A first step towards understanding popularity in YouTube

Gloria Chatzopoulou, Cheng Sheng, Michalis Faloutsos
Department of Computer Science & Engineering
University of California, Riverside
Riverside, California 92521
Email: {chatzopd,shengc,michalis}@cs.ucr.edu

***Abstract*—Being popular in YouTube is becoming a fundamental way of promoting one's self, services or products. In this paper, we conduct an in depth study of fundamental properties of video popularity in YouTube. We collect and study arguably the largest dataset of YouTube videos, roughly 37 million, accounting for 25% of all YouTube videos. We analyze popularity in a comprehensive fashion by looking at properties and patterns in time and considering various popularity metrics. We further study the relationship of the popularity metrics and we find that four of them are highly correlated (viewcount, #comments, #ratings, #favorites) while the fifth one, the average rating, exhibits very little correlation with the other metrics. We also find a "magic number" in the average behavior of videos: for every 400 times a video is viewed, we have one of each of the following user actions: leaving a comment, rating the video and adding to one's favorite set.**

I. Introduction

Becoming popular in YouTube is essential for marketing services and products. For example, there are cases of artists who became "Internet phenomena" via YouTube, thus jump-starting in their careers. Naturally, people have developed ways to boost their visibility in YouTube by increasing the popularity of their videos. Taking this one step further, there are software and services that promise to boost one's video popularity for a fee. At the same time, YouTube is making efforts to address these artificial means of gaming the system. The above prompted us to study the popularity of videos at YouTube.

The overarching goal of our work is to understand fundamental properties of video popularity in YouTube. In fact, defining popularity itself is not as straightforward as one may think. Different aspects of popularity are captured by various "popularity metrics", which we will introduce shortly. We believe that an in depth study of popularity is necessary to understand the relationship and temporal patterns of all these metrics.

We provide a quick overview of the terminology we use in this paper. We use *viewcount* (the number of times a video is watched) as the fundamental parameter of popularity and study the its relationships with other popularity metrics: *number of comments*, *number of favorites*, *number of ratings*, and *average rating*

In fact, these metrics capture the reaction of users to a video, since they go beyond simply watching a video, by representing an action the user takes in response to liking or disliking the video or feeling a need to comment and judge. In the rest of this paper, we will use # to denote "number of", as in #comments. We use the term *categories* as defined by YouTube, that is, each video is assigned to one such category by the author. Besides, YouTube provides *standard feeds*, which are lists of top videos along two dimensions: (a) metric of interest, and (b) interval of interest (today, week, month, all time). For example, there is a feed for "Most viewed" videos of "Today", and a feed for "Most Discussed" of "Today" meaning the videos with highest #comments for that day. The last feature that we explore is the related videos. YouTube using proprietary algorithms provides a list of "related videos" for every video a user watches. We create the related video graph (RVG) which is a directed graph where nodes are videos and an edge e(u,v) represents that video v appears in the related video list of video u. Given this graph we analyze relationships between videos.

Several measurements studies of YouTube have analyzed different statistical and behavioral properties, but none have studied popularity as exhaustively as we do here. Recent works, [1]–[8] study distribution and temporal patterns of viewcount. To the best of our knowledge, our work is the first to study other popularity metrics together with viewcount. [1], [7], [8] propose several solutions to the problem of video spam, while [2], [6], [9], [10] analyze several social networks within YouTube.

In this paper, we conduct an in depth study of the fundamental properties of video popularity in YouTube. We collect arguably the largest dataset of YouTube videos, containing more than 37 millions of video metadata. Our data accounts for roughly 25% of all YouTube videos. We study video popularity in three different aspects: (a) in time, (b) by using multiple popularity metrics, and (c) across different categories of videos as provided and labeled by YouTube.

Our contributions can be summarized in the following key points:

- Four of the popularity metrics (viewcounts, #comments, #ratings, #favorites) are highly correlated. By contrast, the average rating exhibits very little correlation with the other metrics. We develop a simple linear regression model to estimate viewcounts as a function of the other metrics which captures 76.8% of the variability of viewcount, as we discuss later. A potential use of such a model could be to identify candidate videos of abnormal activity (artificial boosting of a video's metrics).

- A user actively "reacts" to a video 1 in every 400 times on average: a video receives a comment, a rating, and is added to someone's favorite list once for every 400 times it is viewed. In a counter-intuitive way, the user response ratio (i.e. number of comments per thousand views) decreases as a function of the number of viewcounts. In other words, as a video is seen by an increasing number of people, it elicits less acute reaction.
- A video does not stay for more than two days in the top 100 of the "Today" standard feeds for all the feeds that we examined (most viewed, most popular, most responded, top rated, most discussed, and top favorites). This indicates that standard feeds are very competitive, and that a video has a small window of opportunity to climb to prominence through those feeds, which is reportedly a common practice. Users exhibit similar daily patterns of accessing videos but potentially different weekly patterns. The daily periodicity of the user behavior (video upload and comment posting) is the same for all types of videos, and the peak time is 1 PM Pacific Time.
- The related video graph exhibits that the related video relationship is reciprocal for 36% of the video pairs. The top-5 videos with highest viewcount are highly connected in the related video graph in other words one video appears in the related video list of the others.

The remainder of the paper is organized as follows. In Section II, we present background information and our crawling method as well as the collected results. Section III discusses the correlations among the metrics of popularity. In Section IV, we report some temporal properties of the popularity metrics and study the standard feeds and their effects in videos' popularity metrics. We construct and analyze the related video graph in Section V. In Section VI we briefly discuss related work and finally, Section VII concludes our work.

II. Background and Data Collection

A. Background

We give a brief introduction of YouTube and review several important features on it. After watching a video, users can give feedback in several ways. They can post one or more text comments to a video, rate the video on a five star scale, or add it to their set of favorite videos. A user is allowed at most once to rate a video and to add a video to his/her favorite set. These feedbacks are measured by several popularity metrics in YouTube, such as viewcount, #comments, #favorites, #rating and avg_rating. YouTube considers viewcount as the fundamental parameter of popularity, and is very careful to not count viewcount multiple times for the same IP in a short period. Users' personal preference are denoted by #favorites, showing whether a video is preferred and will be watched again. A video owner has the freedom to block comments and/or ratings at will. We acknowledge that this capability might distort slightly our results but we believe that the percentage of video owners which impose these restrictions is very small. While regular users can view these metrics from the video web page, the YouTube API [11] provides researchers efficient access through a programming interface.

TABLE I
Popularity Metrics: Basic Statistics of our Crawled Dataset

Metric	Min	Max	Avg	Description
viewcounts	0	114m	11k	Number of views
#comments	0	598k	18.98	Number of comments received
#favorites	0	669k	30.31	Number of times added to other users' favorite lists
#ratings	0	488k	21.95	Number of ratings received
avg_rating	0	5	3.51	Average rating, range from 1 to 5

B. Data Collection

To collect data, we refrained from directly crawling the YouTube site and we used instead the YouTube Data APIs. We populate a local database with their meta-data information collected using the APIs. Similarly, we query the meta-data of users, comments, related videos, subscriptions and many other perspectives using APIs and populate separate tables. The YouTube API limits the related video list to 100 for every video. Once we collect the video information from the standard feeds, we query the related videos of each video in the standard feeds. We recursively go and populate our database with the information of videos in the related list. We also populate metadata for comments, authors and other perspectives from videos information.

Our crawling system contains a data server and several crawlers. Crawlers communicate with YouTube, and send the collected data back to data server. Several crawlers can work in parallel to achieve better scalability and reliability.

By running the crawling system between February to June 2009, we have collected the most exhaustive dataset (as far as we know) about the YouTube site, which contains 20 tables regarding information about various perspectives. The number of records for metadata of videos, users and comments are 37.9 million, 1.4 million and 41.1 million respectively.

C. Metrics

In this study, we focus on five metrics provided by YouTube, viewcount, number of comments, number of favorites, number of ratings and average rating. Although all these metrics reflect the degree of popularity for each video, viewcount is widely regarded as the basic video popularity metric. Table I shows the minimum, maximum and average value of these metrics. For each video, we also record its category. At the time of crawling, YouTube had 29 categories, a combination of 15 regular categories and 14 HD video categories. *Music* and *Entertainment* are the largest two categories, accounting for 45% of videos in our dataset.

III. Relationships and Correlations of Metrics

Using our dataset, we explore the correlations between different popularity metrics. In this section, we first analyze the pairwise correlations among the five popularity metrics. Then,

TABLE II
CORRELATIONS - VIEWCOUNT AND EVERY OTHER POPULARITY METRIC

Variable	Pearson Correlation Coefficient with Viewcount
#comments	0.609
#favorites	0.821
#ratings	0.756
#avg_rating	0.045

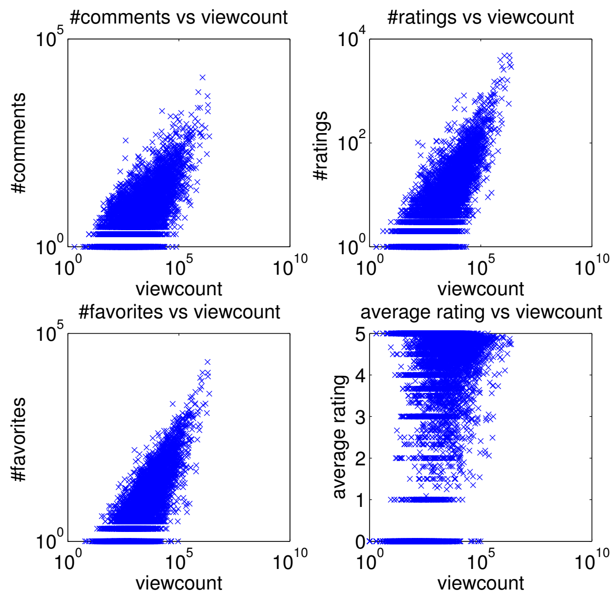

Fig. 1. Visualization of Correlation of viewcount and the other popularity metrics. All pairs exhibit strong correlation apart from viewcount vs average rating.

we present our linear regression model, which can potentially be used to detect video spams. Finally, we study ratios between viewcount and the rest of the metrics.

A. Pairwise Correlations

Strong correlations are observed among most metrics. Viewcount is the index of popularity of a video. We calculate the Pearson Correlation Coefficient (PCC) between viewcount and each of the indicator variables. PCC is the most common metric to measure the dependence between two quantities. A summary of the PCC metric is included in Table II. Viewcount is highly correlated with #comments, #favorites and #ratings. The average rating is the only variable which spans between zero and five. We believe that this peculiarity might be a reason for the low correlation with viewcount. To visualize these correlations, we randomly select 10,000 videos, and plot the values of the popularity metrics in pairs, as shown in Figure 1. We can observe obvious linear trend for correlations between viewcount, #comments, #favorites, and #ratings, while the correlation between viewcount and avg_rating is unclear.

Correlations increase with popularity. It is natural to ask whether the metrics of more popular videos become more correlated among each other. To address this question, we group our videos into four popularity groups in terms of viewcount, and randomly select 100,000 videos from each group. As shown in Table III we calculate the PCC between viewcount and the other four metrics for each group. We observe that pairwise correlations (except those with avg_rating) increase with video popularity. This result can be supported by empirically examining several popular videos. It is not surprising to see a popular video with a large number of comments, ratings and favorites at the same time. This is also confirmed by our study on standard feeds, as we will show in Section IV.

TABLE III
CORRELATION STRENGTH OF VIDEOS GROUPED BY VIEWCOUNT

	Viewcount Ranges			
Popularity metrics	$\leq 1K$	$1K - 10K$	$10K - 100K$	$\geq 100K$
#comments	0.234	0.195	0.236	0.560
#favorites	0.327	0.377	0.452	0.771
#ratings	0.265	0.239	0.324	0.779
avg_rating	0.324	0.092	0.028	0.002

B. Estimation of Viewcount

A widely used statistical tool to model the functional relationship between a dependent variable Y and a set of independent variables X is linear regression. We employ linear regression to build an estimation model of viewcount given a set of indicators, the other popularity metrics. We use R-square [12] as a measure of our model's performance[1]. As a first step, we apply the forward selection stepwise screening method in order to identify the influential variables that should be included in the final model. This method selected the following three variables: #favorites, #comments, #ratings.

The Complete Model. Following the standard methodology, we initially build the full second order model using these three variables. The model contains 9 terms, the first order terms, the second order terms and their pairwise interactions (e.g. the pairwise products). The R-square of the model is 0.7855 which is high enough to trust our model.

The Simplified Model. Going one step further, we reduce the complete model to a simplified model that contains only four terms. We tested and verified the hypothesis that the two models are statistically equivalent. The mathematical equation of the simplified model is given by Equation 1.

$$\begin{aligned} \widehat{viewcount} &= 3201.588F^2 - 0.014R^2 \\ &\quad +14.059F + 391.011R + 5220.755 \quad (1) \end{aligned}$$

where F=#favorites and R=#ratings

The simplified model has an R^2 = 0.768 which is a small change compared to the R^2 of the complete model. All predictors have a significant contribution to the model. We used the F-test to compare the R-square values of the two models. Using this test as hypothesized, the simplified model performs as well as the the complete model.

As future work, we intend to use this model to identify large deviations of the viewcount of videos. These videos can be

[1] R^2 coefficient of determination is the proportion of variability in a dataset that is accounted for by a statistical model.

considered as candidates for further investigation to determine if their popularity has been artificially boosted.

C. Analyzing the ratios between popularity metrics

In this section, we want to quantify the "active response" of users to a video, as explained earlier. Specifically, we define the following metrics:

a) **comment_ratio:** defined as #comments/1K viewcounts. It represents the desire of the users to respond to the video by leaving a comment.
b) **favorite_ratio:** defined as #favorites/1K viewcounts. It captures the desire of the user to become a "fan" of the video.
c) **rating_ratio:** defined as #ratings/1K viewcounts. It represents the desire of the user to evaluate the video.

Since the value of viewcount is usually much larger than other metrics, we count actions per thousand viewcounts.

A video receives one comment, one rating, and is added to someone's favorite list once for every 400 times it is viewed. We first examine the average value of these three ratios. Surprisingly, they are all close to 2.5, which means that a video tends to receive all three user reactions/responses (a comment, a rating and being added to a user's favorite list) each time it is viewed 400 times. This suggests two interesting things. First, responding to a video is an indication of a strong reaction: responding takes more effort than simply watching it. In addition, responding requires the user to login, while watching a video does not require loging in. Second, the probability of a user making a comment, giving a rating and adding a favorite item are equally likely. Note that we don't know if the three actions are taken by the same user, due to the way we collected the data, but it could be an interesting future direction.

The ratios decrease as a function of viewcount. We also study the correlation between the ratios and the absolute number of viewcounts. Intuitively, we tend to believe that more popular videos will be commented, rated, or added into favorite lists more frequently. To verify this, we divide videos into four groups, according to their viewcount. We show the comment_ratio of each group in Figure 2. Surprisingly, we discover that the reaction strength is stronger among less popular (in terms of viewcount) videos. This may suggest that the first viewers of a video are more likely to have active reactions to it. We only have a histogram for comment_ratio here, since the corresponding values for favorite_ratio and rating_ratio are similar.

IV. Temporal Patterns

In this section, we analyze temporal properties of video popularity. We first analyze short term trends. Specifically, we show how popularity metrics evolve daily and weekly. Then, we observe the long time trend to understand how popularity metrics increase or decrease over months or even years. Finally, we focus on very popular videos within a daily timeframe.

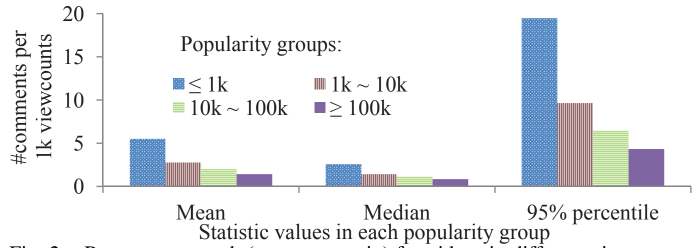

Fig. 2. Response strength (comment_ratio) for videos in different viewcount groups

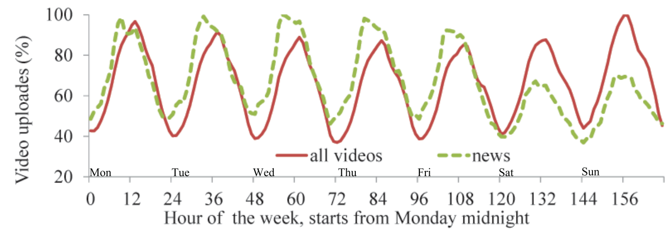

Fig. 3. Hourly video upload rates in a week

A. Daily and Weekly Patterns

To study daily and weekly popularity patterns, we analyze the video upload rate and comment rate on each hour of the week. Due to space limitations, we only present video upload rate, since the comment rate has very similar behavior. We start by grouping the videos into 24 x 7 bins according to the hour of week they were published. E.g. a video published on 3 AM, Monday belongs to bin 3. Then, we normalize the number of uploads in each hour by the maximum value across all the bins. Figure 4 shows the normalized hourly upload rate with a solid line, revealing two patterns, daily and weekly.

Daily peak occurs at 1 PM, and weekly peak is on Sunday for most video categories. On each day, the upload rate reaches the peak at 1 PM and the valley at midnight. Moreover, Sunday is the most active day in a week, followed by Monday. We also delve further into subsets of videos across different categories, and see that these two patterns also apply to many categories, such as *Comedy*, *Film*, and *Music*. Note that most of these categories are entertainment oriented, therefore people visit them more often during their spare time. This is an explanation of the weekly peak on Sunday and daily peak during lunch break. However, several categories exhibit a slightly different weekly pattern, such as *Education*, *Howto*, *News* and *Nonprofit*. We represent the hourly upload rates for videos in *News* by the dashed line in the same figure. We observe a decrease on weekend activity compared to weekdays. We should note that the reported times are based on the Pacific Time Zone. Given that YouTube is globally accessible it is hard to draw conclusions about time and users' location.

B. Long Term Trends

We study the long term popularity trends by analyzing the relationship between video age and each of the popularity metrics. We bin our videos by their age (in days), and then calculate the average value of each of the five metrics. The result is shown in Figure 4. Note that we exclude videos

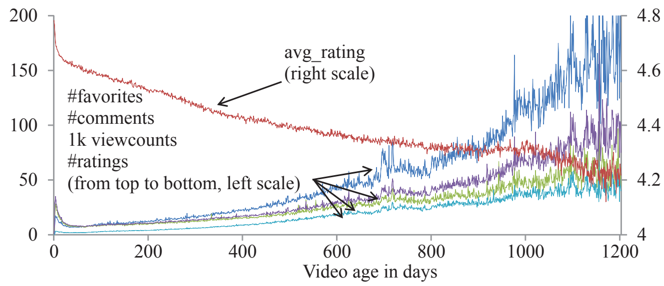

Fig. 4. Popularity metrics vs video age

TABLE IV
THE AVERAGE PERCENTAGE OF VIDEOS THAT APPEAR TWO CONSECUTIVE DAYS IN A STANDARD FEED

Standard Feed	Day by Day similarity
most_viewed	17.2%
most_popular	23.3%
most_responded	54.6%
top_rated	33.1%
most_discussed	30.6%
top_favorites	28%

which are older than 1200 days, because during that early stage there were few videos uploaded per day and thus create meaningless statistics. We observe that all metrics except avg_rating increases with the age of the videos. Interestingly, avg_rating decreases with age, from 4.7 to 4.2, which requires further investigation to lead to an explanation.

C. Standard feeds

The standard feeds contain videos that have attracted users' attention during a particular timeframe. Users react in the view of a video by commenting, responding, forwarding to their friends etc. YouTube logs multiple standard feeds that correspond to different reactions of the users. We examine the "Today" (daily) standard feeds. More precisely these feeds are: most viewed, most popular, most responded, top rated, most discussed, top favorites. We examined the daily standard feeds of 12 consecutive days (10/20/2009-11/1/2009). Every feed contains 100 videos. First, we compute the overlap of the various feeds for every single day. A video that appears in the top rated standard feed with probability 50% will appear in the top favorites standard feed and with probability 62.67% will appear in the most discussed standard feed. In other words, half of the people that give full credit to the video will also add it to their favorite set and post at least a comment about it.

We also analyze how the standard feeds evolve day by day. We define self similarity of one standard feed between two days as the percentage of videos that appear in the feeds of both days. Table IV lists self similarity of standard feeds between two consecutive days. The self similarity of standard feeds during two consecutive days are very consistent for the twelve days that we monitored them. Every video in the standard feed has a rank, the position it appears in the standard feed. The videos that appear in the standard feed for two consecutive days appear between 18 to 33 positions higher the second day in other words these videos gradually attract more attention from the users. The videos in the most popular and most responded feeds deviate from the above behavior since the ranking of these videos seems to drop by 2-4 positions between two consecutive days. The most responded standard feed is the most stable among the rest since more than 50% of the videos remain a position in the feed within the 2 day timeframe and the rank of these videos just falls 4 slots behind. Our explanation for the above behavior is that one video can only be a video response to at most one other video which means that it needs more effort from the users in order for a video to get a long list of video responses. The videos will almost never appear in a standard feed for three consecutive days. The videos in the most_responded standard feed with probability 33% appear in the feed for three consecutive days. Given this finding, we conclude that the users who are aiming to make their videos popular should focus on launching videos that can potentially receive a lot of video responses.

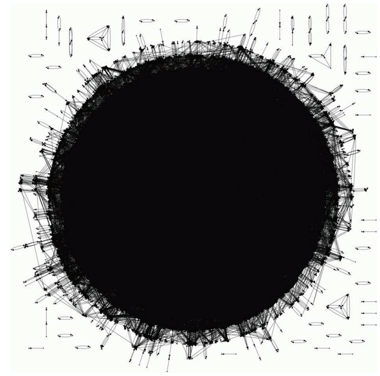

Fig. 5. The Related-Video Graph

TABLE V
RELATED-VIDEO GRAPH STATISTICAL PROPERTIES

Vertex Count	Edge Count	max IN degree	avg. IN degree
6062	40255	142	6.64

V. RELATED VIDEOS NETWORK

In this section, we analyze the related video graph (RVG). Our goal here is to visualize and model the "related videos" relationship of the most popular videos. Clearly, videos that appear as "related videos" to multiple other videos get an advantage of being viewed more.

RVG is a directed graph where nodes are videos and an edge e(u,v) represents that video v appears in the related video list of video u. Since we are interested in the most popular videos, we created the graph using videos with viewcount greater than 1.5M. Figure V shows the visualization of the RVG. Table V shows some basic statistics of the graph. Interestingly, the Giant Weakly Connected Component of the graph contains 98.33% of the nodes.

The "related-video" relationship is reciprocal for 36% of video pairs. We check the hypothesis that the related video relationship is reciprocal. We use the dyade method, which calculates the percentage of video pairs that have a reciprocal relationship over the number of video pairs that are related (with or without reciprocity).

Are the highest-viewcount videos related? We conduct the following measurements. We pick the top-N videos in terms of highest viewcount and checked their connectivity in our graph. We compute the ratio of the existing edges among those videos over the maximum possible number of edges N(N-1) (recall that the edges are directed). The results for top-5, top-10 and top-20 videos are 0.5, 0.36 and 0.195 respectively. For the top-5, we find that 50% of these edges exist (10/20), which we found to be surprisingly high. For top-10 and top-20, the percentage drops rapidly.

Viewcount and related-video in-degree are not strongly correlated. One could expect that popular videos would appear in the related video list of more videos. Interestingly, there is only a moderate degree of correlation between viewcount and the in-degree of a video: the Pearson Correlation Coefficient of the two variables is 0.648. YouTube's algorithm is known to take into consideration title, keyword and description in order to identify related videos, and not just popularity metrics.

The related-video graph exhibits Small World network properties. As defined in the literature [13], a Small World network is a graph with a large clustering coefficient and small characteristic path length. Our graph has 0.326 clustering coefficient and 6.7 characteristic path length, which suggest that the related video graph could be characterized as a Small World network.

VI. Related Work

Recent studies have been focused on understanding YouTube video popularity from different angles. Cha et al. find that video popularity distribution exhibits a power-law pattern with truncated tails, which fits a "fetch-at-most-once" model [1]. Cheng et al. present an "active life span" model to study popularity trends and predict its future growth [2]. [3], [4] measure YouTube traffic from a campus network, and study popularity from a local perspective. Gill et al. [5] identifie and characterize user sessions on YouTube. Benevenuto et al. study video popularity by characterizing video responses [6]–[8]. Most of the above studies consider viewcount as the only popularity metric. Our work, as a important complement, broadens the research scope by studying several other popularity metrics together with viewcount.

The problem of spam detection and prevention has attracted research attention as well. Meeyoung Cha et al. [1] identify the phenomenon of content alias, where multiple identical or very similar copies for a single popular event exist in the system. Benevenuto et al. [7], [8] in particular deal with spams caused by video responses, and various approaches are provided to detect such spams. To extend these studies, we build a linear regression model that can potentially detect anomalies caused by users who exploit the capabilities of the YouTube system.

Researchers also study YouTube by modeling it as social network. [2], [6], [9], [10] study social networks representing video responses, user friendships, related videos, etc. Interesting observations such as small world, power-law distribution and local clustering are identified. Our work differs itself by focusing on a subgraph containing only popular nodes. We discover several properties that have never been reported before.

VII. Conclusion

This work is only the first step in understanding video popularity in YouTube, which provides an initial foundation for further exploring and understanding popularity. We study fundamental properties of video popularity in YouTube using roughly 37 million, accounting for 25% of all YouTube videos. We analyze the relationships of key popularity metrics and we find that four of them are highly correlated (viewcount, #comments, #ratings, #favorites). We also find a "magic number": for every 400 times a video is viewed, we have one of each: a comment, a rating of the video, and an addition to one's favorite set.

Based on this work, we identify several directions to further examine: (a) the co-evolution of the popularity metrics in time, meaning which metric increases first and which metrics follow, (b) the behavior of users in terms of reacting to a video, e.g. whether it is the same user that leaves a comment *and* rates a video, and (c) the effect of standard feeds on long term video popularity.

References

[1] M. Cha et al., "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system", in *IMC '07: Proc. of the 7th ACM SIGCOMM conference on Internet measurement*, 2007.

[2] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos", *IWQoS '08: Proc. of the 16th International Workshop on Quality of Service*, June 2008.

[3] P. Gill et al., "Characterizing user sessions on YouTube", *MMCN 2008: Proc. of the 15th SPIE/ACM Multimedia Computing and Networking*, 2008.

[4] M. Zink et al., "Watch global, cache local: YouTube network traffic at a campus network-measurements and implications", *MMCN 2008: Proc. of the 15th SPIE/ACM Multimedia Computing and Networking*, 2008.

[5] Phillipa Gill et al., "Youtube traffic characterization: a view from the edge", in *IMC '07: Proc. of the 7th ACM SIGCOMM conference on Internet measurement*, 2007.

[6] F. Benevenuto et al., "Video interactions in online video social networks", *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, no. 4, 2009.

[7] F. Benevenuto et al., "Identifying video spammers in online social networks", in *AIRWeb '08: Proc. of the 4th international workshop on Adversarial information retrieval on the web*, 2008.

[8] F. Benevenuto et al., "Detecting spammers and content promoters in online video social networks", in *SIGIR '09: Proc. of the 32nd international conference on Research and development in information retrieval*, 2009.

[9] J. C. Paolillo, "Structure and network in the youtube core", in *HICSS '08: Proc. of the 41st Annual Hawaii International Conference on System Sciences*, 2008.

[10] A. Mislove et al., "Measurement and analysis of online social networks", in *IMC '07: Proc. of the 7th ACM SIGCOMM conference on Internet measurement*, 2007.

[11] "YouTube API", "http://code.google.com/apis/youtube/developers_guide_protocol.html".

[12] Stanton Glantz and Bryan Slinker, *Primer of Applied Regression & Analysis of Variance*, McGraw-Hill Medical, 2 edition, Oct. 2000.

[13] D. J Watts and S. H Strogatz, "Collective dynamics of small-world networks", *Nature*, vol. 393, no. 6684, June 1998.