Chapter 2

Topological parameters of a network and relevant statistical tests

Any graph or network has two basic components - nodes and edges. An unweighted network can be represented as an adjacency matrix (A). Any element of adjacency matrix (A) a_{ij}, is given as

$$a_{ij} = \begin{cases} 1, & \text{if } i \neq j \text{ and i and j nodes are connected by an edge} \\ 0, & \text{if } i \neq j \text{ and i and j nodes are not connected} \\ 0, & \text{if } i = j. \end{cases} \tag{2.1}$$

A toy network is shown in Figure 2.1 and the respective adjacency matrix is presented in Table 2.1.

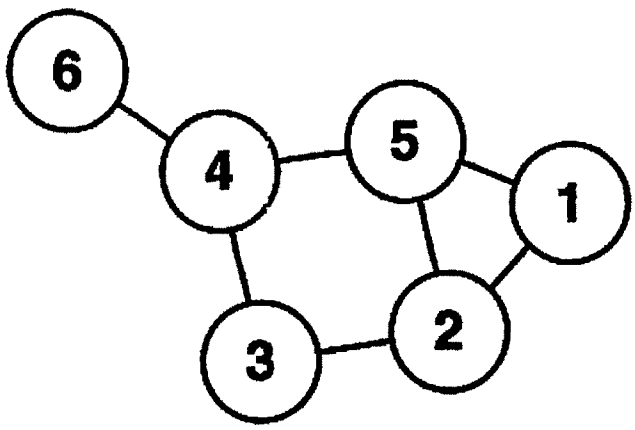

Figure 2.1: *An unweighted graph with six nodes marked as 1, 2, 3, 4, 5 and 6. Here the degrees of the nodes are 2, 3, 2, 3, 3 and 1, respectively.*

Table 2.1: *Adjacency matrix of the graph represented in Figure 2.1*

node	1	2	3	4	5	6
1	0	1	0	0	1	0
2	1	0	1	0	1	0
3	0	1	0	1	0	0
4	0	0	1	0	1	1
5	1	1	0	1	0	0
6	0	0	0	1	0	0

For a weighted network the weight matrix is represented as (W), any element w_{ij} is the weight of the edge between the nodes i and j. A toy weighted network is shown in Figure 2.2 and the weight matrix is presented in Table 2.2.

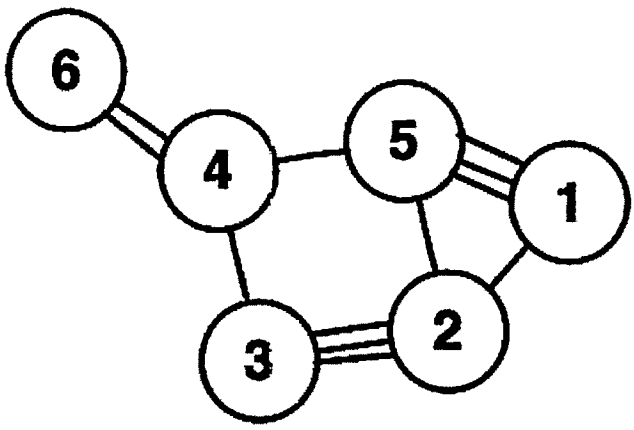

Figure 2.2: *A weighted graph with six nodes marked as 1, 2, 3, 4, 5 and 6. Here the strengths of the nodes are 4, 5, 4, 4, 5 and 2, respectively.*

Table 2.2: *Weight matrix of the graph represented in Figure 2.2*

node	1	2	3	4	5	6
1	0	1	0	0	3	0
2	1	0	3	0	1	0
3	0	3	0	1	0	0
4	0	0	1	0	1	2
5	3	1	0	1	0	0
6	0	0	0	2	0	0

2.1 Degree of a node and its distribution

The degree of any node i is represented by

$$k_i = \sum_j a_{ij}. \tag{2.2}$$

Higher degree of a node implies stronger connectivity of the node in the network.

Table 2.3: *Degrees of nodes of the graph represented in Figure 2.1*

node	1	2	3	4	5	6
k	2	3	2	3	3	1

The probability of degree distribution is represented by

$$P(k) = \frac{N(k)}{\sum N(k)}. \tag{2.3}$$

where $N(k)$ is the number of nodes in the network with degree k.

The average degree of a network is given by

$$< k >= \frac{\sum_i k_i}{N} \tag{2.4}$$

where N the total number of nodes in the network.

Higher average degree implies good inter-connectivity among the nodes in the network.

2.2 Strength of a node and its distribution

If w_{ij} is the number of possible interactions between any i^{th} and j^{th} nodes, then the strength (s_i) of a node i is given by

$$s_i = \sum_j a_{ij} w_{ij}. \tag{2.5}$$

Table 2.4: *Strength of nodes of the graph represented in Figure 2.2*

node	1	2	3	4	5	6
s	4	5	4	4	5	2

The spread in the strength of a node has been characterized by a distribution function $P(s)$; where

$$P(s) = \frac{N(s)}{\sum N(s)} \tag{2.6}$$

$N(s)$ being the number of nodes with strength s. The average strength of a network

is given by

$$< s >= \frac{\sum_i s_i}{N} \tag{2.7}$$

where N the total number of nodes in the network.

2.3 Characteristic path length of a network

The characteristic path length (L) of a network is the shortest path length between two nodes averaged over all pairs of nodes and is given by

$$L = \frac{\sum_i \sum_j L_{i,j}}{N(N-1)} \tag{2.8}$$

where $L_{i,j}$ is the shortest path length between i^{th} node and j^{th} node.

Table 2.5: *Shortest path length matrix of the graph represented in Figure 2.1*

node	1	2	3	4	5	6
1	0	1	2	2	1	3
2	1	0	1	2	1	3
3	2	1	0	1	2	2
4	2	2	1	0	1	1
5	1	1	2	1	0	2
6	3	3	2	1	2	0

Higher characteristic path length implies network is almost in liner chain and lower characteristic path length shows the network is in compact form.

2.4 Clustering coefficient of a network

The clustering coefficient (C) is a measure of local cohesiveness. Traditionally the clustering coefficient C_i of a node i is the ratio between the total number (e_i) of the edges actually connecting its nearest neighbors to the i^{th} node and the total number of all possible edges between all these nearest neighbors $[\frac{k_i(k_i-1)}{2}$; if the i^{th}

vertex has k_i neighbors] and is given by

$$C_i = \frac{2e_i}{k_i(k_i - 1)}. \tag{2.9}$$

where e_i is the total number of edges actually connecting the i^{th} node's nearest neighbors. Then the clustering coefficient of a network is the average of its all individual C_i 's.

Table 2.6: *Clustering coefficients of nodes of the graph represented in Figure 2.1*

node	1	2	3	4	5	6
C	1	1/3	0	0	1/3	-

The average clustering coefficient of a network is given by

$$< C >= \frac{\sum_i C_i}{N}. \tag{2.10}$$

Here, N is the total number of of nodes of the network.

2.5 Small World Property of network

A small-world network is a type of graph in which most nodes can be reached from every other node by a small number of hops or steps. Human social networks, for example, famously connect any two people on Earth - or any player to Deiago Maradona - in six steps or less. This small world property has been observed for many real networks.

To examine if there is any 'Small World' property in a network, one can follow Watts & Strogatz's method [16]. According to them, a network has the small world property if $C >> C_r$ and $L \geq L_r$. Here, C_r and L_r are respectively the clustering coefficient and characteristic path length for the corresponding random network having same number of nodes and edges.

For a random network having N number of nodes with average degree $< k >$, the characteristic path length (L_r) and the clustering coefficient (C_r) can be calculated using the expressions $L_r \approx \frac{lnN}{ln<k>}$ and $C_r \approx \frac{<k>}{N}$ given in [16].

Here we define the ratio

$$p = \frac{C}{C_r} \tag{2.11}$$

and the ratio

$$q = \frac{L}{L_r} \tag{2.12}$$

if $p >> 1$ and $q \approx 1$ then we can say the network has 'Small World' property.

2.6 Weighted clustering coefficient of a network

Combining the topological information with the weight distribution of the network, Barrat et al [65] have introduced an analogous parameter to C and that is known as weighted clustering coefficient, C^w. The weighted clustering coefficient C^w takes into account the importance of the clustered structure on the basis of amount of interaction intensity actually found on the local triplets and is given by

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{w_{ij} + w_{ih}}{2} a_{ij} a_{ih} a_{jh} \tag{2.13}$$

where s_i is strength of i^{th} node, k_i is degree of i^{th} node, w_{ij} is the number of connections between i^{th} and j^{th} nodes of the network.

C and C^w provide global information on the correlation between weights and topology, especially by comparing them with their topological analogs. For a large randomized network (lack of correlations) it is easy to find that $C^w = C$ and $C^w(k) = C(k)$. In real weighted networks, however, two opposite cases are found. If $C^w > C$, then the interconnected triplets are more likely formed by the edges with larger weights. On the other hand, $C^w < C$ signals a network in which the topological clustering is generated by edges with low weight.

The average weighted clustering coefficient of a network is given by

$$< C^w >= \frac{\sum_i C_i^w}{N} \tag{2.14}$$

where N is the total number of nodes of the network.

2.7 Mixing behavior of nodes

Interaction dynamics of a network is a very interesting phenomena. The pattern of connectivity among the nodes of varying degrees affects the interaction dynamics of the network. If the high-degree nodes in a network tend to be connected with other high-degree nodes, then the network is 'assortative'. On the other hand, the network is said to be 'disassortative' if the high-degree nodes tend to be connected with other low-degree nodes.

This mixing behavior of nodes of a network can be understood by two different methods discussed below.

2.7.1 Pearson correlation coefficient of a network

To study the tendency for nodes in networks to be connected to other nodes that are like (or unlike) them, one can calculate the Pearson correlation coefficient of the degrees at either ends of an edge. For our undirected unweighted protein network its value has been calculated using the expression suggested by Newman [18] and is given as

$$r = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i 0.5(j_i + k_i)]^2}{M^{-1}\sum_i 0.5(j_i^2 + k_i^2) - [M^{-1}\sum_i 0.5(j_i + k_i)]^2} \tag{2.15}$$

Here j_i and k_i are the degrees of the vertices at the ends of the i^{th} edge, with $i = 1, ..M$; where M is the total number of edges. The networks having positive r values are assortative in nature.

The coefficient r measures the tendency of degree correlation. This parameter helps us to understand whether the network is assortative or disassortative type. The coefficient r is a normalized coefficient($-1 \leq r \leq 1$). For positive r value one can say that the network has been evolved due to assortative mixing of its nodes. On

the other hand the network having negative r value is evolved due to disassortative mixing of its nodes. One cannot predict the mixing behavior when r is 0.

2.7.2 Average degree of nearest neighbors for vertices of degree k for unweighted (k_{nn}) and weighted network ($k^w_{nn,i}$)

To understand the mixing behavior of nodes in a network one can also calculate average degree of nearest neighbor, k_{nn} for vertices of degree k.

For an unweighted network it is given by

$$k_{nn}(k) = \sum_{\acute{k}} \acute{k} P(\acute{k}/k) \tag{2.16}$$

Here $P(\acute{k}/k)$ is the conditional probability that a given vertex with degree k is connected to vertex of degree $\acute{k}$. When there is a degree correlation, $P(\acute{k}/k)$ does not depend on k neighbor's degree, i.e., $k_{nn}(k)$ is constant. On the other hand, if $k_{nn}(k)$ is an increasing function of k then the network is of assortative type. In a disassortative type of network $k_{nn}(k)$ is a decreasing function of k.
On the other hand, for a weighted network Barrat et al have modified the expression for $k_{nn,i}$ and suggested the equivalent weighted average nearest-neighbors degree $k^w_{nn,i}$ [65] to be defined as

$$k^w_{nn,i} = \frac{1}{s_i} \sum_{j=1}^{N} a_{ij} w_{ij} k_j \tag{2.17}$$

2.8 Whether the network is random or scale-free?

In random networks any node has the same probability to be connected with any other node of the network [66, 67].

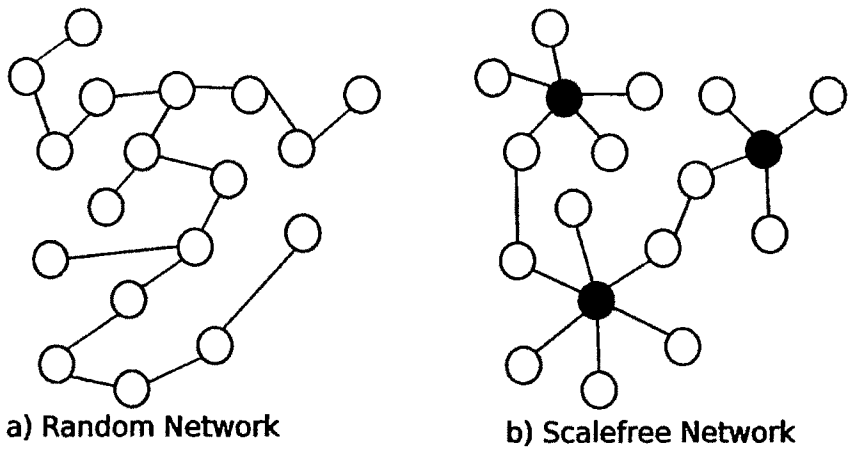

Figure 2.3: *a) Random network and b) Scale-free network.*

On the other hand, in scale-free networks [14], some nodes act as "highly connected hubs" (high degree), although most nodes are of low degree. Scale-free networks' structure and dynamics are independent of the system's size N, the number of nodes the system has. In other words, a network that is scale-free will have the same properties no matter what the number of its nodes is.

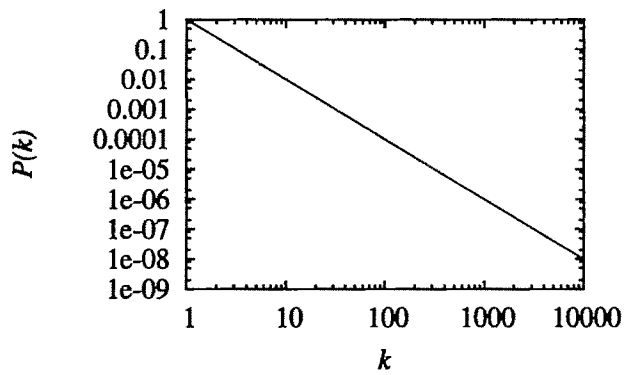

Figure 2.4: *P(k) vs k for scale-free network.*

Scale-free networks are characterized by a power-law degree distribution; the probability that a node has k links follows $P(k) \sim k^{-\beta}$, where β is the degree exponent. The probability that a node is highly connected is statistically more significant

than in a random graph [67], the network's properties often being determined by a relatively small number of highly connected nodes that are known as hubs.

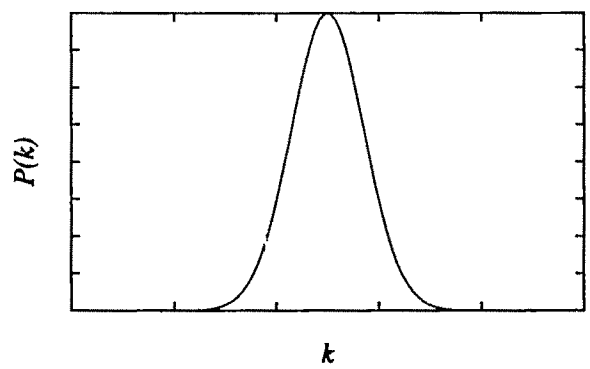

Figure 2.5: *P(k) vs k plot for random network.*

The coefficient β may vary approximately from 2 to 3 for most real networks, however, in some cases it can also take a value between 1 and 2 [68].

The degree distribution of scale-free network is characterized by power law, this is shown in Figure 2.4. On the other hand, the degree distribution for random network is Gaussian in nature and is shown in Figure 2.5.

2.9 Is the network hierarchical?

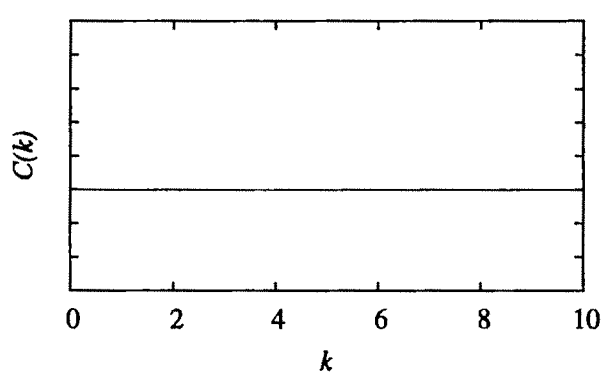

Figure 2.6: *C(k) vs k for random and scale-free network.*

From probability degree distribution one can differentiate the random (Gaussian) and scale-free (follows power-law) networks. Further to understand whether the network is hierarchical or not we have to plot $C(k)$ vs k. For a non hierarchical network $C(k)$ is constant (Figure 2.6). On the other hand for a hierarchical network $C(k) \sim k^{-\alpha}$ where the scaling coefficient $\alpha = 1.0$ [32].

2.10 Residue Centrality

The residue centrality [53] is calculated using the changes of the characteristic path length under removal of node k with its links

$$\Delta L_k = |L - L_{rem,k}| \tag{2.18}$$

where L is the characteristic path length given by Equation 2.8 with $L_{rem,k}$ represents the characteristic path length after the removal of node k and corresponding links from the network.

The statistically significant central nodes are evaluated using the Z-score values [52] of the node centrality defined as

$$Z_k = \frac{\Delta L_k - \overline{\Delta L}}{\sigma} \tag{2.19}$$

where, ΔL_k is the change of characteristic path length under removal of node k; $\overline{\Delta L}$ is the change of characteristic path length under node removal averaged over all nodes in the network, σ is the corresponding standard deviation.

2.11 Closeness Centrality

Closeness centrality is used to find central vertices. It gives higher values to more central vertices. Closeness centrality of a node x, is denoted by $C(x)$ [53] and is calculated as follows

$$C(x) = \frac{N-1}{\sum_{y \epsilon U, y \neq x} d(x,y)} \tag{2.20}$$

where $d(x, y)$ is the geodesic distance between node x and node y. U is the set of all nodes and N is the number of nodes in the network.

The closeness value is therefore the inverse of the average distance between x and other nodes ($\bar{d}$) i.e., $C(x) = 1/\bar{d}$

2.12 Mann Whiteny U Test

The Mann-Whitney U test [69] is a non-parametric test for assessing whether two samples of observations come from the same distribution. The test involves the calculation of a statistic, usually called U, whose distribution under the null hypothesis is known.

The steps are as follows

1. Rank all the observations without regard to which sample they are in.
2. Add up the ranks in first sample. Or the second sample.
3. "U" can be calculated using one of the the followings:

   $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$

   where n_1 is the two sample size for first sample, and R_1 is the sum of the ranks in first sample.

   and

   $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$.

where n_2 is the two sample size for second sample, and R_2 is the sum of the ranks in second sample.