Social Network Final Project Research Report

Agnes Liu, Lanxin Mao, Robin Liu

1. Introduction

Founded in 2005, YouTube is one of the most popular video sharing website on the Web today. Millions of videos can be watched, uploaded and shared through YouTube. Most importantly, YouTube serves as a popular social network and became successful by creating a unique user-to-user social experience. In this research report, we are going to leverage data from the YouTube social network to answer four main research questions:

1. Examine the centrality measures and the correlations between them
2. Examine whether the contact network degree follows Power Law
3. Examine the contact network structure
4. Examine the correlations among five attributes

2. Data Information

2.1 Data Background

The original data set is retrieved from the ASU Social Computing Data Repository. It was crawled on Dec, 2008 from YouTube. (http://www.youtube.com/). *"In particular, we crawled 30, 522 user profiles. To avoid sample selection bias, we choose authors of 100 recently uploaded videos as seed set. This crawling reaches in total 848, 003 users and 1,299,642 videos. However, not all users sharing all kinds of information. After removing those users, we have 15, 088 active user profiles. Based on the crawled information, we construct 5 different interactions between the 15, 088 users."* (R. Zafarani and H. Liu, 2009). The original file contains six files,

including a nodes file and five edge file with each file represents one type of interaction. The original data set contains 15088 nodes, 5574249 edges, 5 attributes and no missing value.

2.2 Data Preparation

To scope for the project, we merged the five edge lists to one file and use it as our network data set. Since not all users sharing all kinds of information, many edges have been removed. Therefore, we now have a relatively smaller data set. The final data set contains: 13723 nodes, 76765 edges, 5 attributes and no missing value.

The five attributes are:

1. The contact network between all users in the dataset
2. The number of shared friends between two users
3. The number of shared subscription between two users
4. The number of shared subscribers between two users
5. The number of shared favorite videos between two users

3. Analysis Method

To answer our research questions, we used the following methods:

- To examine user(vertex) properties, we calculated centrality measurements, and analyzed their distribution and correlation.
- To find out the network structure, we calculated cluster of network and coreness.
- To measure attributes' influence, we also used attribute distribution and correlation. And we apply ERGM model to the subnetwork (central core of the network).

4. Analysis Results

4.1 Centrality Measures

Question 4.1: Is there any correlation among centrality measurements?

Table 1: Denormalized and normalized centrality summary

De-normalized Centrality				
	degree	**betweenness**	**closeness**	**eigenvector**
Min	1	0	5.311e-09	0
Median	6	4052	1.512e-06	0.0003
Max	534	4848487	1.552e-06	1
Normalized Centrality				
	degree	**betweenness**	**closeness**	**eigenvector**
Min	7.288e-05	0	7.288e-05	0
Median	4.372e-04	4.305e-05	2.075e-02	0.0003
Max	3.892e-02	5.150e-02	2.129e-02	1

These four centrality measurements show how central and important the users are:

- Users with high degree connect with more people, so their friends networks are larger;
- Users with high betweenness have significant influence because they control over information passing between others;
- Users with high closeness are closer to all other users;
- Users with high eigenvectors connect with central users.

Table 2: Centrality correlation

	degree	**betweenness**	**closeness**	**eigenvector**
degree	1	0.722	0.104	0.621
betweenness	0.722	1	0.055	0.187
closeness	0.104	0.055	1	0.068
eigenvector	0.621	0.187	0.068	1

Findings from centrality correlations:

- Correlation between degree and betweenness (0.722) is the highest. Users who have more friends usually act as bridges in the network. YouTube can reduce betweenness by recommending one to the other if they have common friends, allowing the whole network becomes denser (now the density is 0.08%).
- Correlation between degree and eigenvector is 0.621. Users with more friends tend to connect with more central users.
- Correlations between closeness and other three centrality are very low. Users with high betweenness don't tend to situated near all other users, so they may act as gatekeepers among different clusters. Users have more friends, or connect with more central users, or as bridges among users are not tend to situated closer to all other users.

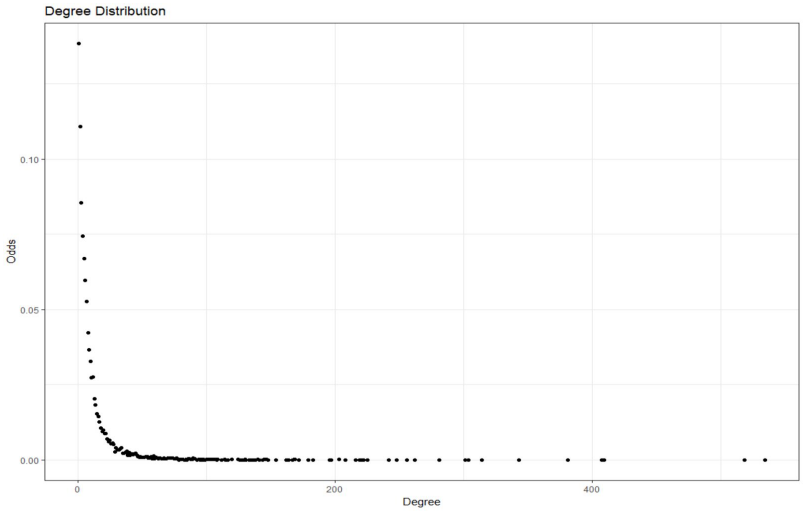
4.2 Power Law

Question 4.2: Does the contact network follows Power Law?

In a real-world scenario, most of the people uses YouTube as entertainment instead of using it as a tool to interact with others regularly; only a small amount of people, namely youtubers/bloggers, would try to gain more popularity through Youtube. Our dataset reflects this phenomenon and follows the Power Law: most of the people would have small degree centralities, and only a handful of people would have large degree centralities. To express this rule more rigorously using mathematical terms: $y = ax^k$

Where y is the fraction of nodes in the network having x connections, k is a parameter typically ranging from -3 to -2, and a is just some constant.

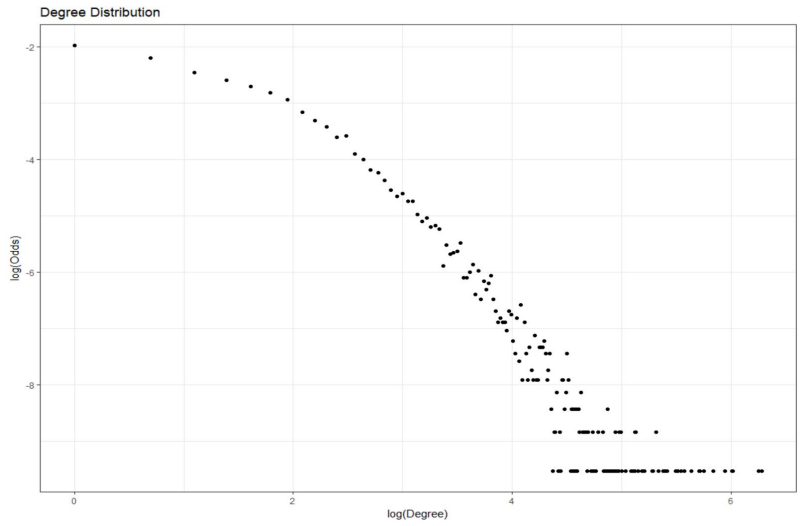
The graph odds vs. degree shows that the odd of a vertex has some inverse relationships with the number of degrees. Next, we can figure out the exact relationship between the two variables using data transformation.



If we take the log of both sides of the equation y = a*(x^k):

$$\text{Log}\,(y) = \text{Log}\,(a\,x^{k})$$
$$\text{Log}\,(y) = \log\,(a) + k\,\log\,(x)$$

We would be able to explain the log odds of a vertex having x degrees with a linear combination of the two terms on the right-hand side of the last equation. Going through the same procedure with our own network, we plotted the log odds against log(degree) and we clearly identify a linear relationship between them, which indicates that the odds and number of degrees follow an exponential decay relationship, and our network follows the power rule.

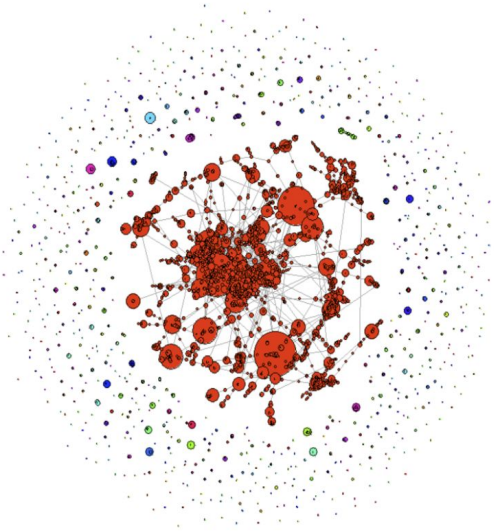


4.3 Network structure

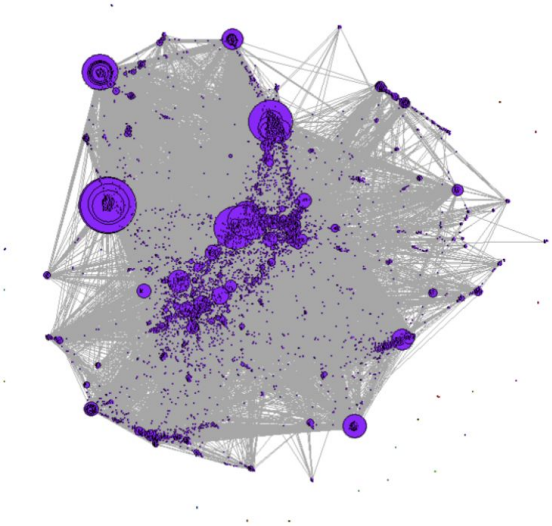
Question 4.3: What is the structure of contact network, core-periphery or clustered/component?

There isn't strong evidence that our network belongs to either option. However, there are some key findings regarding the overall structure.

I. Network Visualization



Graph 1: YouTube Network Sample Graph (Sample=5000)



Graph 2: YouTube Network Graph

Above is the sample and whole graph of network, vertex size is based on degree and components are colored differently.

II. Cluster/Components Analysis

In the entire network, there are 21 clusters in total, including a giant one that contains over 99% nodes. In this giant cluster, average shortest path among users is 4.3, representing a "small

world". Intuitively, the presence of this huge component in our network might lead us to assume that the network has some center nodes, which connect to the most of the nodes.

Table 3: Number and size of cluster in the network

Cluster size	Cluster count
2	4
3	16
13679	1

The result shows that even though there is a giant component in the cluster, the nodes are connected to each other through multiple coordinators instead of a single gatekeeper. These nodes are able to reach to each other by slowly going through others routes but not directly through some well-connected node.

We can take further look at the betweenness of our network. The closer the normalized betweenness to 1, the more star-shaped this network becomes, and every node is connected to a center node. However, the normalized betweenness is only 0.05. This low betweenness means greater accessibility between nodes, which corresponding to our hypothesis.

III. Coreness Evaluation

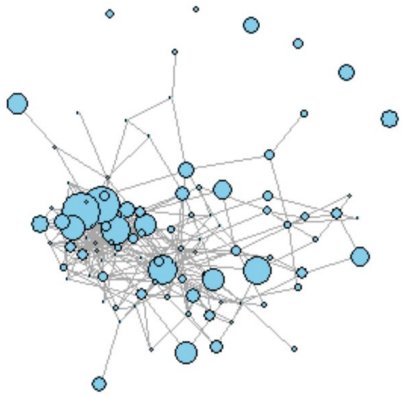
We can also take a look at the distribution of the coreness within our network:

Table 4: Quantile of coreness value

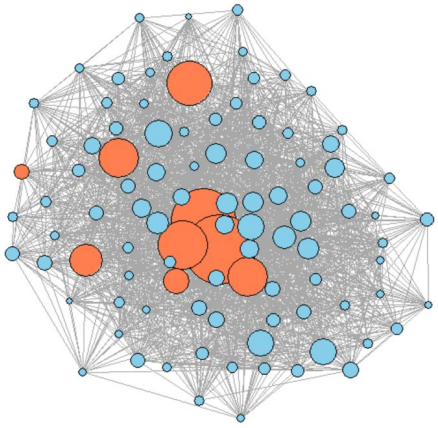
Quantile	0%	25%	50%	75%	100%
Coreness	1	2	5	8	25

Half of nodes could at least form a 5-core subgraph, which verifies our hypothesis that people can connect with others who are not necessarily well connected, managing to reach to everyone and this kind of k-core subgraphs become parts of this giant component.

The maximum coreness is 25; by definition, in 25-core subgraph, there are at least 26 people could be considered in a maximum subgraph where everyone is connected to at least 25 others. In fact, we found that there are 98 nodes with coreness of 25.



Graph 3: Network Graph (Top 100 closeness)



Graph 4: Network Graph (Top 98 coreness)

The above graph illustrates vertexes with the highest coreness in the network, sizing by degree centrality within the entire network. The red vertexes are the ones with top 100 closeness within the network. The large and red vertexes in the graph are both connected to many other nodes and close to others. Therefore, they can be considered as hubs within the entire network and are more central than other vertexes are.

We can also gain some insights by looking at the degree distribution. There are only 1418 nodes (roughly 1/10 of all nodes) that have degree of 25 or more, and we might consider them to be the 'well connected ' nodes within the network. In conclusion, even though there isn't a center node,

there exists a ‘inner circle’ where many well-connected people are connected to each other within the group.

Based on the close connections between well-connected people and low betweenness between nodes, we can tell that YouTube is doing a good job utilizing the strong connections between well-connected people as well as linking other users to the network.

4.4 Attributes

Question 4.4.1: Whether these four attributes vary together? Measure attributes' correlation.

Table 5: Correlation among attributes

	Shared friends	Favorite video	Subscribers	Subscriptions
Shared friends	1	0.146	0.135	0.212
Favorite video	0.146	1	0.540	0.214
Subscribers	0.135	0.540	1	0.139
Subscriptions	0.212	0.214	0.139	1

In general, correlations among these four attributes are low. Among all correlations, the correlation between subscription and favorite videos (0.54) is the highest, showing that they don't vary together.

Question 4.4.2: Whether two users have shared friends/favorite videos/subscribers/subscriptions if they are connected?

We calculated the number of edges in the network which have attributes not equal to zero, indicating that the shared attribute exists. The following are our findings:

- 26.9% connections have shared subscribers.

- 46.8% connections have shared subscriptions.
- 46.5% connections have shared favorite videos.
- 16.1% connections have over 5 shared friends.

There are 50% probabilities that the user and his/her connected friend have shared subscriptions and favorite videos, while only 27% and 16% probabilities that they have shared subscribers and friends respectively.

5. Conclusion

We can conclude from this report that, in general, people have enough connections so that they can reach to others and form a giant component in the network. There also exist a 'inner circle' formed by 'well-connected' people(who might be youtubers/bloggers), who could collaborate and influence others so that people connected to different well-connected people can form connections with each other. Therefore, people who share similar favorites or subscribers don't necessarily have the same subscriptions since everyone is connected to such a group of people. In addition, We found that the distribution of shared friends, subscriptions, subscribers and favorite videos all follow power law. This tells us that most of users are making friends and subscribing to a small group of users. Also, a huge amount of videos are uploading by this small group of users.

References

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.