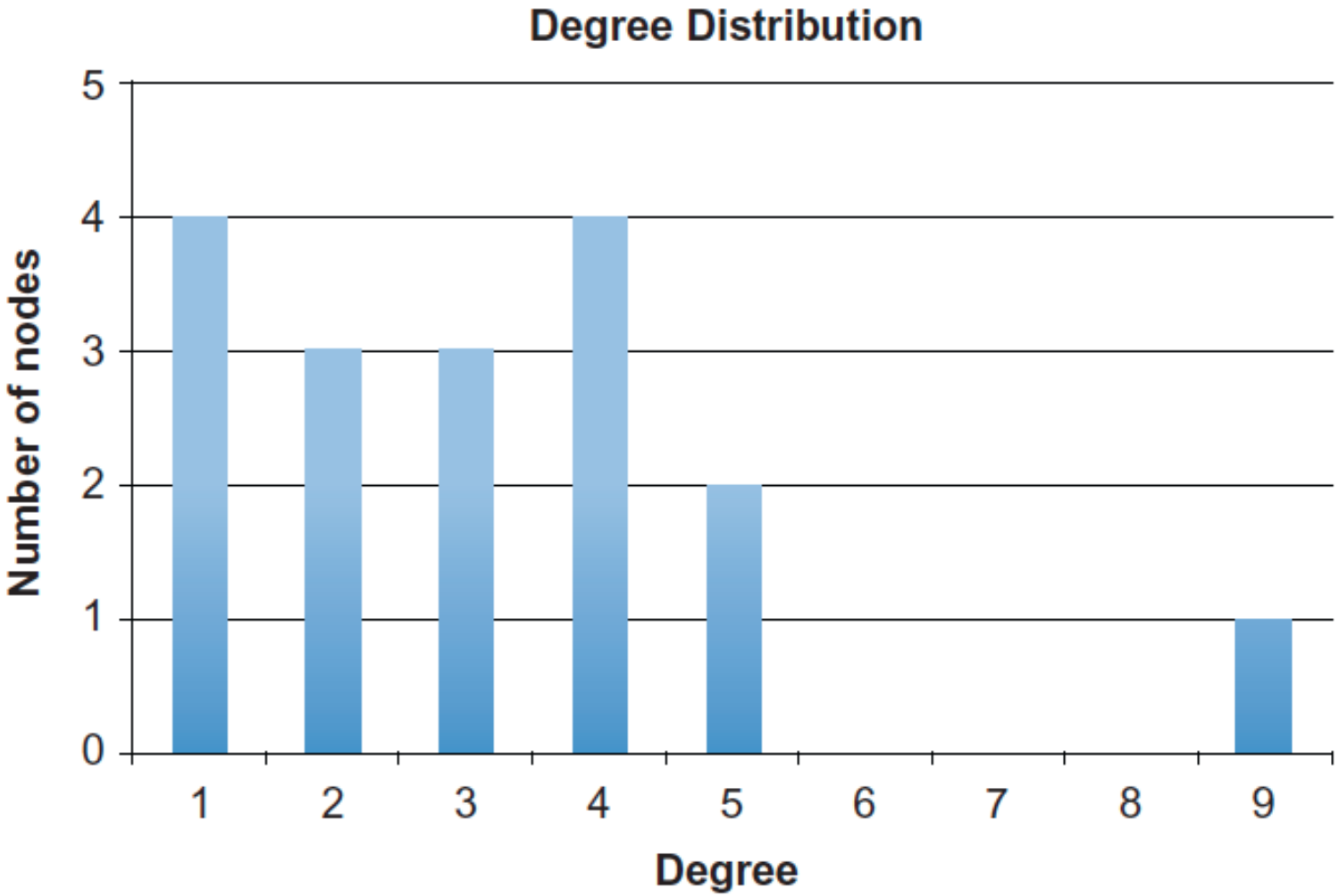
Node	Degree
A	3
B	4
C	4
D	5
E	4
F	4
G	1
H	1
I	1
J	2
K	3
L	3
M	2
N	1
O	2
P	9
Q	5

Step 2: Count how many nodes have each degree

Node	Degree
A	3
B	4
C	4
D	5
E	4
F	4
G	1
H	1
I	1
J	2
K	3
L	3
M	2
N	1
O	2
P	9
Q	5

Degree	Number of Nodes
1	4
2	3
3	3
4	4
5	2
6	0
7	0
8	0
9	1

Visualization as a bar graph

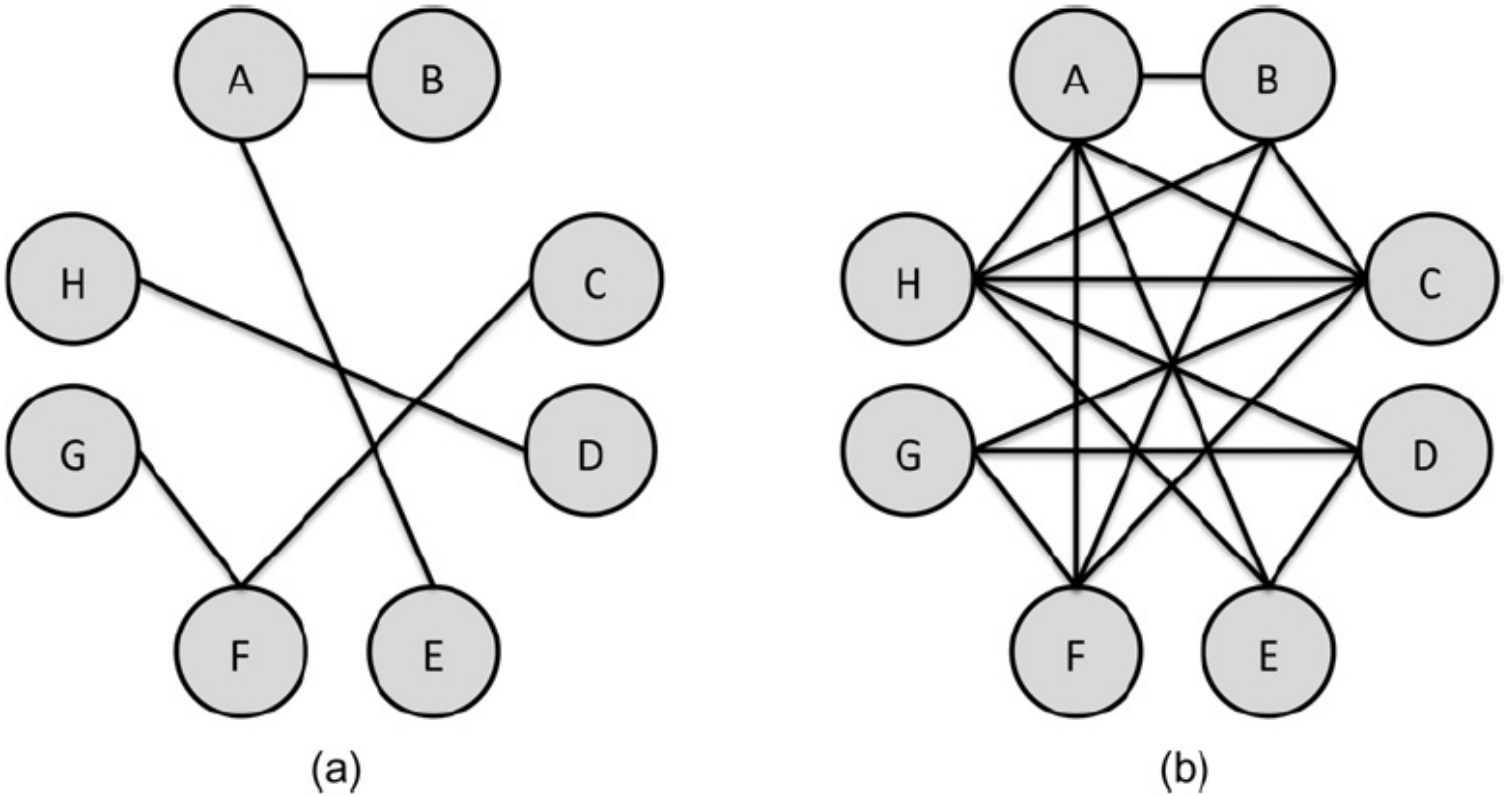


Density

- *Density describes how connected a network is.*

Density

- *Density describes how connected a network is.*



Network (b) has higher density

Calculating Density

Formula

$$density = \frac{\text{number of edges}}{\text{number of possible edges}}$$

Calculating Density

Formula

$$density = \frac{\text{number of edges}}{\text{number of possible edges}}$$

Number of possible edges

Graph with n nodes:

Calculating Density

Formula

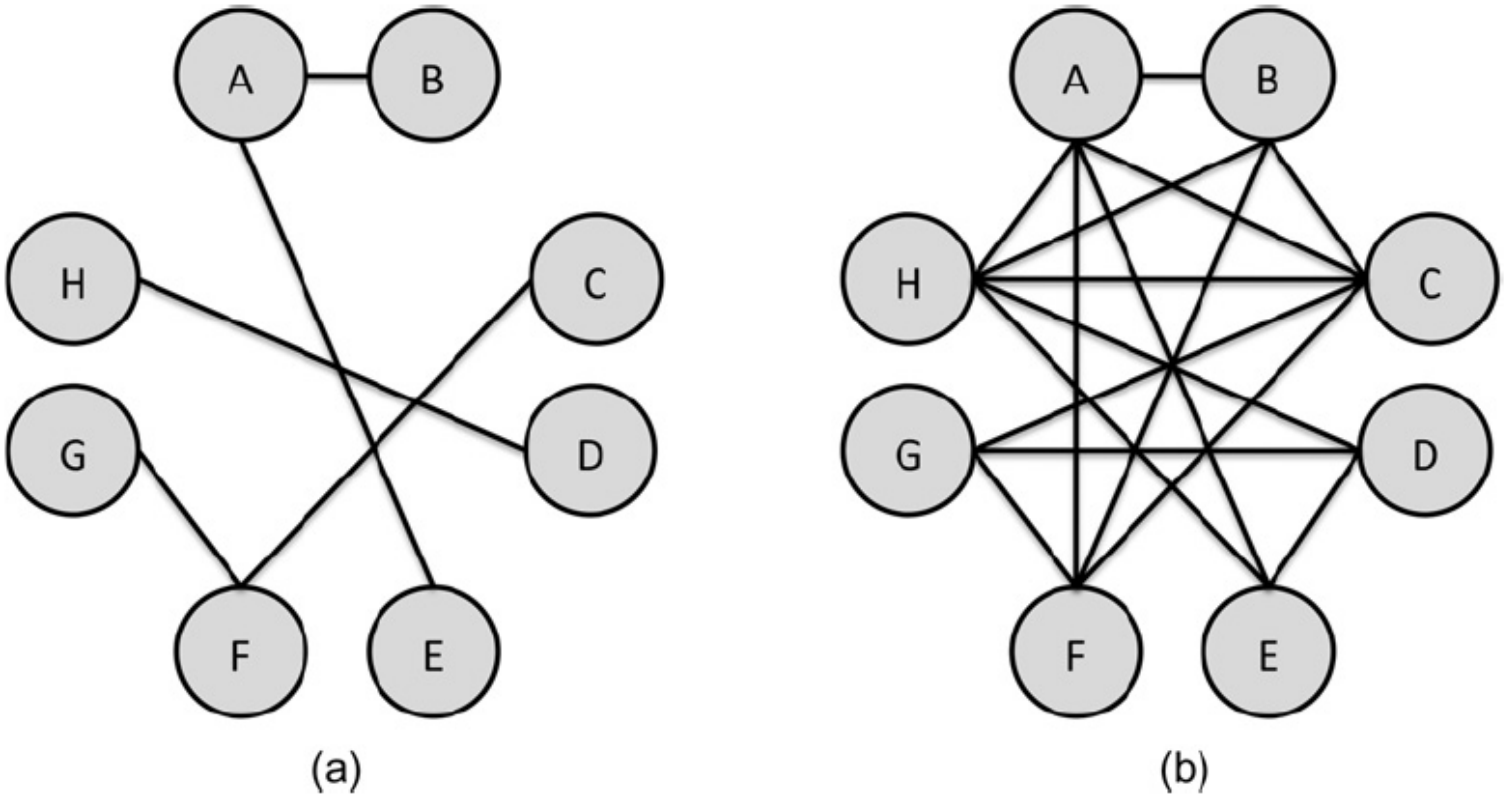
$$density = \frac{\text{number of edges}}{\text{number of possible edges}}$$

Number of possible edges

Graph with n nodes:

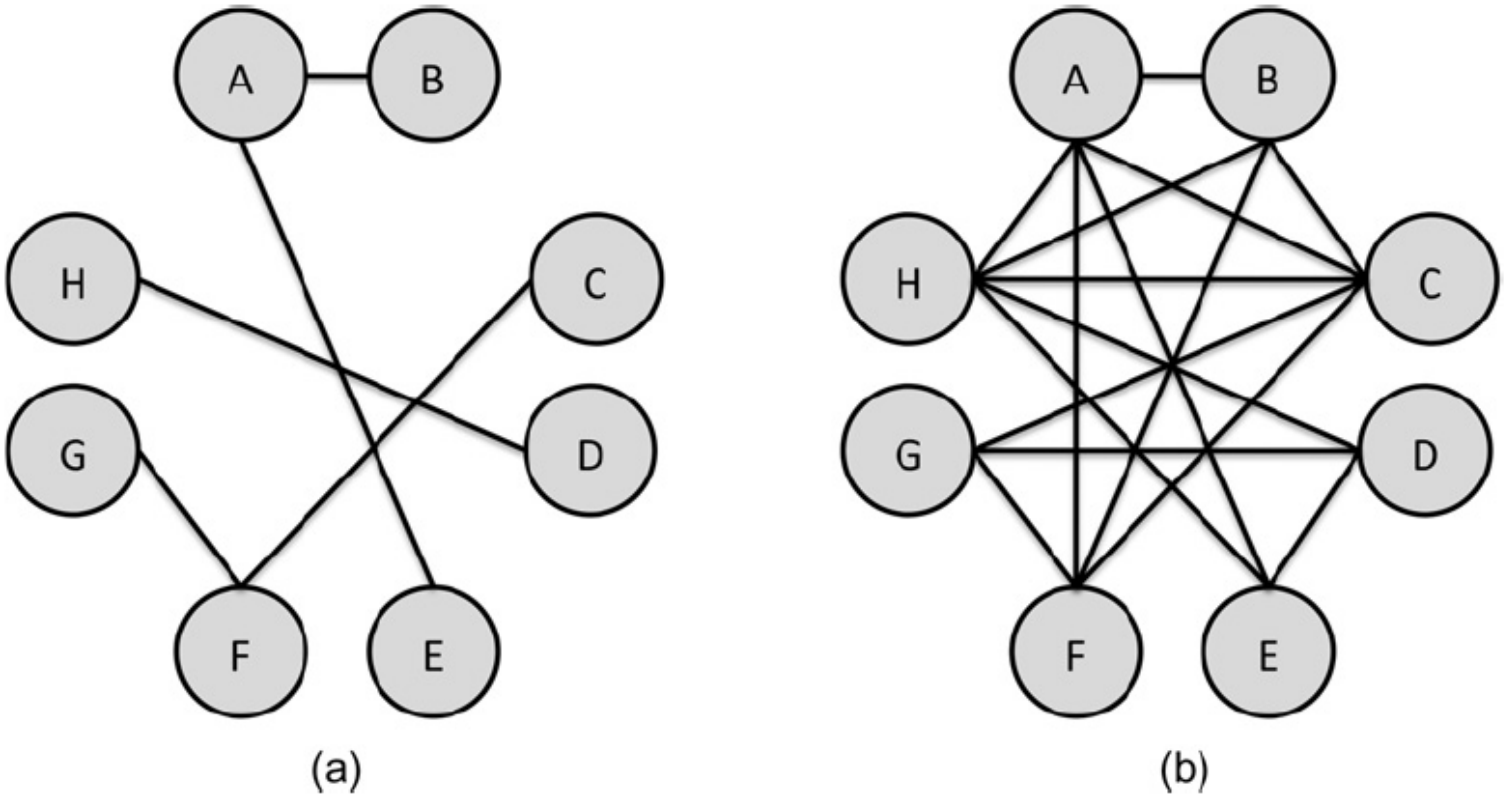
- **Undirected networks:** $\frac{n*(n-1)}{2}$
- **Directed networks:** $n \times (n-1)$

Calculating Density



Densities for Networks (a) and (b)

Calculating Density



Densities for Networks (a) and (b)

- Network (a): 5/28
- Network (b): 16/28

Density

Densest possible network?

Network where all possible edges exist - *a clique*

Density

Densest possible network?

Network where all possible edges exist - *a clique*

Density in Egocentric Networks

- Density is even more commonly used to compare *subnetworks* - especially egocentric networks
- 1.5-degree network is used: node's connection and all the connection between those (node excluded)

Density

Densest possible network?

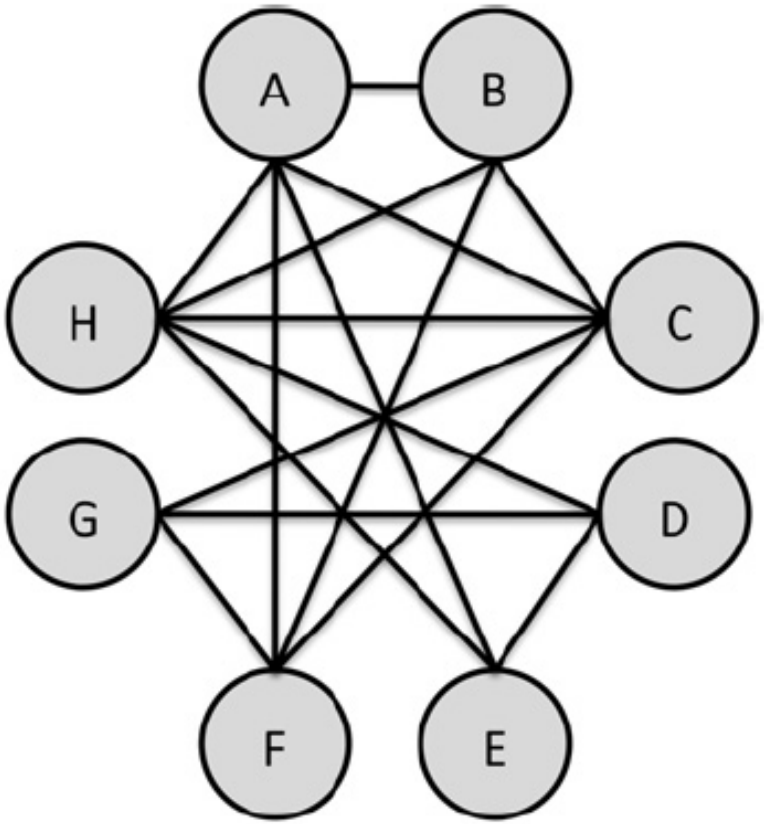
Network where all possible edges exist - *a clique*

Density in Egocentric Networks

- Density is even more commonly used to compare *subnetworks* - especially egocentric networks
- 1.5-degree network is used: node's connection and all the connection between those (node excluded)
- Dense egocentric networks: a lot of their friends know each other
- Sparse egocentric networks: their connections often do not know one another
- Referred to as *local clustering coefficient*.

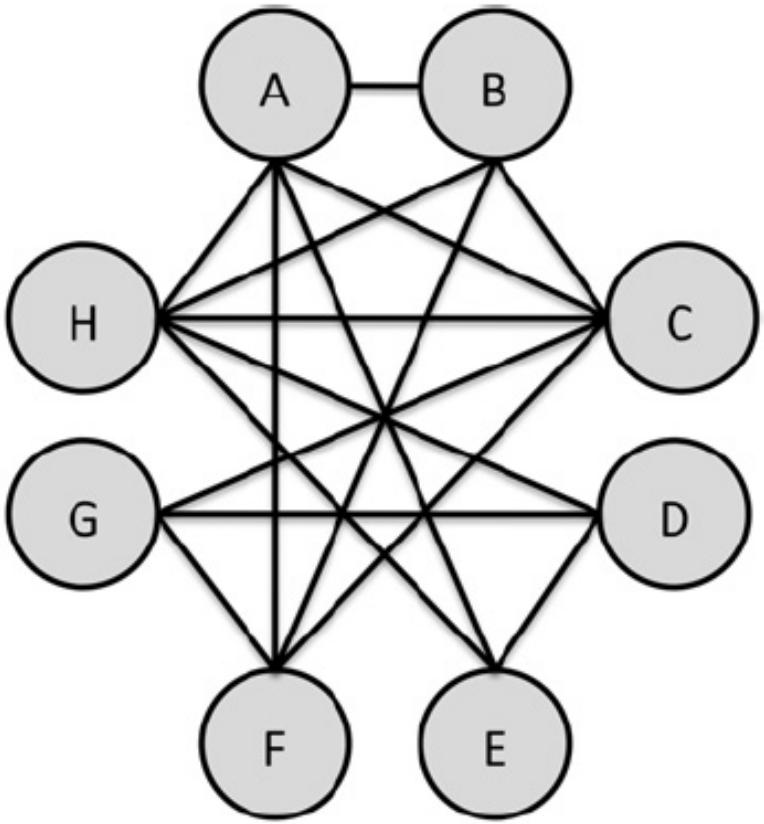
Density in Egocentric Networks

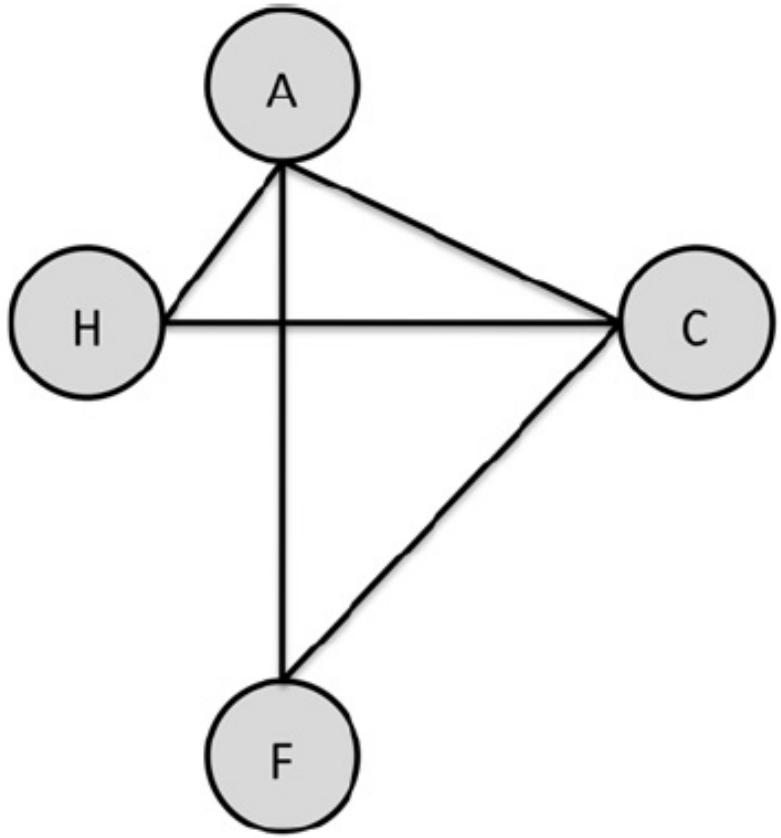
1.5-degree egocentric network for Node B:



Density in Egocentric Networks

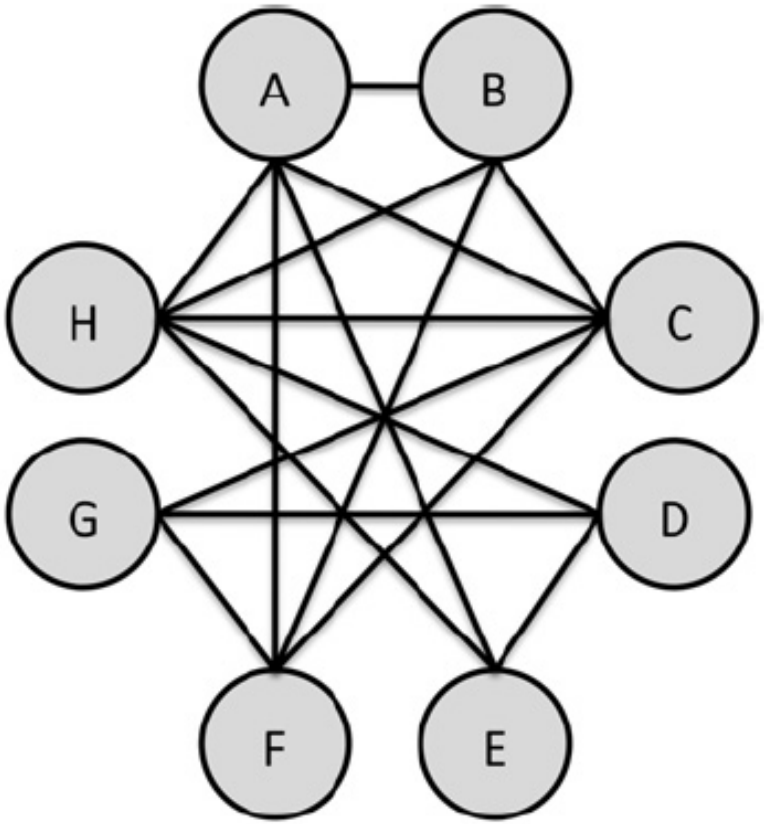
1.5-degree egocentric network for Node B:

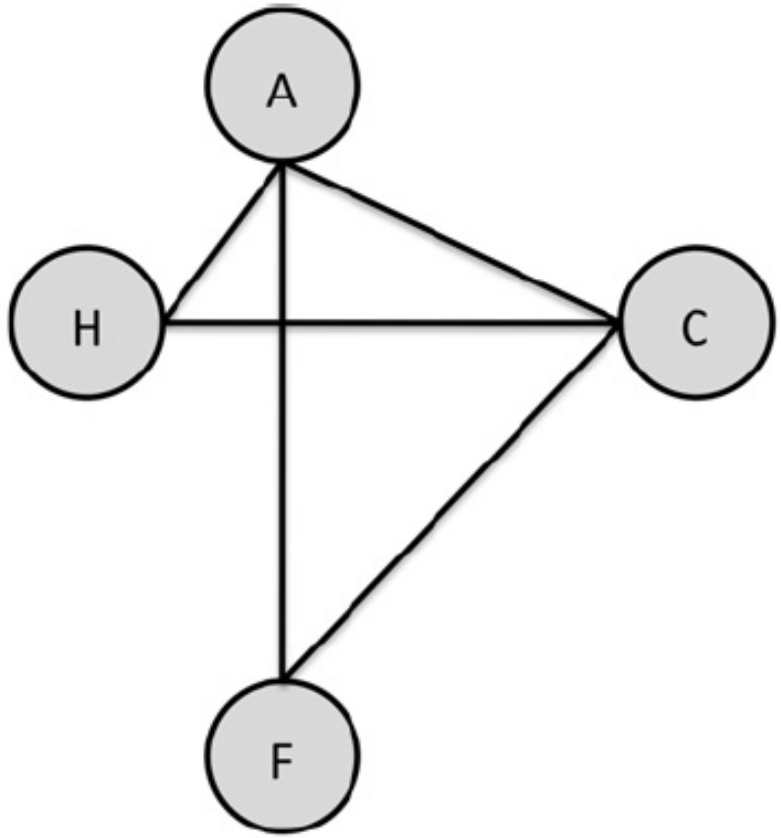




Density in Egocentric Networks

1.5-degree egocentric network for Node B:





Density: 5/6

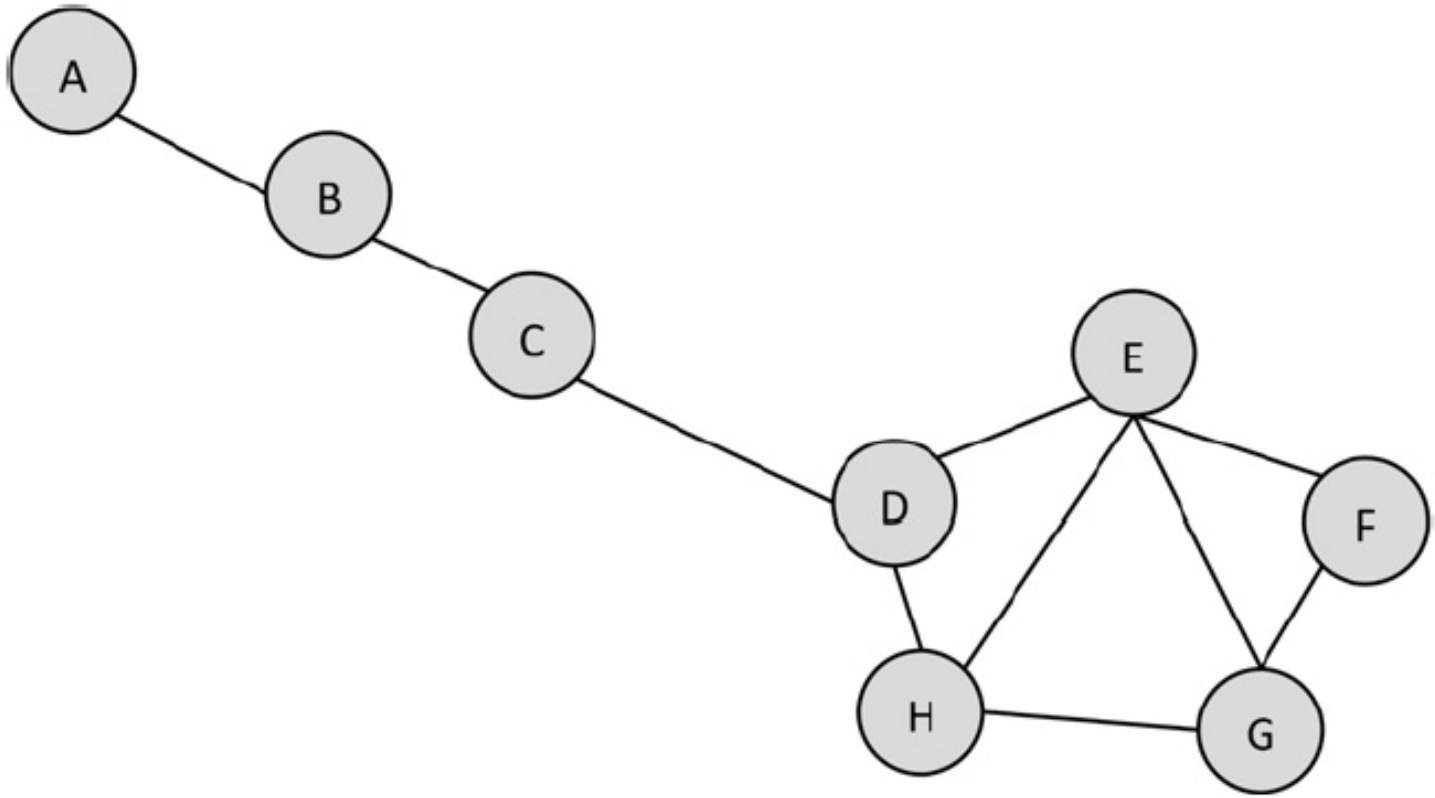
Density in Egocentric Networks

- Having a high egocentric density does not necessarily mean a node is more “popular” or important
- A node with high degree will usually have a lower density

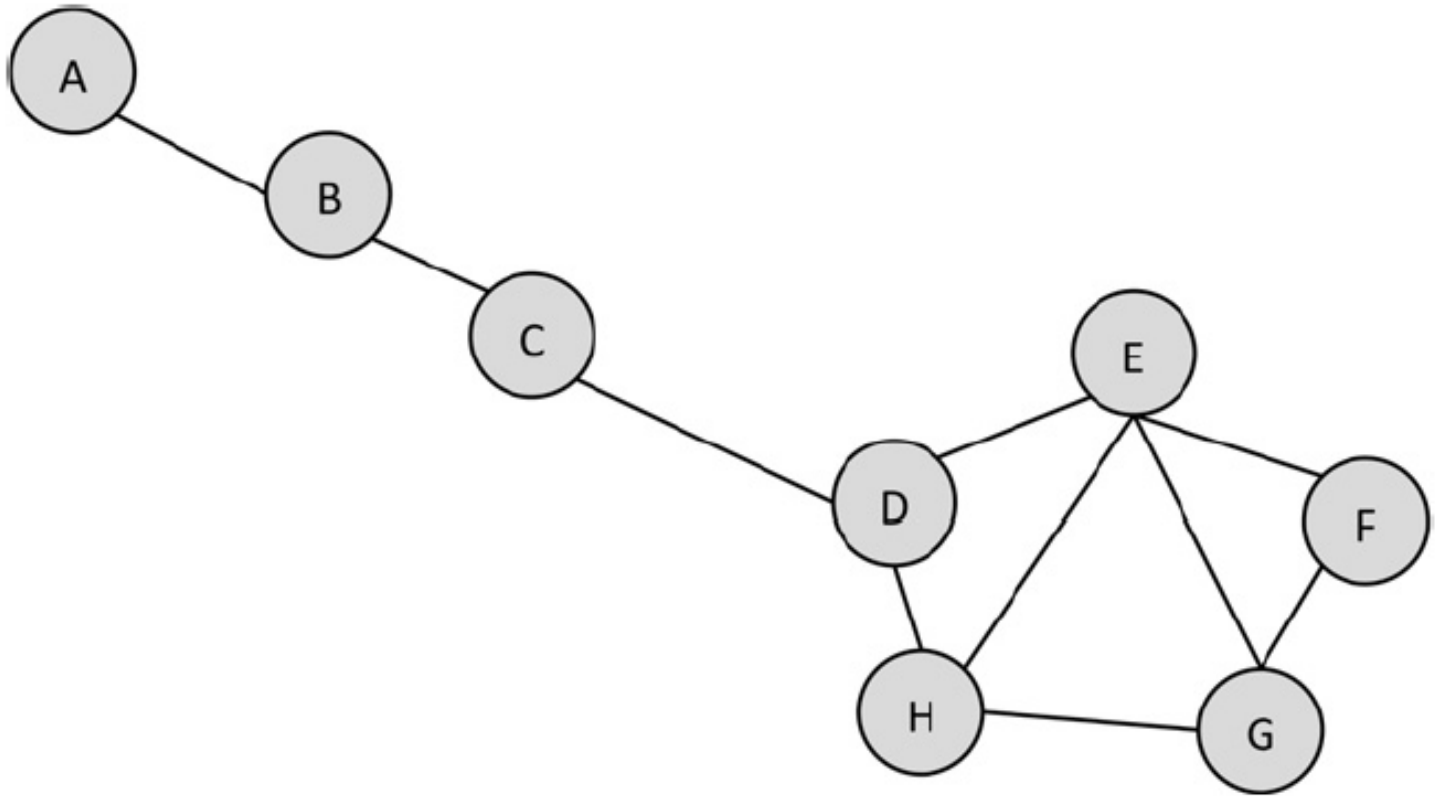
Connectivity

- Also known as *cohesion*, measures how the edges are distributed in the graph
- A count of the minimum number of nodes that would have to be removed before the graph becomes disconnected
- There is no path from each node to every other node

Connectivity Examples

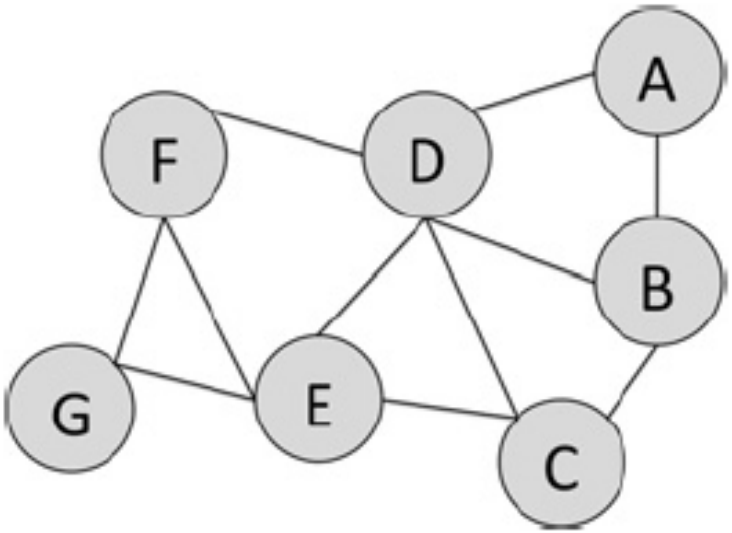


Connectivity Examples

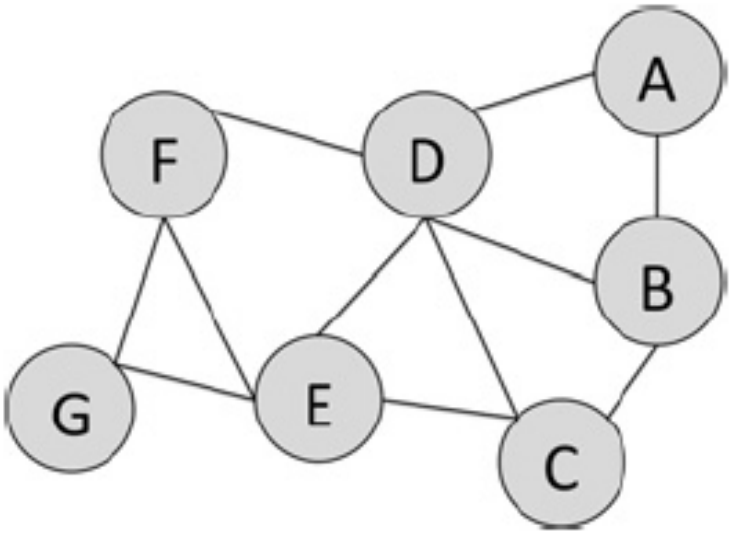


Removing nodes B, C, or D would disconnect the graph, so connectivity is 1.

Connectivity Examples



Connectivity Examples



Removing nodes E and F would separate G from the rest of the graph

Centralization

- *Centralization* uses the distribution of a centrality measure to understand the network as a whole.
- If one node has extremely high centrality while most other nodes have low centrality, the centralization of the network is high
- If centrality is more evenly distributed, the centralization of the network is low

Centralization

- *Centralization* uses the distribution of a centrality measure to understand the network as a whole.
- If one node has extremely high centrality while most other nodes have low centrality, the centralization of the network is high
- If centrality is more evenly distributed, the centralization of the network is low

Centralization of power?

- Betweenness centrality: control one node has in the ability to communicate
- Few nodes with high betweenness: power centralized in those nodes

Computing Centralization

Basic Idea

- Looking at the sum of differences in centrality between the most central node and every other node in the network
- Dividing this by the maximum possible differences in centrality

Computation

- Let $C(n)$ be the centrality of node n
- Let n^* be the most central node
- Sum of differences: $\sum C(n^*) - C(n_i)$
- Maximum possible differences: max $\sum C(n^*) - C(n_i)$

Computing Centralization

Maximum possible differences

- Maximum possible difference is taken over all possible graphs with the same number of nodes.
- **Ex: Beweenness centrality-** Maximum difference would be achieved when any two nodes are connected via the central node (shortest path of length 2)
- $C(n_i) = 0$ for all other nodes, $C(n^*)$: For each of the $(n-1)$ nodes, $(n-2)$ edges go through the node in question: $(n-1) \times (n-2)$ (for directed, divide by 2 for undirected)

Small worlds

Six degrees of separation

- Core idea: Any two people in the world are separated by short paths, on average about six steps
- "small world": people who may be very far apart physically and socially are still connected with relatively small paths

Experiment by Stanley Milgram in 1967

- Sent information packets to people who lived in Omaha, Nebraska and Wichita, Kansas
- Recipients were asked to get the packet to *a specific person* in Boston
- If they knew the contact, they were supposed to send the packet directly to him
- If not, they were supposed to think of someone they did know, who was likely to be closer to the person in Boston, sign their name to a roster, and send the packet to their friend
- Boston contact could examine the roster and see how many steps it took for the letter to arrive
- 64 letters arrived, average number of links: between 5 and 6

Small worlds

Remarkable about the experimental findings

Compared to the number of people in the US, the average shortest path between any two is remarkably short

Small worlds

Remarkable about the experimental findings

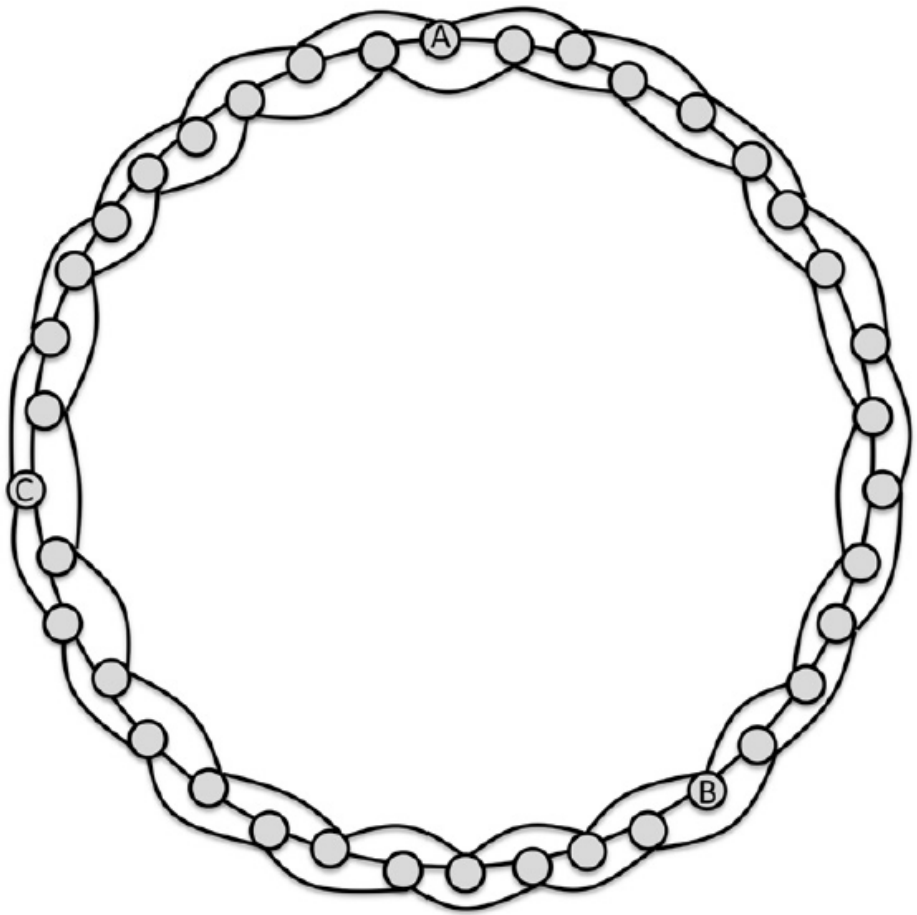
Compared to the number of people in the US, the average shortest path between any two is remarkably short

Small worlds: two primary characteristics

- A short average *shortest path length*
- High clustering (local clustering coefficient)

Example: 36 nodes and 72 edges

- *Regular network:* each node is connected to a fixed number of neighbors on either side



Example: 36 nodes and 72 edges

Regular network:

- Using the edges that move two steps around the ring, nearly 1/4 of the nodes are touched before reaching B

Example: 36 nodes and 72 edges

Regular network:

- Using the edges that move two steps around the ring, nearly 1/4 of the nodes are touched before reaching B
- Expanding the graph to 1000 nodes, path length would be 250

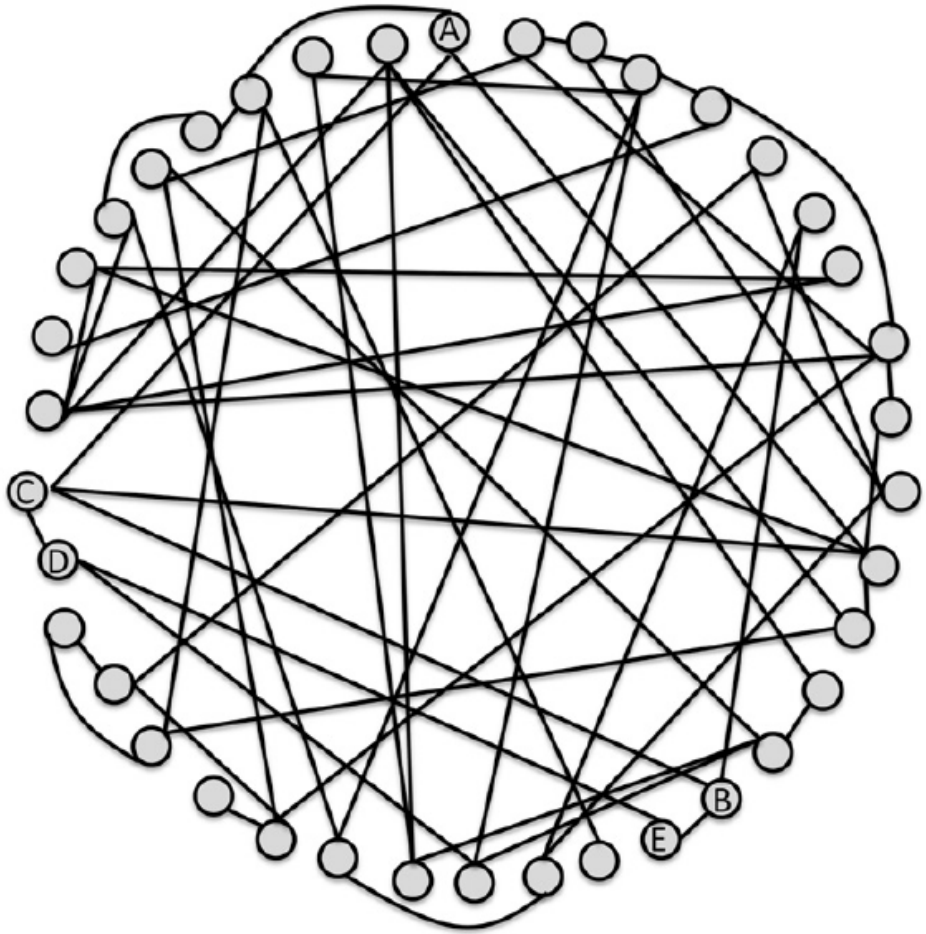
Example: 36 nodes and 72 edges

Regular network:

- Using the edges that move two steps around the ring, nearly 1/4 of the nodes are touched before reaching B
- Expanding the graph to 1000 nodes, path length would be 250
- For a graph with a million nodes, it would be 0.25 million

Example: 36 nodes and 72 edges

- *Random graph:* the edges randomly connect the nodes



Example: 36 nodes and 72 edges

Regular network:

- The shortest path from A to B is much shorter (A to C to B)
- Random edges jump from one side of the network to the other, also connect nearby edges
- Increasing the number of nodes to 1 million (proportional increase in edges), average shortest path length would increase
- But the rate would not be the same as in regular graph

Have the property of a short path length, even when the networks become huge.

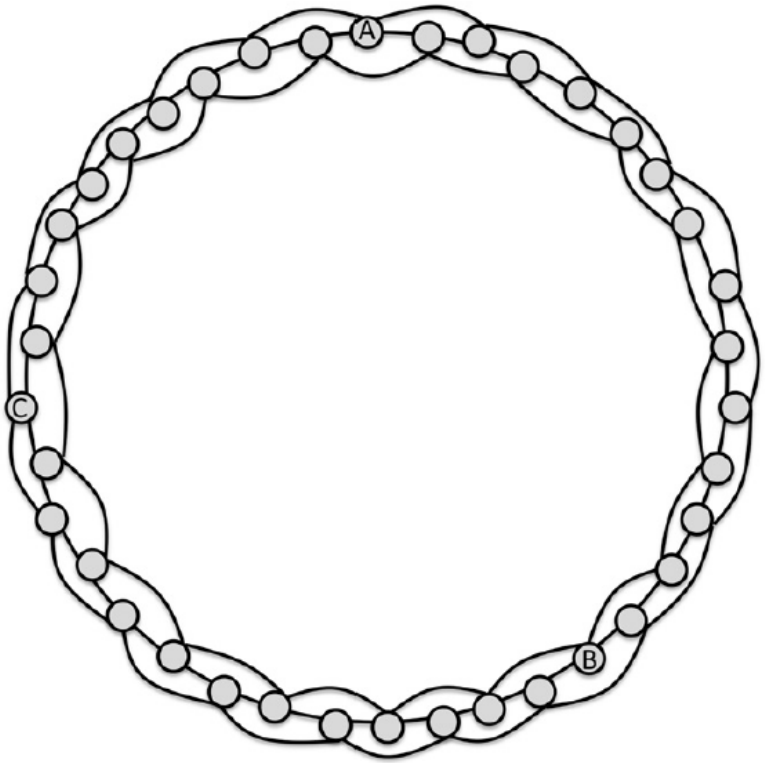
Facebook study, late 2011

- 720 million users
- Average shortest path length: 4.74

Small worlds

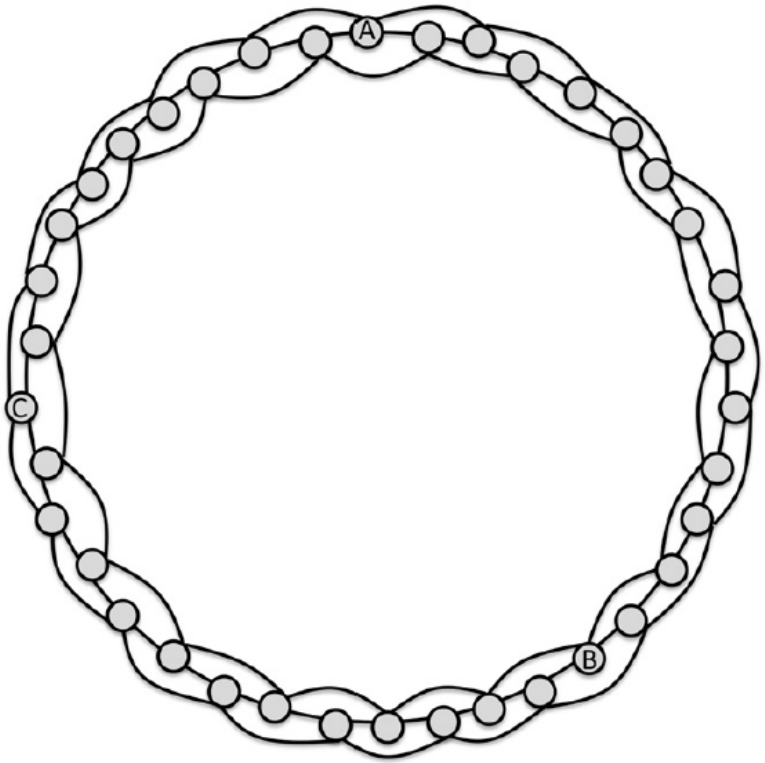
- Another main characteristic: high *clustering*
- A person's friends tend to know one another
- Computed as the average of the nodes' local clustering coefficients

Example: Regular Graph



Node A and B: local clustering coefficient

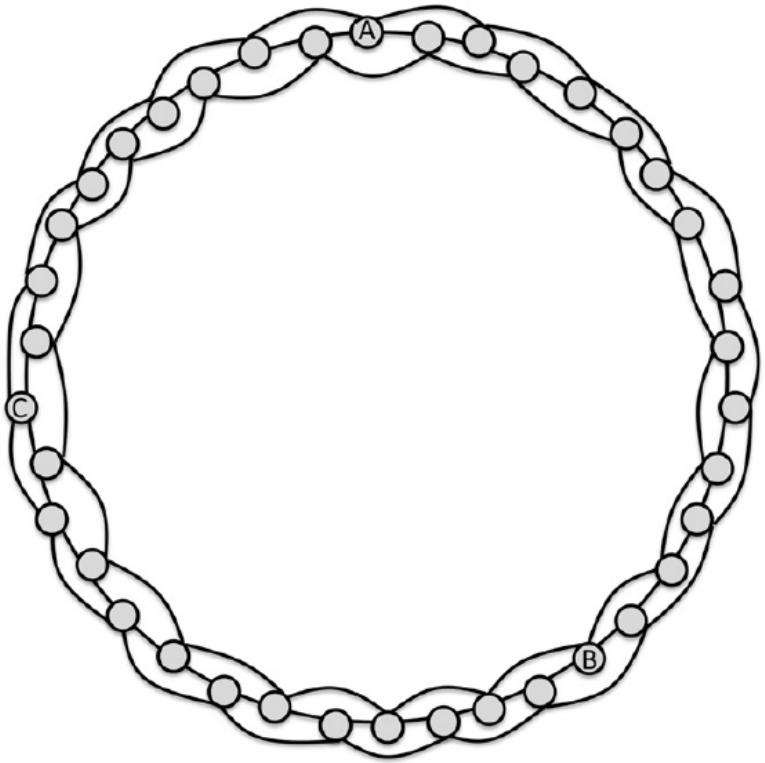
Example: Regular Graph



Node A and B: local clustering coefficient

- Node A: 4 neighbour, 6 possible edges, 3 exist, **0.5**

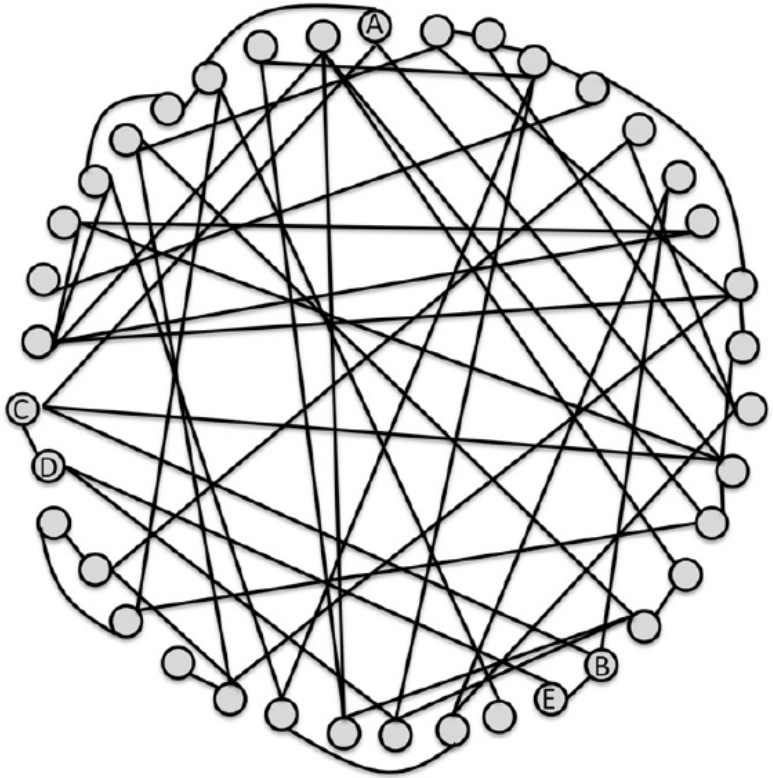
Example: Regular Graph



Node A and B: local clustering coefficient

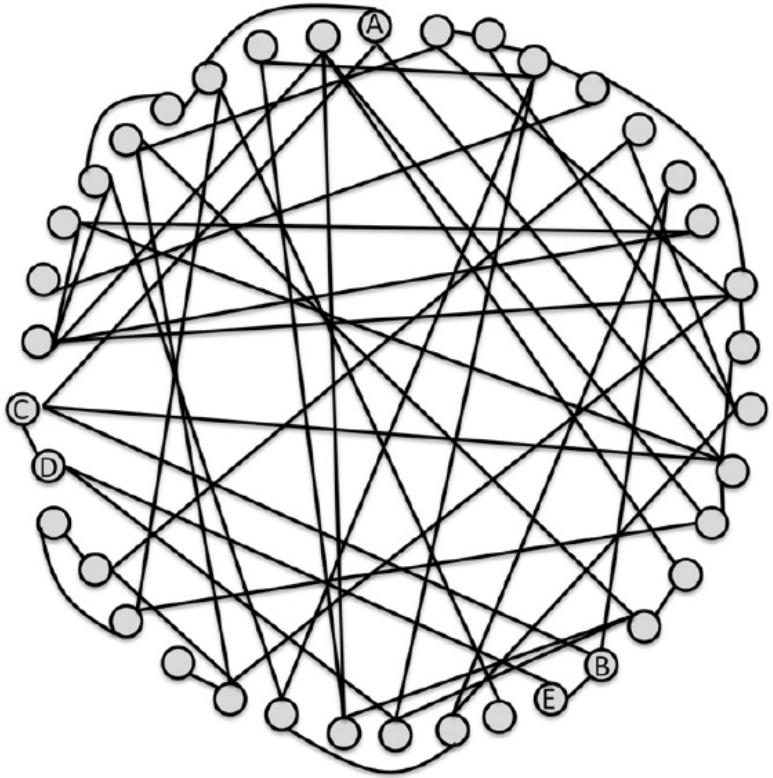
- Node A: 4 neighbour, 6 possible edges, 3 exist, **0.5**
- Node B: Same as Node A

Example: Random Graph



Node A and B: local clustering coefficient

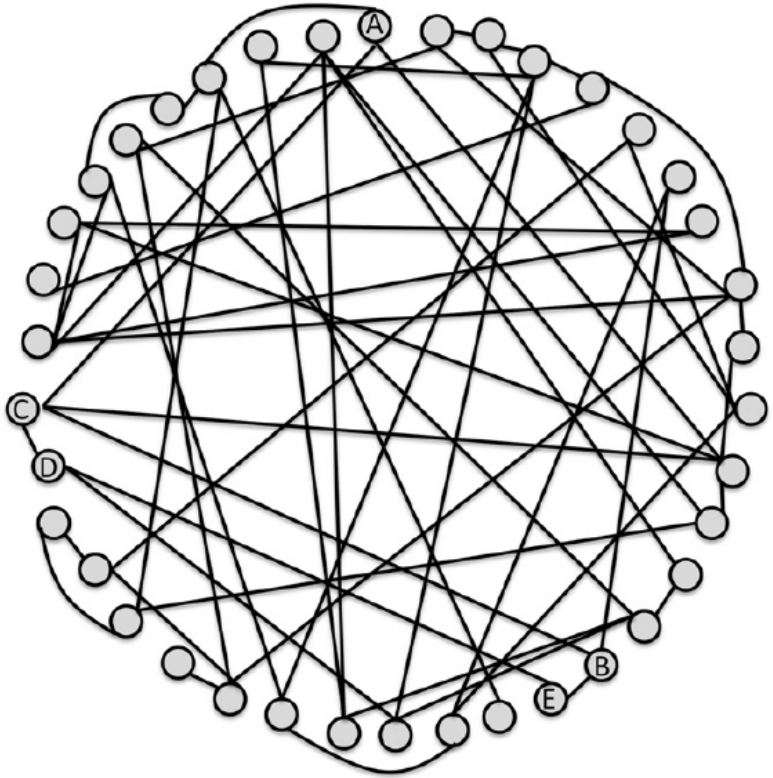
Example: Random Graph



Node A and B: local clustering coefficient

- Node A: 3 neighbour, 3 possible edges, 1 exists, **0.33**

Example: Random Graph



Node A and B: local clustering coefficient

- Node A: 3 neighbour, 3 possible edges, 1 exists, **0.33**
- Node B: 0

Regular vs. Random graphs

Clustering and path length

Regular vs. Random graphs

Clustering and path length

- In regular graphs, the clustering is high
- In random graphs, the shortest path length is small

Regular vs. Random graphs

Clustering and path length

- In regular graphs, the clustering is high
- In random graphs, the shortest path length is small
- Combination of the two?

Combining Regular and Random Graphs

Experiment by Watts and Strogatz, 1988

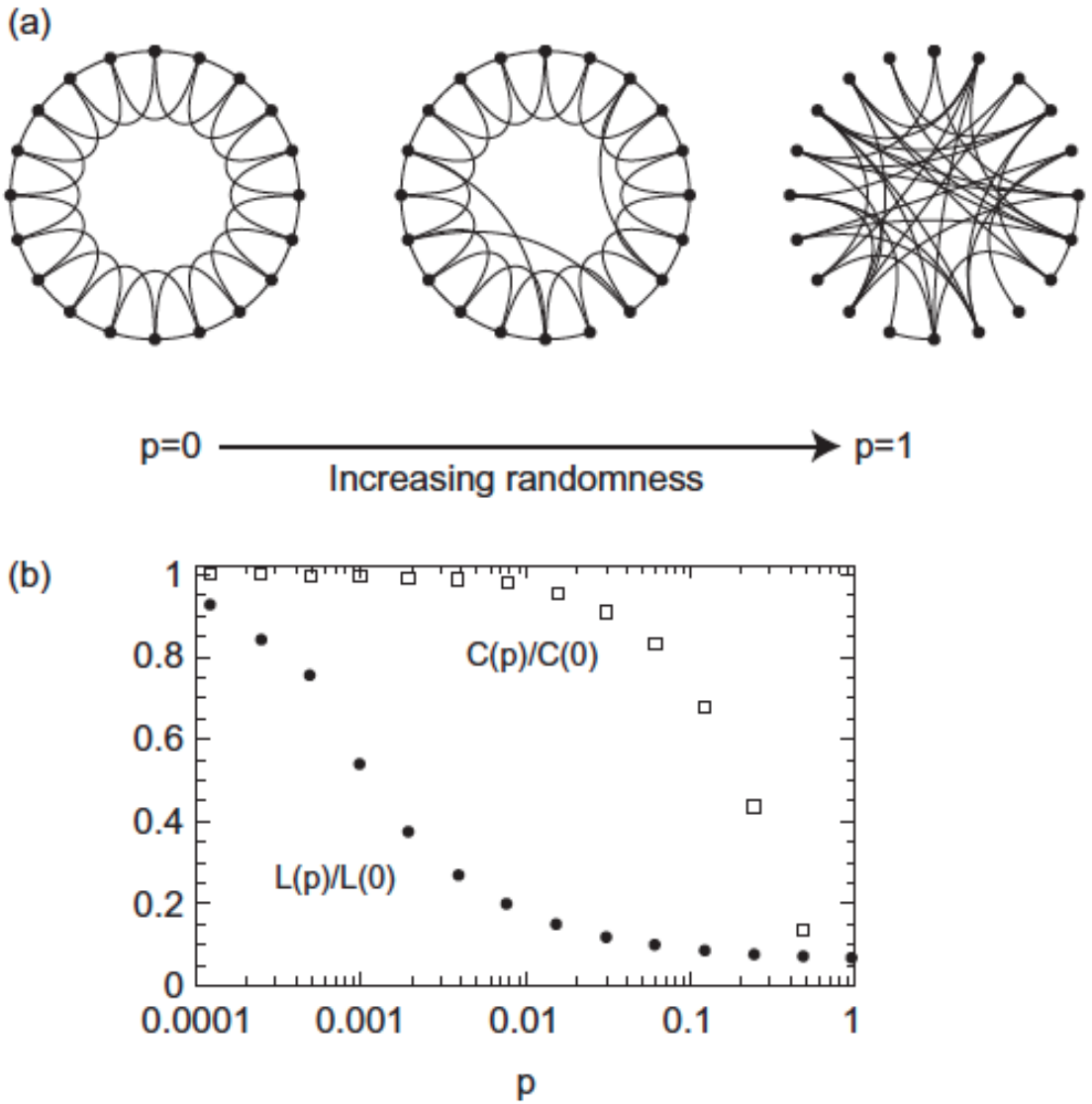
- Took a regular graph and randomly rewired a few edges
- No significant impact on clustering, which remains high

Combining Regular and Random Graphs

Experiment by Watts and Strogatz, 1988

- Took a regular graph and randomly rewired a few edges
- No significant impact on clustering, which remains high
- But significant impact on the path length

Effect of randomness: Clustering (C) and Path length (L)



Social networks: do they evolve in this way?

High clustering coefficient

- Many of our friends know each other
- Looks like a regular graph

Social networks: do they evolve in this way?

High clustering coefficient

- Many of our friends know each other
- Looks like a regular graph

What about short path lengths?

- We know people in different social circles
- Also have connections to people who may be totally outside our social circle
- These correspond to randomly rewired connections, and connect us to otherwise distant social groups

Reference

Jennifer Golbeck. *Analyzing the social web*, Morgan Kaufmann, 2013.
Chapters 2 and 3