Statistics and Social Network of YouTube Videos

Xu Cheng Cameron Dale Jiangchuan Liu
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
Email: {xuc, camerond, jcliu}@cs.sfu.ca

***Abstract*—YouTube has become the most successful Internet website providing a new generation of short video sharing service since its establishment in early 2005. YouTube has a great impact on Internet traffic nowadays, yet itself is suffering from a severe problem of scalability. Therefore, understanding the characteristics of YouTube and similar sites is essential to network traffic engineering and to their sustainable development.**

To this end, we have crawled the YouTube site for four months, collecting more than 3 million YouTube videos' data. In this paper, we present a systematic and in-depth measurement study on the statistics of YouTube videos. We have found that YouTube videos have noticeably different statistics compared to traditional streaming videos, ranging from length and access pattern, to their growth trend and active life span. We investigate the social networking in YouTube videos, as this is a key driving force toward its success. In particular, we find that the links to related videos generated by uploaders' choices have clear small-world characteristics. This indicates that the videos have strong correlations with each other, and creates opportunities for developing novel techniques to enhance the service quality.

I. Introduction

The Internet has witnessed an explosion of networked video sharing as a new killer application in the recent two years. Among them, *YouTube* [1] is the most successful one, with more than 100 million videos being watched every day [2]. The expensive deal by Google [3], as well as the success of similar sites (i.e., Tudou) further confirm the mass market interest. Their great achievement lies in the combination of rich media and, more importantly, their social networks. These sites have created a video community on the web, where anyone can be a star, from lip-synching teenage girls to skateboarding dogs [4]. With no doubt, they are reshaping popular culture and the way people surf the Internet.

Since its establishment in early 2005, YouTube has become one of the fastest-growing websites, and ranks second in traffic among all the websites in the Internet by the survey of Alexa [5]. It has a significant impact on the Internet traffic, but itself is suffering from severe scalability constraints. According to Alexa, the current speed of YouTube is "Slow" (average load time is 3.6 seconds) and is slower than 69% of the surveyed sites. Therefore, understanding the features of YouTube and similar video sharing sites is essential to network traffic engineering and to their sustainable enhancement.

To this end, we have crawled the YouTube site for a four-month period in early 2007, and have obtained 35 datasets totaling 3,269,030 distinct videos' information, which is the largest dataset crawled so far. In this paper, we present a systematic and in-depth measurement study on the statistics of YouTube videos. Using this data, we find that YouTube videos have noticeably different statistics from traditional streaming videos, in aspects from video length to access pattern. We also study some new features that have not been examined by previous measurement studies: the growth trend and active life span of videos.

More importantly, we also investigate the social networking aspect of YouTube, as this is the most unique part and a key driving force towards the success of YouTube and similar sites. In particular, we have found that the graph of YouTube's videos' structure has a clear small-world characteristic. This indicates that the videos have strong correlations with each other, and creates opportunities for developing novel techniques to enhance its service quality.

The remaining part is organized as follows. Section II gives a brief introduction of YouTube. Section III describes our method of gathering information for YouTube videos, and the dataset description is provided in Section IV. The statistics of YouTube videos are analyzed in Section V. We also analyze the social networking aspects and discuss the implications of the results in Section VI. Section VII presents some related works. Finally, Section VIII concludes the paper and notes some future directions.

II. YouTube Primer

A. Web 2.0 and UGC

The technique of *Web 2.0* is a trend in WWW technology, and it marks the new generation of web-based communities such as social networking sites, wikis and blogs, which aim to facilitate creativity, collaboration, and sharing among users. Web 2.0 boosts the development of social networking services, which build online social networks for communities of people who share interests and activities. Facebook, Flickr and YouTube are some notable social networking websites.

Online videos existed long before YouTube entered the scene. However, uploading, managing, sharing and watching videos were cumbersome due to the lack of an easy-to-use integrated platform. More importantly, the videos distributed by traditional media servers and peer-to-peer file downloads like BitTorrent were standalone units of content. The video was not connected to other related videos, for example other

This research was supported in part by a Canada NSERC Discovery Grant and an NSERC Strategic Project Grant.

978-1-4244-2084-1/08/$25.00 ©2008 IEEE

ID	ELhtXtnV3pg
Uploader	Alkarin
Date Added	May 19, 2007
Category	Entertainment
Video Length	268 seconds
Number of Views	596,272
Rating	4.83
Number of Ratings	1,227
Number of Comments	1,475
Related Videos	aUXoekeDIW8, 30MBljXxg3M, 4_Ow2wTuSHE, Sog2k6s7xVQ, ...

TABLE I
META-DATA OF A YOUTUBE VIDEO

episodes of a show that the user had just watched. Also, there was very little in the way of content reviews or ratings.

The new generation of video sharing sites, YouTube and its competitors, have overcome these problems. These new generation sites are also known as *user generated content* (UGC) sites, in which the users are participatory and creative. The systems allow content suppliers to upload video effortlessly, and to tag uploaded videos with keywords. Users can easily share videos by mailing links to them, or embedding them in blogs. Users can also rate and comment on videos, bringing new social aspects to the viewing of videos. Consequently, popular videos can rise to the top in a very organic fashion. The social network existing in YouTube further enables communities and groups, as videos are no longer independent from each other, and neither are users. This has substantially contributed to the success of YouTube and similar sites.

B. YouTube Techniques

YouTube video playback technology is based on Adobe Flash Player and uses the Sorenson Spark H.263 video codec with pixel dimensions of 320 by 240. This technology allows YouTube to display videos with quality comparable to more established video playback technologies (i.e., Windows Media Player, QuickTime and RealPlayer). YouTube officially accepts uploaded videos in .WMV, .AVI, .MOV and .MPG formats, which are converted into .FLV (Adobe Flash Video) format after uploading [6]. It has been recognized that the use of a uniform easily-playable format has been a key in the success of YouTube. Each YouTube video is accompanied by an HTML markup for embedding it in another page, unless the uploader chooses to disable this feature. This simple cut-and-paste feature is especially popular with users of social networking sites, and is also a key in the success of YouTube.

YouTube assigns each video a distinct 11-digit ID composed of 0-9, a-z, A-Z, - and _. Each video contains a series of meta-data: video ID, uploader, date when it was added, category, length, user rating, number of views, ratings and comments, and a list of "related videos". The related videos are links to other videos that have a similar title, description or tags, all of which are chosen by the uploader. A video can have hundreds of related videos, but the webpage only shows at most 20 at once. A typical example of the meta-data is shown in Table I.

III. CRAWLING YOUTUBE

We crawled the YouTube site and obtained information through a combination of the YouTube API and scrapes of YouTube video webpages. The results offer a series of representative partial snapshots of the YouTube video repository.

We consider all the YouTube videos to form a directed graph, where each video is a node in the graph. If a video b is in the related video list (first 20 only) of a video a, then there is a directed edge from a to b. Our crawler uses a breadth-first search to find videos in the graph. We define the initial set of 0-depth video 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 [7], which contains all the meta-data except date added, category and related videos. The crawler then scrapes the video's webpage to obtain the remaining information.

The first normal crawl was on February 22nd, 2007, and started with the initial set of videos from the list of popular videos. The crawl went to more than four depths, finding approximately 750 thousand videos. In the following weeks we ran the crawler every two to three days, finding on average 73 thousand distinct videos each time.

We also use the crawler to update the statistics of some previously found videos to study the growth trend of the video popularity. For this crawl we only retrieve the number of views for relatively new videos. This crawl is performed once a week from March 5th to April 16th 2007, which results in seven datasets.

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

Finally, we have collected the information about YouTube users. The crawler retrieves information on the number of uploaded videos and friends of each user from the YouTube API, for a total of more than 1 million users.

IV. ONLINE DATASET FOR DOWNLOADING

For the normal crawl, we have obtained 35 datasets totaling 3,269,030 distinct videos, from February 22nd to May 18th, 2007. The first crawl started with the initial set of videos from the list of "Recently Featured", "Most Viewed", "Top Rated" and "Most Discussed", for "Today", "This Week", "This Month" and "All Time", which totalled 189 unique videos on that day. For the rest of the normal crawl, the initial set is defined from the list of "Most Viewed", "Top Rated", and "Most Discussed", for "Today" and "This Week", which is about 200 to 300 videos. We crawled all the information listed in Table I, except that instead of the date added we recorded the age of the video (the days counting from Feb. 15, 2005).

For the updating crawl, we only update the statistics for the videos uploaded after February 15th, and only retrieve the statistics from the YouTube API. Specifically, we update the number of views, ratings and comments, as well as the rating of the video.

We have made all the crawled data available online. A more detailed description and the dataset can be found at `http://netsg.cs.sfu.ca/youtubedata.html`.

V. YouTube Videos' Statistics

The crawled data constitute a significant portion of the entire YouTube video repository (A search for "*" as a wildcard character in YouTube returns about 77.1 million videos). Also, since 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 statistics of YouTube videos.

In our measurements, some statistics are static and thus can be measured from the entire dataset (e.g., category, date added and length), some are dynamic that can change from dataset to dataset (e.g., number of views). We consider this dynamic information to be static over a single crawl. In addition, the updated number of views information will be used to measure the growth trend and active life span.

A. Video Category

The user can select from one of 12 categories when uploading the video. Figure 1 plots the distribution of all the categories. In our entire dataset we can see that the distribution is highly skewed: the most popular category is "Music", at about 22.9%; the second is "Entertainment", at about 17.8%; and the third is "Comedy", at about 12.1%. In the figure, we also plot two other categories: "Unavailable" are videos set to private, or videos that have been flagged as inappropriate video, which the crawler can only get information for from the YouTube API; "Removed" are videos that have been deleted by the uploader, or by a YouTube moderator (due to the violation of the terms of use), but are still linked to by other videos.

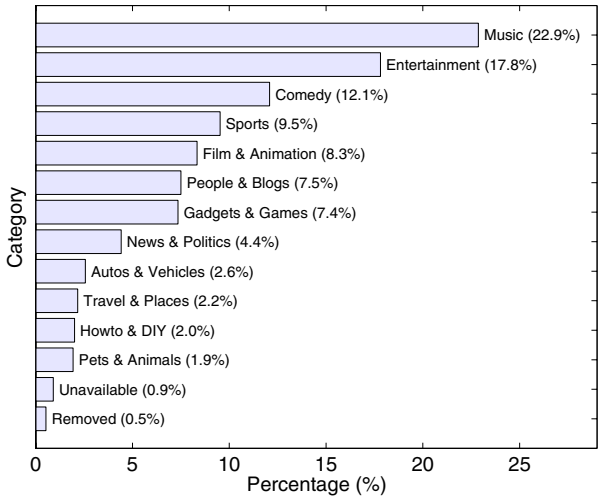

Fig. 1. Distribution of YouTube video categories

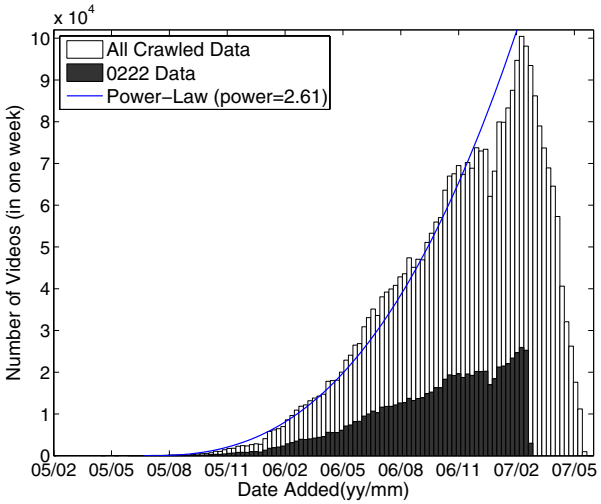

Fig. 2. Uploading trend of YouTube videos

B. Date Added

We record the age of each video in our crawl, so we can study the trend of YouTube uploading. Figure 2 shows the number of new videos added every week in our entire crawled dataset. Our first crawl was on February 22nd, 2007, and therefore we can get the early videos only if they are still very popular videos or are linked to by other videos we crawled. YouTube was established on February 15th, 2005, and we can see there is a slow start, as the earliest video we crawled was uploaded on April 27th, 2005. After 6 months from YouTube's establishment, the number of uploaded videos steeply increases. We use a power law curve to fit this trend.

We can see that in the dataset we collected, the number of uploaded videos decreases steeply starting in March, 2007. However, it does not imply that the uploading rate of YouTube videos has suddenly decreased. The reason is that many recently uploaded videos have not been so popular, and are probably not in other videos related videos' list. Since few videos link to those new videos, they are not likely to be found by our crawler. Nevertheless, as those videos become popular or get linked by others, our crawler may find them and get their information. We also see the same trend in the first and largest dataset crawled on February 22nd, comparing with the entire dataset.

C. Video Length

The most distinguished difference from traditional media content servers is the video length. Most traditional servers contain a small to medium number of long videos, typically 1-2 hour movies (e.g., HPLabs Media Server [8]), whereas YouTube is mostly comprised of short video clips.

We have found that 97.9% video lengths are within 600 seconds, and 99.1% are within 700 seconds in our entire dataset. This is mainly due to the limit of 10 minutes imposed by YouTube on regular users uploads. We do find videos

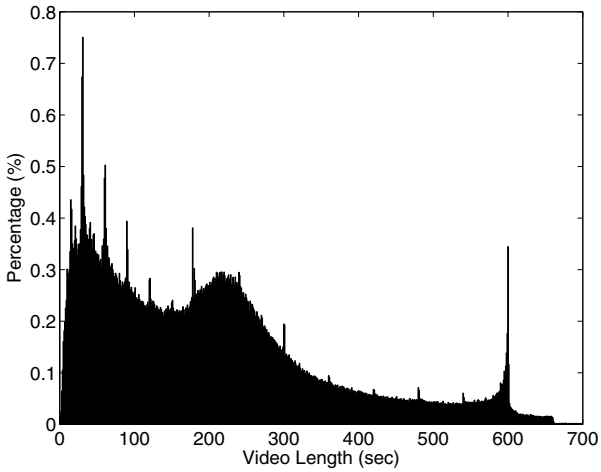

Fig. 3. Distribution of YouTube video length

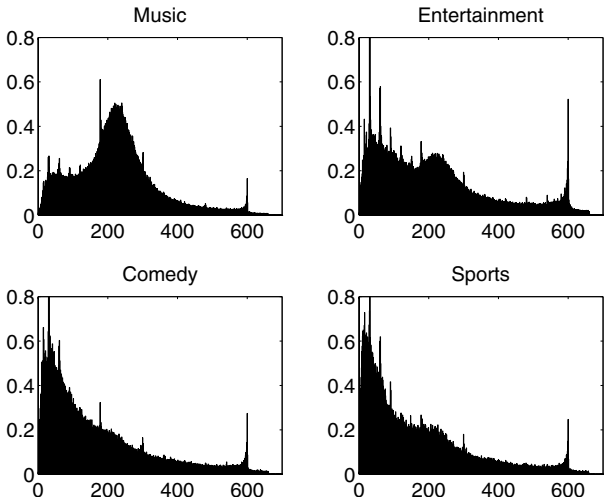

Fig. 4. Distribution of YouTube video length for the four most popular categories

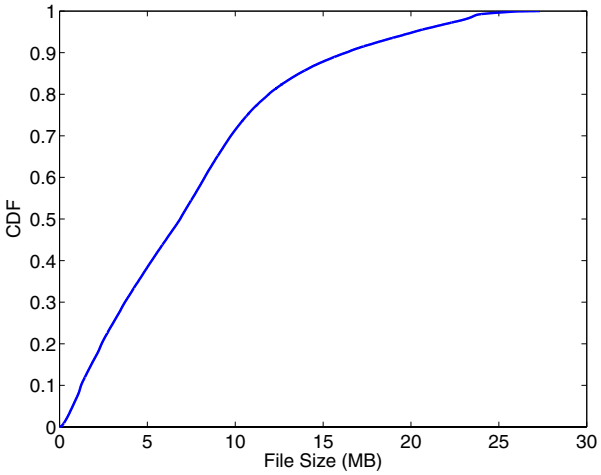

Fig. 5. CDF of YouTube video file size

longer than this limit, because the limit was only established in March, 2006, and also the YouTube Director Program allows a small group of authorized users to upload longer videos [9].

Figure 3 plots the distribution of YouTube video lengths within 700 seconds, which exhibits three main peaks. The first peak is within one minute, containing more than 20.6% of the videos; it shows that YouTube is primarily a site for very short videos. The second peak is between 3 and 4 minutes, containing about 17.1% of the videos; this peak is mainly caused by the large number of videos in the "Music" category, since "Music" is the most popular category, and the typical length of a "Music" video is often within this range (shown in Figure 4). The third peak is near 10 minutes; it is caused by the limit on the uploading video length, which encourages some users to circumvent the length restriction by dividing long videos into several parts, each being near the limit of 10 minutes. We can also observe that there are peaks at around every exact minute.

Figure 4 shows the video length distributions for the top four most popular categories. We can see "Music" videos have a very large peak between three and four minutes (about 27.6%) due to the typical music video length. "Entertainment" videos have the greatest peak at around 10 minutes comparing to other categories, probably because most 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. Probably corresponding to "highlight" type clips, "Comedy" and "Sports" videos have much more videos within two minutes, about 50.9% and 50.4% respectively.

D. Video File Size and Video Bitrate

We also retrieved the file size of about 184 thousand videos, using video IDs from a normal crawl. Not surprisingly, we find that the distribution of video file sizes is very similar to that of video lengths. We plot the cumulative distribution function (CDF) of YouTube video file size in Figure 5. We find that in our crawled data, 98.3% of the videos are less than 25 MB, and the average size is 8.4 MB, which is quite small compared to hundreds of MB movies. However, considering there are over 77.1 million YouTube videos, the total disk space required to store all the videos is nearly 650 TB! Smart storage management is thus quite demanding for such an ultra-huge and still growing site, which we will discuss in Section VI-C.

Among the videos we crawled, we found that 87.6% contain FLV meta-data specifying the video's birate. For the rest of the videos that do not contain this meta-data, we calculate the average bitrate from the file sizes and their lengths. In Figure 6, we observe that the videos' bitrate has three clear peaks. Most videos have a bitrate around 330 kbps (70.6% video bitrate are between 300 kbps and 360 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 streaming rate.

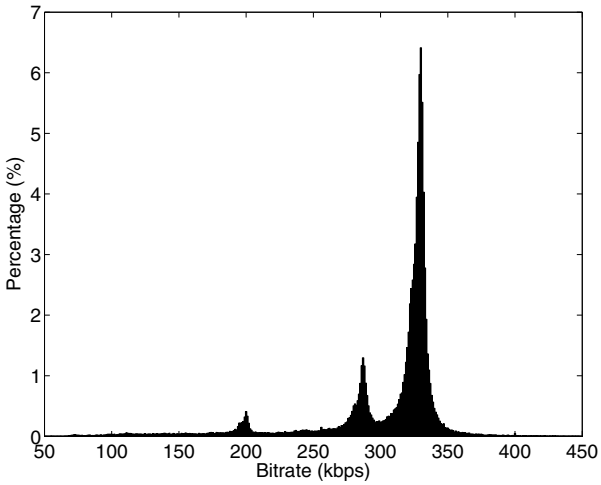

Fig. 6. Distribution of YouTube video bitrate

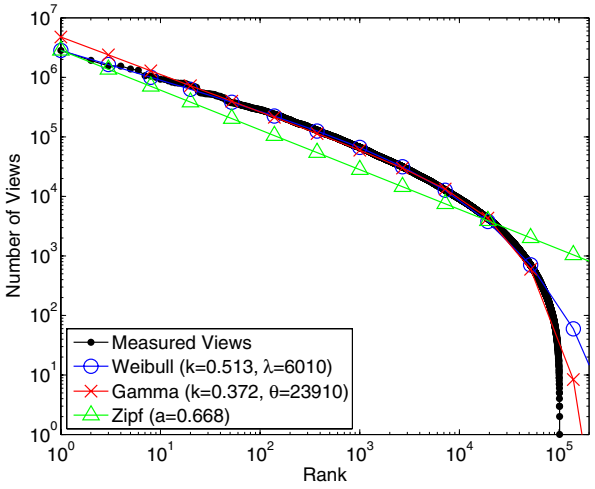

Fig. 7. YouTube videos rank ordered by popularity

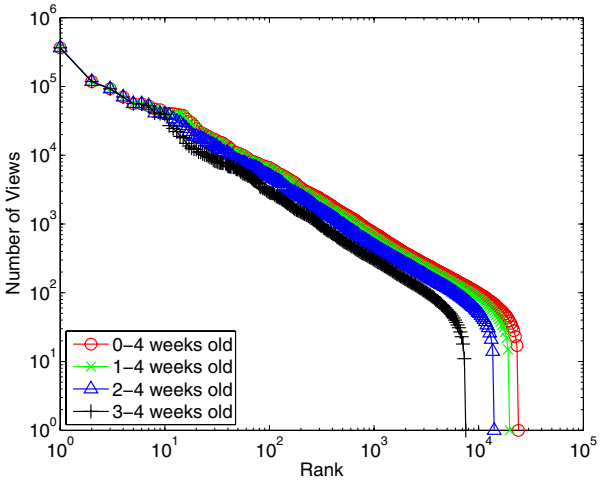

Fig. 8. Recently added YouTube videos rank by popularity

E. Views – Video Popularity and User Access Pattern

The number of views a video has had is the most important statistic we measured, as it reflects the video popularity and user access pattern. Since this statistic is changing over time, we cannot use the entire dataset to measure. We use a single dataset crawled on April 3rd, 2007, containing more than 100 thousand videos. We plot the number of views as a function of the rank of the video by its popularity in Figure 7.

Although the plot has a long tail on the linear scale, it does NOT follow a Zipf distribution, which should be a straight line on a log-log scale. This is consistent with some previous observations [8], [10], [11], [12] which also found that video accesses on a media server does not follow Zipf's law. We can see in the figure, the beginning of the curve is linear, but the tail (after 2×10^3 videos) decreases tremendously, indicating there are not so many unpopular videos as Zipf's law predicts. One explanation is that users are likely to access their own videos several times after uploading it to ensure the videos are uploaded successfully, although the videos will possibly be never accessed again.

This result seems consistent with some results [12], which also have a heavy tail. However, it differs from others [8], [10], [11], in which the curve is skewed from linear from beginning to end. Their results indicate that the popular videos are also not as popular as Zipf's law predicts, which is not the case in our measurement. To fit the skewed curve, some use a concatenation of two Zipf distributions [11], and some use a generalized Zipf distribution [8]. We have used two other distributions, Weibull and Gamma, with a Zipf distribution to compare with. We find that both distributions fit better than Zipf, due to the heavy tail that they have.

We were initially concerned that the crawled data might be biased, as popular videos may be more likely to appear in our BFS crawl than non-popular ones. Since the entire video name space is too large (2^{66}), random sampling directly can be quite difficult. Therefore, we have 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 of these videos, and plot in Figure 8. The leftmost (magenta) plot is only the videos added during the first week (all the 3-4 weeks old videos), while the rightmost (red) plot contains all the videos (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.

We also examine the correlation between video age and number of views, which is shown in Table II. Not surprisingly the video age affects the number of views, because older

age	count	min	max	mean	std
≤ 1 month	22662	0	1556837	2244	17852
1–3 month	22939	0	1922218	2450	17279
3–6 month	24980	0	934062	4163	18639
6–12 month	27570	3	1051432	7670	26165
> 12	3195 month	0	2839893	19095	79142

TABLE II
CORRELATION BETWEEN VIDEO AGE AND NUMBER OF VIEWS

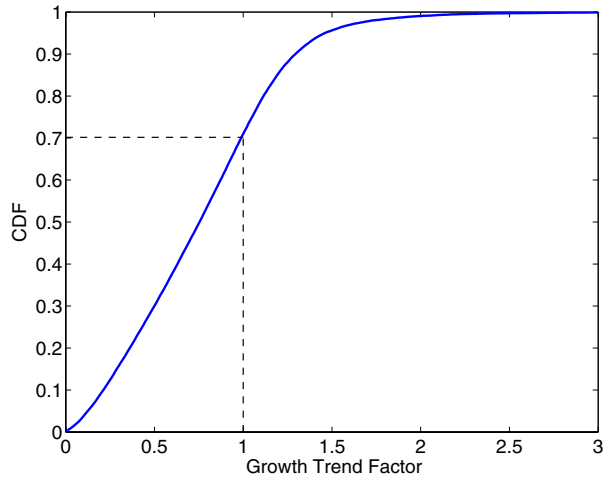

Fig. 9. Distribution of YouTube video growth trend factor

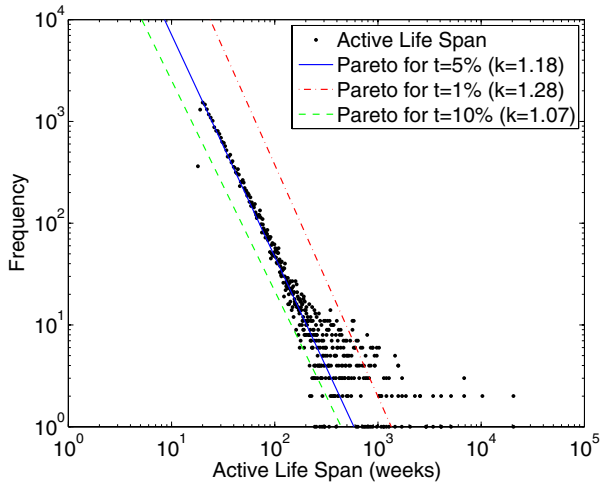

Fig. 10. Distribution of estimated active life span

videos have more opportunity to be accessed. However, we can see in the younger video groups that 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*, making popular videos more popular, and unpopular videos less popular.

F. Growth Trend and Active Life Span

From above, we know that some videos are very popular (their number of views grows very fast), while others are not. Also, after a certain period of time, some videos are almost never accessed. Starting from March 5th, 2007, we updated the number of views of relatively new videos every week for seven weeks. We eliminate the videos that have been removed and thus do not have the full seven data points to ensure the growth trend will be modeled properly, resulting in approximately 43 thousand videos.

To model the growth trend, we have found that a power low can fit better than a linear fit. Therefore, a video can have a increasing growth (if the power is greater than 1), a constant growth trend (power near 1), or a slowing growth (power less than 1). The trend depends on the exponent, which we call the growth trend factor p. We model the number of views after x weeks as

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

where μ is the number of weeks before March 5th that the video has been uploaded, x indicates the week of the crawled data (from 0 to 6), and v_0 is the number of views the video had in the first dataset.

Using Equation 1, we have modeled the 43 thousand videos to get the growth trend factors p, the CDF of which is plotted in Figure 9. We observe that 70% of the videos have a growth trend factor that is less than 1, indicating that most videos grow more and more slowly as time passes.

It is known that YouTube has no policy on removing videos after a period of time or when their popularity declines, so the life span of a YouTube video is almost infinite. However, the video's popularity may grow more and more slowly, and will almost stop growing after some time, which we define as the video's *active life span*. From this active life span, we can extract the feature of a video's temporal locality.

If a video's number of views increases by a factor less than t from the previous week, we define the video's active life span to be over. We prefer this relative comparison to an absolute comparison, since we are only concerned with the shape of the curve instead of the scale. For each video that has a growth trend factor p less than 1, we can compute its active life span l from

$$\frac{v(l)}{v(l-1)} - 1 = t \tag{2}$$

which can be solved for the active life span

$$l = \frac{1}{\sqrt[p]{1+t}-1} + 1 - \mu. \tag{3}$$

Therefore we can see that the active life span depends on the growth trend factor p and the number of weeks the video has been on YouTube, but is independent on the number of views the video had at beginning.

We plot the the probability density function (PDF) for the active life span of the approximately 30 thousand videos (with p less than 1), for a life span factor of $t = 5\%$ in Figure 10. The data is well fitted by a Pareto distribution. From looking at multiple fits with various values of t, we find that they all result in similar parameter k. We also include the Pareto fits for $t = 1\%$ and $t = 10\%$ for comparison.

Since we do not have access to the YouTube server traces, it is difficult to accurately measure a video's temporal locality, which would show whether recently accessed videos are likely to be accessed in the near future. However, the active life span gives us another way to estimate the temporal locality of YouTube videos. The Pareto distribution implies that most videos have a short active life span, which means the videos have been watched frequently in a short span of time, and

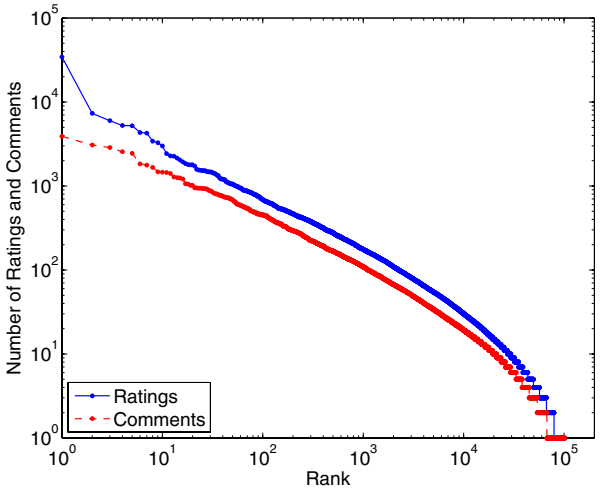

Fig. 11. YouTube videos ratings and comments ranks by the numbers of ratings and comments

	min	max	mean	median	std	zeros
views	0	2839893	4771	741	24892	0.1%
ratings	0	34401	14.5	3	128	21.6%
comments	0	3899	8.7	2	39	33.6%

TABLE III
FACTS OF VIEWS, RATINGS AND COMMENTS

then fewer and fewer people will access them after the video's active life span is complete. This characteristic has good implications for caching and storage. We can design a predictor to forecast the active life span using our model, which can help a proxy or server to make more intelligent decisions. We will discuss this further in Section VI-C.

G. User Behavior

User behavior is also an important feature that is worth studying. We first study the statistics of number of ratings and comments from the same dataset as we did number of views. Whenever a video webpage has been accessed, the number of views increases, whereas users need to log in to rate and comment so that the number of ratings and comments will increase. Therefore, the number of ratings and comments better reflect the user behavior. Figure 11 plots the number of ratings against the rank, and similarly 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.

We list the statistics of ratings and comments in Table III, along with views for comparison. We can see that comments are fewer than ratings, and both are much fewer than views. In fact, a great number of videos do not have a single rating or comment. This indicates that users are more willing to watch videos rather than to log in to rate and make comments.

YouTube currently has about 2.86 million registered users.[1] Users need to login to upload video or watch some limited videos. A user can add another user to their friend list so that it is convenient to watch their friends' videos. From the crawl of user information we performed on May 28th, 2007, we can extract two statistics of YouTube users: the number of uploaded videos and the number of friends. We did this for more than 1 million users found by our crawler in all the crawls performed before this one.

YouTube is a typical UGC website, in that all the videos are uploaded by the users. These active users are a key in the success of YouTube. We have found that many users have uploaded a few videos, and a small number of users have uploaded many videos. Since we collected the user IDs from the previous crawled data, all of these users have uploaded at least one video. However, the information of users that do not upload videos has not been crawled. We believe that there are a huge number of users that have not uploaded a single video.

We calculate the average and median number of friends each user has to be 4.3 and 0, respectively. Interestingly, in over 1 million users data, we have found that 58% of the users have no friends. Therefore, we can see that having friends does not affect the access pattern, so that the correlation between users is not very strong. Thus the social networking existing among users has less impact than that among videos, as we will study in the next section.

VI. THE SOCIAL NETWORK OF YOUTUBE

YouTube is a prominent social media application. Communities and groups exist in YouTube, there are statistics and awards for videos and personal channels, and so videos are no longer independent from each other. It is therefore important to understand the social networking characteristics of YouTube. We next examine the social networking among YouTube **videos**, which is a very unique and interesting aspect of this kind of video sharing sites, as compared to traditional media services.

A. Small-World Phenomenon

One of the most interesting characteristic for social network is the *small-world* phenomenon, which was first introduced by Milgram [13] to refer to the principle that people are linked to all others by short chains of acquaintances (AKA six degrees of separation). It has been found in various real-world situations: URL links in the Web [14], Gnutella's search overlay topology [15], and Freenet's file distribution network [16]. This formulation was used by Watts and Strogatz to describe networks that are neither completely random, nor completely regular, but possess characteristics of both [17]. They introduce a measure of one of these characteristics, the cliquishness of a typical neighborhood, as the *clustering coefficient* of the graph. They define a small-world graph as one in which the clustering coefficient is still large, as in regular graphs, but the measure of the average distance between nodes (the *characteristic path length*) is small, as in random graphs.

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

[1] A channel search for "*" as a wildcard character in YouTube returns about 2.86 million channels, of which a registered user has one.

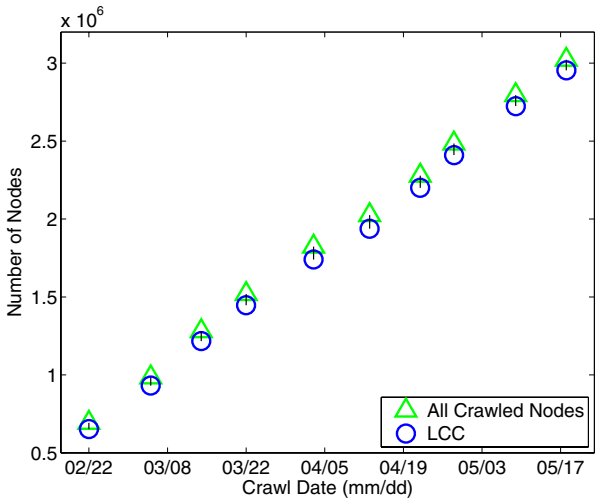

Fig. 12. Size of cumulative datasets

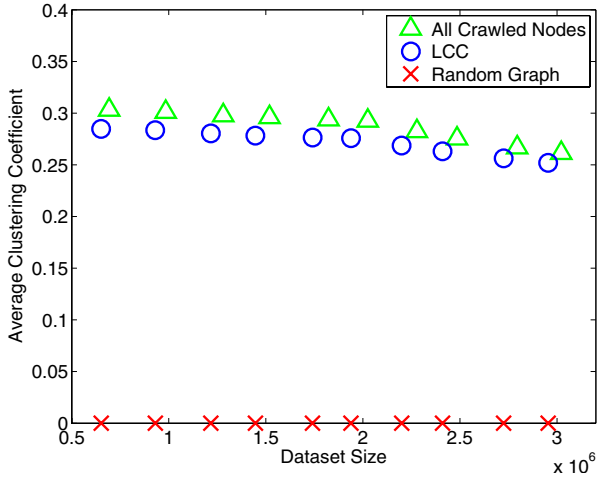

Fig. 13. Clustering coefficient of cumulative datasets

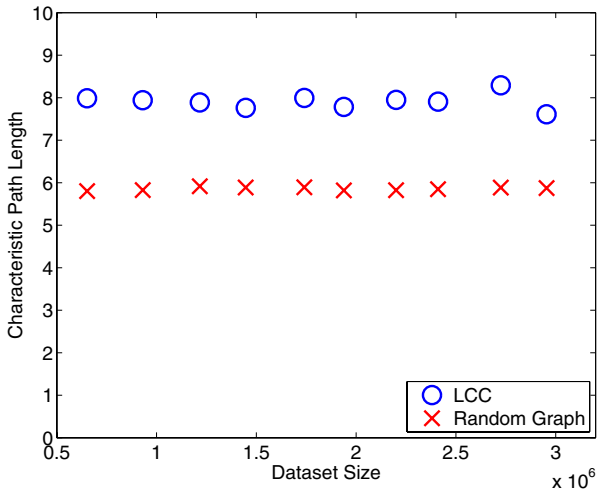

Fig. 14. Characteristic path length of cumulative datasets

possible edges between neighbors of the node that actually exist in the graph. The clustering coefficient of the graph $C(G)$ is then the average of the clustering coefficients of all nodes in the graph. 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 from node i. The characteristic path length of the graph $D(G)$ is then the average of the characteristic path lengths of all nodes in the graph.

B. The Small-World in YouTube

We have obtained ten cumulative datasets, each consisting of all the previously crawled data. The last one thus contains all the data we crawled. We select the ten datasets such that the increment between each consecutive two is similar (about 270 thousand). We measured the graph topology by using the related links in YouTube pages to form directed edges in a video graph for each dataset. Videos that have no outgoing or no incoming links are removed from the analysis.

The graphs are not strongly connected, making it difficult to calculate the characteristic path length. We thus also use the *Largest Strongly Connected Component* (LCC) 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 as the LCC of the crawled dataset, and is similarly limited to a maximum out-degree of 20. Figure 12 shows the sizes of the ten cumulative datasets and their LCC graphs, which are very close to the total dataset sizes.

Figure 13 shows the clustering coefficient for the graph as a function of the dataset size. The clustering coefficient is quite high (around 0.29), especially in comparison to the random graphs (nearly 0). There is a slow drop for larger datasets, showing that there is some inverse dependence on the size of the graph, which is common for some small-world networks [18]. Figure 14 shows the characteristic path length for the graphs. We can see that the characteristic path length (about 8) is only slightly larger than that of a corresponding random graph (about 6), which is quite good considering the still large clustering coefficient of these datasets.

The graph formed by YouTube's related videos has definite small-world characteristics. The clustering coefficients are very large compared to the corresponding random graphs, while the characteristic path lengths are approaching the shortest path lengths measured in the corresponding random graphs. This finding is expected, due to the user-generated nature of the tags, title and description of the videos that is used by YouTube to find related ones.

The results are similar to other real-world user-generated graphs, though 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 [14]. This could possibly be due to the larger number of nodes (8×10^8), but it also indicates that the YouTube network of videos is a much closer group.

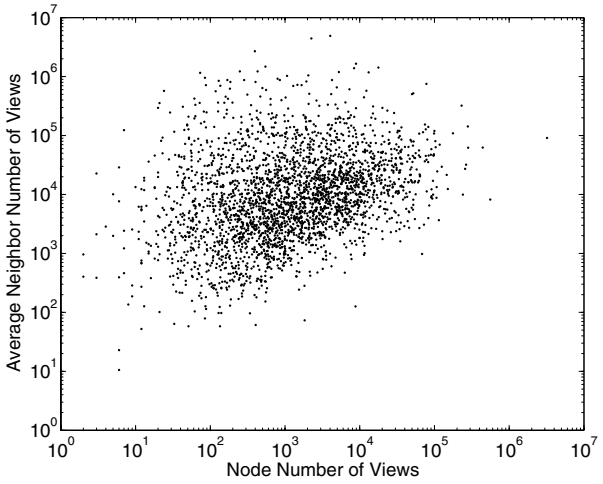

Fig. 15. Correlation of average neighbor views and views

C. Exploring Social Networking for Caching and Storage

According to the study by Alexa [5], the current speed of YouTube is "Slow" and is slower than 69% of the surveyed sites. Caching frequently accessed video files at proxies close to clients is an effective way to prevent users from experiencing excessive access delays. Many strategies have been developed for caching web objects or streaming videos [19]. While we believe that YouTube will benefit from proxy caching, three distinct features call for novel cache designs: first, the number of YouTube videos (77.1 million) is orders of magnitude higher than that of traditional video services (e.g., HPC: 2999, HPL: 412 [8]); second, the YouTube video size is much smaller than a traditional video (a YouTube video of average 8.4 MB versus a typical MPEG movie of 700 MB); and finally, the popularity of YouTube videos does not fit a Zipf distribution, which has important implications on web caching [20].

Considering these factors, prefix caching [21] is probably the best choice. Assume for each video, the proxy will cache a 5 second initial clip (about 200 KB) of the video. Given the Weibull distribution of popularity suggested by our measurements, we have calculated the hit-ratio as a function of the cache size. Assuming the cache is devoted only to the most popular videos, to achieve a 90% hit-ratio the proxy would require about 6 GB of space for the current YouTube video repository, which is acceptable for today's proxy servers.

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 within the group after finishing the first one. This expectation is confirmed by Figure 15, which shows a clear correlation between the number of views and the average of the neighbor videos' views. Once a video is played and cached, the prefixes of its directly related videos can also be pre-fetched and cached.

A remaining issue is when to release the cached prefix due to the limited cache size. We know that the active life span of YouTube videos follows a Pareto distribution, implying that most videos are popular during a relatively short span of time. Therefore, a predictor can be developed to estimate the active life span of a video. The proxy can thus decide which videos have already passed their active life span, and replace them if the cache space is insufficient.

The predictor can also facilitate disk space management on the YouTube server. Videos on the YouTube server will not be removed by the operator unless they violate the terms of service. With a daily $65,000$ new videos introduced [2], the server storage will soon become a problem. A hierarchical storage structure can be built with videos passing their active life span being moved to slower and cheaper storage media.

VII. Related Work

A. Workload Measurement of Traditional Media Servers

There has been a significant research effort into understanding the workloads of traditional media servers, looking at, for example, the video popularity and access locality [10], [11], [8], [12]. We have found that, while sharing similar features, many of the video statistics of these traditional media servers are quite different from YouTube, i.e., the video length and active life span. More importantly, these traditional studies lack a social networking among the videos.

A similar work to ours is the study by Huang et al. [22]. They analyzed a 9-month trace of MSN Video, Microsoft's VoD service, examining the user behavior and popularity distribution of videos. This analysis led to a peer-assisted VoD design for reducing the server's bandwidth costs. The difference to our work is that MSN Video is a more traditional video service, with much fewer videos, most of which are longer than YouTube videos. MSN Video also has no listings of related videos or user information, and thus no social networking aspect.

B. Workload Measurement of New Generation Media Servers

We have seen simultaneous works investigating YouTube and similar Web 2.0 sites in the past two years. The authors in [23] were the first to study the social networking aspect in YouTube. Mislove et al. studied four online social networking sites (Flickr, YouTube, LiveJournal and Orkut) [24], and confirmed the power-law, small-world and scale-free properties of online social networks. Cha et al. studied YouTube and Daum UCC, the most popular UGC service in Korea [25], and also proposed some improvement for UGC design. A YouTube traffic analysis is presented in [26], which tracks YouTube transactions in a campus network, and focused on deriving video access patterns from the network edge perspective. In [27], Paolillo found the social core appearing in YouTube. Zink et al. obtained the trace of YouTube traffic and investigated the caching problem [28].

Different from [23], [24], [27], we investigate the social networking among the videos instead of users. Since people are not required to register and upload videos, the social networking among the users has less impact than that among the videos, and the social networking among the videos

provides us the opportunity for improving the system. Our work complements works in [25], [26], [28] by crawling a much larger set of the videos and thus being able to accurately measure their global properties and, in particular, the social network which they did not consider in their works.

VIII. Conclusion and Future Work

In this paper, we have presented a detailed investigation of the characteristics of YouTube, the most popular Internet short video sharing site to date. Through examining the more than 3 million YouTube video data collected in a four month period, we have demonstrated that, while sharing certain similar features with traditional videos, YouTube exhibits many unique statistics, especially in length distribution. 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-generated content, have both small-world characteristics of a large clustering coefficient indicating the grouping of videos, and a short characteristic path length linking any two videos.

An official report in June 2006 reveals the YouTube's outbound bandwidth is about 20 Gbps with a 20% growth rate per month [29]. Assuming the network traffic cost is \$10 per Mbps, the estimated YouTube transit expense is currently more than \$11 million per month. This high and rising expense for network traffic is probably one reason YouTube was sold to Google. We suggested that the social networks presented among YouTube videos can be explored to enhance the scalability and QoS of YouTube.

Another possible solution is to utilize the peer-to-peer (P2P) technique, which has been quite successful in supporting large-scale live video streaming (e.g., PPLive and CoolStreaming) and on-demand streaming (e.g., GridCast) [30], [31]. Yet the unique characteristics of YouTube videos make P2P delivery quite challenging. In particular, the ultra large quantity of the videos and the short length of each video render the per-video overlay, which has been almost exclusively used in existing systems, suboptimal if not entirely impractical. As an example, for a PPLive or a CoolStreaming client, the initialization time for joining a P2P overlay and locating partners can be over half minute, which is longer than many of the short video clips. We believe that, to guarantee acceptable QoS, a hybrid implementation with complementary servers, proxy caches, and client peers will be an viable solution, and the social networks can again be explored in this context. We are currently implementing a social network based P2P system that assist the servers for short video sharing, and our preliminary results demonstrate that this design effectively addresses the aforementioned challenges [32].

References

[1] "YouTube," http://www.youtube.com.

[2] "YouTube serves up 100 million videos a day online," http://www.usatoday.com/tech/news/2006-07-16-youtube-views_x.htm.

[3] "Google to buy YouTube for \$1.65 billion," http://money.cnn.com/2006/10/09/technology/googleyoutube_deal/index.htm.

[4] "YouTube video-sharing site is changing popular culture," http://www.kcrw.com/news/programs/ww/ww061122youtube_video-sharin.

[5] "Alexa," http://www.alexa.com.

[6] "YouTube: Video Format (Wikipedia)," http://en.wikipedia.org/wiki/Youtube#Video_format.

[7] "API Documentation (YouTube)," http://youtube.com/dev_docs.

[8] W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat, "Long-term Streaming Media Server Workload Analysis and Modeling," HP Labs, Tech. Rep., 2003.

[9] "YouTube Blog," http://youtube.com/blog.

[10] S. Acharya, B. Smith, and P. Parnes, "Characterizing User Access To Videos On The World Wide Web," in *Proc. of ACM/SPIE Multimedia Computing and Networking*, 2000.

[11] J. M. Almeida, J. Krueger, D. L. Eager, and M. K. Vernon, "Analysis of Educational Media Server Workloads," in *Proc. of ACM NOSSDAV*, 2001.

[12] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding User Behavior in Large-Scale Video-on-Demand Systems," *ACM SIGOPS Operating Systems Review*, vol. 40, no. 4, pp. 333–344, 2006.

[13] S. Milgram, "The Small World Problem," *Psychology Today*, vol. 2, no. 1, pp. 60–67, 1967.

[14] R. Albert, H. Jeong, and A.-L. Barabási, "The Diameter of the World Wide Web," *Nature*, 1999.

[15] G. Liu, M. Hu, B. Fang, and H. Zhang, "Measurement and Modeling of Large-Scale Peer-to-Peer Storage System," in *Proc. of Grid and Cooperative Computing Workshops*, 2004.

[16] T. Hong, "Performance," in *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*. O'Reilly & Associates, Inc., 2001, ch. 14, pp. 203–241.

[17] D. J. Watts and S. H. Strogatz, "Collective Dynamics of "Small-World" Networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[18] E. Ravasz and A.-L. Barabási, "Hierarchical Organization in Complex Networks," *Physical Review E*, vol. 67, no. 2, p. 026112, 2003.

[19] J. Liu and J. Xu, "Proxy Caching for Media Streaming over the Internet," *IEEE Communications Magazine*, vol. 42, no. 8, pp. 88–94, 2004.

[20] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," in *Proc. of IEEE INFOCOM*, 1999.

[21] S. Sen, J. Rexford, and D. F. Towsley, "Proxy Prefix Caching for Multimedia Streams," in *Proc. of IEEE INFOCOM*, 1999.

[22] C. Huang, J. Li, and K. W. Ross, "Can Internet Video-on-Demand be Profitable?" in *Proc. of SIGCOMM*, 2007.

[23] M. Halvey and M. Keane, "Exploring Social Dynamics in Online Media Sharing," in *Proc. of the WWW Poster Paper*, 2007.

[24] A. Mislove, M. Marcon, K. P. Gummadi, P. Dreschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *Proc. of ACM IMC*, 2007.

[25] 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. of ACM IMC*, 2007.

[26] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube Traffic Characterization: A View From the Edge," in *Proc. of ACM IMC*, 2007.

[27] J. Paolillo, "Structure and Network in the YouTube Core," in *Proc. of Hawaii International Conference on System Sciences (HICSS)*, 2008.

[28] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Watch Global, Cache Local: YouTube Network Traffic at a Campus Network - Measurements and Implications," in *Proc. of SPIE/ACM Multimedia Computing and Networking (MMCN)*, 2008.

[29] C. Corbett, *Peering of Video*, YouTube, 2006. [Online]. Available: http://www.nanog.org/mtg-0606/pdf/bill.norton.3.pdf

[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. of IEEE INFOCOM*, 2005.

[31] J. Liu, S. G. Rao, B. Li, and H. Zhang, "Opportunities and Challenges of Peer-to-Peer Internet Video Broadcast," *Proceedings of the IEEE*, vol. 96, no. 1, pp. 11–24, 2008.

[32] X. Cheng and J. Liu, "Social Network Based Peer-to-Peer Short Video Sharing," Simon Fraser University, Tech. Rep., 2008.