YouTube Traffic Content Analysis in the Perspective of Clip Category and Duration

Jie Li[1], Andreas Aurelius[1,3], Manxing Du[1,3]
[1]Acreo Swedish ICT AB, Networking and Transmission Laboratory, Electrum 236, SE-164 40 Kista, Sweden.
Email: jie.li@acreo.se

Hantao Wang[2], Åke Arvidsson[2]
[2]Ericsson AB, Stockholm, Sweden

Maria Kihl[3]
[3]Lund University, Department of Electrical and Information Technology, Lund, Sweden

***Abstract*—In this work, we study YouTube traffic characteristics in a medium-sized Swedish residential municipal network that has ~ 2600 mainly FTTH broadband-connected households. YouTube traffic analyses were carried out in the perspective of video clip category and duration, in order to understand their impact on the potential local network caching gains. To the best of our knowledge, this is the first time systematic analysis of YouTube traffic content in the perspective of video clip category and duration in a residential broadband network. Our results show that the requested YouTube video clips from the end users in the studied network were imbalanced in regarding the video categories and durations. The dominating video category was Music, both in terms of the total traffic share as well as the contribution to the overall potential local network caching gain. In addition, most of the requested video clips were between 2-5 min in duration, despite video clips with durations over 15 min were also popular among certain video categories, e.g. film videos.**

Keywords—Internet traffic monitoring, residential Internet traffic pattern, traffic locality, YouTube, video category, video duration, caching gain

I. INTRODUCTION

Internet has now become the global information and communication base for every aspect of people's work and life. Even so, global Internet traffic keeps on growing steadily with the ever-increasing mainly video-oriented content distributions and services, as well as with the ever-increasing new devices that are connected to the network [1]. This puts challenges for network operators and Internet Service Providers (ISP) to deliver broadband IP services that meet the quality-of-service (QoS) requirements of multimedia services needed to provide sufficient quality-of-experience (QoE) to the end users [2].

One important part in meeting this challenge is to understand the Internet traffic characteristics, especially the Internet traffic patterns of residential end users. These users have generated according to [1] about 80% of the global IP traffic in the years 2010-2011, and they are expected to have even larger total traffic share in the next few years. Indeed, many research groups have so far published a large number of internet traffic measurement and analysis results. In particular, it has been pointed out in a long-term study of residential Internet traffic evolution that streaming-based video and audio services are the driving forces for the increase of the total network traffic [3]. Among them, YouTube has become the world's most popular online streaming video service with over one billion monthly users, and nearly one out of every two people on the Internet visiting YouTube's website [4]. Moreover , since its launch in 2005, YouTube has evolved from a pure individual user generated content (UGC) video service to an important business platform for not only advertisers, but also for 'tens of thousands of partners that have created channels and found and built businesses'[4]. Therefore, it is of great interest to investigate the YouTube traffic characteristics and understand its potential impact on both the Internet traffic as well as the end-user experience.

Studies on YouTube traffic characteristics have been reported previously [5-9]. In an early study carried out in 2007 [5], using video clip metadata retrieved from YouTube, analyses on the global popularity evolution of YouTube video clips were carried out. However, only video clips categorized as `Entertainment' and `Science & Technology' were analyzed in this study (1 day for 'Entertainment' and 8 days for `Science & Technology'). Besides, no individual end user behavior was available in this study. In another similar study carried out by Google [6], randomly selected 20 million YouTube videos uploaded between September 2010 and August 2011 were analyzed. The results showed that online video popularity appeared strongly constrained by geographic locality, and the impact of social sharing to the spatial popularity was limited. Again no individual end user behavior was available in this study. In [7-8] measurement studies on YouTube traffic of university campus networks were conducted. The results verified that local and global popularity were not or only slightly correlated. Somewhat differently, in another study [9] based on one week traffic data in February 2011 in two campus networks and three nation-wide ISPs, YouTube user access behaviors were studied, which revealed that users accessed YouTube in a very similar manner, independent of their location, the device they used, and the access network that connected them.

In this work, we study YouTube traffic characteristics in a medium-sized Swedish residential municipal network that has ~ 2600 mainly FTTH broadband-connected households. YouTube traffic analyses were carried out in the perspective of the requested video clip categories and durations and their impact on the potential local network caching gains. Compared to the previously reported studies touching (partly) the video clip category or durations in [5][7][9], the novelty of our work

is that our results are based on the captured (residential) end-user YouTube watch requests, from which we have two sets of data: the requests (including the repeated requests for the same video) and the unique videos watched (while for those published results only the watched video clip statistics were available). Moreover, to the best of our knowledge, this is also the first time systematic analysis of YouTube traffic content in the perspective of video clip category and duration in a residential broadband access network.

II. METHODOLOGY

A. Target network, traffic data collection and processing

The network involved in this study is a medium-sized municipal residential broadband access network in Sweden [10]. There are approximately 2600 households connected to the network. The network is an open network, meaning that there are several ISPs that the connected households can choose from, and each ISP offers a set of subscription types with the maximum symmetric access speed at 100 Mbit/s.

The traffic data collection tool was PacketLogic, a commercial IP traffic monitoring and management equipment featuring currently more than 1600 application signatures in its traffic identification engine [11]. Apart from traffic application identification, PacketLogic can also be used as a traffic data filter to select and dump IP packets of a selected application, e.g., YouTube, using a pre-set filter rule based on the relevant properties of the targeted traffic IP packets. In this work, a filter rule based on the information contained in the uniform resource locator (URL) of the Hypertext Transfer Protocol (HTTP) requests was used to identify and dump the data packets containing YouTube watching requests. More specifically, the filter rule was set as

URL = *youtube.com/watch\?v=*.

Once this filter rule was enabled in PacketLogic, Transmission Control Protocol (TCP) session packets matching this filter rule were dumped to a separate data server. Nevertheless, due to the practical storage limitation of the data server, all the payload packets were dropped in the packet dump, due to the large amount of data of the video clips.

After the packet dump, the raw traffic data was further parsed to extract the relevant information of each captured YouTube 'watch' request record, such as the time of request, the HTTP user agent, IP/MAC addresses of the household etc. In particular, each 'watch' record contained a video clip ID that YouTube assigned, which can be used to retrieve from YouTube the clip property parameters such as the video clip category, duration etc, the so-called 'metadata' of the requested video clips. In our work, the extracted YouTube video 'watch' ID was sent to YouTube Application Programming Interface (API) to retrieve the video metadata using PHP scripts for relatively faster parsing of the retrieved metadata. Worth noting here is that in the process of retrieving the video clip metadata, ~ 5% of the extracted video 'watch' IDs returned empty metadata. Consequently, these video clips (that returned empty metadata) were dropped in the subsequent analyses of the results.

For collecting the traffic data, the PacketLogic probe was installed at the Internet Edge (IE) aggregation point of the residential municipal network, where the service providers are connected to the network (even so, due to the continuous upgrade and re-configuration of the network, not all the households in the network might be covered by PacketLogic during the studied period of time) [10]. After parsing the dumped data packets and retrieving the corresponding metadata from YouTube, the obtained video watch request records and video parameters were transferred into a MySQL data base for final traffic statistics aggregation analyses. Note that when transferring data into the MySQL data base server, all the IP and MAC addresses were hashed, ensuring the protection of the end user integrity.

B. Data trace and terminologies

In this work, a 3 week packet dump during 2012-06-08 – 2012-07-05 was carried out in the Swedish municipal network. Since IP addresses were dynamically allocated to each household during the measurement period, MAC addresses were used to identify each household. In addition, in order to distinguish different users in a household, the combination of the MAC address and the extracted HTTP user agent string was used to represent an end user, i.e., we regard each detected different end user device as a unique end user.

As might already be perceived by attentive readers from the descriptions above, in this work, due to the practical limitation of the data storage server, we used the captured YouTube request records to analyze the YouTube traffic statistics in the studied residential network. Three aggregated statistical parameters were used in the analyses of the obtained results in regarding the video clip categories and durations:

1) Total requests: the total number of captured YouTube watch request records

2) Total videos: the total number of different video clips among the total requested video clips

3) Repeated requests: the difference between the total number of requests and videos

Accordingly, to avoid confusions, we define in this paper the concepts of a 'request' and a 'video' as:

1) Request: a requested YouTube video clip

2) Video: a unique video clip

III. REQUESTED YOUTUBE VIDEO CLIP CATEGORY AND DURATION DISTRIBUTIONS

Table I summarizes the total number of requests, videos and repeated requests captured during the 3 week period of time as well as the number of households and end user devices that generated the corresponding YouTube watch request records. In total, 133941 YouTube watch request records were extracted from the dumped packets, of which 89380 different YouTube video clips were requested, resulting in a total of 44561 repeated requests, exactly 1/3 (33.3%) of the total requests. The captured request records were generated from 2195 households in the studied network by 6321 end user devices (note here that the detected daily household and end user device numbers were at ~ 1/3 of those total numbers), of which 98.4% were PC end user devices. The reason for this is that for most of the home Wi-Fi mobile end user devices the operating systems were not windows based, therefore the preset

filter rule (URL = *youtube.com/watch\?v=*) did not match the URLs in HTTP (YouTube watching) requests sent from non-windows mobile devices that had a slightly different "watch" tag. Hence the results reported in this paper mainly represent YouTube watch request statistics generated by PC end user devices. Worth noting also is that the diurnal time distribution of the recorded YouTube watch requests in this work follows the typical 24-hour diurnal time distribution of the general web-browsing Internet traffic, i.e., the requests steadily increased between 6:00 and 17:00 and reached the peak hours between 17:00 and 22:00, after which the number of requests dropped quickly.

TABLE I. CAPTURED YOUTUBE WATCH REQUEST RECORDS OVERVIEW

Name	Total number
Household (distinct Mac address)	2195
End user device (MAC + user agent)	6321
Total requests	133941
Total videos	89380
Repeated requests	44561

TABLE II. RECORDED REQUEST AND VIDEO CATEGORY SHARES

Category name	Total request share	Total video share	Repeated request share
Music	37.2%	31.8%	48.3%
Entertainment	13.7%	14.4%	12.2%
Film	7.9%	7.6%	8.7%
People	7.9%	8.3%	7.1%
Games	7.2%	8.3%	5.1%
Comedy	6.0%	6.4%	5.2%
Sports	3.1%	3.7%	2.0%
News	2.4%	2.8%	1.6%
Education	2.3%	2.6%	1.7%
Autos	2.1%	2.5%	1.2%
Tech	1.8%	2.1%	1.0%
Howto	1.8%	2.1%	1.0%
Shows	1.4%	1.6%	1.1%
Travel	0.9%	0.9%	0.8%
Animals	0.7%	0.8%	0.5%
Nonprofit	0.5%	0.6%	0.3%
Movies	0.2%	0.1%	0.2%
Trailers	0.02%	0.02%	0.02%
Unknown	3.0%	3.4%	2.1%

A. *Captured YouTube request and video category distributions*

Table II lists the YouTube video categories and their corresponding shares of the recorded total requests, total videos, and repeated requests, respectively. Note that the video categories presented here were derived directly from the retrieved YouTube clip metadata, i.e., the video clip category definition and classification were according to YouTube's practice. From the table one can see clearly that the music videos dominated, accounting for 37.2% of the total recorded YouTube watch requests. Moreover, the repeated requests for music video clips accounted to almost 50% of the total repeated requests, suggesting that the contribution of music video clips to the total potential network caching gain be even more dominant, as discussed in the next section. Apart from Music, other major video clip categories were Entertainment, Film, People, Games and Comedy, with their corresponding total request shares being all over 5%. Another interesting feature is that for music and film videos, the shares of repeated requests were higher (film) or much higher (music) than those of the total requests, suggesting that these two categories have relatively higher potential network caching gains as compared to other major categories, as will also be described in more details in the next section.

B. *Captured YouTube request and video duration distributions*

Fig. 1 shows the video clip duration distributions for all the recorded video clips and their corresponding cumulative distribution function (CDF) curves, for both requests and videos. One can see that in general, YouTube videos with durations between 3 and 4 minutes were mostly popular (accounting indeed for over 17% of all the requested video clips). Besides, the CDF curves of requests and videos were almost identical. Over half (56%) of the requested videos had durations below 5 minutes, and 80% were below 10 minutes. Likewise, over 10% of the requested video clips were longer than 15 minutes. This appears to be largely due to the cancellation of the 15-minute video uploading limit in September 2011 (with only a limit of 20 GB in file size).

Note here again that the video durations presented in this paper were retrieved from the video clip metadata using the extracted video clip ID in the HTTP requests, and hence they might not represent the actual playing time of the corresponding video clips.

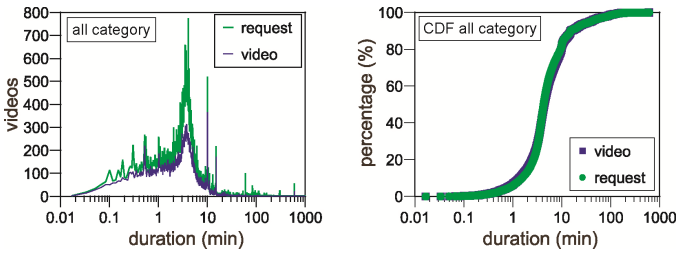

Fig. 1. Recorded video clip duration distribution – all categories.

As a step further, based on the retrieved video clip metadata from YouTube, one may also break down to the level of video categories to gain further insight into the video clip duration distributions and their corresponding CDFs of each individual video category. Fig. 2 shows the corresponding results for the top 6 video categories that had more than 5% of the total requested video shares, for both requests and videos. One can see again that the CDF curves of requests and videos were almost identical even at the video category level. The most striking feature in Fig. 2 is that for music videos, an almost symmetric pulsed duration distributin exhibits, with the peak values at slightly less than 4 minutes. Consequently the corresponding CDF curves feature a normal distribution, showing that ~ 95% of the requested music video clips were less than 10 minutes long. Since music video requests amounted up to 37% of the total video requests, the

impact of music video clips to the overall video duration distribution was also significant, as can be perceived in Fig. 1. Apart from music videos, other major categories show more scattered duration distributions, especially for film and games videos that almost 40% of the reqested video clips were over 10 minutes, while for videos belonging to categories of Entertainment, People and Comedy, 80% of the requested video clips were less than 10 minutes in length.

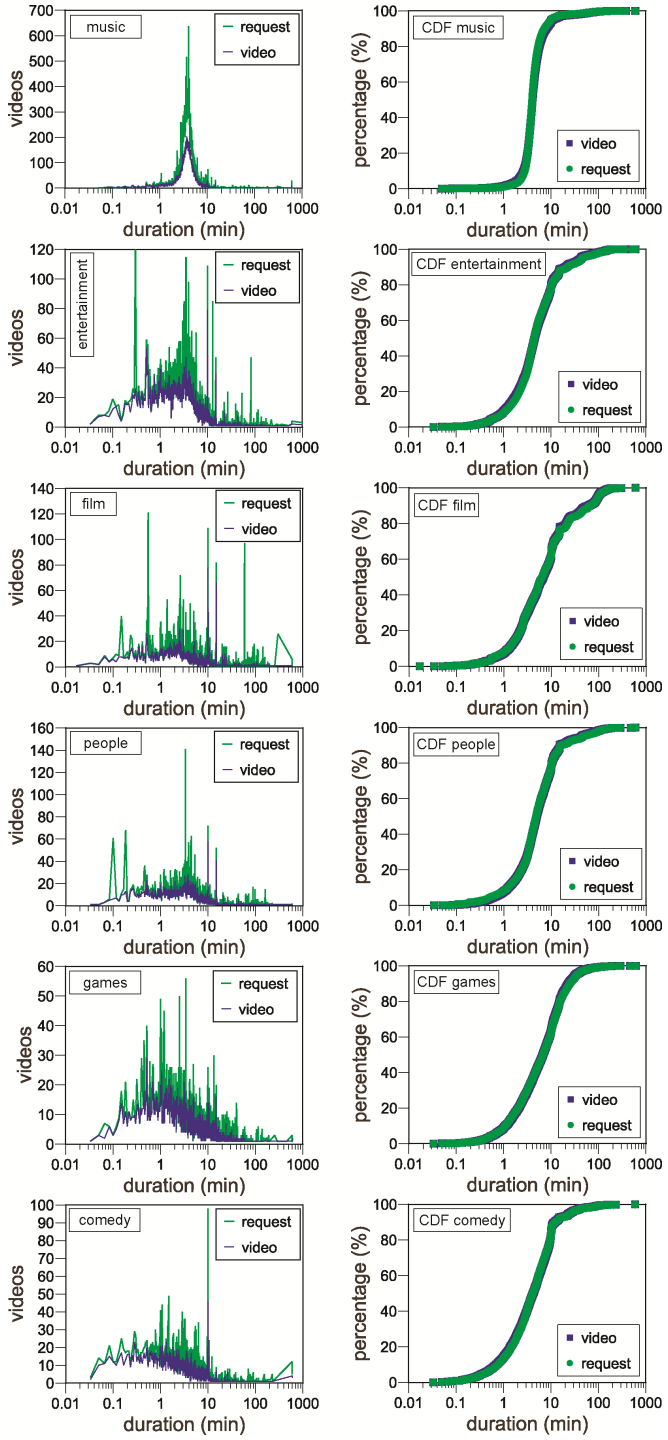

Fig. 2. Recorded video clip duration distributions for categories Music, Entertainment, Film, People, Games and Comedy.

IV. POTENTIAL LOCAL NETWROK CACHING GAINS

In this section, we analyze the recorded YouTube video clip watch requests in regarding the potential local network caching gains. In doing this, in order to understand the cacheabilities of YouTube traffic in the studied network in the perspective of video clip category and durations, we assume that an infinitely sized cache in the studied network was available. An infinitely sized cache means that all the video clips that had been requested by end users would be cached forever in the studied network, and in the meantime, all the YouTube watch requests from the network would be analyzed by the infinitely sized cache to decide whether to forward the request to the YouTube sever or not depending on if a requested video clip could be found in the cache or not. In this way, the maximum or the potential local network caching gain can be calculated simply as

$$caching\ \ gain = \frac{repeated\ \ requests}{total\ \ requests} \qquad (1)$$

for the overall requested YouTube video clips or for a specific video category or for the video clips falling into a certain video duration interval.

A. *Potential network caching gains in regarding YouTube video clip categories*

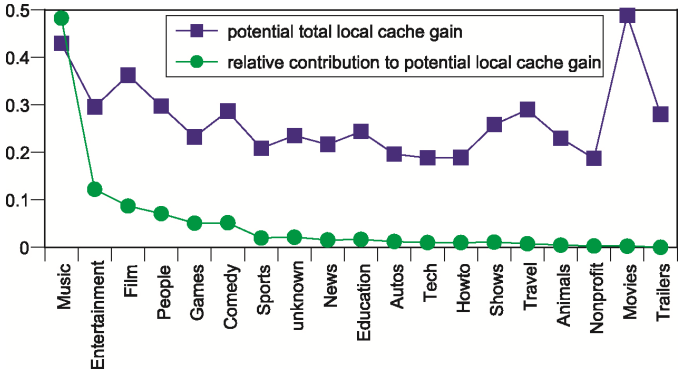

Fig. 3. Potential network caching gains of different video clip categories and their relative contributions to the total potential network caching gain.

Fig. 3 shows the overall potential local network caching gains in regard to video categories, calculated using Eq. (1). For reference, the repeated request shares of each category in Table II are also shown here. The repeated request share of a category represents basically the *relative contribution to the overall network potential caching gain* defined in Eq. (1), as denoted in the figure. One can again see immediately that music videos played the dominant role contributing to almost 50% to the overall network potential caching gain. In the meantime, the music videos, together with film and movie videos, also had relatively higher potential caching gains, despite the movie videos indeed contributed little to the overall potential network caching gain. On the other hand, the Sports, News and those knowledge oriented categories (Autos, Tech and Howto) show relatively lower potential caching gains.

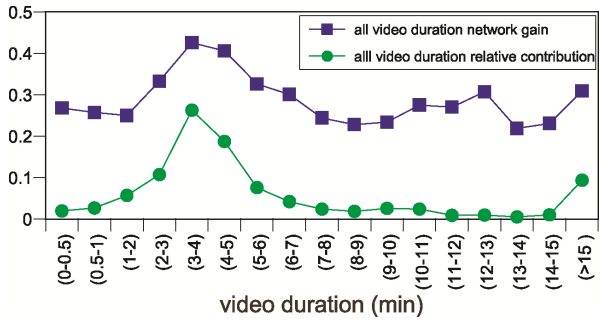

Fig. 4. Time distributed potential network caching gains of all video categories and their relative contributions to the total potential caching gain.

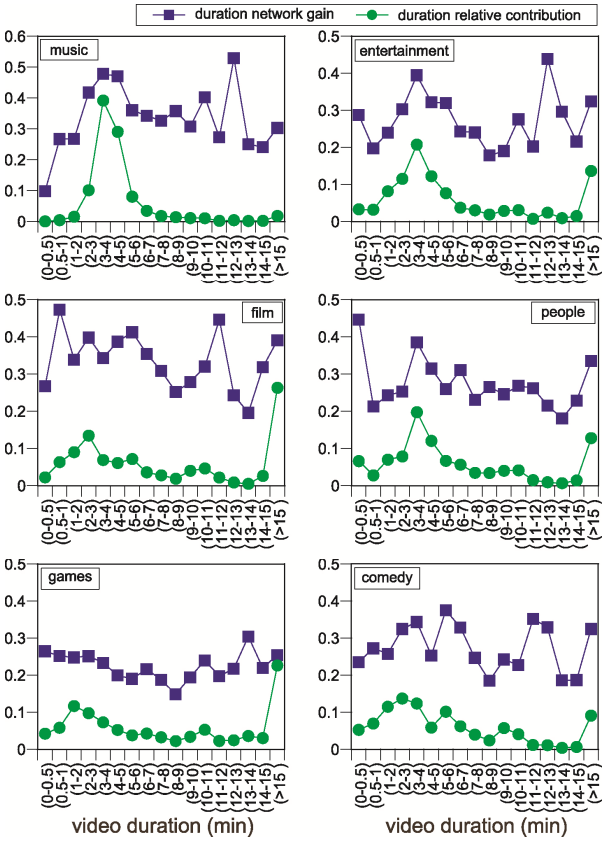

Fig. 5. Time distributed potential network caching gains of different video categories and their relative contributions to the total potential network gain.

B. Potential network caching gains in regarding YouTube video clip durations

We now analyze the potential network caching gains in regarding the video clip durations. Firstly, Fig. 4 shows the potential network caching gains versus the requested video clip durations aggregated at one-minute interval for videos below 15 minutes in duration and for all the videos longer than 15 minutes. For easy reference and comparison, the relative contributions, i.e., the repeated request shares at each time interval are also shown. From this figure one can see that video clips within the time intervals of 2-5 minutes not only dominated the relative contributions to the overall network potential caching gain, but also had higher caching gains themselves. In addition, video clips over 15 minutes also had relatively higher caching gains, attributed mainly to the film video clips as shown in Fig. 2.

Down to the video individual category level, Fig. 5 illustrates the potential network caching gains versus video duration intervals for the top 6 video categories (corresponding to those in Fig. 2), together with the corresponding relative contributions to the potential caching gains of respective categories. One can see again that for the dominant music videos, time intervals between 3-5 minutes contributed most of the category potential caching gain, while for film and games videos the time interval over 15 minutes had the largest share.

V. CONCLUSIONS

In this work, YouTube traffic characteristics in a medium-sized Swedish residential municipal network were studied in the perspective of video clip categories and durations, and analyses on the corresponding potential local network caching gains were carried out. Our results show that YouTube video clips that were requested from the end users in the studied network were imbalanced in regarding the video clip categories and durations. The characteristically short-duration (~ 4 min) music videos dominated the YouTube traffic, both in terms of the total traffic share as well as the contribution to the overall potential network caching gain. We believe that our findings provide fundamental knowledge for the development of local caching strategies to offload significantly amount of YouTube traffic from the backbone transport networks, which is the next step of our research along this track.

ACKNOWLEDGEMENT

This work was supported by the Swedish Governmental Agency for Innovation Systems (Vinnova) in the European CelticPlus projects IPNQSIS & NOTTS and in the project EFRAIM, and by the Swedish National Strategic Research Area (SRA) in the project eWIN.

REFERENCES

[1] White Paper, “Cisco Visual Networking Index: Forecast and Methodology, 2010–2015”, June 2011, Cisco
[2] IP Network Monitoring for Quality of Service Intelligent Support (IPNQSIS), http://projects.celtic-initiative.org/ipnqsis/
[3] J. Li, A. Aurelius, V. Nordell, M. Du, Å. Arvidsson, and M. Kihl, “A five year perspective of traffic pattern evolution in a residential broadband access network”, Future Network & Mobile Summit 2012, 4 - 6 July 2012, Berlin, Germany
[4] http://youtube-global.blogspot.se/
[5] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, S. Moon, “I Tube, You Tube, Everybody Tubes: Analyzing the World’s Largest User Generated Content Video System”, IMC’07, San Diego, California, USA
[6] A. Brodersen, S. Scellato, M. Wattenhofer, “YouTube Around the World: Geographic Popularity of Videos”, WWW 2012, Lyon, France
[7] P. Gill, M. Arlitt, Z. Li, A. Mahanti, “YouTube Traffic Characterization: A View From the Edge”, IMC’07, San Diego, California, USA
[8] M. Zink, K. Suh, Y. Gu, J. Kurose, “Characteristics of YouTube network traffic at a campus network – Measurements, models, and implications”, Journal of Computer Networks, Volume 53 Issue 4, March, 2009, Pages 501-514
[9] A. Finamore, M. Mellia, M. Munafò, “YouTube Everywhere: Impact of Device and Infrastructure Synergies on User Experience”, IMC’11, November 2–4, 2011, Berlin, Germany
[10] M. Kihl, P. Odling, C. Lagerstedt, and A. Aurelius, “Traffic analysis and characterization of Internet user behavior”, ICUMT2010, pp. 224 -231
[11] Procera Networks, http://www.proceranetworks.com