MoF-DAC
Ministry of Finance
Data Analytics Community

SOCIAL NETWORK ANALYSIS

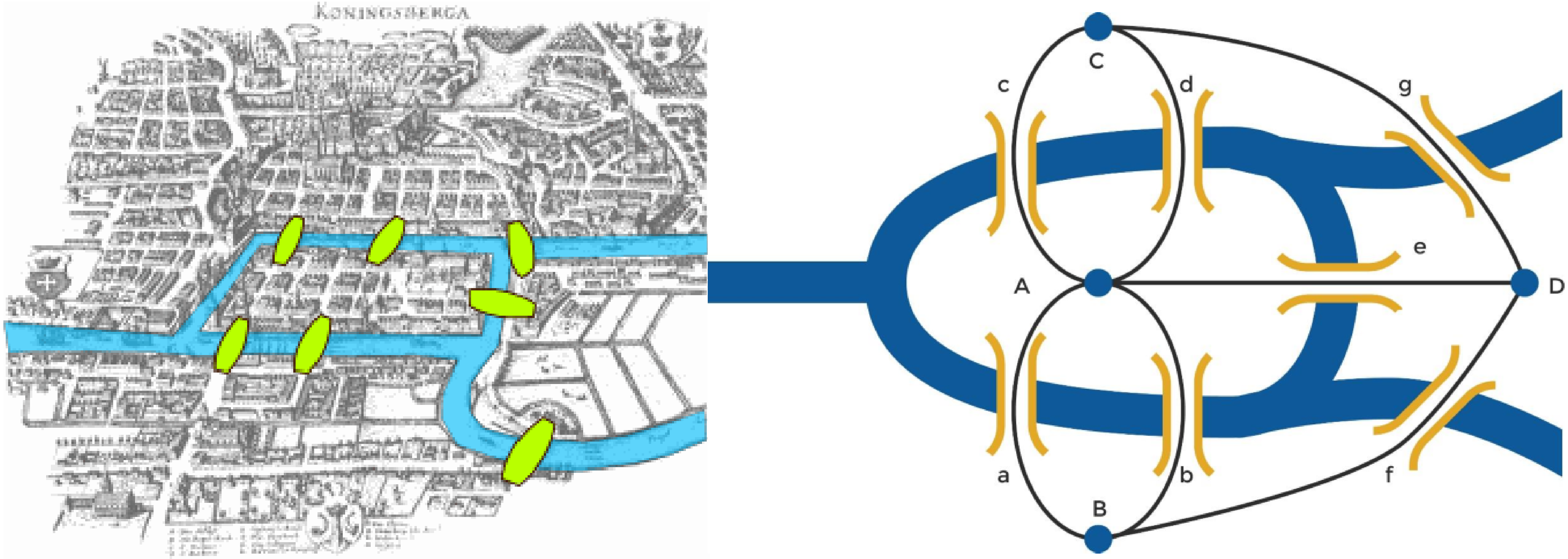
Ade Satya Wahana
Aris Budi Santoso
Leonard Yulianus

Social Network Definition

- A social network is a social structure made up of a set of social actors (such as individuals or organizations), sets of dyadic ties, and other social interactions between actors.
- The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures.
- The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics.

(Wikipedia)

Network Analysis



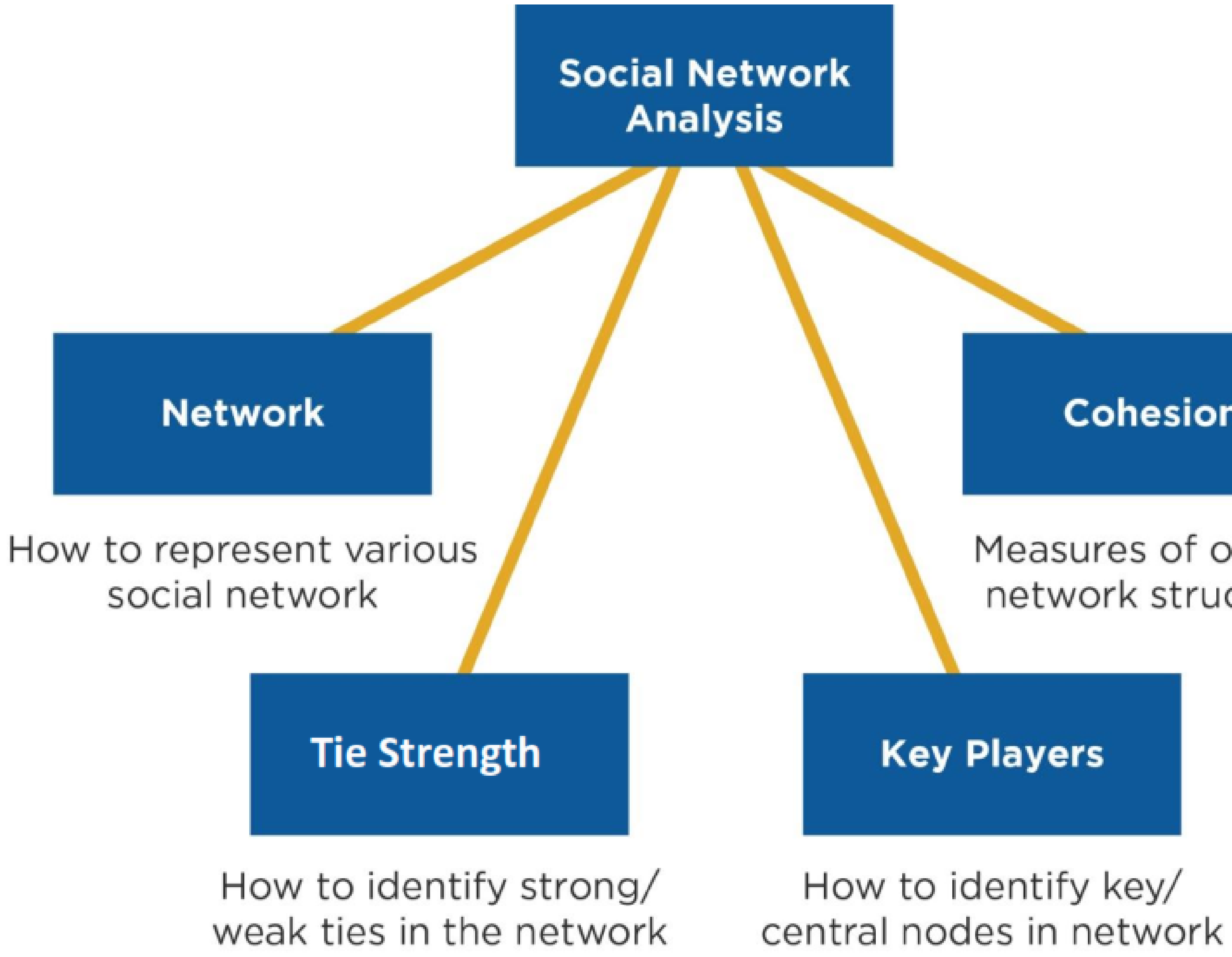
Konigsberg Bridge

The problem was to devise a walk through the city that would cross each of those bridges once and only once.
Proven to have no solution and became foundation of graph theory

Network Analysis History

- SNA origins come from social science and network analysis (graph theory)
- Network analysis concerns with the formulation and solution of problems that have a network structure; such structure is usually captured in a graph
- Graph theory provides a set of abstract concepts and methods for the analysis of graphs. These, in combination with other analytical tools and with methods for the visualization and analysis of social networks, form the basis of what we call SNA methods.
- SNA is not just a methodology; it is a unique perspective on how society functions. Instead of focusing on individuals and their attributes, it centers on relations between individuals, groups, or social institutions

SNA Basic Concept



- Graph Representation

Present Relations as Graph Network

Communication

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

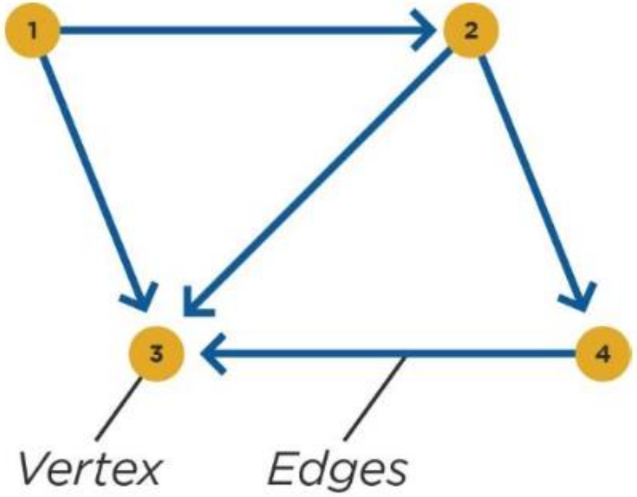
Jim : Mary, you and your dad should come for dinner

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

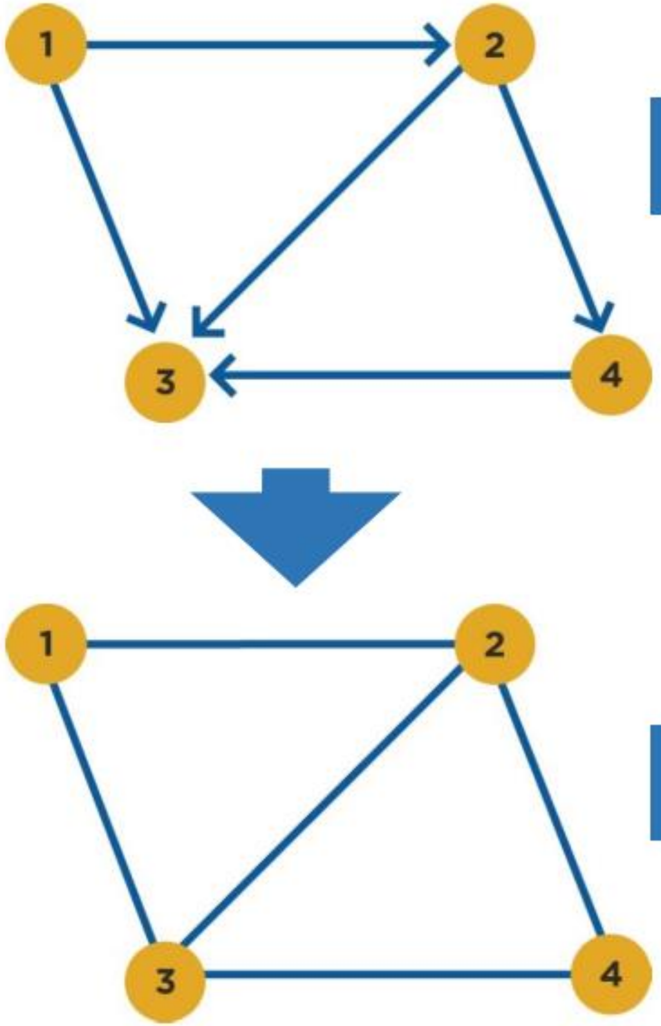
Can we study their interactions as a network ?



Nodes

Relations

Types of Graphs



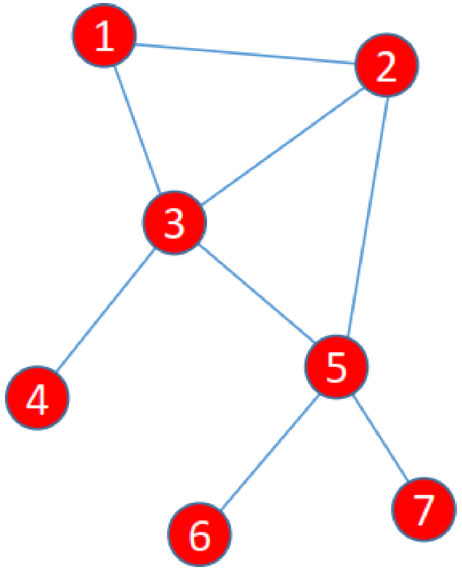
Directed
(who contact whom)

Twitter Follower / Following

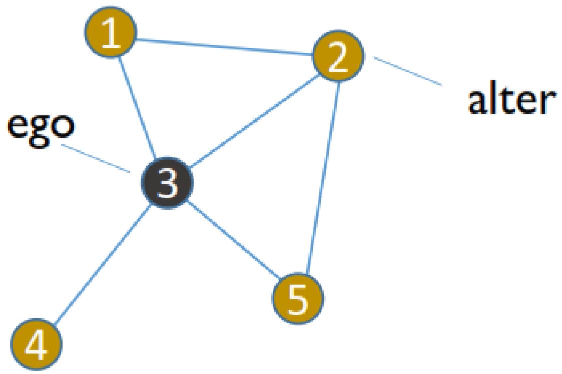
Undirected
(who knows whom)

Facebook Friends

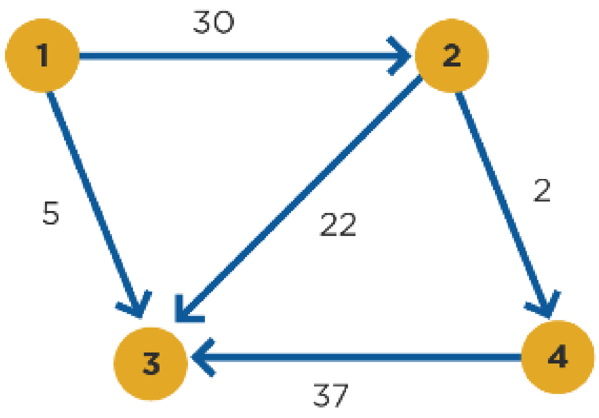
'whole' network*



Ego Network



Tie Strength



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

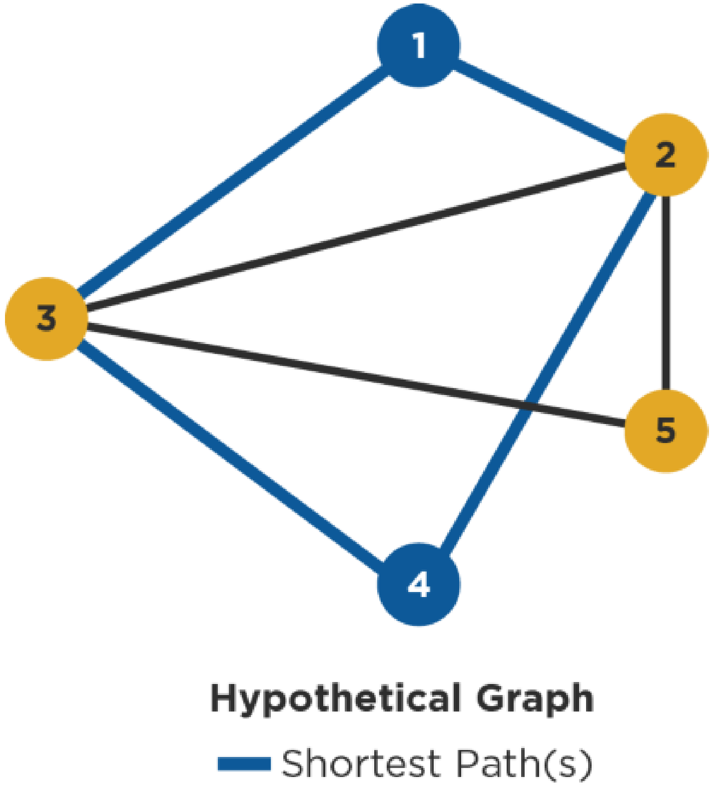
Edge List

Vertex	Vertex	Weight
1	2	30
1	3	5
2	3	22
2	4	2
4	3	27

***Adjacency Matrix* (Weight)**

Vertex	1	2	3	4
1	-	30	5	0
2	30	-	22	2
3	5	22	-	37
4	0	2	37	-

Path & Shortest Path



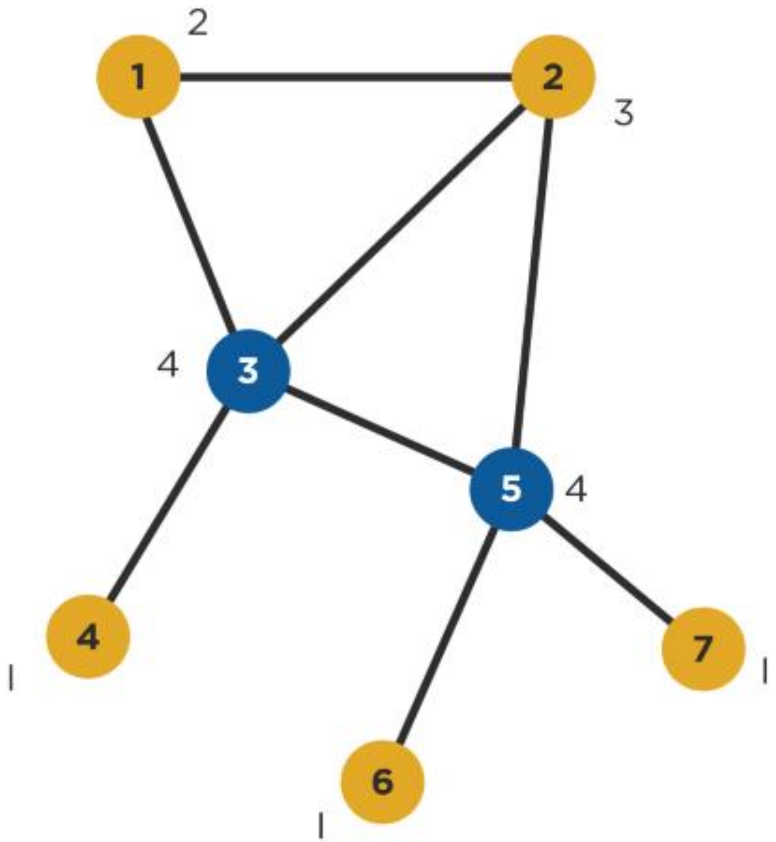
- A path between two nodes is any sequence of non-repeating nodes that connects the two nodes
- The shortest path between two nodes is the path that connects the two nodes with the shortest number of edges (also called the distance between the nodes)
- In the example to the right, between nodes 1 and 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)

- Network Analytic Measures

Node Centrality Measures

- Ukuran dalam *network analysis* yang digunakan untuk menentukan node yang paling penting dalam sebuah *network/ graph*
- Ukuran Centrality :
  - Degree Centrality
    - Angka jumlah banyak koneksi dari/ke sebuah node
  - Closeness Centrality
    - Nilai yang menunjukan rata-rata shortest path suatu node dengan seluruh node pada network
  - Betweeness Centrality
    - Nilai jumlah berapa kali node masuk pada shortest path antar seluruh node
  - Eigenvector Centrality
    - Metrik yang menunjukan seberapa terkoneksi sebuah node dengan node penting lain

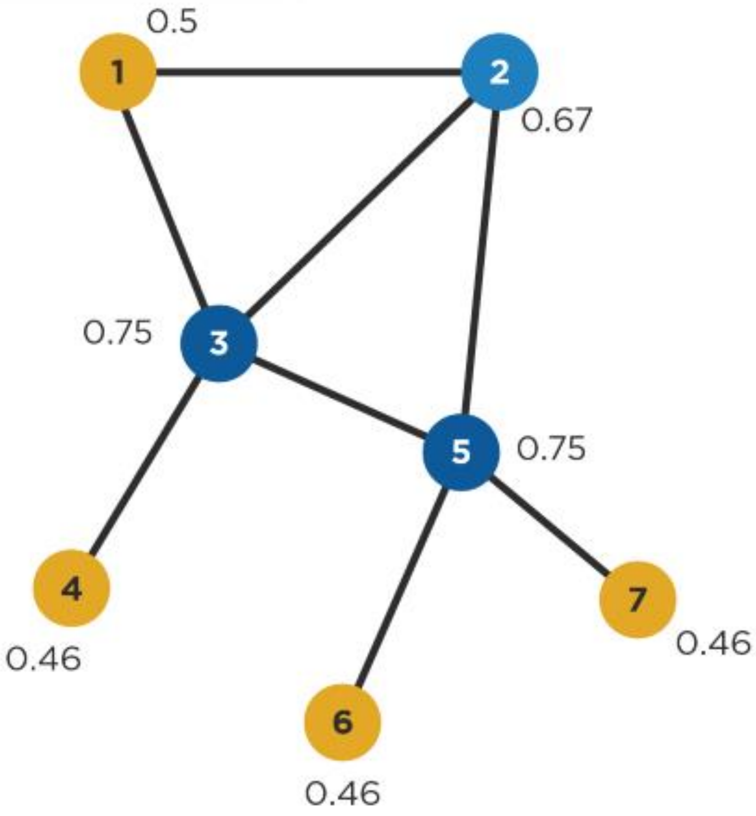
Degree



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'

Closeness

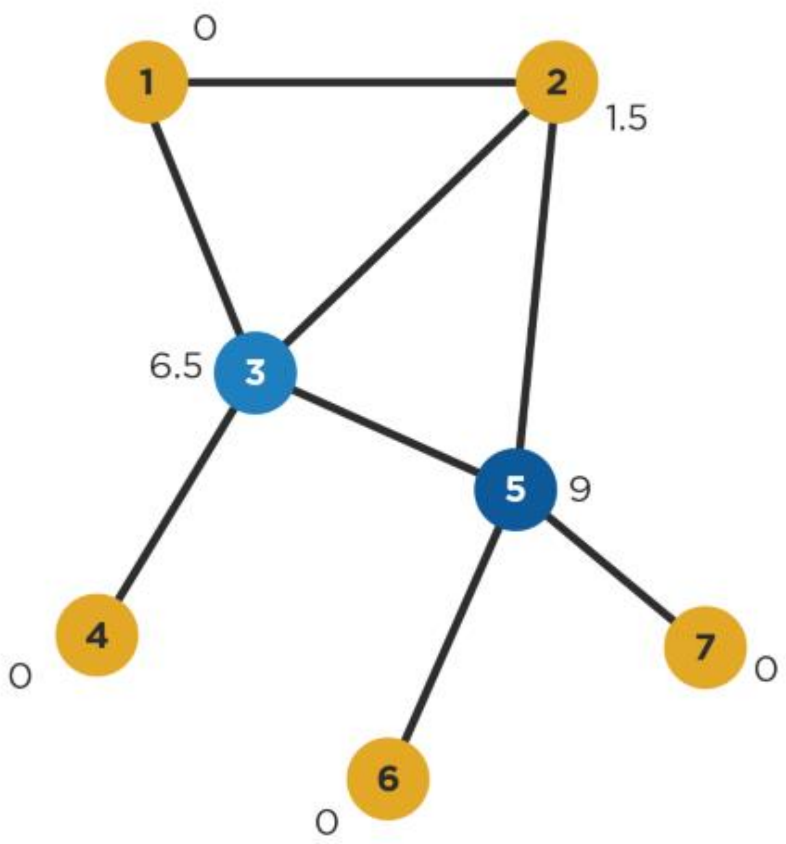


Note: Sometimes closeness is calculated without taking the reciprocal of the mean shortest path length. Then lower values are 'better'.

- 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

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

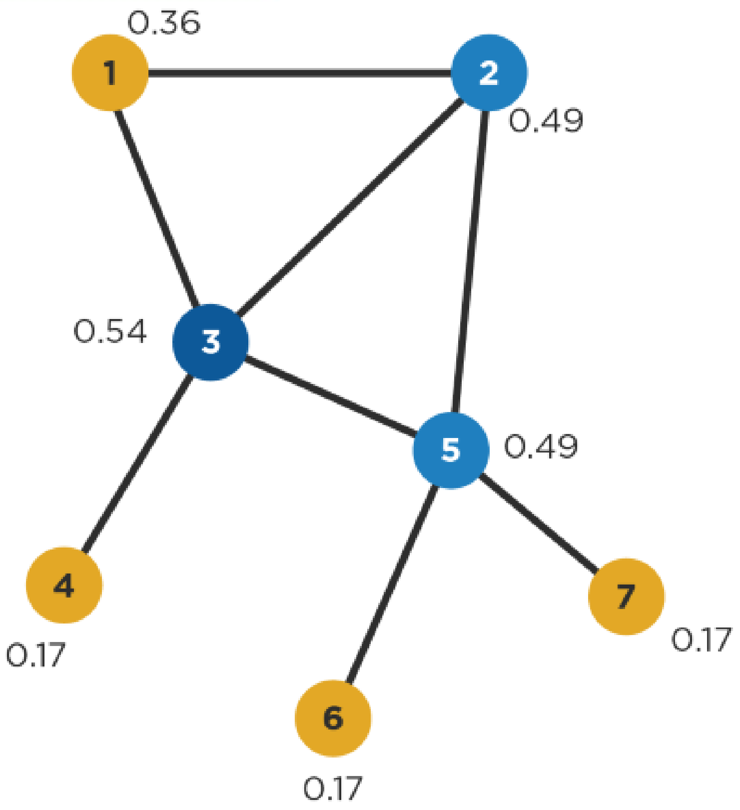
Betweenness



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)

Eigenvector



- 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

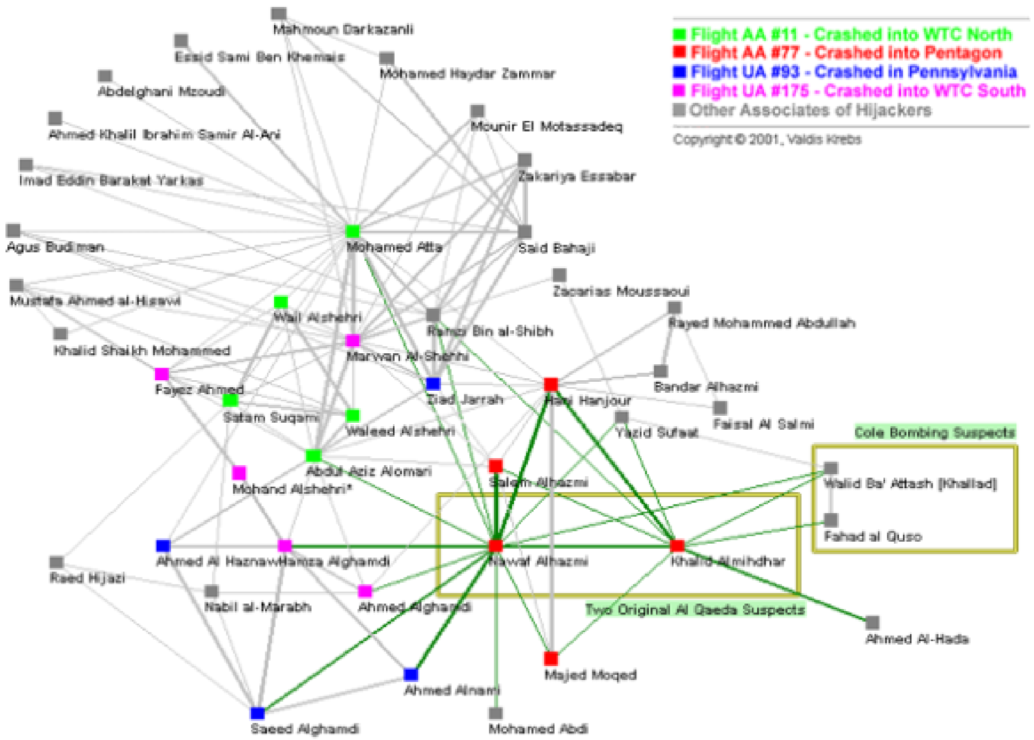
Note: The term 'eigenvector' comes from mathematics (matrix algebra), but it is not necessary for understanding how to interpret this measure

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

Centrality Interpretation

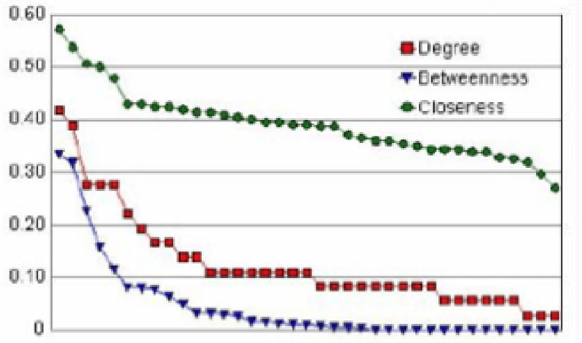
Centrality measure	Interpretation in social networks	Other possible interpretations…
Degree	How many people can this person reach directly?	In network of music collaborations: how many people has this person collaborated with?
Betweenness	How likely is this person to be the most direct route between two people in the network?	In network of spies: who is the spy though whom most of the confidential information is likely to flow? The JBs. [1]
Closeness	How fast can this person reach everyone in the network?	In network of sexual relations: how fast will an STD spread from this person to the rest of the network?
Eigenvector	How well is this person connected to other well-connected people?	In network of paper citations: who is the author that is most cited by other well-cited authors?

Application Example



Group Size	37
Potential Ties	1332
Actual Ties	170
Density	13%

Geodesics	
length	#
1	170
2	626
3	982
4	558
5	136
6	0



Degrees		Betweenness		Closeness	
0.417	Mohamed Atta	0.334	Nawaf Alhazmi	0.571	Mohamed Atta
0.389	Marwan Al-Shehhi	0.318	Mohamed Atta	0.537	Nawaf Alhazmi
0.278	Hani Hanjour	0.227	Hani Hanjour	0.507	Hani Hanjour
0.278	Nawaf Alhazmi	0.158	Marwan Al-Shehhi	0.500	Marwan Al-Shehhi
0.278	Ziad Jarrah	0.116	Saeed Alghamdi*	0.480	Ziad Jarrah
0.222	Ramzi Bin al-Shibh	0.081	Hamza Alghamdi	0.429	Mustafa al-Hisawi

paper: Mapping Network of Terrorist Cells - V.E.Krebs (2002)

Network Measures

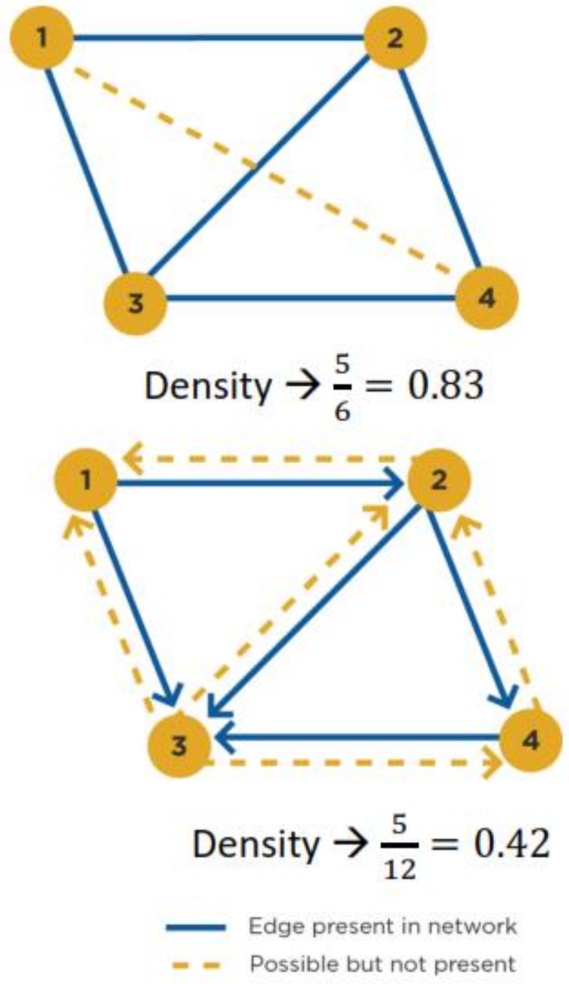
Network Density

Reciprocity

Network Diameter

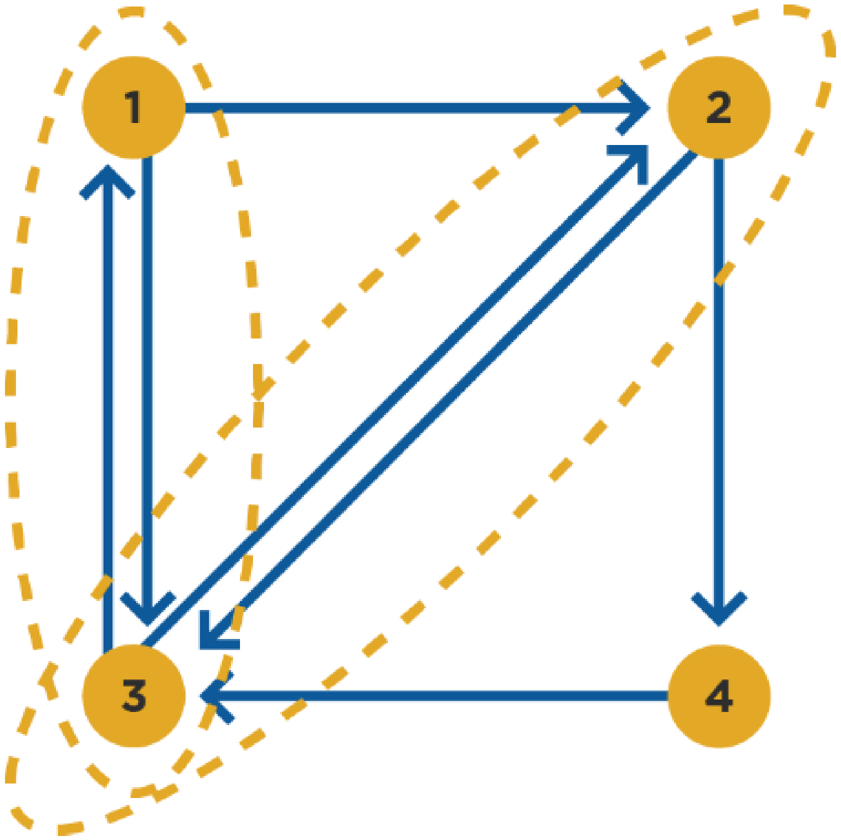
Clustering Coefficient

Density



- 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

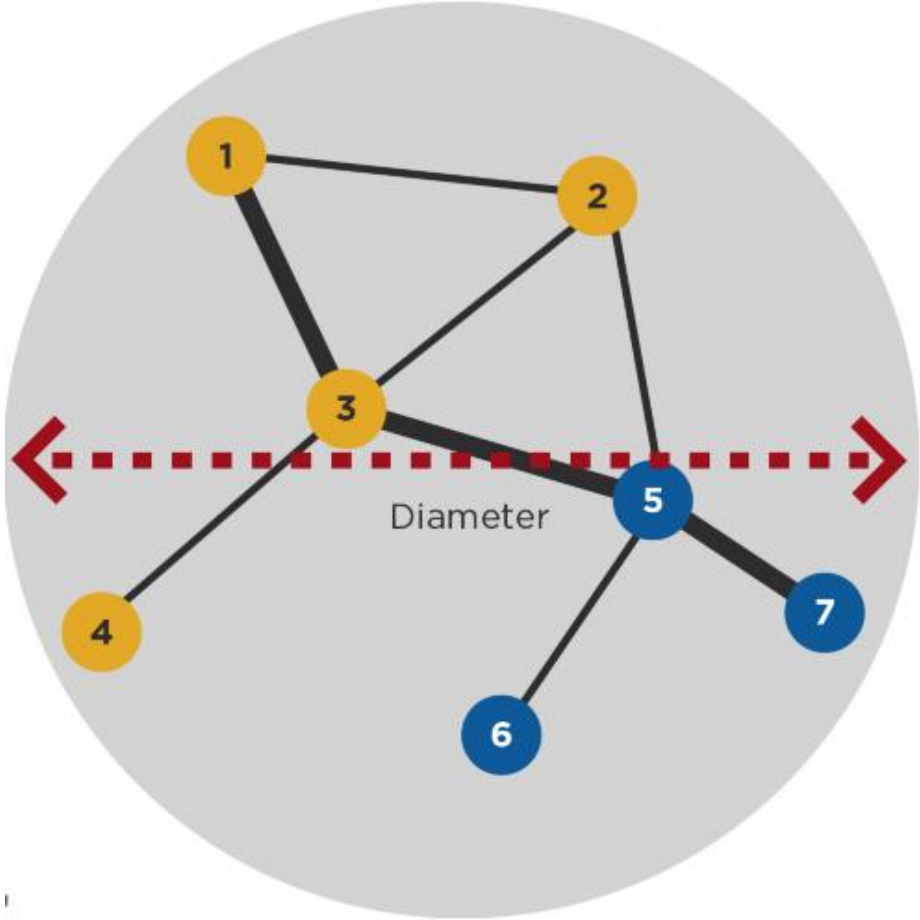
Reciprocity



Reciprocity for network = 0.4

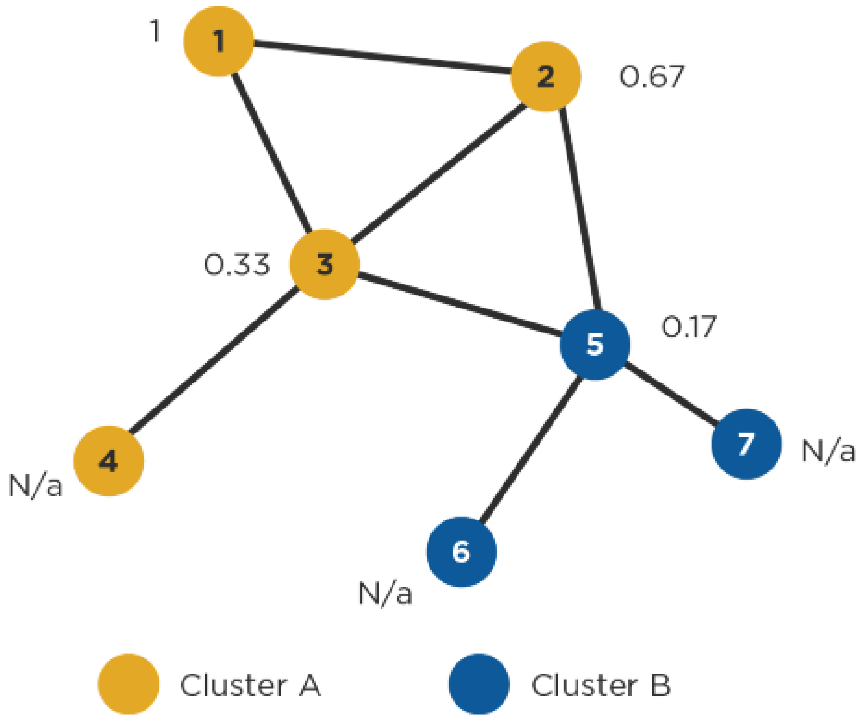
- The ratio of the number of relations which are reciprocated (i.e. there is an edge in both directions) over the total number of relations in the network
- ...where two vertices are said to be related if there is at least one edge between them
- In the example to the right this would be 2/5=0.4 (whether 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

Network Diameter



- 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*)

Clustering Coefficient

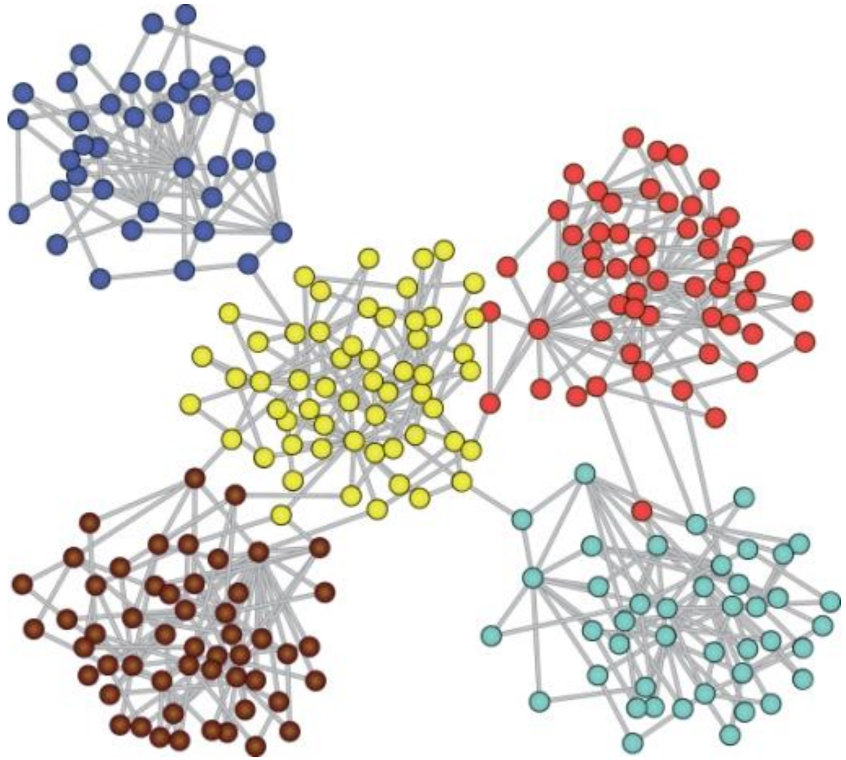


Network clustering coefficient = 0.375
(3 nodes in each triangle x 2 triangles = 6 closed triplets divided by 16 total)

- 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)

Community Detection

- Community adalah kumpulan individu yang memiliki interaksi yang tinggi
  - Interaksi antar individu dalam komunitas tinggi
  - Interaksi dengan individu di luar komunitas rendah
- Community Detection adalah cara untuk menemukan kelompok-kelompok dalam sebuah jaringan
- Algoritma
  - Hierarchical (Divisive)
    - Girvan Newman Method
  - Modular
    - Louvain



CONTOH PENERAPAN PADA PYTHON

TERIMA KASIH