Data Analysis & Visualization

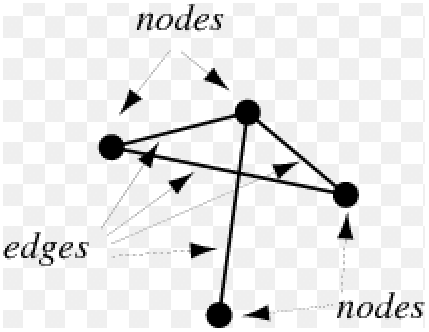
LVC 2: Network Analysis

Network analysis is a technique that uses **graph theory** to study complex real-life problems like marketing, neuroscience, computational biology, etc. In real-life scenarios, there are problems that are quite easily solvable by traditional methods but in the case of problems where a multitude of inter-connections are present between the existing entities, it becomes quite necessary to take the help of network analysis.

Network analysis enables visualization, interpretation, and understanding of the relation and flow of information or signals. Using graph theory, network analysis is capable of solving such problems in a more efficient and organized manner in comparison to traditional methods. As an effect, network analysis has found extensive use in the modern era. Let us begin with understanding the **network** and its **components**.

Network

A network is a **collection of nodes and edges** that are interconnected with each other. It operates by flowing some signal or matter through the edges between the nodes. A network builds a complete representation of the entire transmission process. The below picture depicts a sample network where nodes and edges are shown clearly.



Node: A node is a vertex in the network. It represents the elements of the network and can be a person, a place, or an object. A node is generally associated with some features that describe the characteristics of that node and the features also help in developing a connection with a specific node that possesses the same or similar features. For example, considering Facebook as a network, the people having an account on Facebook are the nodes while the message services, friend requests, etc. can be used to develop connection between two persons / nodes. A person on Facebook does have many features (profile and friends related) associated with him / her.

Edge: An edge is a connection between two nodes that represents the transmission of a signal, object, or information from one node to another. A flight passing from one airport to another (one node to another) is creating an edge between the nodes.

In **symbolic** terms, a network is represented as **G(V, E),** where **G** stands for **Graph** or network, **V** stands for vertices or nodes, and **E** stands for edges or connections between the nodes.

Let us use a **few examples** to clearly understand this.

1. Facebook is a network of people throughout the world. People who are having an account on Facebook are the nodes of the network while an edge is a connection between people through friendship.
2. The Internet is a network where nodes are the computers, routers, etc. while the edge is the signal path that works as a communication channel between the nodes.
3. In neural networks, neurons are the nodes while synapses are the edges between the neurons.
4. In airline transport, airports are the nodes while the flights connecting them are the edges.
5. In a power grid, a substation can be considered a node while the transmission lines are the edges between the nodes.

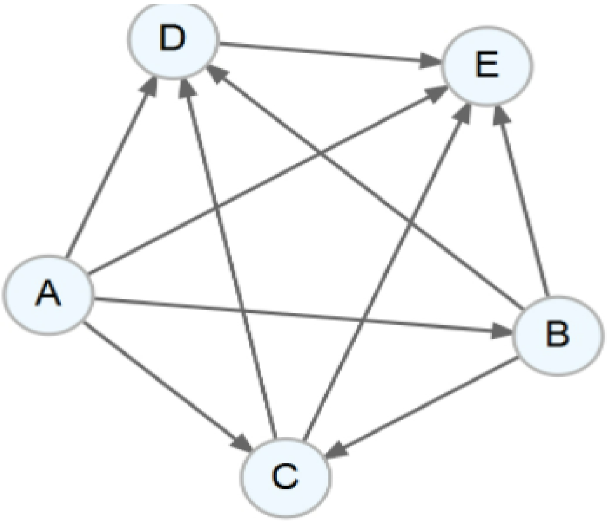
Below are some other examples of networks with **vertices and edges**.

Network	**Vertex**	**Edge**
World Wide Web	Web page	Hyperlink
Gene regulatory network	Gene	Regulatory effect
Food web	Species	Who-eats-who
Phylogenetic tree	Species	Evolution
Netflix	Person / movie	Rating

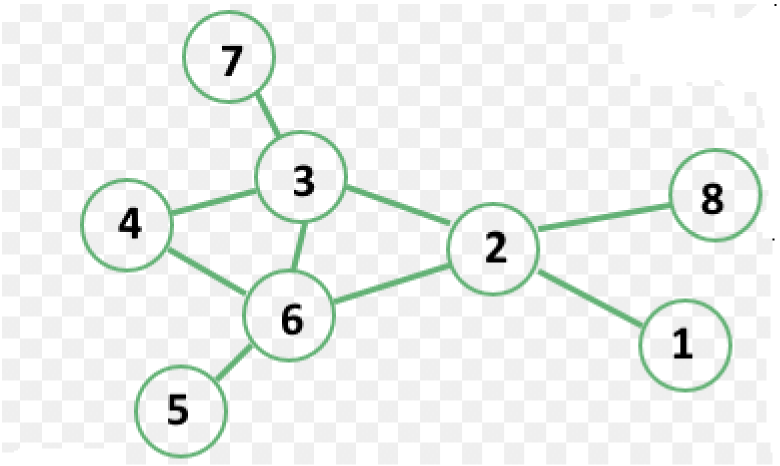
Based on different types of real-life applications and in consideration of the corresponding requirements, networks can be of many different types. Below is a detailed view of the classification of networks.

Classification of Networks

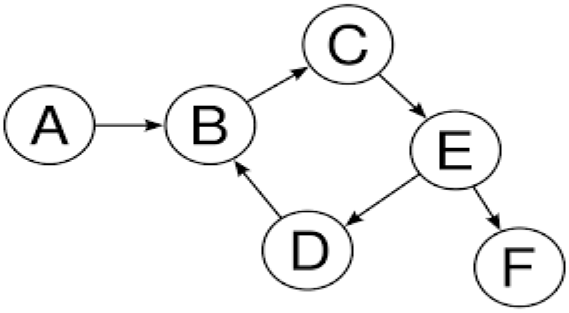
1. **Directed and Undirected Networks:** This is a distinction of networks based on whether the flow between the nodes is occurring in a unique direction or in an unspecified direction.
    a. **Directed Networks:** These are networks where the transmission / flow between the nodes occurs in a **specific direction**. For example, the food chain of animals is a directed network. If a cat eats a rat, then the reverse is not possible (the rat cannot eat the cat). In the below plot A, B, C, D, and E are the nodes that are connected with directed edges.



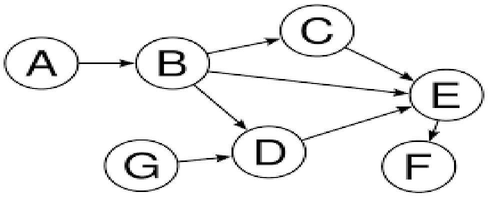
    b. **Undirected Networks:** In these networks, the transmission / flow between the nodes does **not** occur in a fixed direction. If a transaction occurs from node A to B, then it may occur from node B to A as well. In the transmission network, a vehicle can go from node A to node B but the reverse is also possible, i.e. vehicles can go from B to A as well. So, the direction of transmission is **not fixed,** hence, it is an undirected network. Another example is social media networks because the capability of sending a message (rather than who sent the message to whom) can go either way - if Person A and Person B are connected on a social network, either of them can send a message to the other by default. In the below figure nodes are connected to each other but the edges are not in a specific direction between any two nodes. The transaction is possible both ways.



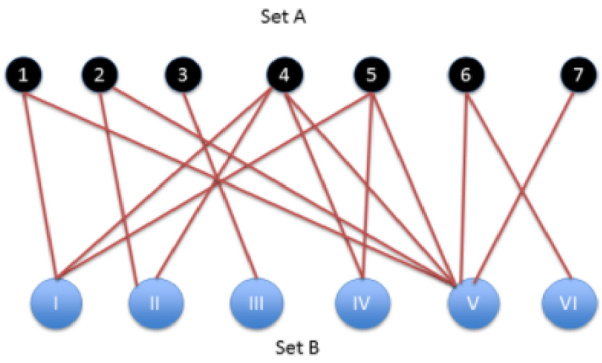
2. **Cyclic and Acyclic Networks:** This is a distinction based on whether the network makes a complete cycle or not. A cycle is a structure formed by starting and ending at the same node in the network.
    a. **Cyclic Networks:** These are the networks that make a complete cycle of nodes and edges. For example, Gene networks are cyclic networks. The nodes in the below graph are clearly making a directed cycle (B-C-E-D-B). Hence, it is a cyclic network.



    b. **Acyclic Networks**: These are the networks that do not contain a complete cycle of nodes and edges in one direction.



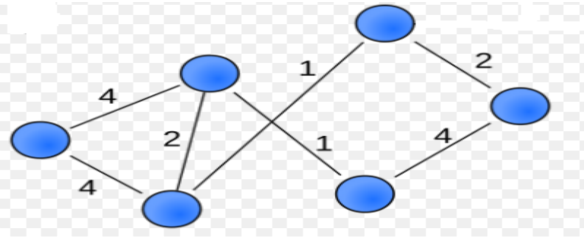
3. **Bipartite Network**: It is a network where two classes of nodes exist, say A and B. Nodes of class A will be connected only with nodes of B and vice versa. So, all links of the network connect a node in A with a node in B. There will not be any connections between two nodes belonging to the same class. For example, in the movie rating example, a movie is associated with a certain rating and a rating does refer to a certain movie or set of movies. But a movie does not refer to another movie or a rating count does not refer to another rating data. In the below figure, nodes are in two sets, namely A and B. Each node of set A connected to a node of set B, but not with any node of set A and the same with set B of the network.



4. **Weighted and Unweighted Network:**

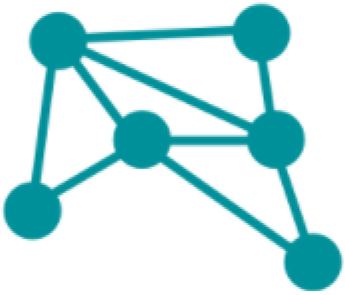
   In a diverse range of networks, there might be an association of **weights** to the edges of the network. For example, in transport system, a certain path (road / airway) is possibly used more than the others in the same network. Such edges can be given more weight than the others.

   a. **Weighted Network:** It is a type of network where each edge is associated with a certain weight parameter. For example, in the neural network, every edge is having some numeric weight that shows the contribution of the corresponding node in the output. In the below graph, the number on each edge is showing the corresponding weight of that edge.

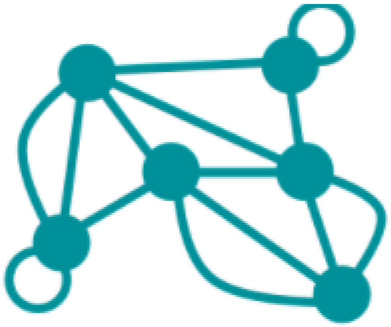


b. **Unweighted Network:** It is a type of network where no consideration of weights is done.

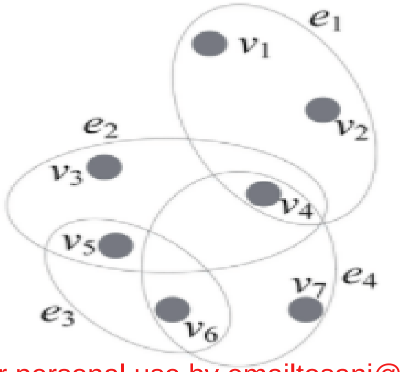
5. **Simple Network:** It is an **undirected** network with **at most one edge** between any pair of vertices. Also, it does not have any **self-loops** (an edge that originates and terminates at the same node). Examples of such networks are the internet, power grid, etc. Below is a simple graph as there is no self-loop and it fulfills the criterion of a simple network.



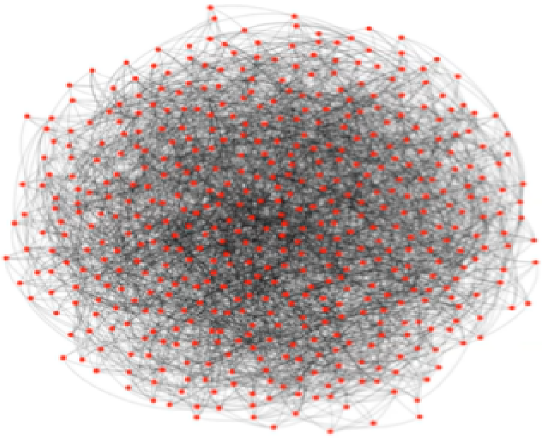
6. **Multigraph:** In a multigraph, self-loops and multiple links between two nodes are possible. For example, in road networks, it is possible that between two points / places two roads are going through different routes. In the below figure, there exist self-loops and multiple links, hence it is a multigraph.



7. **Hypergraph:** It is the **generalization** of a graph where an edge can join any number of nodes (>2). For example, the protein interaction network is a hypergraph. In the below graph, the edge e1 is connected to v1, v2, and v4 vertices, and so on. So, it is a hypergraph.



Networks are represented in a computer like a complex graph that is not interpretable when the number of nodes is high. They become like a hairball and just look fuzzy. Below is a diagram of a very large network.



Therefore, we need a method to represent networks.

Representation of a Network

To utilize network analysis properly, it becomes important to visualize them and represent them in a useful symbolic way. Generally, there are two ways of representing a network.

1. **Adjacency list:** It is useful in the case of very big networks where only a small count of nodes actually need to be represented. Since only connected nodes are included, it is generally a small representation. A sample representation of the adjacency list is as follows:

   $$Undirected\ graph\colon\ 1 - 2 - 3 \ = \ \{\{1,2\},\{2,3\}\}$$

   Only the connected nodes are shown here.

2. **Adjacency matrix:** It is more useful when the number of nodes is small because it considers all the existing nodes and includes them in the matrix. Symbolically, it can be represented as follows:

   $$A_{ij} = \ \{1,\ if\ (i,j)\ \in\ E,$$
   $$0,\ otherwise\,\}$$

   In general, the diagonal of the adjacency matrix are all zeros, but this is not mandatory, especially whenever there is a **self-loop.** The adjacency matrix always **symmetric** for undirected networks.

Adjacency lists and adjacency matrices are different from each other in terms of handling nodes that are not connected. The adjacency list is a representation that handles all the connected nodes while in the adjacency matrix all the nodes are considered.

The adjacency matrix has found a multitude of utilizations in today's world of computing. One of its beautiful applications is to find common friends of people on Facebook. To do this, we look for two ones in a single row or column of the adjacency matrix. Taking the matrix product of the adjacency matrix with itself, that is $A \cdot A$ (matrix product of A with A), can tell the number of common friends any pair of people have. For such cases, the matrix approach is very useful.

Among the nodes and edges, a network possesses some patterns that are associated with the structure (distribution of nodes and edge) of the network. To understand such existing patterns of a network, it becomes customary to define some quantitative features that can describe useful features / patterns of the network. Such measures also help in comparing the utilization of two networks. Below are some quantitative features of networks.

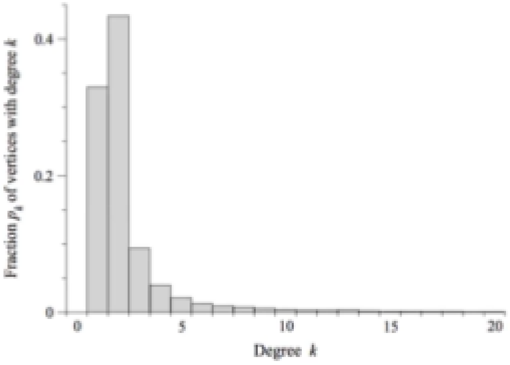
Quantitative Measures of Networks

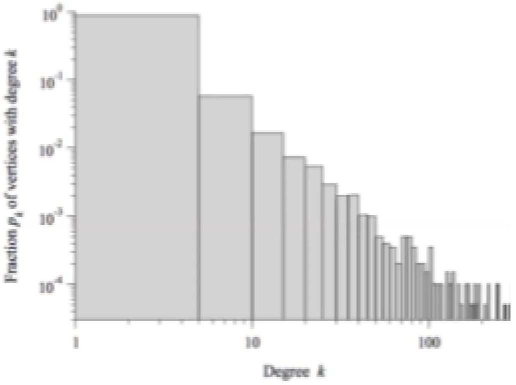
There are certain quantitative measures in network analysis that are explained below

1. **Connected Components:** It is a **subsection** of the entire network that exists and functions as a unit. Within a connected component, there exists a way to transmit from one node to **any** other node. The number of existing nodes in a connected component is called the **component size**. The higher the component size, the more complex the connected component is.

2. **Degree Distribution:** The degree of a node is the number of edges connected to that node. The degree distribution gives the detail of the number of edges that are originating from a certain node. It is a feature associated with a specific node.
   The **fraction of nodes** with degree k in the graph is the total number of nodes with degree k divided by the total number of nodes. The plot of "fraction of nodes" vs "degree k" gives the **degree distribution** of the network. The degree distribution captures only a small amount of information about the network. But that information still gives important clues into the structure of the network.

To understand this, let us use the example of the Facebook network. It is possible that there are people who have a different number of friends associated with them. Below are the plots that describe the fraction of total nodes with degree k. The first graphs shows that the maximum fraction of nodes are having a degree of 2 (nodes where two edges are connected)





The second plot is all about the logarithmic transformation of the first one. This plot shows an approximately fixed slope. As the graph depicts, at the tails the graph is fat, which shows there are many nodes with a high degree of distribution.

$$Log\ P_k \ = \ -\ \alpha\ log\ (k)\ +\ c \text{ , For some c>0}$$

$Log\ P_k$ is the logarithm of the fraction of nodes corresponding to the log of degree k. The slope is negative which shows that if the degree of freedom **increases,** the fraction of nodes will decrease, i.e., there are very few people with a very high number of friends in the Facebook network. This equation is called **power-law distribution.**

Power-law Distribution: In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another.

3. **Diameter and Average Path Length** - In most networks, it is important to understand the distribution of the distance between two nodes. It affects the transmission time between the nodes and is helpful in solving real-world problems.

a. **Diameter:** It is the distance between the two farthest nodes in the network.

$$D = max(d_{ij})$$

Where, d_{ij} denotes the length of the **geodesic path** (or the shortest path) between nodes i and j.

b. The **average path length** is the average distance between any two nodes in the network. Mathematically, it can be given as follows:

$$apl = \frac{1}{nC_2}\sum_{i\leq j}(d_{ij})$$

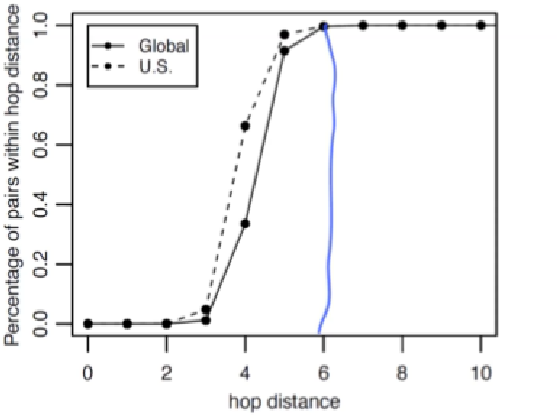
Sometimes, the average path length is small while sometimes it is large too. The average path length can be interpreted as, if it is small then the message or signal or any possible transmission among two random nodes can be done very quickly through the network. If it is large, then it will take more time to transmit from one node to another. In real-life scenarios, sometimes it also happens that the network is not connected. In such a case, we find the largest **component** of the network and find the diameter and **average possible length** of the same. It is interpreted as for the whole network.

To understand this, let us go through the famous experiment named **"The six degrees of separation"**. It depicts the real-life interpretation of the average possible length in a network.

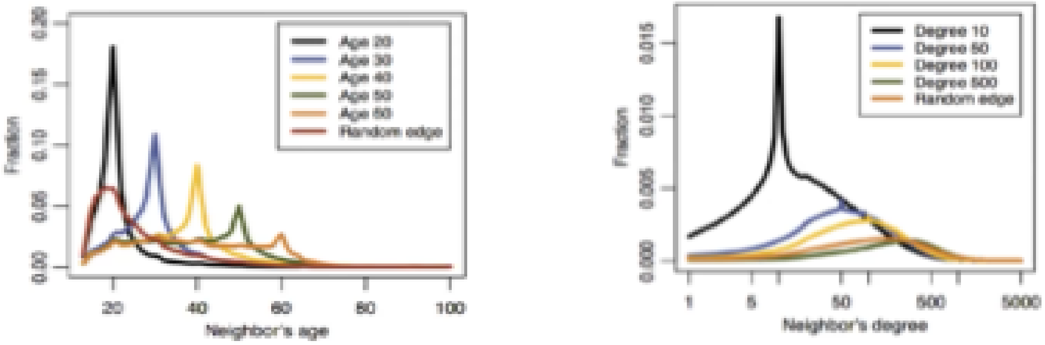
c. **Small world and 6 degrees of separation:** It is related to **the small world problem (1967)** done by Stanley Milgram. In his experiment, participants from a particular town were asked to get a letter to a particular person in a different town by passing it from acquaintance to acquaintance. 18 out of 96 letters made it in an average of 5.9 steps, suggestingthat the diameter of the social network in the US is 6.

However, is this true or should we take the conclusion of 6 degrees of separation with a grain of salt? Let's consider another example to answer this question.

Diameter of Facebook (2011): To understand the diameter of the Facebook network, let us take the help of the following plot. The Y-axis presents the percentage of pairs within hop distance while on the x-axis, we have the hop distance (hop is the logical distance between networks based on the number of routers that must be traversed by packets sent between them) is there. The plot contains data at the global level shown without the dotted line and the data of the United States with dotted lines. According to the figure, the diameter of the Facebook network is almost 6. This is because the percentage data is not growing any further once the hop distance is close to 6.



4. **Homophily or Assortative Mixing:** Homophily is the tendency of people to associate with others that are similar to them. It is also observed that people of a certain age group have more friends in the same age group. Considering nation to be criteria of similarity, someone who is from the **United States** is supposed to have more friends in Facebook from the United States than any other nation. Taking the number of friends as another criterion, it is found that people who have more friends are friends with those who also have more friends. So in Facebook, homophily is seen to exist as well.



In the left plot of the above figure, it can be seen that people with age 20 have the maximum fraction of friends in the same age group and so is the case with people of other age groups. In the right plot, people with a high degree of distribution have friends who alos have a high degree.

Working with a network with a large number of nodes and edges has never been an easy task. When it comes to interpretation, it is quite wise if we could identify the significant nodes. It plays a vital role from resource-saving to an effective utilization perspective, if the nodes where action has to be taken are identified before. To do so, it is needed to understand how to find important nodes in a network.

Finding Important Nodes in Networks

While working with networks it is very critical to understand the importance of nodes present in the network. This helps in making decisions that are useful for the purpose of the network. From an interpretation point of view, any node that seems to be central is called an important node. A node being central is an intuition that is defined below.

Centrality Measure: It is a measure that captures the importance of a node's position in the network. One possible intuition can be that it is a node that is close to all other nodes. For example, to build an airport that is close to other cities. In the case of a criminal network, the person or node that is connected to most of the nodes might be the most important person through whom all the information is passing.

To measure the centrality of a node, there are many possible ways.

1. **Degree Centrality:** This is the most intuitive way to measure the centrality of a node in a network. According to degree centrality, a node is supposed to be more central / important if it is a high degree node, i.e., if it has more outgoing edges originating from it. For a certain node in an undirected network, it is defined mathematically as follows:

$$K_i = \sum_j (A_{ij}).$$

Where, K_i is the degree of the i^{th} node of the network while A_{ij} is the element corresponding to the i^{th} row and j^{th} column in the adjacency matrix A.

For a certain node in a directed network, it is defined mathematically as follows:

$$K_i^{out} = \sum_j (A_{ij})$$

Where K_i^{out} is the number of nodes originating from the i^{th} node of the directed network.

2. **Eigenvector Centrality:** It is a different intuition for measuring centrality. According to Eigenvector centrality, each node is given a score that is proportional to the sum of scores of all its neighbors (nodes that are directly connected to a certain node). The intuition is that if the neighboring nodes are important the central node is also important.

   The process starts with giving some common importance to all the nodes and then start computing for individual nodes. Then, update the centrality of each node by centrality of neighbors as follows:

   $$x_i^{(1)} = \sum_{j=1}^{n} A_{ij} x_j^{(0)} \qquad \text{where, } x^{(0)} = 1 \text{ for all nodes}$$

   Iterate this process as follows:

   $$x^{(k)} = A^{(k)} X^{(0)}$$

   After some number of iterations, it will start to converge to a constant. Then, we get the Eigenvector corresponding to the largest eigenvalue of the adjacency matrix of the network.

3. **Closeness Centrality:** Track how close a node is to any other node. The closer it is, the higher the importance of the node. Mathematically, it is given as follows:

   $$C_i = \left(\frac{1}{n-1}\sum_{j\neq i} d_{ij}\right)^{-1}$$

   Where d_{ij} is the distance between nodes i and j.

4. **Betweenness Centrality**: It measures the extent to which a node lies on paths between other nodes. Mathematically, it is given as follows:

   $$B_i = \frac{1}{n^2}\sum_{s,t}\frac{n_{st}^{i}}{g_{st}}$$

   Where n_{st}^{i} is the number of shortest paths between s and t that pass through i, and g_{st} is the total number of the shortest paths between s and t.

Which centrality measure to use?

It seems that we have a multitude of centrality measures that can be applied based on the suitability of the application. To understand this, let us take an example of a friendship network:

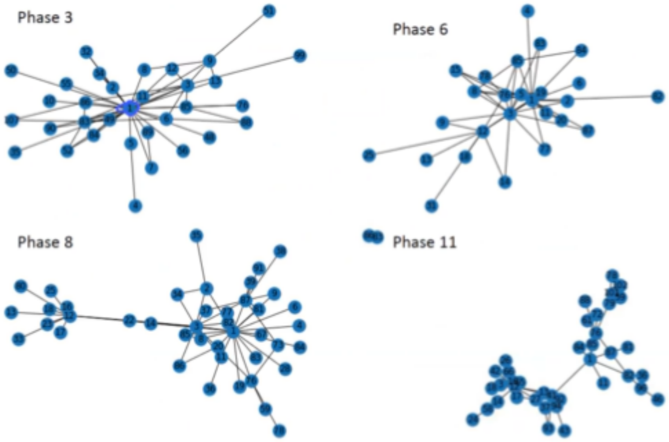
a. To identify the most popular person, **degree centrality** should be used. Higher the degree centrality more popular is the person.
b. To identify the most popular person that is friends with popular people, we need to use the **Eigenvector centrality**.
c. To identify the person that could best inform the group, **closeness centrality** should be used. The higher the value of closeness centrality, the higher the delivery of information through that person.
d. To identify the person that should be removed in order to best break the network, **betweenness centrality** should be used. The higher the centrality value, the more the likelihood that the network will break on their removal.

Now that the importance of nodes is covered properly, let us see a real-life scenario where making a slight change in the structure of the network plays a vital role in changing the importance of the nodes in the network.

Case study: CAVIAR (Criminal Network in Montreal)

Consider the CAVIAR case study, which is basically related to criminal network analysis in Montreal (a city in Canada). There existed a network of criminals in the city that supplied drugs from city to city. The police department was working to control such activities. The mandate was to seize the drugs being supplied instead of arresting the criminals.

To do this, the police department got the existing network details of the criminals. In this network, nodes are the **people** who were involved in the **supply chain** and edges are **"who is calling whom"** to make the supply process continuous. The police **wiretapped** (one can get the legal authentication to wiretap a person) criminals that helped them to know who is calling whom and when. Due to the legal verifications, it took 11 seizures (phases) to track them. The observation is whenever there is a seizing of drugs the criminals reoriented the network to do the supply. And accordingly, the most important node / criminal was also changing. This is because of reorienting, the information lead through which all the information passes also changes.



- The network consisted of 110 numbered players: 1-82 were traffickers, 83-110 were non-traffickers (Financial investors, accountants, owners of various importation businesses).

- According to the above plot, it can be observed that in different phases of drug seizure, there is a reorientation in the network and hence a change in the structure. As the reorientation is taking place, there is a change in the most important node as well. One example of reorientation is traffickers reoriented to **cocaine** import from **Colombia**, transiting through the **United States**. As they got caught, they re-oriented.

References

Chapters 1 - 10 (but mostly chapters 6 - 8) in

M. E. J. Newman. *Networks: An Introduction*. 2010.

For an analysis of the Facebook network:

J. Ugander, B. Karrer, L. Backstrom and C. Marlow. *The Anatomy of the Facebook Social Graph*. 2011.

For more information on the CAVIAR network:

C. Morselli. Inside Criminal Networks (Springer, New York). Chapter 6: Law-enforcement disruption of a drug-importation network). 2009.