YouTube All Around: Characterizing YouTube from Mobile and Fixed-line Network Vantage Points

Pedro Casas, Pierdomenico Fiadino, Arian Bär, Alessandro D'Alconzo
FTW - Telecommunications Research Center Vienna
{surname}@ftw.at

Alessandro Finamore, Marco Mellia
Politecnico di Torino
{surname}@tlc.polito.it

***Abstract*—YouTube is the most popular service in today's Internet. Its own success forces Google to constantly evolve its functioning to cope with the ever growing number of users watching YouTube. Understanding the characteristics of YouTube's traffic as well as the way YouTube flows are served from the massive Google CDN is paramount for ISPs, specially for mobile operators, who must handle the huge surge of traffic with the capacity constraints of mobile networks. This papers presents a characterization of the YouTube traffic accessed through mobile and fixed-line networks. The analysis specially considers the YouTube content provisioning, studying the characteristics of the hosting servers as seen from both types of networks. To the best of our knowledge, this is the first paper presenting such a simultaneous characterization from mobile and fixed-line vantage points.**

***Keywords*—YouTube; Google; Content Delivery Networks; Mobile Networks; Traffic Measurements; EU project mPlane.**

I. Introduction

YouTube is the most popular video streaming service in today's Internet, and is responsible for more than 30% of the overall Internet traffic [1], [2]. Every minute, 100 hours of video content are uploaded, and more than one billion users visit YouTube each month[1]. This enormous popularity poses complex challenges to network operators, who need to design their systems properly to cope with the high volume of traffic and the large number of users. The challenges are bigger for mobile operators, who have to deal with an ever-increasing traffic volume with the capacity constraints of mobile networks, and in a much more competitive market. Indeed, mobile makes up to almost 40% of todays YouTube's global watch time, and video traffic accounts for more than 30% of the downstream peak traffic in large-scale cellular networks such as AT&T in the US [4]. Finally, the provisioning of YouTube through the massive Google CDN [10] makes the overall picture even more complicated for ISPs, as the video requests are served from different servers at different times. The highly distributed architecture and dynamic behavior of large CDNs allow achieving high availability and performance; however, content delivery policies can cause significant traffic shifts in just minutes, resulting in large fluctuations on the traffic volume carried through the ISP network paths.

These observations have motivated a large research effort on understanding how YouTube works and performs [5]–[8], covering aspects such as content delivery mechanisms, video popularity, caching strategies, and CDN server selection policies among others. These papers focus exclusively on YouTube as observed in fixed-line networks. In this paper we take a step further on the characterization of YouTube, additionally considering the impact of the type of network on the specific flow characteristics and provisioning behavior of the underlying servers. In particular, we perform a comparison of how YouTube is provisioned in fixed-line and mobile networks, analyzing three days of YouTube traffic traces collected in both networks. The insights of our analysis are particularly useful for the ISP, who usually has a hard time in figuring out where are the problems of the service delivery when their customers experience poor performance with YouTube. In the EU project mPlane[2] we are developing a global Internet-scale measurement platform to better understand and diagnose performance degradation events in large-scale services such as YouTube, and this study provides rich input to better develop the measurement and analysis processes.

The main contribution of this paper is providing a first analysis of YouTube from both fixed-line and mobile vantage points. To the best of our knowledge, this is the first paper presenting such a simultaneous characterization of YouTube. In particular, we find out that the wide-spread usage of caching in mobile networks provides high benefits in terms of delay to the contents as well as downlink throughput. In addition, we identified marked variations on the delay from the fixed-line vantage point to the YouTube servers, suggesting either a widely spread and heterogeneous server farm behind the YouTube front-ends, or the presence of a highly dynamic path-changes policy in the interconnection to the preferred YouTube servers.

The remainder of this paper is organized as follows: Sec. II describes the datasets we use, and reports the analysis on the servers infrastructure providing YouTube in both networks, particularly studying the latency to the video contents and the provisioning dynamics. Sec. III analyzes the characteristics of the YouTube traffic as observed in both networks, as well as the delivery performance in terms of downlink throughput from the different Autonomous Systems (ASes) hosting YouTube videos. Finally, Sec. IV concludes this paper.

II. YouTube Hosting Infrastructure

Google replicates YouTube content across geographically distributed data-centers worldwide, pushing content as close to end-users as possible to improve the overall performance

The research leading to these results has received funding from the European Union under the FP7 Grant Agreement n. 318627, "mPlane".

[1] http://www.youtube.com/yt/press/statistics.html

[2] http://www.ict-mplane.eu/

978-1-4799-5280-9/14/$31.00©2014 IEEE

Autonomous System	# IPs	#/24	#/16
All server IPs fixed-line	3646	97	22
15169 (Google)	2272	60	2
43515 (YouTube)	1222	12	1
36040 (YouTube)	43	2	2
All server IPs mobile	2030	63	10
15169 (Google)	1121	38	2
43515 (YouTube)	844	15	2
LISP	35	4	3
36040 (Google)	26	5	3

Table I. NUMBER OF IPS AND PREFIXES HOSTING YOUTUBE, AS OBSERVED IN BOTH FIXED-LINE AND MOBILE NETWORKS.

of the video content provisioning, minimizing the effects of peering point congestion and enhancing the user experience. Google maintains a latency map [10] between its servers and network prefixes aggregating geographically co-located users, in order to redirect their requests to the closest server in terms of latency. In addition, it employs dynamic cache selection strategies to balance the load among its servers, handle internal outages, manage scheduled updates and migrations, etc.. In this section we study this complex infrastructure for the case of YouTube, as observed from two different vantage points, located in a fixed-line network and a mobile network.

A. Traffic Datasets

The two datasets correspond to almost 90 hours (from Monday till Thursday) of YouTube flows collected at two major European ISPs during the second quarter of 2013. In the fixed-line network, the monitored link aggregates 20,000 residential customers accessing to the Internet through ADSL connections. Flows are captured using the Tstat passive monitoring system [14]. Using Tstat filtering and classification modules, we only keep those flows carrying YouTube videos. In the mobile network, flows are captured at the well known Gn interface, and filtered using the HTTPTag traffic classification tool [13] to keep only YouTube flows. To preserve user privacy, any user related data (e.g., IMSI, MSISDN) are removed on the fly, whereas any payload content beyond HTTP headers is discarded. Both datasets are imported and analyzed on-line through the data stream warehouse DBStream [15]. Finally, using the server IP addresses of the flows, the complete dataset is complemented with the name of the ASes hosting the content, extracted from the MaxMind GeoCity databases[3].

B. Server Infrastructure Hosting YouTube

Table I reports the number of unique server IPs serving YouTube in both networks, as well as the ASes holding the major shares of servers. To understand how these IPs are grouped, the table additionally shows the number of IPs per different network prefix. Even if the number of customers associated to the mobile network traces is much larger than in the fixed-line network, the number of unique server IPs observed in the latter is almost the double, with more than 3600 different IPs in the 90 hours. In both cases, two Google ASes hold the majority of the IPs (i.e., AS 15169 and AS 43515), grouped in a small number of /16 subnets. In the mobile network we also include the observed IPs of the Local ISP (LISP), which plays a key role in the delivery of YouTube, due to the extensive usage of content caching. Indeed, it is very common in mobile networks to have forwarding caches at the edge of the network to reduce latency and speed up content delivery [4]. Even though the impact of video caching on the Radio Access Network is limited, ISPs might prefer to reduce the load on the transport network to both reduce peering costs and improve closeness to the content.

(Network) Autonomous System	% bytes	% flows
(FL) 15169 (Google)	80.8	77.3
(FL) 43515 (YouTube)	19.1	22.5
(M) LISP	69.3	66.7
(M) 15169 (Google)	30	32.7

Table II. NUMBER OF BYTES AND FLOWS PER AS HOSTING YOUTUBE IN FIXED-LINE (FL) AND MOBILE (M) NETWORKS.

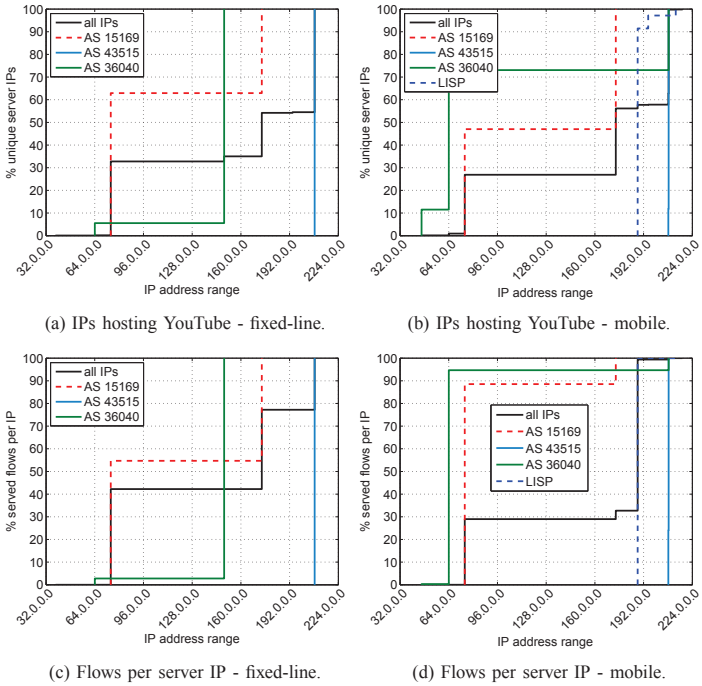
Table II shows that about 80% of the YouTube volume and number of flows are served by the AS 15169 in the fixed network, and up to 70% of the traffic is served by IPs owned by the LISP in the mobile network. This correlates pretty well with the fact that about 65% of the HTTP video content observed in the mobile network of AT&T in the US can be cached at the edge in standard forwarding proxies [4]. Still, we can not say from our analysis whether these IPs correspond to content caching performed by the LISP or also to Google servers deployed inside the ISP, which is a common approach followed by Google to improve end-user experience, known as Google Global Cache (GGC)[4]. In fact, a large share of YouTube content is normally transparent to middle boxes, as videos are marked as "no-cache". We plan to further study this in the future.

To appreciate which of the aforementioned IP blocks host the majority of the YouTube flows, figure 1 depicts the distribution of the IP ranges and the flows per server IP. According to figures 1(c) and 1(d), the majority of the YouTube flows are served by two or three well separated /16 blocks in the fixed-line and mobile networks respectively. Interestingly enough, only a limited fraction of YouTube traffic is served from AS 43515 in the mobile network. Figure 2 additionally depicts the number of flows served per IP in both networks. Separated steps on the distributions evidences the presence of preferred IPs or caches serving a big number of flows, which are most probably selected by their low latency towards the end customers.

Finally, we study the dynamics of the traffic provisioning from the aforementioned ASes. Figure 3 depicts (a,b) the number of active IPs and (c,d) the flow counts per hour (normalized) in both networks during three consecutive days. In both networks, the active IPs from either AS 43515 or AS 15169 show an abrupt increase at specific times of the day; for example, about 200 IPs from AS 43515 become active daily at about 10:00 in the fixed-line network, whereas IPs from AS 15169 almost triple at peak hours (between 17:00 and 23:00) in the mobile network. Note that the number of active IPs from the LISP is constant during the whole period, showing their main role in the delivery of YouTube flows. In terms of flow counts, figure 3(c) evidences a very spiky behavior in the flows served from AS 43515, and some of

[3]MaxMIND GeoIP Databases, http://www.maxmind.com.

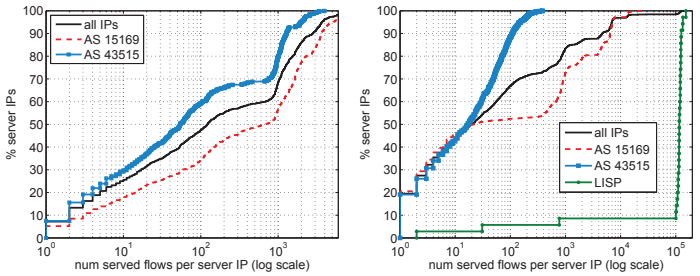
[4]https://peering.google.com/about/ggc.html



(a) IPs hosting YouTube - fixed-line. (b) IPs hosting YouTube - mobile.

(c) Flows per server IP - fixed-line. (d) Flows per server IP - mobile.

Figure 1. IP ranges distribution and flows per server IP hosting YouTube. The majority of the YouTube flows are server by very localized IP blocks.



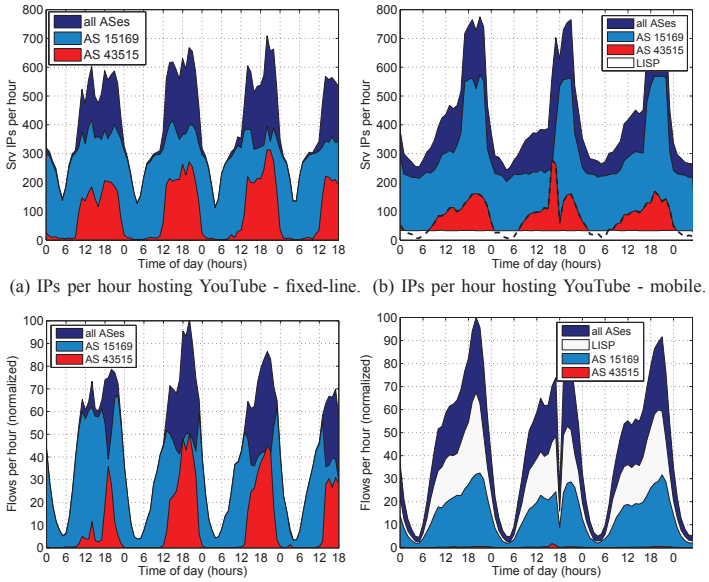
(a) Flows per IP in fixed-line. (b) Flows per IP in mobile.

Figure 2. Flows per IP and per AS. Clear sets of IPs serve a large share of the flows, evidencing the presence of preferred caches.

the load balancing policies followed by Google in the region of the fixed-line ISP, e.g., a drastic switch from AS 15169 to AS 43515 of the flows served at about 18:00. In the mobile network, the LISP servers handle the majority of the flows daily, and as a consequence, the dynamics of the flow counts are much smoother. This indirectly implies that the load forecasting from each of the servers is much straightforward in the mobile network, resulting in a potentially much easier traffic management at the core network.

C. How Far are YouTube Videos?

We investigate now the latency and the location of the previously identified servers, considering the distance to the vantage points in terms of Round Trip Time (RTT). The RTT to any specific IP address consists of both the propagation delay and the processing delay, both at destination as well as at every intermediate node. Given a large number of RTT samples to a specific IP address, the minimum RTT values are an approximated measure of the propagation delay, which is directly related to the location of the underlying server. It



(a) IPs per hour hosting YouTube - fixed-line. (b) IPs per hour hosting YouTube - mobile.

(c) Flow counts per hour - fixed-line. (d) Flow counts per hour - mobile.

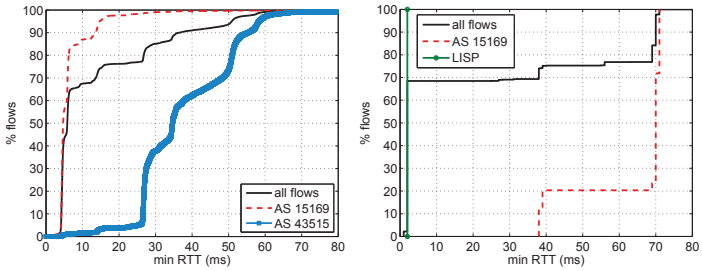
Figure 3. IPs and flows per hour during 90 hs. The glitch in the flow counts in the mobile network is caused by maintenance of the monitoring probe.

follows immediately that IPs exposing similar min RTT are likely to be located at a similar distance from the vantage point, whereas IPs with very different min RTTs are located in different locations.

RTT measurements are passively performed on top of the YouTube flows in the fixed-line network. Mobile networks usually employ Performance Enhancement Proxies (PEPs) to speed-up HTTP traffic, and therefore, passive min RTT measurements on top of HTTP traffic provide incorrect results [12]. We therefore consider an active measurement approach in the mobile network, running standard pings from the vantage point to get an estimation of the min RTT to the servers. We then weight the obtained min RTT values by the number of flows served by each IP to get a rough picture of where the flows are coming from, similar to [11].

Figure 4 shows the distribution of the min RTT values for the flows observed in both networks. Steps in the CDF suggest the presence of different data-centers or clusters of co-located servers. Figure 4(a) shows that about 65% of the flows in the fixed-line network come from servers most probably located in the same country of the ISP, as min RTT < 5 ms. This is coherent with the fact that Google selects the servers with lower latency to the clients. A further differentiation by AS reveals that the most used servers in AS 15169 are located much closer than the most used servers in AS 43515. As depicted in figure 4(b), the lion share of the flows in the mobile network comes from the LISP servers, which are located inside the ISP (min RTT < 2 ms). The rest of the flows served from AS 15169 are located at potentially two geographically different locations, one closer at around 40 ms from the vantage point, and one farther at about 70 ms.

The richness of the passive RTT measurements performed in the fixed-line network permits to further study the dynamic behavior of the servers' selection and load balancing strategies used by Google to choose the servers. Figure 5(a) depicts the variation of the distribution of min RTT measured on the YouTube flows for a complete day, considering contiguous



(a) min RTT (passive) in fixed-line. (b) min RTT (active) in mobile.

Figure 4. min RTT to servers in different ASes. Latency is passively measured on top of the YouTube flows in the fixed-line network, whereas active RTT measurements are performed in the mobile network.

time bins of 3 hours length. Correlating these results with those in figure 3(c) permits to better understand the daily variations. Whereas the majority of the flows are served from very close servers until mid-day, mainly corresponding to AS 15169, servers in farther locations are additionally selected from 14:00 on, corresponding to the increase in the number of flows served from AS 43515.

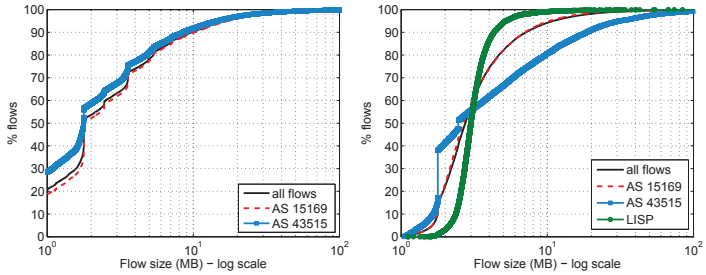
Finally, figures 5(b) and 5(c) reveal a very interesting pattern which could be potentially harmful for the performance of the video delivery, but that we were not able to diagnose in current paper. The figures depict the min RTT values observed during a complete day for flows hosted at different IPs in two /24 subnets at AS 15169 and AS 43515, namely 74.125.13.0/24 and 208.117.250.0/24 respectively. The interesting observation is that the min RTT to the same set of IPs varies with a very structured pattern, presenting different clusters of min RTT values in both subnets. For example, min RTT values of 5, 9, and 14 ms are systematically observed for the flows served from IPs at 208.117.250.0/24.

These marked variations could be the result of strong and very periodic congestion events, which is in fact very unlikely. We tend to believe that either a very spread and heterogeneous server farm behind the YouTube front-end servers in the corresponding IPs, or the presence of a highly dynamic path-changes policy in the interconnection to the specific YouTube servers is the origin of such a behavior. A deeper study of these patterns is left for future work.

III. YouTube Traffic and Performance

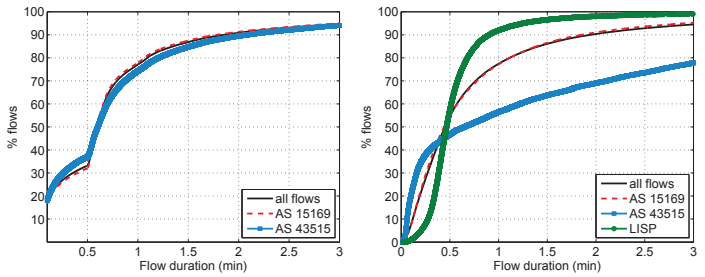
We study now the characteristics of the YouTube flows as observed from both vantage points, as well as the performance achieved in terms of downlink throughput. Figure 6 depicts the distribution of flow size for the different hosting ASes. Figure 6(a) shows that about 20% of the flows served in the fixed-line network are smaller than 1 MB, and that flows served by the AS 43515 are slightly smaller than those provided by the AS 15169 in this network.

The CDF reveals a set of marked steps at specific flow sizes, for example at 1.8 MB and 2.5 MB. Our measurements and studies performed in [3] reveal that YouTube currently delivers 240p and 360p videos in chunks of exactly these sizes, explaining such steps. A similar behavior is observed for chunks of bigger sizes. About 75% of the flows are smaller than 4 MB, 90% of the flows are smaller than 10 MB, and a very small fraction of flows are elephant flows, with sizes higher than 100 MB.



(a) YouTube flow size - fixed-line. (b) YouTube flow size - mobile.

Figure 6. YouTube flows sizes. The steps in the CDF at sizes 1.8 MB, 2.5 MB, 3.7 MB, etc. correspond to the fixed chunk-size used by YouTube to deliver videos of different resolutions and bitrate.

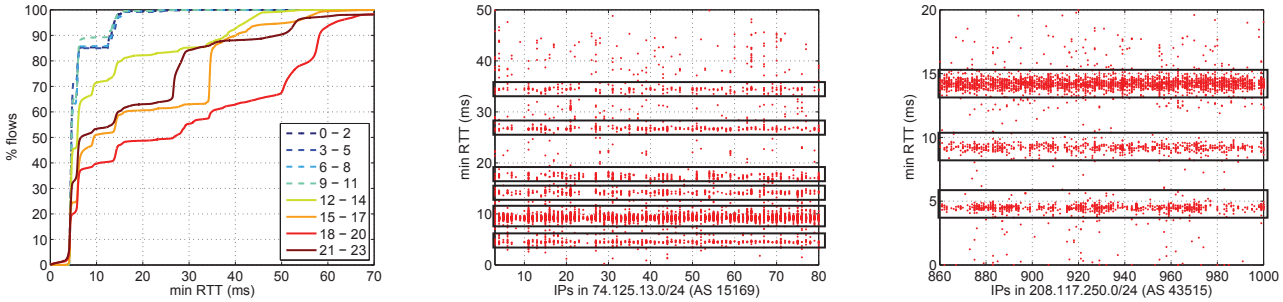


(a) YouTube flow duration - fixed-line. (b) YouTube flow duration - mobile.

Figure 7. YouTube flows duration. About 85% of the flows observed in both networks are shorter than 90 seconds. A large share of flows have an average duration of about 30 seconds.

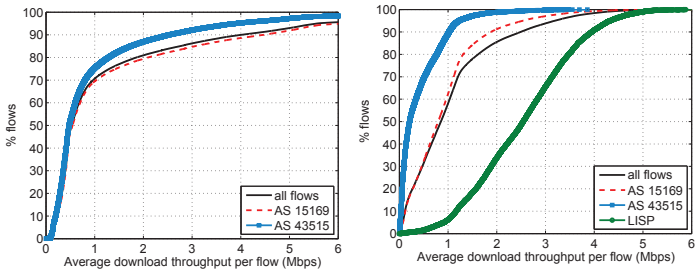
The flows considered in figure 6(b) for the mobile network are only those with a size bigger than 1 MB. This filtering is performed as a means to improve the estimation of the downlink throughput in our traces. Surprisingly, the flows served by the AS 43515 in the mobile network tend to be rather larger than those provided by the other ASes, and more than 20% of the flows served by this AS are bigger than 10 MB. The interesting observation comes when analyzing the size of the flows served by the LISP. The CDF reveals a very concentrated flow size between 2 MB and 4 MB, suggesting that the cached contents (or those served by YouTube servers inside the ISP) could potentially cover, at least in terms of flows size, 75% of the flows observed in the fixed-line network. We have not investigated the characteristics of the YouTube videos hosted by the LISP IPs and those served in the fixed-line network, which would provide further insights about the type of contents that are potentially cacheable. We plan to do so in future studies, following the approach used in [4].

Figure 7 depicts the distribution of the flows duration, in minutes. The flow duration in both networks is below 3 minutes for about 95% of the total flows. The abrupt step in the CDF of the flows observed in the fixed-line network at about 30 seconds is most probably linked to the aforementioned video chunk sizes, but we were not able to verify this observation. About 85% of the flows observed in both networks are shorter than 90 seconds. Similar to the flow size, the flows served from AS 43515 are rather longer in the mobile network, with more than 20% of the flows lasting more than 3 minutes. Given the small fraction of traffic served from AS 43515 in the mobile network, we can not say for sure that the behavior of the servers in this AS is different when it comes to different types of networks. Still, the important differences in the



(a) min RTT - daily variation. (b) min RTT - servers in AS 15169. (c) min RTT - servers in AS 43515.

Figure 5. min RTT dynamics in the fixed-line network. (a) The server selection strategies performed by Google are not only based on closest servers. (b,c) Strong variations on the min RTT to the same Google IPs suggest the presence of path changes or very heterogeneous latencies inside Google's datacenters.



(a) YouTube throughput- fixed-line. (b) YouTube throughput - mobile.

Figure 8. Average YouTube flow downlink throughput per AS. Flows served by the LISP are the ones achieving the highest performance, evidencing the benefits of local content caching and low-latency servers.

flow characteristics coming from AS 43515 in both networks might suggest some kind of network (or device) awareness on the way YouTube video is provisioned, as observed in [8]. Finally, and also correlating with previous observations, the distribution of the duration of the flows served by the LISP IPs is concentrated around 30 seconds, matching pretty well the aforementioned abrupt step in the CDF of the flow duration in fixed-line networks.

To conclude the study, figure 8 reports the distribution of the average downlink throughput per flow measured in both networks, discriminating by hosting AS. The downlink throughput is the main network performance indicator that dictates the experience of a user watching YouTube videos [9]. Both figures 8(a) and 8(b) consider only flows bigger than 1 MB, to provide more reliable and stable results (i.e., avoid spurious variations due to the TCP protocol start-up). The downlink throughputs achieved in both networks are rather similar, with more than 15% of the flows achieving a throughput above 2 Mbps. This suggests that the downlink throughput is partially governed by the specific video encoding bitrates and the flow control mechanisms of YouTube and not exclusively by the specific access technology. Still, when analyzing the performance results per AS, it is evident that the flows served by the LISP are the ones achieving the highest performance, with an average flow downlink throughput of 2.7 Mbps. This out-performance evidences the benefits of local content caching and low-latency servers for provisioning the YouTube flows.

IV. Conclusions

In this paper we have presented a characterization of the YouTube service from traffic traces captured at both mobile and fixed-line networks. To the best of our knowledge, this is the first paper presenting such a simultaneous characterization of YouTube. Besides describing and analyzing the servers infrastructure hosting the flows, as well as the characteristics of the traffic itself, we have shown that the usage of caching in mobile networks provides high benefits in terms of delay to the contents as well as downlink throughput. We have also identified a very interesting behavior on the latency to the YouTube servers in the fixed-line network, which we are planning to further investigate as part of our future work. We believe that the insights provided in this paper will improve the capabilities of measurements-based frameworks to better diagnose performance issues in large-scale Internet services such as YouTube. In particular, we are applying the insights of this work into the EU project mPlane, developing a large-scale anomaly detection and root cause analysis approach for CDN-based services.

References

[1] C. Labovitz et al., "Internet Inter-domain Traffic", in *SIGCOMM*, 2010.

[2] V. Gehlen et al., "Uncovering the Big Players of the Web", in *TMA*, 2012.

[3] F. Wamser et al., "YouTube Modeling: From Packets to User-Perceived Quality", FTW-TECHREPORT-138, 2014.

[4] J. Erman et al., "Over The Top Video: the Gorilla in Cellular Networks", in *IMC'11*.

[5] P. Gill et al., "YouTube Traffic Characterization: A View From the Edge" in *IMC'07*.

[6] M. Zink et al., "Characteristics of YouTube Network Traffic at a Campus Network - Measurements, Models, and Implications", in *Computer Networks*, 2009.

[7] R. Torres et al., "Dissecting Video Server Selection Strategies in the YouTube CDN", in *ICDCS*, 2011.

[8] A. Finamore et al., "YouTube Everywhere: Impact of Device and Infrastructure Synergies on User Experience", in *IMC'11*.

[9] P. Casas et al., "YOUQMON: A System for On-line Monitoring of YouTube QoE in Operational 3G Networks", in *IFIP Performance*, 2013.

[10] R. Krishnan et al., "Moving Beyond End-to-End Path Information to Optimize CDN Performance", in *IMC'09*.

[11] P. Casas et al., "IP Mining: Extracting Knowledge from the Dynamics of the Internet Addressing Space", in *ITC 25*, 2013.

[12] A. Botta et al., "Monitoring and Measuring Wireless Network Performance in the Presence of Middleboxes,", in *WONS*, 2011.

[13] P. Fiadino et al., "HTTPTag: A Flexible On-line HTTP Classification System for Operational 3G Networks", in *INFOCOM*, 2013.

[14] A. Finamore et al., "Experiences of Internet Traffic Monitoring with Tstat", in *IEEE Network*, vol. 25(3), 2011.

[15] A. Bär et al., "DBStream: an Online Aggregation, Filtering and Processing System for Network Traffic Monitoring", in *TRAC*, 2014.