9

Characteristics and Potentials of YouTube: A Measurement Study

Xu Cheng, Cameron Dale, Jiangchuan Liu

Simon Fraser University

Introduction

Established in 2005, YouTube is one of the fastest-growing websites, and has become one of the most accessed sites in the Internet. It has a significant impact on the Internet traffic distribution, but itself is suffering from severe scalability constraints. Understanding the features of YouTube and similar video sharing sites is thus crucial to network traffic engineering and to sustainable development of this new generation of services.

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 3-month period in early 2007, and obtained more than 2 million distinct videos. This constitutes a significant portion of the entire YouTube video repository. Using this collection of datasets, we find that YouTube videos have noticeably different statistics from traditional streaming videos, such as video length.

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 uploaders' choices form a small-world network. This suggests that the videos have strong correlations with each other, and creates opportunities for developing novel Peer-to-Peer distribution schemes to efficiently deliver videos to end users.

E.M. Noam and L.M. Pupillo (eds.), *Peer-to-Peer Video*, doi: 10.1007/978-0-387-76450-4_9,

The rest of the paper is organized as follows. Next is a section presenting some background information and other related work. Following is a section which first describes our method of gathering information about YouTube videos, which is then analyzed generally, while the social networking aspects are analyzed separately in the subsequent section. The last section discusses the implications of the results, and suggests ways that the YouTube service could be improved. Finally, we draw our conclusions.

Background and Related Work

Internet Video Sharing

Online videos existed long before YouTube entered the scene. However, uploading videos, managing, sharing, and watching them was very cumbersome due to a 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. Each single video was not connected in any way to other related video clips, for example, to other 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, formed by YouTube and its competitors, has overcome these problems as they allow content suppliers to upload videos effortlessly, automatically converting them from many different formats, and to tag uploaded videos with keywords. Users can easily share videos by mailing links to them, or embedding them on web pages or in blogs. Users can also rate and comment videos, bringing new social aspects to the viewing of videos. Consequently, popular videos can rise to the top in a very organized fashion.

The social network existing in YouTube further enables the development of communities and groups. Videos are no longer independent from each other, and neither are users. This has substantially contributed to the success of YouTube and similar sites.

Workload Measurement of 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 [2][8]. We have found that, while sharing similar features, many of the video statistics of these traditional media

servers are quite different from those of YouTube; for example, the video length distribution. More importantly, these traditional studies lack a social network among the videos.

A similar work to ours is the study by Huang et al. [5]. 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 far less videos, most of which are also longer than YouTube videos. MSN Video also has no listings of related videos or user information, and thus no social networking aspect.

We have seen simultaneous works investigating social networks in popular Web 2.0 sites, including Flicker, Orkut, and LiveJournal [7]. While YouTube is also one of the targeted sites in their studies, a thorough understanding of the unique characteristics of short video sharing has yet to be gained, particularly considering that YouTube has a much higher impact. Recently, a YouTube traffic analysis was presented which tracks YouTube transactions in a campus network [4]. The research focus was on deriving video access patterns from the network edge perspective. Our work complements it by crawling a much larger set of the videos and thus being able to accurately measure their global properties, and in particular, the social networks.

Characteristics of YouTube Video

In this paper, we focus on the access patterns and social networks in YouTube. To this end, we crawled the YouTube site for a 3-month period and obtained information on its videos through a combination of the YouTube API and scrapes of YouTube video web pages. The results offer a series of representative partial snapshots of the YouTube video repository.

Methodology of Measurement

Video Meta-data

YouTube randomly assigns each video an 11-digit ID. Each video contains the following intuitive meta-data: user who uploaded it, date when it was uploaded, category, length, number of views, number of ratings, number of comments, and a list of "related videos." The related videos are links to

other videos that have a similar title, description, or related 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, so we limit our scrape to these top 20 related videos. A typical example of the meta-data is shown in Table 9.1.

Table 9.1 Meta-data of a YouTube video

ID	2AYAY2TLves
Uploader	GrimSanto
Added Date	May 19, 2007
Category	Gadgets & Games
Video Length	268 seconds
Number of Views	185,615
Number of Ratings	546
Number of Comments	588
Related Videos	aUXoekeDIW8, Sog2k6s7xVQ, …

YouTube Crawler

We consider all the YouTube videos to form a directed graph, where each video is represented by a node in the graph. If video *b* is in the related video list (only among the first 20) of 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.

Our first crawl was carried out on February 22, 2007, and found approximately 750 thousand videos in about 5 days. In the following weeks we ran the crawler every two to three days. On average, the crawl found 80,000 distinct videos each time. We also crawled other statistics such as the file size and bitrate information. By the end of April 2007, we had obtained 27 datasets totaling 2,676,388 distinct videos. This constitutes a significant portion of the entire YouTube video repository.[1] Also, 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.

Video Category

One of twelve categories is selected by the user when uploading the video. Table 9.2 lists the count numbers and percentages of all the categories. In our entire dataset we note that 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%.

Table 9.2 List of YouTube video categories

Category	Count	%
Autos and Vehicles	66,878	2.5
Comedy	323,814	12.1
Entertainment	475,821	17.8
Film and Animation	225,817	8.4
Gadgets and Games	196,026	7.3
Howto and DIY	53,291	2.0
Music	613,754	22.9
News and Politics	116,153	4.3
People and Blogs	199,014	7.4
Pets and Animals	50,092	1.9
Sports	258,375	9.7
Travel and Places	58,678	2.2
Unavailable	24,068	0.9
Removed	14,607	0.5

In the table, we also list two other categories. “Unavailable” are the videos set to private, or videos that have been flagged as inappropriate, which the crawler can only get information for from the YouTube API, whilst “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.

Video Length

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

In our entire dataset, 97.8% of the videos last less than 600 s, and 99.1% are under 700 s. This is mainly due to the limit of 10 min imposed by YouTube on uploads by regular users. We do find videos longer than this limit though, as the limit was only established in March 2006, and also the YouTube Director Program allows a small group of authorized users to upload videos that are longer than 10 min.[2]

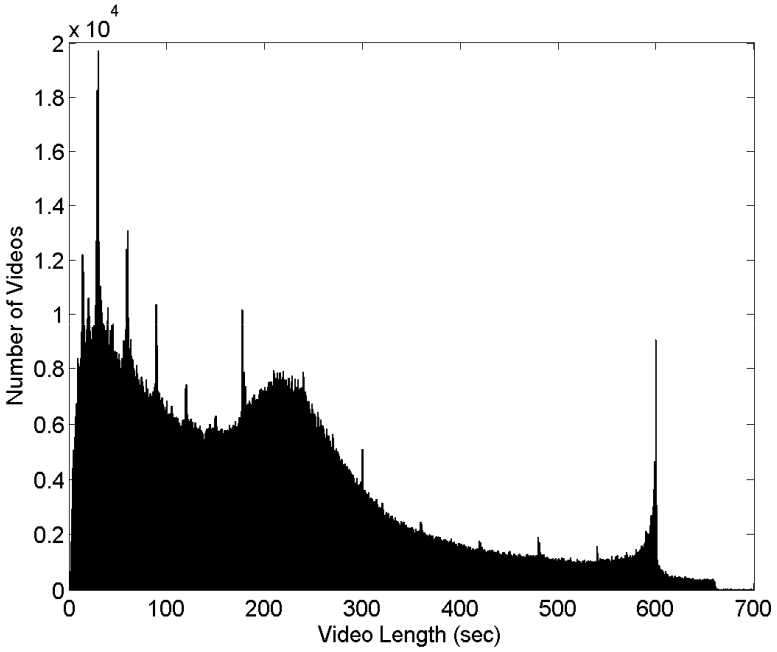

Fig. 9.1 Distribution of video length

Figure 9.1 shows the distribution of YouTube videos' lengths of less than 700 s, which exhibits three peaks. The first peak is for videos that last less than a minute, and contains more than 20% of all videos, which clearly demonstrates that YouTube is primarily a site for very short videos. The second peak is between 3 and 4 min, and contains about 16.7% of the videos. This peak is mainly caused by the large number of videos in the "Music" category. "Music" is the most popular category for YouTube, and the typical length of a music video is often within this range. The third peak is close to the maximum of 10 min, 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 min.

File Size and Bitrate

We retrieved the file size of nearly 190,000 videos. In our crawled data, 98.8% of the videos are smaller than 30 MB size. Not surprisingly, we find that the distribution of video sizes is very similar to the distribution of video lengths. We calculate an average video file size to be about 8.4 MB.

Considering there are over 42.5 million YouTube videos, the total disk space required to store all the videos is more than 357 terabytes! Smart storage management is thus quite demanding for such an ultra-huge and still growing site, which we discuss in another paper [3].

We found that the videos' bitrate has three clear peaks. Most videos have a bitrate around 330 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.

Date Added – Growth Trend of Uploading

During our crawl we recorded the date that each video was uploaded, so that we could study the growth trend of YouTube. Figure 9.2 shows the number of new videos added every 2 weeks in our entire crawled dataset.

February 15, 2005 is the day that YouTube was established. Our first crawl was on February 22, 2007; this meant that we could only find early videos if they were still very popular videos or are linked to by other videos we crawled. We can see there is a slow start, the earliest video we crawled was uploaded on April 27, 2005. Six months after YouTube's establishment, the number of uploaded videos increases steeply.

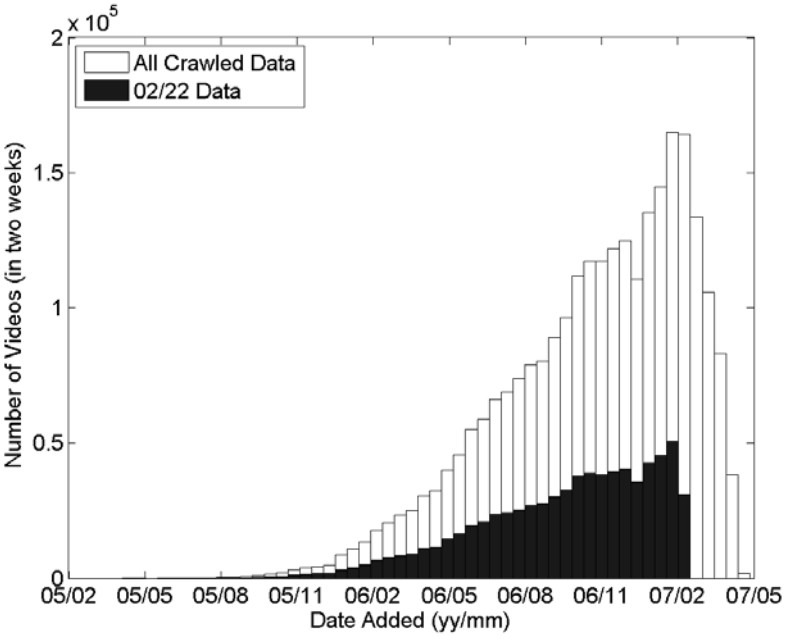

Fig. 9.2 Uploading trend of YouTube videos

Views – User Access Pattern

The number of views a video has had is the most important characteristic we measured, as it reflects the popularity and access patterns of the videos. We use a single dataset containing more than 100,000, which is considered to be relatively static.

Figure 9.3 shows the number of views as a function of the rank of the video by its number of views. The plot has a long tail on the linear scale (not shown), which means there are a few videos that have been watched millions of times, and there are also a great number of videos that are seldom watched. However, unlike website visitors distribution, web caching and Peer-to-Peer file sharing workload, the access pattern does not follow a Zipf distribution, which should be a straight line on a log–log scale. The figure shows that the beginning of the curve is linear on a log–log scale, but the tail (after the 2 × 103 video) decreases tremendously, indicating there are not so many less popular videos as Zipf's law predicts.

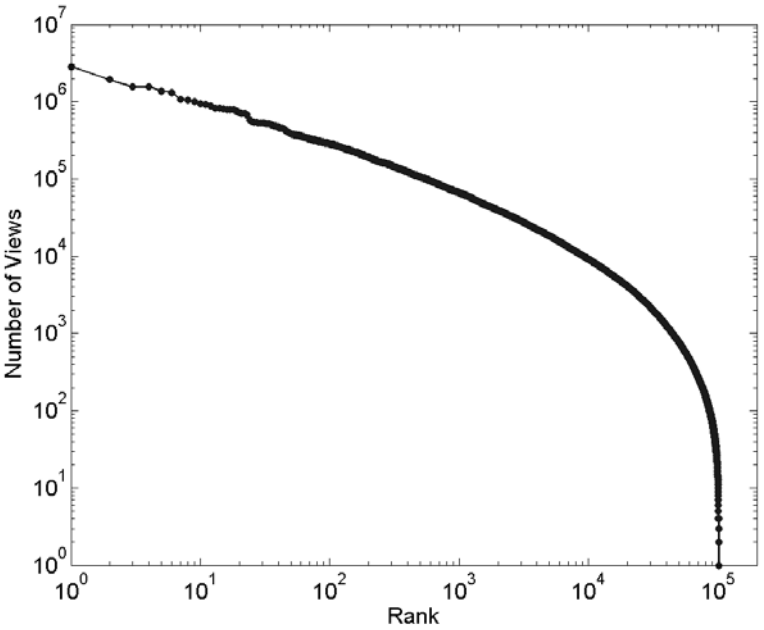

Fig. 9.3 YouTube videos rank ordered by popularity

The Social Network in YouTube

YouTube is a prominent social media application: there are communities and groups in YouTube, there are statistics and awards for videos and personal channels. Videos are no longer independent from each other, and neither are the users. It is therefore important to understand the social network characteristics of YouTube. We next examine the social network among YouTube users and videos, which is a unique and interesting aspect of this kind of video sharing sites, as compared to traditional media services.

Small-World Phenomenon

The small-world network phenomenon is probably the most interesting characteristic for social networks, and has been found in various real-world situations, such as URL links in the Web [1].

The concept of a small-world was first introduced by Milgram to refer to the principle that people are linked to all others by short chains of acquaintances (popularly known as six degrees of separation) [6]. This formulation was used by Watts and Strogatz to describe networks that are neither completely random, nor completely regular, but possess characteristics of both [9]. 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.

The Small-World in YouTube

We measured the graph topology for all the YouTube data gathered, by using the related links in YouTube pages to form directed edges in a video graph for the entire dataset. For comparison, we also generate random graph with the same number of nodes and average node degree of the crawled dataset.

We found that the clustering coefficient of our YouTube dataset is quite high, about 0.3, and is especially large in comparison to the random graphs, which are nearly 0. We also found that the characteristic path length is about 8, which is only slightly larger than that of the corresponding random graph. This is quite good, considering the still large clustering coefficient of these datasets.

The network formed by YouTube's related videos list has definite small-world characteristics. The clustering coefficient is very large compared to a similar sized random graph, while the characteristic path length is approaching the short path lengths measured in the 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.

These results are similar to other real-world user-generated graphs, 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 [1]. This could possibly be due to the larger number of nodes (8×108 in the web), but it may also indicate that the YouTube network of videos is a much closer group.

Further Discussions

Can Peer-to-Peer Save YouTube?

Short video sharing and Peer-to-Peer streaming have been widely cited as two key driving forces to Internet video distribution, yet their development remains largely separated. The Peer-to-Peer technology has been quite successful in supporting large-scale live video streaming (e.g., TV programs like PPLive and CoolStreaming) and even on-demand streaming (e.g., GridCast). Since each peer contributes its bandwidth to serve others, a Peer-to-Peer overlay scales extremely well with larger user bases. YouTube and similar sites still use the traditional client-server architecture, restricting their scalability.

Unfortunately, our YouTube measurement results suggest that using Peer-to-Peer delivery for YouTube could be quite challenging. In particular, the length of a YouTube video is quite short (many are shorter than the typical connection time in a Peer-to-Peer overlay), and a user often

quickly loads another video when finishing a previous one, so the overlay will suffer from an extremely high churn rate. Moreover, there is a large number of videos, so the Peer-to-Peer overlays will appear very small.

Our social network findings again could be exploited by considering a group of related videos as a single large video, with each video in the group being a portion of the large one. Therefore, the overlay would be much larger and more stable. Although a user may only watch one video from the group, he/she can download the other portions of the large video from the server when there is enough bandwidth and space, and upload those downloaded portions to other clients who are interested in them. This behavior can significantly reduce the bandwidth consumption from the server and greatly increase the scalability of the system.

Finally, another benefit of using a Peer-to-Peer model is to avoid single-point of failures and enhance data availability. While this is in general attractive, it is worth noting that timely removal of videos that violate the terms of use (e.g., copyright-protected or illegal content, referred to by the "Removed" category above) have constantly been one of the most annoying issues for YouTube and similar sites. Peer-to-Peer delivery will clearly make the situation even worse, which must be well addressed before we shift such sites to the Peer-to-Peer communication paradigm.

A Peer-to-Peer Simulation

Our ongoing work is to design a Peer-to-Peer structured short video sharing system. In this system, peers are responsible for redistributing the videos they have already downloaded. Therefore, the workload traffic of the server is significantly reduced. We conduct a simulation and plot the results in Fig. 9.4.

In Fig. 9.4, the topmost line represents the server bandwidth in client-server structure; the lowest line represents the server bandwidth in optimal Peer-to-Peer structure, in which the peer has unlimited uploading bandwidth, unlimited storage to store all the downloaded video, and exists all the time. The optimal situation is impossible to implement, thus we limit the peer's uploading bandwidth, storage, and existing time. In this case, the server bandwidth is represented by the second lowest line, and the total peer uploading bandwidth is represented by the second highest line. From the figure, we can easily find out that the server bandwidth is greatly reduced in Peer-to-Peer structure, amounting to approximately 39.8% of that in the client-server structure; the contribution of all the peers is more than that of the server.

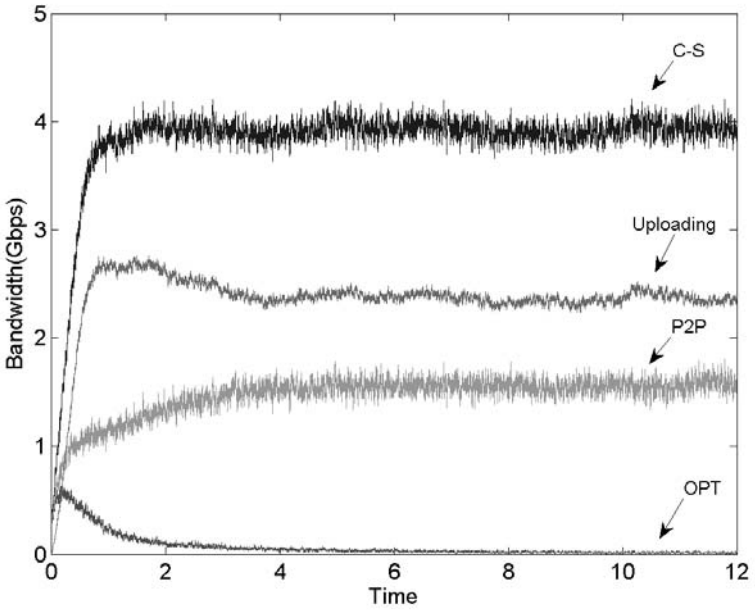

Fig. 9.4 Bandwidth comparison of the Peer-to-Peer experiment

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 massive amounts of data collected in a 3-month period, we have demonstrated that, while sharing certain similar features with traditional video repositories, YouTube exhibits many unique characteristics, 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 among YouTube videos, which is probably its most unique and interesting aspect, and has substantially contributed to the success of this new generation of services. 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. We have suggested that these features can be exploited to facilitate the design of novel Peer-to-Peer strategies for short video sharing.[3]

Notes

1. There are an estimated 42.5 million videos on YouTube: http://googlesystem.blogspot.com/2007/06/google-videos-new-frame.html
2. YouTube Blog: http://youtube.com/blog
3. Xu Cheng is a Ph.D. student in the School of Computing Science at Simon Fraser University, British Columbia, Canada. Email: xuc@sfu.ca. Cameron Dale is a M.Sc. student in the School of Computing Science Simon Fraser University. Email: camerond@cs.sfu.ca. Jiangchuan Liu is Assistant Professor in the School of Computing Science, Simon Fraser University, British Columbia, Canada. Email: jcliu@cs.sfu.ca.

References

Albert R., Jeong H., and Barabasi A. The Diameter of the World Wide Web. Nature, vol. 401, pp. 130, 1999.

Almeida J.M., Krueger J., Eager D.L., and Vernon M.K. Analysis of Educational Media Server Workloads. In Proc. of NOSSDAV, 2001.

Cheng X., Dale C., and Liu J. Statistics and Social Network of YouTube Videos. In Proc. of IWQoS, 2008.

Gill P., Arlitt M., Li Z., and Mahanti A. YouTube Traffic Characterization: A View From the Edge. In Proc. of IMC, 2007.

Huang C., Li J., and Ross K.W. Can Internet Video-on-Demand be Profitable? In Proc. of SIGCOMM'07.

Milgram S. The Small World Problem. Psychology Today, vol. 2, no. 1, pp. 60–67, 1967.

Mislove A., Marcon M.,. Gummadi K.P, Dreschel P., and Bhattacharjee B. Measurement and Analysis of Online Social Networks. In Proc. of IMC, 2007.

Tang W., Fu Y., Cherkasova L., and Vahdat A. Long-term Streaming Media Server Workload Analysis and Modeling. Technical report, HP Labs, 2003.

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