Understanding the Characteristics of Internet Short Video Sharing: A YouTube-Based Measurement Study

Xu Cheng, *Student Member, IEEE*, Jiangchuan Liu, *Senior Member, IEEE*, and Cameron Dale

***Abstract*—Established in 2005, YouTube has become the most successful Internet website providing a new generation of short video sharing service. Today, YouTube alone consumes as much bandwidth as did the entire Internet in year 2000 [1]. Understanding the features of YouTube and similar video sharing sites is thus crucial to their sustainable development and to network traffic engineering. In this paper, using traces crawled in a 1.5-year span (from February 2007 to September 2008), we present an in-depth and systematic measurement study on the characteristics of YouTube videos. We find that YouTube videos have noticeably different statistics compared to traditional streaming videos, ranging from length, access pattern, to their active life span. The series of datasets also allow us to identify the growth trend of this fast evolving Internet site, which has seldom been explored before. We also look closely at the social networking aspect of YouTube, as this is a key driving force toward its success. In particular, we find that the links to related videos generated by uploaders' choices form a small-world network. This suggests that the videos have strong correlations with each other, and creates opportunities for developing novel caching and peer-to-peer distribution schemes to efficiently deliver videos to end users.**

***Index Terms*—Measurement, peer-to-peer, social network, YouTube.**

I. Introduction

THE recent four years have witnessed an explosion of networked video sharing as a new killer Internet application. The most successful site, YouTube, now enjoys more than 6 billion videos being watched every month [2]. The success of similar sites like the new Yahoo! Video and Youku (the most popular video sharing site in China), and the expensive acquisition of YouTube by Google [3], further confirm the mass market interest. Their great achievement lies in the combination of the content-rich videos and, equally or even more importantly, the establishment of a social network. The systems allow content suppliers to upload videos effortlessly, and to tag uploaded videos with keywords and links to other videos. Users can easily share videos by mailing links to them, or embedding them in blogs. The videos are no longer independent from each other with the clients browsing them following the links. Consequently, popular videos can rise to the top in a very organic fashion. With no doubt, these sites are changing the content distribution landscape and even the popular culture.

Established in 2005, YouTube is one of the fastest-growing websites, and has become the third most accessed site in the Internet, as surveyed by Alexa [4]. An April 2008 report estimated that YouTube consumed as much bandwidth as did the entire Internet in year 2000 [1], and industry insiders estimate that YouTube spends roughly $1 million a day to pay for its server bandwidth [5]. On the other hand, a recent study revealed that the performance of YouTube is much worse than many other measured sites [6]. Therefore, understanding the features of YouTube-like sites is crucial to network traffic engineering and to the sustainable development of this new generation of service.

In this paper, we present an in-depth and systematic measurement study on the characteristics of YouTube videos. We crawled the YouTube site for a four-month period in early 2007, collecting three million distinct videos' information. We conducted a second round of crawling for a seven-month period in 2008, and have obtained 59 datasets totaling 5, 043, 082 distinct videos' information, which is, to our knowledge, the largest dataset crawled so far. From this large collection of datasets, we find that YouTube videos have noticeably different statistics from traditional streaming videos, in aspects from video length to access pattern. The long span of the two rounds of measurement also enables us to examine new features that have not been addressed in previous measurement studies, for example, the growth trend and active life span.

We also look closely at the social networking aspect of YouTube, as this is a key driving force toward the success of YouTube and similar sites. In particular, we find that the links to related videos generated by uploader's choices form a small-world network. This suggests that the videos have strong correlations with each other, and creates opportunities for developing novel caching and peer-to-peer distribution schemes to efficiently deliver videos to end users.

The rest of the paper is organized as follows. Section II presents the related work. Section III describes our method of gathering information of YouTube videos, which is then analyzed in Section IV. Section V further analyzes the social networking aspect. Section VI discusses the implications of the

Manuscript received January 30, 2009; revised May 13, 2009; accepted September 12, 2009. Date of publication May 31, 2013; date of current version July 15, 2013. This work was supported in part by a Canada NSERC Discovery Grant and an NSERC Strategic Project Grant. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pal Halvorsen.

X. Cheng and J. Liu are with the School of Computing Science, Simon Fraser University, Burnaby, BC, Canada (e-mail: xuc@cs.sfu.ca; jcliu@cs.sfu.ca).

C. Dale was with the School of Computing Science, Simon Fraser University, Burnaby, BC, Canada. He is now with IBM Canada, Vancouver, BC, Canada (e-mail: camerond@cs.sfu.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2265531

results, and suggests ways that the YouTube service could be improved. Finally, Section VII concludes the paper.

II. Related Work

There have been significant research efforts into understanding the workloads of traditional media servers, looking at, for example, the video popularity and access locality [7]–[10]. We have carefully compared their measurement results with ours, and have found that, while sharing similar features, many of the video statistics of these traditional media servers are quite different from YouTube-like sites, e.g., the video length distribution, user access pattern and active life span. More importantly, these traditional studies lack a social network among the videos.

We have seen simultaneous works investigating YouTube and similar Web 2.0 sites in the past four years. Cha *et al.* studied YouTube and Daum UCC, the most popular user-generated content (UGC) service in Korea. They examined the user behavior, identified the key elements that shape the popularity distribution, and also proposed some improvement for UGC design [11]. Gill *et al.* tracked YouTube transactions in a campus network, focusing on deriving video access patterns from the network edge perspective, and also discussed the improvement approaches such as caching and CDNs [12]. Our work complements theirs by crawling a much larger set of the videos and thus being able to more accurately measure their global properties, in particular, the social network, which was not addressed in those works.

Halvey *et al.* were the first to study the social network aspect in YouTube, focusing mainly on users [13]. Mislove *et al.* studied four online social networking sites (Flickr, YouTube, LiveJournal and Orkut), and confirmed the power-law, small-world and scale-free properties of online social networks [14]. Our study complements these existing works, including our previous work [15], by the long-term measurement that spans 1.5 years. It facilitates our understanding of the evolution and the latest development of this rapidly evolving service. We focus on the networks of YouTube videos, which are indirectly formed by user interactions yet have more significant implication than the networks of users. We also present initial attempts to exploring the social networks for accelerating content distribution.

III. Methodology of Measurement

We have built a YouTube crawler and collected the YouTube videos' information through a combination of the YouTube API and scrapes of YouTube video web pages. In this section, we first briefly introduce the YouTube techniques, and then describe our YouTube crawler and the crawled datasets.

A. YouTube Video Format and Meta-Data

YouTube's video playback technology is based on Adobe's Flash Player, which allows YouTube to display videos with quality comparable to well established video playback technologies (such as Windows Media Player, QuickTime and Realplayer). YouTube accepts uploaded videos in many formats, which are converted into the .FLV (Adobe Flash Video) format after uploading. It is well recognized that the use of a uniform easily-playable format is critical in the success of YouTube. During the course of our measurement, YouTube used the H.263 video codec, and it introduced "high quality" format that uses the H.264 codec for better viewing quality in late 2008 [16]. Our measurement and conclusions however are largely independent of these changes.

TABLE I
Meta-Data of a YouTube Video

ID	YiQu4gpoa6k
Uploader	NewAgeEnlightenment
Date Added	August 08, 2008
Category	Sports
Video Length	270 seconds
Number of Views	924,691
Number of Ratings	1,039
Number of Comments	212
Related Videos	ri1h2_jrVjU, 0JdQlaQpOuU, ...

YouTube assigns each video a distinct 11-digit ID composed of 0–9, a-z, A-Z, -, and _. Each video contains the following intuitive meta-data: video ID, user who uploaded it, date when it was added, category, length, number of views, ratings and comments, and a list of "related videos". The related videos are links to other videos that have similar titles, descriptions, or tags, all of which are chosen by the uploader. A YouTube page only shows at most 20 related videos at once, so we also limit our scrape to these top 20 related videos. A typical example of the meta-data is shown in Table I.

B. YouTube Crawler

Given the links between the videos, we consider all the YouTube videos to form a directed graph, where each video is a node in the graph. If video b is in the related video list of video a, then there is a directed edge from a to b. Our crawler uses a breadth-first search (BFS) to find videos in the graph.[1] We define the initial set of a list of IDs, which the crawler reads in to a queue at the beginning of the crawl. When processing each video, it checks the list of related videos and adds any new ones to the queue. Given a video ID, the crawler first extracts information from the YouTube API [17], which contains all the meta-data except for date added, category and related videos. The crawler then scrapes the video's webpage to obtain the remaining information.

We ran our crawler every two days, thus obtaining a number of datasets. In most cases, the crawl ended when it finished crawling the fourth depth each time. We started our crawl on February 22nd, 2007, and the first round ended on May 18th, 2007, collecting 2,994,947 videos. We started the second round of crawling on March 27th, 2008. On average, the crawler found 81 thousand distinct videos each time. The crawl ended on September 8th, 2008, collecting 5,043,082 videos, in which only 8.3% of the data were also crawled in the first round, suggesting that YouTube is rapidly growing.

To study the growth trend of the video popularity, we also used the crawler to update the statistics of some previously

[1]We use BFS because it can easily find the active videos that are close to the YouTube entry page with no bias. It also facilitates our control of the crawling scale, so that each crawl of a dataset will not last for a long time, which will be unsuitable for measurement of some dynamic characteristics, e.g., number of views.

found videos. For this crawl we only retrieved the number of views for relatively new videos. In 2007, we obtained 7 datasets in a two-month period. In 2008, we re-collected this information, crawling once a week from April to September 2008, which resulted in 21 datasets. We will be focusing on the 2008 data, which represents the latest development of YouTube, and we will point out noticeable and interesting differences between the 2008 and 2007 data.

We also separately crawled the file size and bitrate information. To get the file size, the crawler retrieved the response information from the server when requesting to download the video file and extracted the information on the size of the download. Some videos also have the bitrate embedded in the FLV video meta-data, which the crawler extracted after downloading the meta-data of the video file.

Finally, the crawler retrieved information on the number of uploaded videos and friends of each user from the YouTube API, for a total of more than 2 million users.

All the crawled data are available online at http://netsg.cs.sfu.ca/youtubedata.html.

IV. Characteristics of YouTube Video

Our crawled videos constitute a good portion of the entire YouTube video repository (around 120 million videos as of September, 2008). Because most of these videos can be accessed from the YouTube homepage in less than 10 clicks, they are generally active and thus representative for measuring characteristics of the repository. We will also show later that our crawled datasets are not biased.

In the measurements, some characteristics of a video are static and can be measured once from the entire dataset (e.g., category, length and date added). Some characteristics are dynamic, which change from dataset to dataset (e.g., number of views). We consider this dynamic information to be static over a single crawl. Later, the updated number of views information will be used to measure the growth trend and active life span of videos.

A. Video Category

In YouTube, one of the 15 categories is selected by a user when uploading a video. Table II lists the number and percentage of all the categories. In our entire dataset, we can see that the distribution is highly skewed: the most popular category is "Entertainment", at about 25.4%, and the second is "Music", at about 24.8%. These two categories of videos constitute half of the entire YouTube videos, suggesting that YouTube is mainly an entertainment-like site.

B. Video Length

The length of YouTube videos is the most distinguished difference from traditional media contents. Whereas most traditional servers contain a significant portion of long videos, typically 1–2 hour movies (e.g., HPLabs Media Server [9] and OnlineTVRecorder [18]), YouTube is mostly comprised of short video clips. In our entire dataset, 98.0% of the videos' lengths are within 600 seconds. This is mainly due to the limit of 10 minutes imposed by YouTube on regular users uploads. We do find videos longer than this limit though, because the YouTube Director Program allows a small group of authorized users to upload videos longer than 10 minutes [19].

TABLE II
List of YouTube Video Categories

Rank	Category	Count	Pct.	Pct. (2007)
1	Entertainment	1,304,724	25.4%	25.2%
2	Music	1,274,825	24.8%	22.9%
3	Comedy	449,652	8.7%	12.1%
4	People & Blogs	447,581	8.7%	7.5%
5	Film & Animation	442,109	8.6%	8.3%
6	Sports	390,619	7.6%	9.5%
7	News & Politics	186,753	3.6%	4.4%
8	Autos & Vehicles	169,883	3.3%	2.6%
9	Howto & Style	124,885	2.4%	2.0%
10	Pets & Animals	86,444	1.7%	1.9%
11	Travel & Events	82,068	1.6%	2.2%
12	Education	54,133	1.1%	–
13	Science & Technology	50,925	1.0%	–
14	Unavailable	42,928	0.8%	0.9%
15	Nonprofits & Activism	16,925	0.3%	–
16	Gaming	10,182	0.2%	–
17	Removed	9,131	0.2%	0.5%

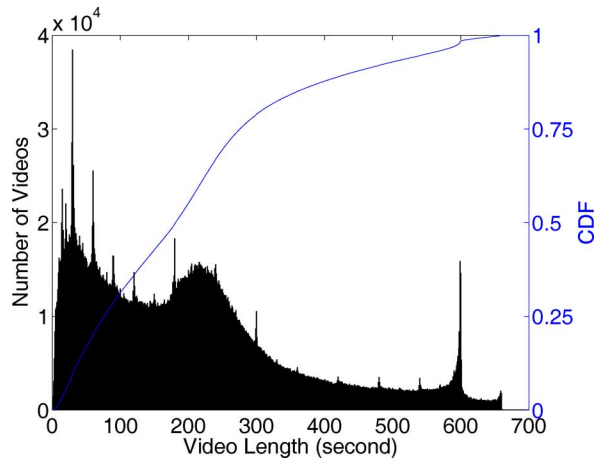

Fig. 1. Distribution of YouTube video length.

Fig. 1 shows the histogram and cumulative distribution function (CDF) of YouTube videos' lengths within 700 seconds, which exhibits three peaks. The first peak is within one minute, and contains 20.0% of the videos, which shows that YouTube is primarily a site for very short videos. The second peak is between 3 and 4 minutes, and contains about 17.4% of the videos. As shown in Fig. 2, this peak corresponds to the videos in the "Music" category, which is the second most popular category for YouTube. The third peak is near the maximum of 10 minutes, and is caused by the limit on the length of uploaded videos. This encourages some users to circumvent the length restriction by dividing long videos into several parts, each being near the limit of 10 minutes. Similar reason also explains the peaks at around every exact minute.

Fig. 2 shows the video length distributions for the top four most popular categories. "Entertainment" videos have a similar distribution as the entire videos', and have the greatest peak at around 10 minutes. This is because a great portion of these videos are talk shows, which are typically a half hour to several hours in length, but have been cut into several parts near 10 minutes. "Music" videos have a very large peak between three and four minutes (29.1%). "Comedy" and "People & Blogs"

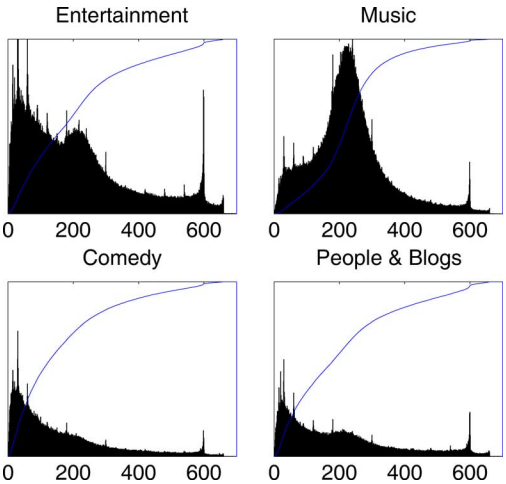

Fig. 2. Length distributions for the four top categories.

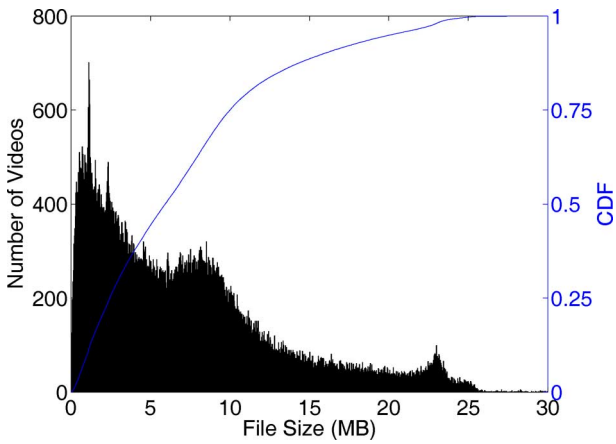

Fig. 3. Distribution of YouTube video file size.

videos have more videos within two minutes (53.1% and 41.7% respectively), likely corresponding to "highlight" type of clips.

C. Video File Size and Video Bitrate

Using video IDs from a normal crawl, we retrieved the file size of more than 130 thousand videos. Not surprisingly, we find that the distribution of video sizes is very similar to the distribution of video lengths, due to the constant bitrate (CBR) coding mode used in YouTube. We plot the histogram and CDF of YouTube video file sizes in Fig. 3. In our crawled data, 99.1% of the videos are less than 25 MB. For the 2008 dataset, we calculate an average video file size to be 7.6 MB, which is smaller than that of the 2007 dataset (8.4 MB), so there are more and more short videos uploaded. Nevertheless, considering there are nearly 120 million YouTube videos, the total disk space required to store all the videos is nearly 900 TB! Smart storage management is thus quite demanding for such an ultra-huge and still growing site, which we will discuss in more detail in Section VI.

We found that 99.6% of the videos we crawled contained FLV meta-data specifying the video's bitrate in the beginning of the file. For the rest of the videos, we calculate an average bitrate from the file size and its length. As shown in Fig. 4, the videos' bitrate has a clear peak at around 320 kbps, with two other peaks at around 285 kbps and 200 kbps. This implies that YouTube videos have a moderate bitrate that balances quality and bandwidth. Table III lists the statistics of video length, file size and video bitrate.

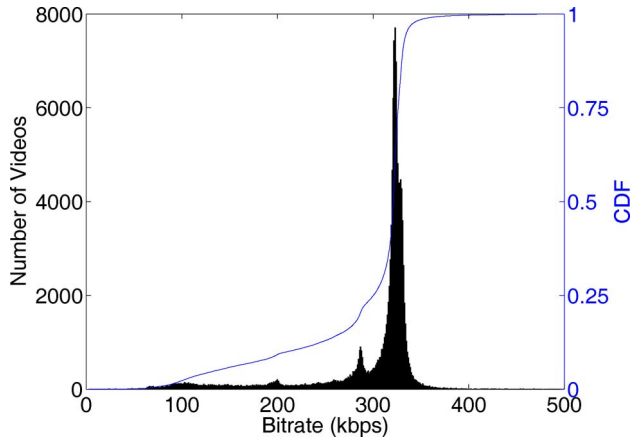

Fig. 4. Distribution of YouTube video bitrate.

TABLE III
STATISTICS OF VIDEO LENGTH, FILE SIZE AND BITRATE

	Min	Max	Mean	Median	Std. Dev.
length(s)	0	10799	210	182	196.6
size(MB)	0.003	338	7.6	6.0	8.8
bitrate(kbps)	11	1005	299	322	60.7

D. Date Added—Uploading Trend

During our crawl, we recorded the date that each video was uploaded, so as to study the YouTube's uploading trend. Fig. 5 shows the number of new videos added every two weeks in our entire crawled dataset of 2008. YouTube was established on February 15th, 2005, and we can see there is a slow start, as the earliest video we crawled was uploaded 8 days after that day. Note that, we can get the early videos only if they are still quite active or are linked to by other videos we crawled. After 6 months from YouTube's establishment, the number of uploaded videos increases steeply. This trend can be well fitted by a power law curve, as shown in Fig. 5.

In the dataset we collected, the number of uploaded videos decreases steeply starting in March, 2008. However, this does not imply that the uploading rate of YouTube videos has suddenly decreased. The reason is that, by that time, many newly uploaded videos had yet to be included in other videos' related list, and thus could not be found by our crawler unless they were very popular right after uploading. We have found that the 2007 data also shows this feature, and the 2008 data indeed confirms that the uploading trend does not decrease.

E. Views—User Access Pattern

The number of views a video has had is another important characteristic we measured, as it reflects the popularity and access patterns of the videos. Because this property is changing over time, we cannot measure it from the entire dataset that combines all the data together. Therefore we use a single dataset

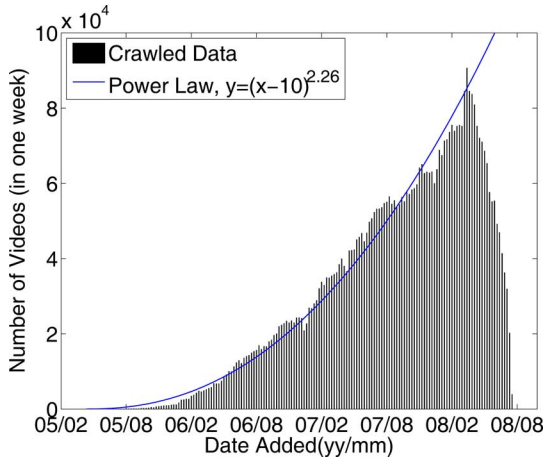

Fig. 5. Number of YouTube videos added in crawled data.

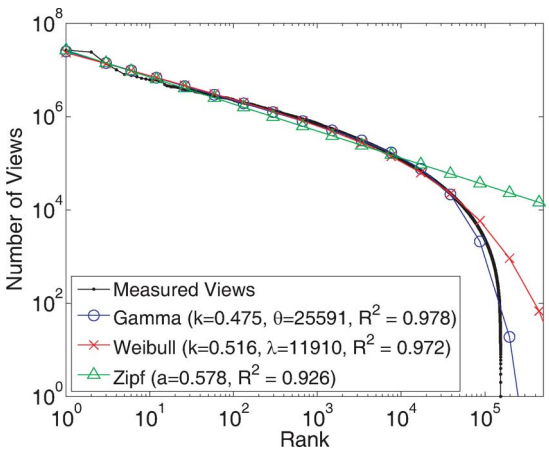

Fig. 6. YouTube videos' views against rank.

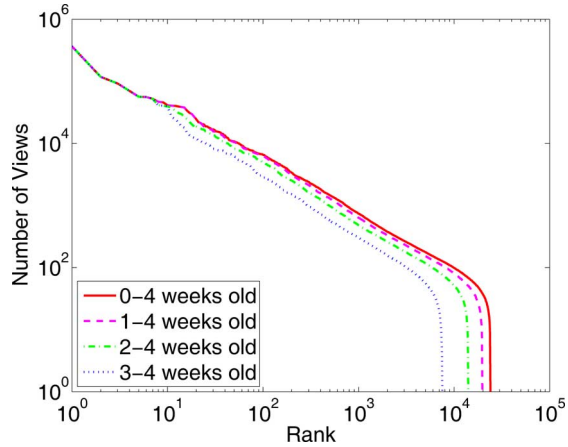

Fig. 7. Recently added videos' views against rank.

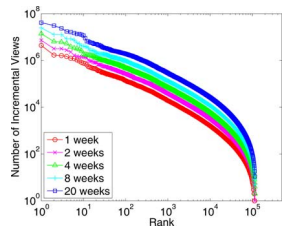

Fig. 8. YouTube videos' incremental views against rank.

containing more than 150 thousand videos, which can be considered relatively static. We have examined all the datasets, and others also show the same result.

Fig. 6 shows the number of views as a function of the rank of the video by its popularity. Though the plot has a long tail on the linear scale, it does not follow the well-known Zipf distribution $y = x^{-\alpha}$ [20], which should be a straight line on a log-log scale. We can see in the figure, the beginning of the curve is linear on a log-log scale, but the tail (after 10^4th videos) decreases tremendously, indicating there are not so many unpopular videos as Zipf's law predicts. We believe that this is because the links among videos enable each of them being browsed by interested viewers through various paths. Another reason might be a user will access to his/her own video several times after uploading it to check if it is successfully uploaded, and thus there are few videos that have never been accessed or only once.

While previous measurements on traditional media servers also found that video accesses to a media server does not follow the Zipf's law [7]–[10], our result differs from theirs in which the curve either is skewed from beginning to end or does not have the heavy tail. Their results indicate that the popular videos are also not as popular as Zipf's law predicts, which is not the case in YouTube.

To fit the skewed curve, previous studies have used a concatenation of two Zipf distributions [8] or a generalized Zipf distribution [9]. However, we find that the Gamma and Weibull distributions both fit better than the Zipf, due to the heavy tail (in log-log scale) that they have. We also calculate the coefficient of determination with R^2 to measure the fitness, and as the figure shows, the Gamma distribution is better than the Weibull, and both are much better than the Zipf.

We were initially concerned that the crawled data might be biased, as popular videos may appear in our BFS more likely than non-popular ones. Since the entire video name space is too large (2^{66}), a direct random sampling can be quite difficult. Therefore, we had been saving the recently added videos from the YouTube RSS feed for four weeks, as sampling from these is close to random. We update the views counts of these videos, and plot in Fig. 7. The leftmost (blue) plot is the videos added during the first week only (i.e., all the 3–4 weeks old videos), while the rightmost (red) plot contains all the videos (i.e., all the 0–4 weeks old videos). There is a clear heavy tail in all the plots, verifying that our BFS crawl does find non-popular videos just as well as it finds popular ones. Using the 21 datasets of updated views, we calculate the incremental views of the videos for different spans of time, as shown in Fig. 8. We can see as the time passes, the curve has a more and more heavy tail, also confirming this.

Next, we investigate the correlation between video's length and number of views. We divide the dataset into five groups and calculate the statistics of views, as list the statistics in Table IV. We can see that medium-length videos (151 s–240 s and 241 s–420 s) are relatively more popular than very short videos and long videos. However, we find that the deviations in all of the

TABLE IV
CORRELATION BETWEEN LENGTH AND VIEWS

Length	Min	Max (k)	Mean	Median	Std. Dev.
$\leq 60s$	0	5,696	27,066	3,675	108,881
61s–150s	0	7,179	38,286	4,562	127,795
151s–240s	0	26,601	46,118	5,973	237,230
241s–420s	0	24,192	46,224	7,101	212,712
$> 420s$	0	4,386	19,095	6,399	107,909

TABLE V
CORRELATION BETWEEN VIDEO AGE AND VIEWS

Age	Min	Max (k)	Mean	Median	Std. Dev.
≤ 1 month	0	768	5,169	449	28,134
1–3 month	3	7,179	8,677	1,198	66,324
3–6 month	7	3,398	14,549	2,829	62,925
6–12 month	3	6,624	31,852	6,514	122,968
> 12 month	4	26,601	74,045	17,829	275,270

five groups are very large, and the correlation coefficient of video length and number of views is 0.007, implying that the correlation is indeed pretty weak.

Finally, we examine the correlation between video age and number of views, which is shown in Table V. Not surprisingly, the video age affects the number of views (correlation coefficient being 0.166), because older videos have more opportunity to be accessed. However, we can see in the younger video groups there are very popular videos, and in the older video group there are very unpopular videos. In fact, the deviations in all the groups are quite large. This indicates that different videos have different *growth trends*, i.e., the videos' popularities increase in various speeds.

F. Growth Trend of Number of Views and Active Life Span

To model the growth trend, starting from April 16th, 2008, we updated the number of views of relatively new videos every week for 21 weeks. We eliminate the videos that have been removed, resulting in 120 thousand videos for evaluating.

We have found that the growth trend can be modeled better by a power law than a linear fit. A video can have an increasing growth (if the power is greater than 1), a constant growth trend (power close to 1), or a slowing growth (power less than 1). The trend depends on the exponent, which we call the growth trend factor p. Let μ be the number of weeks that the video has been uploaded before the first update, and v_0 be the number of views the video had in the first dataset. We have

$$v(x) = v_0 \times \frac{(x+\mu)^p}{\mu^p}. \tag{1}$$

We evaluate the 120 thousand videos using (1) to get the distribution of their growth trend factors p, which is shown in Fig. 9. We also plot the growth trend factor of the 2007 data. For the 2008 data, over 80% of the videos have a factor being less than 1, indicating that most videos grow more and more slowly as time passes. We notice that this number is greater than the 2007 data, implying that YouTube is growing fast and new videos are uploaded more frequently.

Since YouTube has no policy on removing videos after a period of time, the life span of YouTube videos is generally infinite. However, for a video of growth trend factor less than 1, its popularity will almost stop growing after a certain time; more

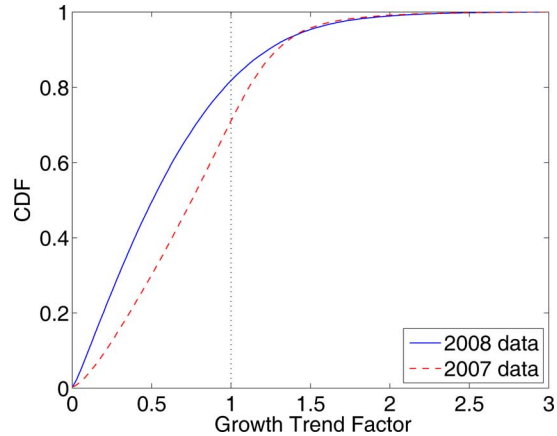

Fig. 9. Distribution of video growth trend factor.

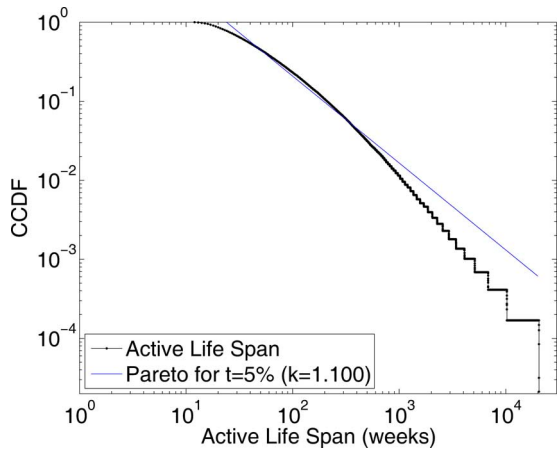

Fig. 10. Distribution of estimated active life span.

formally, if its number of views increases by a factor less than a threshold t from the previous week, we define its *active life span* to be over. We can compute this active life span l from $v(l)/v(l-1) - 1 = t$, that is,

$$l = \frac{1}{\sqrt[p]{1+t}-1} + 1 - \mu. \tag{2}$$

Fig. 10 shows the complementary cumulative distribution function (CCDF) for the active life span of the approximately 95 thousand videos (with p less than 1), for a threshold of $t = 5\%$. It can be roughly fitted by a Pareto CCDF $(x_m/x)^k$, where x_m indicates the offset of the fitted line, and the the parameter k is about 1.100. From looking at multiple fits with various values of t, we find that they all result in similar parameter k, and the only difference is the offset (x_m) of the line, because the model reflects the relative increase (threshold t) instead of the absolute increase.

Since the server logs of YouTube are not publicly available, we cannot directly measure the temporal locality that shows whether recently accessed videos are likely to be accessed in the near future. Fortunately, the active life span gives us a way to estimate the temporal locality. Specifically, the fitted Pareto distribution in Fig. 10 implies that most of the videos have been watched frequently only in a short span of time, and after a video's active life span, fewer people will access it. This

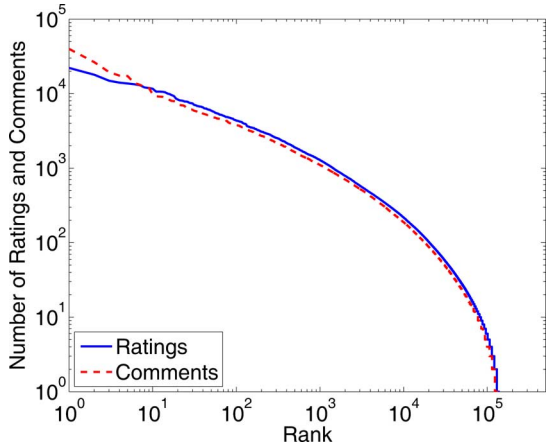

Fig. 11. Numbers of ratings and comments against ranks.

TABLE VI
STATISTICS OF VIEWS, RATINGS AND COMMENTS

	Mean	Median	Std. Dev.	Zeros
views	37,595	5,357	174,780	0.1%
ratings	68	12	280	8.5%
comments	59	10	283	12.2%
views (2007)	4,771	741	24,892	0.1%
ratings (2007)	15	3	128	21.6%
comments (2007)	9	2	39	33.6%

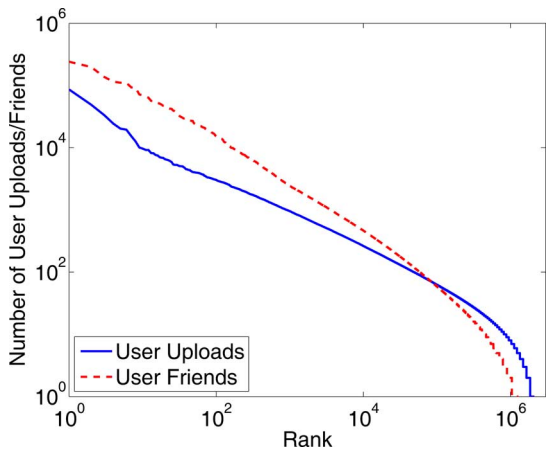

Fig. 12. Number of uploads and friends against ranks.

characteristic has good implications for web caching and server storage. We can design a predictor to forecast the active life span using our active life span model, which can help a proxy or server to make more intelligent decisions, such as when to drop a video from the cache.

V. THE SOCIAL NETWORK IN YOUTUBE

YouTube is a prominent social media service: there are communities and groups in YouTube, and so videos are no longer independent from each other. We next examine the social networks among YouTube users and videos, which is a very unique and interesting aspect of this new generation of video sharing sites.

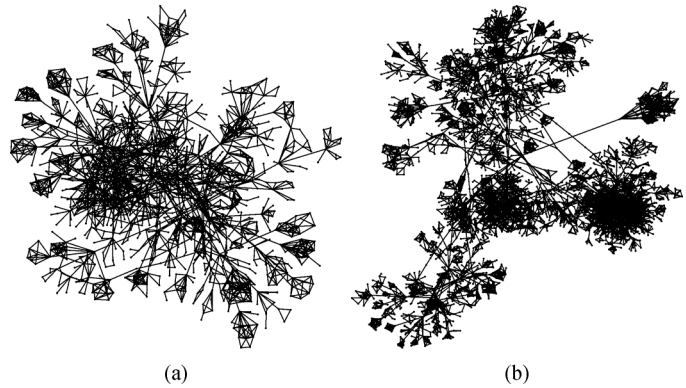

Fig. 13. Two sample graphs of YouTube videos and their links. (a) 1000 nodes. (b) 4000 nodes.

A. User Ratings & Comments, Friends and Upload

YouTube currently has more than 11 million registered users,[2] who can login to upload video or watch some restricted videos. A registered user can also add another user to their friend list so as to conveniently watch their friends' videos. Our crawler has found 2.1 million distinct registered users from all the crawls, which constitute a quite good portion of the existing registered users.

We first study the statistics of number of ratings and comments from the same dataset as we did number of views. Since a user needs to log in to rate and comment a video, the number of ratings and comments partially reflect the user behavior. Fig. 11 plots the number of ratings against the rank, and also for the comments. The two have a similar distribution, and we note that the tails do not drop so quickly as that of the number of views, indicating that they are more Zipf-like. This is because the ratings and comments cannot be accessed from the other videos. We also list the statistics of ratings and comments in Table VI, along with views, and the statistics of the 2007 data for comparison. We can see that both comments and ratings are much fewer than views, and many videos do not have a single rating or comment. Interestingly, the number of videos that have no rating or comment decreases compared with the 2007 data, indicating that users are more willing to rate and make comments than before.

In Fig. 12, the blue line plots the rank of user uploaded videos' count. The distribution is similar to those of views, ratings and comments. We find that many users have uploaded a few videos, and a small number of users have uploaded many videos. Since the user IDs we collected are from the previous crawled video data, all of these users have uploaded at least one video. However, the information of users that do not upload videos cannot be directly crawled, though we believe that there are a huge number of such users.

We also plot the rank of number of friends each user has in Fig. 12 as the red line. Interestingly, we find that 40.3% of the users have no friends in all these user data. Therefore, we can see that having friends does not affect the access pattern, so that the relationships between users are not very strong. This suggests

[2]A channel search for "*" as a wildcard character in YouTube returns more than 11 million channels, of which each registered user has one.

that the social network existing among users is of less impact. We will thus focus on the networks among videos, which are indirectly formed with user interactions.

B. The Small-World in YouTube Videos

Small-world network phenomenon is probably the most interesting characteristic for social networks. Milgram [21] initiated the study of small-world networks when investigating the phenomenon that people are linked by short chains of acquaintances (a.k.a. six degrees of separation). Such networks possess characteristics of both random and regular graphs [22]. More formally, given the network as a graph $G = (V, E)$, the *clustering coefficient* C_i of a node $i \in V$ is the proportion of all the possible edges between neighbors of the node that actually exist in the graph, and the clustering coefficient of the graph, $C(G)$, is the average of the clustering coefficients of all nodes. The *characteristic path length* d_i of a node $i \in V$ is the average of the minimum number of hops it takes to reach all other nodes in V and the characteristic path length of the graph, $D(G)$, is then the average of the characteristic path lengths of all nodes. A small-world network has a large clustering coefficient like a regular graph, but also has a small characteristic path length like a random graph.

We measured the graph topology for the network of YouTube videos, by using the related links in YouTube pages to form directed edges in a video graph. Videos that have no outgoing or no incoming links are removed from the analysis. Some visual illustrations for parts of the network (about 1000 and 4000 nodes) are shown in Fig. 13.

From the entire crawled data, we obtain four datasets for measurement, each consisting different order of magnitude number of videos. Since not all of the YouTube videos were crawled, the graphs are not strongly connected, making it difficult to calculate the path length. Thus we use the largest strongly connected component of each graph for the measurements. For comparison, we also generate random graphs that are strongly connected. Each of the random graphs has the same number of nodes and average node degree of the datasets.

Fig. 14(a) shows the clustering coefficient for the graph, as a function of the size of the dataset. The clustering coefficient is quite high (between 0.2 and 0.3), especially in comparison to the random graphs (nearly 0). There is a slow decreasing trend in the clustering coefficient, showing that there is some inverse dependence on the size of the graph, which is common for small-world networks [23]. Fig. 14(b) shows the characteristic path length for the graphs. We can see that the average diameter (between 10 and 15) is only slightly larger than the diameter of a random graph (between 4 and 8), which is quite good considering the still large clustering coefficient of these datasets. Moreover, as the size of graph increases, the characteristic path length decreases for the YouTube video graph, but increases for the random graph. This phenomena further verifies that the YouTube graph is a small-world network.

The small-world characteristics of the video graph can also be observed from the their visual illustrations (see Fig. 13). The clustering coefficient is large compared to the same sized random graph, while the characteristic path lengths are approaching the short path lengths in the random graphs. We

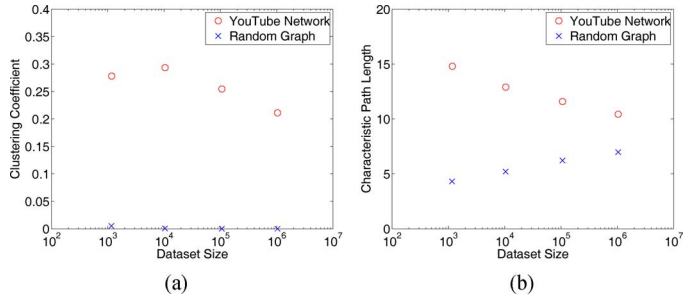

Fig. 14. Small-world characteristic of YouTube videos. (a) Clustering coefficient. (b) Characteristic path length.

believe this is due to the user-generated nature of the tags, titles, and descriptions of the videos that is used by YouTube to find related ones. The results are similar to other real-world user-generated graphs that exist, yet their parameters can be quite different. For example, the graph formed by URL links in the World Wide Web exhibits a much longer characteristic path length of 18.59 [24]. This is likely due to the larger number of nodes (8×10^8 in the web), but it also indicates that the YouTube network of videos is a much closer group.

VI. Further Discussions

We next discuss the implications of our measurement results toward improving the YouTube service. Given that YouTube is suffering from the huge bandwidth cost but it indeed survives only by attracting more users, scalability is no doubt the biggest challenge it faces, and is thus also our focus.

A. Implications on Proxy Caching and Storage Management

Caching frequently used data at proxies close to clients is an effective way to save backbone bandwidth and prevent users from experiencing excessive access delays. Numerous algorithms have been developed for caching web objects or streaming videos [25]. While we believe that YouTube will benefit from proxy caching, three distinct features call for novel cache designs. First, the number of YouTube videos (120 million) is orders of magnitude higher than that of traditional video services (e.g., 2999 in HPC and 412 in HPL [9]). Second, the size of YouTube videos (99.1% being less than 25 MB) is much smaller than a traditional video (a typical MPEG movie of 700 MB). Finally, the view frequencies of YouTube videos do not well fit a Zipf distribution, which has important implications on web caching [26].

Considering these factors, full-object caching for web or segment caching for streaming video are not practical solutions for YouTube. Prefix caching [27] is probably the best choice. Assume for each video, the proxy will cache a 10 second initial prefix, i.e., about 400 KB of the video. Given the Gamma distribution of view frequency suggested by our measurements, we calculate and plot the hit-ratio as a function of the cache size in Fig. 15, assuming that the cache space is devoted only to the most popular videos. To achieve an 80% hit-ratio, the proxy would require less than 8 GB of disk space for the current YouTube video repository, which is acceptable for today's proxy servers.

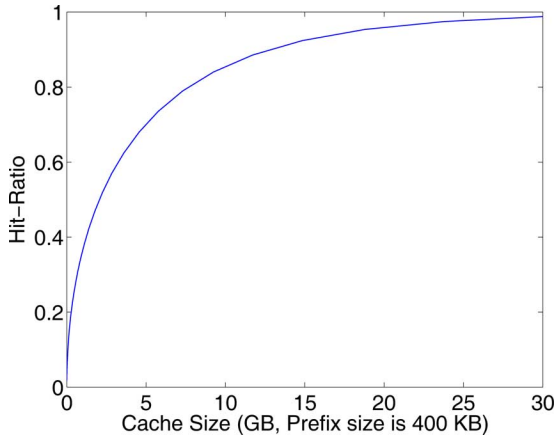

Fig. 15. Prefix caching hit-ratio as a function of cache size.

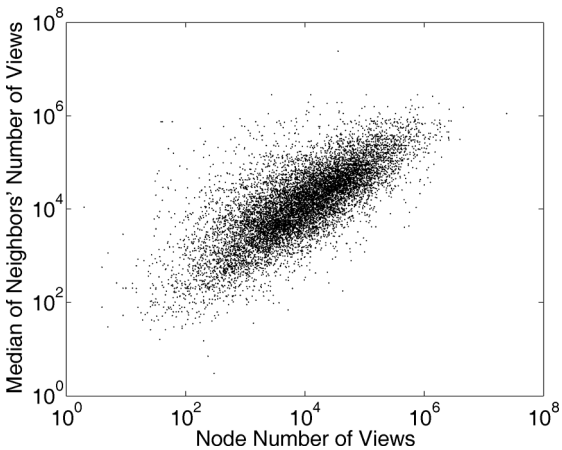

Fig. 16. Correlation of views and median of neighbors' views.

The cache efficiency can be further improved by exploring the small-world characteristic of the related video links. That is, if a group of videos have a tight relation, then a user is likely to watch another video in the group after finishing the first one. This expectation is confirmed by Fig. 16, which shows a clear correlation (correlation coefficient being 0.749) between the number of views for a videos and the median views of its related videos. Once a video is played and cached, the prefixes of its directly related videos can also be pre-fetched and cached, if the cache space allows.

A remaining critical issue is when to release the space for a cached video, given the constant evolution of YouTube's video repository. Currently, videos on a YouTube server will not be removed by the operator unless they violate the terms of service. With around 15 hours of video being uploaded every minute (approximate 370,000 videos every day) [28], the server storage will soon become a problem. We have found in Section IV-F that the active life span of YouTube videos roughly follows a Pareto distribution, implying that most videos are popular during a relatively short span of time. Therefore, a predictor can be developed to forecast the active life span of a video (using (2)). With the predictor, the proxy can decide which videos have already passed their life span, and replace it if the cache space is insufficient. Also a hierarchical storage structure can be built with videos passing their active life span being moved to slower and cheaper storage media.

B. Can Peer-to-Peer Save YouTube?

In the mean time of the booming of YouTube-like sites, peer-to-peer has evolved into a promising communication paradigm for large-scale content sharing. With each peer contributing its bandwidth to serve others, a peer-to-peer overlay scales extremely well with larger user bases. Besides file sharing, it has been quite successful in supporting large-scale live streaming [29] (e.g., CoolStreaming [30] and PPLive [31]) and on-demand video streaming, thus naturally being believed as an accelerator of great potentials for YouTube-like video sharing services.

Unfortunately, using peer-to-peer delivery for short video clips can be quite challenging, evidently from our measurement results on YouTube. In particular, the length of a YouTube video is short, so the long startup delay becomes unacceptable. A user often quickly loads another video when finishing the previous one, so the overlay will suffer from an extremely high churn rate. Moreover, there are a huge number of videos with highly skewed popularity, thus many of the peer-to-peer overlays will be too small to function well. They together make existing per-file based overlay design inefficient. Previous study on MSN Video has suggested a peer-assisted VoD (PA-VoD) [10]. However, we notice that the statistics for MSN Video are quite different from YouTube, particularly in terms of lengths, and consequently PA-VoD may not perform well. On the other hand, our social network finding could be explored to address the above challenges.

We envision a social-network assisted peer-to-peer system customized for short video sharing [32]. A distinct feature is that a peer caches all its previously played videos, and makes them available for re-distributing. All the peers that are watching a particular video or that have previously cached it and are serving as potential suppliers form an overlay for this video. The system further introduces an upper-layer overlay on top of the overlays of individual videos, as shown in Fig. 17. When a peer finishes watching a video, it checks its neighbors about the availability of the next video, and the neighbors will further check with theirs. Intuitively, this upper-layer overlay exploits the clustering property of the social network, which brings these peers of similar interests closer. We also further introduce a social network assisted pre-fetching mechanism to achieve fast and smooth transition [32].

We have performed both simulations and PlanetLab [33] prototype experiments to verify the effectiveness of our proposed design. The simulation is based on the latest crawled dataset, and we also emulate the real Internet environment by considering different bandwidth capacity. In the PlanetLab experiment, we have conducted a series experiment with the number of online nodes ranging from 50 to 235 (the maximum number of active nodes on PlanetLab during our experiment). As comparison, we have also implemented the client/server system and PA-VoD, a state-of-the-art system that relieves server stress of MSN videos [10]. In PA-VoD, the clients also serve as peers to relay traffic, but unlike our design, they do not cache and share

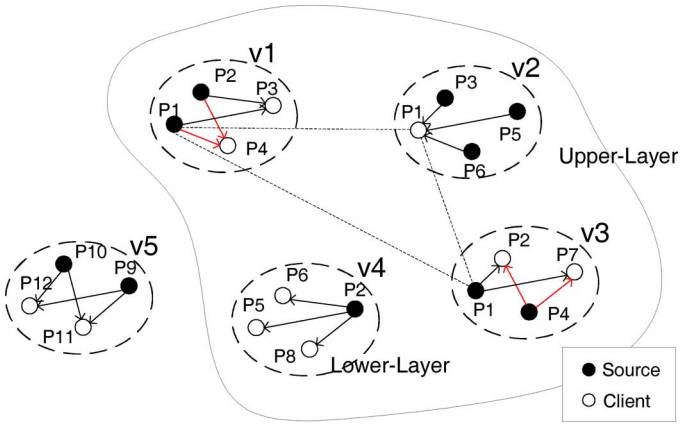

Fig. 17. Illustration of a bi-layer overlay.

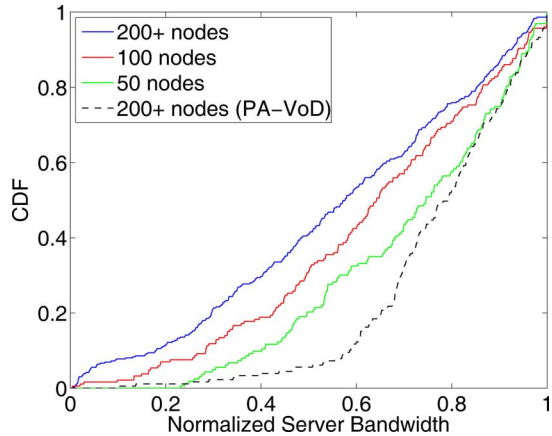

Fig. 19. CDF of normalized server traffic (PlanetLab).

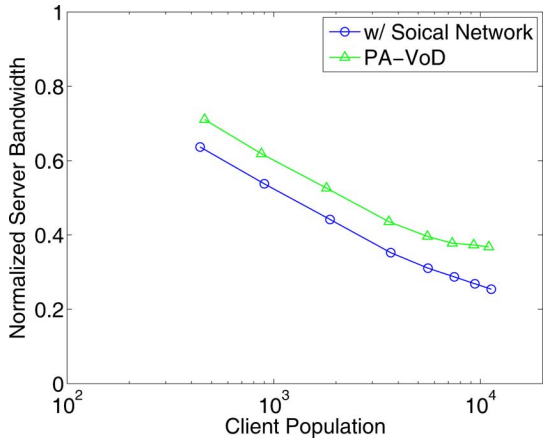

Fig. 18. Server bandwidth consumptions (simulation).

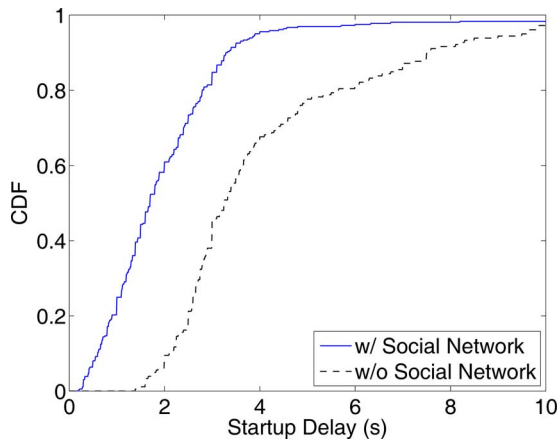

Fig. 20. CDF of startup delays (PlanetLab).

videos that they have downloaded, and it is also blind to the existence of social networks.

Fig. 18 compares the server bandwidth consumptions, where the results are normalized by the total bandwidth of the pure client/server system. It is obvious that our system saves more server bandwidth than PA-VoD does for all client populations. More importantly, the consumption drops more quickly than that of PA-VoD when the number of clients increases, suggesting that our design is quite scalable with client population. Fig. 19 shows the CDF of the normalized server bandwidth for PlanetLab experiment. Clearly, our system saves more server bandwidth: for 200 concurrent clients, near 55% of them have downloaded less than three fifths traffic from server; in contrast, only 10% peers save that much bandwidth in PA-VoD. We also examine the startup delay which is the interval from selecting a new video to starting play this video, as shown in Fig. 20. We can see more than 95% peers have an average startup delay shorter than 4 seconds with social network assisted pre-fetching, whereas only 70% peers can achieve this without it.

VII. Conclusion

This paper has presented a detailed investigation of the characteristics of YouTube, the most popular Internet short video sharing site to date. Through examining millions of long-term YouTube video data, we have demonstrated that, while sharing certain similar features with traditional video system, YouTube exhibits many unique characteristics, especially in length distribution, access pattern and growth trend. These characteristics introduce novel challenges and opportunities for optimizing the performance of short video sharing services.

We have also investigated the social network of YouTube videos, which is probably the most unique and interesting aspect, and has substantially contributed to the success of this new generation of service. We have found that the networks of related videos, which are chosen based on user preferences, have noticeable small-world characteristics, namely a large clustering coefficient indicating the grouping of videos, and a short characteristic path length linking any two videos.

We have suggested that these features can be exploited to facilitate the design of novel caching and peer-to-peer strategies for short video sharing. We have presented some initial results of a social network assisted peer-to-peer short video streaming system, and shown that it effectively reduces the workload of server, improves the quality of playback, and scales well to large client populations.

References

[1] "Web Could Collapse as Video Demand Soars", 2008. [Online]. Available: http://www.telegraph.co.uk/news/uknews/1584230/Web-could-collapse-as-video-demand-soars.html.

[2] "YouTube Surpasses 100 Million U.S. Viewers for the First Time", 2009. [Online]. Available: http://www.comscore.com/press/release.asp?press=2741.
[3] "Google to Buy YouTube for $1.65 Billion", 2006. [Online]. Available: http://money.cnn.com/2006/10/09/technology/googley-outube_deal/index.htm.
[4] "YouTube.com—Traffic Details from Alexa". [Online]. Available: http://www.alexa.com/siteinfo/youtube.com.
[5] "YouTube Looks for the Money Clip", 2008. [Online]. Available: http://techland.blogs.fortune.cnn.com/2008/03/25/youtube-looks-for-the-money-clip/.
[6] M. Saxena, U. Sharan, and S. Fahmy, "Analyzing video services in web 2.0: A global perspective," in *Proc. ACM Int. Workshop Network and Operating Syst. Support for Digital Audio and Video (NOSSDAV)*, May 2008, pp. 39–44.
[7] S. Acharya, B. Smith, and P. Parnes, "Characterizing user access to videos on the world wide web," in *Proc. ACM/SPIE Multimedia Comput. and Network (MMCN)*, Jan. 2000.
[8] J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon, "Analysis of educational media server workloads," in *Proc. ACM Int. Workshop Network and Operat. Syst. Support for Digital Audio and Video (NOSSDAV)*, Jun. 2001, pp. 21–30.
[9] W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat, Long-Term Streaming Media Server Workload Analysis and Modeling, HP Labs, 2003, Tech. Rep. HPL-2003-23.
[10] C. Huang, J. Li, and K. W. Ross, "Can Internet video-on-demand be profitable?," in *Proc. ACM SIGCOMM*, Aug. 2007, pp. 133–144.
[11] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM Conf. Internet Measurement (IMC)*, Oct. 2007, pp. 1–14.
[12] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the edge," in *Proc. ACM SIGCOMM Conf. Internet Measurement (IMC)*, Oct. 2007, pp. 15–28.
[13] M. Halvey and M. Keane, "Exploring social dynamics in online media sharing," in *Proc. ACM Int. Conf. World Wide Web (WWW) Poster*, May 2007, pp. 1273–1274.
[14] A. Mislove, M. Marcon, K. P. Gummadi, P. Dreschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. ACM SIGCOMM Conf. Internet Measurement (IMC)*, Oct. 2007, pp. 29–42.
[15] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of YouTube videos," in *Proc. IEEE Int. Workshop Quality of Service (IWQoS)*, Jun. 2008, pp. 229–238.
[16] "YouTube Technical Notes (Wikipedia)". [Online]. Available: http://en.wikipedia.org/wiki/YouTube #Technical_notes.
[17] "YouTube API". [Online]. Available: http://youtube.com/dev_docs.
[18] T. Hoßfeld and K. Leibnitz, "A qualitative measurement survey of popular Internet-based IPTV Systems," in *Proc. Int. Conf. Commun. and Electron. (ICCE)*, Jun. 2008, pp. 156–161.
[19] "YouTube Blog". [Online]. Available: http://youtube.com/blog.
[20] G. K. Zipf, *The Psychobiology of Language*. Boston, MA, USA: Houghton-Mifflin, 1935.
[21] S. Milgram, "The small world problem," *Psychol. Today*, vol. 2, no. 1, pp. 60–67, 1967.
[22] D. J. Watts and S. H. Strogatz, "Collective dynamics of "Small-World" networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.
[23] E. Ravasz and A.-L. Barabási, "Hierarchical organization in complex networks," *Phys. Rev. E*, vol. 67, no. 2, p. 026112, Feb. 2003.
[24] R. Albert, H. Jeong, and A.-L. Barabási, "The diameter of the world wide web," *Nature*, vol. 401, pp. 130–131, Sep. 1999.
[25] J. Liu and J. Xu, "Proxy caching for media streaming over the Internet," *IEEE Commun. Mag.*, vol. 42, no. 8, pp. 88–94, Aug. 2004.
[26] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, Mar. 1999, vol. 1, pp. 126–134.
[27] S. Sen, J. Rexford, and D. F. Towsley, "Proxy prefix caching for multimedia streams," in *Proc. IEEE INFOCOM*, Mar. 1999, vol. 3, pp. 1310–1319.
[28] "YouTube's Chad Hurley: "We Have The Largest Library of HD Video on the Internet"," 2009. [Online]. Available: http://www.techcrunch.com/2009/01/30/youtubes-chad-hurley-we-have-the-largest-library-of-hd-video-on-the-internet.
[29] J. Liu, S. G. Rao, B. Li, and H. Zhang, "Opportunities and challenges of peer-to-peer Internet video broadcast," *Proc. IEEE*, vol. 96, no. 1, pp. 11–24, Jan. 2008.
[30] X. Zhang, J. Liu, B. Li, and T.-S. P. Yum, "CoolStreaming/DONet: A data-driven overlay network for peer-to-peer live media streaming," in *Proc. IEEE INFOCOM*, Mar. 2005, vol. 3, pp. 2102–2111.
[31] "PPLive". [Online]. Available: http://www.pplive.com.
[32] X. Cheng and J. Liu, "NetTube: Exploring social networks for peer-to-peer short video sharing," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 1152–1160.
[33] "PlanetLab". [Online]. Available: http://www.planet-lab.org.

Xu Cheng (S'08) received the B.Sc. degree from Peking University, Beijing, China, in 2006, and the M.Sc. degree from Simon Fraser University, Burnaby, BC, Canada, in 2008, both in computer science. He is currently a Ph.D. student and research assistant under the supervision of Dr. Jiangchuan Liu, at Simon Fraser University. His research interests include peer-to-peer networks, video streaming, wireless sensor networks and social networks. He is a Student Member of IEEE Communications. He will intern in the Department of Computing at Hong Kong Polytechnic University in 2010.

He was awarded SFU-CS Graduate Entrance Scholarship in 2006, CS Graduate Fellowship, Computing Science Travel and Minor Research Award, and Applied Science Conference Travel Support Award in 2008, SFU Graduate Fellowship and Graduate (International) Research Travel Award in 2010, at Simon Fraser University. He also won the BCNET Broadband Innovation Challenge in 2009.

Jiangchuan Liu (S'01–M'03–SM'08) received the B.Eng. degree (cum laude) from Tsinghua University, Beijing, China, in 1999, and the Ph.D. degree from The Hong Kong University of Science and Technology in 2003, both in computer science. He was a recipient of Microsoft Research Fellowship (2000), a recipient of Hong Kong Young Scientist Award (2003), and a co-inventor of one European patent and two US patents. He co-authored the Best Student Paper of IWQoS'08 and the Best Paper (2009) of IEEE Multimedia Communications Technical Committee (MMTC).

He is currently an Associate Professor in the School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, and was an Assistant Professor in the Department of Computer Science and Engineering at The Chinese University of Hong Kong from 2003 to 2004. He is a Visiting Associate Professor in the Department of Computing at Hong Kong Polytechnic University from 2009 to 2010.

His research interests include multimedia systems and networks, wireless ad hoc and sensor networks, and peer-to-peer and overlay networks. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, and an editor of IEEE Communications Surveys and Tutorials. He is a member of Sigma Xi.

Cameron Dale received his B.Sc. from Simon Fraser University in Burnaby, BC, Canada in 2000, majoring in Physics with a minor in Computing Science. After 5 years working in Fiber Optics, he returned to Simon Fraser University to receive his Master of Science in 2008, in the field of Computing Science.

He published papers in 4 conferences during his work on his Master's degree. One paper was awarded the Best Student Paper award at IWQoS in 2008. All are in the field of peer-to-peer networking.

He is currently an Application Developer working at IBM's Centre for Solution Excellence in Vancouver, BC, Canada.