Traffic produced by YouTube has had a significant impact on both fixed and mobile networks. The study and evaluation of YouTube content features can benefit network traffic engineering by supporting the development of sustainable video delivery services and regulation of network traffic. Such evaluations are particularly useful to network operators who aim to refine and optimize existing cache algorithms to better adapt to YouTube video traffic patterns.

The main objective of this article is to depict the latest YouTube traffic profiles and deliver updated and valuable information for future researchers. There has been a healthy amount of research on YouTube video analysis, but most of it was conducted prior to 2007. Following Google’s acquisition of YouTube in 2008, several major aspects of the network and service framework were restructured, leading to changes in the user policy and service infrastructure—including new limits on video duration, size, and resolution. Here, we reflect on YouTube video characteristics in light of such changes. Given that it has been more than six years since similar studies were conducted, our findings provide comparative insight into how YouTube videos have developed over recent years.

Measurement Methodology
We compare statistics from the works of Xu Cheng, Cameron Dale, and Jiangchuan Liu; Phillipa Gill and her colleagues; and Abdolreza Abhari and Mojgan Soraya with our own findings, because these works present similar scopes of investigation, covering video categories, video durations, file sizes, and bit rates. All four publications were based on statistics from prior to Google’s acquisition of YouTube, so we use a comparative analysis to depict how YouTube videos have developed in recent years based on the collective statistics (see Table 1).

In terms of YouTube traffic analysis, a variety of research has been performed with different approaches and objectives. Pablo Ameigeiras and his colleagues collected statistics from the most viewed video clips, studied the impact of video encoding rates on traffic generation, and proposed a traffic model that could be used for feasible and effective simulation. However, we don’t use this work as a reference point, because it was based on a total trace of 32,860 videos, and it analyzed traffic from a different perspective.

YouTube Video Metadata
After a video is uploaded and converted, YouTube randomly assigns it a unique 64-bit number, which is represented in a base-64 encoding algorithm by an 11-character alphanumeric ID. For all video records we collected for this study, a check function was implemented to confirm every record in the dataset was distinct by removing records with duplicate IDs. Each record in the dataset contains intuitive metadata and the first 2 Kbytes of the content.

A typical example of the metadata for a YouTube video is as follows:
- YouTube ID: aZpD0btOZx8
- Video title: Super Mario
- Video category: Music
- Content length: 4,163,902 bytes
- Resolutions: 34/640, 360/9/0/115, 5/320, 240/7/0
- Video duration: 64 seconds
- Total views: 36,383,401

The main difference between YouTube’s video service and other traditional video services is that YouTube’s videos are not streamed to
users; instead, they’re downloaded over normal HTTP-over-TCP connections. Except for possibly stopping the download, YouTube users have no control over the download speed. The data rate of HTTP connections is not controlled at the same pace as the video playback rate; rather, video content is sent at the maximum data rate of the network capacity, which might overload the underlying network.

**Customized Web Spider**

Crawling has been a popular methodology for collecting and characterizing YouTube videos.1–4 We developed a customized Web spider to collect metadata information from YouTube video content. The YouTube content item (that is, the video webpage) is linked to other content items that have similar titles, descriptions, or tags, chosen by the uploader. A YouTube content item might have hundreds of YouTube content-related links, although the YouTube webpage only shows the top 20 related links at any given time. Consequently, the relationship between YouTube videos can be considered as a directed spider-Web graph, where each video is a node on the graph and videos are linked to each other via the top 20 related links. The Web spider followed the recursive links among YouTube videos and captured a video dataset using a breadth-first search technique.6

The sample space for this project focused primarily on popular and active YouTube content. This is pertinent, because rarely accessed or unpopular YouTube content has a relatively low impact on the traffic optimization strategy. Thus, the initial YouTube URLs were loaded from www.youtube.com/videos/?s=pop, which is the front page for popular content on YouTube. The Web spider operates with several different IP addresses, which were changed on a regular basis and restarted once a day to bypass YouTube’s restrictions. This was necessary, because YouTube blocks access from the same IP address through which a large number of YouTube content items have been accessed in a given time period. (We discuss the legitimacy issue later.)

The YouTube Web spider worked on 11-thread actions to achieve maximum performance. Two URL tables were maintained by the system: one recorded visited URLs and the other listed new URLs. Upon system startup, the spider loaded all stored YouTube IDs into the visited URL table to avoid URL duplication. Thread 0 added a new URL retrieved from the YouTube front page to the new URL table every 30 minutes, because new links are regularly updated by the YouTube server. The remaining 10 threads (threads 1–10) were worker threads for YouTube information retrieval, each of which carried out the following process repeatedly:

- visit one URL that is randomly selected from the new URL table,
- retrieve content metadata,
- download the first 2 Kbytes of YouTube content,
- save the metadata and 2 Kbytes of content to database,
- add 20 new URLs to the new URL table and filter out duplicated URLs, and
- add the visited URL to the visited URL table.

This process involved three HTTP transactions.

The first transaction acquired the YouTube metadata video information. YouTube only lets users view videos online; the videos can’t be downloaded. Thus, the actual links for HTTP video streaming are encoded in YouTube’s HTML and JavaScript pages and are updated

---

Table 1. Collective statistics overview.

<table>
<thead>
<tr>
<th>Publication</th>
<th>Collection timeframe</th>
<th>Total collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study”¹</td>
<td>Early 2007</td>
<td>2,676,388 videos</td>
</tr>
<tr>
<td>“YouTube Traffic Characterization: A View from the Edge”²</td>
<td>Jan.–Apr. 2007</td>
<td>23,250,438 HTTP transactions</td>
</tr>
<tr>
<td>“Statistics and Social Network of YouTube Videos”³</td>
<td>Early 2007</td>
<td>3,269,030 videos</td>
</tr>
<tr>
<td>“Workload Generation for YouTube”⁴</td>
<td>2007 and 2008</td>
<td>60,544 videos</td>
</tr>
<tr>
<td>This article</td>
<td>Apr.–May 2013</td>
<td>1,245,700 videos</td>
</tr>
</tbody>
</table>

---

March–April 2015
periodically. Consequently, a major challenge for this project was to decode HTTP streaming links for the required data to be retrieved for analysis. (We discuss copyright issues in the next section.)

The second HTTP transaction is responsible for downloading actual video content. To save bandwidth and storage space, only the first 2 Kbytes of YouTube content is downloaded and stored in a MySQL database for further processing, which was sufficient for the final analysis. The video downloading was necessary for two reasons. First, the length of the video is a critical parameter for content analysis, which does not exist in the video’s HTML webpage—it’s only in the header of the HTTP download stream. Second, the audio and video encoding schema only exists within video content.

The final transaction retrieved the viewing history for the relevant videos. The viewing history represents a data summary of daily visits to the video, from the day that it was uploaded to the day that the Web spider accessed it. The history also indicates how the video’s popularity has grown (or declined) and the lifespan of the YouTube video (for videos removed from YouTube). The view history is returned in the body of the HTTP response.

Copyright and Access Issues
Factors regarding copyright and access have scarcely been addressed in existing academic publications, so we consider such issues here. Numerous debates and discussions are available in various selections of texts and online sources with regard to the legality of downloading and the potential copyright infringement of YouTube videos.7,8

YouTube provides free video content for users who, in return, register website hits to increase the popularity and, ultimately, profits of the site. By downloading videos for offline viewing, users would be circumventing their exposure to advertising placed on YouTube beyond the initial viewing, thereby removing any money-making potential in subsequent viewings. Naturally, YouTube does not want users to sidestep advertising by engaging in video downloads. However, avoiding advertising is not inherently illegal, in the same sense that it is legal to use in-browser advert blockers and skip ads on TV. These discussions reflect the common arguments on the issue.

Despite the ongoing confusion and debate, the research described in this article can be justified for two major reasons. First, we didn’t fully download YouTube video content—we extracted only the first 2 Kbytes of content, because the header contains all necessary metadata information for this research. Second, the research conducted complies with the fair use policy of the World Intellectual Property Organization, which states “the fair use of a copyright work, for purposes such as criticism, comment, news reporting, teaching, scholarship, or research, is not an infringement of copyright.”9

YouTube Video Characteristics
All 1,245,700 dataset records were saved as raw data in a MySQL database. Each record represents a unique YouTube video content item. A data processing application was developed using the C# programming language to analyze each record.

Video Category Distribution
Upon uploading a video, the YouTube server selects one of 18 predefined video categories. Table 2 compares the video category rankings for 2007 and 2013. The data for 2007 is taken from Cheng, Dale, and Liu’s “Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study,”1 while 2013 represents the primary data we collected using the customized Web spider. As can be seen, a few popular categories account for a greater percentage of the content, whereas numerous less popular categories represent a much smaller share of the content.

Compared to 2007, although Music and Entertainment remained the two most popular categories in 2013, two new categories were added: Science & Technology and Education. The third largest category in 2007 was Comedy with 12.1 percent, but that dropped to 5.9 percent in 2013 and was replaced by the Gaming category, which rose from 7.3 to 8.5 percent. In the other categories, two decreased slightly: Sports (from 9.7 to 8.0 percent) and Film & Animation (from 8.4 to 5.9 percent), while other growing categories included People & Blogs (from 7.4 to 8.1 percent), Cars & Vehicles (from 2.5 to 3.9 percent), and, most markedly, How To & Style (from 2.0 to 5.1 percent).

Video Duration Distribution
Generally speaking, YouTube videos are shorter than traditional media videos (such as films and TV programs). The results from the dataset reveal that more than 96 percent of the videos

---

IEEE MultiMedia

58
retrieved in this analysis are less than 600 seconds in length. This is mainly due to the limit of 10 minutes imposed by YouTube in March 2006, which was increased to 15 minutes in July 2010.10 Partner users of YouTube and users with verified status can upload videos longer than the set limit.10

Figure 1a depicts the histogram of the distribution of YouTube video duration for 2007 and 2013, where the x-axis represents the video duration range, and the y-axis refers to the number of videos for all categories. Again, the 2007 data is taken from the existing research of Cheng, Dale, and Lui,1 and the 2013 data was collected via the customized Web spider. Figure 1b illustrates and compares the YouTube video duration distribution of the four most popular categories of 2013: Music, Entertainment, Gaming, and People & Blogs. The x-axis represents the video duration range every 20 seconds, while the y-axis refers to the percentage of the number of videos in the category.

From the 2013 data alone, the duration distribution exhibits four peaks, and the overlapped contour of Figure 1b resembles the shape of Figure 1a. The 2007 dataset only shows the first three peaks. The first peak is at 1 minute. Note that in 2007, approximately 21 percent of the videos were shorter than one minute; in 2013, 16 percent were shorter than one minute. YouTube has been viewed as an outlet for short videos since 2005, although the decrease in the percentage of short videos indicates that YouTube is gradually trying to cater to those wishing to upload longer videos as well.

The second and third peaks are consistent over the past five years, as shown in the comparison of 2007 and 2013 figures. The second peak, which is within the range of 200 to 240 seconds, occurs because the Music category has been a long-standing popular category on YouTube, and the typical length of music videos is often within this range. The third peak is near the duration of between 580 and 600 seconds due to the duration limit imposed by YouTube. Users often tend to divide long videos into several pieces with each fitting the boundary of 10 minutes.

The fourth peak in the 2013 figure is caused by the number of videos that exceed 700 seconds in length. This peak is missing from the 2007 figure because of the previous 10-minute boundary. As noted earlier, in July 2010, YouTube raised the video uploading limit to 15 minutes, and five months later it allowed verified users to upload videos longer than 15 minutes. Users have clearly started to take advantage of this new facility.

Statistics in this research indicate that Gaming is the main category for long videos (that is, those longer than 700 seconds). As shown in Figure 1b, 6.3 percent of videos in this category are 700–900 seconds in length, while other categories all have less than 1 percent in this duration (except for the Entertainment category, which has 1.2 percent). There are barely any

---

**Table 2. YouTube video category rankings, 2007 and 2013.**

<table>
<thead>
<tr>
<th>Rank</th>
<th>2007 Category</th>
<th>Percentage</th>
<th>2013 Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Music</td>
<td>22.9</td>
<td>Music</td>
<td>22.8</td>
</tr>
<tr>
<td>2</td>
<td>Entertainment</td>
<td>17.8</td>
<td>Entertainment</td>
<td>16.0</td>
</tr>
<tr>
<td>3</td>
<td>Comedy</td>
<td>12.1</td>
<td>Gaming</td>
<td>8.5</td>
</tr>
<tr>
<td>4</td>
<td>Sports</td>
<td>9.7</td>
<td>People &amp; Blogs</td>
<td>8.1</td>
</tr>
<tr>
<td>5</td>
<td>Film &amp; Animation</td>
<td>8.4</td>
<td>Sports</td>
<td>8.0</td>
</tr>
<tr>
<td>6</td>
<td>People &amp; Blogs</td>
<td>7.4</td>
<td>Comedy</td>
<td>5.9</td>
</tr>
<tr>
<td>7</td>
<td>Gaming</td>
<td>7.3</td>
<td>Film &amp; Animation</td>
<td>5.9</td>
</tr>
<tr>
<td>8</td>
<td>News &amp; Politics</td>
<td>4.3</td>
<td>How To &amp; Style</td>
<td>5.1</td>
</tr>
<tr>
<td>9</td>
<td>Autos &amp; Vehicles</td>
<td>2.5</td>
<td>News &amp; Politics</td>
<td>4.6</td>
</tr>
<tr>
<td>10</td>
<td>Travel &amp; Places</td>
<td>2.2</td>
<td>Cars &amp; Vehicles</td>
<td>3.9</td>
</tr>
<tr>
<td>11</td>
<td>How To &amp; DIY</td>
<td>2.0</td>
<td>Science &amp; Technology</td>
<td>2.9</td>
</tr>
<tr>
<td>12</td>
<td>Pets &amp; Animals</td>
<td>1.9</td>
<td>Education</td>
<td>2.9</td>
</tr>
<tr>
<td>13</td>
<td>—</td>
<td>—</td>
<td>Travel Events</td>
<td>2.2</td>
</tr>
<tr>
<td>14</td>
<td>—</td>
<td>—</td>
<td>Pets &amp; Animals</td>
<td>1.8</td>
</tr>
</tbody>
</table>
videos between 15 and 20 minutes for the four categories outlined in Figure 1b, with a maximum of 0.65 percent from the Music category. For videos that are longer than 20 minutes, the order of the four major categories is Gaming (1.3 percent), People & Blogs (0.6 percent), Entertainment (0.5 percent), and Music (0.15 percent).

Resolution Distribution

There is no published research data regarding YouTube video resolution. The analysis of YouTube video resolutions is important because it gives a definitive indication of video qualities in the system, and it offers an indicative reflection of users’ uploading capabilities. The original YouTube service only offered videos with one resolution level, at 320 × 240, which solely used Macromedia’s proprietary Flash technology. With the dramatic growth in smartphone use in recent years, YouTube started offering support for the MP4 format in 2007 for devices that do not offer Flash, such as Apple’s iPhone and iPad. YouTube also started letting users upload higher quality videos. Furthermore, starting in March 2008, it permitted a wider range of resolutions.10 Such expansions of YouTube services illustrates the company’s willingness to adapt to an evolving market, letting users choose suitable video resolutions and formats according to their available bandwidth, requirements, and devices.

The current observation in 2013 of YouTube HTML source code reveals that YouTube’s video playback technology is based on both Flash (FLV) and MPEG4 (MP4). In fact, YouTube accepts a variety of video formats, such as WMV, AVI, MOV, and MPEG, which are automatically converted into the FLV and MP4 formats in different resolutions upon uploading. Many YouTube videos will also play using HTML5 in supported browsers, where formats such as WebM VP8 and H.264 are permitted.

When the same video content is being uploaded to YouTube servers, each unique upload is transcoded into a variety of formats and resolutions to support streaming requirements,11 so several sources (that is, files) are saved on the server, each corresponding to one resolution. A total of 1,245,700 unique YouTube video are recorded in this dataset, and the sum of all resolution sources for all of these videos is 4,264,696, which means, on average, each YouTube content item has approximately 3.4 resolution sources. The ranking of various YouTube video resolutions in 2013 is

1. 320 × 240 (FLV)—100 percent,
2. 640 × 360 (MP4)—74 percent,
3. 640 × 360 (FLV)—66 percent,
4. 854 × 480 (FLV)—40 percent,
5. 320 × 240 (MP4)—26 percent,
6. 320 × 240 (others)—20 percent,
7. 1,280 × 720 (total)—14 percent, and
8. 1,920 × 1080 (total)—3 percent.

These percentages refer to the proportion of YouTube video content that has enabled the corresponding resolution. For example, the percentage for 320 × 240 (FLV) is 100 percent, which means that all YouTube contents contain

---

Figure 1. YouTube video duration distribution: (a) duration distribution (2007 vs. 2013) and (b) duration distribution of major categories (2013).
a resolution of 320 × 240 in Macromedia’s Flash format. The statistics show that the original resolution 320 × 240, which was originally introduced by YouTube, is still by far the leading resolution. Currently, the two most popular resolutions in YouTube are 320 × 240 and 640 × 360.

**Video File Size Distribution**

Video file size distributions have been investigated, where the analyses of file sizes were based on all collected YouTube videos and did not offer a perspective on specific video resolutions. Figure 2 shows the video file size distributions for video contents with resolutions of 640 × 360 (FLV) and 320 × 240 (FLV) for 2013. The x-axis represents the range of YouTube video file sizes, and the y-axis refers to the proportion of YouTube videos within the respective ranges. The statistical results and analysis focus on two popular resolutions in Flash format—640 × 360 and 320 × 240, which represent high- and low-quality videos.

Table 3 compares the file size ranges we collected against statistics from existing research. YouTube’s policy on the size limit of video files was 100 Mbytes when previous studies were carried out prior to 2008. The current file size limit is 2 Gbytes for uploading via YouTube Web or 20 Gbytes if up-to-date browser versions are used. Because of this policy change, the average file size has increased over the past few years.

Cheng, Dale, and Lui state that the distribution of file sizes is similar to that of video lengths, and Abhari and Soraya discovered that the file size distribution can be modeled by a gamma distribution. Both articles have given an approximate estimation of YouTube file size distribution using all collected YouTube videos in the dataset. A more precise analysis should consider video resolutions. As discussed in the previous section, a range of resolution options have been offered by YouTube since 2008, which has had a direct impact on file sizes. When the same YouTube content is uploaded with different resolutions, file size varies accordingly. As shown in Figure 2, file sizes of resolution 640 × 360 (FLV) are notably higher than those of 320 × 240 (FLV).

One important goal of investigating YouTube file sizes is to help network carriers with cache management. The average YouTube video file size is approximately 17.6 Mbytes for a resolution of 640 × 360 (FLV) and 6.5 Mbytes for a resolution of 320 × 240 (FLV). Therefore, if 1 million YouTube videos were to be cached, the total disk space required for storage would be approximately 17.6 Tbytes for resolution 640 × 360 (FLV) and 6.5 Tbytes for a resolution 320 × 240 (FLV).

The first peak (also the main peak) of file size distribution can still be described using the gamma model. There is a slight rise of percentage after the long tail. Around 0.69 percent of YouTube video files that are 640 × 360 (FLV) are more than 100 Mbytes, and around 0.34 percent of YouTube video files in 320 × 240 (FLV) are more than 50 Mbytes. Despite the remarkable increase of file size allowance from YouTube, the maximum file size that users can upload is still affected by the video resolution choice and video duration limit. Take 640 × 360, for example. The maximum video bit rate of standard quality uploads is 1,000 kbps, and the maximum audio bit rate is 197 Kbps, so the maximum possible file size for this resolution would be \(1 + 0.197\) Mbits, multiplied by 15 minutes, which equals roughly 134 Mbytes.

**Data Rate Distribution**

Gill and her colleagues estimated the YouTube video playback data rate by dividing the video file size by the video duration. In the dataset we collected, the data rate can be easily observed from the FLV header retrieved from the metadata. The value of the total data rate is in the format of IEEE 754-2008, IEEE Standard for Floating-Point Arithmetic.

More than 99.9 percent of the YouTube videos accessed contained FLV metadata specifying the content’s total data rate, video data rate, and audio data rate. (The FLV headers of fewer than 1,000 YouTube records from a total of
1,245,700 records in the dataset were corrupt and thus couldn’t be parsed. This indicates that virtually all YouTube videos are transmitted at a constant bit rate (CBR). Compared with a variable bit rate (VBR), which is suitable for high-quality video download and certain high-bandwidth streaming environments, CBR is a more reliable choice for streaming videos than any bandwidth that users might have. Figure 3 depicts the distributions of the playback total data rate and audio data rate for the two most popular resolutions: 320 x 240 (FLV) and 640 x 360 (FLV). The x-axis represents the bit rate of YouTube content in Kbps, and the y-axis refers to the proportion of YouTube content.

As shown in Figure 3a, the YouTube video total data rate in 320 x 240 (FLV) format reaches its peak at 300 Kbps and 320 Kbps in 2013 data, which is consistent with the result of Cheng, Dale, and Liu, where 300 Kbps and 360 Kbps data rates were observed as a peak period in 2007. The total data rate for 640 x 360 (FLV) also reaches its peak at 320 Kbps. The peak data rates for two different resolutions fall into the same range, which implies the most common data rate budget for today’s users is approximately 320 Kbps. The total data rate for more than 98 percent of YouTube videos in the 320 x 240 (FLV) format is within 360 Kbps, while the total data rate for 640 x 360 (FLV) is up to 960 Kbps. Both ranges are within the data rate limits set by YouTube.

In terms of the audio data rate, Figure 3b indicates that both distributions of the two resolutions result in two peaks that coincide with the two audio bit rates that YouTube recommends: the mono audio bit rate is 64 Kbps, and the stereo audio bit rate is 128 Kbps. Although no data is shown for the recommended data rate of 320 x 240, based on Figure 4b, the mono audio bit rate should be 16 Kbps, and the stereo audio bit rate should be 64 Kbps.

Compared with traditional video repositories, YouTube exhibits many unique characteristics that can introduce novel challenges and opportunities for optimizing the performance of short video sharing services. The findings of this article can benefit telecommunication
providers in the evaluation and estimation of cache usage in network services. Beyond highlighting some of the core characteristics of YouTube’s content storage and distribution, the outcomes of this work might also benefit other researchers in the field by informing them of the likely bandwidth and data compression requirements for popular and emerging technologies, such as smartphones and wearable devices—where YouTube remains a comparatively pervasive and popular application. The methodological approach adopted in this work in the form of a customized Web spider can also be used to analyze similar levels of mass provision within digital media services. Future surveys of this type could be useful in exploring distinctions between content provision when accessed via different platforms—such as via workstations, tablets, or smartphones—as this will demonstrate the impact of different user platforms on service performance.

References


Xianhui Che is an assistant professor in the Department of Electrical and Electronic Engineering at the University of Nottingham Ningbo, China. Her research interests include network communication, embedded system, and digital media. Che obtained her PhD in electronic systems engineering from the University of Essex, UK. Contact her at cherry.che@nottingham.edu.cn.

Barry Ip is an associate professor in the School of International Communications at the University of Nottingham Ningbo, China. His research interests include computer and video game design and the use of learning technology in higher education. He also has a keen interest in a variety of business-, health-, and research-related topics, and his latest research examines game retail in China, visual language in games, and longitudinal studies into game quality. Ip obtained his PhD in digital media from Swansea University, UK. Contact him at barry.ip@nottingham.edu.cn.

Ling Lin is a chartered engineer and technical director of Ningbo Xiangshan Elevator Components Ltd, China, and he often offers research consultancy at the University of Nottingham Ningbo, China. He has expertise in telecommunications, network protocols, multimedia, and motor control. Lin received his PhD in electronic systems engineering from the University of Essex, UK. Contact him at linglin_unnc@outlook.com.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.