Social Network Analysis

Ifeoma Adaji

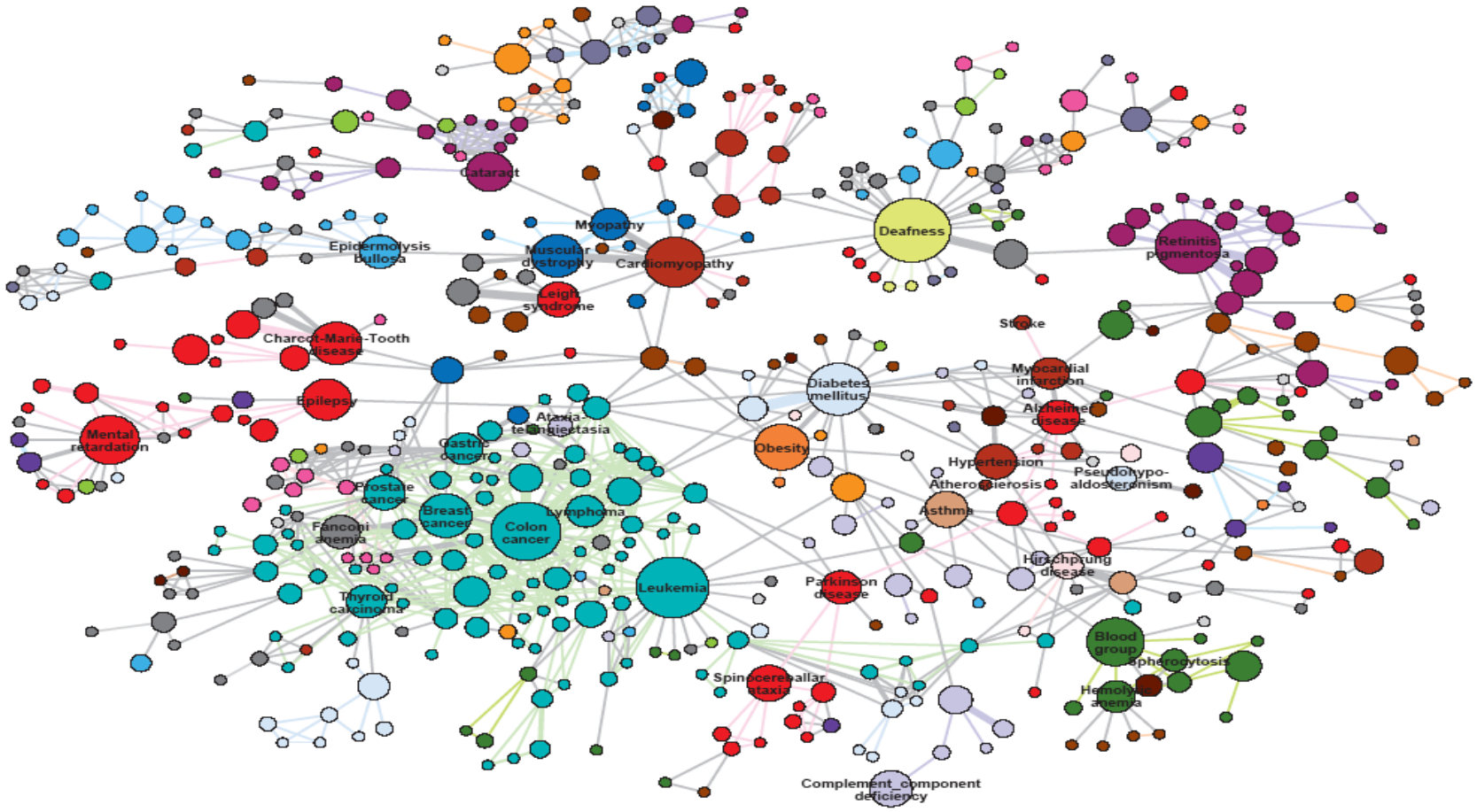
Learning objectives

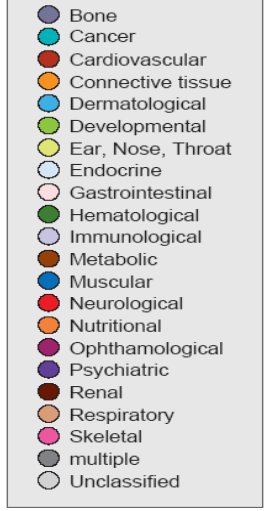
- What is social network analysis
  - Why is it important
- Network elements
  - Nodes, edges
- Network metrics
  - Centrality (degree, closeness, betweenness)
- Application of social network analysis
  - Movie galaxy
- Tools for social network analysis
  - NodeXL
  - Gephi

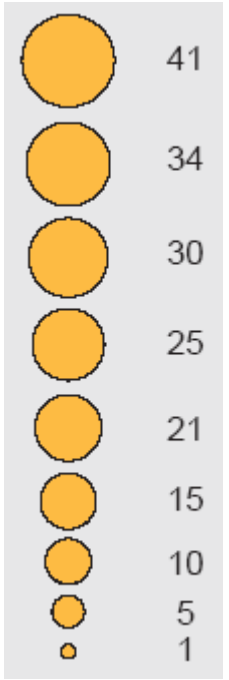
Social Network Analysis

- Analysis of the structure of relationships between social entities
- Process of investigating social structures in networks
- SNA uses mathematical concepts & graph theory
- Consists of nodes and edges

Human disease network

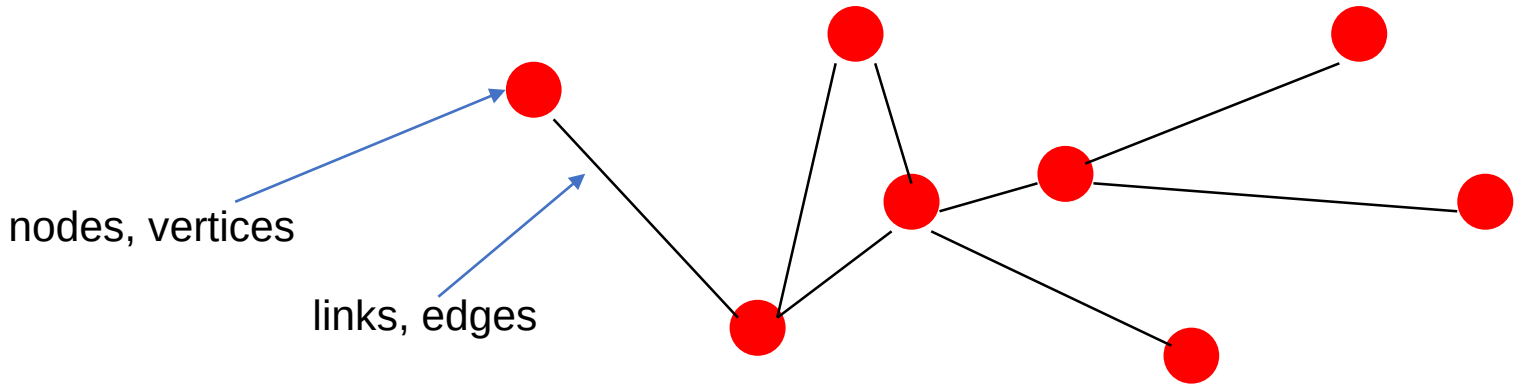






Networks

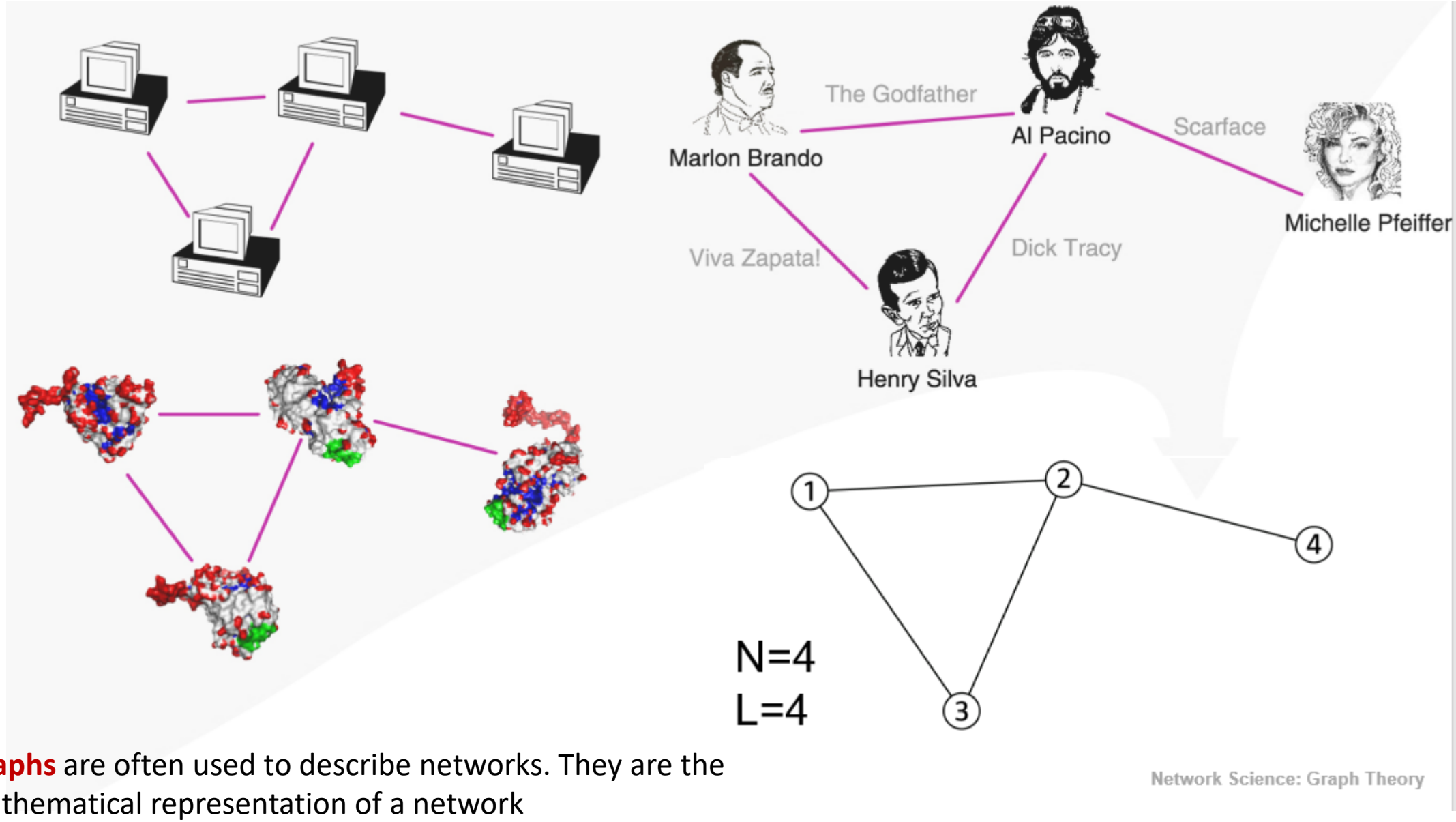
- Networks are at the heart of complex systems
  - Composed of many interconnected parts
- A network is a catalog of a system's components (nodes, N) and the direct interactions between them (links or edges, L)
- Network often refers to real systems
  - The world wide web, social network, network of actors



Nodes can be individuals, actors, people or things within a network (books in a library)

Points	Lines	Discipline
Vertices	Edges, arcs	Math
Nodes	Links	Computer science
Sites	Bonds	Physics
Actors	Ties, relations	Sociology

Networks

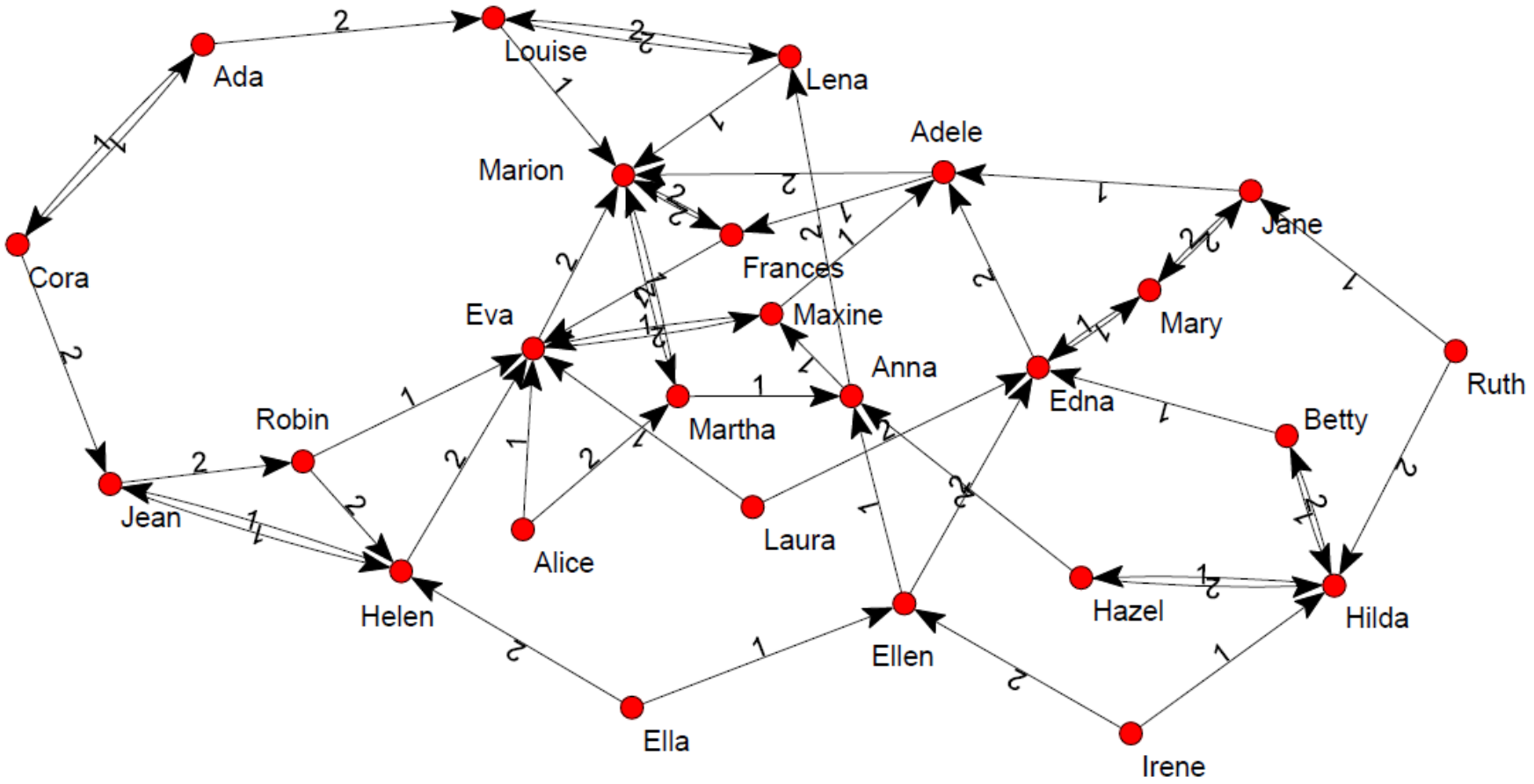


Graphs are often used to describe networks. They are the mathematical representation of a network

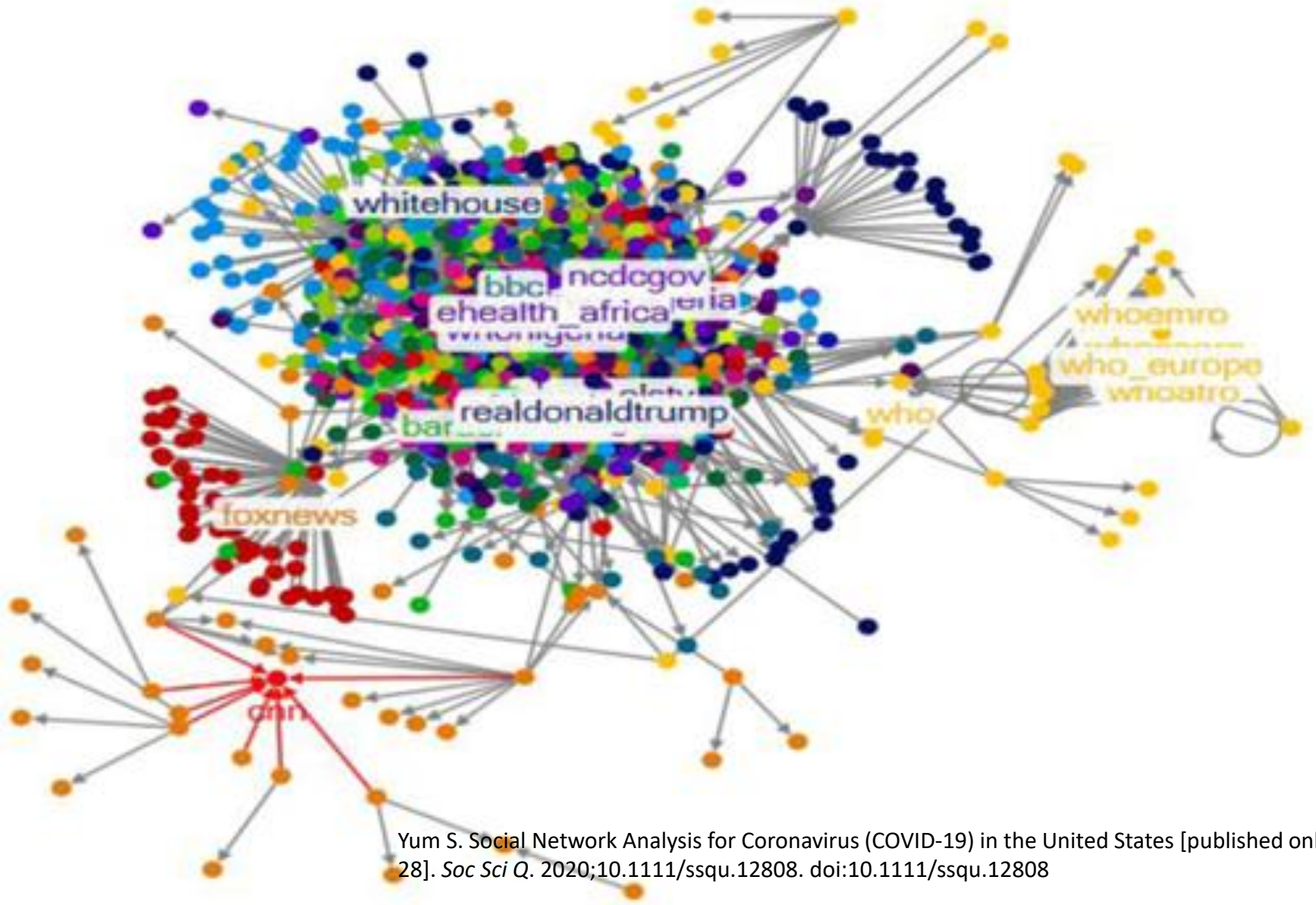
Networks

- Edge/links can be directed or undirected
- A network is directed if all of its links are directed
  - Retweet, likes, following on social media
  - Jane follows John; John might not follow Jane (Jane -> John)
- A network is undirected if all its links are undirected
  - Jane and John follow each other (Jane <-> John or Jane – John)

Example of directed network

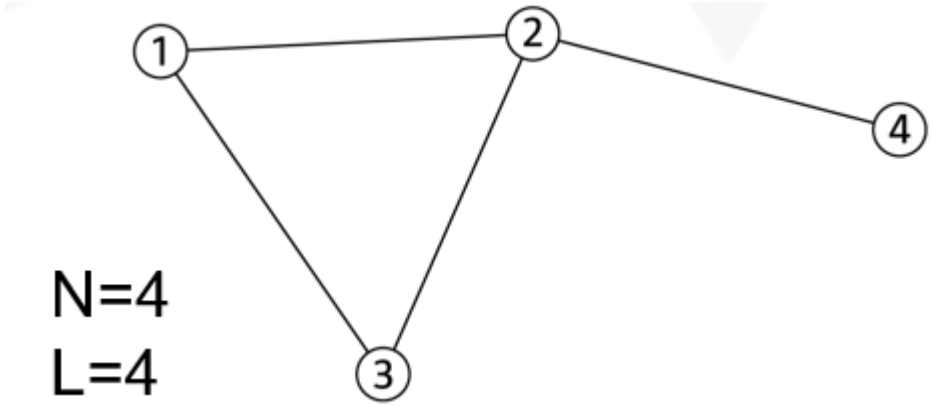


Choosing , 1st and 2nd dining table partners in a girls' school dormitory (Moreno, *The sociometry reader*, 1960)



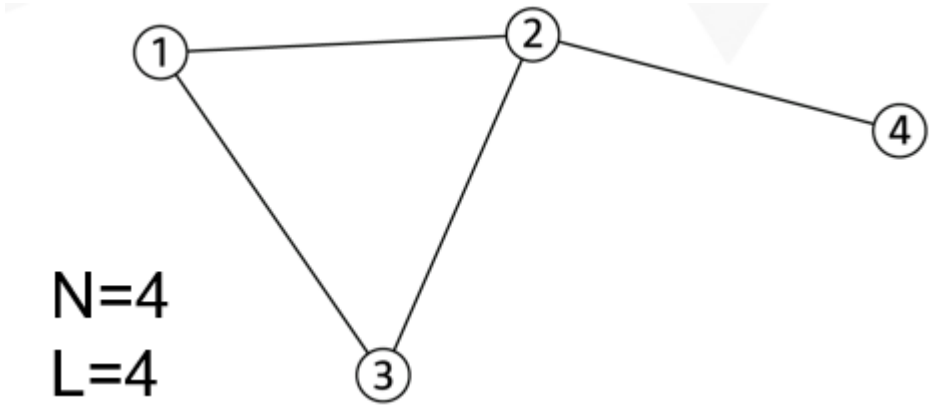
Yum S. Social Network Analysis for Coronavirus (COVID-19) in the United States [published online ahead of print, 2020 May 28]. *Soc Sci Q*. 2020;10.1111/ssqu.12808. doi:10.1111/ssqu.12808

Adjacency matrix



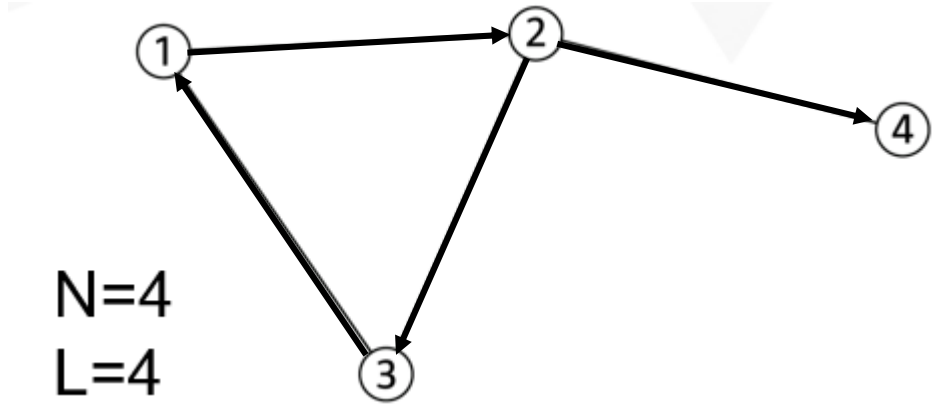
- The adjacency matrix of a directed network of N nodes has N rows and N columns, its elements being:
  - A_{ij} = 1 if there is a link pointing from node i to node j
  - A_{ij} = 0 if nodes i and j are not connected to each other
- The adjacency matrix for an undirected network has two entries for each link, for example, link (1,2) is represented as A_{12} = 1 and A_{21}=1
- Hence, the adjacency matrix of an undirected network is symmetric, $A_{ij} = A_{ji}$
- A_{ii}=0 unless the node has self loops

Adjacency matrix



A_{ij} =

	1	2	3	4
1	0	1	1	0
2	1	0	1	1
3	1	1	0	0
4	0	1	0	0

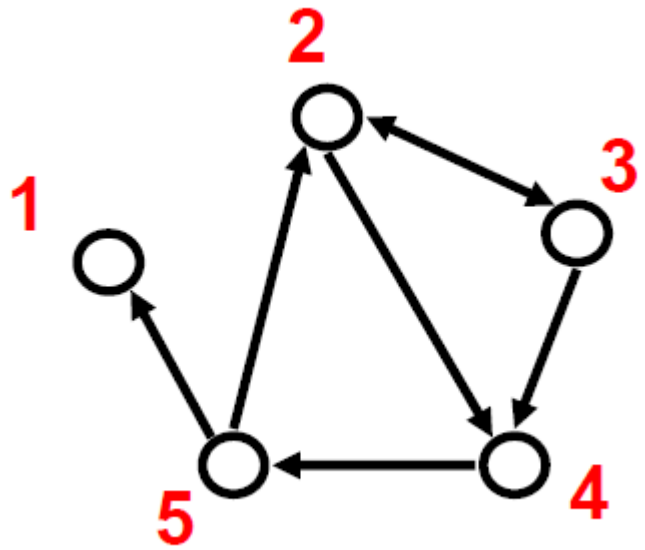


A_{ij} =

	1	2	3	4
1	0	1	0	0
2	0	0	1	1
3	1	0	0	0
4	0	0	0	0

- A_{ij} = 1 if there is a link pointing from node i to node j
- A_{ij} = 0 if nodes i and j are not connected to each other
- A_{ii}=0 unless the node has self loops
- $A_{ij} = A_{ji}$ if the network is undirected

Adjacency matrix of the network below

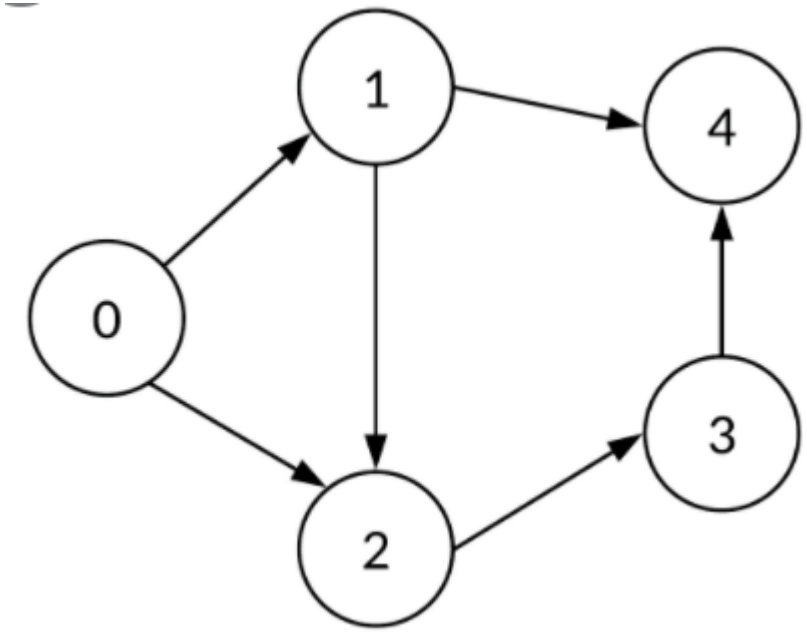


$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Draw the corresponding directed graph for the adjacency matrix below

A_{ij} =

	1	2	3	4	5
1	0	1	1	0	0
2	0	0	1	0	1
3	0	0	0	1	0
4	0	0	0	0	1
5	0	0	0	0	0



Network metrics – Indices of centrality

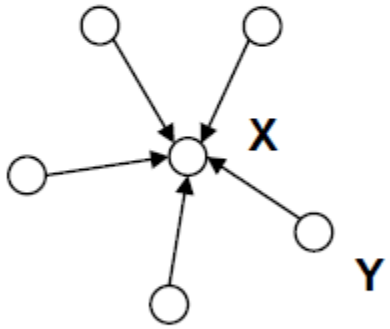
- Centrality indices are answers to the question

  "What characterizes an important node?"

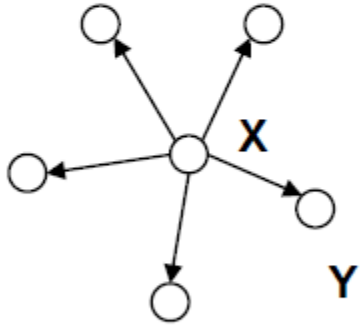
- They are measures of nodes and their relationship to the overall network
  - E.g. The most important/influential person in a network, spreaders of disease, topic discussed the most.
- Degree –Indegree, Outdegree: How many people can this person reach directly?
- Betweenness: How likely is this person to be the most direct route between two people in the network?
- Closeness: How fast can this person reach everyone in the network?

Network metrics – Indices of centrality

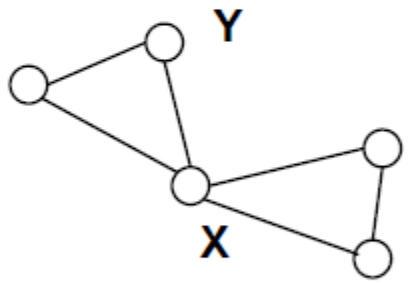
In each of the following networks, X has higher centrality than Y according to a particular measure



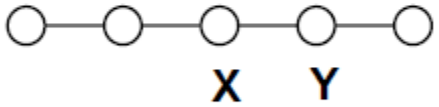
indegree



outdegree



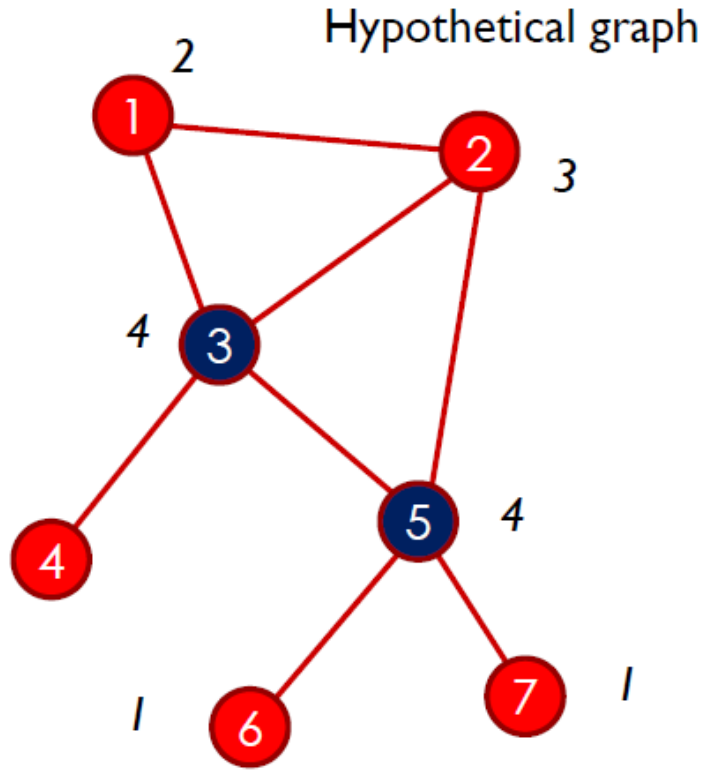
betweenness



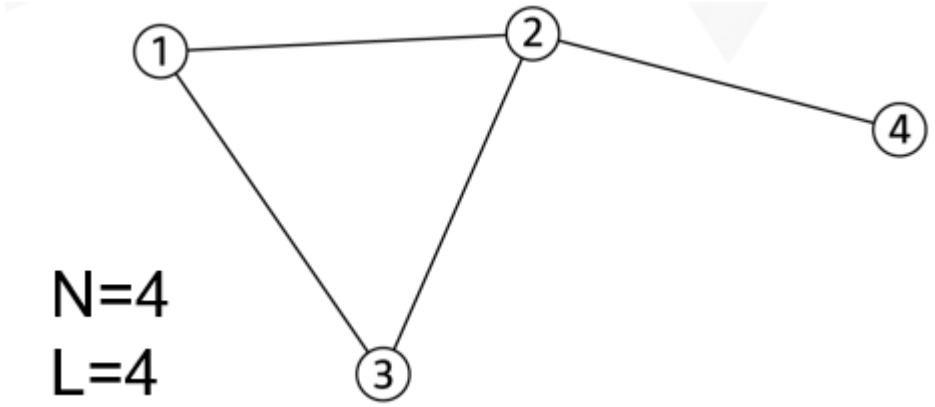
closeness

Degree centrality

- Representation of a system's components (N)
  - Ex. Actors, people, things within a network
- Node network properties
  - Degree of a node: Number of links that lead into or
  - Indegree: how many directed edges (arcs) are incid
  - Outdegree: how many directed edges (arcs) origina

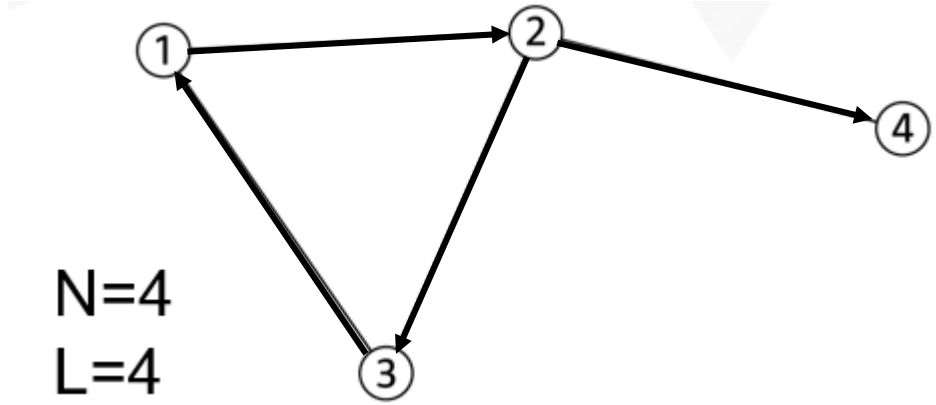


Calculating degree from adjacency matrix



A_{ij} =

	1	2	3	4
1	0	1	1	0
2	1	0	1	1
3	1	1	0	0
4	0	1	0	0



Outdegree

Indegree

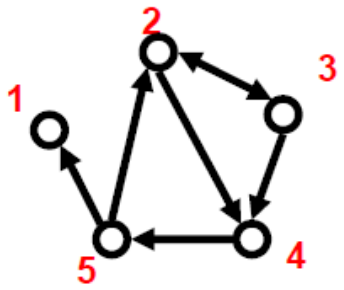
A_{ij} =

	1	2	3	4
1	0	1	0	0
2	0	0	1	1
3	1	0	0	0
4	0	0	0	0

- A_{ij} = 1 if there is a link pointing from node i to node j
- A_{ij} = 0 if nodes i and j are not connected to each other
- A_{ii}=0 unless the node has self loops
- $A_{ij} = A_{ji}$ if the network is undirected

Calculating degree from adjacency matrix

- Outdegree = $\sum_{j=1}^{n} A_{ij}$



A =

	1	2	3	4	5
1	0	0	0	0	0
2	0	0	1	1	0
3	0	1	0	1	0
4	0	0	0	0	1
5	1	1	0	0	0

example: outdegree for node 3 is 2, which we obtain by summing the number of non-zero entries in the 3rd *row*

$$\sum_{j=1}^{n} A_{3j}$$

- Indegree = $\sum_{i=1}^{n} A_{ij}$

A =

	1	2	3	4	5
1	0	0	0	0	0
2	0	0	1	1	0
3	0	1	0	1	0
4	0	0	0	0	1
5	1	1	0	0	0

example: the indegree for node 3 is 1, which we obtain by summing the number of non-zero entries in the 3rd *column*

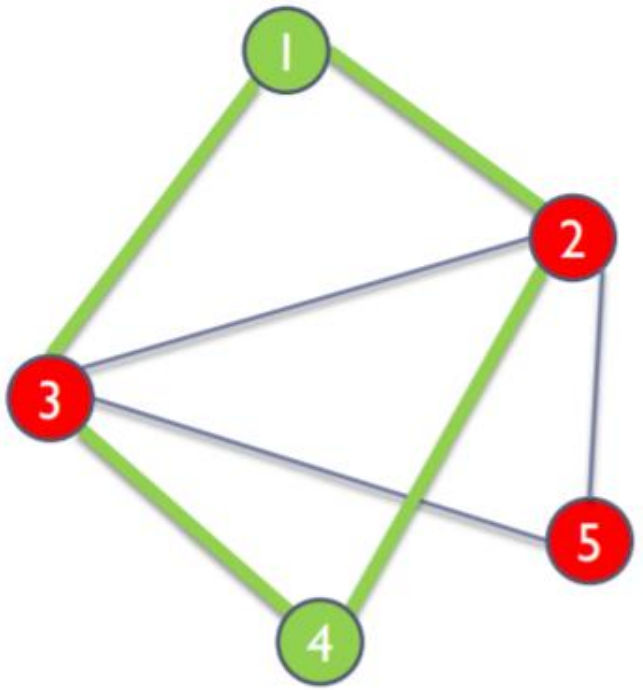
$$\sum_{i=1}^{n} A_{i3}$$

Paths and shortest paths

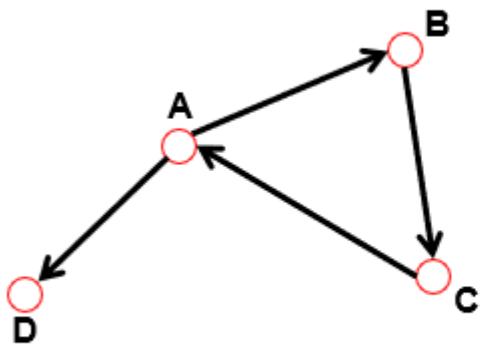
Shortest path

- A *path* between two nodes is any sequence of non-repeating nodes that connects the two nodes
- The *shortest path* between two nodes is the path that connects the two nodes with the shortest number of edges (also called the *distance* between the nodes)
- In the example to the right, between nodes 1 and 4 there are two shortest paths of length 2: {1,2,4} and {1,3,4}
- Other, longer paths between the two nodes are {1,2,3,4}, {1,3,2,4}, {1,2,5,3,4} and {1,3,5,2,4} (the longest paths)
- Shorter paths are desirable when speed of communication or exchange is desired (often the case in many studies, but sometimes not, e.g. in networks that spread disease)

Hypothetical graph



Paths and shortest paths

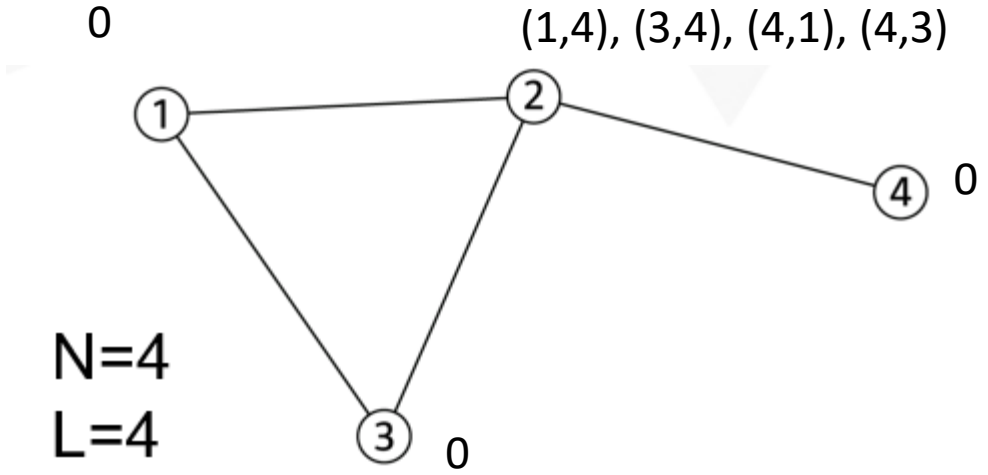


In directed graphs each path needs to follow the direction of the arrows.

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).

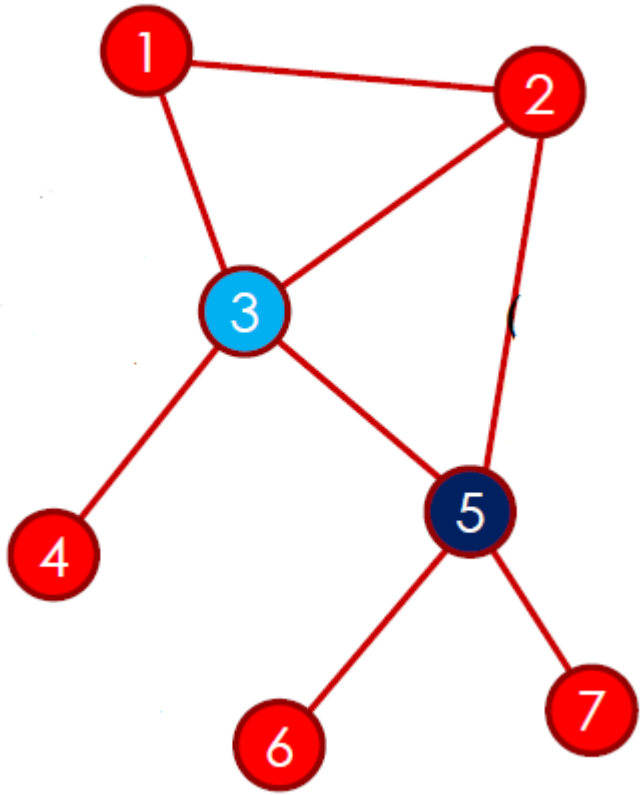
Betweenness: captures brokerage

- Quantifies the number of times a node acts as a bridge along the shortest path between 2 other nodes.
- How many pairs of individuals in the network would be connected through you on their shortest path to each other
- Shows which nodes are more likely to be in communication paths between other nodes
- How likely is this person/node to be the most direct route between two people in the network? – Information control



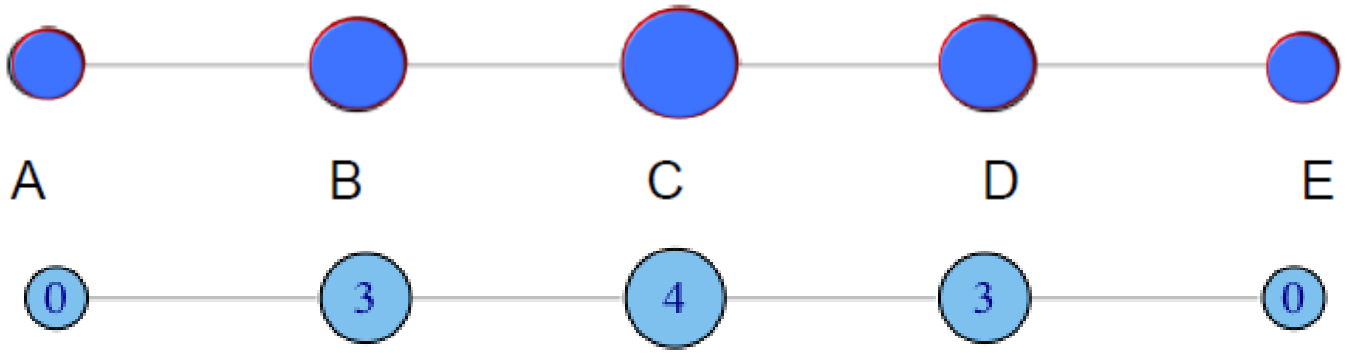
What node has the highest betweenness centrality?

- How many node pairs will go through each node on their shortest path to each other?
- Which nodes are more likely to be in communication paths between other nodes
- Useful in determining points where the network would break apart (think who would be cut off if nodes 3 or 5 would disappear)



Node 5 has higher betweenness centrality than 3
No node pairs need to go through 1,4,6 or 7 to get to each other

What node has the highest betweenness centrality?



- A lies between no two other vertices
- B lies between 3 pairs of vertices: (A,C), (A,D), (A,E)
- C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)
- D lies between 3 pairs of vertices (A,E), (B,E), (C,E)
- E lies between no two other vertices

Closeness centrality

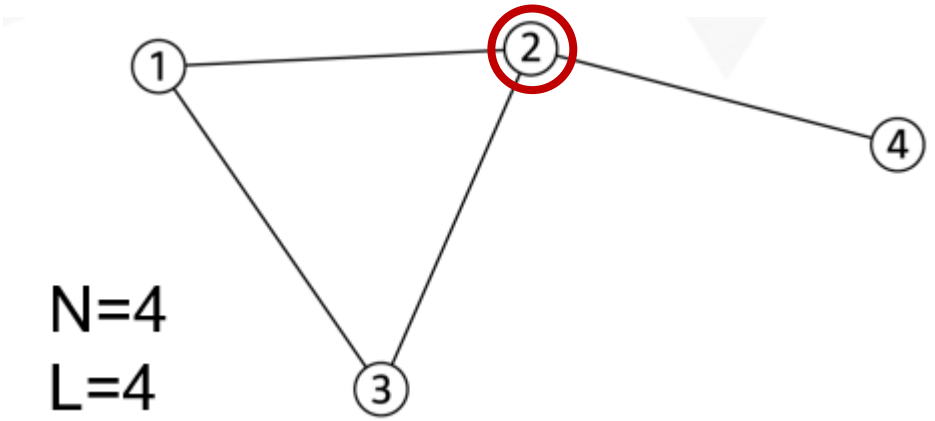
- Closeness is a measure of reach
  - The speed with which information can reach other nodes from a starting node
- Useful in cases where speed of information dissemination is main concern
- Calculate the mean length of all shortest paths from a node to all other nodes in the network (i.e. how many hops on average it takes to reach every other node)
- Take the reciprocal of the above value so that higher values are 'better' (indicate higher closeness) like in other measures of centrality

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

Closeness centrality

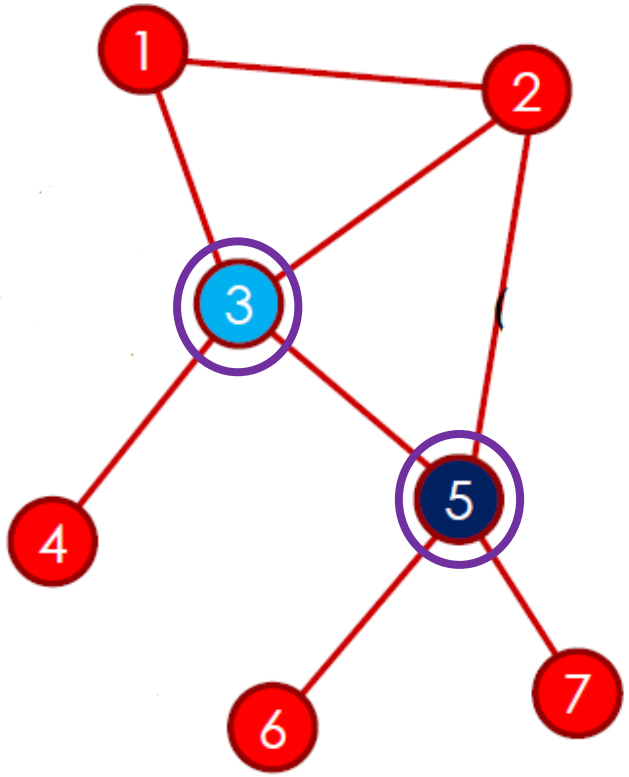
- Calculate the mean length of all shortest paths from a node to all other nodes in the network (i.e. how many hops on average it takes to reach every other node)
- Take the reciprocal of the above value so that higher values are 'better' (indicate higher closeness) like in other measures of centrality

	1	2	3	4	Sum	Avg	Rec
1							
2							
3							
4							



What node has the highest Closeness centrality

	1	2	3	4	5	6	7	Sum	Avg	Rec
1	-	1	1	2	2	3	3	12	2	0.5
2	1	-	1	2	1	2	2	9	1.5	0.67
3	1	1	-	1	1	2	2	8	1.3	0.75
4	2	2	1	-	2	3	3	13	2.17	0.46
5	2	1	1	2	-	1	1	8	1.33	0.75
6	3	2	2	3	1	-	2	13	2.17	0.46
7	3	2	2	3	1	2	-	13	2.17	0.46



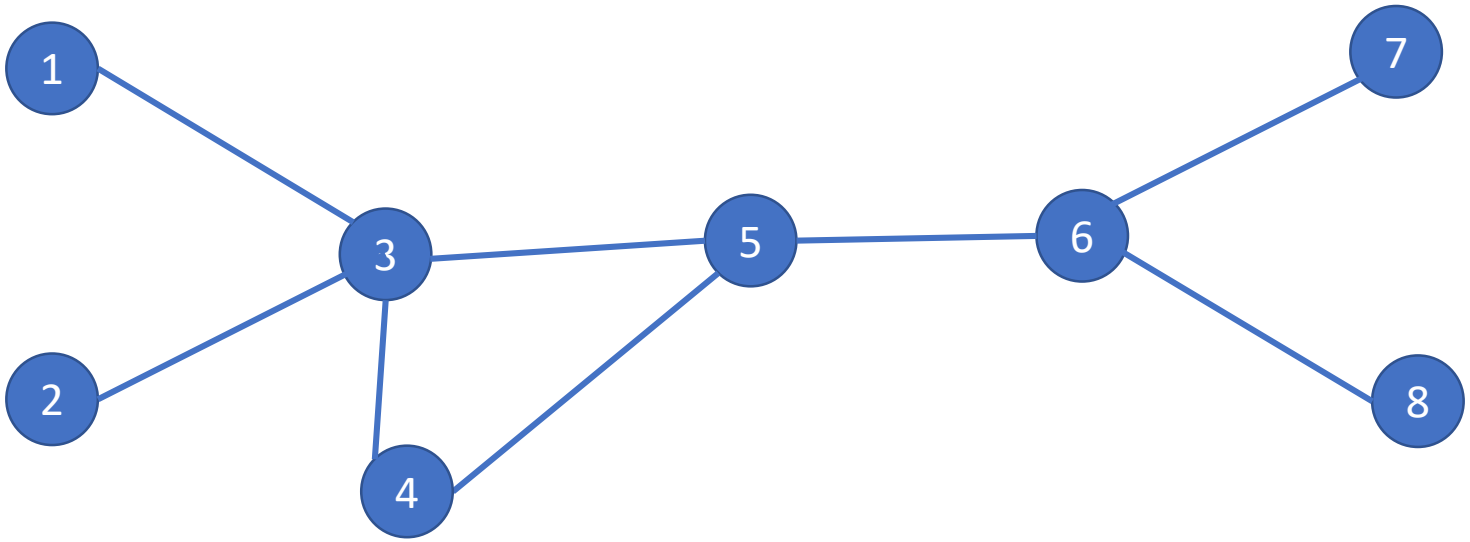
Centrality measures

Centrality measure	Interpretation in social networks
Degree	How many people can this person reach directly?
Betweenness	How likely is this person to be the most direct route between two people in the network?
Closeness	How fast can this person reach everyone in the network?

Centrality measures

Centrality measure	**Other possible interpretations…**
Degree	In network of music collaborations: how many people has this person collaborated with?
Betweenness	In network of spies: who is the spy through whom most of the confidential information is likely to flow?
Closeness	In network of relations work/family: how fast will Tuberculosis spread from this person to the rest of the network?

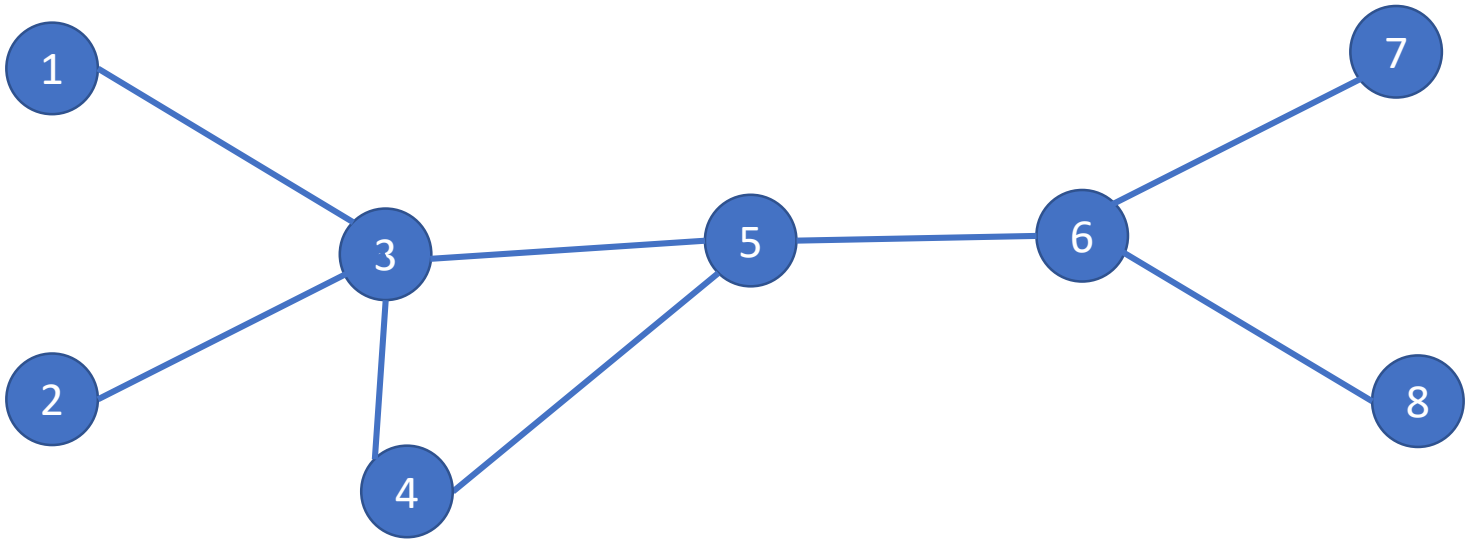
Exercise



- Let's assume the above network is a friendship network. Each node represents a person and the edge represents friendship between people. If one is interested in the node(s) that can most efficiently (quickly) get information from every other node in the network, what centrality measure will be most appropriate? Why?

- Closeness centrality

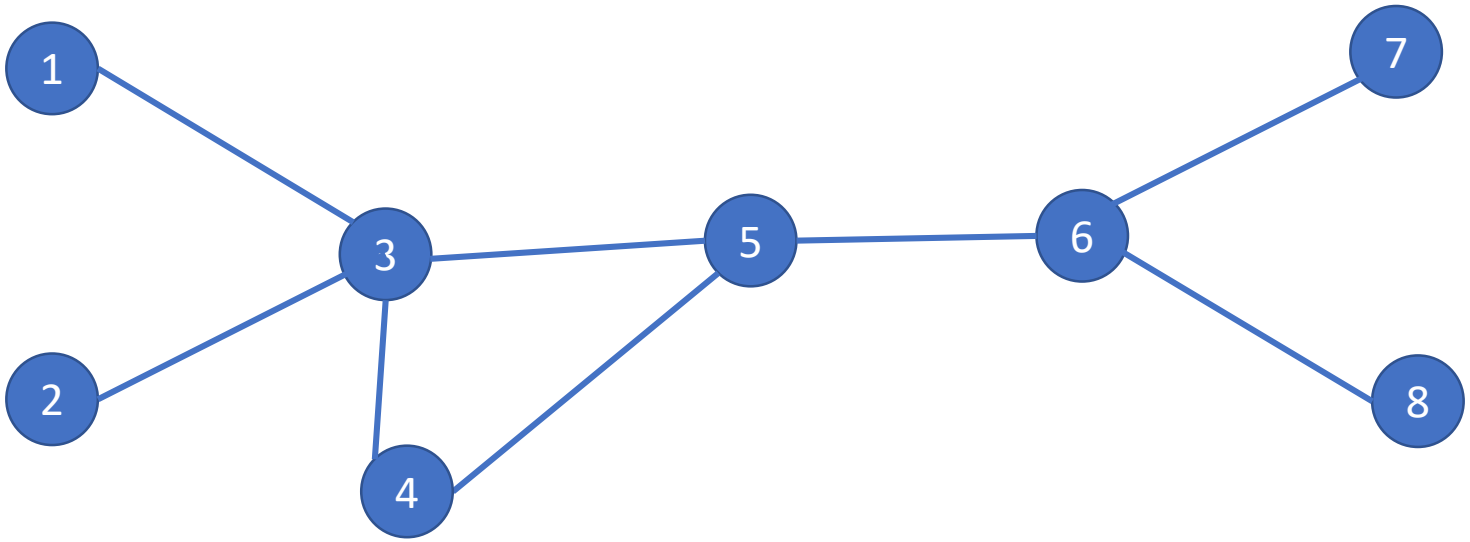
Exercise



- Let’s assume the above network is a friendship network. If one is interested in finding the node that can most frequently control information flow in the network, which centrality measure will be most appropriate? Why?

- Betweenness centrality

Exercise



- Let's assume the above network is a friendship network. Each node represents a person and the edge represents friendship between people. If one is interested in finding the most popular person in the network, which centrality measure will be most appropriate? Why?

- Degree centrality

Application of SNA – Movie Galaxy

- Movie Galaxy uses SNA to represent the character interaction and relationships between actors in movies.
- Nodes = actors, Links/edges = co-appearance (same screen appearance)
- A relationship exists if 2 people appeared in the same screen.
- http://moviegalaxies.com/movies/828-Titanic

Tools for SNA – NodeXL and Gephi

- NodeXL
  - Tool for SNA; network overview, discovery and exploration
  - Implemented as add-in to MS Excel
  - Allows import from FB, Twitter and other networks
  - Represents network as an edge list
- Gephi
  - Tool for visualizing graphs and networks
  - Open source and free

Why is SNA important

- Marketing insights
- Detecting crime
- Detecting epidemics
- Understanding relationship structures

Conclusion

- Social network analysis describes the relationship between entities in a network
- Networks are represented as nodes and edges
- Centrality metrics
  - Degree
  - Closeness
  - Betweenness

References

- *Network science* by Barabási, Albert-László. Cambridge University Press, 2016.
- *Linked. How everything is connected to everything else and what it means for business, science and everyday life* by Barabási, Albert-László, 2002.
- Lectures notes in Social Computing. Julita Vassileva.