Social Media and Network Analysis

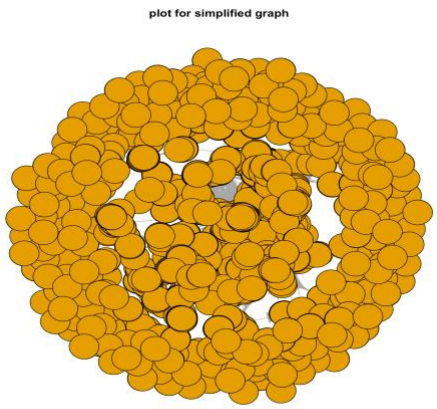
Pawanjeet Kaur

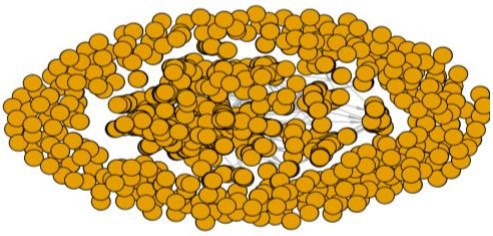
Advanced Lab 2: Implications of Network Structure of the SAP Online Knowledge Community Platform

SAP community dataset is one of the large datasets, providing the community interactions of users. The left column of the csv provides us the users who have answered the questions in threads posted. As part of the first network analysis, we will be exploring the network structure of SAP community. Dataset contains the directed graph data connecting left column to right column. It represents an induced subgraph with 10 percent of nodes from complete dataset collected.

As part of the analysis, first we will look into the overall structure of the graph visually and by the numbers. This community structure shows us that it is not a simple graph, means, users are connected multiple times and are connected to self, i.e. have loops, with **6090** edges and **3415** vertices.

We have simplified the graph in R, which means we took the count of the number of links between two nodes and assigned it as weight of the edge. After simplifying the graph, the edges reduce to **4120** and vertices remain same **3415**.Below are the images of the graphs with inv-weight:





A network can be described using many metrics like adjacency matrix, centrality, betweenness, closeness etc. In a sense, we are analyzing here local structure of the graph as we are using some part of the huge network for our analysis.

Reciprocity: One of the most important measures to analyze any social network is reciprocity. Reciprocity is used to find the importance of the nodes in a network. Especially, for directed graphs we can see how much two nodes reciprocate i.e. if nodes are mutually linked. SAP community graph shows reciprocity of **0.005825243.** Quantitatively, this ratio defines the number of relations which are reciprocated over the total number of relations in the network. This metric is only used in directed graphs as we are seeing the bi-directional relation between nodes to see how much each node reciprocates. This small number shows that the network is unidirectional most of the relations are from left column to right column.

Connected or Not? A graph is connected if there exists a path from every node to every other node. Network under consideration is not connected graph, which means every user is not connected to every other user i.e. there exists some users who have not answered questions posted by some users.

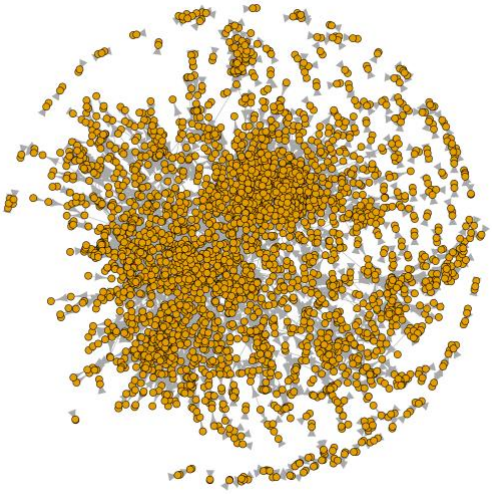
Strong/Week: A strong component is a maximal subgraph in which there is a path from every point to every point following the directed path. A weak component is a maximal subgraph which would be connected if we ignored the direction. As per our analysis, the network is neither a strong network nor a weak network. This shows that there are some users in the left column which have not participated in the forum at all.

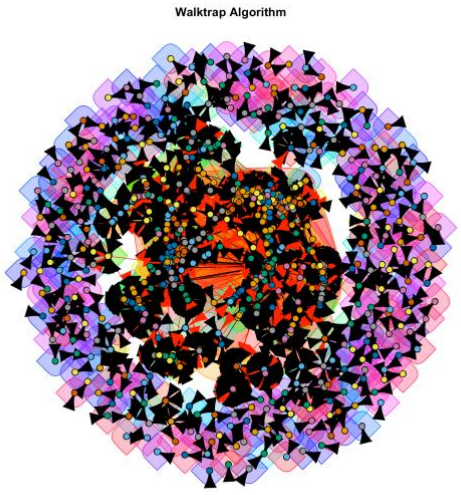
Transitivity: It is the property of the graph which states if node A is connected to node B and node B is connected to node C then node A is related to node C as well. It is also a quantitative measure which gives the extent to which friends of mine are friends to each other. For network in discussion, transitivity **0.009985725** gives the fraction of 3 times the number of triangles in the network with number of nodes with edges to unordered pairs of nodes.
Extent of triadic closure can be expressed using clustering coefficient and transitivity.

Average Path Length: It is the average number of steps along the shortest path for all possible pairs of network nodes. For SAP community network, the average path length is **3.982714** i.e. to reach from one node to another at average we have to take 4 steps.

Diameter is the longest path in the network. If we pick two distant nodes in the graph the shortest path between these two nodes is called diameter. For graph in discussion, if we take simplified graph, two users are at length 26 whereas with inverted weight the distance reduces to **14.2**

Now, to analyze the **community structure** in the network we will use different layout algorithms. First, we use '**Fruchterman layout**' to see the overall structure of the network. Also, tried Circle layout whereas there is no interpretation using that layout. We will see the overall structure using the "**Walk-trap algorithm**" as fast greedy one of the good algorithms to detect community is only used for undirected graphs and spring glass can also not be used as it does not work for unconnected networks.

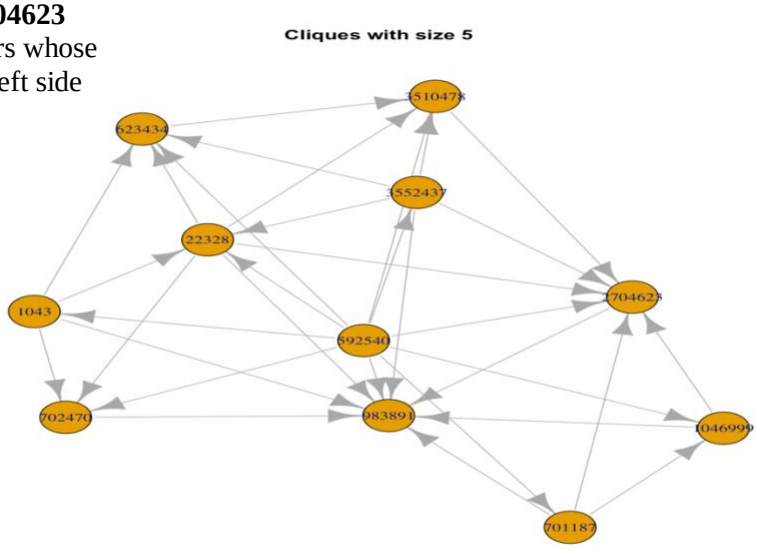


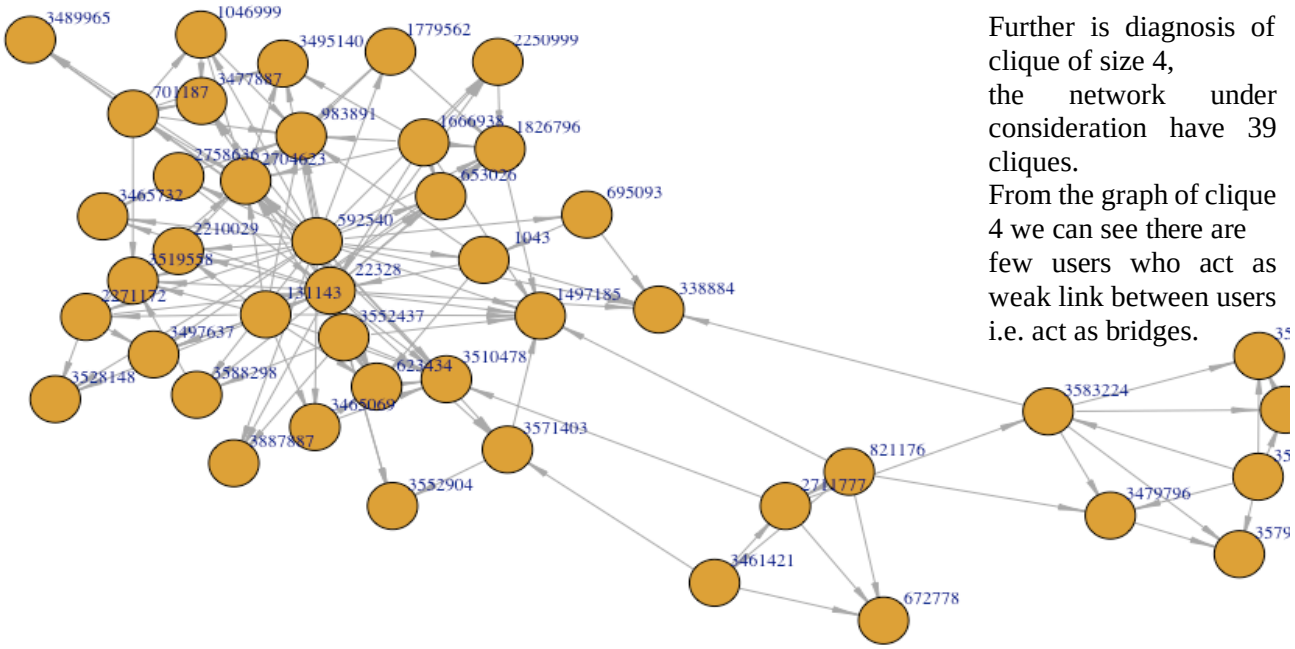


Clique: To further analyze graphs we will use a very important metric known as clique. A clique of size n gives us the fully connected graph with nodes n from a set of m nodes. We will look into the clique with size of 2,3 ,4 and 5.

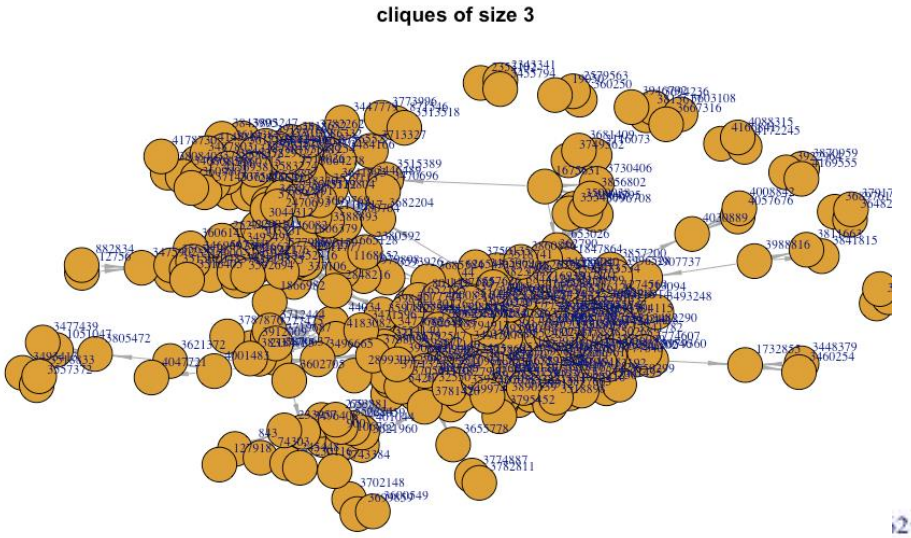
Size	Number of Clique
2	3320
3	335
4	39
5	5

It has 5 cliques of size 5 with nodes **983891 and 2704623** with highest inward connection means these are users whose questions are answered by most of the users on the left side of the dataset.





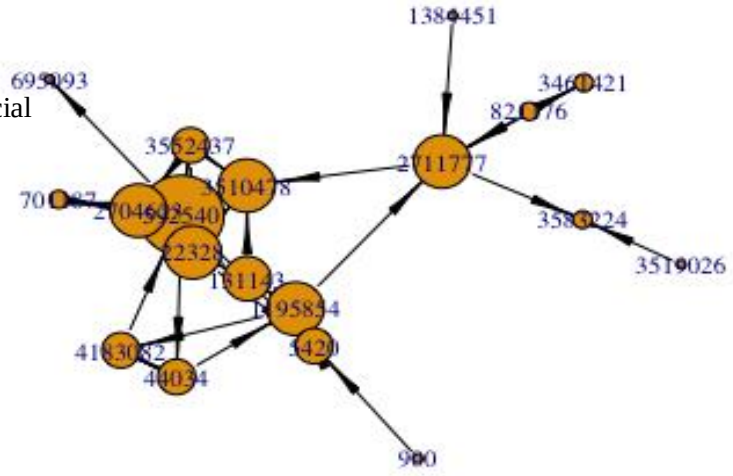
Further is diagnosis of clique of size 4, the network under consideration have 39 cliques.
From the graph of clique 4 we can see there are few users who act as weak link between users i.e. act as bridges.

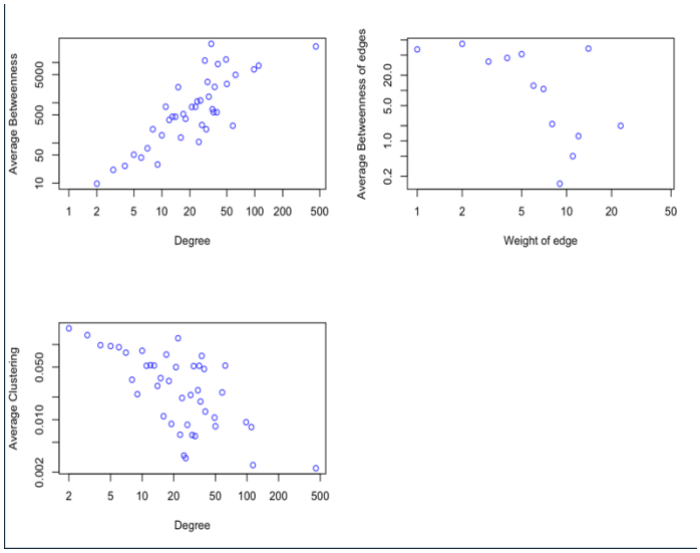


Left is graph of clique size 3, we have 335 components of clique 3 and 3320 of size 2.

Clique 3, as 3 nodes are participating, also defines the extent of triadic closure. There are 335 strong triadic closures present in the network.

Centrality is the most important metric in any social network analysis as it tells about central node i.e. the most important node in the graph or in social sense who is the most important person.



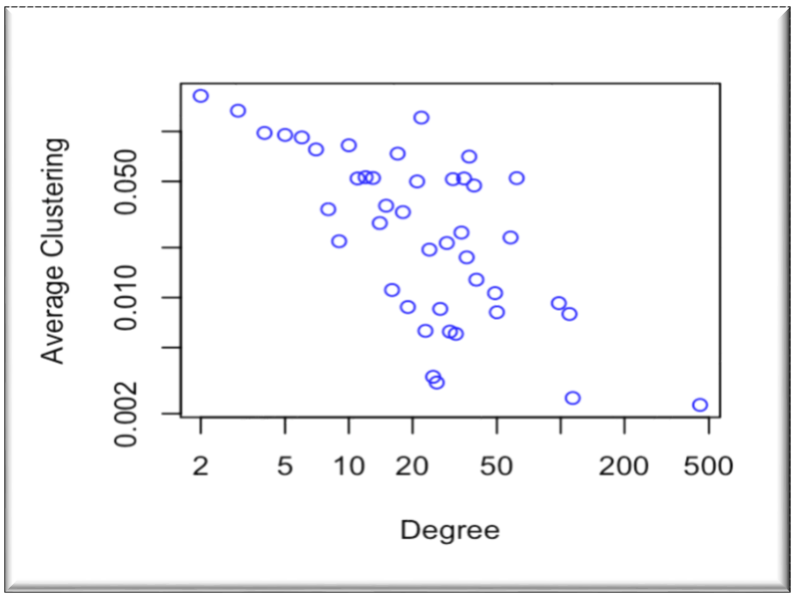


Degree Centrality: Degree of any node is the number of edges it has. A node with a higher number of edges interprets to a more important node in the network. Network in discussion is a directed graph for which the in-degree i.e. the incoming edges gives us important inference about the users.

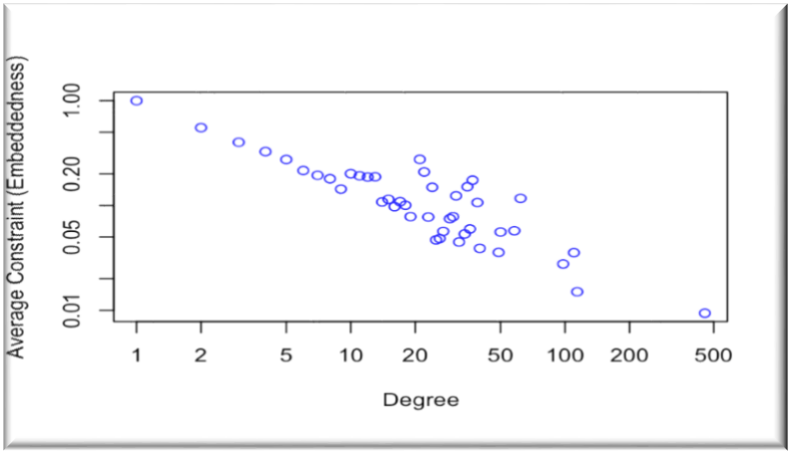
Betweenness Centrality: It is a measure of how many shortest paths pass by given edge/node namely, edge and node betweenness. High value of betweenness centrality gives us important nodes in the network.

Authority Centrality: There can be users/nodes which are not central but are an important part of the network. Authority gives us how much knowledge is possessed by a specific node and hub is how well a node knows where to find information on a given topic.

Closeness Centrality: Importance of any node is defined by how close a node is to another node.



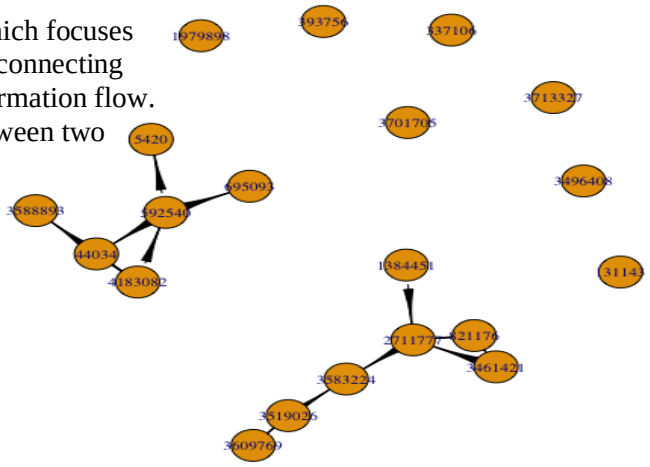
Clustering coefficients: "In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together" **Source: Wikipedia**
It gives the quantitative measure of how likely the nodes tend to cluster together. whereas, the degree of the node defines how many neighbors it connected. So, nodes or users answering questions to the same questions are more likely to be in one cluster and have high clustering coefficient. When the degree increases, the average clustering decreases.



Embeddedness: It is the measure of lack of overlap in the network. High overlap of nodes means there is high content flow from one node to another. i.e. nodes with high degree will have low embeddedness. From the plot on left, with increasing degree, the embeddedness is decreasing.

Structural Holes: This is one of the measures which focuses on control overflows within the network. A node connecting to other vertices in the network is a source to information flow. A structural hole is the absence of connection between two neighbors of an actor of specific node.

User with ID **592540 and 2711777** are structural nodes in SAP community graph.



Degree Distribution: At last to see the overall network structure we can also look into the degree distribution of the SAP community network. Degree distribution is the probability distribution of nodes in the entire network. As the network consists of a huge number of nodes, we take the log on both axes to shrink. Below two graphs show relation of degree with its intensity and vertex degree with average neighbor degree.