YouTube Network Analysis

Final Project Report

Jiaqi SHI
MSc in Data Science and Business Analytics
b00713537@essec.edu

Kimya DHADE
MSc in Data Science and Business Analytics
b00718577@essec.edu

Xiaofeng XU
MSc in Data Science and Business Analytics
b00717754@essec.edu

ABSTRACT

As one of the most influential online websites, YouTube has more than 1,300,000,000 hours' videos and more than 5 million videos are watched by its users every day. In this project, we are going to analyze the YouTube video network based on the data we crawled online. This dataset is randomly generated on a random day, so it is representative of the whole YouTube dataset. We will visualize the connection among tremendous videos in the dataset by network graph, as well as calculating several important attributes of the network. After that, we are going to build a simple but effective channel recommendation system. We started from the users' watching history, and recommend different channels users may be interested in. The recommendation is content-based, and it is based on the YouTube's built-in video recommendation system. Therefore, our goal is to apply our knowledge in a real business case rather than implementing a fantastic YouTube recommendation system, which is out of our ability. Besides, as a commercial website, applying technologies into the business case to generate profit is the most significant element for YouTube. Therefore, in our project, we are also going to talk about the monetization ecosystem, which helps advertisers choose what type of ads must be advertised. We are going to analyze how the channel recommendation system can be implemented with the advertisement, and how the system can improve the advertisement effect.

CCS CONCEPTS

• **Computing methodologies** → **Network Science** → Network Analysis

KEYWORDS

Network analysis, YouTube Channel Recommender systems, Monetization Ecosystem

1 INTRODUCTION AND PROBLEM DEFINITION

Network analysis is prominent in many practical areas and widely used in e-commerce, social network, and advertising recommendation. It is an effective tool to do the analysis for both technical and business sides. As the largest video-sharing website, YouTube has a tremendous number of videos, and the relations among different videos are complex but informative. Our team believes that analyzing the YouTube video network is one of the best implements of network science in the real world. We are going to do basic network science analysis and implement our other technical skills regarding the real commercial world applications.

The goal of our project is to analyze the YouTube video dataset, build a simple but representative channel recommendation system, and analyze the importance of monetization ecosystem in the real business world.

This report is organized as follows: In the second part of the project, we talk about related YouTube network analysis projects. In the third part, the basic description of the dataset is provided. Then, in section four, we conduct the basic network analysis. In section five, the implement of a channel recommendation system is described in detailed. After that, we will introduce the monetization ecosystem in YouTube, including its importance and application. Last but not least, we talk about the further research that can be done with this subject.

2 RELATED WORK

Since the rise of User generated content on internet and its exponential increase in popularity (Reiffers, Alexandre, et al 2015). More and more advertisers are now shifting away from traditional media such as TV radio and newspapers and turning to online social media networks (Michael Zink et.al. 2008). Google Display Network is a massive network of google services users and this power of user data is tapped through Google AdSense and AdWords. In this study we target how the advertising works on YouTube and how uploader segmentation according to the advertisers needs and requirements can help us replicate YouTube's internal channel recommendation systems. In the first part we apply the methods to analyze the structure of YouTube's network. We detect the strongly and weakly connected components (John C. Paolillo 2008). Our graph-based findings were similar to the ones proposed by the (Torres et.al 2015) in concept despite the difference in the age of the dataset. This is followed by segmentation of channels we work with detecting Influencers in the segments that will be the key in audience reach out maximization. Influence Maximization algorithms help in detecting the diffusion of information in

social networks, news viral marketing and outbreak of diseases (Maria-Evgenia Rossi et.al. 2017). In the application of our project we explore the use of influencer detection in the targeted ad segmentation to maximize.

3 YouTube VIDEOS' STATISTICS

2.1 Dataset Introduction

In this project we use dataset that was extracted using YouTube API . Each record is presented by the video ID, with nine features of this video. The definition of the nine features is explained in the following table:

Table 1: Information of the dataset

Video ID	An 11-digit string, which is unique
uploader	A string of the video uploader's username
age	an integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment)
category	a string of the video category chosen by the uploader
length	an integer number of the video length
views	an integer number of the views
rate	a float number of the video rate
ratings	an integer number of the rating
comments	an integer number of the comments
related ID	up to 20 strings of the related video IDs

Regarding the dataset size, limited by the computation capability of our laptops, we only used the data at 1007-03-02, which was collected using 3-depth BFS (Breath First Search) approach and contained 10,324 distinct videos.

3.2 Video Category

In this section, we segment videos according to their own attributes like categories, and then to see the relationship between number of views, and number of comments with categories

. The purpose if this analysis is to find out the trend as well as the users preference in the YouTube community

For the video category, the user can select from one of 12 categories when uploading the video. We plot the histogram of each category.

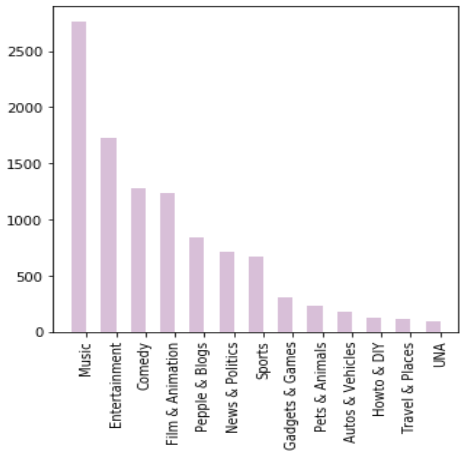

Fig 1. Distribution of numbers

In Fig.1 we can see that the distribution is highly skewed: the most popular category is "Music", at about 24%; the second is "Entertainment", at about 17%; and the third is "Comedy", at about 15%. It shows that the top three categories counted more than 55% of the total videos, which proves that YouTube is more a life-entertainment video platform rather than a very professional community.

Next, we analyze the number of views and comments. The number of views is an important indicator to describe the video attribute, which shows the popularity as well as the model if how users watching videos. Since the view and comments number are changing with time passes, but out data is collected at a certain period. Therefore, we are performing under the premise that the numbers of views and comments are stationary at a short period.

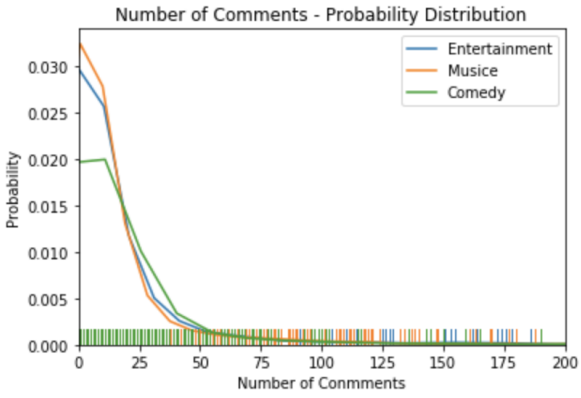

Fig 2. Distribution of Comments by Category

From Fig.2, the Entertainment and Music have similar comments distribution, which has very high probability around 0 to 25 comments. Meanwhile, the Comedy category videos have many comments than others with lower probability distribution at 0 to 25 comments per videos.

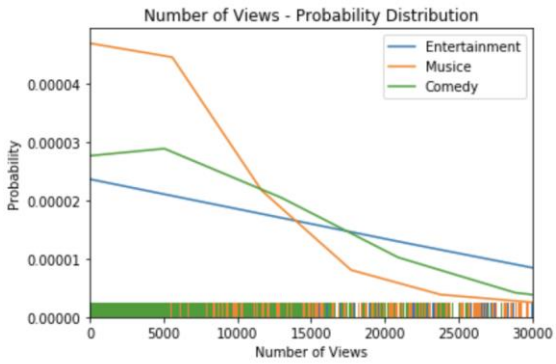

Fig 3. Distribution of views by category

The Views distribution graph shows that Music videos have the lowest number of views in these top-three categories. The Entertainment category videos are the most popular videos with low probability distribution in the low number of views range.

Comparing the previous two graphs, we can say that Number of Views graph show a strong linear relationship with the Probability than the number of comments. It shows the two different patterns of watching and commenting. Users have high chance to randomly watch a lot of videos, while, under most circumstances, users only leave comments when they are very interested in videos. The property leads to the highly non-linearity between the number of comments and probability.

3.3 Video Length

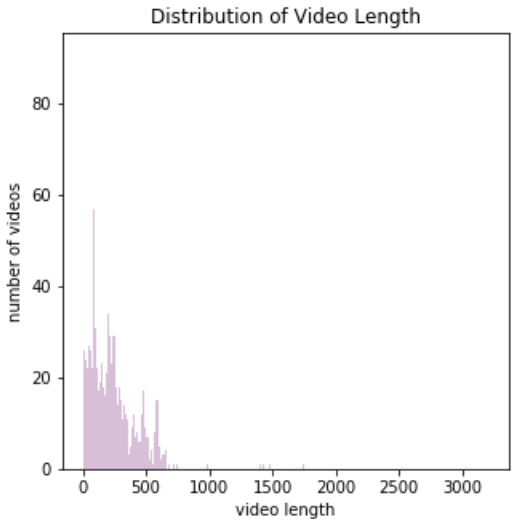

Fig 4. Distribution of video length

The most distinguished difference from traditional media content servers is the video length. Because YouTube is a UGC website, known as user-generated content, there would be a large part of short and medium length videos on YouTube. We have found that over 98% video length is within 600 seconds in our dataset. This is mainly due to the limit of 10 minutes imposed by YouTube on regular users uploads (The limit was raised to 15 minutes from 2010). We do find videos longer than this limit, because the limit was only established in March 2016, and also the YouTube Director Program allows a small group of authorized users to upload longer videos.

4 YouTube NETWORK PROPERTIES

The first step to constructing a network graph on YouTube. Nodes are videos in the dataset. To generate edges, if video *b* appears in the related_id list of video *a* then we consider there is a directed edge from *a* to *b*. The YouTube network can be shown as below:

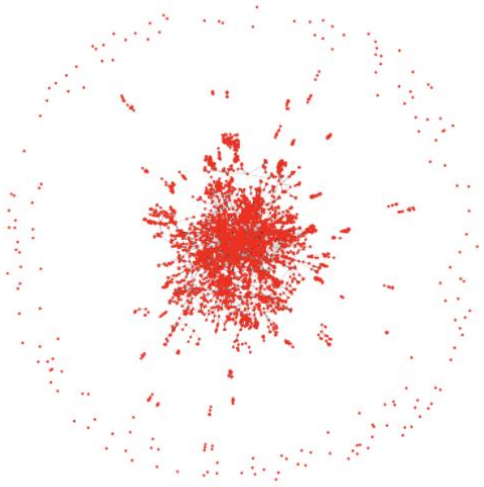

Fig 5. Network Visualization

It shows that the YouTube network is highly concentrated in the centroid part. And there are also many "outliers" that don't have strong connections with the major graph part. It is very similar to other real-life social networks, like Facebook friends' relation graph and twitter following relations graph.

After that, we implement the knowledge learn from the course and analyze how these videos are related to others. We calculate several network attributes of the YouTube dataset, and got the following outcome:

Table 2. Some results of graph properties

number of nodes	10,324
number of edges	34,969
average degree	6.7743
number of clusters	238
clustering coefficient	0.56

diameter	45

We can say from the table that although there are more than 10,000 videos in our dataset, and many of them are outliers, the diameter of the graph is only 45, which means that we can connect only two videos through less than 45 videos. It is exactly similar to the small world concept we talked about. These attributes prove our previous analysis, and it shows that the network nodes tend to create tightly knit groups characterized by a relatively high density of ties.

We also did the degrees distribution analysis. It turns out that degree distribution is a power-law distribution. Fitting into the model, the estimated alpha is 28 and the minimum X value is also 28. The number of degrees - count plot shows that the majority videos have only 1 or 2 degrees, and the count is decreasing with the growing number of degree. The power-law property could be explained by the fact that most of the traffic is generated by those most popular or most viewed videos.

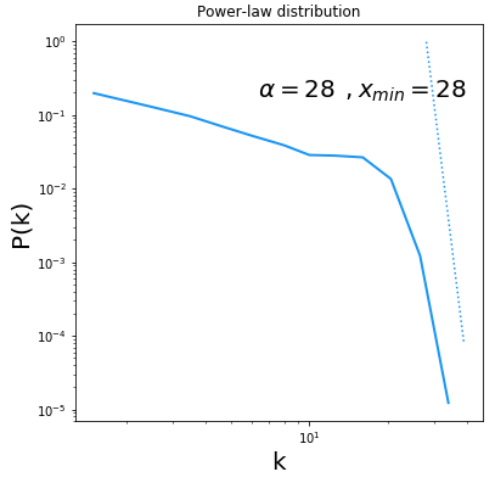

Fig 6. Model Distribution

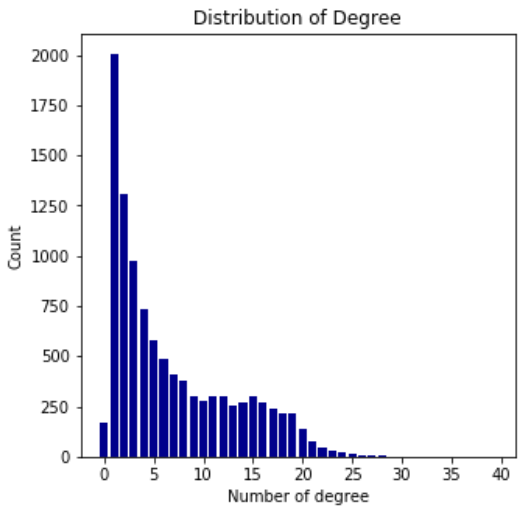

Fig 7. Number of Degree

5 YouTube NETWORK APPLICATION

5.1 Channel Recommendation

One of the significant business application of this network analysis is to build a channel recommendation system. While ordinary video recommendation system focuses on videos, we believe that if a user is interested in several videos uploaded by one same uploader, there is a very high probability that the user will also be interested in this uploader's other videos. Since our dataset doesn't contain each user's watch history, so we are not able to build a User-Based collaborative filtering algorithm. We decided to approach it with an indirectly but effective method.

we start with a video id. In the dataset, we have the information of video relation, which is based on video categories and real users' watching history. After getting the related videos' id, we find out the uploaders of these videos and count the frequency of different uploaders in this "related videos' uploaders" list. If a certain uploader appears more than or equal to twice in the list, we assume that these uploaders should to recommended to the user. If the related videos don't have common uploaders, we rate them by their score and recommend the uploader of highest score video. The score is not an attribute in the given dataset, but we calculate it by multiplying the rating by the numbers of views. We believe it's a balanced measure of video quality, which takes both the rating and popularity into account. The following graph also presents the pipeline:

Fig 8. Channel Recommendation System

To give a specific example. If a user watches a video with ID **[8ud8Mcmxo1M].** We extract the related videos list:

['PVBABmTh-pE',
'NUT3Lvn3iDM',
'R1cUwae72QY',
'EPIzEonZpgo',
'ydLOYXkEBic'
,'j6I7G7NdujA',
'lklKQ9h6JvQ',
'1LNbIHN9Rtk',
'm2UPsxrWL3M',
'q8rbwoQun1I',
'MHScD9ieUvQ',
'qBJybe46ukM',
'7cYAqoJ89sk',
'DurwYjia91w',
'fnfyIkpraec',
'lgmZZx3jnik',
'Yul7FmCPJRM',
'HqprbhLKRV8',
'A6nCzO5_OYA',
'DLghP1pjYJY'].

Among all uploaders of these videos, we found 7 uploaders appearing equal to or more than twice:

[loved30', 'VamRipper', 'ARAO13', 'grigoku' ,'THEMJAYZYO', 'Kyuubi2Naruto','forgotten38']

Therefore, we will recommend these channels to the user.

Of course, the realistic condition will be much more complicated. For instance, it will not be possible to recommend so many channels to a user just because of one single video watch, but we believe the logic behind our hypothetical and real algorithm should be similar.

Due to the limitation of our dataset, we are not able to evaluate the accuracy of our channel recommendation system, since we don't have the user watching history. However, we found a very interesting fact that, there are usually several common uploaders for a video, which implies that uploader is good indicators of user's preference.

5.2 Advertisement Application

The YouTube monetization ecosystem consists of three key players namely: Viewers, Uploader and Advertisers. Viewers or Users are the end audience for the content that is created by the uploaders and advertisers. Their response helps in customer targeting by gathering information on them. Uploaders or Channels are the ones who create content on YouTube and reach out to viewers.

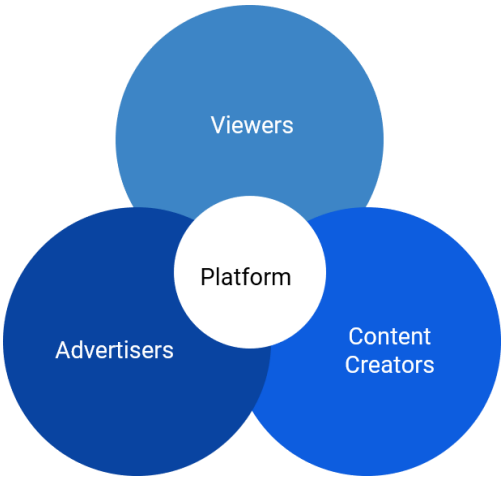

Fig 9. YouTube Monetization Ecosystem

The YouTube Monetization system, essentially, works by placing ads in the beginning, middle or the end of the videos or in the video popups. This model of advertisement is called pre-roll ads. The Uploader gets paid each time the user watches an ad that is featured in the video all the way through. We also consider product placement in the videos as a possible alternative application for this service. In this the Advertiser can select the channels that they want their products to be promoted by in the video.

In order to this model to work from an advertiser's perspective, we need to maximize the outreach of the product. Maximizing outreach includes promoting the product in the channel that consistently has more views and on the videos that are trending. For building a recommendation system we first need to know how ad targeting works in the real world.

Fig 10. Advertisers

1. Budget: Knowing the Advertisers budget will help us in segmentation of the channels. So for advertisers that Pay higher price we can offer the videos that have high latest popularity as well as channels that consistently perform better on YouTube metrics.
2. Product Type: We can also divide the audience base by the type of product the company wants to promote. So for example beauty products will be associated with DIY, Beauty and Style, How To's and Lifestyle channels.
3. Target Demographic: YouTube collects analytic data on its users by using Google AdSense and AdWords. Every video has detailed information that pertains to how often the video was watched since its upload, the trend lines, video activity: ratings, comments likes and subscription driven by the video. This section also gives information on the demographic that has watched the video and where the viewers hail from.
4. Geography: Geography is another factor that is of significant importance in detecting where the target audience is from and how to increase outreach of product in the given domain. Since YouTube ad network is served by Google, the outreach can include up to 1 billion users and 90% of the internet. This is why it is important to segregate the channels by the geographical location/language they cater to.
5. Other Preferences: This section caters to any specific preferences the advertisers may have such as in case they want to promote more on certain channels or not include some in their campaign especially the ones that don't follow the community guidelines. This also helps regulate in case of Mature ad content or brand guidelines.

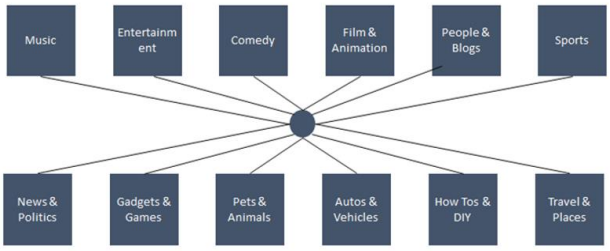

Fig 11. Graph of YouTube categories with Platform at the Center

Graph Network Analysis of the network shows that YouTube has the above 12 categories in which the videos are uploaded.

Since the dataset is from early 2007, we can expect some channels to be defunct and some videos to no longer be available. But the

Fig 12. Top video uploaders/channels of 2007 of each category

5.2 Demo

We conclude this project with a demo application that asks the user to input the advertising details along with a the VideoID of the video that they are most interested in.

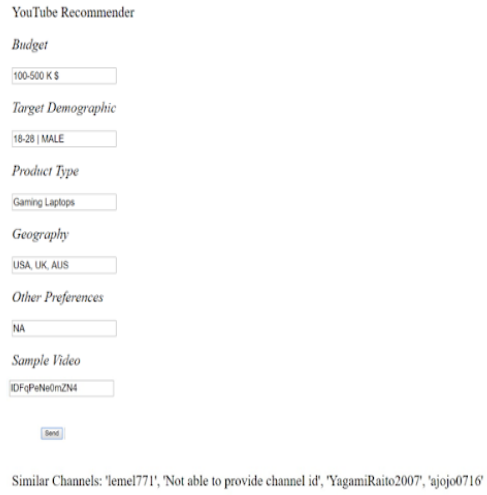

Fig 13. Demo Structure

6 FURTHER RESEARCH

Two interesting ideas for further research that we think about are the network analysis of the video comments and YouTube analysis regarding user location. For instance, we can implement the natural language processing technologies into the network analysis and conduct the sentiment analysis. It will be interesting to analyze what kind of video get positive as well as negative comments, and people's behaviors on YouTube. Besides, that it's also significant to analyze user regarding their location. We can build a comprehensive network combining both the videos and real user geography. Through it, we will be able to analyze people's preference in different locations as well as the trendy videos in different locations, which will give us the ability to build a better recommendation system and advertisement system.

REFERENCES

1. Cheng, Xu, et al. Dataset for "Statistics and Social Network of YouTube Videos". 2008, netsg.cs.sfu.ca/youtubedata/.
2. Ma, Xiaoqiang, et al. "Exploring sharing patterns for video recommendation on YouTube-like social media." Multimedia Systems, vol. 20, no. 6, 2013, pp. 675–691., doi:10.1007/s00530-013-0309-1.
3. Pitas, Iōannēs. Graph-Based social media analysis. CRC Press, 2016.
4. Ro, Yonghyun, et al. Youtube Graph Network Model and Analysis. http://snap.stanford.edu/class/cs224w-2014/projects2014/cs224w-37-final.pdf
5. Reiffers, Alexandre, et al. "A Study of YouTube recommendation graph based on measurements and stochastic tools." HAL, https://hal.inria.fr/hal-01217047/document.
6. Malliaros, Fragkiskos D., et al. "Locating influential nodes in complex networks." Scientific Reports, vol. 6, no. 1, 2016, doi:10.1038/srep19307.
7. Rossi, Fragkiskos D, et al. "Spread it Good, Spread it Fast." Proceedings of the 24th International Conference on World Wide Web - WWW 15 Companion, 2015, doi:http://fragkiskos.me/papers/influential_spreaders_WWW_2015.pdf.
8. Torres, Ian, and Jacob Cornard Trinidad. CS 224W Project Milestone Analysis of the YouTube Channel Recommendation Network. 8 Dec. 2015,snap.stanford.edu/class/cs224w-2015/projects_2015/Analysis_of_the_YouTube_Channel_Recommendation_Network.pdf.
9. Merrer, Erwan Le, and Gilles Trédan. "The topological face of recommendation: models and application to bias detection." CORE, 1 Jan. 1970, core.ac.uk/display/83853960.
10. Tsingalis, Ioannis, et al. A statistical and clustering study on Youtube 2D and 3D video recommendation graph . ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6877872.

11. T., Tejal, and Sheetal A. “YouTube Video Recommendation via Cross-Network Collaboration.” International Journal of Computer Applications, vol. 146, no. 11, 2016, pp. 9–17., doi:10.5120/ijca2016910896.
12. https://www.sciencedirect.com/science/article/pii/S1389128608003423