Youtube Data Analysis

Avishek Bose* Biplav Timalsina† Naresh Kumar Giri‡
Department of Computer Science and Engineering
University of Nevada, Reno
Email: *bose.avishek@nevada.unr.edu, †btimalsina@nevada.unr.edu ‡ngiri@nevada.unr.edu

***Abstract*—YouTube is the largest Online Video Sharing Sites(OVSS) in the world; it is one of the prominent sites to disseminate information in form of videos. Since, little is known about the properties of OVSS, the factors that shape the structure, interest of individuals on YouTube, we intend to analyze YouTube dataset of 3 months. This work examines the social network aspect of YouTube by studying comments, rates, views graphs along with structure of the YouTube and reveal insights on how the videos are distributed in YouTube. At first, we study degree distribution of the dataset and we also plotted in graph to see its nature. Then, we study nature of videos based on the categorization statistics. Finally, we perform Principal Component Analysis(PCA), Factor Analysis(FA), Growth trend, and check for normality for the dataset.**

We found that, degree distribution follows power log normal nature. And we also found that certain categories of data are more prominent in YouTube than others, and compare the different categories of data as well. There is strong correlation between Ratings and Comments for videos, and no other peculiar relations between the attributes. None of the attributes were normally distributed. PCA revealed the interplay of User's Interest contributes maximum to the variability in the data. FA also reveals rate received by videos are not related to any other attributes available. Clustering analysis revealed that there were no distinct clusters and all videos were randomly distributed. Growth trend analysis revealed uniform growth trend for rate, very high variability for views, comments and number of rates.

I. Introduction

Recent decade has seen tremendous sharing of information in online sharing platforms in the form of images, blogs and videos. With the advent of latest technology to support storage of large amount of data, popularity of videos has increased dramatically. It can be felt by the commotion it has generated at sites like YouTube, Facebook, DailyMotion, Vimeo and likewise. While all of these sites have their own followers, YouTube is one of the most sought after OVSS today. As of 2017, the total number of people using YouTube has reached a staggering 1.3 Billion and almost 4 billion videos are watched every single day on YouTube. Almost 300 hours of video are uploaded every minute, and the visitor count is almost 300 Million every day. These are some astonishing numbers. The dynamics of video popularity is dramatically changing every day. This work intends to better understand the hidden dynamics and reveal latent variables in play in the dataset. We use the YouTube dataset uploaded by Xu Cheng [1] which consists of data from 22 Feb, 2017 to May 18th, 2017 for a period of 3 months . We study the graph structure in terms of indegree, outdegree and total degree distribution, categorization statistics, and perform numerous test on the dataset to reveal essential information about the data. Questions addressed in this paper are:

1) What is nature of degree distribution?
2) What is the frequency of videos in different category?
3) Is there a direct correlation between rates, number of ratings, views and comments?
4) Is there any particular pattern in those attributes for various categories of data?
5) Can the youtube videos be clustered and specific characteristic features can be extrapolated?
6) Are there any hidden variables in play that explain the variability in the data?
7) What is the growth trend of rates, comments, views and number of ratings with time for different category of video?

The results show that the distribution of attributes of videos are rather randomly distributed, and do not particularly follow a pattern.

TABLE I: YouTube Metadata

Video ID	11 Digit unique identifier
Uploader	Uploader's Username
Age	Days the video has been on youtube
Category	Video Category
Length	Video Length
Views	Number of views
Rate	Video Rating
Ratings	Number of ratings
Comments	Number of Comments
Related ID	upto 20 related video IDs

The organization of the paper is as follows. Section II discuss some previous works. Section III explains the rationale behind this project. Different approaches used for the project are elaborated in Section IV and how those were implemented is elaborated in Section V. Section VI discuss the findings of the project. Section VII and Section VIII explains the specific contribution made by project and discusses on future work. Finally, we conclude with Section IX.

II. Background

In 2005, YouTube launched and had been emerged a major video sharing website where users can upload, watch and share videos with others. There has been significant study in the field of social network analysis.

The authors in [2] studied the user behavior and popularity distribution of the videos hosted on MSN. In [3] authors

have studied the analysis of social interaction on the video sharing and search services on youtube. Authors in [4], studied YouTube data. They gathered information of uploaded video on YouTube, and did network analysis and power law distribution analysis. Cheng et al. [5], have crawled data for the period of four months, and presented in-depth measurement study of statistics of 3 million YouTube videos.

In 2012, Mirjam et al. [6] examined the full-scale YouTube subscription graph, comment graph, and video content corpus. They also found that YouTube is deviating significantly from network characteristics that mark traditional online and social networks, such as homophily, reciprocative linking, and associativity.

Wattenhofer et al. in [7] studied YouTube network aspect measuring YouTube subscription graph, comment graph, and video content corpus. This work was carried by persons within YouTube itself. They find YouTube deviate from other traditional social online networks.

Sara et al. [**?**] did her research on aggregation techniques for characterization to ease computation on the large social network and to gain insight into subgroup interactions. She contributed to several analysis techniques for a wide variety of social network structures and available data using two methods for subgroup detection and application of two network measures.

Authors in [8] studied characteristics of YouTube, focusing on short video sharing. They found that the related videos on related field are chosen based on user preferences, have noticeable small-world characteristics.

In [9] authors studied problem of video recommendation methods for videos in online social networks that does not meet the needs of the users. To overcome this problem they propose a new model of trust based video recommendation approach.

Paper [10] discusses about popularity dynamics of videos in video sharing sites like YouTube, based on views, ratings and comments. They also present an emulator to model video viewing and sharing behavior of users.

III. Motivation/Significance/Challenge

In modern day, YouTube is the largest user-driven video content reservoir in the world. Although some research present in this domain that we have discussed in Section II, the network analysis of YouTube data has not been fully explored and there is much more information to be mined. Analyzing YouTube dataset will not only be useful for big data analysts such as market analysis, recommendation system, user preferences, degree analysis but also from economic perspective such as product advertisement criteria, channel subscription promotion, user-friendly customization of YouTube itself. It is a fact that the digital world is producing massive data. Big data analysis will help us gain better insights in our approach to business. Additionally, working with the dataset of social networks like YouTube can be a good initial to go deep in big data analysis.

IV. Approach

Categorization and degree distribution is studied by MapReduce framework. Analysis of data is done in R. For preprocessing, the data with null value is erased and analysis is done.

A. MapReduce Framework

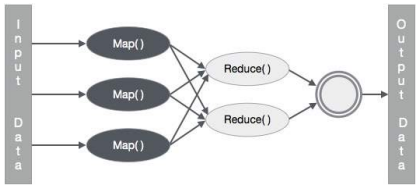


Fig. 1: Simple Diagram of MapReduce Framework.

MapReduce is a framework for big data analysis which is used to write programs that processes huge amount of data, in parallel on large commodity hardware cluster in a reliable manner. This framework contains two tasks. First, Map which takes a set of data and converts it into another set of intermediate data and second, Reduce which takes an intermediate set of data as input and combines those sets into a smaller set of tuples as output. Fig. 1 shows the basic layout of MapReduce Framework.

B. Key Value Pair

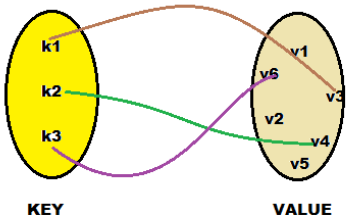


Fig. 2: Depiction of Key Value Pair.

Key-value connected pair is two values usually synthesized in such a way that the value is accessed using the key. In the case of big data analysis, it is a pivotal issue to implement on any module. In this work, we use single or multiple attributes as key and did same for the value as well if its needed to do. Fig. 2 shows the basic diagram of key-value pair arrangement.

C. Degree Distribution

Degree distribution is one of the key factor while studying networks. Degree distribution of the network represents the connection among the nodes. It represents how nodes are connected to other nodes. It also answers to the questions like, do all the nodes have equal amount of connection to other nodes, or some nodes have many connections while other have few connections? By analyzing these, we get to know if the network is centralized or decentralized. For each YouTube

video there are other related video that youtube suggests us to watch. We want to know the nature of the suggestions given by YouTube.

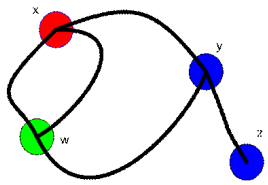


Fig. 3: Degree distribution of an Undirected Graph.

[11] For an undirected network, degree distribution of a node can be defined as total number of neighboring nodes to that nodes. If graph is represented in adjacency matrix, the degree of i^{th} node is represented as the sum of i^{th} row in that matrix.

$$D_i = \sum_j a_{ij} \tag{1}$$

where,D_i is degree of node i, a_{ij} is an element with i^{th} row and j^{th} column, j is total number of columns in the matrix. This can be understood from the Fig. 3. In the figure, we can see in node 'z' has only one edge or link and it's degree is one. In node 'y' has three edges and it's degree is three.

we can see that node 'z' has only one edge or link has degree one and node y has three edges or links has degree three.

In directed graph, we need to consider both edges coming into the node and some edges going out from the node. The degree distribution is calculated as summation of indegree and outdegree. Indegree is total count of incoming nodes whereas, outdegree is total count of outgoing nodes. The distinction of indegree and outdegree also provides some more information of the graph.

$$D_i{}^{in} = \sum_j a_{ij} \tag{2}$$

where $D_i{}^{in}$ is indegree of a node D_i.

$$D_i{}^{out} = \sum_j a_{ij} \tag{3}$$

where, $D_i{}^{out}$ is outdegree of node D_i.

$$D_i{}^{total} = D_i{}^{in} + D_i{}^{out} \tag{4}$$

where, $D_i{}^{total}$ is total degree of node D_i.

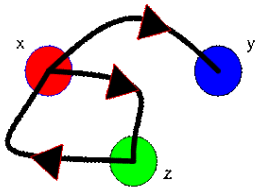


Fig. 4: Degree distribution of an Directed Graph.

Degree of the directed graph can be described form Fig. 3. Node 'y' has one incoming edge from node 'x' so node 'y' has one indegree. There are no outgoing edge from node 'y', so it has total degree one. Node 'x' has one incoming edge from edge 'z', so node 'x' has one indegree. And node 'x' has two outgoing edges and has two outdegree. So adding indegree and outdegree, total degree of node 'x' is three.

V. Implementation

Indegree and outdegree are calculated from the database using map and reduce function. For calculating outdegree, the videoID is used as key and number of related videos is used as value. As the database contains maximum 20 related videos, so the outdegree of any video can be only up to 20. Indegree is calculated as, for any videoID in the related field section is key and value is 1. For the calculation of total degree, the output of indegree file and outdegree file are given as input for map and reduce. The key is videoID for both files and value is value after videoID (indegree for output of indegree file, and outdegree for output of outdegree file).

We implement MapReduce framework using Java and used IDE named Spring Suite Tool in Linux environment. We create ten individual java files to make the analysis easy over the dataset. Following is the list of codes written to categorize dataset

- 1_mean_cmnt_cnt_by_ctgr.java: Java code in this file calculates the average number of comment by each video category.
- 2_mean_ratng_cnt_by_ctgr.java: Java code in this file calculates the average number of ratings by each video category.
- 3_mean_view_cnt_by_ctgr.java: Java code in this file calculates the average number of view count by each video category.
- 4_No_of_video_by_ctgr.java: Java code in this file calculates the number of videos by each video category.
- 5_size_ctgr_rng_id_cnt.java: Java code in this file calculates the id count belongs to particular size range by providing a category name and two integer values for ranges as user input.
- 6_size_with_id_count.java: Java code in this file calculates the id count belongs to particular size range by providing two integer values for ranges as user input.
- 7_view_cnt_ctgr.java: Java code in this file calculates the frequency of view count of videos by sequential ranges of view-count.
- 8_size_cnt_ctgr.java: Java code in this file calculates the frequency of videos by sequential ranges of size.
- 9_top_X_rtd_video_Id.java: Java code in this file calculates top X-rated video ID. First, it calculates the ratings of each video. Then, using Linux command (sort–n –k7 –r | head –nX), we get the IDs of top X-rated videos.
- 10_top_Y_pop_video_id.java: Java code in this file calculates top Y-rated video ID. First, it calculates the ratings of each video. Then, using Linux command (sort –n –k5 –r | head –nY), we get the IDs of top Y-rated videos.

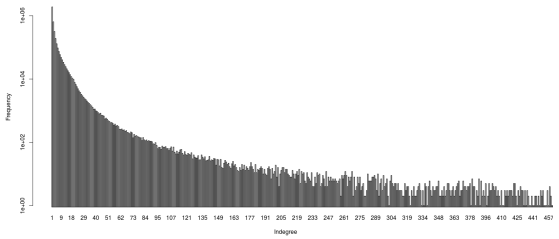


Fig. 5: Indegree distribution

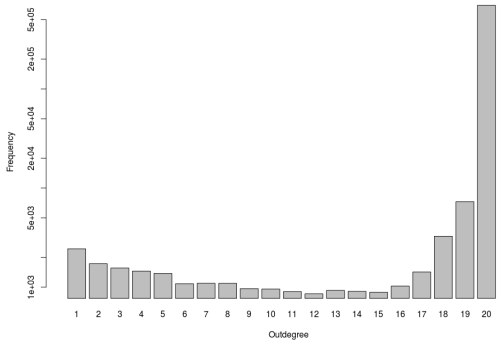


Fig. 6: Outdegree distribution

For the analysis of the dataset, we used R studio. First, the data was pre-processed to remove inconsistent values and no values. After we cleaned the data we performed analysis.

VI. Evaluation

A. Degree Distribution and Graph Plot

The above graph shown in Fig. 5, shows the indegree distribution. The x-axis is indegree and y-axis is number of videos having that particular indegree. We can see that the number of videos having few indegree is much more which is around more than one million. And as the indegree increases the frequency of video decreases exponentially. In this distribution, we can observe the power law distribution. Most of videos have indegree less than 20. For total data of $3,772,776$ the maximum indegree found is 1882 and minimum indegree is 1. The average indegree value is 3.800597.

The graph shown in Fig. 6 shows the outdegree distribution. Outdegree is shown in x-axis and frequency is shown in y-axis. In this plot we can see that maximum frequency is of the video having outdegree as 20. This is because database includes only 20 related videos which is represented as outdegree. There is no particular patter in the outdegree but most of the videos have maximum outdegree. For total dataset of $730,119$ the maximum outdegree is found to be 20 and minimum to be 1. The average outdegree is 19.63899.

The graph shown in Fig 7 shows the total degree distribution. Large number of videos have very less total degree distribution. Only few videos have more number of connections with other videos. From the figure, we can see that the

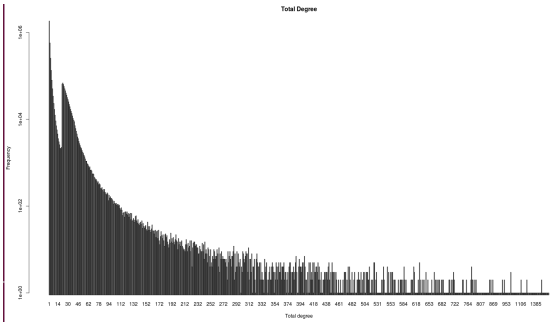


Fig. 7: Total distribution

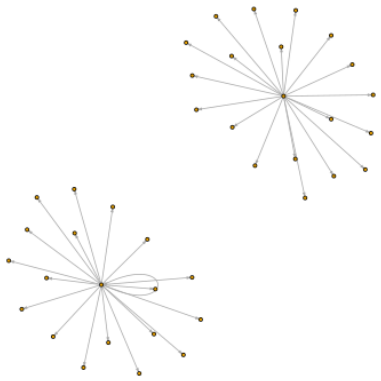


Fig. 8: Plot of two nodes of videos

total degree distribution decays exponentially for increasing degree distribution.

The above graphs shown in Fig 8 and Fig 9 gives the pictorial representation of outdegree of the dataset we used. From the graphical representation we can see the going from one node to another node. Most of the nodes in the outer range do not have connection with other nodes. Those node contain some related nodes, which are not connected with any other nodes.

B. Categorization Implementation

Fig. 10, Fig. 11, and Fig. 12 shows the graphs of the mean value of comment, mean ratings, and mean view count by video categories respectively. Fig. 13 shows frequencies of videos by categories. Fig. 14 shows frequencies of videos by ranges of size.

C. Correlation Analysis

Correlation is one of the most common and most useful statistics. A correlation is single number that describes the degree of relationship between the two variables. A covariance matrix (also known as dispersion matrix or variance-covariance matrix) is a matrix whose element in the i,j position

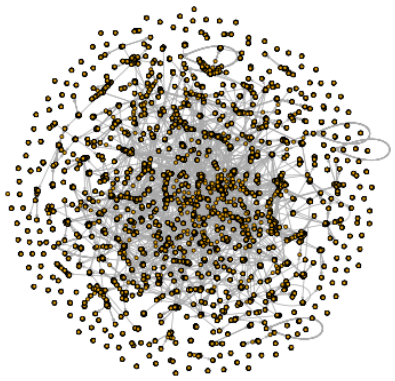


Fig. 9: Plot of all nodes of a file

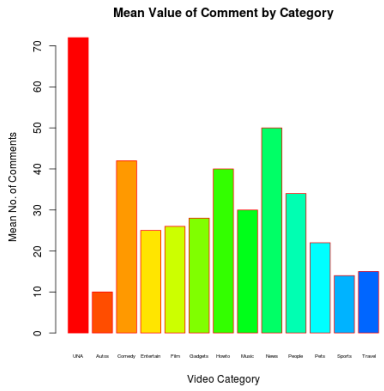


Fig. 10: Mean Value of Comment by Video Category.

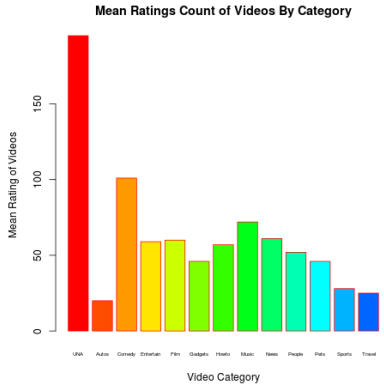


Fig. 11: Mean Value of Ratings by Video Category.

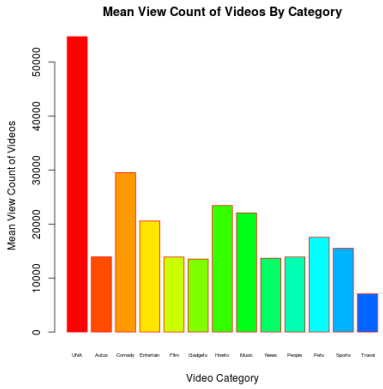


Fig. 12: Mean Value of View Count by Video Category.

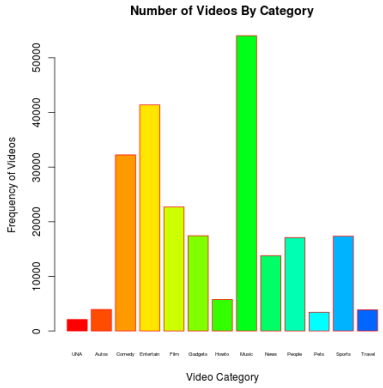


Fig. 13: Frequencies of Videos by Categories.

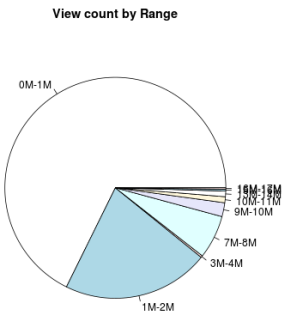


Fig. 14: Frequencies of Videos by Ranges of Size.

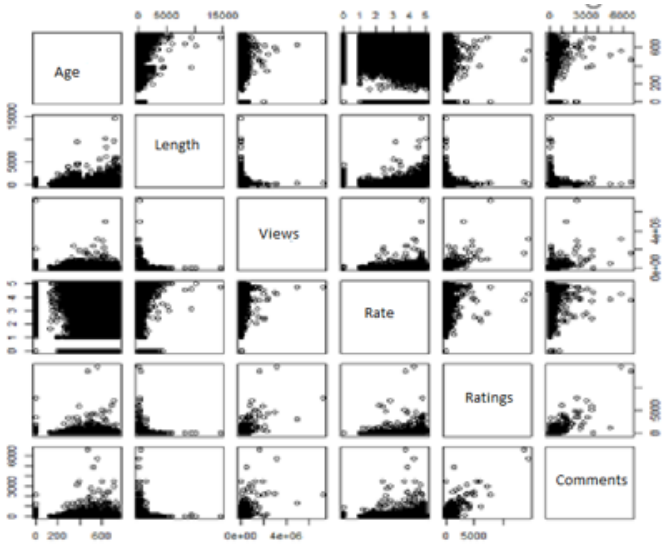


Fig. 15: Scatter Plot

is the covariance between the ith and jth elements of a random vector. A vectors is a random variable with multiple dimensions. Variance,

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \tag{5}$$

Covariance,

$$Cov(x,y) = \frac{\Sigma(X_i-\mu_i)(y_j-\mu_j)}{N-1} \tag{6}$$

Variance indicates how much a random variable varies w.r.to its mean position or the variability of a random variable.Co-variance is an index that represents how much two variables vary w.r.to their means, when looked in a combined approach. Correlation matrix is also a kind of co-variance matrix, but it is normalized to avoid the contribution of different scales the data may be in.

TABLE II: Correlation Matrix

	Age	Length	View	Rate	Rating	Comments
Age	1					
Length	0	1				
Views	-0.12	0	1			
Rate	-0.12	0	0.7	1		
Ratings	-0.08	0	0.5	0.5	1	
Comments	-0.09	0	0.6	0.07	0.58	1

The maximum correlation that can be seen is 0.7 between views and rate, and 0.58 between comments and ratings, which is definitely intuitive. A video that has a large number of comments would also receive larger number of ratings, and a video which receives a large number of views would also receive large number of ratings. Age is negatively correlated with almost every variables. This is quite counter-intuitive, as videos that are on youtube for a longer period of time should receive more views, rate, ratings and comments. This phenomenon may reveal the information that newly updated videos are most sought after than the old ones, and that metrics related to old videos in general dont grow as highly as the newly posted videos. Similarly, length of the videos has nothing to do with any other metrics. The number of views a video receives is highly correlated with the number of ratings, or rather the number of ratings a video receives is highly correlated with the number of views it receives, which is quite obvious. Similar observations can be made about other metrics in study.

From the Fig. 15 , we can see that there is no definite correlation between any variables except ratings and comments, which is also not very high.

D. Normality Tests

While doing any kind of analysis, assuming normality enables us to perform various kind of rigorous mathematical analysis, and come up with suitable understanding of the underlying sample domain. Therefore, for preliminary understanding of the data, we performed normality test for the various variables available. Normality test tries to measure if the data sample in our data table is normally distributed or not, and compute how likely it is for a random variable underlying the data set to be normally distributed. And, it can also be used for model selection measures for different analysis. The red line in the figures are the line of mean for the corresponding variables. A random variable is said to be normal if the data points are distributed around it uniformly. From our observation, length is almost normally distributed as compared to other variables. Views, Ratings, Rate and comments (in the next page) are not normally distributed. This gives us some insight that the videos uploaded in Youtube as a whole are not uniformly distributed, and that some videos receive more views, ratings, rate and comments than other videos while the length of video is almost normal. Now, we will also check if our data is multivariate normally distributed. A random vectors is said to be multivariate normally distributed if every linear combination of its components

$$y = a_1X_1 + a_2X_2 + ... + a_nX_n \tag{7}$$

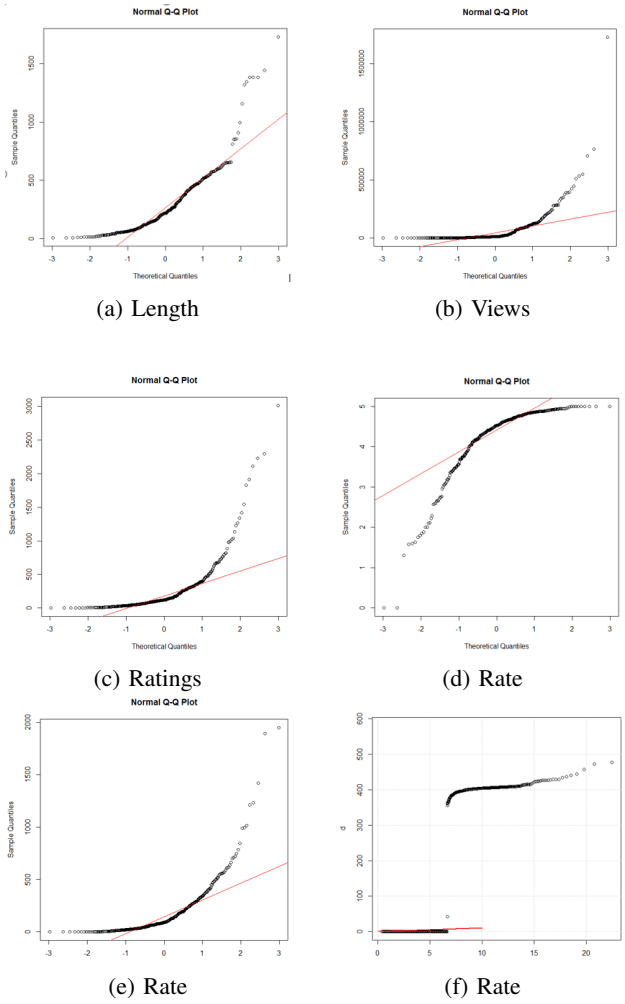
is normally distributed.

We can test for multivariate normality by using 2 test of independence. In order to establish that 2 categorical variables are dependent, the chi-squared statistic should be above a certain cutoff. This cutoff increases as the number of classes within the variable increases. Alternatively, we could just perform a chi-square test and check the p-values. Like all statictical tests, chi-squared test assumes a null hypothesis and an alternate hypothesis. The general practice is, if the p-values that comes out in the result is less than a pre-determined significance level, which is 0.05 usually, then we reject the null hypothesis.

H0: The variables are independent.

H1: The variables are related.

The null hypothesis is that the variables are independent, and



(a) Length (b) Views

(c) Ratings (d) Rate

(e) Rate (f) Rate

Fig. 16: Normality tests

the alternate hypothesis is that they are related. After analysis, the following values were obtained: P=2.2e-16 Since this is $<$0.05, we reject the Null Hypothesis and concluded that the variables are related to each other. From Fig. 16f we see that the data is not multivariate normal.

E. Principal Component Analysis

Principal Component Analysis (PCA) is a dimension reduction tool that can be used to reduce large set of variables to a small set that still contains most of the information in the large set. It is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. PCA can be considered as a rotation of the axes of the original variable coordinate system to new orthogonal axes, called principal axes, such that the new axes coincide with directions of maximum variation of original observations.

The first component accounts for as much of the variability in the data as possible and each succeeding component accounts for as much of the remaining variability as possible. The property of maximum variation of the projected points defines the first principal axis; it is the line or direction with maximum variation of the projected values of the original data points.

TABLE III: Principal Components

	Comp1	Comp2	Comp3	Comp4	Comp5
Age		-.72	0.69		
Length	0.54			0.79	0.27
Views	0.11	-0.69	-0.72		
Rate	0.6			-0.15	-.78
Ratings	0.61			-0.15	-0.78
Comments	0.57			-0.59	0.57
Proportion	0.49	0.23	0.17	0.09	0.03

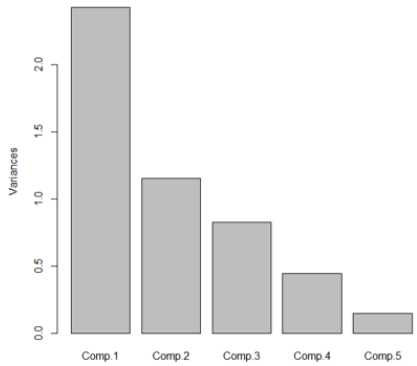


Fig. 17: Variance/Component Diagram

We can see that the from Table III and Fig. 17 that the proportion of variance contributed by the first principal component (PC) is 48.5%, by second PC is 23%, third PC is 16.5% and likewise.

It is obvious that the variance of the variables is not concentrated in a few principal components, and we at least need two components to account for 71.6% of the total variance in the dataset.

From Table III, we can see that the first principal component has strong correlation with ratings, views and comments, and this component can be described as Users interest in the videos. Similarly, the second principal component has strong negative correlation with length and rate, and this component can describe the videos that have long length but still received bad ratings.

F. Factor Analysis

It is a statistical method used to describe the variability among observed variables in terms of potentially lower number of unobserved variables, called factors. The observed variables are modelled as linear combinations of the potential factors, plus error terms. By using factor analysis, we can reduce a dataset with large number of variables to one with relatively small number of variables. If there are p random variable $x_1, ...x_p$ with mean $\mu_1 ... \mu_p$. Then suppose for some unknown constants l_{ij} and k unobserved random variables F_j called common factors (because they influence all the observed

random variables), where i=1,2,3... p and j= 1,2,... k, such that k<p, we have

$$x_i - \mu_i = l_{i1}F_1 + \ + l_{ik}F_k + \varepsilon_i \tag{8}$$

In matrix terms, we can represent them as,

$$x - \mu = LF + \varepsilon \tag{9}$$

where, L is matrix of factor loadings, F is factors, ϵ is the vector of errors. We also assume that F and ϵ are independent, expected value of F is zero and that all the factors are uncorrelated. If cov(x- μ)= σ, then

$$= LL^T + \mu \tag{10}$$

$$Var(xi) = l_{i1}^2 + ... + l_{im}^2 + \Psi_i \tag{11}$$

Finally, with this relation we will be able to explain the variability of our data in terms of factor loadings. ψ is the covariance matrix of errors, which is a diagonal matrix. One of the hardest things to do during Factor Analysis is how many factors to settle on. One of the way is to consider all those factors with eigen value >=1. This is because a factor with an eigenvalue of 1 accounts for as much variance as a single variable, and the logic is that only factors that explain at least the same amount of variance as a single variable is worth keeping. But this should be taken only as a tool, not a hard and fast rule to select number of factors. Another option is using Scree plot. A scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. It always should be a downward curve. The point where the slope of the curve is leveling of indicated the number of factor that should be generated by the analysis. But it can also be just used as a tool rather than a hard and fast rule.

Another important metric is to keep in mind is the total amount of variability of the original variables explained by each variables explained by each factor solution. We cannot be using more than two factors in this Factor Analysis, because from the formula:

$$M <= (p-1)/2 \tag{12}$$

Where, M=no of factors P=no of variables For our case, p=5, therefore, m =1,2. Hence, we cannot work with greater than 2 number of factors. We will now observe the value of loading, communalities and uniqueness of the different variable of our dataset, for different kinds of rotation and for different number of factors. By doing this brute force approach (and not using approaches of eigen values, scree plot) we are trying to select which model of FA would be best for understanding our data while looking at all of our possibilities.

TABLE IV: One Factor Analysis

Variables	None	Varimax	Promax	Communality	Unique
Length				0.01	0.99
Views	0.72	0.72	0.72	0.52	0.49
Rate				0.01	-.99
Ratings	0.99	0.99	0.99	1	0.005
Comment	0.83	0.83	0.83	0.68	0.32

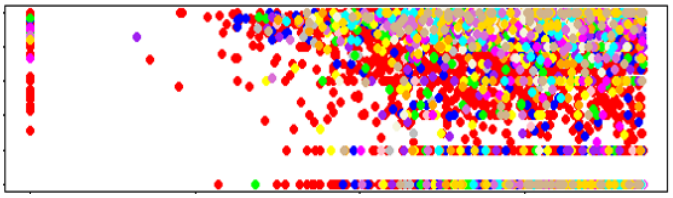
Fig. 18: Factor Loadings for two factors

	Factor Loadings (2 Factor)							
Variables	Unrotated		Varimax		Promax		Communalities	Uniqueness
	F1	F2	F1	F2	F1	F2		
Length	0.675	-0.735		0.995		0.998	1	0.00
Views	0.541	0.468	0.712		0.716		0.52	0.489
Rate	0.182			0.178		0.168	0.04	0.962
Ratings	0.771	0.633	0.99	0.121	0.998		1	0.00
Comments	0.640	0.519	0.817	0.105	0.824		0.68	0.321

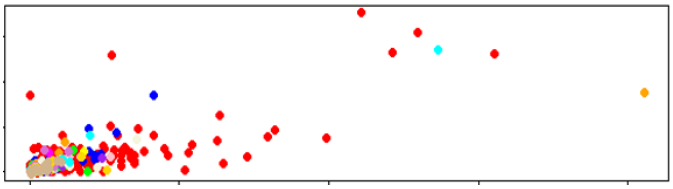
In one factor case, only 43% of the total variability was covered by the factor, which is very low. Hence, we could say we cannot purely represent the variability in the data using 1 factor. In this case, length and rate of the videos were unaccounted for, and they behaved separately contributing differently to the variance of the dataset. The first factor related well with ratings, comments and views, which can again be viewed as General Interest of users on the videos. It is congruent with out insight from our Principal Component Analysis. The cumulative variance covered by two factors were almost similar for various rotations, and it accounted for only 64.4% of the total variance of data in two factor case. The representation of the variables in the factor was good for all variables except the Rate, which couldnt be addressed by any factors, as its uniqueness is 96.2%. From all these observations, we can derive that rate received by the videos cannot be predicted or is not related to any other variables in play. It is natural as ratings received by any video depends on the quality of the video, content and other things in play, and not on length, or views, number of ratings or comments.

G. Clustering

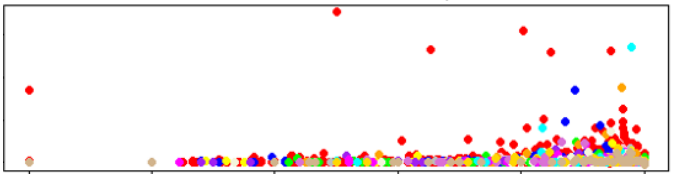
Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute best criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.We performed cluster analysis in our data with the intention of finding out if there are specific behavior within the data samples, and how they are related with each other. If there are definite clusters in our data, we would be able to derive the characteristic features and gain new insight in the data.



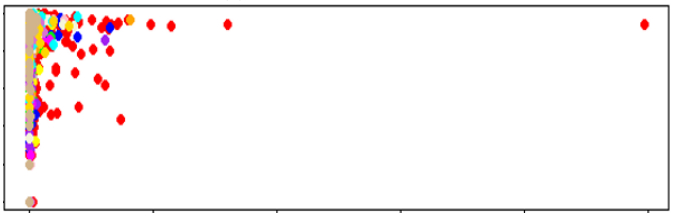
(a) Rate vs Age



(b) Comments vs Ratings

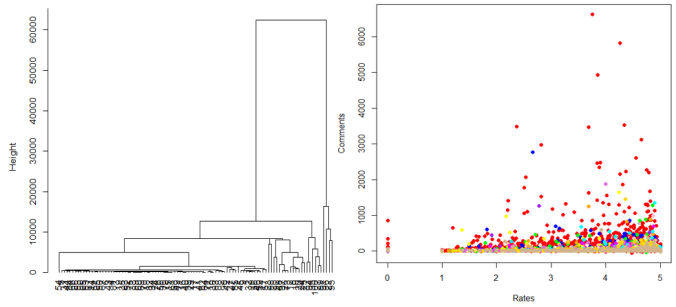


(c) Comments vs Views



(d) Rates vs Views

Fig. 19: Cluster Analysis



(a) Classification Dendogram (b) Clustering Comments/Rates

Fig. 20: Cluster Analysis

We clustered our data on the basis of different features, and as can be seen in the adjacent figures, there is not presence of definite cluster in the data. The different colors show the different category of data they were from.

In the Fig. 20, we can see the dendogram for the clustering with the attributes Comments/Rates. To the right of it, we can see that there is no definite clusters within the data. Hence, we are unable to derive any clusters in our dataset. And, we can say the distribution of attribute values is random in different categories of data, ie. videos belonging to different categories dont show characteristic behavior, rather they have the similar features and we cannot predict the category of the videos simply looking at the attributes.

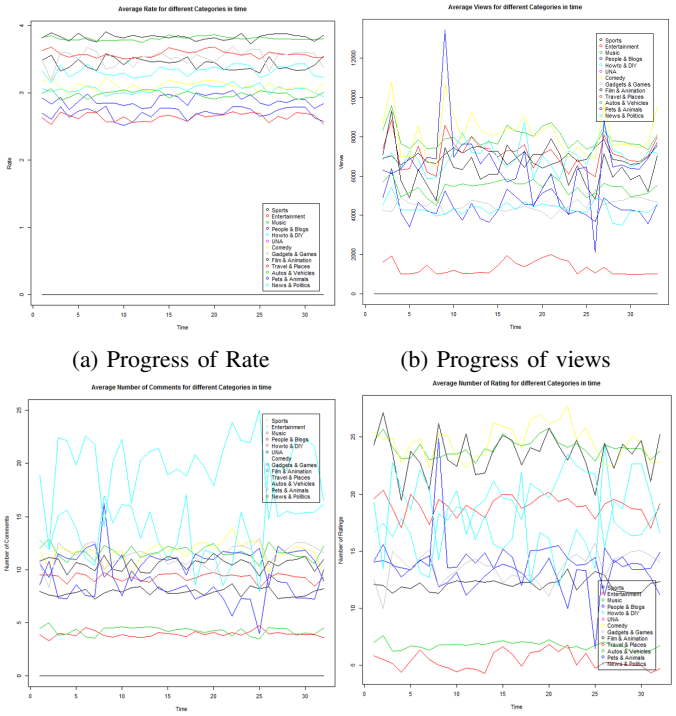
H. Growth Analysis

We now plan to observe the behavior trend of videos categorized into various innate categories and compare them with each other with data available for 3 months. We will be able to compare videos attributes and general trend according to category and gain a bird's eye vision of the general behavior of data in each category. In Fig. 21 progress of average rate, views, comments, and number of rates of various categories are plotted against time.

Fig 21a tells us that ratings received on a scale of 1 to 5 are generally stable for all categories with minor fluctuations here and there. Its almost as if they are normally distributed along a central mean line, with different variability for different categories. Music and Film & Animation category video received consistently higher rating that other categories. The ratings curve of music is really smooth. This can be explained as there is a lot of interest in music industry, and people do rate a large number of videos, and that could have resulted in the smoothing of the curve. While Film & Animation also receives good ratings, but it might be affected by certain influx of interest in films once released. We could also say that the spikes could have been caused for days with new releases, but we would have to check more on that. Also, it could be caused because it receives less number of ratings, and a sudden rise in interest in videos could dominate the trend of the ratings. The lowest ratings are received by the Travel & Places. It is obvious as such videos are far less compared to other types of videos and fall less in the trending videos, and thus are viewed less and rated less. It also may be because these videos are usually individually created, and thus are less professional, or cover only one aspect of the destination, and people like rate it less.

Fig 21b shows the growth trend of average views for different category of data. This analysis shows that the views for the videos for each category is very random with spikes. The most prominently varying views is received by Pets & Animals section, ranging at a highest of 13440 views at 16th May and lowest of 2014 views on 30 June. This strange sudden increase in interest and sudden downfall could have been because the a lot of video fell in the top section on 16th May, and the opposite in June 30. Other prominent curve is the curve for Travel and Places that received consistently low views throughout. This insight is congruent with our previous reasoning.Other popular videos categories are Comedy, Music and Entertainment which share almost similar curves.

Fig. 21c and 21b shows the trend for comments and views. Comments for News & Politics is seconded by HowTo & DIY, which is obvious. The fluctuation seen can be addressed to the volatile nature of events happening in the world, and also release of new DIY videos that has sparked the interest of viewers. Autos and Vehicles and Travel & Places receive the lowest of comments. As these videos are viewed less



(a) Progress of Rate

(b) Progress of views

(c) Progress of Comments

(d) Progress of Number of rates

Fig. 21: Growth Analysis

often, lower number of comments is obvious. This is because the coverage of such videos are to a select few interested, while categories like music, entertainment, sports covers a wide audience, and hence enjoys larger interest. Similarly, number of ratings received are really fluctuating and can be used to distinguish different categories. Sports,comedy and music receive highest ratings, while travel and places receive the lowest number of ratings. Thus, we can verify that Youtube viewers are more interested in some categories of videos than others.

VII. Contribution

This article performed rigorous analysis on YouTube Dataset available to us. First, we studied the structure of distribution of videos in the dataset; we also visualised the data by plotting them in graphs. Futher, we studied the indegree, outdegree and total degree for the dataset. We verified power law nature of the data. Then, we performed categorization for various search condition. We then plotted the result using graphs. Further, we studied if the attributes of the data were normally distributed, performed chi-square tests to check multi-normality. We found that the data were not normally distributed. Then, from PCA we found that the variability of the data were explained by User's General Interest. Factor Analysis also revealed the same hidden variables. When we performed cluster analysis on the dataset, we were expecting to find some hidden cluster that related to particular category of data, but surprisingly, videos were randomly distributed and couldn't be clustered. Study of growth trend of data revealed that the users were particularly interested in certain category of data such as Music, Entertainment and Sports, and viewed and commented on such videos. However, there was no specific pattern for the rates received by videos. It can be explained by the fact that people tend to give more rates to high quality videos rather than being affected by category.

VIII. Discussion

This project was able to work on Degree Distribution, Categorization statistics and various analysis. Given the nature of the data and the time frame, we were able to generate some essential pattern in the dataset which will be helpful to people interested in understanding the hidden dynamics of YouTube. The data we used was crawled in 2007 and it would have been better if we used latest dataset. In that way, the analysis would have resulted in more relevant information. Apart from that, we could work on social network behaviour on the basis of user information, and calculating popularity index that characterizes the videos trend in YouTube. Further, we could work on the uploading pattern of uploaders.

IX. Conclusion

Degree distributions are essential tools to understand the hidden structure of a system. Similarly, categorization statistics provides various insights on the videos in the dataset. Further, analyzing the dataset for normality, multinormality and using Principal Component Analysis and Factor Analysis to understand the hidden interplay of variables helps us to understand the very nature the data are distributed. Similarly, cluster analysis and growth trend analysis help us to find pattern in the data set, and can even help in modeling the data. Our work sheds light on trying to know more about an OVSS, but it can be used on any kind of websites. As we progress in technology, understanding the social behaviour and technological aspects of such system will enable us to better understand, analyse and predict about such systems, which is further essential part of growing technology. Our work is an effort to do the same, and we believe we have done so by performing various kind of rigorous mathematical analysis.

References

[1] J. L. Xu Cheng, Cameron Dale. (2008) Dataset for "statistics and social network for youtube videos". [Accessed 3-April-2017]. [Online]. Available: http://netsg.cs.sfu.ca/youtubedata/

[2] C. Huang, J. Li, and K. W. Ross, "Can internet video-on-demand be profitable?" *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 4, pp. 133–144, 2007.

[3] M. J. Halvey and M. T. Keane, "Exploring social dynamics in online media sharing," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 1273–1274.

[4] P. Yu, M. Hu, and N. Kim, "Social network analysis," 2008.

[5] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos," in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*. IEEE, 2008, pp. 229–238.

[6] M. Wattenhofer, R. Wattenhofer, and Z. Zhu, Eds., *The YouTube Social Network*, 2012.

[7] M. Wattenhofer, R. Wattenhofer, and Z. Zhu, "The youtube social network," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[8] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of internet short video sharing: A youtube-based measurement study," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1184–1194, 2013.

[9] L. Cui, L. Sun, X. Fu, N. Lu, and G. Zhang, "Exploring a trust based recommendation approach for videos in online social network," *Journal of Signal Processing Systems*, pp. 1–13, 2016.

[10] S. Ghosh and S. Kumar, "Video popularity distribution and propagation in social networks," *Int. J. Emerg. Trends Technol. Comput. Sci.(IJETTCS)*, vol. 6, no. 1, pp. 001–005, 2017.

[11] T. Kardi. (2015) Pictorial introduction to graph theory. [Accessed 11-May-2017]. [Online]. Available: http://people.revoledu.com/kardi