Social network Analysis

Agarwal Prateek 05CS1021

Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India
prateek.iitkgp@gmail.com

1 Introduction

A social network is a set of actors (or points, or nodes, or agents) that may have relationships (or edges, or ties) with one another. The basic idea of a social network is very simple. Networks can have few or many actors, and one or more kinds of relations between pairs of actors. To build a useful understanding of a social network, a complete and rigorous description of a pattern of social relationships is a necessary starting point for analysis. That is, ideally we will know about all of the relationships between each pair of actors in the population.

In its simplest form, a social network is a map of all of the relevant ties between the nodes being studied. The network can also be used to determine the social capital of individual actors. These concepts are often displayed in a social network diagram, where nodes are the points and ties are the lines.

One reason for using analysis techniques in social networks is to represent the descriptions of networks compactly and systematically. This also enables us to use computers to store and manipulate the information quickly and more accurately than we can by hand. For small populations of actors (e.g. the people in a neighborhood, or the business firms in an industry), we can describe the pattern of social relationships that connect the actors rather completely and effectively using words. To make sure that our description is complete, however, we might want to list all logically possible pairs of actors, and describe each kind of possible relationship for each pair. This can get pretty tedious if the number of actors and/or number of kinds of relations is large. Formal representations ensure that all the necessary information is systematically represented, and provides rules for doing so in ways that are much more efficient than lists.

Social network analysis has now moved from being a suggestive metaphor to an analytic approach to a paradigm, with its own theoretical statements, methods and research tribes. Analysts reason from whole to part; from structure to relation to individual; from behavior to attitude. They either study

whole networks, all of the ties containing specified relations in a defined population, or *personal networks*, the ties that specified people have, such as their "personal communities".

The shape of a social network helps determine a network's usefulness to its individuals. Smaller, tighter networks can be less useful to their members than networks with lots of loose connections (weak ties) to individuals outside the main network. More open networks, with many weak ties and social connections, are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties. In other words, a group of friends who only do things with each other already share the same knowledge and opportunities. A group of individuals with connections to other social worlds is likely to have access to a wider range of information. It is better for individual success to have connections to a variety of networks rather than many connections within a single network. Similarly, individuals can exercise influence or act as brokers within their social networks by bridging two networks that are not directly linked (called filling structural holes).

The power of social network analysis stems from its difference from traditional social scientific studies, which assume that it is the attributes of individual actors – whether they are friendly or unfriendly, smart or dumb, etc. – that matter. Social network analysis produces an alternate view, where the attributes of individuals are less important than their relationships and ties with other actors within the network. This approach has turned out to be useful for explaining many real-world phenomena, but leaves less room for individual agency, the ability for individuals to influence their success, because so much of it rests within the structure of their network.

Social networks have also been used to examine how organizations interact with each other, characterizing the many informal connections that link executives together, as well as associations and connections between individual employees at different organizations. For example, power within organizations often comes more from the degree to which an individual within a network is at the center of many relationships than actual job title. Social networks also play a key role in hiring, in business success, and in job performance. Networks provide ways for companies to gather information, deter competition, and collude in setting prices or policies.

2 Properties of Social Network

1. **Milgram Experiment:** The experiments probed the distribution of path lengths in an acquaintance network by asking participants to pass a letter2 to one of their first-name acquaintances in an attempt to get it to an assigned target individual. Most of the letters in the experiment

were lost, but about a quarter reached the target and passed on average through the hands of only about six people in doing so . This experiment was the origin of the popular concept of the "six degrees of separation,". Traditional social network studies often suffer from problems of inaccuracy, subjectivity, and small sample size. Because of these problems other methods were required for probing social networks.
One source of copious and relatively reliable data is collaboration networks. These are typically affiliation networks in which participants collaborate in groups of one kind or another, and links between pairs of individuals are established by common group membership. A classic, though rather frivolous, example of such a network is the collaboration network of film actors, which is thoroughly documented in the online Internet Movie Database. In this network actors collaborate in films and two actors are considered connected if they have appeared in a film together.

2. **Small World Effect :**
Many networks exhibit an interesting behavior known as the *small world effect*, discovered by Milgram in 1969. In the original experiment(discussed above), Milgram sent 60 letters to various people, asking them to forward the letter to one of two unacquainted targets. The senders only knew the recipient's name, occupation and general location. They were instructed to send the card to a person whom they knew on a first-name basis who they thought was most likely to know the target personally. That person would do the same, and so on, until it was delivered to the target himself. Of all the letters that were sent, only about a fourth of them reached the target. Of these, 80% were delivered in four or fewer steps. Almost all the chains were less than 6 steps long. Thus it led to the general hypothesis, that any two people in the US were connected by a chain of less than 6 people. This kind of behavior was also observed in communities involving mathematicians and actors. The small world effect observed in Milgram's experiment has not gone unchallenged. The chief contention being that the small world effect did not take into account the large percentage of undelivered messages. Nevertheless, the experiment did help in establishing the small world effect as an important characteristic of many complex networks. More formally, it can be said that if a network shows the small world effect, then the diameter of the network grows logarithmically with the number of nodes in the network.

   a) **Finding whether a network shows small world effect :**
   For an undirected network of n nodes the mean smallest path (l) between vertices can be expressed as

$$l = \frac{1}{\frac{n(n+1)}{2}} \sum_{i \geq j} d_{ij} \tag{1}$$

where *dij* is the smallest distance from vertex i to vertex j. Most of the networks observed in real life (*e.g. film actors, company directors, emails, internet and electronic circuits*) have $l \leq 6$. The shortest between vertices can be found using standard BFS for unweighted graphs. Since the time required for performing BFS from a single node is $O(m)$, the total time required to find l would be $O(mn)$. For networks, with more than one component, we can find l for each connected component, and take the harmonic mean of these ls to get the overall mean smallest path length.

3 Transitivity or Clustering

In many networks it is found that if vertex A is connected to vertex B and vertex B to vertex C, then there is a heightened probability that vertex A will also be connected to vertex C. In the language of social networks, the friend of your friend is likely also to be your friend. In terms of network topology, clustering coefficient of the network means the presence of a heightened number of triangles in the network—sets of three vertices each of which is connected to each of the others.
It can be quantified by defining a clustering coefficient C thus:

$$C = \frac{3 \times number\ of\ triangles\ in\ the\ network}{number\ of\ connected\ triples\ of\ vertices} \tag{2}$$

where a "connected triple" means a single vertex with edges running to an unordered pair of others.

Example 1. Now, let us calculate the clustering coefficient for the figure 1 using eq. 2. It is easy to see that there are 2 triangles {123},{134} in the graph. And the connected triplets will be {123}, {231}, {312}, {134}, {341}, {413}, {234}, {135}, {214}, {534} and {532}.

So,

$$C = \frac{3 \times 2}{11} = 0.54$$

For directed graphs, triangles can be considered as cycles of length three and the denominator will be the number of connected triplets.

Example 2. It will be more clear with an illustration. Consider fig. 2 for example. Here, only one triangle {453} is present. And triplets present are {453}, {534}, {345}, {234}, {456} and {653}.

Here,

$$C = \frac{3 \times 1}{6} = 0.50$$

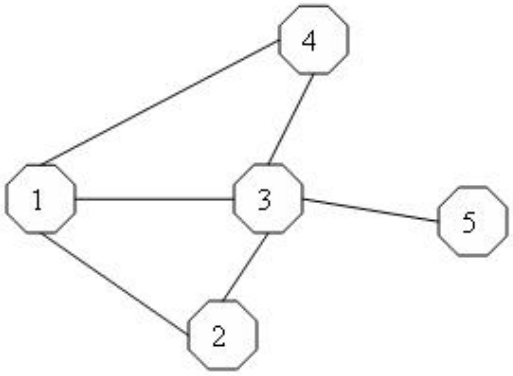

Fig. 1. Sample undirected network for calculation of clustering coefficient

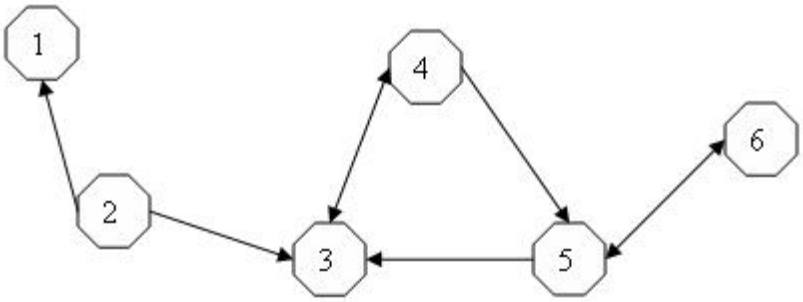

Fig. 2. Directed network for calculation of clustering coefficient

In effect, C measures the fraction of triples that have their third edge filled in to complete the triangle. The factor of three in the numerator accounts for the fact that each triangle contributes to three triples and ensures that C lies in the range $0 \leq C \leq 1$. In simple terms, C is the mean probability that two vertices that are neighbor of a common third vertex in the network will themselves be neighbors. It can also be written in the form

$$C = \frac{6 \times number\ of\ trianles\ in\ the\ network}{number\ of\ paths\ of\ length\ two} \tag{3}$$

Example 3. For example the clustering coefficient for the figure 1 using eq. 5 is calculated as:

Calculating Clustering Coefficient for each vertex:-

$C_A = \frac{\text{No. of triangles in which A is the vertex}}{\text{No of triples where A is the center}} = \frac{2}{3}$
Similarly,
$C_B = \frac{1}{1}$
$C_C = \frac{2}{6} = \frac{1}{3}$
$C_D = \frac{1}{1}$
$C_E = 0$
$Thus$ C= $\frac{C_A+C_B+C_C+C_D+C_E}{\text{No of vertices}} = \frac{\left(\frac{2}{3}+1+\frac{1}{3}+1+0\right)}{5}$
$= 0.60$

An alternative definition of the clustering coefficient, also widely used, has been given by Watts and Strogatz. The following definition finds clustering coefficient considering clustering coefficient for each vertices.

$$C_i = \frac{number\ of\ triangles\ with\ i\ as\ a\ vertex}{number\ of\ triples\ with\ i\ as\ center} \tag{4}$$

For vertices with degree 0 or 1, for which both numerator and denominator are zero, we put $C_i = 0$. Then the clustering coefficient for the whole network is the average

$$C = \frac{1}{n}\sum_{i=1}^{n} c_i \tag{5}$$

This definition effectively reverses the order of the operations of taking the ratio of triangles to triples and of averaging over verticesone here calculates the mean of the ratio, rather than the ratio of the means. It tends to weight the contributions of low-degree vertices more heavily, because such vertices have a small denominator in latter equation and hence can give quite different results from the former.

So, we see the result is different from that calculated from eq. 2. Lets now see for a directed graph,

The clustering coefficient for the figure 2 using eq. 5 is calculated as:

Example 4. Again calculating Clustering Coefficient for each vertex:-

$C_3 = \frac{\text{No. of triangles in which 3 is the vertex}}{\text{No of triples where 3 is the center}} = \frac{1}{2}$
Similarly,
$C_1 = 0$
$C_2 = 0$
$C_4 = \frac{1}{1}$
$C_5 = \frac{1}{3}$
$C_6 = 0$
$Thus$ C= $\frac{C_1+C_2+C_3+C_4+C_5+C_6}{\text{No of total vertices}} = \frac{\left(0+0+\frac{1}{2}+\frac{1}{1}+\frac{1}{3}+0\right)}{6}$
$= \frac{11}{36} = 0.31$

As we see, in this case coefficient turns out to be greater in eq. 2. So, it totally depends on the structure of the network.

4 Cohesive Subsets of nodes

4.1 Introduction

A well-known concept, more or less corresponding to that of the peer group is the clique: a group all members of which are in contact with each other or are friends, know each other, etc. A formal definition can be given as:

1. A clique in a graph is a maximal complete subgraph of three or more nodes. All nodes in the clique are adjacent to each other. No other nodes are adjacent to all nodes in the clique.
2. A clique is a very stringent definition of a cohesive subgroup.

1. **Subgroups based on Reachability :**
   a) ***n*-Clique**

   Cohesive subgroups based on reachability require that geodesic distances[1] among members of the subgroups be small. Thus, we can specify some cutoff value, n, as the maximum length of geodesics connecting pairs of actors within the cohesive subgroup. Restricting geodesic distance among subgroup members is the basis for the definition of a n-clique.
   An n-clique is a maximal subgraph in which the largest geodesic distance between any two nodes u, v is not greater than n. Here maximal means that there is no other node present in graph such that adding that node can still preserve the n-clique property.

[1]Geodesic distance : The geodesic distance between two nodes, denoted by $d(i, j)$ is the length of the shortest path between them.

$$d(u, v) \leq n, \forall u, v \in \mathrm{S}$$

and there are no additional nodes that are also at distance n or less from all nodes in the subgraph. But note that the path of length n or less linking a member of the n-clique to another member may pass through an intermediary who is not in the group.

When n = 1, the subgraphs are cliques, since all nodes are adjacent. 2-cliques are subgraphs in which all members are not adjacent, but all members are reachable through at most one intermediary.

Example 5. For example in the fig. 3 {1,3,5}, {2,6,4}, {1,2,6} are some of the 2-cliques. It is important to note that in {1,3,5}, the geodesic distance between two nodes 1 and 3 contains node 2 which is not a part of 2-clique.

Thus an n-clique can be defined on the basis of geodesic that pass through nodes not in n-clique. The n-clique may not even be connected.

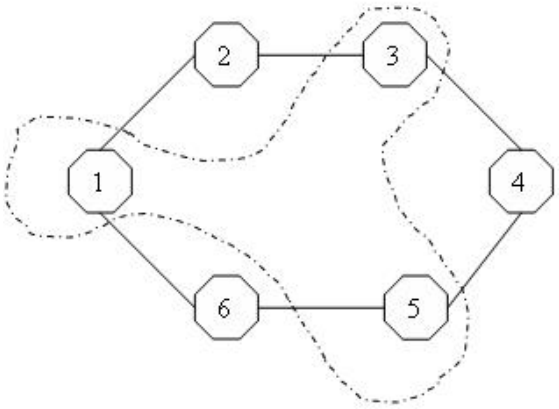

Fig. 3. Graph to demonstrate N-Cliques. Here {1,3,5}, {2,6,4} and {1,2,6} are some of the 2-cliques of the graph.

Example 6. Considering fig. 4, all the stars of degree 3: e.g. {1, 2, 3, 8} etc.; all the cycles C,: e.g. {1, 2, 8, 7} etc.; and the two cycles C,: {1, 2, 3,4, 5} and {6, 7, 8, 9, 10} form 2-cliques.

Example 7. Also, taking another fig. 5, {1, 2, 3, 4, 5, 7} and {1, 3, 5, 6, 7, 8} form 2-cliques.

b) **n-clan**
An n-clan is a n-clique in which the geodesic distance, $d(i, j)$ between all nodes in the subgraph is no greater than n for paths taken within the subgraph. The n-clans in a graph can be found by examining all n-cliques and excluding those that have diameter greater than n, for example in fig. 3, {1,3,5} is not a 2-clan because the diameter of this subset is ∞. Any n-cliques that require non-subgraph members are excluded from consideration. An n-clan is an n-clique with diameter (D) less than or equal to n.
For the fig. 3, among the mentioned 2-cliques, only {3,4,5} and {2,1,6} are 2-clans.
Also, interestingly in fig. 4, all the 2-cliques also form 2-clans. It is mainly because the intermediate nodes in the shortest path between any two nodes in the clique are also present in the clique. But, in fig. 5, no 2-clique forms a 2-clan. {1, 2, 3, 4, 5, 7} is not a 2-clan because the diameter of the subgraph induced is 3 which is larger than n i.e. 2. The same reason holds for the other clique.

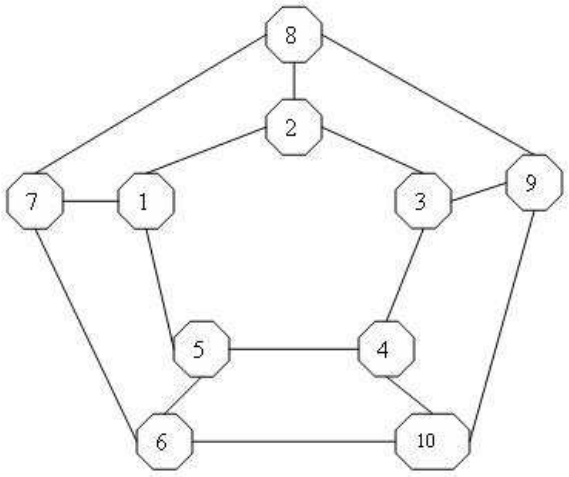

Fig. 4. Sample graph where all 2-cliques are 2-clans, namely all the stars of degree 3: e.g. {1, 2, 3, 8} etc.; all the cycles C,: e.g. {1, 2, 8, 7} etc.; and the two cycles C,: {1, 2, 3,4, 5} and {6, 7, 8, 9, 10}.

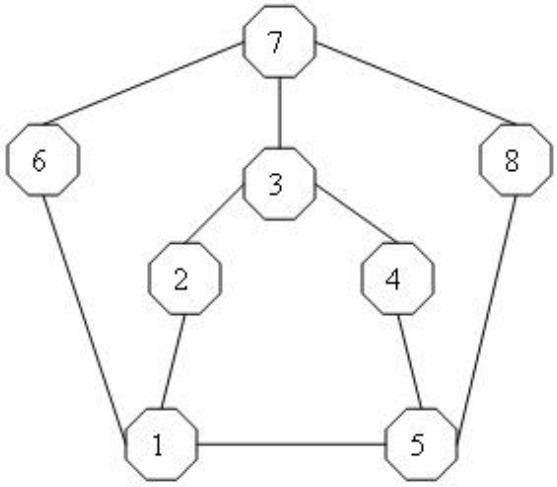

Fig. 5. Graph G where there are no 2-clans.

4.2 N-club

An n-club is defined as maximal subgraph of diameter n. That is an n-club is a subgraph in which diameter between all nodes within the subgraph is less that or equal to n and no nodes can be added that have geodesic distance n or less from all members of the subgraph. It is important to note that an n-club need not to be an n-clique but an n-clan has to be.Also, every n-clan is both an n-club and an n-clique.
For the fig. 6, 2-cliques:- {1,2,3,4,6} and{1,3,4,5,6}
2-clan:- {1,3,4,5,6}
2-club:- {1,2,3,4} ,{1,2,3,6} and {1,3,4,5,6} Note here, {1,2,3,4} ,{1,2,3,6} are not n-clans but n-cliques.

2. **Subgroups Based on Degree**
Whereas n-cliques, n-clans and n-clubs all generalize the notion of clique via relaxing distance, the k-plex generalizes the clique by relaxing density.

4.3 *k*-plex :

A k-plex is a maximal subgraph of n nodes where each node is adjacent to at least n - k nodes in the subgraph. So, 1-plex represents a clique.

Example 8. For example in Fig. 6, the set {1, 2, 3, 4} fails to be a 2-plex because each member must have at least 4-2=2 ties to other members of the set, yet 4 has only one tie within the group. In Fig. 8, the set {1,2,4,5}

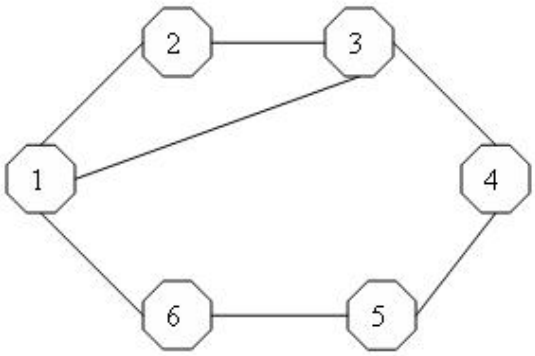

Fig. 6. Sample Network for calculation of N-club for the network.

is a 2-plex, as each of the nodes in the set have degree ≥ 2. In fig. 7, {1, 2, 3, 4} form a 1-plex as the subgraph induced by them forms a clique.

Property: If $k < \frac{(n+2)}{2}$, then diameter $D \not> 2$.
Proof: We will use the proof by contradiction to prove it. Let S be the set of nodes in k-plex, where $k < \frac{(n+2)}{2}$, and $u, v \epsilon S$ be at a distance > 2 i.e. there is no common neighbor of u and v.
$So, N(u) \bigcap \mathrm{N(v)} = 0$
Let $N(u)$ be the number of neighbours of u and $N(v)$ be the number of neighbours of v.
For any vertex t ϵ S,

$$
\begin{aligned}
&N(t) > n - \tfrac{(n+2)}{2} \\
&\Rightarrow N(t) > \tfrac{(n-2)}{2} \\
&\Rightarrow N(t) > (\tfrac{n}{2} - 1) \\
&\Rightarrow N(t) \geq \tfrac{n}{2} \\
&Also, \\
&N(u) \cup N(v) = N(u) + N(v) - N(u) \bigcap N(v) \\
&\text{so } \mathrm{N(u)} \bigcap \mathrm{N(v)} = 0 \\
&N(u) + N(v) = \tfrac{n}{2} + \tfrac{n}{2} = n
\end{aligned}
$$

On the basis of above acquired result we obtain that there are at least $(n+2)$ vertices including u & v, which is a contradiction.

4.4 *k*-core

Another approach to cohesive subgroups based on nodal degree is the k-core. A k-core is a subgraph in which each node is adjacent to at least a minimum number k of the other nodes in the subgraph. In contrast to k-plex which specifies the acceptable number of lines that can be absent from each node, the k-core specifies the required number of lines that must be present from each node to others within the subgraph.
A k-core is defined as maximal subgraph where each node is adjacent to at least k nodes in the subgraph. Hence, every member of a 2-core is connected to at least 2 other members, and no node outside the 2-core is connected to 2 or more members of the core (otherwise it would not be maximal). Every k-core contains at least $k + 1$ vertices, and vertices in different k-cores cannot share edges. Because if they share a common edge, then we can join both of the subgraphs to form a subgraph of greater size.
A 1-core is simply a component. Also, the minimum size of maximal subset k-core is $(k+1)$.k-cores can be described as loosely cohesive regions which will contain more cohesive subsets. For example, every k-plex is contained in a k-core.

Example 9. In the fig. 7 {1,2,3,4} forms 3-core.

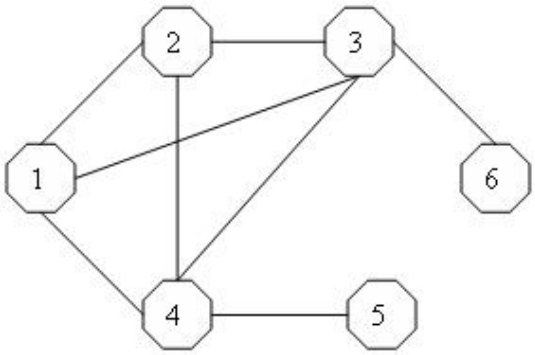

Fig. 7. Figure Illustration for calculation of K-core

4.5 LS-Sets

An LS-set is a subgraph definition that compares ties within a subgraph to ties outside the subgraph by focusing on greater frequency of ties among subgroup member compared to ties from subgroup member to outside.
Definition:
Given a graph G (V, E), let H be the subset of V and K be any proper subset of H then
H is LS if

$$\alpha(K, H-K) > \alpha(K, V-H),\ \forall K \subset H \tag{6}$$

where α(S1,S2) denotes the number of ties from the set S1 to the member of set S2.
i.e. All subsets of LS-set are more connected to LS –members than nodes outside of LS-set.

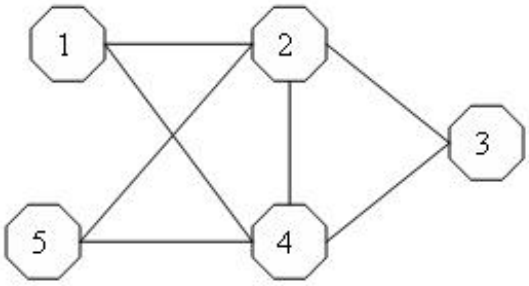

Fig. 8. Initial network configuration for calculation of LS-sets

Example 10. Like in Fig. 8, the set {1,2,4,5} is not an LS set since α({2,4,5},{1}) $\not>$ α({2,4,5},{3}).Here,putting H = {1,2,4,5}, K = {2,4,5} and V = {3} in Eq. 6, we get α({2,4,5},{1}) = 2 whereas α({2,4,5},{3}) = 3.

Example 11. In contrast, the set {1,2,4,5} in Fig. 9 does qualify as an LS set.

Taking another example for the figure given below {1,2,3,4} also does not form a LS-set. Convince yourself with taking H = {1, 2, 3, 4}, K = {2, 3, 4} and V = {5, 6, 7, 8} and putting them in Eq. 6.

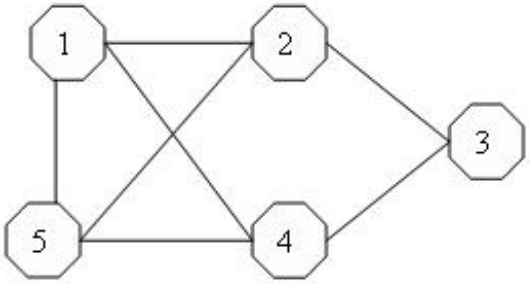

Fig. 9. Modified form of previous graph. Note {1,2,4,5} form LS sets in this modified graph but not in the original.

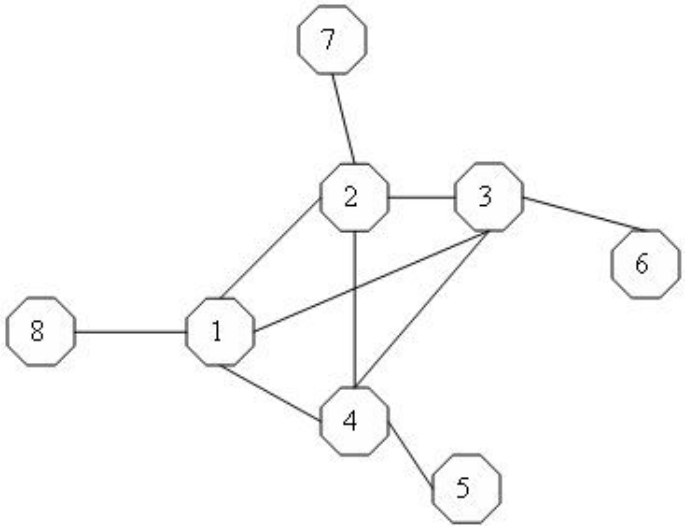

Fig. 10. Figure Illustration of a network for finding LS-sets

5 Centrality

Within graph theory and network analysis, there are various measures of the **centrality** of a vertex within a graph that determine the relative importance of a vertex within the graph (for example, how important a person is within a social network or, in the theory of space syntax, how important a room is within a building or how well-used a road is within an urban network). Or some social problems like, in a hypothetical sexual interaction network, who are

most likely to be affected by STDs such as AIDS. In simple words, Centrality of a node measures its Power, Prestige, Prominence and imPortance.
There are four measures of centrality that are widely used in network analysis: degree centrality, betweenness, closeness, and eigenvector centrality.

1. **Degree Centrality**
   The first, and simplest, is **degree centrality**. Degree centrality is defined as the number of links incident upon a node (i.e., the number of ties that a node has). Degree is often interpreted in terms of the immediate risk of node for catching whatever is flowing through the network (such as a virus, or some information). If the network is directed (meaning that ties have direction), then we usually define two separate measures of degree centrality, namely indegree and outdegree. Indegree is a count of the number of ties directed to the node, and outdegree is the number of ties that the node directs to others. For positive relations such as friendship or advice, we normally interpret indegree as a form of popularity, and outdegree as gregariousness.

$$DC(v) = \frac{d(v)}{E(G)}$$

   a) Find the neignbors of a node.
   b) Repeat for all nodes.

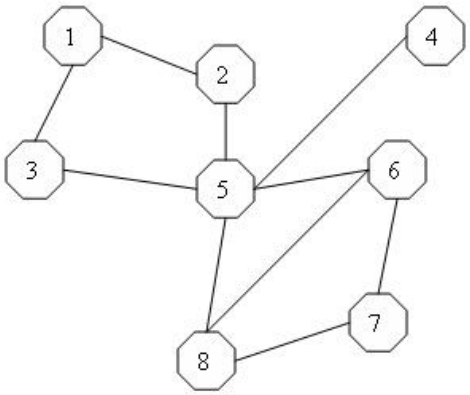

Fig. 11. Figure Illustration of a network for finding centralities

Example 12. For fig. 11, the degree coefficients for each of the nodes are given as:

	1	2	3	4	5	6	7	8
D(v)	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{4}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{3}{20}$

So Node 5 has highest degree coefficient of 0.2, hence it is the most central node in the network according to degree centrality.

Calculating degree centrality for all nodes in a graph takes $\Theta(\mathrm{V}^2)$ in a dense adjacency matrix representation of the graph(step 1 takes $\Theta(\mathrm{V})$ and it repeats n times), and $\Theta(E)$ in a sparse list representation(only have to visit neighbors of all nodes).

2. **Betweenness Centrality**
Betweenness is a **centrality** measure of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.
For a graph $G: = (V, E)$ with n vertices, the betweenness $CB(v)$ for vertex v is:

$$CB(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the number of shortest geodesic paths from s to t, and $\sigma\_st(v)$ is the number of shortest geodesic paths from s to t that pass through a vertex v.

Example 13. For fig. 11, Lets try to find betweenness centrality of node 1, from node 3 to node 2 there are two shortest paths, namely 3-¿1¿2 and 3-¿5-¿2. So out of the two, one passes through node 1. Similarly from node 3 to node node 4, one of the two shortest paths passes through node 1. Hence,$CB(1) = \frac{1}{2} \times 2$. Similarly, the betweenness coefficients for each of the nodes are given as:

	1	2	3	4	5	6	7	8
CB(v)	$\frac{1}{2} \times 2$	$\frac{1}{1} \times 6 + \frac{1}{2} \times 4$	$\frac{1}{2} \times 4$	0	$\frac{1}{1} \times 12$	$\frac{1}{2} \times 5$	0	$\frac{1}{2} \times 5$

So Node 5 has highest betweenness coefficient of 12, hence it is the most central node in the network according to betweenness centrality.

The values may be normalized by dividing through by the number of pairs of vertices not including v, which is ${}^{n-1}C_2 = (n - 1)(n - 2)/2$. As this can be the possible maximum numerator for any node. The normalizing factor in the above example will be (8-1)(8-2)/2 = 21.
The normalized coefficients of each of the nodes are given as:

	1	2	3	4	5	6	7	8
CB(v)	$\frac{1}{21}$	$\frac{8}{21}$	$\frac{2}{21}$	0	$\frac{12}{21}$	$\frac{5}{42}$	0	$\frac{5}{42}$

Calculating the betweenness and closeness centralities of all the vertices in a graph involves calculating the shortest paths between all pairs of vertices on a graph. This takes $\Theta(V^3)$ time with the Floyd–Warshall algorithm. On a sparse graph, Johnson's algorithm may be more efficient, taking $O(V^2 \log V + VE)$ time.

3. **Closeness Centrality**
Closeness is a centrality measure of a vertex within a graph. Vertices that are 'shallow' to other vertices (that is, those that tend to have short geodesic distances to other vertices within the graph) have higher closeness. In the network theory, **closeness** is a sophisticated measure of centrality. It is defined as the reciprocal of sum of geodesic distances (i.e the shortest path) between a vertex v to all other vertices reachable from it:

$$C(v) = \frac{1}{\sum_{t \in V \setminus v} dG(v,t)}$$

Example 14. For fig. 11, closeness coefficient for node 1 = $\frac{1}{1+1+2+2+3+4+3}$
Similarly, working for each of the nodes in fig. 11, we get:

	1	2	3	4	5	6	7	8
CB(v)	$\frac{1}{16}$	$\frac{1}{12}$	$\frac{1}{14}$	$\frac{1}{18}$	$\frac{1}{10}$	$\frac{1}{13}$	$\frac{1}{18}$	$\frac{1}{13}$

From the table, we find that CB(5) = 0.1 is the highest in the graph, so it is the most central node according to closeness centrality criteria.

More the closeness coefficient of a vertex, more the node will central in the network. Closeness can be regarded as a measure of how long it will take information to spread from a given vertex to other reachable vertices in the network.

4. **Eigenvector Centrality**
Eigenvector centrality is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Google's Page Rank is a variant of the Eigenvector centrality measure.
For the *ith* node, let the centrality score be proportional to the sum of the scores of all nodes which are connected to it. Hence

$$x_i = \frac{1}{\lambda} \sum_{j \in M(i)}^{n} x_j = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} x_j$$

where $M(i)$ is the set of nodes that are connected to the *ith* node, N is the total number of nodes and λ is a constant. In vector notation this can be rewritten as

$$X = \frac{1}{\lambda} AX \text{ or as the eigen vector equation } X\lambda = AX$$

In general, there will be many different eigenvalues λ for which an eigenvector solution exists. However, the additional requirement that all the entries in the eigenvector be positive which implies (by the Perron–Frobenius theorem) that only the greatest eigenvector corresponding to the eigenvalue results in the desired centrality measure. The i^{th} component of the related eigenvector then gives the centrality score of the i^{th} node in the network. Power iteration is one of many eigenvalue algorithms that may be used to find this dominant eigenvector.

a) **Procedure to find Eigenvector Centrality**
   i. Set $e_i = 1,\ \forall i \in n$.
   ii. Calculate $e_i^* = \sum_{j=0}^{n} a_{ij} e_j,\ \forall i \in n$
   iii. $\lambda = \sum_{i=0}^{n} e^*$.
   iv. Then e_i is updated to $e_i = \frac{e_i}{\lambda},\ \forall i$.
   v. Step 2 to 4 are repeated until λ stops changing.

Example 15. For fig. 11, we will work out the eigenvector centrality coefficients for each of the nodes. Each of the rows constitute of a round, and next round is worked on previous round's values.

	1	2	3	4	5	6	7	8
L1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
L2	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{4}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{3}{20}$
L3	$\frac{5}{56}$	$\frac{7}{56}$	$\frac{6}{56}$	$\frac{3}{56}$	$\frac{11}{56}$	$\frac{9}{56}$	$\frac{6}{56}$	$\frac{9}{56}$
L4	$\frac{13}{162}$	$\frac{25}{162}$	$\frac{16}{162}$	$\frac{7}{162}$	$\frac{31}{162}$	$\frac{26}{162}$	$\frac{18}{162}$	$\frac{26}{162}$

So after Round 4, we can see that the eigenvector centrality coefficient for node 5 is the highest in the graph. So it is the most central node according to eigenvalue centrality criteria.

6 Equivalence

Network analysis most broadly defines two nodes (or other more elaborate structures) as similar if they fall in the same "equivalence class." There are many ways in which actors could be defined as equivalent based on their relations with others. For example, we could create two equivalence classes of actors with out-degree of zero, and actors with out-degree of more than zero. Equivalence have been particularly useful in applying graph theory to the understanding of social roles and structural positions.

1. **Structural Equivalence**

   Two nodes are structurally equivalent if they have same relationships to all other nodes. In real life complete structural equivalence is not feasible.

Consequently we measure the degree of similarity between two nodes. We can measure degree of structural equivalence from the adjacency matrix of the network.

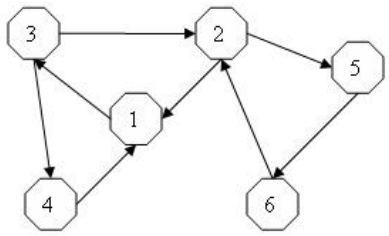

Fig. 12. Sample directed Graph for illustration of structural equivalence.

The adjacency matrix of above figure is:

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

From the above matrix, the outdegree matrix A can be written as:

$$A = \begin{pmatrix} 0\,0\,1\,0\,0\,0 \\ 1\,0\,0\,0\,1\,0 \\ 0\,1\,0\,1\,0\,0 \\ 1\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,1 \\ 0\,1\,0\,0\,0\,0 \end{pmatrix}$$

The indegree matrix A^T can be written as:

$$A^T = \begin{pmatrix} 0\,1\,0\,1\,0\,0 \\ 0\,0\,1\,0\,0\,1 \\ 1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,1\,0 \end{pmatrix}$$

Combining A and A^T, we get A':

$$A' = \begin{pmatrix} A \\ A^T \end{pmatrix} = \begin{pmatrix} 0\,0\,1\,0\,0\,0 \\ 1\,0\,0\,0\,1\,0 \\ 0\,1\,0\,1\,0\,0 \\ 1\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,1 \\ 0\,1\,0\,0\,0\,0 \\ 0\,1\,0\,1\,0\,0 \\ 0\,0\,1\,0\,0\,1 \\ 1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,1\,0 \end{pmatrix}$$

The equivalence is now calculated by measuring the co-relation among the columns. Each column signifies a node. Now there are different strategies to measure the relation among the columns. We discuss some of them below.

a) **Pearson's Correlation Coefficient**

Pearson's correlation coefficient between two nodes can be measured as follows.

$$\gamma_{xy} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_j}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_j^2 - \frac{\left(\sum y_j\right)^2}{n}\right)}}$$

where x and y are nodes represented as columns in matrix A' (described before), and n is the number of rows in each column. In order to calculate Pearson's correlation coefficient (PCC) the following steps are followed:

Calculate A which contains the outdegrees of each nodes, and then the indegrees of each nodes. Indegree can be actually calculated from A, by taking its transpose i.e. A^T

Next the following matrix is created:

$$\begin{pmatrix} \gamma_{11} & & & & & \\ \gamma_{12} & \gamma_{22} & & & & \\ \gamma_{13} & \gamma_{23} & \gamma_{33} & & & \\ \gamma_{14} & \gamma_{24} & \gamma_{34} & \gamma_{44} & & \\ \gamma_{15} & \gamma_{25} & \gamma_{35} & \gamma_{45} & \gamma_{55} & \\ \gamma_{16} & \gamma_{26} & \gamma_{36} & \gamma_{46} & \gamma_{56} & \gamma_{66} \end{pmatrix}$$

Example 16. Since Pearson correlation coefficient (PCC) between same nodes is 1. For fig. 12, the PCC matrix will be as follows:

$$\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 1 & & & & & \\ -\sqrt{2} & 1 & & & & \\ -1 & -\sqrt{2} & 1 & & & \\ -\frac{1}{\sqrt{2}} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 1 & & \\ 0 & -1 & 0 & -\frac{1}{2} & 1 & \\ -\frac{1}{\sqrt{2}} & -1 & -\frac{1}{\sqrt{2}} & -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}$$

Since γ_{BD} is maximum (neglecting the diagonal elements) so nodes B and D are said to have structural equivalence. Actually, more the γ value between the nodes, more is the equivalence between them.

b) **Euclidian Distance**

Euclidian distance is the measure of dissimilarity between two nodes. It uses Geodesic Distance[2] to calculate the shortest distance between two nodes.

We construct matrices G and G^T, where an entry (x,y) shows the shortest distance cost from x to y.

G =

	A	B	C	D
A	0	2	1	2
B	1	0	2	3
C	2	1	0	1
D	1	3	2	0

G^T=

	A	B	C	D
A	0	1	2	1
B	2	0	1	3
C	1	2	0	2
D	2	3	1	0

[2]Geodesic Distance is the shortest distance between two nodes.

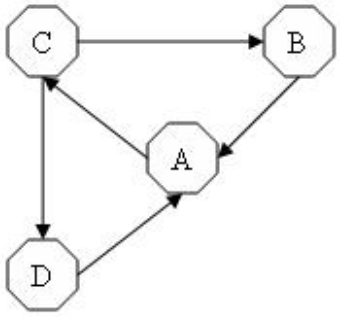

Fig. 13. Euclidian distance is calculated for the above network, to illustrate structural equivalence.

Now Putting G and G^T together we get :

$$D = \begin{pmatrix} G \\ G^T \end{pmatrix} = \begin{pmatrix} 0\,2\,1\,2 \\ 1\,0\,2\,3 \\ 2\,1\,0\,1 \\ 1\,3\,2\,0 \\ 0\,1\,2\,1 \\ 2\,0\,1\,3 \\ 1\,2\,0\,2 \\ 2\,3\,1\,0 \end{pmatrix}$$

Next the following matrix is created: $P_{ED}M =$

$$\begin{pmatrix} D_{AA} & & & \\ D_{AB} & D_{BB} & & \\ D_{AC} & D_{BC} & D_{CC} & \\ D_{AD} & D_{BD} & D_{CD} & D_{DD} \end{pmatrix}$$

Euclidian Distance $= D_{AB} = \sqrt{\sum_{i=1}^{2n}(A_i - B_i)^2}$

For fig. 13, calculating the ED matrix:

$$\begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{pmatrix} 0 & & & \\ \sqrt{17} & 0 & & \\ \sqrt{14} & \sqrt{17} & 0 & \\ \sqrt{17} & \sqrt{36} & \sqrt{17} & 0 \end{pmatrix}$$

If Euclidian Distance between two nodes is relatively more then it implies that those two nodes are relatively more different. Hence, for fig. 13 we see that ED_{AD} is the minimum ($\sqrt{14}$), and so they are the most equivalent nodes.

c) **Percentage of Exact Match**

In some cases the nodes we are measuring may be measured at the nominal level. In such data applying Euclidian distance or Pearson's correlation coefficient will be misleading. For binary/categorical/discrete data this measure is useful. Here if we want to measure percent exact match between two nodes A an B, then we shall see how many row values of these columns are matching. For example, let column corresponding to A is (0 1 0 1 0 0 1 0) and column corresponding to B is (0 0 1 0 1 0 0 0). So total matches are three out of eight i.e. 37.5 %.

$$EM_{xy} = number\ of\ i's\ s.t.\ x_i = y_i, \forall i$$

Example 17. Calculating $P_{EM}M$ for fig. 13, and using the same A' matrix used earlier,

$$\begin{array}{c} A \\ B \\ C \\ D \end{array} \begin{pmatrix} 1 & & & \\ \frac{3}{8} & 1 & & \\ \frac{1}{4} & \frac{3}{8} & 1 & \\ \frac{3}{8} & 1 & \frac{3}{8} & 1 \end{pmatrix}$$

Here too, Percentage of exact match turns out highest for nodes B and D. Problem of percent match is if the data is too sparse then more absence (0) indicates percent matches higher. For solving this problem we can use Jaccard Coefficients(discussed next).

d) **Jaccard's Coefficients**

In this method, we ignore absence of data and consider number of matches with respect to presence of data.
Let, S is No. of positive matches
$P1$ is Non Null elements in A
$P2$ is Non Null elements in B
Then $JC = \frac{S}{S+P1+P2}$
In other words it can also be represented as, For any nodes A and B,

$$JC_{AB} = \frac{A \bigcap B}{A \bigcup B}$$

where the numerator represents the number of matching non-null entries of A and B, and the denominator represents the number of entries which are non-null in at least one of them.

Example 18. Like for fig. 13, $JC_{AB} = \frac{0}{5} = 0$ Calculating the entire matrix for the figure,

$$\begin{array}{c} A \\ B \\ C \\ D \end{array} \begin{pmatrix} 1 & & & \\ \frac{0}{5} & 1 & & \\ \frac{0}{6} & \frac{0}{5} & 1 & \\ \frac{0}{5} & \frac{2}{2} & \frac{0}{5} & 1 \end{pmatrix}$$

Here, JC_{BD} has the maximum Jaccard's Coefficient, showing B and D are the most equivalent nodes.

2. **Regular Equivalence**
Two nodes are said to be *regularly equivalent* if they have the same profile of ties with other nodes who are also regularly equivalent. This is also known as automorphic equivalence. For example, in the fig. 14 nodes (1, 2), (3, 4) and (5, 6) are regular equivalent nodes.

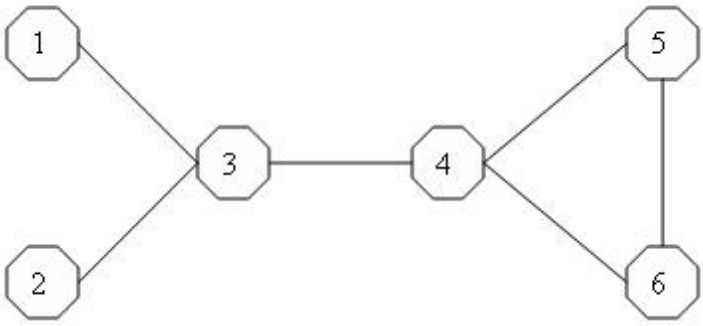

Fig. 14. Network Illustration showing regular equivalance of nodes (1, 2), (3, 4) and (5, 6)

Measuring Similarity (Regular Equivalence):
We make profile of each node to measure their regular equivalence.
An $n \times n$ matrix is made and an entry (x,y) indicates the shortest distance between x and y.
Referring figure 14:
Calculating the profiling matrix :

	1	**2**	**3**	**4**	**5**	**6**
1	0	2	1	2	3	3
2	2	0	1	2	3	3
3	1	1	0	1	2	2
4	2	2	1	0	1	1
5	3	3	2	1	0	1
6	3	3	2	1	1	0

The profile columns are sorted and group the nodes which are same.

	1	**2**	**3**	**4**	**5**	**6**
1	0	0	0	0	0	0
2	1	1	1	1	1	1
3	1	2	1	1	1	1
4	2	2	1	1	2	2
5	3	3	2	2	3	3
6	3	3	2	2	3	3

The profile columns of 1 and 2 are same. Similarly, (3, 4) and (5, 6) pairs show regular equivalence.

7 Hierarchical Clustering

Hierarchical cluster analysis is a statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. It starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. When there are N cases, this involves N-1 clustering steps, or fusions.
Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process of Johnson's (1967) hierarchical clustering is this:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 can be done in different ways, which is what distinguishes *single-link* from *complete-link* and *average-link* clustering. In *single-link* clustering (also called the *connectedness* or *minimum* method), we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If the

data consist of similarities, we consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster. In *complete-link* clustering (also called the *diameter* or *maximum* method), we consider the distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster. In *average-link* clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.

Example 19. For fig. 13,suppose the threshold is 0.95. Now, calculating the Pearson's coefficients for each of the nodes

$$\begin{matrix} A \\ B \\ C \\ D \end{matrix} \begin{pmatrix} 1 & & & \\ -0.45 & 1 & & \\ 0.60 & -0.45 & 1 & \\ -0.45 & 1 & -0.45 & 1 \end{pmatrix}$$

Since $\gamma_{BD}(1.0)$ is the only one above the threshold(0.95) so nodes B and D are collapsed into one. Corresponding new value of correlation coefficients are put as:

$$\begin{matrix} A \\ B/D \\ C \end{matrix} \begin{pmatrix} 1 & & \\ -0.46 & 1 & \\ -0.6 & -0.7 & 1 \end{pmatrix}$$

Now suppose the threshold value is -0.5. The only value greater than this is $\gamma_{A\&B/D}$. So, next A and B/D are merged together:

$$\begin{matrix} A/B/D \\ C \end{matrix} \begin{pmatrix} 1 & \\ 0.7 & 1 \end{pmatrix}$$

If we choose threshold lesser than 0.7, say 0.6, then A/B/D and C can be merged together. Finally,

$$A/B/D/C\,(1)$$

Example. The following pages trace a hierarchical clustering of distances in kms between Indian cities. The method of clustering is *single-link*.

Input distance matrix:

	mum	kgp	kol	hyd	del
mum	0	2500	1960	800	1750
kgp	2500	0	120	2000	1800
kol	1960	120	0	2100	2500
hyd	800	2000	2100	0	3900
del	1750	1800	2500	3900	0

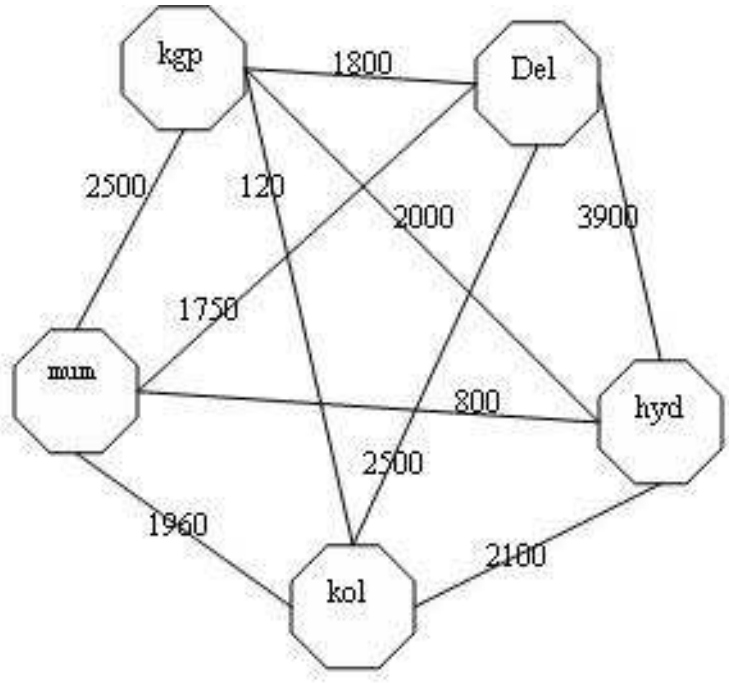

Fig. 15. Network of cities with inter city distance as edge costs.

The nearest pair of cities is kgp and kol, at distance 120. These are merged into a single cluster called "kgp/kol" or "k/k".
Then we compute the distance from this new compound object to all other objects. In single link clustering the rule is that the distance from the compound object to another object is equal to the shortest distance from any member of the cluster to the outside object. So the distance from k/k to mum is chosen to be 1960, which is the distance from kol to mum. Similarly, the distance from k/k to del is chosen to be 1800.

After merging kgp with kol:

	mum	k/k	hyd	del
mum	0	1960	800	1750
k/k	1960	0	2000	1800
hyd	800	2000	0	3900
del	1750	1800	3900	0

The nearest pair of objects is mum and hyd, at distance 800. These are merged into a single cluster called "mum/hyd" or "m/h". Then we compute the distance from this new cluster to all other clusters, to get a new distance matrix:

After merging mum with hyd:

	m/h	k/k	del
m/h	0	1960	1750
k/k	1960	0	1800
del	1750	1800	0

Now, the nearest pair of objects is del and m/h, at distance 1750. These are merged into a single cluster called "d/m/h". Then we compute the distance from this new cluster to all other objects, to get a new distance matrix:

After merging del with m/h:

	d/m/h	k/k
d/m/h	0	1800
k/k	1800	0

Finally, we merge the last two clusters at level 1800. The process is summarized by the following clustering diagram.

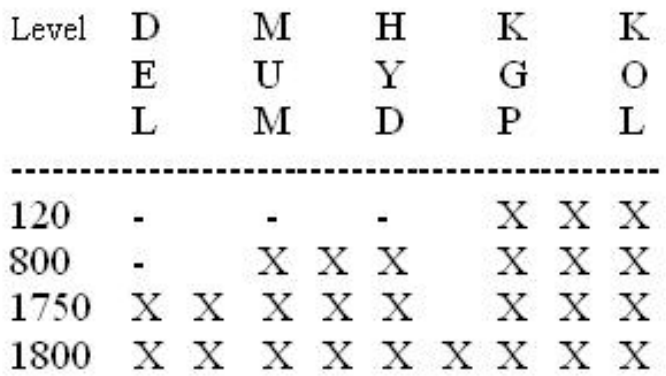

Fig. 16. Final resultant table after each level calculations.

In the diagram 16, the columns are associated with the items and the rows are associated with levels (stages) of clustering. An 'X' is placed between two columns in a given row if the corresponding items are merged at that stage in the clustering.

8 Assortativity

8.1 Introduction

Out of several different possible explanation of the formation of communities in the social network, assortativity is most prominent. It states "rich mixes with rich". Many such networks including social networks, computer networks, and biological networks show assortative mixing on their degrees, i.e., a preference for high-degree vertices to attach to other high-degree vertices. Others show disassortative mixinghigh-degree vertices attach to low degree ones.

8.2 Detailed Example

Consider the following men-women marriage relation depending on their race. Here $e_{(i,j)}$ element of the matrix gives the fraction of men of i race marry women of j race.

	black	hispanic	white	others	$a_i = \sum_j e_{ij}$
black	0.258	.016	.035	.013	.322
hispanic	.012	.157	.058	.019	.246
white	.013	.023	.306	.035	.377
others	.005	.007	.024	.016	.052
$b_j = \sum_i e_{ij}$	.289	.204	.423	.084	

Diagonal elements represent the fraction of couples in partnership with members of their own group and off-diagonal those in partnership with members of other group. Inspection of the matrix shows that the matrix has considerably more weight along its diagonal than off it indicating that assortative mixing does take place. The amount of assortative mixing in a network can be quantified by measuring how much of the weight in the mixing matrix falls on the diagonal and how much off it. Let us define e_{ij} to be the fraction of all edges in a network that joins the vertex of type i with type j. According to the matrix, we can say that index i represents man and j represents female which makes the matrix asymmetric. The matrix should satisfies these equations,

$$\sum_{ij} e_{ij} = 1,$$

$$\sum_i a_i = 1,$$

$$\sum_j b_j = 1,$$

where a_i and b_j are the fraction of each type of end of an edge that is attached to vertices of type i. The assortativity measure Q can be defined as,

$$Q = \frac{\sum_i e_{ii} - 1}{N - 1}$$

So for completely assortative network, $\sum_i e_{ii}$ will be $\sum_i^N (1) = N$.
Hence, $Q = \frac{N-1}{N-1} = 1$
And for completely random networks, e_{ii} can be written as e_i.
And, $\sum_i^N e_i = 1$. So, $Q = \frac{1-1}{N-1} = 0$
Now we define a quantitative measure r of the level of assortative mixing in the network,

$$r = \frac{Tr(e) - ||e^2||}{1 - ||e^2||}$$

Trace of a matrix represented by Tr(e), is the sum of all diagonal elements. Thus,

$$r = \sum_i \frac{(e_{ii} - a_i b_i)}{(1 - a_i b_i)}$$

Finally,

$$r = \frac{\sum_{i=1}^{n} e_{ii} - \sum_{i=1}^{n} a_i b_i}{1 - \sum_{i=1}^{n} a_i b_i}$$

It takes the value 1 for the perfectly assortative network, since in that case, the entire weight of the matrix lies along its diagonal. Conversely, if there is no assortative mixing at all, then r becomes quite low. Networks can also be disassimilative: vertices may associate preferentially with others of different types - the opposite attracts phenomenon. In that case, r becomes negative.

8.3 Assortativity mixing in networks

Many of the models that have been successful in reproducing features of networks in the real world are mainly based upon preferential attachment model in which the probability of a given source vertex forming a connection to a

target vertex is some function of the degree of the target vertex. Many such networks show "assortative mixing" on their degrees, i.e., a preference for high-degree vertices to attach to other high-degree vertices.

Now, we consider a simple undirected network of N vertices and M edges. And j_i, k_i be the degrees of the vertices at the ends of the i^{th} edge, with $i = 1...M$
Then, Assortativity measure r can be defined as the definition of Pearson Coefficient as

$$r = \frac{1}{\sigma_q^2} \sum_{jk} jk(e_{jk} - q_j q_k)$$

where,

$$q_k = \frac{(k+1)p_k + 1}{\sum\limits_j jp_j}$$

and p_k is the degree distribution of the graph.
Also, definition of variance $\sigma_q^2 is$:

$$\sigma_q^2 = \sum_k k^2 q_k - [\sum_k kq_k]^2$$

Finally, we get the equation for r:

$$r = \frac{M^{-1} \sum\limits_i j_i k_i - [M^{-1} \sum\limits_i \frac{1}{2}(j_i + k_i)]^2}{M^{-1} \sum\limits_i \frac{1}{2}(j_i^2 + k_i^2) - [M^{-1} \sum\limits_i \frac{1}{2}(j_i + k_i)]^2}$$

8.4 Assortativity Clustering

Suppose we are given the probability distribution matrix of a network i.e. e_{ij} represents the probability of having an edge between any two nodes i and j. And also we have the total number of edges m.
Then, we will be able to calculate the total number of edges from node i to rest of the graph by multiplying e_{ij} with m, we call it m_i. We now consider all the edges from node i in one Group, say group I.

Our aim is to calculate the number of nodes in each group. Let group I has n_i number of nodes.
We know that average degree of a node $Z = \sum kp_k$.
So, total edges from a node i will be

$$m_i = n_i.Z$$

Or,

$$n_i = \frac{m_i}{Z}$$

Now, for each group we know the nodes in that group and degree distribution in the group. Now, according to the degree distribution, create the stubs for each of the nodes in the group. We know the number of edges between i-i, i-j and so on. So, m_{ii} times join the stubs internally in the I group, m_{ij} times from I group to J group, and so on.

Continuing for all groups, we will get cluster graph between all the groups using the method of assortative clustering.

8.5 Assortativity and Epidemics

Assortativity mixing can have many practical applications, for example, for the spread of disease on social networks social networks being assortatively mixed in many cases. The core group of an assortatively mixed network could form a "reservoir" for disease, sustaining an epidemic even in cases in which the network is not sufficiently dense on average for the disease to persist. On the other hand, one would expect the disease to be restricted to a smaller segment of the population in such cases than for diseases spreading on neutral or disassortative networks.
Assortative mixing also has implications for questions of network resilience. It has been found that the connectivity of many networks (i.e., the existence of paths between pairs of vertices) can be destroyed by the removal of just a few of the highest degree vertices, a result that may have applications in, for example, vaccination strategies. In assortatively mixed networks, however,it has been found that removing high-degree vertices is a relatively inefficient strategy for destroying network connectivity, presumably because these vertices tend to cluster together in the core group, so removing them becomes somewhat redundant.

9 Networks in Real World

Network in real world can be divided into four loose categories:

1. Social Network
2. Information Network
3. Technological Network
4. Biological Network

1. **Social Network**

A social network is a set of people or groups of people with some pattern of contacts or interactions between them. The patterns of friendships between individuals, business relationships between companies, and intermarriages between families are all examples of social networks. The properties of these networks will be discussed in detail later in the section.

2. **Information Network**
   Our second network category is what we will call information networks (also sometimes called "knowledge networks"). The classic example of an information network is the network of citations between academic papers. These citations form a network in which the vertices are articles and a directed edge from article A to article B indicates that A cites B.
   a) **Citation Network :**
      The citations form a network in which the vertices are articles and a directed edge from article A to article B indicates that A cites B. Citation networks are acyclic because papers can only cite other papers that have already been written, not those that have yet to be written. Thus all edges in the network point backwards in time, making closed loops impossible or at least extremely rare (see Fig. 17).
      Alfred Lotka's groundbreaking 1926 discovery of the so-called Law of Scientific Productivity, states that the distribution of the numbers of papers written by individual scientists follows a power law (see Fig. 18). That is, the number of scientists who have written k papers falls off as k-$\propto$ for some constant $\propto$.
   b) **World Wide Web Network :** Another very important example of an information network is the World Wide Web, which is a network of Web pages containing information, linked together by hyperlinks from one page to another.
      Unlike a citation network, the World Wide Web is cyclic; there is no natural ordering of sites and no constraints that prevent the appearance of closed loops (See fig. 19). The Web also appears to have power-law in- and out-degree distributions.
      From the figure below it would be clear that {1,2,3,6,4,1} forms a cycle.

3. **Technological Network**

   Our third class of networks is technological networks, man-made networks designed typically for distribution of some commodity or resource, such as electricity or information. The electric power grid is a good example. Another very widely studied technological network is the Internet, i.e., the network of physical connections between computers.

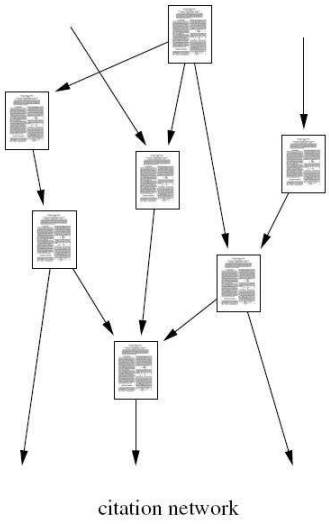

Fig. 17. The citation network of academic papers in which the vertices are papers and the directed edges are citations of one paper by another.

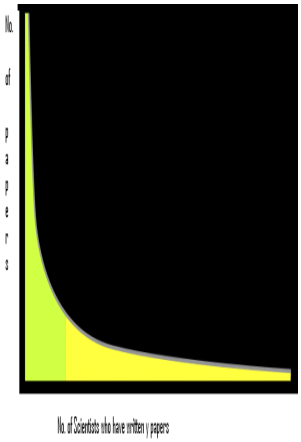

Fig. 18. Graph showing power law distribution for the numbers of scientists versus the number of k papers they have written.

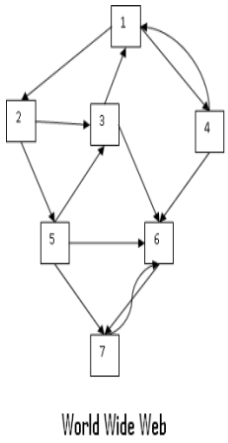

Fig. 19. The World Wide Web, a network of text pages accessible over the Internet, in which the vertices are pages and the directed edges are hyperlinks. There are no constraints on the Web that forbid cycles and hence it is in general cyclic.

4. **Biological Network**

An important class of biological network is the genetic regulatory network. The expression of a gene, i.e., the production by transcription and translation of the protein for which the gene codes, can be controlled by the presence of other proteins, both activators and inhibitors, so that the genome itself forms a switching network with vertices representing the proteins and directed edges representing dependence of protein production on the proteins at other vertices. Another much studied example of a biological network is the food web, in which the vertices represent species in an ecosystem and a directed edge from species A to species B indicates that A preys on B.