YouTube Network Analysis

SOCIAL MEDIA AND NETWORK ANALYSIS – IDS 564

BALA ROHIT YERUVA, SANJAY DASARI, LAAVANYA GANESH, VIJETA SHAH

YouTube is the largest video sharing website and the second most popular website in the internet domain. It allows users to watch and share videos with others. With the growing impact of digital presence, several corporate firms have created their own brand channels in YouTube and run various marketing campaigns. While YouTube promotes these marketing campaigns to the users at a premium, firms also need to engage with the user community for better returns. In order to engage with the user community, one needs to understand this network better. In this submission, we made an attempt to build a network graph and draw several inferences that can create significant business value.

1. Data Description:

The data is obtained from the Social Computing Laboratory repository of Arizona State University[1]. The overall network has 15088 nodes and approximately 1 million to 3 million edges. In building the community graphs, the network has been downscaled in order to get better visualizations.

The data set has five files as listed below

1.1 Contact Network Data: It lists out friends of each individual obtained from Google circles. There are two columns in the data set. 'X' in first column and 'Y' in second column suggests that X is a friend of Y.

1.2 Shared Friends Data: This file mentions the number of common friends between two friends. This file has 3 columns where the first two columns mentions the node names of the friends and the third column gives out the total number of common friends between two users.

1.3 Shared Subscriptions Data: Similar to the shared friend's data, this file captures the number of common subscriptions between two users.

1.4 Shared subscribers Data: This file helps to identify, the total number of common followers for two particular users.

1.5 Shared Favorite Videos Data: In this data, we get the details of number of video that are commonly liked and shared by two individuals.

Additional Notes on Data Selection: For the project, we have built different network diagrams using these details instead of one composite network. The reason is that each of the network diagrams would reveal a different characteristic of the network. For example, the shared subscriptions and favorite videos network would reveal how similar the interests of two individuals are but the contact network would reveal how vast the reach of an individual is.

2. Research Goal:

This project has 2 main research goals as mentioned below.

There are 2 main features that any social network requires:

- a strong engagement platform for users to keep them linked to the network and hence help the network to grow
- a good source of revenue generation.

This can be achieved through advertisements and campaigns. With this end in mind, we have 3 research goals:

Study the network: We aimed at building a social network and understanding it thoroughly through various centrality measures. At each step, we understood the significance of these parameters and tried to interpret the business impact of the same.

2.1 User Centric network structure analysis: Understand how users react to subscriptions and interact with each other in order to know what recommendations to give them thus increasing their satisfaction index improving the network quality. This in turn will attract more users to the network and enable passive users to in turn become more active

2.2 Business Centric Advertisement and Campaign strategy: How to place Ad's in the network. who are the users who can propagate the information? which in turn would make a campaign successful.

3. Summary Statistics:

We used the standard summary statistics of the network to draw a few inferences about the nature of the interaction present in the network.

3.1 Assortativity Co-efficient: This statistic measures the ability of higher degree nodes connecting with other higher degree nodes. Our initial guess was that this measure would be higher for the Shared Subscriptions and Favorite Videos network. However, the measure for both these networks is 0.129 and 0.167. One reason could be that the network is large and the number of higher degree nodes are comparatively small relative to the total network size which has more than 15000 nodes. Hence, a separate iteration of this measure has been performed on the largest components of the

"Shared Subscriptions" and "Favorite Videos" network. As expected, the assortativity degree was more than 0.6 for both the networks.

3.2 Clustering Co-efficient: The clustering co-efficient for all the networks is tabulated below. For a network of this size, the clustering co-efficient is also higher (close to .5). This shows the adherence of triadic closure property in the network to the most extent. Naturally, if two users are sharing a common friend with whom their interests in subscriptions and favorite videos match then the two users in question are highly likely to be online friends. Please note that, we are not generalizing this behavior for all the networks. This is being stated as an observation as it is interesting to see such a clustering co-efficient even though the network size is huge.

3.3 Other Network Measures

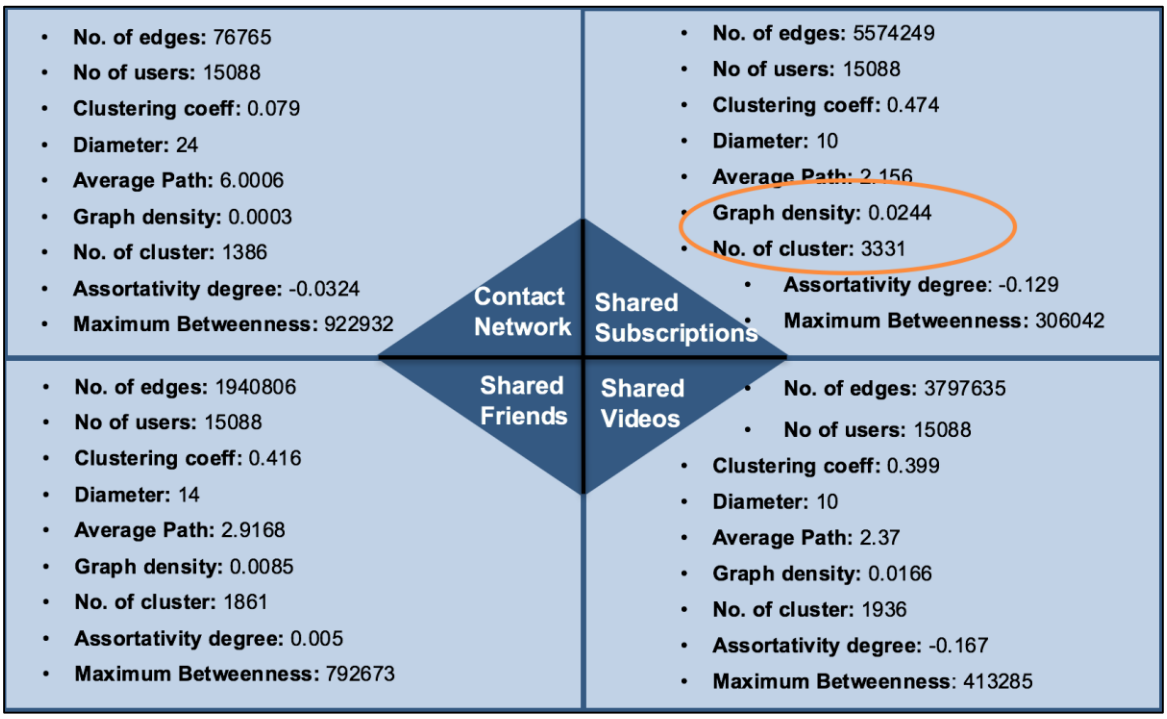

Figure 3.3: Other Network Measures

We ran the above summary statistics on the 4 networks namely Contact Network Data, Shared Friends Data, Shared Subscriptions data & Shared Favorite Videos Data.

We observed that the Network of Shared Subscriptions has the highest Graph density and the highest number of clusters. Hence we felt that this network is of much importance for YouTube's advertisement placement strategy and marketing campaigns.

4. Construction of Network:

4.1 Shared Friends Network:

The below diagram shows the network structure of various users. An edge between two nodes indicate that the corresponding users are friends.

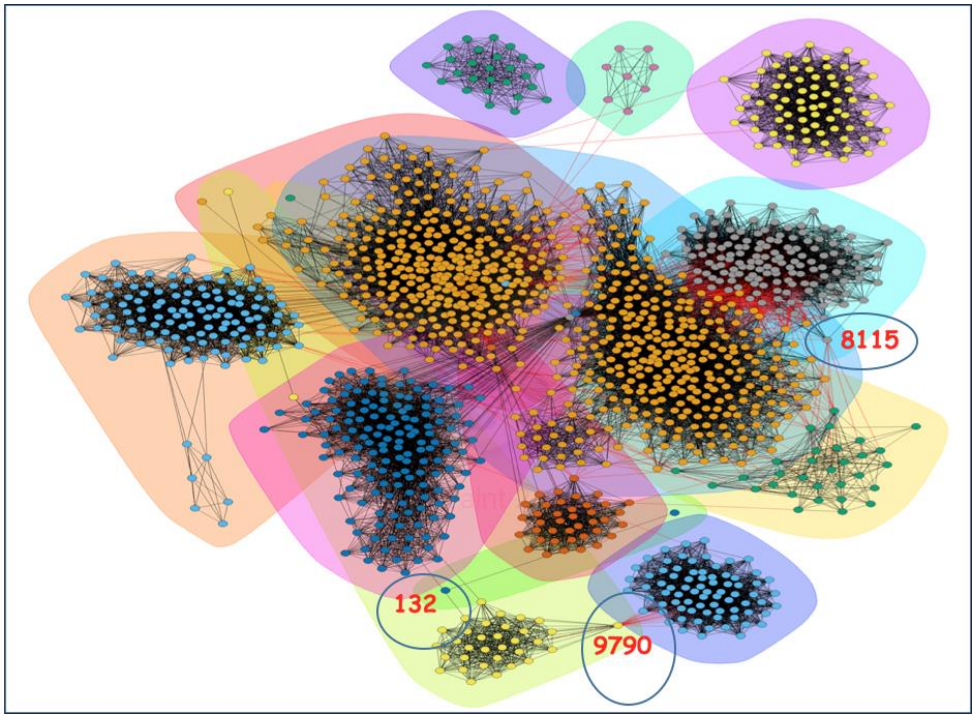

Figure 4.1: Shared Friends Network

4.1.1 The following inferences can be made about the above network:

a. The contact network and shared friends network are similar in nature. The shared friends network has one additional detail namely the number of common friends between two people. For this reason, we skipped the analysis on contact network and preferred to analyze shared friends network as it is rich in analysis.
b. The nodes have a few "giant components" along with a few isolated nodes. The components may have formed according to a geographic location. At a global scale it is more likely that friends are mostly from the same country or region. For example, all the users from India would have formed a giant component.
c. The isolated nodes may refer to users who have recently joined the network or they may be passive users who are not interested in adding or building friendships online.
d. A few nodes such as 24, 162, 4930, 218, 4971,106 are highly influential nodes in the network.

4.1.2 Cliques v/s Frequency

These nodes are among the top nodes in the entire network in terms of degree.

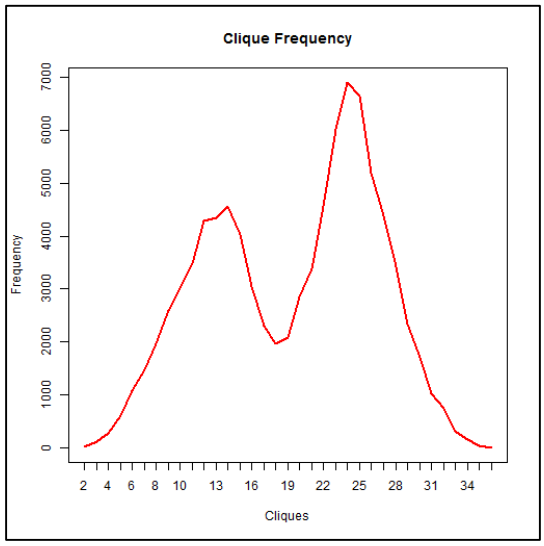

Figure 4.1.2: Cliques v/s Frequency for Shared Friends Network

From the clique frequency graph we can see that, the highest number of cliques have 25 shared friends followed by 14 shared friends. These individuals within maximal cliques would have similar preferences. Hence learning the preferences of the high degree individuals within the maximal cliques would give an appropriate prediction of other users within the cliques. In order to recommend appropriate videos and aim at maximum propagation of a marketing campaign, we need to target these cliques of 25 shared friends in different communities.

To diffuse information across different regions (different communities) videos need to be pushed through the shared friends connected across different communities (local bridges like nodes 132, 9790 & 8115 illustrated in Figure 4.1). Hence in the shared friends network it is important to target nodes with high degree within maximal cliques for knowing user preferences for learning recommendations & local bridges for information diffusion across communities.

4.1.3 Degree v/s Clustering Coefficient

It is also observed that as the degree of individual's increase, that is as they have more connections (in this case it would be followers), the clustering coefficient decreases. These high degrees will be related to popularity of an individual. These popular individuals would not have a lot of friends connected to each other

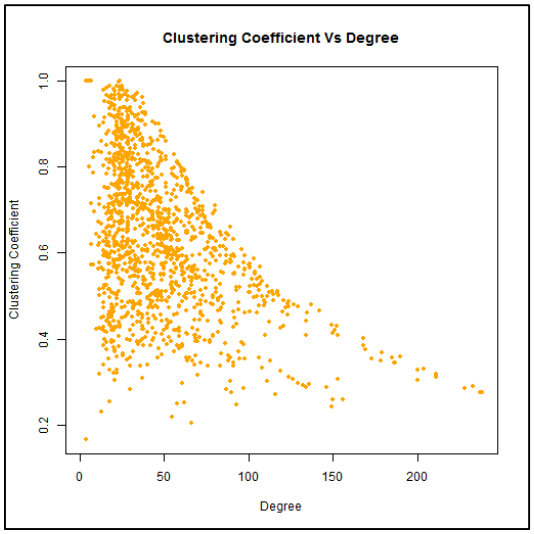

Figure 4.1.3: Degree v/s Clustering Coefficient for Shared Friends Network

4.2 Shared Subscriptions Network:

This network structure implies the relationship among different nodes based on the number of shared subscriptions among them. An edge indicates that both the nodes have a common shared subscription. The weight of the edge is defined by the number of shared subscriptions.

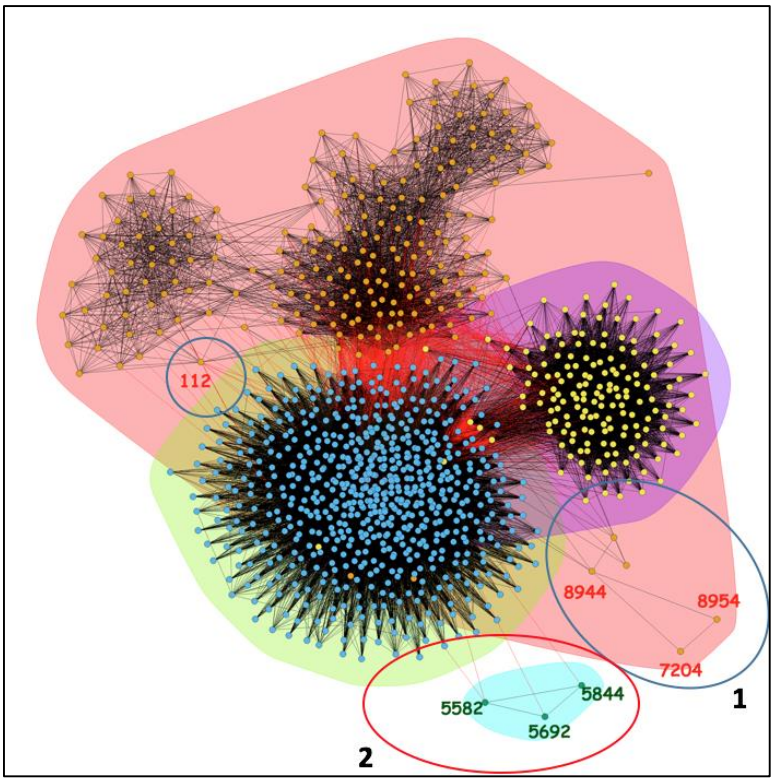

Figure 4.2: Shared Subscriptions Network

4.2.1 A few key observations on the below network are as follows:

The community structure of this network is divided into large communities and small communities. Large communities refer to the people who share generic common interests. Most of the people would follow the top news channels, celebrity or music channels. Hence, we would find huge community sizes comprised of common people. Top degree nodes comprising of these news, celebrity or music channels are: 153,8856,8460,163,6583,8652,6810,8473,5703,8760,5592,8722, 6004,132,8747,162,118,5634,6967,151. Hence, if we want to target a large chunk of viewers for a particular advertisement or a marketing/social campaign, we would recommend to target these above nodes.

The small communities may attribute users who have special interest in niche fields. These are indicated by Nodes circled in 1 & 2 in Figure 4.2. For example, a YouTube channel named "Living Thru Glass" explicitly discuss only about google glass technology. This is a very niche channel with only 34 subscribers. If this technology grows, then it attracts more users and becomes a large community and if it doesn't the community may vanish altogether.

Figure 4.2.1: Youtube channel named "Living Thru Glass"

Even though, the user base is large, only a few nodes are connected across communities. These nodes like node 112 in Figure 4.2 act as local bridges of information in the respective communities. Such nodes would often help in transmitting a diverse information into the community, improving overall knowledge level of individuals across an array of topics. For example, encouraging technology focused individuals towards music and vice versa. however, this kind of intercommunity mix needs to be handled with caution by YouTube as they might risk losing out on potential customers if they push this kind of an agenda. A solution to this would be in terms of video recommendations. In order to diversify knowledge, they might recommend a video out of a user's general search scope, but in the next search pane, they should have some of the user's favorite videos

4.2.2 Cliques v/s Frequency

The below clique frequency graph indicates that this network has the highest maximum sized cliques. This indicates that there are few users that have very high degree of subscriptions while most users have low subscriptions. This is an important fact, as many celebrity, movie, song and technology

videos would have a high level of subscriptions while common users may have few or no subscriptions.

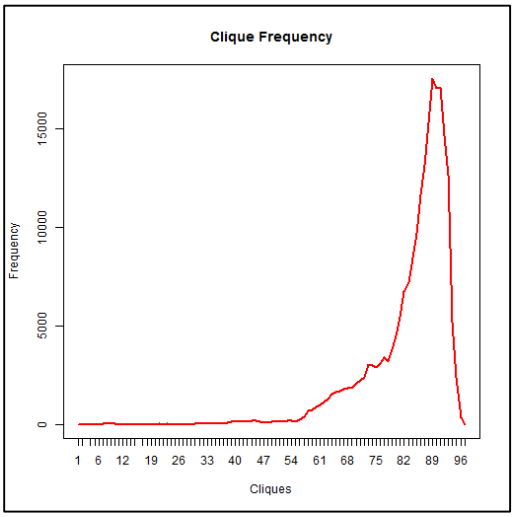

Figure 4.2.2 Cliques v/s Frequency for Shared Subscriptions Network

Hence, promotions or advertisement of any kind should be linked to these high shared viewership videos. Again the videos should be selected carefully based on the similarity of topic concept between the advertisement and the video. For example, a cosmetic advertisement may be linked to a music video and a new phone advertisement to a technology video. Also, if a particular celebrity has endorsed a product, the advertisement for that product may appear with the music/ movie video of that celebrity.

4.2.3 Degree v/s Clustering Coefficient

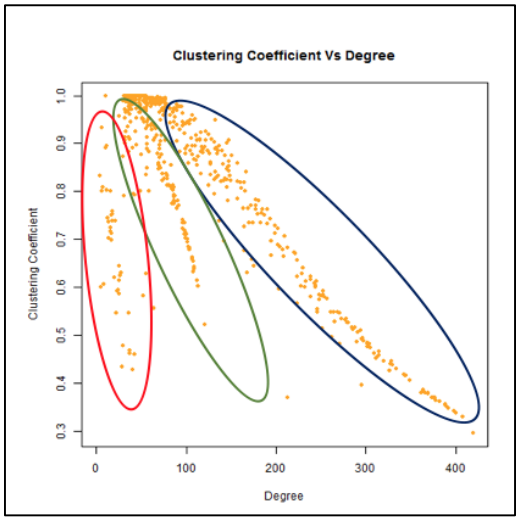

Figure 4.2.3: Degree v/s Clustering coefficient for Shared Subscriptions Network

The graph of clustering coefficient v/s degree has interesting insights for this particular graph. It shows 3 distinct behaviors.

Common Folks: These have few connections implying low degree. This is the group of people who do not upload videos on YouTube but watch videos. This group of people would not ideally be subscribed to, but might subscribe to others. Here we see the clustering coefficient is 0.4 and above showing that this group has a decent level of connectivity with each other's friends.

Budding Stars: These are the group of people that have a higher degree than the commoners and we would like to call them budding stars. They have a decent level of degree. As the degree increases, the clustering coefficient decreases. In the initial stages of growth of an artist, they would get their friends to subscribe to the videos. These friends would in turn be subscribed to each other as well indicating high clustering coefficient. At this stage, it is good to know the most influential nodes which is obtained from the shared friends network (5926, 9002, 7039, 5834, 9898, 10247, 5657, 5854, 8215, 11432, 5618, 10890, 6788, 6739, 6741, 8445, 9450, 10335, 5703, 8652) as they would influence their common friends. As the person become popular, their followers will increase but most of them could come from different regions and may not be mutually connected implying low clustering coefficient.

Celebrity Status: These famous videos have a similar growth path as budding stars but their degree increases at a much faster rate with and clustering coefficient decreases much faster as well.

4.3 Favorite Videos Network:

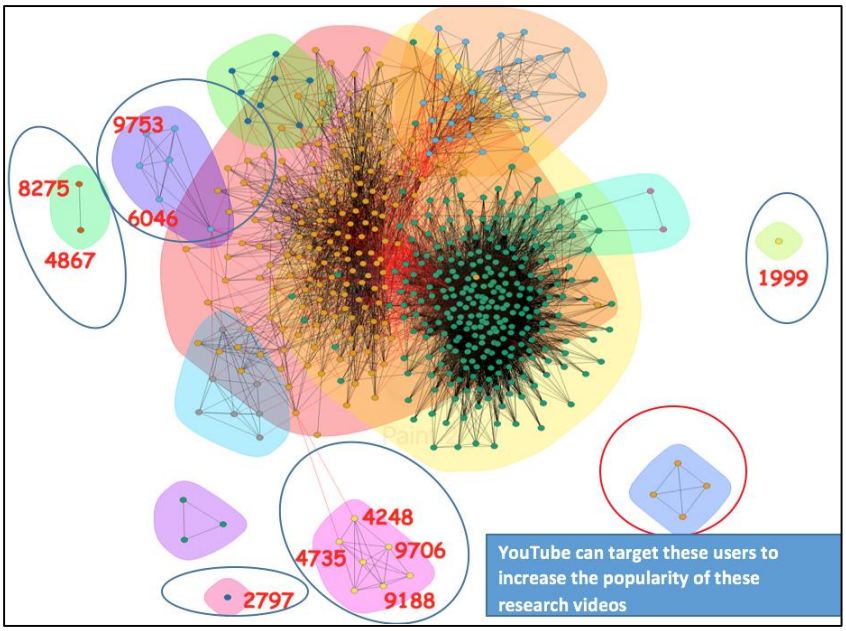

Figure 4.3 Favorite Videos Network

This network tells us about the total number of common videos between two users. This network presents us more qualitative information because it is about the video content consumption of the users. Broad patterns of this network can be generalized to any other video sharing platform.

4.3.1 Key inferences of this network are presented below

a. As expected, based on the favorite videos-users have aligned themselves into large communities and very small communities. Again this is because, most of the popular videos are liked by many people. For example, videos like Gangnam style are a favorite for thousands of users.

b. The small isolated communities like the one having nodes 4735, 4248, 9706, 9188 also give us actionable insights. The striking example for this research group channels or academic institutions. For example, The MIT Open courseware and the Harvard Classes channels in YouTube would have a similar kind of favorite videos. It is very important for YouTube to see whether such nodes are promoted appropriately. Because, though they may have a very few direct connections but these nodes are influential nodes which pull so many users towards YouTube and indirectly increase the revenue for the firm. Such isolated communities (like the ones circled in Figure 4.3) can be communities focusing on key research areas or political interest. Such videos join the overall network through local bridges as their popularity increases.

4.3.2 Degree v/s Clustering Coefficient

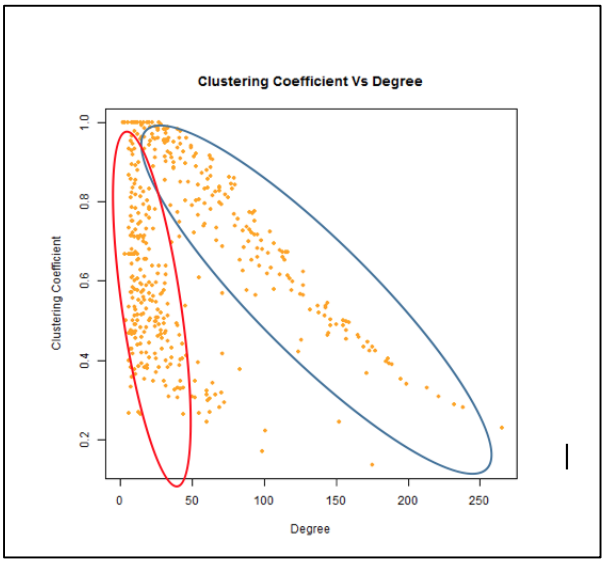

Figure 4.3.2 Degree v/s Clustering Coefficient for Favorite Videos Network

Shared videos show mainly 2 types of patterns. Connecting it to the shared subscriptions, shared videos show only commoners and celebrity status. The budding stars type of individuals seem to have an increasing number of shared subscriptions but the videos are not shared as often.

5. Analysis

5.1 Scale Free Network:

The "Shared Friends" network and the "Shared Videos" network have exhibited a scale free distribution. Both the networks have a long-tail phenomenon when the degree of the nodes is plotted against the frequency of the degree. The Log-Log plot of the same also showed linear trend. From this, we can infer that the networks have more number of higher degree and lower degree nodes than the normal random network.

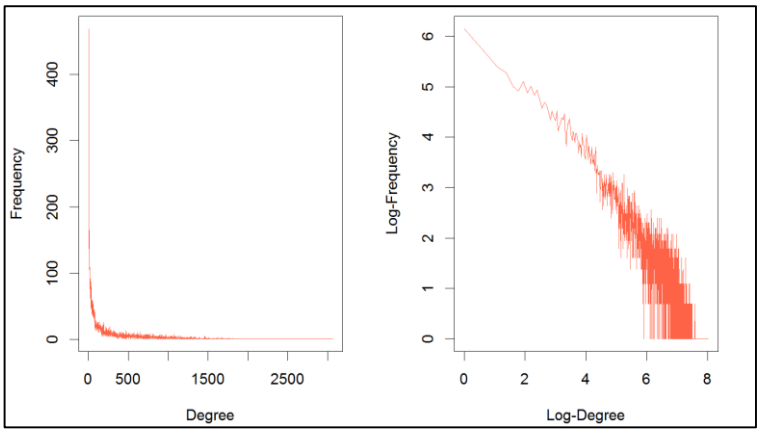

Figure 5.1.1 Scale Free Network for Shared Friends Network

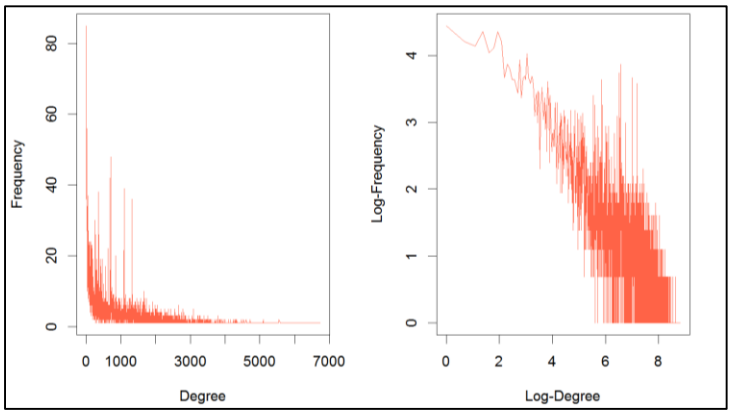

Figure 5.1.2 Scale Free Network for Favorite Videos Network

5.2 Degree Versus Betweenness:

As the degree of the nodes is increasing then the Betweenness of the nodes is also increasing. This effectively translates to the fact that the shortest distance between the nodes is most often connected by the nodes of the higher degree in this network. This inference has a business significance. If YouTube wants to propagate any information quickly throughout the network, then it is a must that the higher degree nodes should be given the information first. In this way, the information exchange would happen with less time and friction.

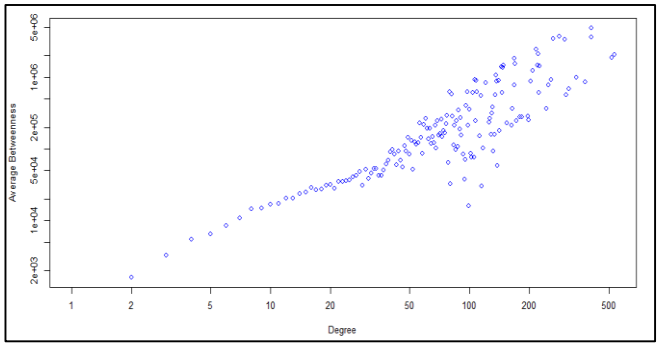

Figure 5.2 Degree v/s Betweenness

5.3 Degree Versus Embeddedness:

Degree is inversely proportional to the Embeddedness of the nodes. This means that the higher degree nodes have the risk of being exposed to similar kind of information always. Hence, to maintain the diversity in the feed content of the higher degree nodes, the YouTube algorithm has to include an additional parameter that can give new topic suggestions to the higher degree nodes. If this feature is not implemented, then it is highly likely that the higher degree nodes would always have a homogeneous information in their feed.

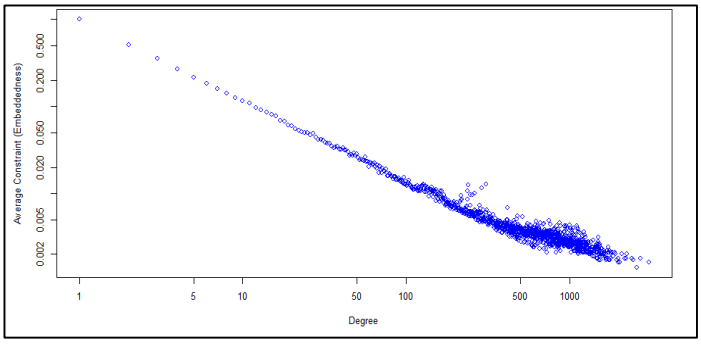

Figure 5.3 Degree v/s Embeddedness

6. Social Contagion

the phenomenon of epidemics is not only constrained to infectious diseases but can also represent anything that can spread across a community such as information, norms, gossips and habits. This notion of spreading information through networks of people is known as Social Contagion. In our context, the information here represents a YouTube channel or the videos that they upload and the community represents the network of shared subscriptions. A study of how information spreads across this network could give insights for a YouTube channel on whom to target and what are the properties of the channel that could attract subscribers.

We have then created an SIS Epidemic Model with information spreading across the network at random because different source of things spread across different source of ties. For instance, smoking and drinking habits might be influenced by their friends rather than your parents. From the simulation we have observed that the information circulates predominantly in the center of the network and it doesn't always reach the peripheral nodes due to the closely knitted community at the center which acts as an information barrier. Further, in order to study the spread of information based on the initial node (patient zero), we have run the simulation taking the most (8652) and the least (11415) influential nodes, based on the highest and the lowest degree, as the starting nodes and the following were the observations:

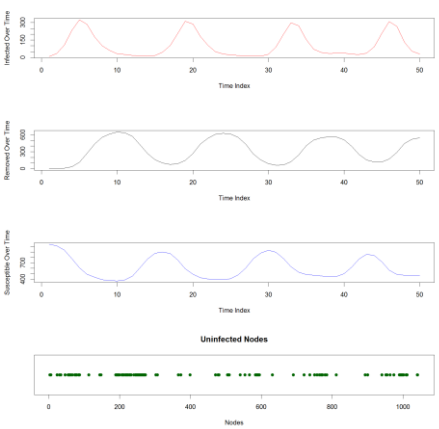

Figure 6.1 Most Influential Node

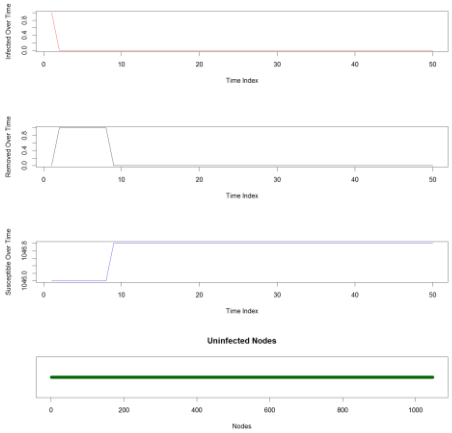

Figure 6.2 Least Influential Node

In both the above graphs, the first plot indicates the total infected nodes. The second and the third plots indicates the number of nodes in the immune state and the number of susceptible nodes respectively. The last plot shows the node, if any, that were never infected throughout the simulation. With all the other parameters including the network parameters being constant, we could observe

that the information spreads continuously throughout the network when the starting node is more influential and it dies out almost immediately when the least influential node was taken as the initial node. Therefore, in order to increase the channel subscriptions, we need to identify influential targets which are capable of spreading the information throughout the network.

After identifying the target nodes, we need to determine the properties of the channel to makes it popular. In fig. 6.3, we could observe that the information died out after a finite amount of time. This could represent the fact that the new channel only uploaded one viral video and eventually its popularity decreased with time. We could also see that it didn't spread across a large number of users. Fig. 6.4 represents the fact that the channel remained popular indefinitely with just one video. This is practically not possible because at some point of time people do lose interest if there is no new content. Fig. 6.5 represents ideal scenario where the channel is ready to upload another new video as soon as its previous video's popularity decreases and thus keeping the users interested throughout the simulation. We could also observe that in such a scenario, the information was able to spread to almost the entire network except for a few users who might be the dormant users of the community.

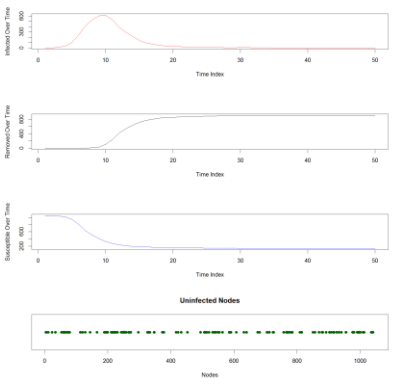

Figure 6.3

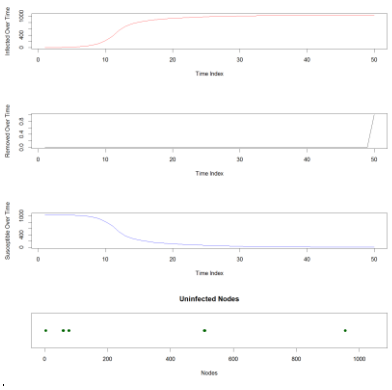

Figure 6.4

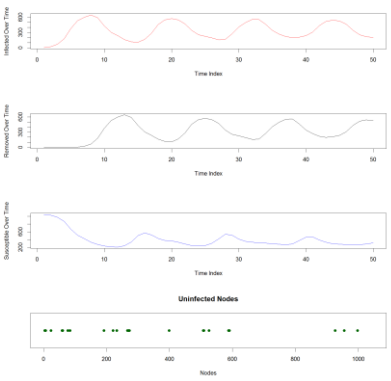

Figure 6.5

7. Conclusions

The final recommendations from our project are as below

1. The Network has relatively higher number of nodes with high degree or low degree than the normal random network. These high degree nodes exercise a higher influence in spreading information. For example, a video shared by a person with 1000 friends is more likely to be seen by many people. Hence, based on the subscription list of these nodes, YouTube can get a fair idea on the likes and dislikes of this person. Based on this appropriate ad can be placed on the nodes.

YouTube can target specific users as shown in the paper based on the category they are separated into (degree vs clustering coefficient). YouTube is not a network where links between friends of friends make as much an impact as propagation of information via high degree nodes like celebrity, brands and so on.

2. **Maintaining heterogeneity of information on the Network:** High degree nodes run the risk of being surrounded with homogenous information (as their embeddedness is low). Such users, if they see similar content always would slowly lose interest on the website. Hence, it is critical to have special parameters that broadcast complementary sets of information to such people.
3. For a marketing campaign video to be appreciated or viewed by many people, ensure that the high degree nodes have access to the campaign. This would facilitate a quicker information propagation in the network
4. **Recommendations:** Influential nodes within a region or genre can be used to learn user patterns of likes/ subscriptions and hence recommend future videos based on these patterns and the friendships between users
5. Information Transfer across genres and regions: Local bridges are important for information transfer between communities especially between regions (shared friends) and to maintain heterogeneity of interests in the network (shared subscriptions and shared videos).
6. **New information growth:** In some cases, subscriptions/videos are shared by a small group of close knit users. Based on the context and content, Youtube may want to plant an edge in order to make a less know topic more widely spread. This might be one of the strategies YouTube adopts to diffuse new information into the system and keep the network relevant.

References:

1. R. Zafarani and H. Liu, (2009). Social Computing Data Repository at ASU [http://socialcomputing.asu.edu]. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering.

2. http://www-personal.umich.edu/~ladamic/courses/networks/si508f07/projects/youtube.pdf

3.