Social Network Analysis 1

Dian Ramadhani

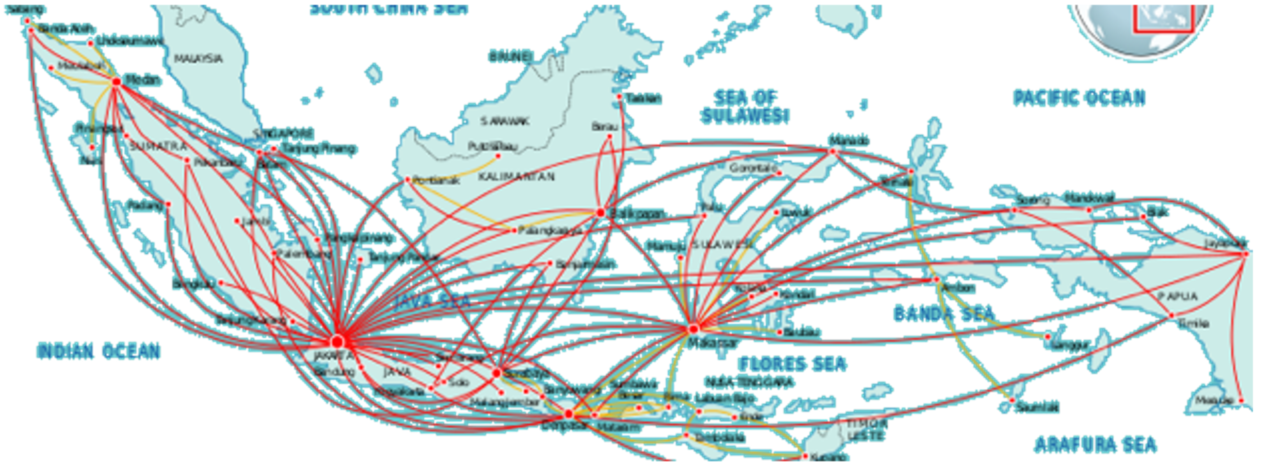
Learning Outline

- Handshake with Social Network Analysis
- Network Representation
- Tie Strength Identification
- Key Player Identification
- Measuring Overall Social Network Structure

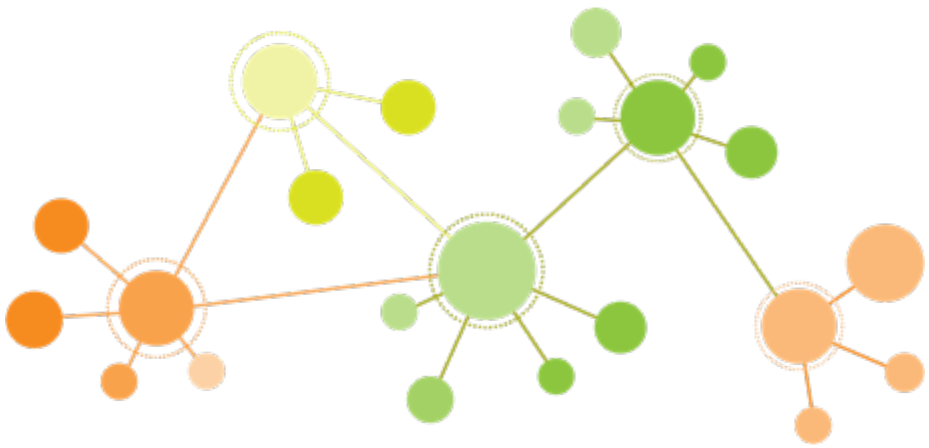
Handshake with
Social Network Analysis

What is a Network?

- Powerful for describing complex systems by explaining interconnection between elements.
- Can be adopted universally in various fields e.g. mathematics, computer science, economics, sociology, chemistry, biology, etc.



Social Network



- Build upon:
  - **Nodes as users**
  - **Edges as interaction flow between users**
  - **Graph type to indicate the nature of interaction**
  - **Weight to indicate the importance level of an interaction**
- Focused on relationships and interconnected behaviors of various entities e.g. objects, people and organizations.
- The more specific area associated with SNA is the dynamic network behavior study formally known as Dynamic Network Analysis (DNA).

Network Representation

Let’s study their interactions as a network

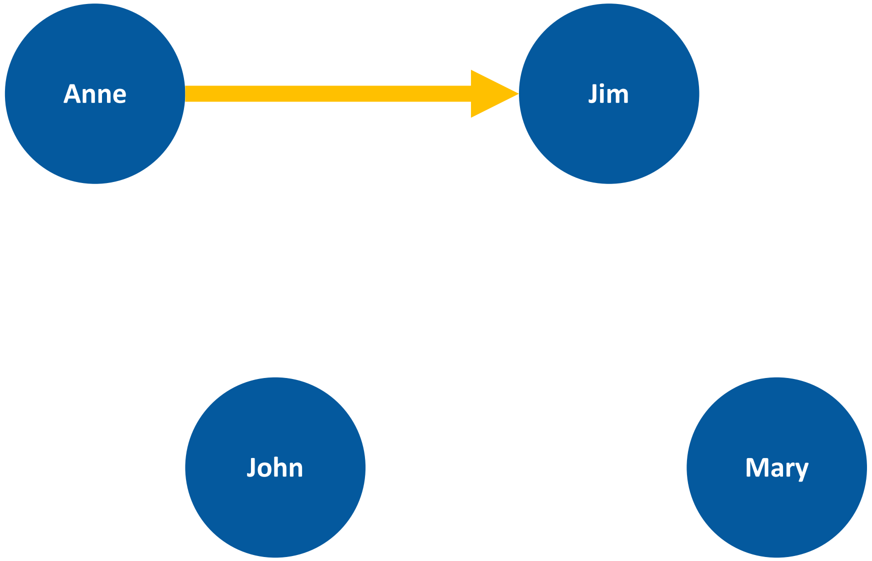
Anne

Jim

John

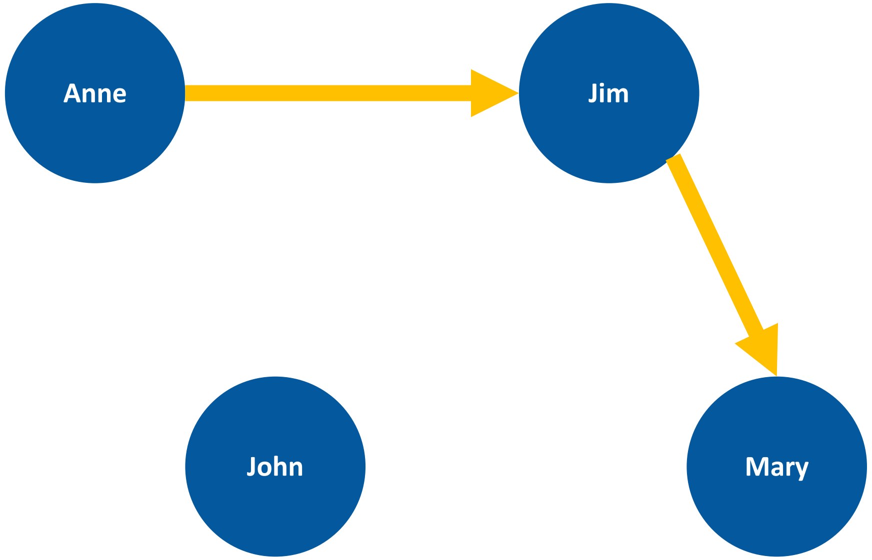
Mary

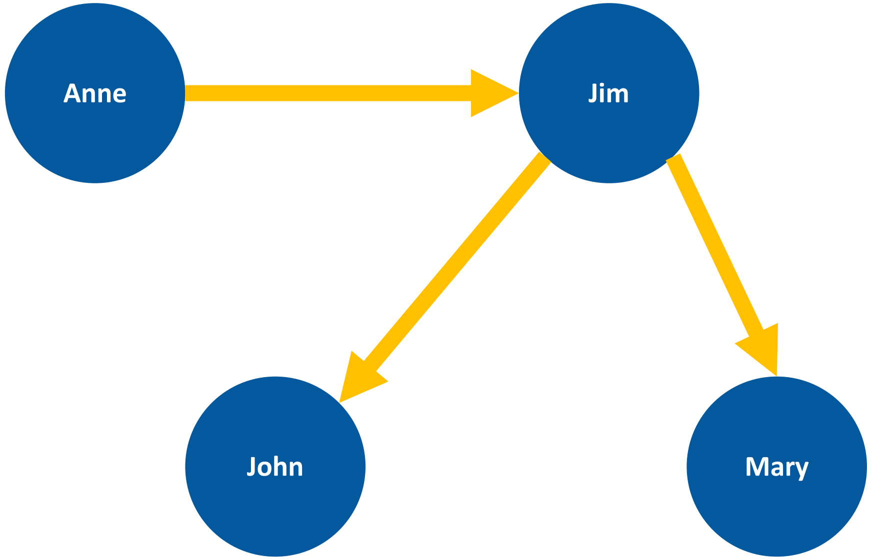
Anne : Jim, tell Mary and John they're invited



Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

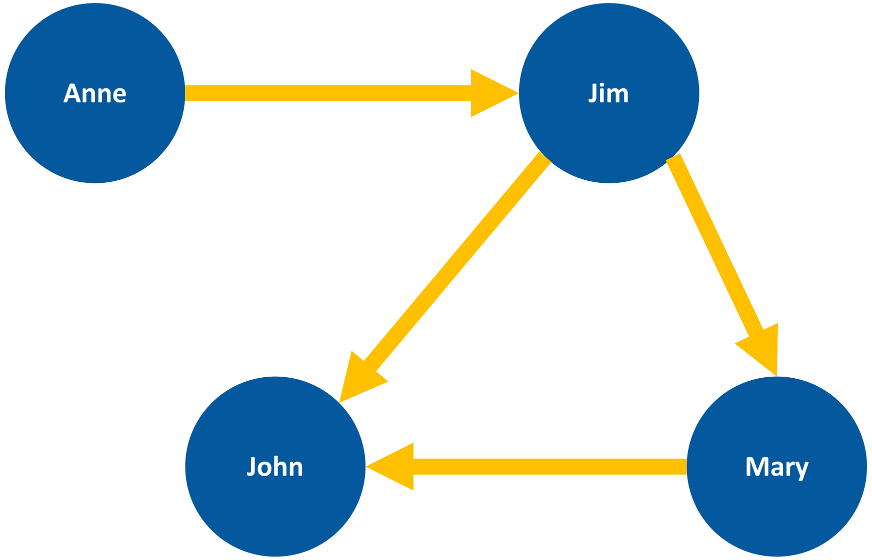




Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

Jim : Mr. John, you should both come for dinner



Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

Jim : Mr. John, you should both come for dinner

Mary : Dad, we are invited for tonight

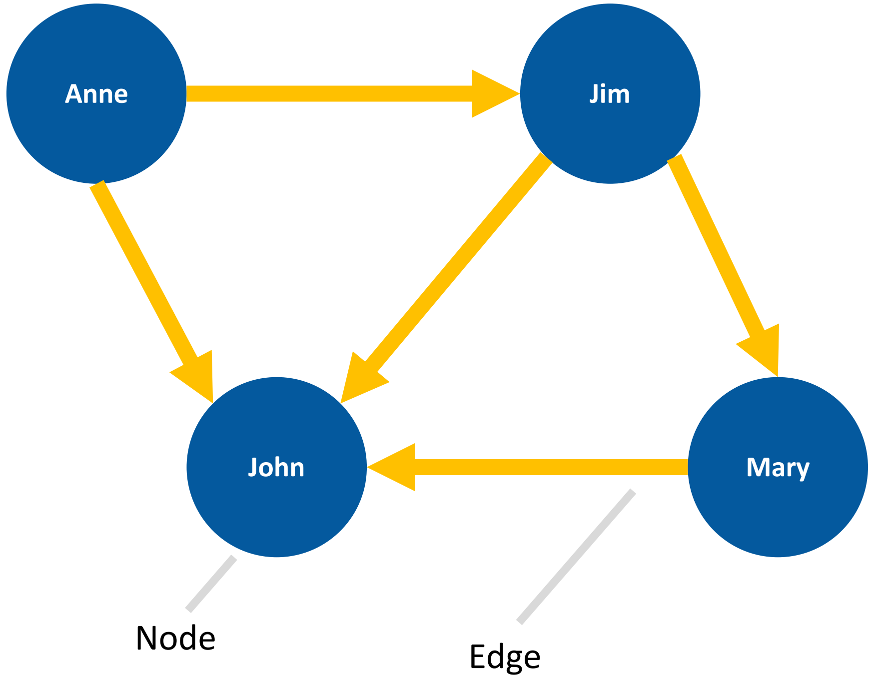
Anne : Jim, tell Mary and John they're invited

Jim : Mary, you and your dad should come for dinner

Jim : Mr. John, you should both come for dinner

Mary : Dad, we are invited for tonight

Anne : John, did Jim tell you about the dinner? You must come



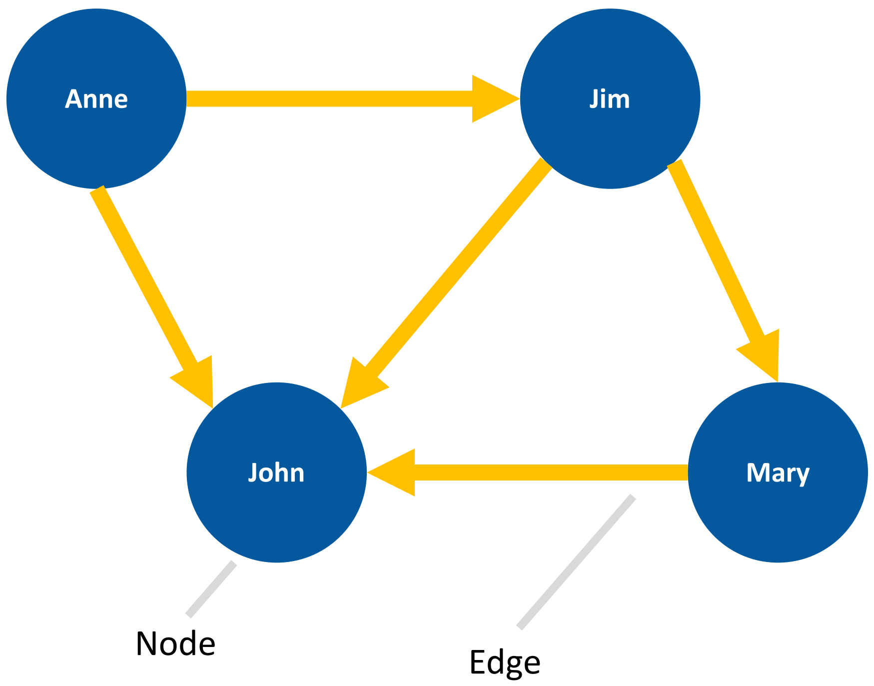
Network Type: Directed

Edge List

Vertex	Vertex
Anne	Jim
Anne	John
Jim	John
Jim	Mary
Mary	John

Adjacency Matrix

Vertex	Anne	Jim	John	Mary
Anne	-	1	1	0
Jim	0	-	1	1
John	0	0	-	0
Mary	0	0	1	-



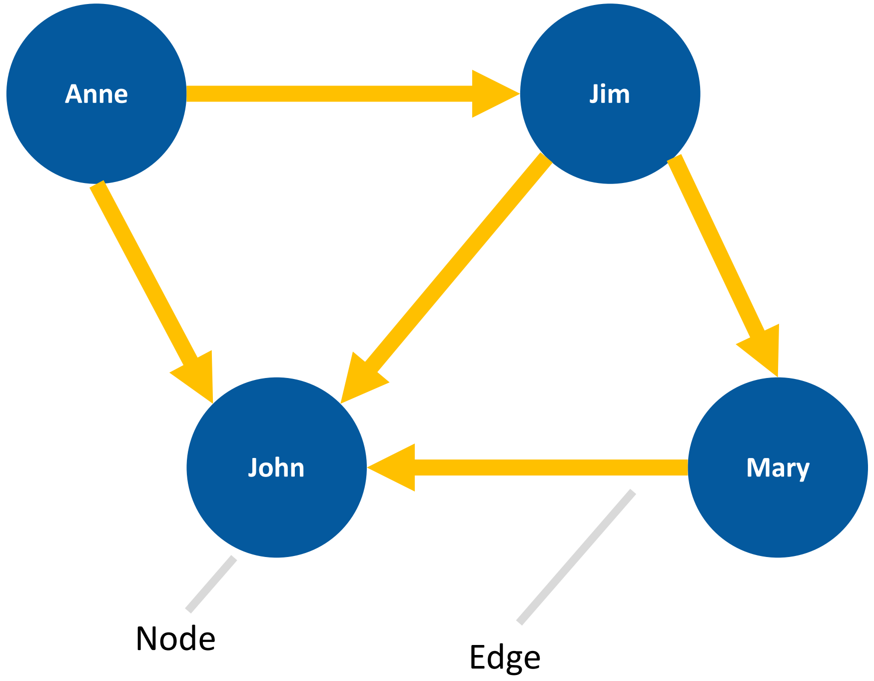
Network Type: Undirected

Edge List

Vertex	Vertex
Anne	Jim
Anne	John
Jim	John
Jim	Mary
Mary	John

Adjacency Matrix

Vertex	Anne	Jim	John	Mary
Anne	-	1	1	0
Jim	1	-	1	1
John	1	1	-	0
Mary	0	1	1	-

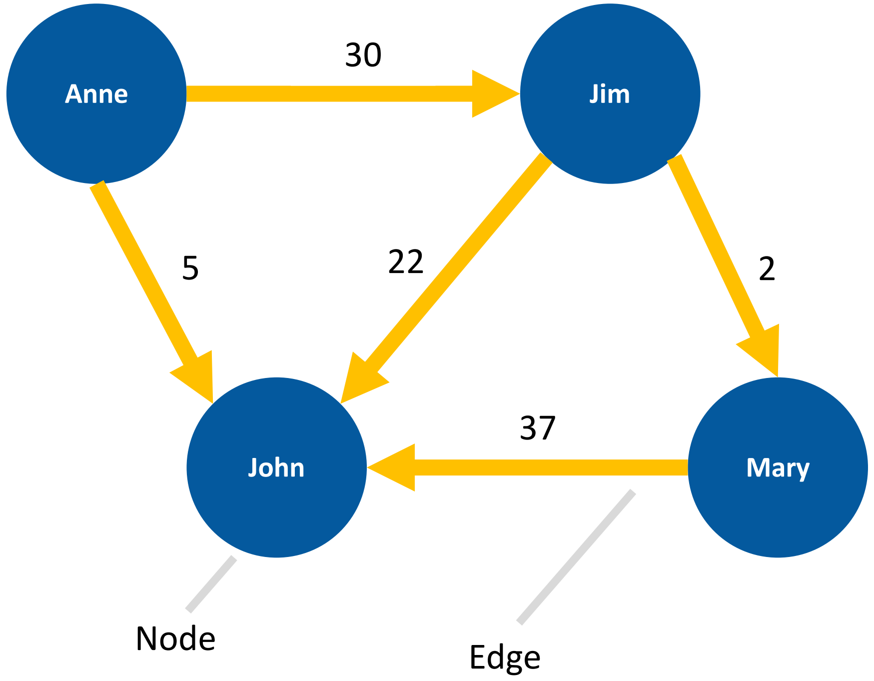


Network Type: Weighted

Edge List

Vertex	Vertex	Weight
Anne	Jim	30
Anne	John	5
Jim	John	22
Jim	Mary	2
Mary	John	27

Weight could be
- Frequency of interactions in period of observation
- Number of items exchanged in period
- Individual perceptions of strength of relationship
- Cost of communications or exchange, e.g. distance



Adjacency Matrix

Vertex	Anne	Jim	John	Mary
Anne	-	30	5	0
Jim	30	-	22	2
John	5	22	-	1
Mary	0	2	37	-

Tie Strength

Edge Weight as Relationship Strength

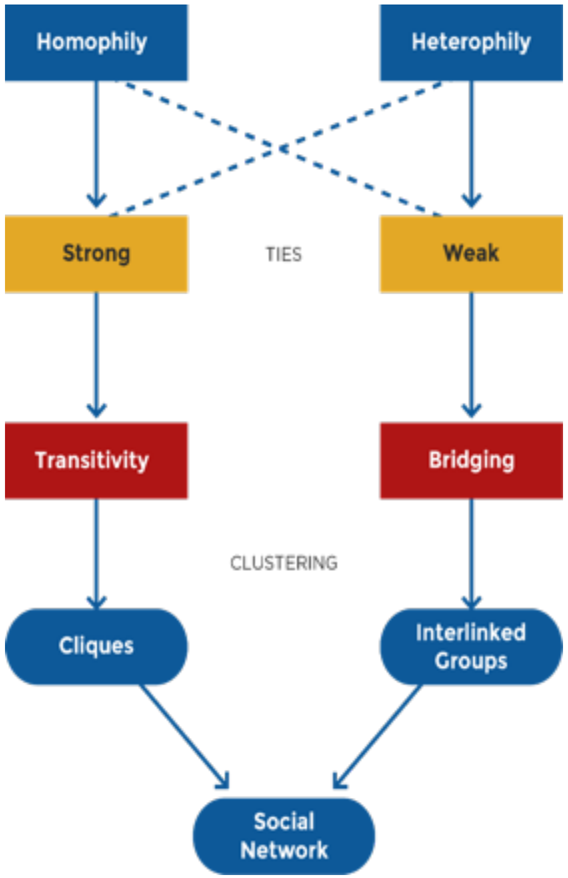
Edges can represent **interactions, flows of information or goods, similarities/affiliations,** or **social relations**

Specifically for social relations, a ‘proxy’ for the strength of a tie can be:

- The frequency of interaction or the amount of exchange
- Reciprocity in interaction or flow
- The type of interaction or flow between the two parties (e.G., Intimate or not)
- Other attributesof the nodes or ties (e.G., Kin relationships)
- The structure of the nodes’ neighborhood (e.G. Many mutual ‘friends’)

Surveys and interviews allows us to establish the existence of mutual or one-sided strength/affection with greater certainty, but proxies above are also useful

Homophily, Transitivity, and Bridge

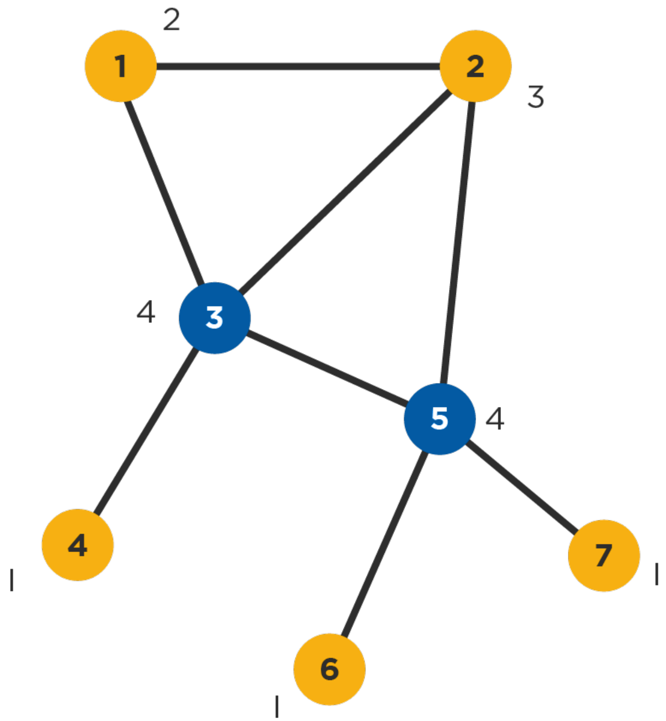


- Homophily is the tendency to relate to people with similar characteristics (status, beliefs, etc.)
- Transitivity in SNA is a property of ties: if there is a tie between A and B and one between B and C, then in a transitive network A and C will also be connected
- Bridges are nodes and edges that connect across groups

Key Player Identification

Who is the Most Important Person?

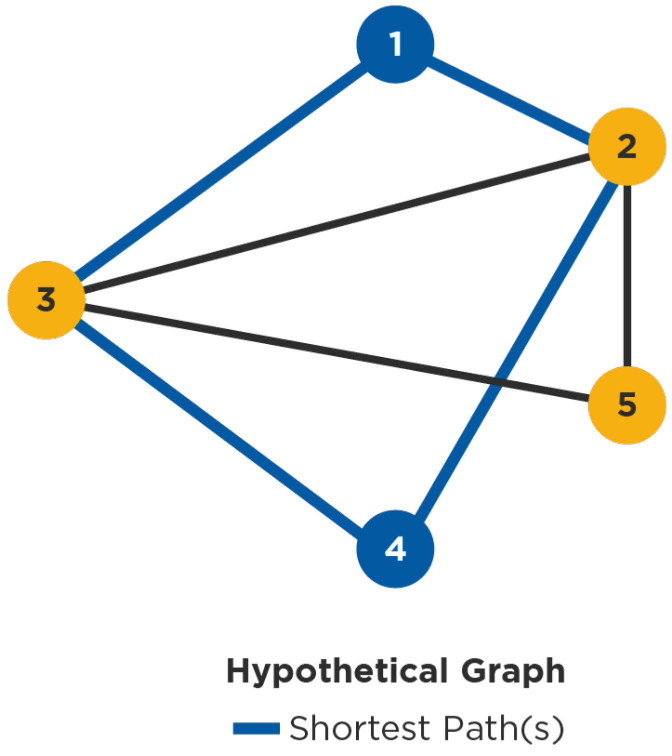
Degree Centrality



Nodes 3 and 5 have the highest degree (4)

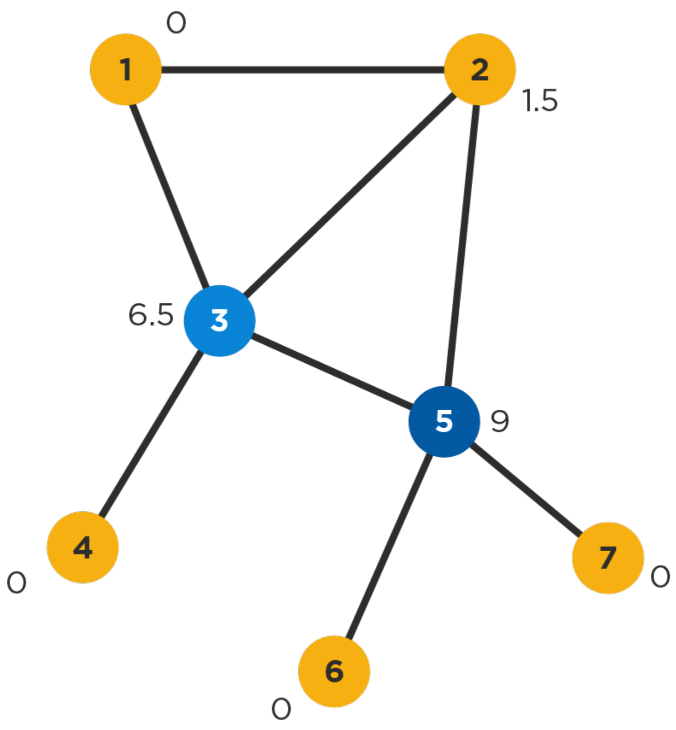
- A node’s (in-) or (out-) degree is the number of links that lead into or out of the node
- In an undirected graph they are of course identical
- Often used as measure of a node’s degree of connectedness and hence also influence and/or popularity
- Useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate ‘neighborhood’

Paths and Shortest Paths



- 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)

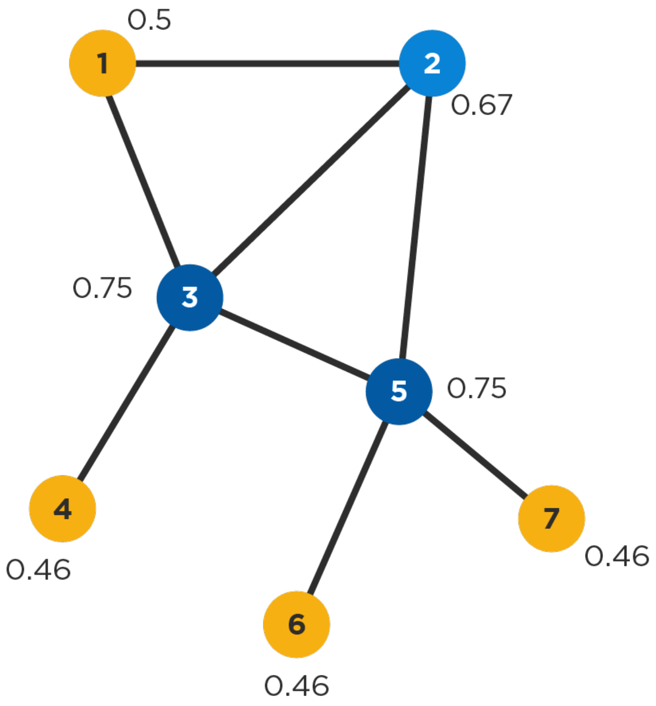
Betweenness Centrality



Node 5 has higher betweenness centrality than 3

- For a given node v, calculate the number of shortest paths between nodes i and j that pass through v, and divide by all shortest paths between nodes i and j
- Sum the above values for all node pairs i,j
- Sometimes normalized such that the highest value is 1 or that the sum of all betweenness centralities in the network is 1
- Shows which nodes are more likely to be in communication paths between other nodes
- Also useful in determining points where the network would break apart (think who would be cut off if nodes 3 or 5 would disappear)

Closeness Centrality

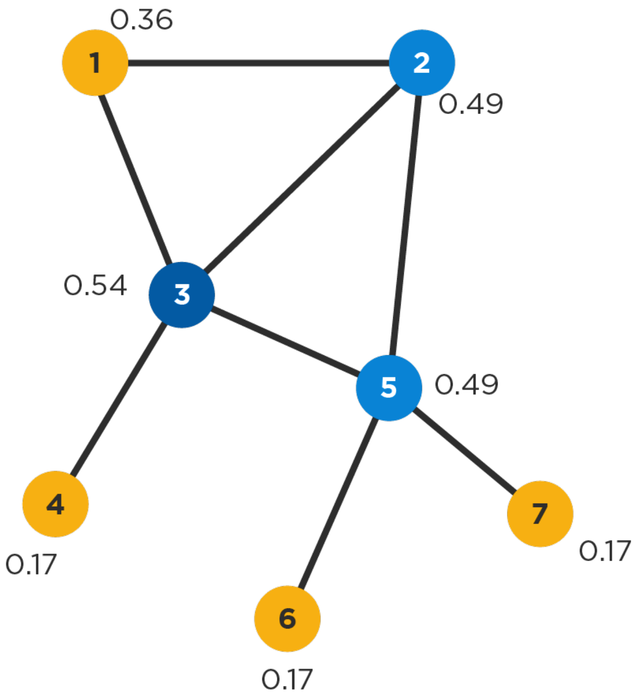


- Calculate the mean length of all shortest paths from a node to all other nodes in the network (i.e. how many hops on average it takes to reach every other node)
- Take the reciprocal of the above value so that higher values are 'better' (indicate higher closeness) like in other measures of centrality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

Note: Sometimes closeness is calculated without taking the reciprocal of the mean shortest path length. Then lower values are 'better'.

Nodes 3 and 5 have the highest (i.e. best) closeness, while node 2 fares almost as well

Eigenvector Centrality



- A node's eigenvector centrality is proportional to the sum of the eigenvector centralities of all nodes directly connected to it
- In other words, a node with a high eigenvector centrality is connected to other nodes with high eigenvector centrality
- This is similar to how Google ranks web pages: links from highly linked-to pages count more
- Useful in determining who is connected to the most connected nodes

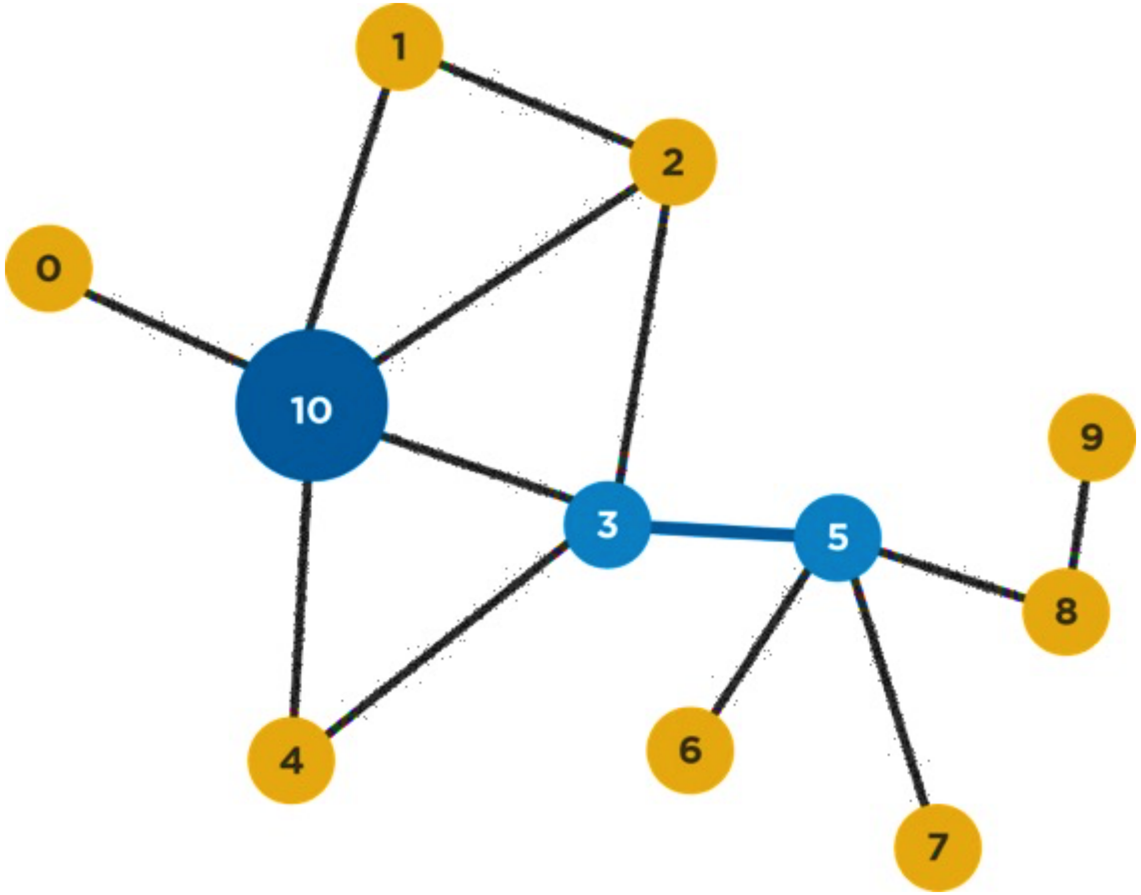
Note: The term 'eigenvector' comes from mathematics (matrix algebra), but it is not necessary for understanding how to interpret this measure

Node 3 has the highest eigenvector centrality, closely followed by 2 and 5

Interpretation of Measures

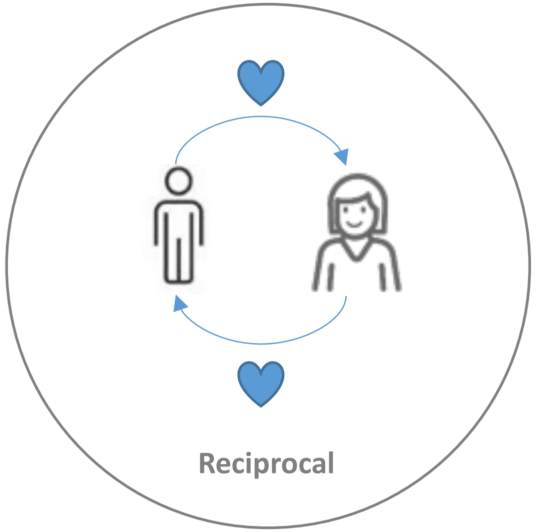
Centrality measure	Interpretation in social networks
Degree	How many people can this person reach directly?
Betweenness	How likely is this person to be the most direct route between two people in the network?
Closeness	How fast can this person reach everyone in the network?
Eigenvector	How well is this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

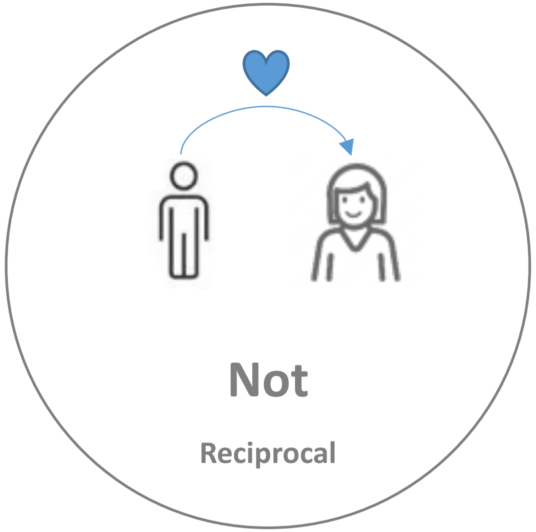
Sets of Key Players



Overall Social Network Structure

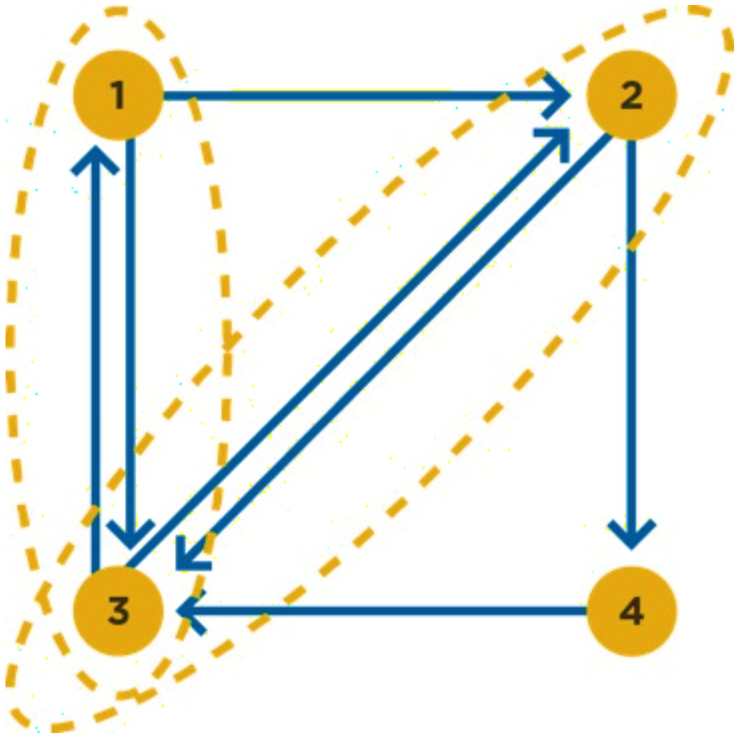
Reciprocity (degree of)





The ratio of the number of reciprocated relations over the total number of relations

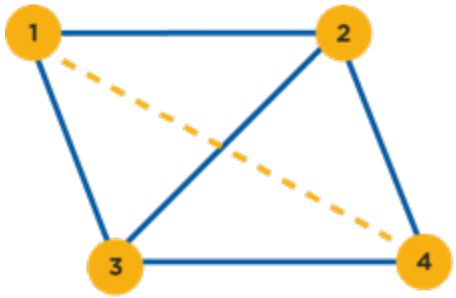
Reciprocity (degree of)



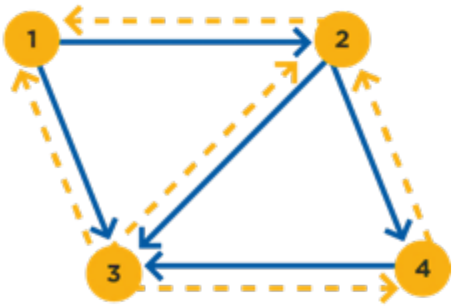
Reciprocity for network = 0.4

- The ratio of the number of relations which are reciprocated (i.e. there is an edge in both directions) over the total number of relations in the network
- …where two vertices are said to be related if there is at least one edge between them
- In the example to the right this would be 2/5=0.4 (whether this is considered high or low depends on the context)
- A useful indicator of the degree of mutuality and reciprocal exchange in a network, which relate to social cohesion
- Only makes sense in directed graphs

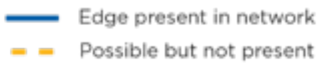
Density



Density → $\frac{5}{6} = 0.83$

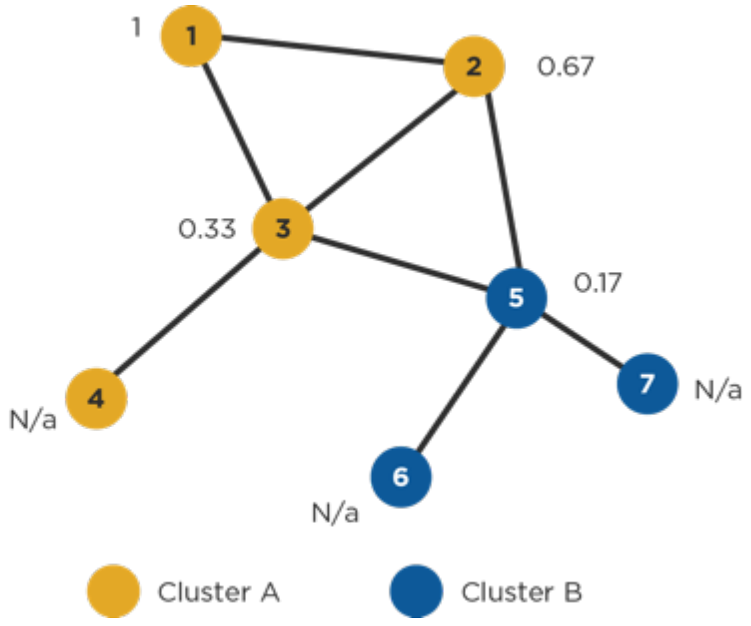


Density → $\frac{5}{12} = 0.42$



- A network's *density* is the ratio of the number of edges in the network over the total number of possible edges between all pairs of nodes (which is *n(n-1)/2*, where *n* is the number of vertices, for an undirected graph)
- In the example network to the right density=5/6=0.83 (i.e. it is a fairly *dense* network; opposite would be a *sparse* network)
- It is a common measure of how well connected a network is (in other words, how closely knit it is) – a perfectly connected network is called a *clique* and has density=1
- A directed graph will have half the density of its undirected equivalent, because there are twice as many possible edges, i.e. *n(n-1)*
- Density is useful in comparing networks against each other, or in doing the same for different regions within a single network

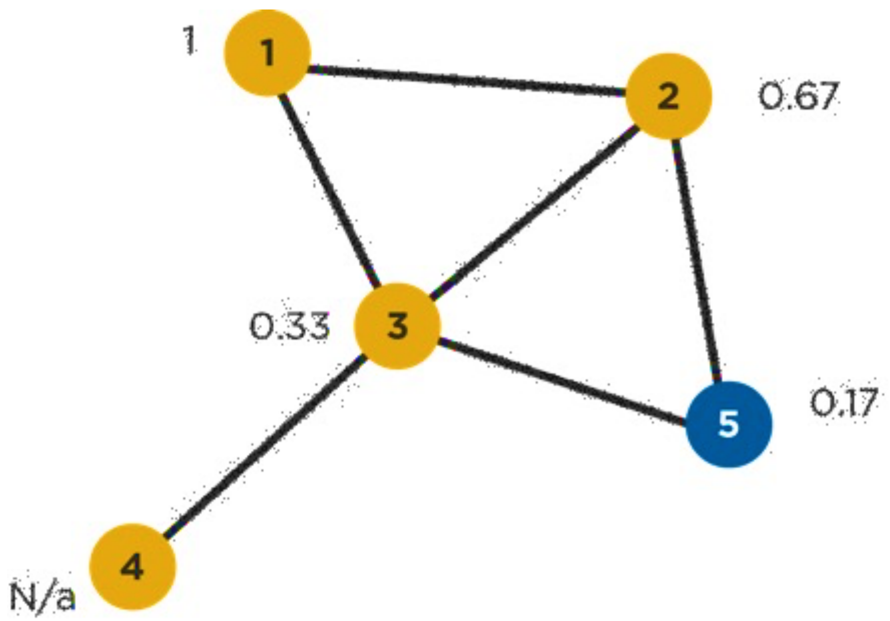
Clustering



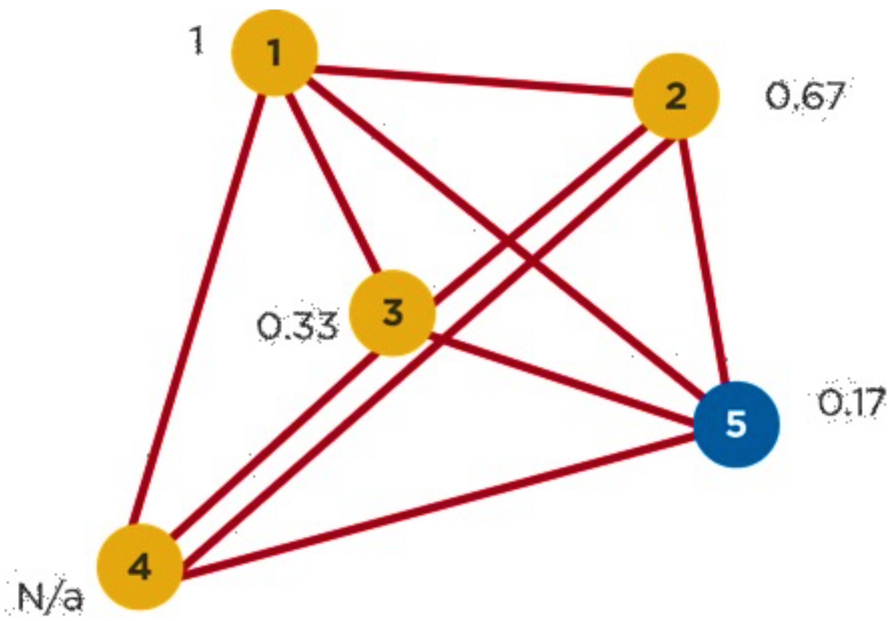
Network clustering coefficient = 0.375
(3 nodes in each triangle x 2 triangles = 6 closed triplets divided by 16 total)

- A node's *clustering coefficient* is the number of closed triplets in the node's neighborhood over the total number of triplets in the neighborhood. It is also known as *transitivity*.
- E.g., node 1 to the right has a value of 1 because it is only connected to 2 and 3, and these nodes are also connected to one another (i.e. the only triplet in the neighborhood of 1 is closed). We say that nodes 1,2, and 3 form a *clique*.
- Clustering algorithms identify clusters or 'communities' within networks based on network structure and specific clustering criteria (example shown to the right with two clusters is based on *edge betweenness*, an equivalent for edges of the betweenness centrality presented earlier for nodes)

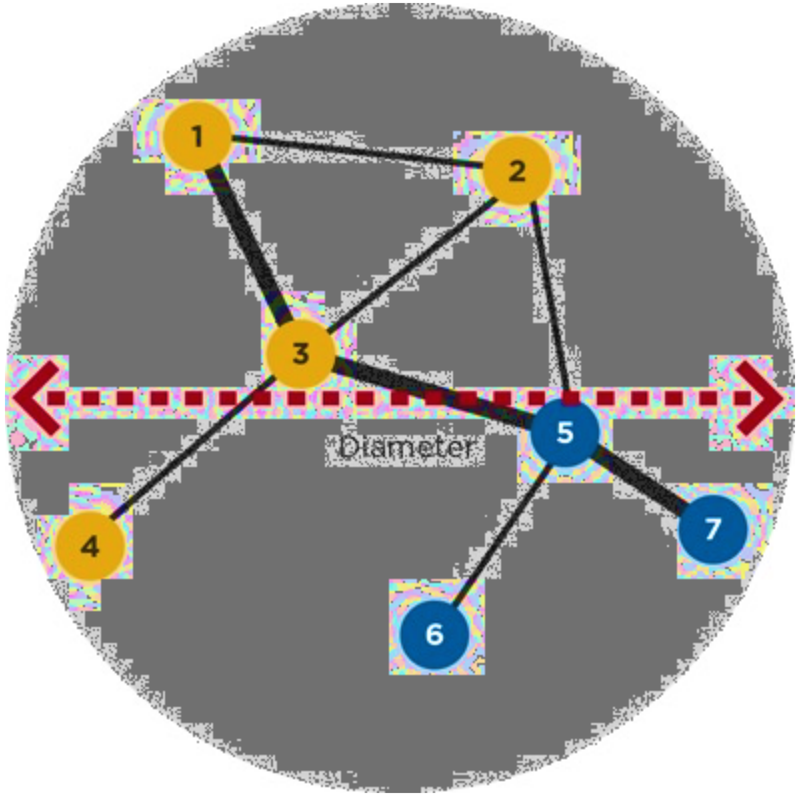
Triplets of Node 3



Possible Number of Triplets



Average and Longest Distance



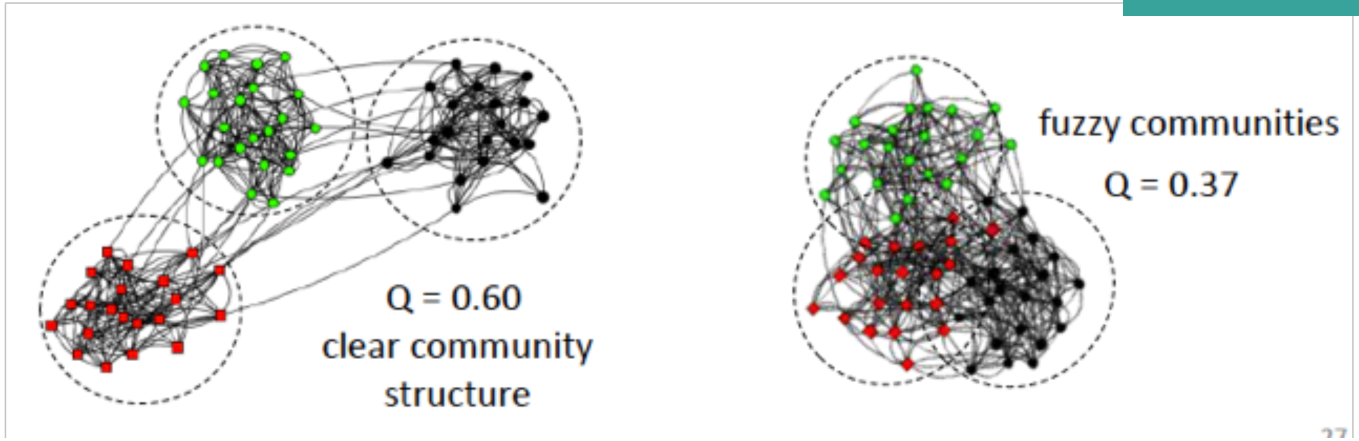
- The longest shortest path (distance) between any two nodes in a network is called the network's *diameter*
- The diameter of the network on the right is 3; it is a useful measure of the *reach* of the network (as opposed to looking only at the total number of vertices or edges)
- It also indicates how long it will take at most to reach any node in the network (sparser networks will generally have greater diameters)
- The average of all shortest paths in a network is also interesting because it indicates how far apart any two nodes will be on average (average *distance*)

Modularity

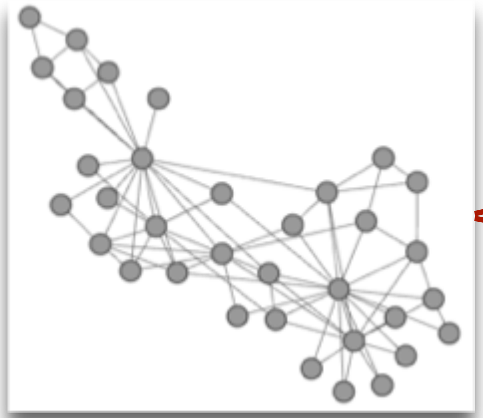
$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j)$$

Edges inside the community

Expected number of edges if i,j places at random



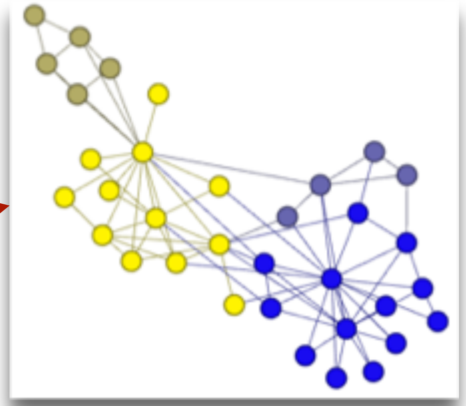
Metric Simulation



model network

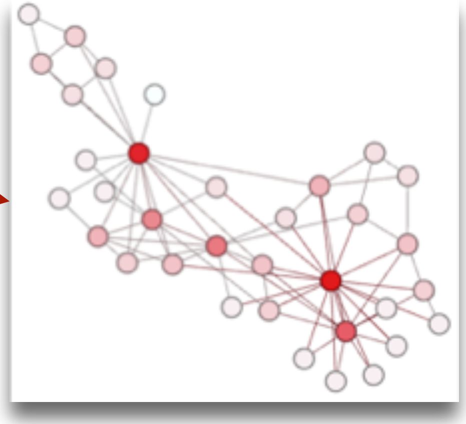
modularity

community detection



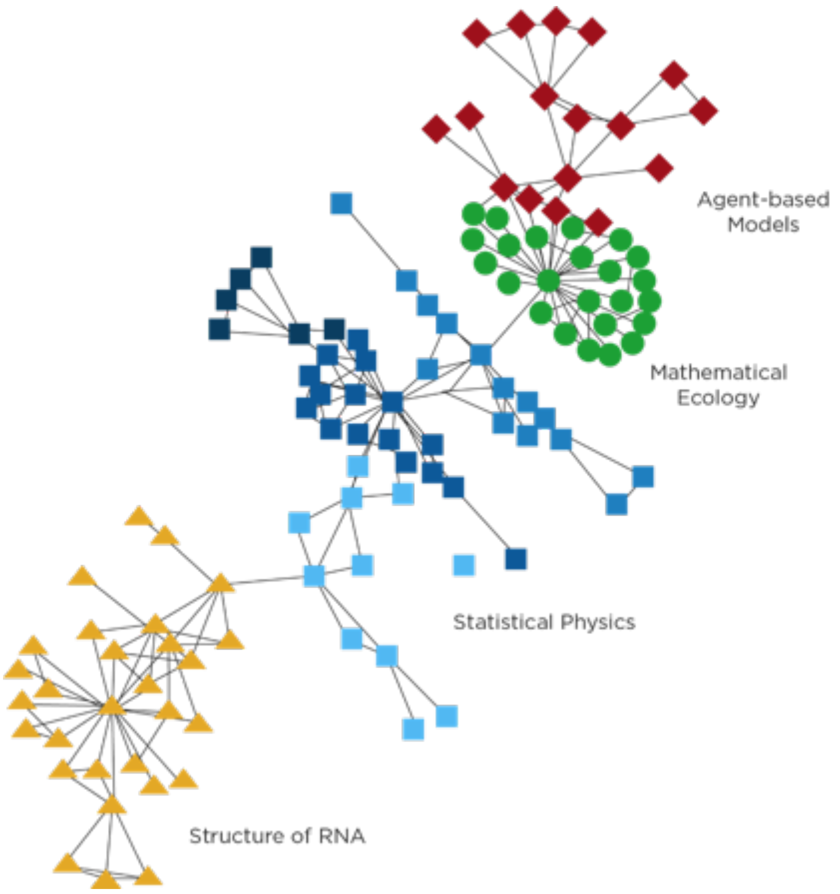
community detection result

centrality



degree centrality result

Example : Finding Community



- Collaboration network of scientist at Santa Fe Institut (Girvan & Nirwan)
- 27| scientist (vertices) / 1 | 8 nodes from largest component edge = scientist coauthor one of more publications
- Komunitas : kumpulan titik titik dimana jumlah hubungan internal antar titik lebih besar dari pada jumlah hubungan dengan titik eksternal

I just need my friend,

friend of my friend, friend of friend of my friend, friend of friend of friend of my friend, friend of friend of friend of friend of my friend, friend of friend of friend of friend of friend of my friend,

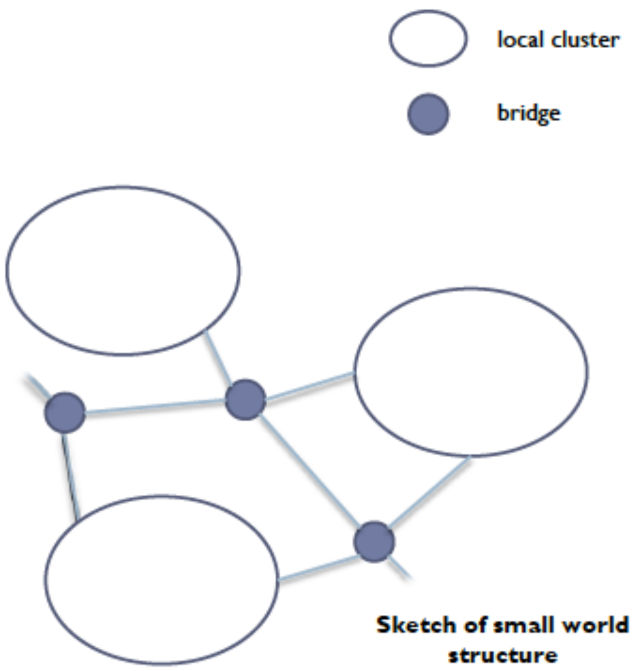
to contact him

This is Dian

This is Peter Parker

Small Worlds

- A small world is a network that looks almost random but exhibits a significantly high clustering coefficient(nodes tend to cluster locally) and a relatively short average path length (nodes can be reached in a few steps
- It is a very common structure in social networks because of transitivity in strong social ties and the ability of weak ties to reach across clusters
- Such a network will have many clusters but also many bridges between clusters that help shorten the average distance between nodes



Sketch of small world structure

You may have heard of the famous “6 degrees” of separations

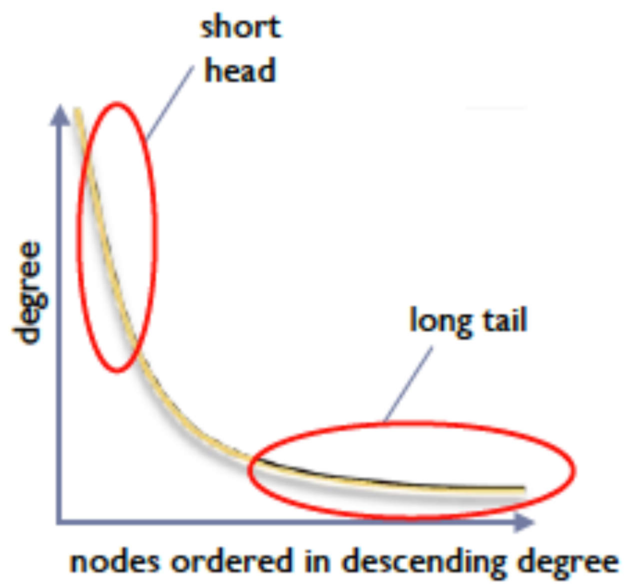
Preferential Attachment

A property of some networks, where, during their evolution and growth in time, a the great majority of new edges are to nodes with an already high degree; the degree of these nodes thus increases disproportionately, compared to most other nodes in the network

- The result is a network with few very highly connected nodes and many nodes with a low degree
- Such networks are said to exhibit a long-tailed degree distribution
- And they tend to have a small-world structure!

(it turns out, transitivity and strong/weak tie characteristics are not neccessary to explain small world structures, but they are common and can also lead to such structures)

Example of network with preferential attachment



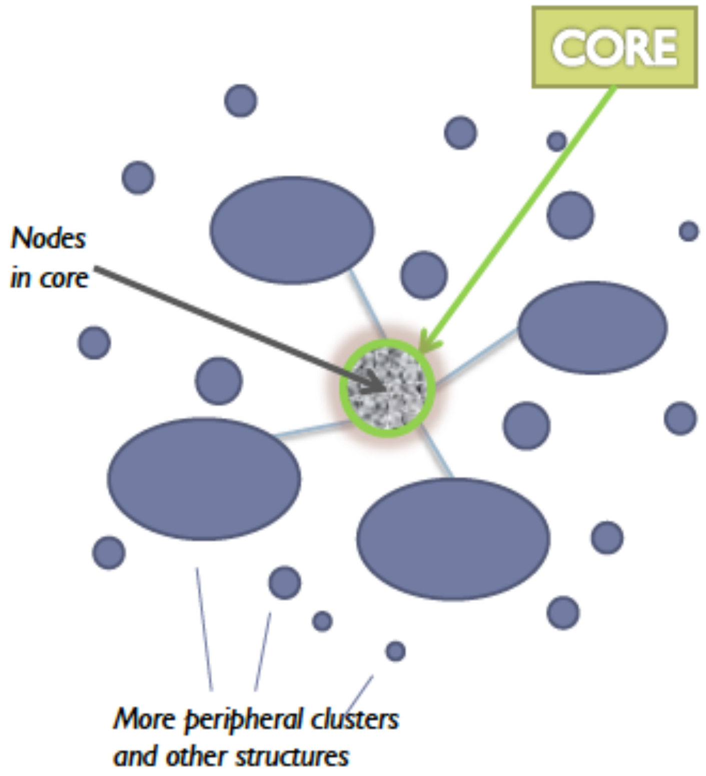
Skectch of long-tailed degree distributions

Reasons for Preferential Attachment

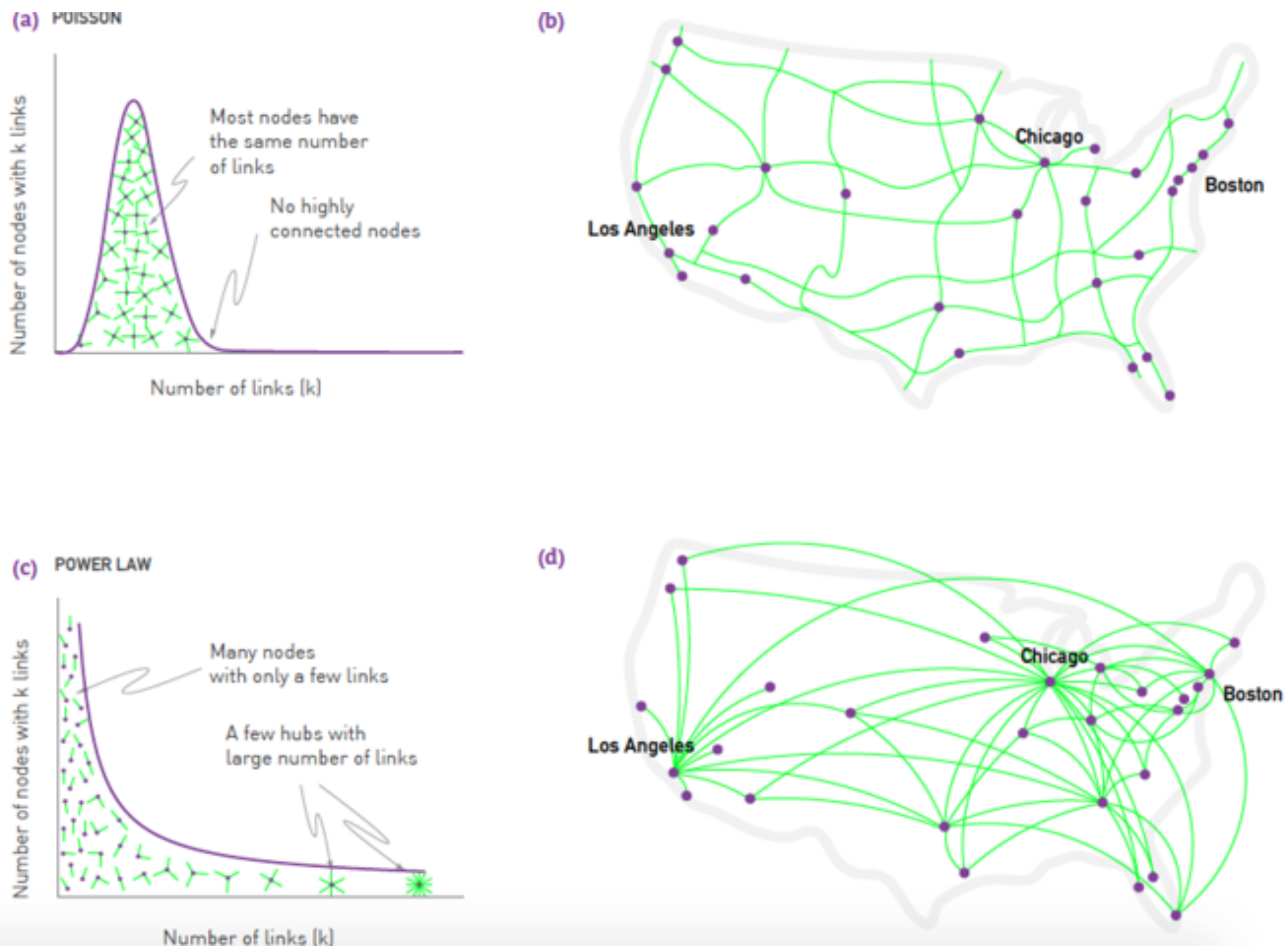
POPULARITY	QUALITY	MIXED MODEL
We want to be associated with popular people, ideas, items, thus further increasing their popularity, irrespective of any objective measurable characteristics	We evaluate people and everything else based on objective quality criteria, so higher quality nodes will naturally attract more attention faster	Among nodes of similar attributes, those that reach critical mass first will become 'star' with many friends and followers ('halo effect')
also known as 'the rich get richer'	also known as 'the good get better'	may be impossible to predict who will become a star, even if quality matters

Core-Periphery Structures

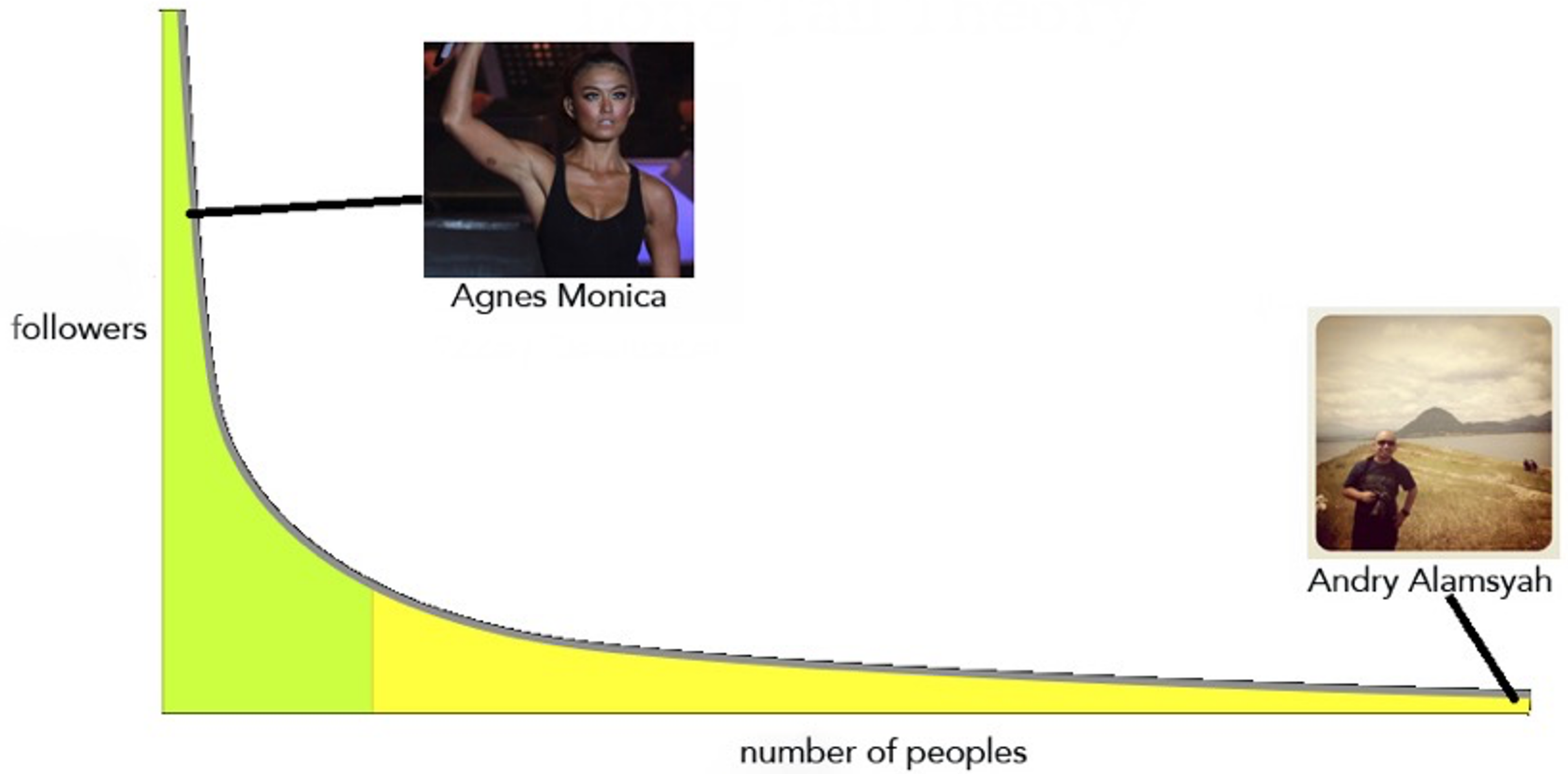
- A useful and relatively simple metric of the degree to which a social network is centralized or decentralized, is the centralization measure (usually normalized such that it takes values between 0 and 1)
  - It is based on calculating the differences in degrees between nodes; a network that greatly depends on 1-2 highly connected nodes (as a result for example of preferential attachment) will exhibit greater differences in degree centrality between nodes
  - Centralized structures can perform better at some tasks (like team-based problem-solving requiring coordination), but are more prone to failure if key players disconnect
- In addition to centralization, many large groups and online communities have a core of densely connected users that are critical for connecting a much larger periphery
  - Cores can be identified visually, or by examining the location of high-degree nodes and their joint degree distributions (do high-degree nodes tend to connect to other high-degree nodes?)
  - Bow-tie analysis, famously used to analyze the structure of the Web, can also be used to distinguish between the core and other, more peripheral elements in a network



Social Network Characteristic



Social Network Characteristic: Power Law



power law distributions

Thank You

Dian Ramadhani