SOCIAL ANALYTICS

IDS 564 – ADVANCED LAB 1

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

Submission By-

Shivam Duseja

UIN: 678895805

In this report, we will analyze the SAP Online Knowledge Community dataset. This dataset provides information about the users posting questions on the SAP Online forum and the users responding to these questions. In the analysis of this social network, the users are represented by nodes and edges are directed from the user who provides an answer to the posted question. This analysis is based on 10% users (nodes) drawn from the dataset provided by Prof. Peng Huang, of the University of Maryland.

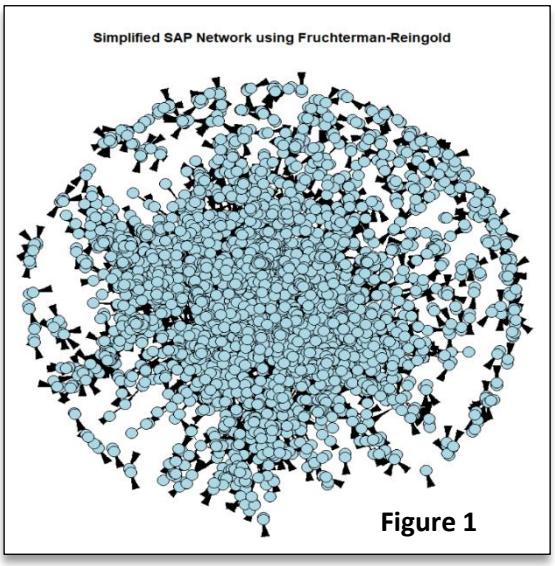


Figure 1

We will start our analysis by looking at the number of nodes – 3415 and number of edges – 6090. From the community structure, we can see that a single user can answer multiple questions posted by other users. This means that there might be a pair of users having multiple connections (cyclic loop) i.e. the graph is not simplified. We have then simplified the graph by assigning the number of links between two nodes as the weight of their edge. The simplified graph using Fruchterman-Reingold is displayed in **Figure 1.**

We will further analyze this community network using different metrics like reciprocity, transitivity, betweenness, degree etc. and drive useful insights from the same.

From **Table 1**, we can see that this community network is not strongly/weakly connected i.e. the graph is not connected even if we ignore the edge directions. Hence, we can conclude that there are some users within the community that have not answered any question posted on the SAP user forum.

Viewer Zoom

Table 1: R Output

Metric_Name	Metric_Val
Reciprocity	0.005825
Transitivity	0.009986
IsConnected	✖ No
IsConnected-Strong	✖ No
IsConnected-Weak	✖ No

Reciprocity has a very small value which implies that the network is more of unidirectional. This is evident from the fact that the node direction represents an answer provided to a question and generally the users providing answers on the SAP forum will be very less when compared to the total users accessing the forum.

Transitivity – the global clustering coefficient (3 x # of closed triplets / # of triplets) also has a very small value i.e. the degree to which the users in the community tend to form groups/clusters together is less. A general thing to notice in the social networks is that people/users are experts in their respective fields/departments which leads to the formation of clusters of such departments/fields. From the transitivity value, we can infer that there will be a smaller number of users who will respond to most of the questions posted within the forum across different departments/offices (common nodes across the entire network will be small).

Avg. Path Length: From **Table 2**, average path length is in general the average distance between two nodes, which in our case will be the avg. distance between two users. This is a famous property mentioned by S Milgram – 'Six Degrees of Separation'. This property says that all individuals are six or

fewer connections(social) away from each other. As a general concept, this (Low Avg. Path length) is also one of the two main conditions of the 'Small World' concept – the other condition being high value of clustering coefficient. For our case, we can say that the average distance between the users accessing the SAP community forum is 3.98 i.e. ~4 steps for information to reach from 1 user to another.

Viewer Zoom

Table 2: R Output

Path_Metric	Path_Metric_Vals
Avg. Path Length	3.982714
Diameter	14.272278
Diam.- numeric weights	26.000000
Diam.- Inv Weights	14.272278

Diameter: Generally, diameter represents the shortest distance between two farthest nodes. In our case, we can say that two distant users are at a distance of 26 whereas this distance reduces to 14.2 if we use inverted weights.

Cliques and their Structures:

From **Table 3**, we can see that there are **5 cliques of size 5** i.e. these nodes represent the users that have received multiple answers to the question they posted on the forum.

Viewer Zoom

Table 3: R Output

clique_size	clique_number
2	3320
3	335
4	39
5	5

Cliques of Size 4: From the table, we can see that there are 39 cliques of size 4. **Figure 2** - displays subgraph showing the cliques of size 4. From the graph, we can see that there are some nodes that act as bridges in transferring information to the neighboring nodes (concept of structural holes).

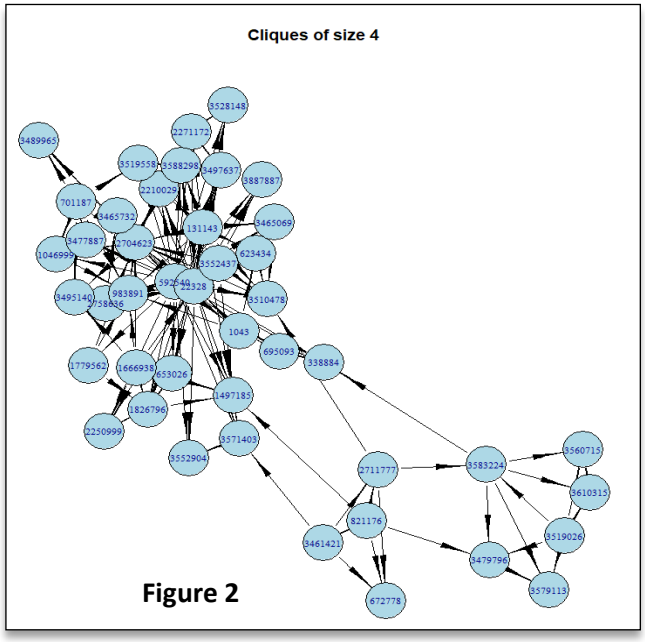


Figure 2

Cliques of Size 3: There are 355 cliques of size 3. These cliques also display the strength of strong triadic closure existing in the network (355 cliques).

Cliques of Size 2: The network has 3320 cliques of size 2 i.e. 3320 users are connected to atleast one user on the SAP forum.

Structural Holes:

In **Figure 3,** we have plotted Top-25 nodes with the highest structural Hole access i.e. 25 nodes with the lowest constraint value. In general, the structural holes concept, developed by Ronald Burt, relies on the fact that novel information is generally higher within a group of people rather than between the groups. The theory says that – any individual acting as a mediator or bridge between two connected groups is important in a sense that this individual will be transferring information from one group to another. Apart from that this individual can also combine information/ideas coming from different groups and can come up with the most innovative/idea among all the users(nodes). In our case, the nodes – 592540, 821176, 131143 are the users transferring important information among the different groups within the network.

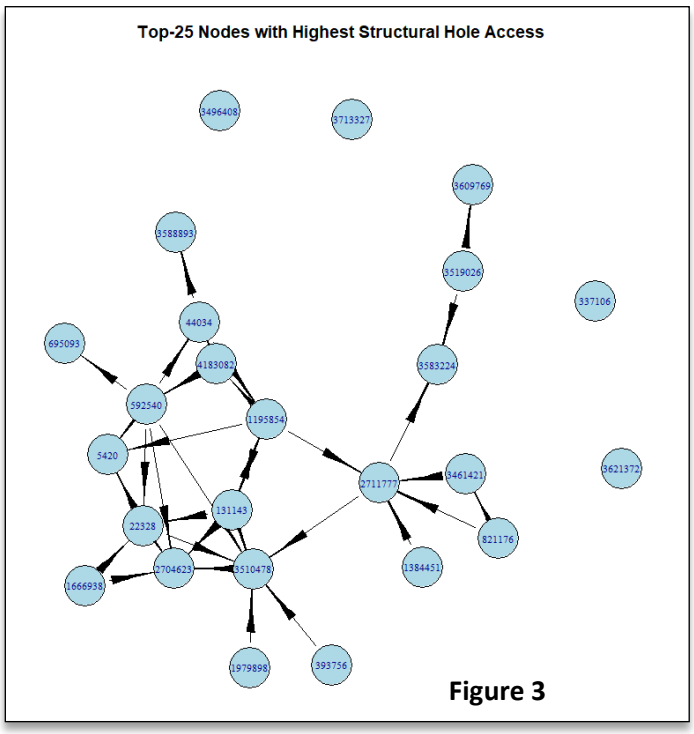


Figure 3

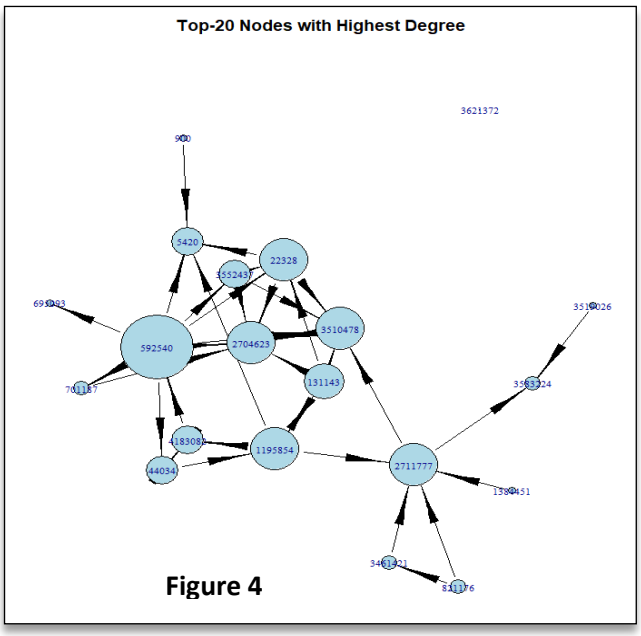


Figure 4

Degree

Generally, degree of a node is analogous to its interaction with the neighboring nodes. **Figure 4**, on the left displays the top-20 nodes with their sizes corresponding to their degree. Hence, we can say that the user(node) 592540 answers the-most number of questions or is most active in the SAP Knowledge forum.

Authority: Let's say – If Node 1 is linked to Node 2 which in turn is linked to many other nodes, we can say that Node 1 will have a higher authority score. This Node 2 will thus help in propagating information from Node 1 to other nodes of the network. Hence, Node 1 will have a higher authority score.

Node Betweenness:

Generally, node betweenness measures the amount of time a node lies between other nodes. Nodes with higher betweenness are really important in a social network setting as these nodes are the ones that transfer most information throughout the network. From **Figure 5**, we can say that user/node – 1195854 is most important as it is transferring the most amount of information in the SAP network. Information transfer within the network will get affected if these nodes with highest betweenness are removed.

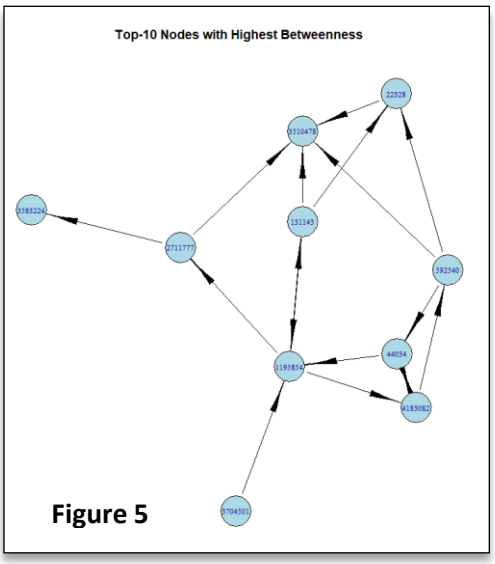


Figure 5

Edge Betweenness:

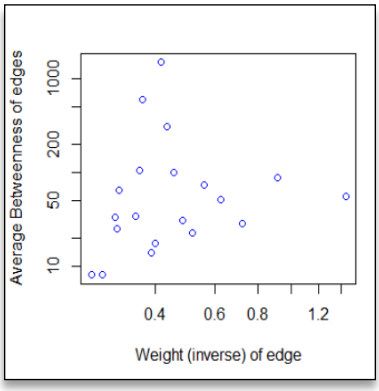
Similar to Node betweenness, Edge betweenness, is the number of times an edge occurs between different nodes/vertices. In our case, we can say that the edges with the highest betweenness values will be transferring the highest information within the network. Edges – 2153, 10, 1272 are the edges with highest betweenness values i.e. Edge 2153 is the most important edge in terms if flow of information in the network.

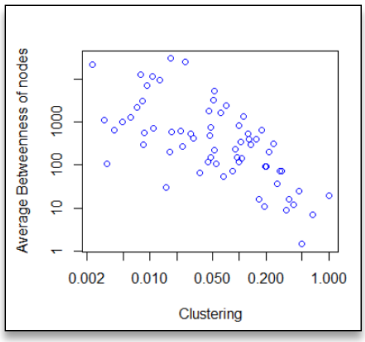
Local Clustering Coefficients:

Generally, local clustering coefficient of a node tells us about how close the node's neighbors are to a complete graph(clique). This metric value also tells us about the strength of clusters in the cliques. Local Clustering coefficient value of NaN means that the node has only 1 neighbor whereas a value of 1 means that the node is strongly present in the cluster.

Avg. Betweenness of Edges v/s Weight of Edge

From the plot, we can observe that on increasing the inverse weight of the edges (except a couple of outliers), the Avg. betweenness of edges increase. Hence, we can say that the Avg. betweenness of the edges has a negative relationship with the weight of the edges.



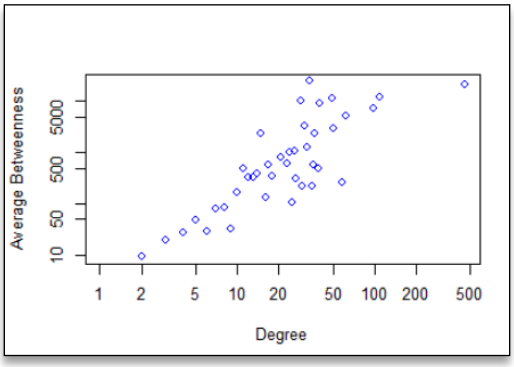


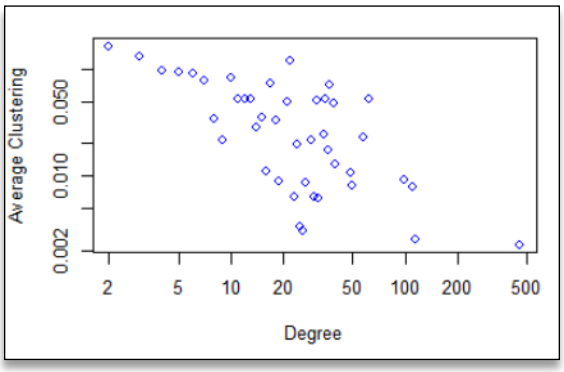
Avg. Betweenness of Nodes v/s Clustering

From the graph, we can observe that as the clustering values increase the Avg. Betweenness of the nodes decrease. This can also be seen from the fact that the nodes/users that are not a part of strong clusters/groups will generally act as a structural hole. Hence, the people with low clustering values will have a high Avg. Betweenness of nodes.

Avg. Betweenness v/s Degree

Avg. Betweenness of a node measures its importance in transferring important information whereas Degree measures the interaction of nodes with its neighboring nodes. Generally, a node with a higher degree might have a high or low Avg. Betweenness, however, cumulated (Avg) value of betweenness will have a positive relationship with the degree of node.



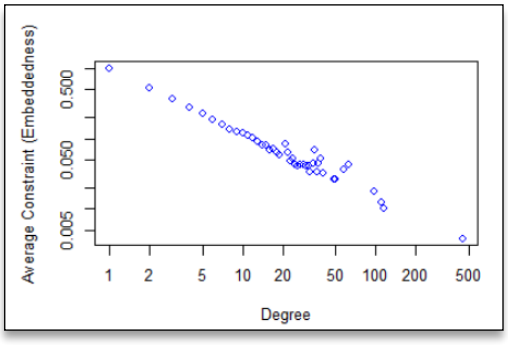


Avg. Clustering v/s Degree

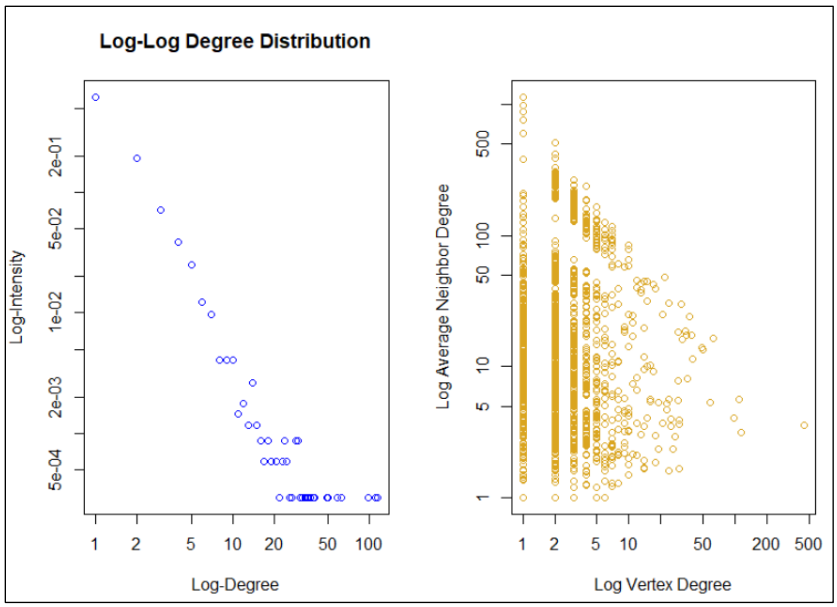
Clustering coefficient tells us about the strength of nodes within a cluster whereas degree tells us about the number of nodes a particular vertex/node is connected to. Generally, people within a single department form stronger groups/clusters and will be connected to lesser number of people/nodes outside the department/cluster. This is clearly evident from the graph above that shows a negative relationship of clustering coefficient and degree.

Avg. Constraint v/s Degree

Embeddedness in terms of social network setting can be defined as the measure of lack of overlap in a particular network. Local bridges for instance, have an embeddedness of zero. In case of a high overlap within the network, the information flow increases, hence, the nodes with high degree will have low embeddedness. The same is visible from the graph on the right.



Log – Log Degree Distribution:



The SAP community comprises of a large number of nodes and in order to visualize these nodes, we will use log axis to view the degree distribution of the SAP community. The above two graphs – show that there is a negative relationship between log degree and its log intensity.

Community Detection Algorithms

As this is a directed graph, we have used Walk-Trap and Girvan Newman algorithm to detect different communities within the network.