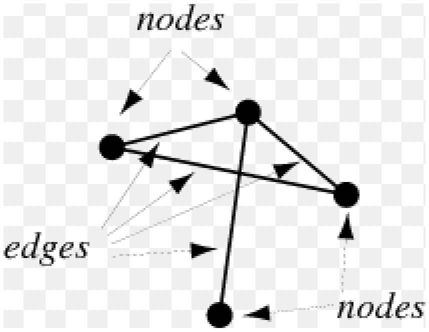
Post Session Document - Network Analysis

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 it might be helpful in developing a connection with a specific node that possesses the same or similar feature. For example, considering Facebook as a network the people having an account on Facebook are nodes while the message services, friend requests, etc acts like an edge between two persons/nodes. A person on Facebook does have many features (profile and friends related) associated with him. It can be seen easily in the above network.

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. It can be seen in the above network.

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 in terms of relation.

Let us use a **few examples** to clearly understand it -

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 as a **node** while the **transmission lines** are the **edges** between such nodes.

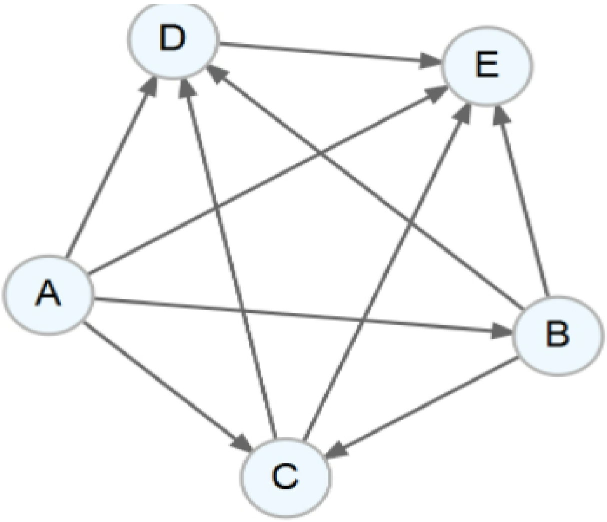
Below are some other examples of networks with **vertices and edges** in them -

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

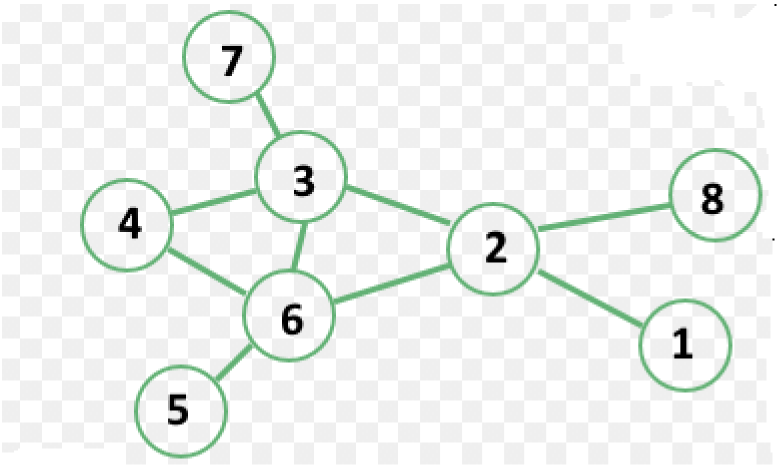
Based on different types of real-life applications and in consideration of the corresponding requirements, networks can be of many types. Below is a detailed view of the classification of the network -

Classification of network

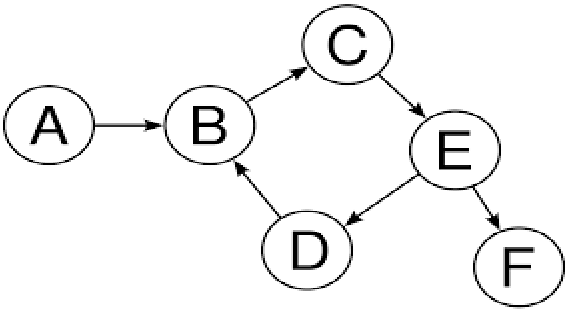
1. **Directed and Undirected network** - This is a distinction of networks based on whether the flow between nodes is occurring in a unique direction or in an unspecified direction.
    a. **Directed network** - It is a network where the transmission/flow between the nodes occurs in a **specific direction**. If a transaction goes from node A to node B then it won't occur from B to A. For example, the food chain of animals is a directed network. If a cat eats a rat, then the reverse is not possible (rat can not eat the cat). In the below plot A, B, C, D, and E are the nodes that are connected with directed edges.



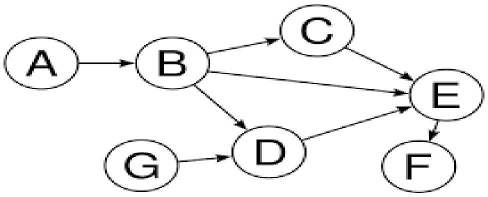
    b. **Undirected network** - It is a network where the transmission/flow between the nodes does **not** occur in a fixed direction. If a transaction occurs from node A to B then it may occur from node B to A as well. In the transmission network from node A to node B (from place A to place B) a vehicle can go but the reverse is also possible that some vehicles can go from B to A. So the direction of transmission is **not fixed** hence it is an undirected network. Another example is social media networks. A message can go either way. 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 network** - 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.

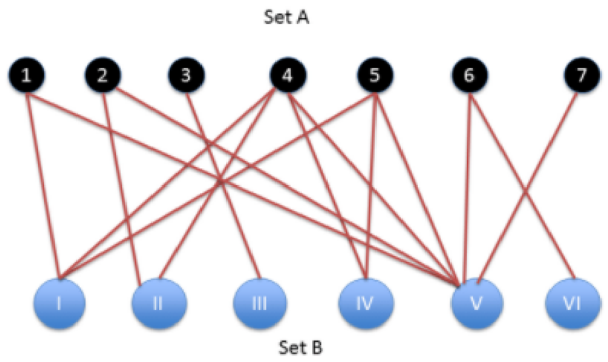


    b. **Acyclic network** - These are the networks that do not complete a proper cycle of nodes and edges in one direction.



3. **Bipartite network** - It is a network where two classes of nodes exist i.e. say A and B. Nodes of A will be connected only with nodes of B and vice versa. So all links of the network connect a

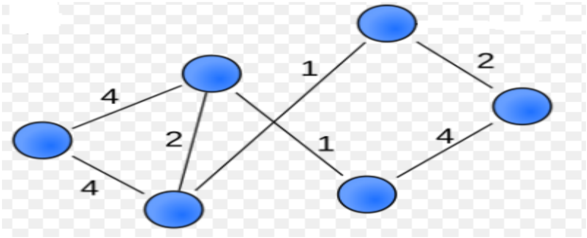
node in A with a node in B. There will not be any connection between two nodes belonging to either A and B only. For example, in the movie rating example, a movie is associated with a certain rating and a rating does refer to a certain movie or set of movies. But a movie does not refer to another movie 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 another node of set B but with any node of set A and the same with set B of the network.



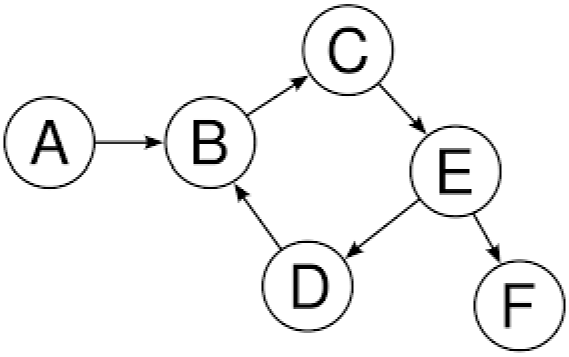
4. **Weighted and non-weighted network** -

   In a diverse range of networks, there might be an association of **weights** to the edges of the network. To understand, let us take an example of a transport system. A certain path/road/airway is possibly used more than the others in the same network. Such edges will be given more weight than the others.

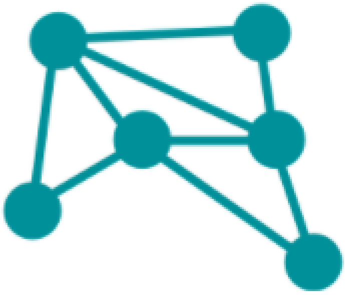
   a. **Weighted network** - It is a type of network where each edge is associated with a certain weight parameter. For example, in the neural network, every edge is having some numeric weight that shows the contribution of the corresponding node in the output. In the graph on every edge, the number present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



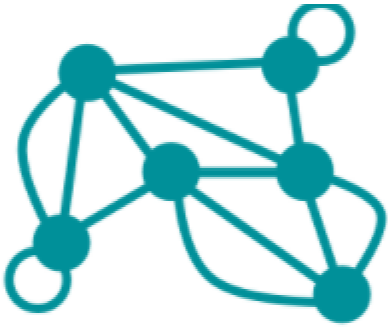
   b. **Non-weighted network** - It is a type of network where no consideration of weights is done. The below graph shows a non-weighted network. None of the nodes are associated with any weight that can change the value of their contribution for some purpose.



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-loop**. 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.

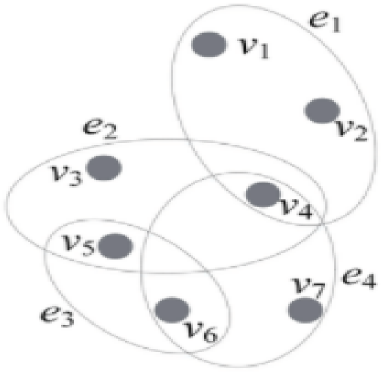


6. **Multigraph** - In a multigraph self-loops and multiple links between two nodes are possible. For example, in Road networks, it is possible that between two points/places two roads are going through different routes. In the below figure there exist self-loops and multiple links, hence it is a multigraph.

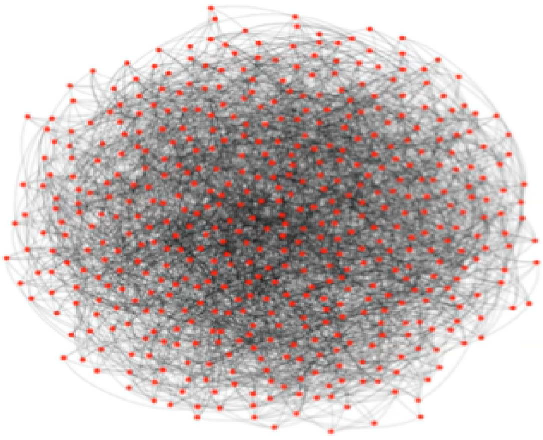


7. **Hypergraph** - It is the **generalization** of a graph where a node can join any number of nodes(>2). For example, the protein interaction network is a hypergraph. In the below graph

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.



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 nodes with zero entry are not included, it is generally a small representation. A sample representation of the adjacency list is as follows -

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

Only the connected edges are shown here.

2. **Adjacency matrix** - It is more useful in the case where the number of nodes is low because it considers all the existing zeros 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.** A self-loop can be defined as a structure where an edge originates and terminates at the same node. In this kind of network, the diagonal value will be 1 and not 0. The network is symmetric especially when it is **undirected**.

Adjacency lists and adjacency matrices are different from each other in terms of **handling non-zero** nodes. An adjacency list is a representation that handles **all the non-zero** nodes while in the adjacency matrix all the nodes irrespective of whether it is a zero or non-zero node are taken under consideration.

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 in the adjacency matrix in a single row or column we look for two ones. It makes sure that the two are having common friends. 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 two people have. For such cases, the matrix approach is very useful.

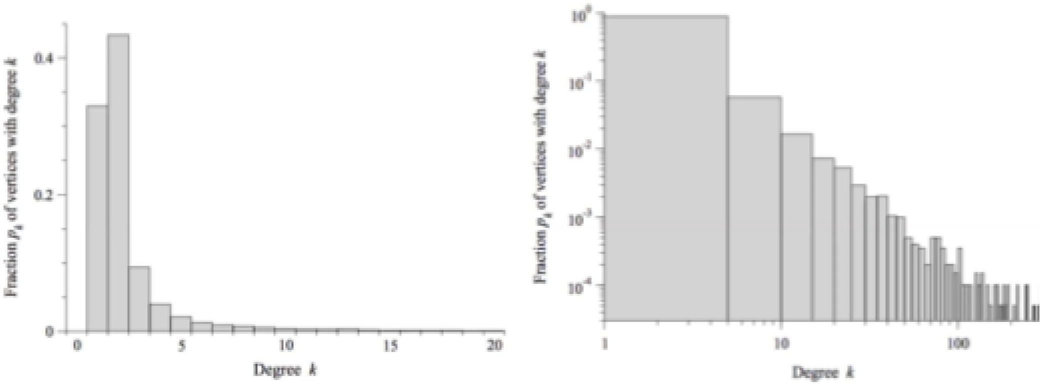
Among the nodes and edges, a network possesses some patterns that are associated with the structure (distribution of nodes and edges in the network) of the network. To understand such existing patterns of a network it becomes customary to define some quantitative features that can describe useful features/patterns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

Quantitative measures of the network - There are certain quantitative measures of network analysis that are explained below -

1. **Connected components** - Connected components are the set of nodes where one node is reachable to the other. 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.

2. **Degree distribution** - 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 **average degree distribution** is the total number of outgoing nodes divided by the total number of nodes. The average degree distribution is not a very informative term.

   To understand this, let us use an example of the Facebook network. It is possible that there are people who have a different number of friends associated with them and hence a different **fraction** of the entire network with different degrees of distribution. Below are the plots that describe the fraction of **total vertices** with the degree k. To interpret the first graph well, the maximum fraction of vertices 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 approximately shows a 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 vertices corresponding to the log of the degree of distribution. The slope is negative which shows that if the degree of freedom **increases** then the fraction of people will decrease i.e. there are very fewer people with a very high number of friends in the Facebook network. This equation is called **power-law distribution** because the quantity P_k is varying exponentially with k.

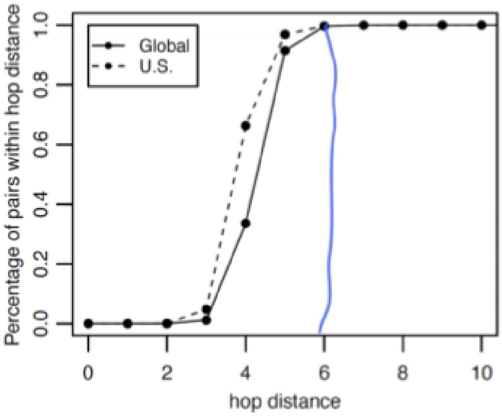
Power-law distribution - In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another.

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-life problems. To understand this, we need to define certain terminologies -

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

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

Diameter of Facebook (2011) - To understand the diameter of the Facebook network, let us take the help of the following plot. Y-axis presents the percentage of pairs within hop distance while on the x-axis **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 it can be interpreted as 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.



Geodesic path or shortest path - It is the smallest distance between two specific nodes namely i and j in a network.

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.

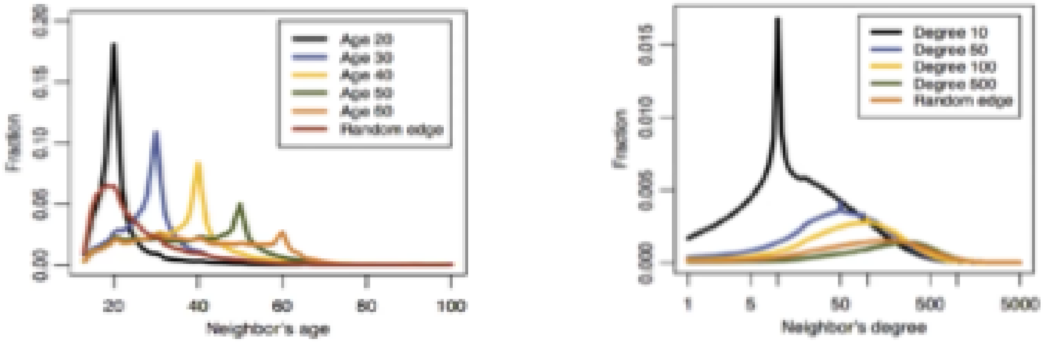
To understand this let us go through the famous experiment named **"The six degrees of separation"**. It depicts the real-life interpretation of the average possible length in a network.

Small world and 6 degrees of separation

It is related to **the small world problem (1967)** done by Stanley Milgram. In this problem, there were 96 people in different places of a country. They had to send one letter each to a **fixed destination**. So a total of 96 letters are to be sent to a specific destination. It was observed that out of 96 only 18 letters could reach the destination. So an average of 5.9 steps was taken by each letter to reach the destination.

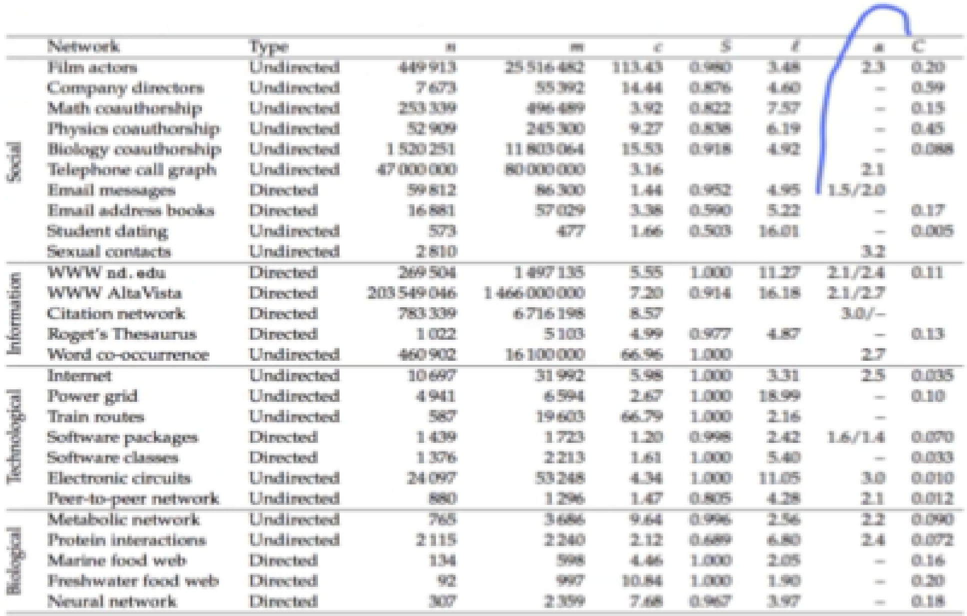
4. **Homophily or assortative mixing** - Homophily is the tendency of people to associate with others that are **similar** to them. It is also observed that people of a certain age group have more friends in the **same** age group. Considering nation to be criteria of similarity, someone who is from the **United States** on Facebook is supposed to have more friends 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 top left plot in the above figure, it can be seen that people with age 20 have the maximum fraction of friends in the same age group i.e. around 20. And so is the case with people of other age groups. In the bottom plot, people with a high degree of distribution have friends who are similar(in people who are a high degree of distribution).

Characteristics of different networks - Now let us get more familiar with the characteristics of different networks that are available in today's world. It will help us get more acquainted with the features of networks that are needed to interpret while understanding them.



	Network	Type	n	m	c	S	ℓ	α	C
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1	
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0	
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005
	Sexual contacts	Undirected	2 810					3.2	
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7	
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–	
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7	
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–	
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18

The above table describes some of the characteristics of different networks available. From the table, below are some descriptions that will help to understand it properly -

m= It is the number of edges, Higher the number of edges more complex is the network.

c= It is the mean degree of distribution of the network. The higher the value of $c$more is the complexity of the network.

S = proportion of Largest component of the network.

L = mean geodesic distance in the network. It is the average distance between shortest paths between pairs of existing nodes in the network. Lowe the value of L higher is the mobility in the network and hence easier it is to transmit through the network.

α= exp power-law degree distribution -

C = It is the number of triangles present in the network divided by the number of any three nodes that are connected. This quantity is also called the clustering **coefficient.** It tells about the tendency of nodes to cluster together in the network. Higher the value of Cmore is the tendency of nodes to cluster together to build clusters.

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 more 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 which is explained below.

Finding important nodes in a network -

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 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 criminality network, the person or node that is connected to most of the nodes might be the most dangerous 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 the undirected network, it is defined mathematically as follows -

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

   Here K_i is the degree of 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 K.

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

   $$K_i^{out} = \sum_j (Aij)$$

   Where K_i^{out} is the degree of centrality of a 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 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 starts computing for individual nodes. Then update each centrality by centrality of neighbors as follows -

   $$x_i^{(1)} = \sum_{j=1}^{n} A_{ij} X_j^{(0)}$$

   Iterate this process as follows -

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

   After a huge 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.

If there exist m>0 such that $A^m > 0$ then one can show that -

$$k \to \infty X^{k} = \alpha\lambda_{max} v$$

Where λ_{max} is the largest eigenvalue and $v \geq 0$ is the corresponding eigenvector. Alpha depends on the choice of $X^{(0)}$. This is the **Perron-Frobenius** theorem. The Eigenvalue approach can be interpreted as follows -

- A node is important if it has **important** neighbors. i.e. A person might be understood as important if he has friends who are important.
- A node is important if it has **many** neighbors. A person might be understood as important if he has a huge count of friends.
- The eigenvector corresponding to the largest eigenvalue of A provides a ranking of all nodes. For every eigenvalue, there is an eigenvector. But for the max eigenvalue, there is an eigenvector that has to be picked.

3. **Closeness centrality** - Track how close a node is to any other nodes. 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 $dij$is the distance between nodes i and j. In disconnected networks, the average is taken over the nodes in the same component as i. Apart from this harmonic centrality is also considered for closeness centrality as follows

$$H_{i} = \frac{1}{n-1}\sum_{J\neq i}\frac{1}{d_{ij}}$$

4. **Betweenness centrality** - It measures the extent to which a node lies on paths between other nodes. It is mathematically given as follows -

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

Where $n^{i}st$ 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

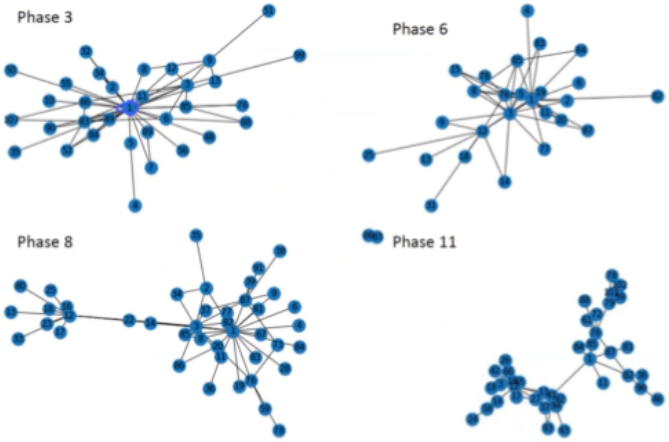
It seems that we have a multitude of centrality measures that can be applied based on the suitability of the application. To understand this let us take an example of a friendship network -

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 centrality, the higher the delivery of information to the group.
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 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 too 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 exists a network of criminals in the city that does supply drugs from city to city. The police department was working to control such activities. The mandate is to seize the drugs being supplied but **not to arrest the criminals**. To do this the police department has the existing network details of the criminals. In this network, nodes are the **people** who are involved in the **supply chain** and edges are **"who is calling whom"** to make the supply process continuous. The police have **wiretapped** (One can get the legal authentication to wiretap a person) the warrants that help them know who is calling whom and when. Due to the legal verifications, it took 11 periods 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 by reorienting, the information lead through which all the information passes also changes.



- The network consists of 110 numbered players: 1-82 are traffickers, 83-110 are 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 also. One example of reorientation is traffickers reoriented to **cocaine** import from **Colombia**, transiting through the **United States**. As they are always getting caught, they re-orient.

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.