Multi-dimensional analysis of YouTube user information

Shuchit Gandhi
Information and Communication Technology
minors in Computational Sciences
DA-IICT
Email: 201301441@daiict.ac.in

Yashodhan Mohan Bhatnagar
Information and Communication Technology
minors in Computational Sciences
DA-IICT
Email: 201301225@daiict.ac.in

***Abstract*—Popular sites like YouTube allow users to comment shared content (bookmark, photos, videos), and users can tag their own favorite content. Users can also connect to each other, and subscribe to or become a fan or a follower of others. YouTube is a video sharing site where various interactions occur between users. In the data set that we have selected, for each user, his/her contacts, subscriptions and favorite videos were crawled resulting in 15,088 active user profiles. Based on the crawled information, 5 different interactions were constructed between the 15,088 users which were - the contact network, the number of shared friends between two users, the number of shared subscriptions between two users, the number of shared subscribers between two users, the number of shared favorite videos. From this data set, we'll be trying to answer questions like community formation in a video sharing website like YouTube, influence of liked videos and subscriptions on contacts, and the factors that determine the influence of a user in the network.**

I. Introduction

YouTube is a video streaming website allowing users to upload, share, like, dislike videos, create channels and subscribe to other accounts. It takes the concept of a social network and mirrors it into the video viewing habits of the users. Thus, common intuition leads to believe that people with similar viewing habits are likely to bump into each other more often in the comments section. Also, creators of channels with similar content are expected to have more common subscribers due to the suggestion feature included in the website which suggests various channels of similar interests to the users.

The YouTube community is significantly distinct from a normal online social network on the fundamental fact that while a social network almost usually connects two people who know each other in the physical world, the YouTube community captures abstract features of people's interest via likes and subscriptions in the form of networks. It is also able to capture phenomena such as viral videos via time series networks. Since it also creates a semblance of forum under each video in the form of comments section, it allows users from vast geographical spreads to connect and exchange interests and views by recommending videos.

In this study, we shall be considering five types of information collected from a sample Youtube network. The information types are:

1) Contact Network: This unweighted network contains a binary edge between two users if they are friends on YouTube.
2) Common Friends Network: This weighted, undirected network contains an edge between two users if they have mutual friends on YouTube and the edge weight is proportionate to the number of friends they have in common.
3) Common Subscriptions Network: This weighted, undirected network contains the number of common channels or users subscribed by any two users with the edge weight equal to the number of common channels. It corresponds to a persons similarity in interest with another person.
4) Common Subscribers Network: This weighted, undirected network focuses on content creators and contains an edge between any two creators if they have any common subscribers with the edge weight equal to the number of common subscribers.
5) Common Favorite Videos Network: This again correlates to the shared interest between any two users by counting the number of liked videos they have in common.

II. Exploratory analysis

We begin the analysis of the data by performing an exploratory study of the five kinds of networks by simply ignoring the weights of the edges for the time being. This creates five binary networks where the interactions, instead of being weighted, are converted to whether or not two users interact. For instance, for the common friends network, two users are connected by an edge by the sole criteria of whether or not they share atleast one mutual friend or not.

A. Degree distribution

The primary point of analysis for any network is distribution of the degrees of the nodes in the network. Having said this, we can find the degree distribution of the five unweighted networks in Figure 1. As can be seen in the figures, existence of fat-tailed distributions can be visualized with quite certainty. This signals towards existence of "hubs" (not in a technical sense of its definition) or the presence of highly connected nodes since there are substantial amount of nodes with degrees

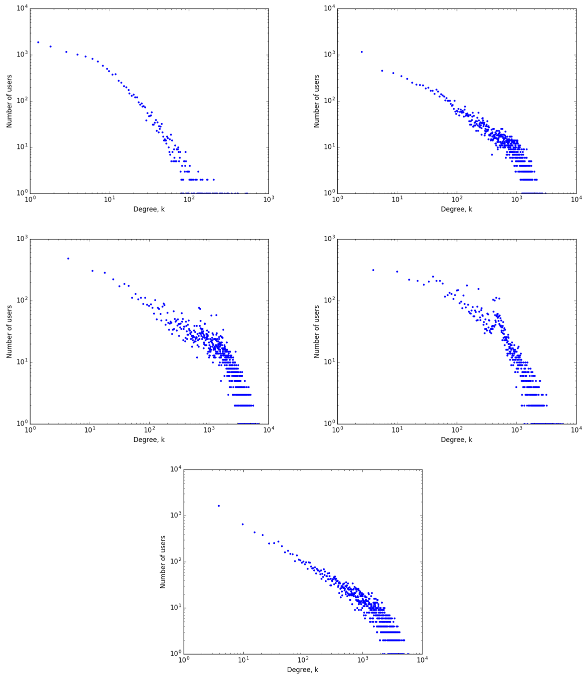

Fig. 1. Degree distribution for each of the 5 networks without considering their weights, from left-to-right, top-to-bottom: 1) Contact network, 2) Common friends network, 3) Common subscriptions network 4) Common subscribers network 5) Common favorite videos network

much higher than the mean degree of the network which can be seen in Table I. For the four networks, the implication is as follows:

1) Contact: There exist users who have mutual friends with a lot of other users. This allows for ”people who know people” phenomena where there exists users who can be used as stepping stones to shorten their reach to another user to contact someone. This affects in shortening the characteristic path length of the network.
2) Common friends: The existence of a hub in this network signals towards users who are strongly connected to many users. Since the user has mutual friends with a lot of other users, this strengthens a user's claim to belong to another user's friend circle, i.e., the neighbours of that user are closely bound.
3) Common subscriptions and common favorite videos: The hubs in these networks signal to similar implications as in common friends with the sole difference that the commonality with the other user comes in the form of interests and likes rather than other users. This also indicates the presence of users who have a diverse range of interests allowing him/her to have common interests with a large number of user. This also, in real domain, verifies the presence of spam accounts which like, subscribe and make connections for the purpose of maximising reach for third-party websites.

Network	Average degree
Contact	11.18
Common friends	293.12
Common subscriptions	947.59
Common subscribers	428.39
Common Favorite Videos	577.14

TABLE I
AVERAGE DEGREES

B. Clustering coefficient distribution

It is quite common to measure the average clustering coefficient of a network to measure the degree to which a network is close to being a clique. But while the average clustering coefficient provides a lot of information about the network, a lot of information about the local clustering is lost such as its distribution and the existence of isolated nodes as well as clusters within a cluster or a hierarchical structure. We now consider the distribution of the local clustering coefficient of each vertex in all the networks (as in Figure 2) to see if any significant patterns emerge. Except for the contact network's anomalous behaviour, rest of the plots show similar behaviour and hence we can take them into two cases:

1) Contact network: The clustering coefficient of a vertex signals the probability of finding two of its neighbours themselves to be connected. In other words, it measures the degree to which the induced network of its neighbours is a clique. In this case, we see a distribution similar to a power law similar distribution but since the

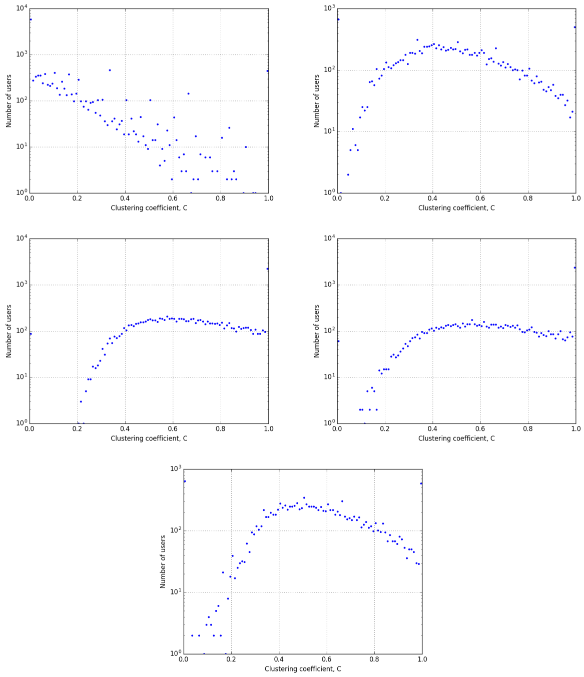

Fig. 2. Clustering coefficient distribution for each of the 5 networks without considering their weights, from left-to-right, top-to-bottom: 1) Contact network, 2) Common friends network, 3) Common subscriptions network 4) Common subscribers network 5) Common favorite videos network

range of values is small, it may be prudent to make any assumption on that basis. What we can make an inference of is the fact that it has a large number of vertices with zero or negligible clustering coefficient. The number of vertices decreases as the clustering coefficient increases indicating low proportion of clusters or cliques inside the network.

2) Common friends, subscriptions, subscribers and favorite videos: In these plots, we see a behaviour which is most prominent in the plot for common subscriptions. We notice that the number of vertices increases exponentially and then stagnates for a range before decreasing subtly as the clustering coefficient increases. We also notice the initial increase only after a threshold C value. One could possible infer that the clusters formed inside the network prefer to be neither isolated nor complete but somewhere in between near the average clustering coefficient (as seen in Table II). This was, ofcourse, expected for in log scale, these plots are not much unlike an inverted parabola which on an ordinary scale translate to normal distributions albeit slightly right skewed.

One interesting thing that one can notice here is that the amount of stars (C=0) and n-pyramids (C=1) are unnaturaly high than any other C value. Of these, the n-pyramids indicate, admittingly weakly, an existence of closely bound communities. The n-pyramids are also skewed by triangles and complete K_4 graphs. The unnaturally high amount of stars on the other hand probably indicate spam accounts. $C = 0$ can also be attributed to isolated nodes.

Network	Average degree
Contact	0.07952
Common friends	0.4159
Common subscriptions	0.4742
Common subscribers	0.4205
Common Favorite Videos	0.3990

TABLE II
AVERAGE CLUSTERING COEFFICIENT

C. Clustering coefficient versus degree

We now explore the comparison between the clustering coefficient and the degree of a vertex to infer, if anything, about a person's degree's effect on its neighbours' connectivity, or vice versa. In terms of the domain, we will look for patterns as to whether if a user has a high number of people with which it has common favorite videos or subscription, then does those neigbours with common interests have any common interactions. We compare the two variables using a scatter plot of the clustering coefficient and the degree.

The first thing visible is the inverse power law with $\gamma = -1$. We can approximate the γ to be -1 by the following logic. Clustering coefficient for a vertex v, C_v, is defined as:

$$C_v = \frac{|\{e_{ij} : e_{vj}, e_{vi} \epsilon E\}|}{k_v(k_v - 1)} \quad (1)$$

We can define the numerator as a new quantity, N_v, such that:

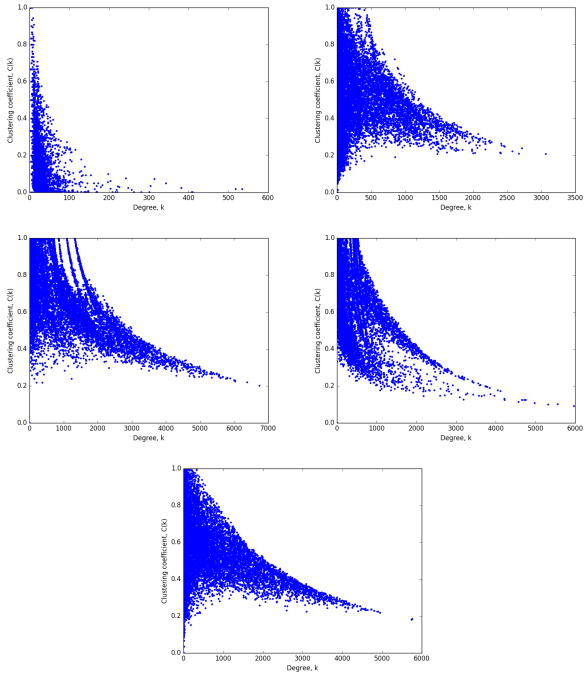

Fig. 3. Clustering coefficient versus degree distribution for each of the 5 networks without considering their weights, from left-to-right, top-to-bottom: 1) Contact network, 2) Common friends network, 3) Common subscriptions network 4) Common subscribers network 5) Common favorite videos network

$$N_v = C_v k_v(k_v - 1) \quad (2)$$

We tried to plot N_v versus the degree of vertex v, and found plots as in Figure 4. Clearly, N_v grows linearly with the degree of a vertex in this network. Now, looking back at equation 1, we see that the numerator grows linearly with degree, k, whereas the denominator grows at a quadratic rate. Hence the clustering coefficient forms a curve that has a ceiling or general trend which follows the power law with $\gamma = -1$. In the domain, this mirrors to the fact that as the number of users with which you have anything in common grow, the probability that any two of those users themselves have anything in common grows low. This can be understood due to two reasons:

1) At lower degrees, activities, such as sharing or liking videos, quickly spread to each of the neighbour allowing for higher clustering coefficients. This can be seen as the effect of fast and efficient spread of information in a smaller network. We also see a larger range of clustering coefficients at lower degrees which is simply due to the fact that there are a large number of nodes with low degrees, each having a different clustering coefficient efficiently spreading out over the entire range from 0 to 1. The higher degrees have lower clustering coefficient due to the fact that as the number of users with which a user has something in common increases, the probability that its neighbours belong to the same interest circle decreases.

2) The lower bound seen in the figure can be attributed to the fact that even if a user tries to connect to people with completely different tastes such that a star network may be formed, there is a certain probability that the neighbours themselves have something in common. This random connection between neighbours due to a common interest is what attributes to the lower bound for the figures. This lower bound increases as the degree increases initially, but is unable to dominate the $\gamma = -1$ effect, and thus eventually decreases.

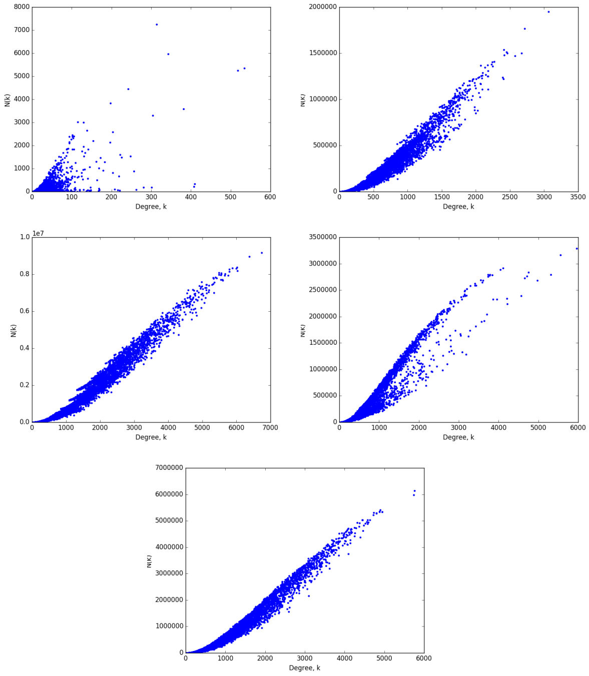

Fig. 4. N_v versus degree distribution for each of the 5 networks without considering their weights, from left-to-right, top-to-bottom: 1) Contact network, 2) Common friends network, 3) Common subscriptions network 4) Common subscribers network 5) Common favorite videos network

III. Multi-dimensional Analysis Methods

Having, qualitatively argued about the implications of the individual network in a collective domain, we now look at one of the suggested methods to mathematically combine various dimensions of a social network which cannot be defined satisfactorily in a single dimension. Youtube is a perfect example for multi-interaction domain where people connect and display their choices and interests via various forms of expression such as subscription, sharing, liking and commenting. We have been considering each form of interaction separately up till now and supporting the observations with arguments from the real world. We can now mathematically combine the network parameters to create new observations which take into account two or more of the networks to bring out patterns arising out of a more complete look of the system since more facets are being considered. We will also be looking at noise in networks which affect observations when considering a single network alone but can be avoided or their effect minimized by considering two or more networks together. The feature we will be attempting to extract will be the group structure in a network. Group or cluster extraction in a multi-dimensional network becomes difficult when some people interact with other members within the same group in one form of activity consistently, but are inactive in another. Modularity is a feature which allows one to check how dense the interaction is between groups within themselves but lesser interaction with members of another group forming a ghetto-like structure.

The modularity of a network is defined as :

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta(s_i, s_j) \tag{3}$$

We look at two methods to calculate the modularity of the network taking into account all the five dimension of the dataset:

A. *Average Modularity Maximization*[4]

The simplest strategy to handle multiple dimensions is to create a network which averages over all interactions to create a single network. One can then calculate the modularity for this average network. The averaging techniques can be various - from weighted averaging to completely unbiased averaging of the weights of the edges. Averaging multiple networks implies averaging the weights of the edges as well as the weights of the vertices.

B. *Total Modularity Maximization*[4]

This strategy calculates the modularity for each dimension and tries to maximize the average of these modularities. In the case of the YouTube network, one calculates the modularity for all the five networks and then averages these modularities. The method then tries to maximize this modularity.

IV. Conclusion

When a multi-dimensional network with various interactions is available, it might be insufficient to extract information and features from only one type of interaction. This becomes especially important in social networks with other major features like video sharing where having a contact is not the primary action of a contact. Instead, integrating combining various formes of information can compensate for incomplete information in each dimension as well as reduce the noise from elements like spam accounts. Intuitively, with a multi-dimensional network, one can use richer information to infer more accurate latent clusters present in the underlying community.

References

[1] M. E. J. Newman *Analysis of weighted networks*. Disordered Systems and Neural Networks, Phys. Rev. E 70, 056131, 2004.

[2] A. Ioannis, T. Eleni *Statistical Analysis of Weighted Networks* Physics and Society, 2007.

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

[4] L. Tang, X. Wang, H. Liu *Uncovering Groups via Heterogeneous Interaction Analysis* IEEE International Conference on Data Mining (ICDM09), 2009.