Youtube Graph Network Model and Analysis

Yonghyun Ro, Han Lee, Dennis Won

Introduction

A countless number of contents gets posted on the YouTube everyday. YouTube keeps its competitiveness by maximizing click-through-rate with their cutting-edge video recommendation algorithms. This enables about 20 recommended videos for each video on YouTube. Each video can be thought as a node and two videos (or nodes) are connected by an edge if one recommends the other, which creates a large directed graph. This paper studies the characteristics of YouTube videos using diverse techniques such as PageRank, modularity, and statistical estimation. This allows to identify *influence sets* of YouTube videos, which have potential to have advance applications such as measuring effectiveness to add advertisements. By running these analysis algorithm, it is possible to observe a pattern for a popular videos for different categories.

Related Works

[1] *Xu Cheng, Cameron Dale, Jiangchuan Liu, "Statistics and Social Network of YouTube Videos." IEEE, 2008.*
gives a great overview of the dataset we will be using in our research. It describes the interesting facts about the dataset for each characteristic collected, such as video length, number of views, etc. It also goes briefly into the structure of the Youtube network, comparing it to a small world phenomenon. We will go further into the details and characteristics of the Youtube network in our analysis.

[2] *Jonah Berger, Katherine Milkman, "What Makes Online Content Viral?" Journal of Marketing Research, 2011.*
gives one way to consider the virality of the content using only the characteristics of the content. The algorithm it uses focuses on the emotional characteristics of the content, and calculates the score of each content to justify virality of the video. In our research, we will use some motivations of this paper to vary the weights and probability of infection.

[3] *John C. Paolillo, "Structure and Network in the Youtube Core." Hawaii ICSS, 2008.*
gives basic overview of the Youtube network as a social network, describing the degree distribution and the shape of the network defining some more characteristics using color, but it does not go into detail on how the structure of the network affects the virality and views of other videos. We will try and use some of the facts introduced in this paper to best simulate the virality.

Models and Methods

We use the dataset on the Youtube network crawled on Feb. 22nd, 2007 by Xu Cheng. There are some differences between modern day Youtube and the past Youtube in that there is no restraint on the length of the video anymore. As such, we will not consider the absolute measurement length of the video, and will even try to avoid using the fractional measurement length of the video.

As shown by [1], the growth rate of views falls almost completely to zero after some time, and since this timeframe is hard to define, we instead look only at the total number of views the video has received to determine the video's virality.

1 Pagerank vs. Video Feature Correlation Analysis

This part of the project analyzed the pagerank score of about 750,000 total Youtube videos by representing each video as a node with a pagerank score. The edges between nodes were formed to represent the relatedness of the two videos. That is, when video A is one of the 20 related videos of video B, then there exists a directed edge from B to A. Edges were not weighted. Using the pagerank score, this project evaluated the influence degree of each video. Thus, if a video has a high pagerank score, that means that video is related to many videos in the graph network, thus have a high influence within the network because it is pointed by many videos in the network.

In this pagerank analysis section, this project mainly examines the correlation between pagerank and multiple different statistical measures for a particular video. Specifically, we looked at the correlation of pagerank with the following six features:

- age (how long ago was the video posted on Youtube)
- number of views (how many people viewed the video)
- number of comments (how many people posted comments for the video)
- length of the video
- rate (the number of likes and dislikes for the video)
- ratings (the rating scored by viewers out of 5.0 rating system)

The purpose of such analysis is to examine what features contribute to what degree in determining the pagerank score, or the influence level, of the video with in the Youtube network. Using snap.py, the Youtube video graph was ran with the pagerank algorithm to compute the pagerank score of each video, and we analyzed the top 100 videos with highest pagerank score to analyze the characteristics of high-influence videos, and ultimately suggest or backtrack the Youtube's video recommendation algorithm parameters. After calculating the pagerank scores of 100 high-influence videos, those scores were graphed with different video feature stats above. The result correlation graphs can be found in the results section below.

2 Computing Influence Set

2.1 Motivation

Each video corresponds to only one of the thirteen categories. In contrast, viewers have multiple interests; a viewer watching a video from a comedy category sometimes will also click on videos from a music category and watch them, should they be present in a relevant video list. Since each viewer's actual preference is not given to us and cannot be estimated easily, we will instead attempt to figure out which videos the viewer can potentially watch through the graph structure. We will do this by computing an influence set for a subset of videos, specifically the ones with high number of views in each category, to see what other videos viewers get to through relevant video lists.

2.2 Modeling Assumptions

We will define a few assumptions here, and attempt to find few models that fit our assumptions the best.

i. Average clicks

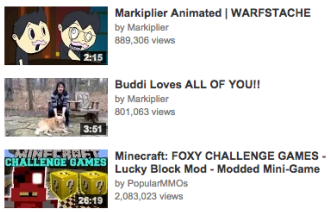
We will assume that on the first view, given a certain video list, a viewer is likely to click on a video from that list about 2 out of 3 times.

ii Clickthrough rate

In a single session, people will click through the relevant video list at most 3-4 times starting from a certain video.

iii Viewers' preference

Viewers have multiple interests; they are not interested in just one category. Also, on relevant videos section, the only information people can see of the video is the video length, uploader, title, and number of views - since only the number of views can be used for statistical analysis, we will only consider this as a feature that affects people's preferences.



2.3 Edge Probability Estimation Models

In all three following models, we run a random process similar to random walk and conclude that the videos are in the influence set of a certain video if they have appeared more than *N* times out of *I* iterations. Our experiments used *N = 25* and *I = 50.*

i. Bernoulli Uniform Edge Probability Model

In this model, we will assume that all edges follow the same bernoulli distribution. The probability of people clicking on one video from the relevant videos list two out of three times can be modelled by finding the edge probability that satisfies the following equation:

N = *number of videos in the relevant video list*

$$1 - P(\textit{no videos were clicked}) = 1 - (1 - P(\textit{A video was clicked}))^{N} = 0.66$$

There are about 10.5 outgoing edges per node on average. So the probability of a video for the edge being clicked is around 0.1. One other assumption that we satisfied for the model was the clickthrough rate; since there are about 10.5 outgoing edges, we expect to see between 10^3 and 10^4 videos at in an influence set.
We tried three different edge probabilities, namely $0.1,\ 0.105,\ 0.11,$ for the random process, and found that 0.105 satisfies our assumptions well.

ii. Independent Edge Probabilities Weighted by Number of Views

Now, we try to capture viewers' preferences by assigning different edge probabilities to each video based on its number of views. The algorithm to compute the edge probability is as follows:

M = *a maximum probability for a video*
A = *a probability for the videos with average number of views*
V = *number of views for a video*
V_{Max} = *maximum number of views among all videos*
V_{Avg} = *Average number of views among all videos*
V_{src} = *number of views for a source video*
V_{dst} = *number of views for a destination video*

$$P(\textit{video}) = a * log(V/V_{Max}) + b$$

$$a = \frac{(A-M)}{log(\frac{V_{Avg}}{V_{Max}})}$$

$$b = M$$

$$P(\textit{source video to destination video}) = (0.5 * (a * log(\frac{V_{src}}{V_{max}}) + b) + 0.5 * (a * log(\frac{V_{dst}}{V_{max}}) + b))$$

A and M are the parameters of this model. Each video is given a probability, and the final edge probability will be a weighted sum of the source video probability and destination video probability. Using the same assumption from before (people click on relevant video about 1 in 3 times), M should not be so high and A and M should not differ by much. This means that as the number of views gets higher, there should be a diminishing gain in terms of click probability. Therefore, we use the logarithmic regression to estimate the probabilities, and the resulting equation is simply a result of fitting a logarithmic regression using two (x,y) observations, $(1,\ M)$ and $(\frac{V_{Avg}}{V_{Max}}, A)$. One thing to note about this model is that the fraction of nodes that have lower view counts than the average number of views is 0.857. This means that most of the videos will have a noticeably lower probability than those above the average.

Once again, several combinations of (A, M) were tried, ranging from $(0.15,\ 0.1)$ to $(0.18,\ 0.11)$, and $(0.15,\ 0.11)$ was chosen to satisfy our clickthrough rate assumption.

iii. Dependent edge probabilities weighted by neighbors in the relevant video list

Now, we consider all the videos in the same relevant video list together to compute the influence set. This means that the viewers' preferences do not just depend on the video's view counts but also on the neighboring videos. The algorithm to compute the influence set in this case was the following:

1) This follows a BFS algorithm; dequeue a node n from the queue Q.
2) we generate a random number in $(0, 1)$; if it is > 0.33, we go to step 3). Otherwise, we stop and do not visit any videos from the considered node, and go back to step 1). This is to be consistent with our average number of clicks assumption.
3) Give each video in a list a weight, defined by: $Weight = log(V)^2$
4) Add all weights and get C
5) Generate a random number between $(0, C)$
6) Pick a video that corresponds to the random number. This choice is done by the weights defined in previous steps.
7) If the chosen video were already visited or were chosen during current iteration, we stop adding and go to step 1).
8) If the chosen video were not considered before, add it to Q, and go to step 5).

Step 2) ensures that we choose at least one edge from the relevant video list 2 out of 3 times. We give different probabilities to each video based on its view counts in step 3), but whether or not they get chosen is an independent process, just as we have done in previous models.

2.4 Brief Analysis of the Models

The bernoulli uniform edge probability model should serve as a good baseline for our analysis, but is not the model that perfectly fits our assumptions; it does not use view counts at all in formulating edge probabilities. In contrast, our independent edge model takes advantage of the view count, but does not capture our first assumption well (2 out of 3 views on average). Finally, our dependent edge model takes all assumptions into consideration, so it is the most complete model. Our standard model therefore is the dependent edge model, and our analysis will focus mainly on the results from this model.

Results and Findings

1 Influence Set Results

1.1 Confirming Influence Set Sizes and Analysis

To fit assumptions set by our models, we should have gotten influence sets with size 10^4 at maximum. Our results satisfy this assumption.

	Uniform	Independent	Dependent
People & Blogs	3668	4992	3133
Music	5845	5211	7010
Howto & DIY	1953	2095	1138
Gadgets & Games	3701	3650	4206
News & Politics	34	3065	3804
Autos & Vehicles	601	4147	3065

Table. Category (Subset) and Maximum Size of an Influence Set in the Corresponding Category

The most notable difference between the uniform model and the rest is the increase in the size of the influence set for the categories "News & Politics" and "Autos & Vehicles." All three models conclude that the same video has the biggest influence set in "News & Politics" category, but the discrepancy is very high. Looking at this video, we see that it has 20 edges, and 15 of those videos have more than 30,000 views; this means that the edge probabilities will generally be higher in the first clickthrough. The same is true for subsequent videos as well, as exhibited by the following video's relevant video list:

Category	# views	# edges
Howto & DIY	41123	19
Sports	61824	19 ...

Each video picking is an independent process but total probability required to get to the picking point gets exponentially smaller, so we are able to stop branching out around the fourth order.

1.2 Category Maximizing Videos Given their Influence Sets

Figuring out which videos maximize the viewers in a certain category is helpful for many applications such as targeted advertising. For example, an auto company might be tempted to put an ad on a video with 40 million views from a "Comedy" category, but putting an ad on a video with 800K views from an "Auto & Vehicles" category might turn out to be more beneficial. Measuring this effect through influence sets can be helpful for cases such as viewer-targeted advertising (guess the viewer's preferences from the viewing patterns), and predicting what type of advertisable categories, such as "Auto & Vehicles," the viewers of a video in a non-advertisable category such as "People & Blogs" would be interested in. Is it better for the advertisers to chase the highest number of viewers, or is there a more optimal strategy? We try to find evidence of this phenomenon through the influence sets we obtained by running our models. We first estimate by just the number of videos. For the second way, we first add up the number of viewers for all videos in the influence set, compute the categorical ratio, which is the ratio of number of viewers who watched a video from a certain category to the total number of viewers, then multiply the resulting ratio by the original video's viewer counts. This is essentially to estimate the number of viewers interested in a certain category among the viewers of the original video.

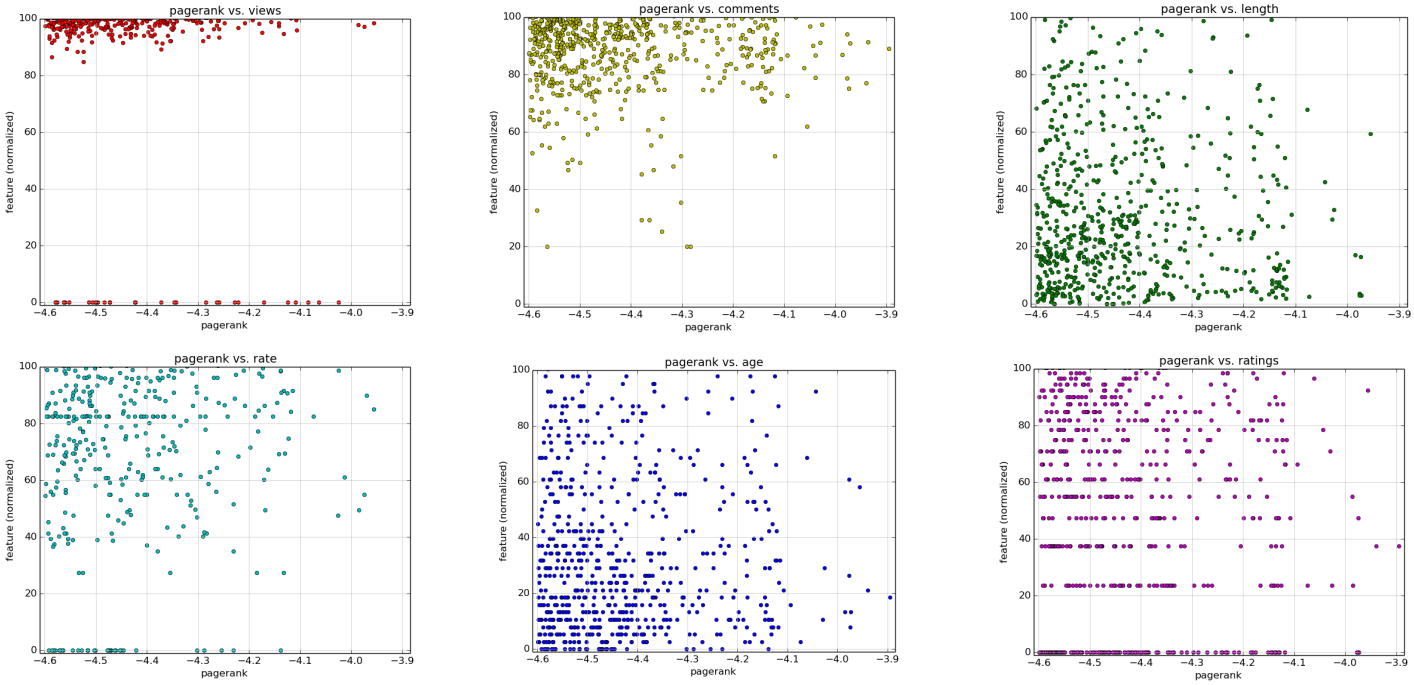
Considered Category	**Uniform**		**Dependent**		**independent**	
	C1	**C2**	**C1**	**C2**	**C1**	**C2**
People & Blogs	Music	Comedy	Music	People & Blogs	People & Blogs	People & Blogs
Entertainment	Music	Entertainment	Music	Comedy	Music	Entertainment
Howto & DIY	Comedy	News & Politics	Music	News & Politics	Music	News & Politics
Gadgets & Games	Music	Comedy	Music	Comedy	Sports	Comedy
News & Politics	Music	News & Politics	Music	Music	Sports	Comedy
Sports	Music	Sports	Music	Entertainment	Sports	Comedy
Pets & Animals	Music	Pets & Animals	Sports	Pets & Animals	Sports	Pets & Animals
Music	Comedy	Comedy	Music	Comedy	Music	Comedy
Travel & Places	Comedy	Comedy	Music	Travel & Places	Sports	Music
Autos & Vehicles	Comedy	Autos & Vehicles	Music	Autos & Vehicles	Travel & Places	Auto & Vehicles
Film & Animation	Comedy	Film & Animation	Music	Film & Animation	Comedy	Film & Animation

Table. C1 = Maximizing Video's category (By count of videos in the considered category), C2 = Maximizing Video's category (By categorical ratio * numViewers). Cells colored in orange are in agreement throughout three models

It is interesting to node that the category-maximizing video computed using the first method tends to belong to either "Comedy" or "Music" category. This because, as evidenced by previous analysis using network structure, videos with a high number of views tend to have stronger relationships with more videos than videos with a lower number of views. Since the number of connections is much larger for videos with a high number of views, videos in a "Comedy" category or a "Music" category will dominate other videos in this regard, so the first method is hard to trust. We instead look at the second method. The second method computed that targeting a video in its own category is the most optimal for "Pets & Animals," "Autos & Vehicles," and "Film & Animation" for all three models. This is an interesting result; it means that the guaranteed number of

people interested in a certain category is lower for a video with a high number of views from a different category than for a video with a relatively low number of views from the considered category. In contrast, we see that for “Gadgets & Games,” a video from a “Comedy” category has a higher guaranteed number of viewers who are interested in “Gadgets & Games” than a video that has the highest number of viewers among a “Gadgets & Games” category. This is probably due to the fact that two categories are highly correlated; gaming videos tend to be along the borderline of “Comedy” so viewers for the two categories will mostly overlap. Their correlation is shown by the correlation table in part 3 of this section.

2 Pagerank vs. Video Feature Correlation Analysis



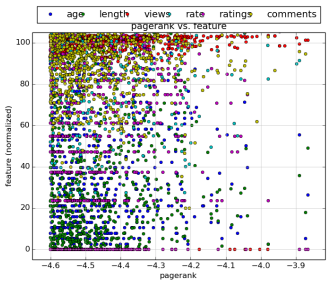
The graphs above have x-axis as the pagerank score in log10 base, and the y-axis as the normalized percentile of each video feature stat for all 6 different video statistics (views, comments, length, rate, age, ratings). All 6 graphs have 1000 data points which correspond to the 1000 Youtube videos with the highest value of pagerank score.

Note that from pagerank vs. views scatter plot, almost all 1000 videos have high percentile that is very close to 100 percentile, meaning that there is a very high correlation between the pagerank score and the number of views the videos have. Similarly, from pagerank vs. comments scatter plot, almost all 1000 videos have high percentile that is very close to 100 percentile, meaning that there is a very high correlation between the pagerank score and the number of comments the videos have.
On the other hand, however, the length of video has not only a relatively low correlation with pagerank, but also almost an inverse correlation. This signifies that most of the high-influence videos with high pagerank scores are usually at a short length compared to other videos in the network. This explains why the data points on the pagerank vs. length scatter plot has most of its points at the low percentile range. This might be because many of the “hot” and popular videos on Youtube with lots of views and comments tend to be short, suggesting that shorter videos (like short music videos) have more tendency to be viral and popular than long videos (like full movies). This is similar to the pagerank vs. age correlation, which also has most of its data points clustered at the low percentile range, although with lower correlation than the length of the video. This suggests that the Youtube recommendation algorithm tries to put more emphasis on the newer and trendier videos than older videos so that the new videos have better chance of being viral and popular.

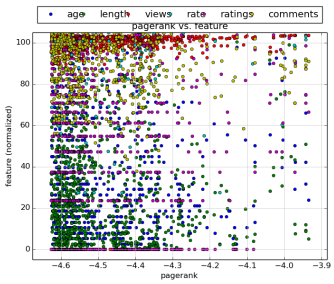
Lastly, from the pagerank vs. rate scatter plot, we can induce that videos with high rate value (the number of likes and dislikes) are tend to be more influential in the Youtube network than the ones with low rate values. This might be closely related to the number of views and the number of comments because high number of views and comments are likely to

induce high number of rates from users. The fact that the users not only view but also either comment or rate the video suggests that the video is popular and engaging enough to make users to conduct online activities for that video, which might be why the videos with high rate values tend to have high pagerank score. However, what was interesting was the scatter plot for pagerank vs. ratings, the video score given by the viewers by the 5.0 rating system. The scatter plot for pagerank vs. ratings has a very low correlation. It is a bit geared toward the top percentile, weakly suggesting that videos with high rating values tend to have high pagerank values, but still the data points on the graph is very sparsely distributed over all percentile range. This hints that the rating given by the viewers do not have a strong signal in representing how popular, engaging and influential the video is within the Youtube network.

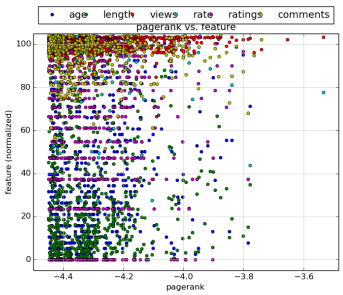
Comedy



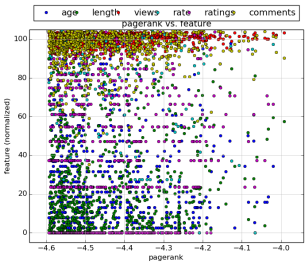
Entertainment



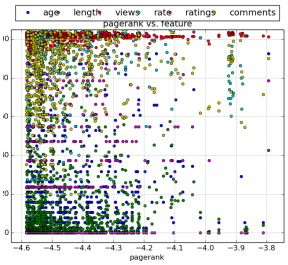
Film & Animation



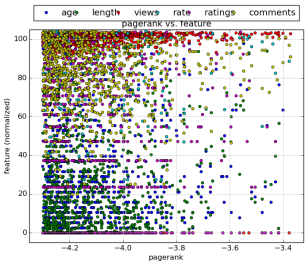
Music



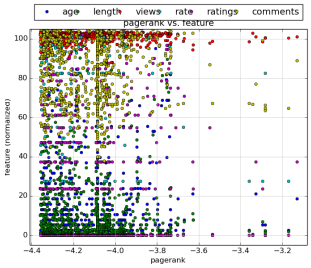
People & Blogs



Sports



Travel



This project also performed the same pagerank vs. video feature correlation analysis on each subgraph extracted out from the entire Youtube graph network by each different category for the 7 main Video categories defined at Youtube. The above 7 scatter plots represent each category pagerank vs. video features distribution. Note that to save space, we overlapped each different video feature data points on a same scatter plot graph with labels provided to indicate which color data points correspond to which video feature data points. The correlation analysis results were not too different compared to the analysis done for the entire Youtube network. They all have strong positive correlation with the video features of the number of views, comments, and the rate value while have strong negative or inverse correlation with the features of the age and length of the videos. Lastly, the ratings value again seemed quite weakly correlated to the pagerank scores of videos.

3 Category Analysis

Each YouTube video belongs to one of 12 categories as listed in below table. Each video recommends 20 videos. Intuitively, these 20 videos are likely to be in the same category but can sometimes contain videos from other categories if they are related in some sense. In this part of the paper, we examine the graph analysis by looking at category interdependency and category intradependency. The former will observe at how categories are related to one another and the latter will study the special features of each set of videos of same category as a perspective of graph.

3.1 Category Interdependency

Below table represents the interdependency between all possible category pairs. For example, of all recommended videos from all videos in Autos category, 52.22% are in autos category and 5.46% are in comedy category. This means that about 10.444 videos out of 20 recommended videos of Autos video are expected to be Autos and 1.092 videos are expected to be Comedy.

Each entries are highlighted with darker red color if values are high. Diagonal entries have the most red color since each category tend to recommend videos in the same category. This aligns with our expectation. Also, it is interesting to see that Sports and Games category tend to recommend videos within the same category with the expected probability of about 70%. This can be explained by the activities of viewers in these category tend to only watch videos within the same category. However, videos in People category tend to leave the category because only about 29.79% of recommended videos from People category are recommending videos in People category. Other categories that have relatively high correlation with People category are Entertainment and Music, which is not surprising.

<Category Interdependency Table>

from \ to	Autos	Comed	Entertain	Film	Games	DIY	Music	News	People	Pets	Sport	Travel
Autos	52.22%	5.47%	9.16%	2.25%	3.52%	4.59%	6.21%	2.19%	2.59%	0.39%	8.74%	2.66%
Comedy	0.94%	41.18%	17.30%	6.01%	3.92%	1.87%	13.51%	3.49%	5.38%	1.13%	4.24%	1.03%
Entertain	0.94%	10.91%	45.33%	8.57%	3.75%	1.83%	15.78%	2.55%	4.59%	0.54%	4.08%	1.12%
Film	0.40%	6.63%	14.69%	53.72%	5.24%	1.18%	12.11%	1.36%	2.21%	0.31%	1.52%	0.62%
Games	0.70%	4.36%	6.57%	5.42%	69.47%	1.02%	5.83%	0.67%	1.04%	0.22%	4.36%	0.34%
DIY	4.15%	6.92%	11.11%	4.15%	3.36%	42.52%	8.53%	8.51%	5.70%	0.85%	2.37%	1.84%
Music	0.35%	5.29%	11.19%	5.17%	2.40%	0.92%	67.84%	1.41%	2.78%	0.30%	1.70%	0.65%
News	0.66%	7.75%	8.71%	2.66%	1.44%	4.27%	8.23%	50.96%	10.70%	0.66%	2.12%	1.82%
People	0.85%	11.55%	15.45%	4.10%	2.19%	2.96%	15.24%	9.88%	29.79%	0.93%	4.79%	2.28%
Pets	0.97%	12.57%	10.45%	4.18%	3.14%	2.19%	10.79%	2.68%	4.24%	41.98%	3.65%	3.16%
Sports	1.45%	4.55%	7.30%	1.44%	2.45%	0.81%	5.25%	1.02%	2.62%	0.32%	71.84%	0.94%
Travel	2.82%	6.62%	11.67%	3.87%	2.02%	3.40%	12.87%	5.73%	7.59%	2.53%	6.36%	34.52%

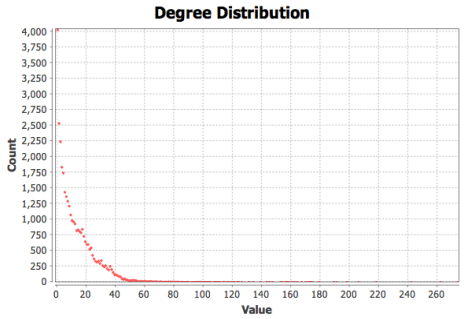
3. 2 Category Intradependency

Previous part of the paper explained how videos in different categories correlate. This part will focus on the videos in each category as known as intradependency. In order to do so, we took one category at once and removed all other videos belonging to a category and calculated average degree, average clustering coefficient, modularity, and number of communities.

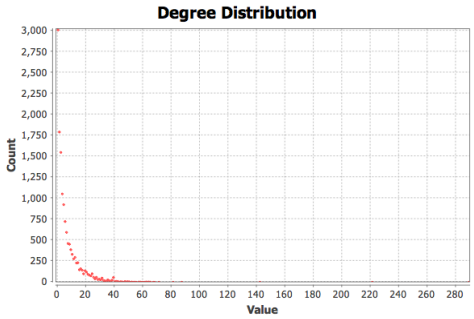
The result, as shown below, aligns with the previous discussion. Sports and Games have highest average degrees at around 7.5, which is expected from the category interdependency. Also, People, Pets and Travel have the least average degrees as they tend to have less intra-correlation than other categories.

	Average Degree	Average Clustering Coefficient	Modularity	# of Communities
Autos	4.457	0.270	0.892	1,943
Comedy	4.533	0.248	0.861	10,046
Entertainment	4.732	0.269	0.924	15,020
Film	5.927	0.226	0.858	7,942
Games	7.455	0.260	0.907	2,975

DIY	4.712	0.288	0.907	3,320
Music	7.074	0.318	0.886	8,067
News	6.093	0.289	0.848	4,881
People	3.133	0.247	0.948	13,068
Pets	3.886	0.244	0.886	1,931
Sports	7.582	0.267	0.857	3,756
Travel	3.132	0.266	0.940	3,746

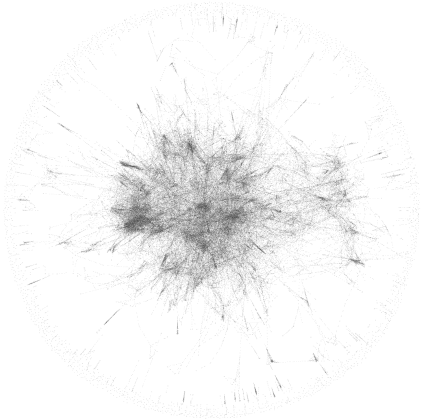


<Travel>



<News>

Above two plots of degree distributions show the difference between Travel Category and News Category. Other degree distribution data for each category shows similar result. The reason data goes beyond 20, which is the total number of recommended video per video is some popular videos are recommended from many other videos, causing the degree to be greater than 20. All category graphs follow preferential attachment with different alpha values. For example, Travel videos showed alpha at 5.8, which compares to the alpha for News at 2.9. This analysis is also comparable with above result where 50% of News video recommends other News videos versus Travel recommending only 35%. The preferential attachment can be visualized using Gephi, a graph visualization software.



<Visualization of Autos & Vehicle Graph>

Above is the visualization of Autos & Vehicles graph data, small dots that forms vague circular shape around are lone videos that has no edge to other Autos & Vehicles videos. Darker region in the center of the plot are the ones with higher degree distribution. As we can observe, this visualization intuitively aligns with our expectation of preferential attachment.

Conclusion

Influence set calculation models gave results that were consistent with our assumed viewer behaviors and should lay good foundations for future studies in influence sets in YouTube data. Our categorical analysis on the returned influence sets returned interesting results but will need further proofs and better estimation methods of viewer preferences. Category analysis showed the correlation between videos in each category. In addition, category intra-dependency analysis showed the details of each subgraph.

The pagerank score correlation analysis against each video feature (age, length, views, comments, rate, and rating) suggests the common characteristics of the highly influential videos on Youtube graph network, when the influence of the video on the network is represented by the pagerank scores of the videos. We examined that high-influence videos share the common traits of having high number of views, high number of comments, and high rate values, while having short length, young age (more recent). The ratings values did not seem to have much correlation with the pagerank scores of the videos. This result hints the Youtube's video recommendation algorithm parameters for each video features examined here. For practical application of such result, online video advertisement companies could target videos with high number of views, comments, and rate values that is quite short and recently posted on the website in order to maximize the advertisement effect.

Contribution

Our contributions to this project are equal.