

Large Scale Characterisation of YouTube Requests in a Cellular Network

Fehmi Ben Abdesslem and Anders Lindgren

SICS Swedish ICT

Stockholm, Sweden

{fehmi, andersl}@sics.se

Abstract—Traffic from wireless and mobile devices is expected to soon exceed traffic from fixed devices. Understanding the behaviour of users on mobile devices is important in order to improve the offered services and the provision of the underlying network. Globally, more than 60% of consumer Internet traffic is estimated to be video traffic, and the most popular video website, YouTube, estimates that mobile access makes up nearly 40% of the global watch time. This paper presents the first work to study the characteristics of YouTube user requests on a nationwide cellular network. This study is based on the analysis of a large dataset generated by 3 million users and collected by a major telecom operator. We show for instance that 20% of the users generate 78% of the requests, and that over 80% of the requests target only 20% of the distinct videos accessed during the data collection period. Our results provide a comprehensive insight into the way people use YouTube on mobile devices, and show a very high potential for video cacheability on the cellular network.

I. INTRODUCTION

We are living in the YouTube generation. People are more and more moving away from watching scheduled broadcast TV and consume video in short on-demand online snippets. This is now reaching the mobile world as well. Over the past few years, network traffic from wireless and mobile devices has grown at a rapid rate and is expected to exceed traffic from fixed devices by 2016, and mobile data traffic will increase 13-fold between 2012 and 2017 [1]. This increase in traffic is to a large extent driven by video content, which is expected to make up 69 percent of all consumer Internet traffic in 2017 (up from 57 percent in 2012). This is becoming an increasing problem for cellular network operators as a large part of their available bandwidth is consumed by video traffic. For these reasons, understanding the video consumption patterns on mobile devices is becoming more and more important, in order to better adapt the network infrastructure and mechanisms to users' needs and behaviours.

A large part of the video content consumed on the Internet comes from sites with user-generated content. YouTube is the largest such user-generated content website, that allows any Internet user to upload their videos and watch videos uploaded by other users. In terms of average number of daily visitors and page views, this website is ranked in the top 5 most popular websites in most countries.¹ This popularity has now been extended to mobile phone users, as most mobile phones are now designed to watch videos on YouTube: featuring a large screen, dedicated applications, advanced browsers supporting

video streaming, and fast cellular standards such as UMTS, HSPA, or LTE. According to statistics from YouTube's website², mobile access makes up almost 40% of YouTube's global watch time, and YouTube is available on hundreds of millions of devices. It is reasonable to estimate that videos watched on mobile phones are generally shorter than the ones watched on a desktop computer, actually making 40% a lower bound on the proportion of requests generated by mobile phones.

In this paper, we characterise the usage of YouTube in cellular networks. Our study is based on a large dataset containing HTTP requests collected by a major telecom operator from all their mobile customers over the country for 41 days. We filtered out requests sent to YouTube servers to analyse them and study how YouTube is used by mobile users. To the best of our knowledge, this is the first large-scale study of YouTube usage on a nationwide cellular network.

The main contribution of this paper is three-fold. First, we provide an insight into the video access patterns of users in a nationwide cellular network, and how this depends on temporal aspects, type of video, and other factors. This complements previous studies based on smaller scale datasets or traditional Internet access. Further, we study how video popularity trends can be classified into different categories and see that we are able to identify videos that *go viral* from sharing in social media. Finally, we discuss the above findings and how such results can be of use for designing new network protocols and mechanisms (such as caching algorithms) and when doing network provisioning to ensure sufficient performance in the network.

The remainder of this paper is organised as follows. Before explaining our methodology to collect the data in Section III, we review related research works characterising online video traffic in Section II. An analysis of the data is then presented in Section IV, where we study in detail the demand patterns of users and the videos requested. Finally we conclude the paper in Section V by showing the implications of our results and their potential applications.

II. RELATED WORK

With the increased popularity of online video services, more attention within the scientific community has been drawn towards these services. Both user generated content (UGC) websites and streaming websites operated by major content providers have been analysed and characterised from different

¹<http://www.alexacom>

²<http://www.youtube.com/yt/press/en-GB/statistics.html>

angles, that are all complementary. In this section, we outline some previous related works and how our study complements them.

As for streaming websites such as Video on Demand (VoD) websites, Yu *et al.* [2] collected and studied the citywide traffic generated by more than 150,000 users over 219 days. Data was collected from the server logs. Catch-up TV is another form of VoD that has been more recently studied. It allows on-demand access of previously broadcast TV content. Nencioni *et al.* [3] analysed consumption patterns of nearly 6 million users of a nationwide deployment of a catch-up TV service (BBC iPlayer) for 8 weeks. Again, data was collected from server logs.

As for UGC websites, Cha *et al.* [4] crawled two popular websites, namely YouTube and Daum, to collect meta information such as the number of views or the ratings of the videos. The data covers several years and contains information on more than 2 million videos. However, there is no information on individual user and access. Zink *et al.* [5] collected data in a different way, by capturing HTTP headers in a campus network, between clients in the campus and YouTube servers. The data covers 8 days and around 5,000 unique clients. Gill *et al.* [6] combined both approaches, by collecting YouTube traffic generated by students and staff of a campus, accessing more than 600,000 videos over 3 years, and crawling the YouTube website for meta information.

Instead of collecting data from the YouTube website or from the local network, some studies relied on data collected in-between by the network operator. This generally allows to study a larger scale of users than when the data is collected from local networks, and provides a better insight on individual behaviours and demand patterns than when the data is crawled from the YouTube website. For instance, a recent study by Arvidsson *et al.* analysed the demand patterns for YouTube of 35,000 clients over several weeks [7].

All these works share a similar methodology of analysis with our study, as they all aim to provide a data-driven characterisation of the demand patterns to better understand the resulting traffic generated by the users. However, there are much less studies focusing on YouTube access from mobile phones, which is becoming more and more popular with the technological improvements of mobile devices.

The most related work involving mobile phones is the study by Finamore *et al.* [8], that compares YouTube traffic generated by mobile devices to the one generated by desktop computers, using a dataset of more than 35,000 unique users and 900,000 videos. This dataset was collected both from local networks (university campuses), and from core networks by ISPs. However, this dataset was generated from WiFi traffic instead of cellular network traffic.

III. DATASET AND METHODOLOGY

The analysis performed in this paper is based on a large-scale dataset of cellular network traffic traces from a major European operator that were collected on a national scale over a period of 41 days between December 2011 and January 2012. The traces contain the URL of around 30 billion HTTP requests together with timestamps and an anonymized identifier of the user sending it.

In this study, we focus on the usage behaviour and access patterns of users viewing YouTube videos over the cellular network. We are thus only interested in users who sent at least one request to YouTube during the collection period. We extract the video identifiers from the request URLs to identify the different videos requested by users. There are 3 million YouTube users (roughly 30% of the total number of users appearing in the dataset). These YouTube users have sent around 75 million requests for 10 million distinct videos.

Out of the 961 hours between the first and last request collected, 9 hours are missing from the dataset: two hours on the 21st of December, three hours on the 2nd of January, two hours on the 22nd of January, and one hour on 23rd and 25th of January. Although those missing hours are noticeable, for instance, when looking at the amount of requests on the days with missing hours, we believe that they are not changing any of our conclusions.

IV. ANALYSIS

What are YouTube video access patterns like for users in a cellular network? Is it similar or different from users in fixed networks? Do most users even access YouTube on a frequent basis from their mobile device, and are there obvious temporal patterns in the way users access YouTube content over the cellular network? Do these accesses follow the same popularity trends as for YouTube views globally, such as viral spread of videos among users? Does a majority of popular videos exhibits some common characteristics that can be useful for predicting the popularity of future videos?

In this section, we answer these and many more questions by analysing the dataset described in the previous section from different angles. First we take a detailed look at user activity in general, including understanding the regularity and intensity of usage for different users. We then analyse the popularity of the videos and draw conclusions regarding the distribution of video popularity. Finally, we take a closer look at the properties of the videos requested (such as video category or age), and study popularity trends and viral growth of videos.

A. User Activity

For a given time period, we define a unique user as a user who sent at least one request for any YouTube video during that time period. Figure 1 shows the number of unique users and requests for each day. The number of requests sent by these users per day is roughly proportional to the number of unique users. Note that the first and last day are not complete, as the data collection started in the afternoon of the first day and stopped in the afternoon of the last day, explaining the lower number of requests collected on these two days. The missing hours in our dataset also explain some gaps observed, on 2nd of January for instance. We observed that although no hours are missing from the 11th to the 16th of January, the dataset was lighter on these days, explaining the gaps observed on these days.

Despite covering winter holidays, the dataset does not show any decrease nor increase of activity during the whole collection period, in terms of number of users or number of requests. All gaps observed are due to missing data in the dataset. A similar observation can be made for the bank

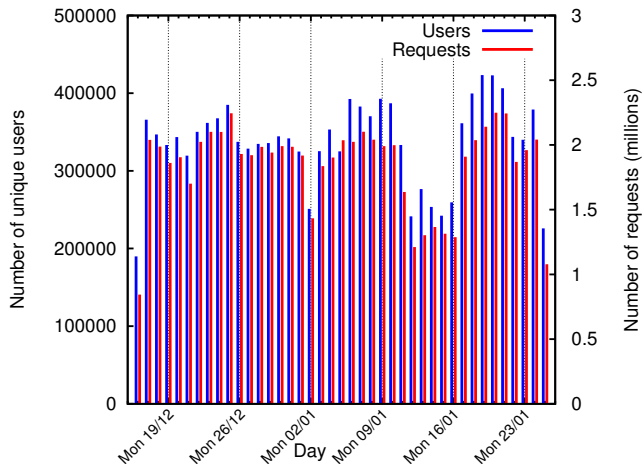


Fig. 1. Number of unique users sending at least one YouTube request for each day, and number of requests per day.

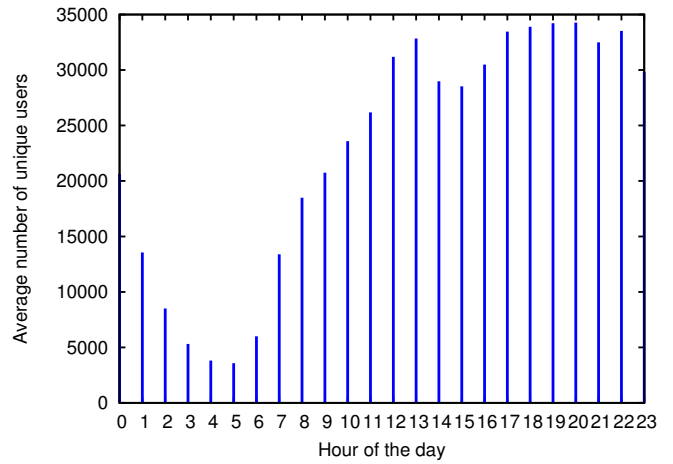


Fig. 3. Average number of unique users sending at least one YouTube request for each hour of the day. In addition to observing the normal diurnality of users, it seems that YouTube usage is most prevalent during the lunch break and in the evening.

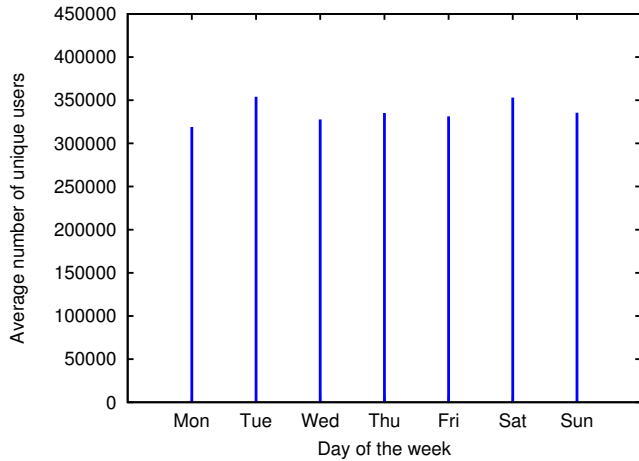


Fig. 2. Average number of users sending at least one request to YouTube for each day of the week. No clear difference between different days of the week is observable, contrary to some other services which have noticeable difference between weekdays and weekends.

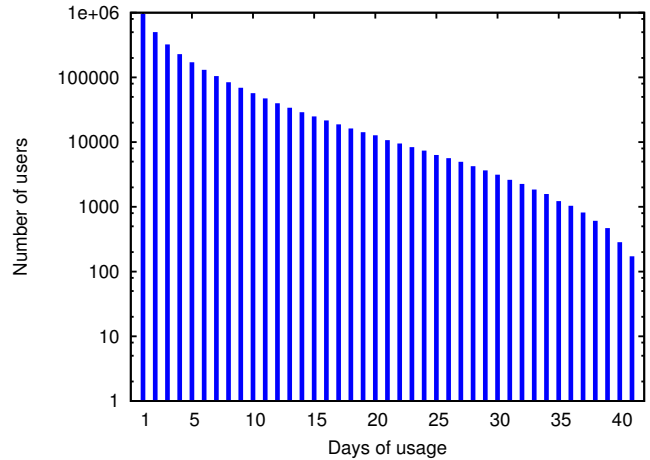


Fig. 4. Distribution of users depending on the number of days of activity.

holidays (25th December and 1st January). Figure 2 shows that there is no significant difference between weekdays and weekends.

However, the number of unique users varies significantly during the day, as we expected. Figure 3 shows the average number of active users for each hour of the day. As expected, there is a major reduction in user activity at night after midnight when most people are sleeping. During the daytime, we can see that the most activity happens during the evenings, and around noon, indicating that people use YouTube as a recreational activity during their lunch break.

While we have a similar number of users active most days, all users are however not accessing YouTube content every day. Figure 4 shows the distribution of the users against the number of days they are active, highlighting that only 172 users used YouTube on their cellular device every day. The cumulative distribution shown in Figure 5 reveals that approximately 32% of the users only used Youtube on their mobile phone for only one day, and 81.5% of them used YouTube less than one week

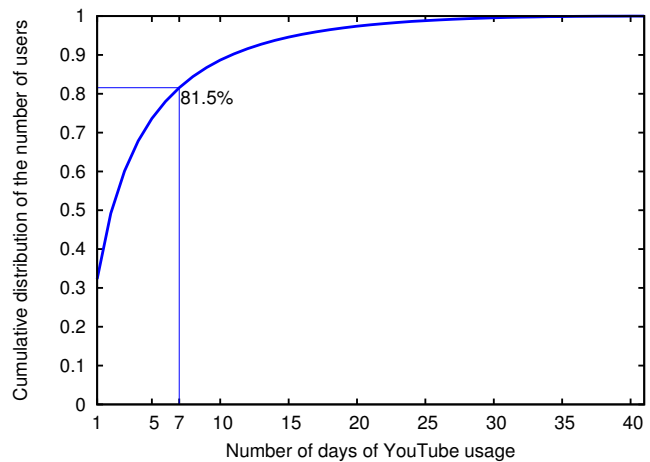


Fig. 5. Cumulative distribution function of the users per number of days of activity. 81.5% of the users used their mobile device on less than 7 days during the collection period.

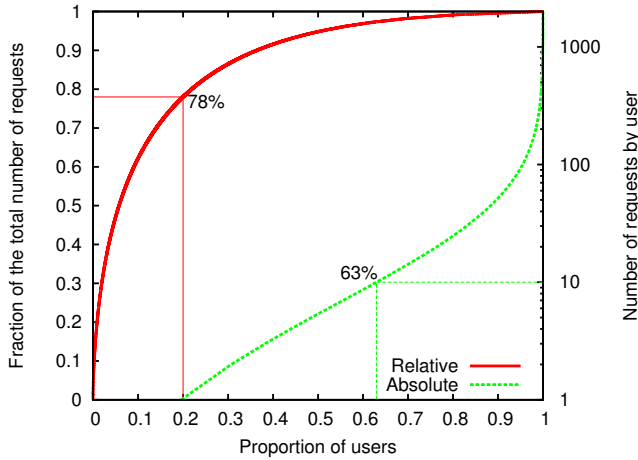


Fig. 6. Amount of requests generated by users, expressed with both their absolute value, and their relative value (compared to the total number of requests). A mere 20% of the users are responsible for 78% of the total number of requests to YouTube, and a majority (63%) of the users made at most 10 requests for YouTube videos during the collection period.

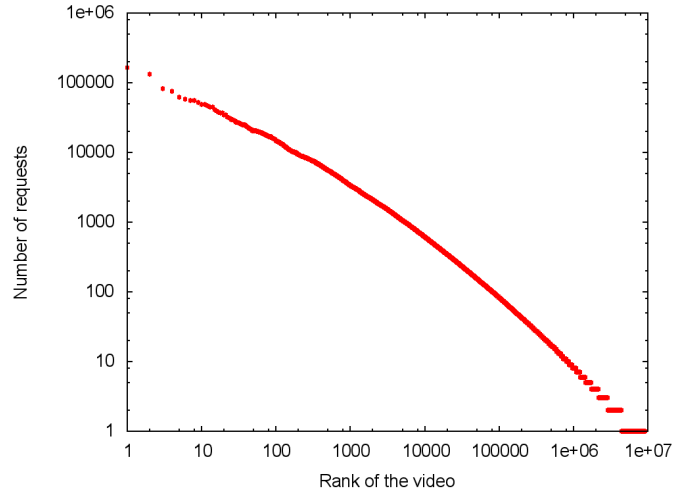


Fig. 8. Distribution of the number of requests by users. The almost straight line shown in this log-log plot indicates that the distribution follows a power law. We also numerically verify that the distribution is fitting a Zipf distribution.

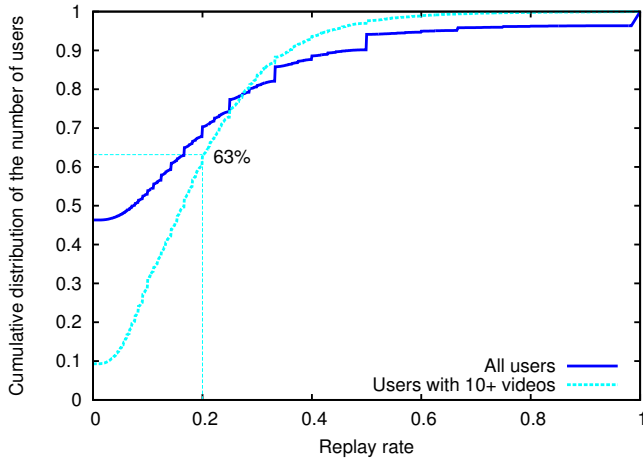


Fig. 7. Distribution of users depending on their replay rate. Out of the users requesting at least 10 different videos, 37% of the users replayed more than 20% of their videos.

during the collection period of our dataset. This shows that watching YouTube videos on a mobile device is generally not a frequent daily activity for a large majority of users.

There is also a large variation of the number of requests depending on the users. Figure 6 shows that 20% of the users generate 78% of the total number of requests. In the same graph, we also plot the distribution of the absolute number of requests at most generated by a fraction of the users. This shows us that many users only make a few requests for video content, and as many as 63% of the users make 10 or fewer requests for YouTube videos during the collection period.

Users requesting the same piece of content multiple times cause unnecessary network traffic. If such content can be stored locally in the user device instead of fetching it from the source every time, there is potential to save bandwidth and provide the user with an improved quality of experience as the latency to start the video will be lower and the risk of annoying

disruptions in playback will be reduced or eliminated. To understand how often users access the same video multiple times, and thus how large the potential waste of network resources by sending the same bytes multiple time is, we study the replay rate of users. For a given user, we define a replayed video as a video that has been played more than once by that user. We then define the replay rate as the proportion of replayed videos for that user. We observe that nearly half of the users request replayed videos. Figure 7 shows the distribution of the users against the proportion of replayed videos. However, as we could see in Figure 6 many users only make a few requests for videos (and with some of these possibly being replays, the number of distinct videos per user might be even lower). For these users, the potential gain from any bandwidth savings possible is very small. Therefore, it is more interesting to focus on users requesting at least 10 different videos. For this subset of the users, around 90% of the users replayed at least one video, and 37% of the users replay more than 20% of the videos at least once.

B. Popularity Distribution

Web content popularity is known [9] to follow a Zipf law in many different contexts, and the same behaviour is expected for other types of content as well. Thus, in this section, we study this distribution to provide an insight on the popularity of individual videos. We first rank the 10 million videos in decreasing order of popularity (view count in our dataset). For a Zipf distribution, the frequency of the video of rank k should follow:

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^N 1/n^s}$$

Where N is the total number of videos and s the value of the exponent characterising the distribution. This function follows a straight line when plotted on a log-log scale. Figure 8 shows that when videos are ranked in decreasing order of their number of views in our dataset, their popularity follows a Zipf distribution, characterised by a straight line. This is confirmed

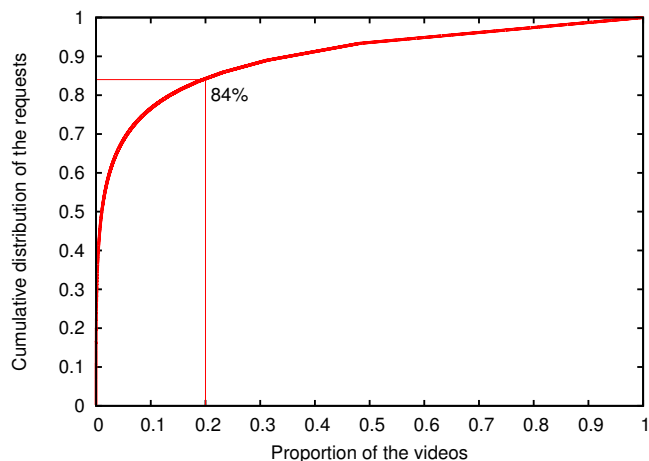


Fig. 9. Cumulative distribution of the videos per number of requests. A small fraction of the videos contribute to the vast majority of the total number of requests.

by running a regression, that gives a coefficient of power law of 1.07. We note that the Zipf law does not hold in other YouTube datasets [10] (obtained by crawling the website) or other UGC video websites such as the one studied in [11].

The Pareto principle [12] (also known as the 80–20 rule) is a phenomenon that can be observed in many real-life situation. This principle states that, for many events, roughly 80% of the effects come from 20% of the causes. To test if the principle also holds for YouTube requests in cellular networks, we show in Figure 9 the distribution of the requests. In particular, we observe that 20% of the videos are targetted by 84% of the requests. Together with the similar result already observed in Figure 6, this confirms previous results [2] showing that the Pareto principle is verified.

C. Videos

Videos on YouTube are classified in pre-defined categories. When uploading a video, users choose the category that describes the best their video. Using the publicly available YouTube API, we collect a set of metadata, including categories and creation time, for all the videos in our dataset. The videos requested in our dataset belong to 21 different categories.

Figure 10 shows the proportion of the videos for some popular categories. The Music category is the most popular, followed by the Entertainment category. Table I shows further statistics about these categories. Nearly 30% of videos requested on YouTube are music videos. Private videos are video for which access has been restricted by the uploader. However, the API also returns this value for videos that have been removed due to either the violation of the terms of use, or to the termination of the uploder’s ccount. We believe that a majority of the videos listed as private are indeed videos that have been removed due to copyright violations rather than actual private videos. Similarly, a tenth of the videos could not be found when querying the API two years after the

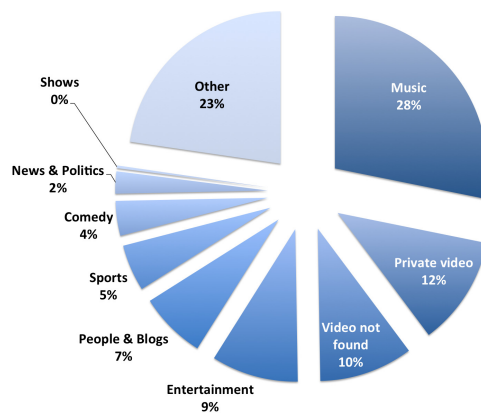


Fig. 10. Pie chart representing the proportion of videos per category. A large number of videos belong to the Music category.

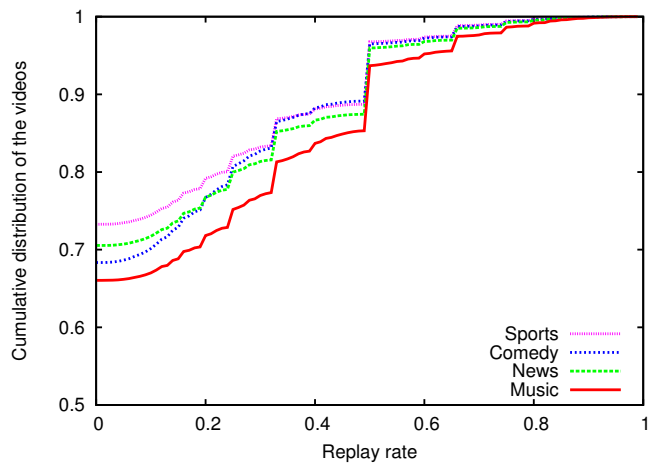


Fig. 11. Cumulative distribution of videos for each category, depending on their replay rate.

data collection period. These videos have either been removed by the uploader, or by YouTube for breaching the terms of usage. For these two categories, we cannot determine the real category that those videos had when available during the time of our data collection, but they account for approximately 22% of the videos requested. The overall distribution of categories indicate that people mainly come to YouTube to be entertained in various manners – only a very small fraction of the videos are in categories such as Education, or News and Politics.

As seen above, almost 30% of the distinct videos were music videos, but when looking at what fraction of the total number of requests the different categories contribute with (as seen in Table I), the Music category gets an even larger share and almost 40% of video requests are targetting a music video. The ranking of requests follows approximately the one of the videos IDs, with some exceptions. For instance, while less than 4% of videos are in the Comedy category, more that 6.5% of the requests are targetting these videos. This suggests that the average number of requests for a comedy video is higher than a sport video for instance, that represent more than 5% of the videos but account for less than 3.6% of the requests.

TABLE I. CHARACTERISTICS OF VIDEOS PER CATEGORY.

Category	Unique IDs	Requests	Replay rate	Median age (days)
Music	29.25%	38.27%	34.91%	457.14
Private video	12.05%	11.65%	29.48%	N/A
Video not found	10.33%	9.52%	29.55%	N/A
Entertainment	9.58%	7.32%	26.08%	435.30
People & Blogs	7.35%	5.16%	27.67%	428.95
Sports	5.13%	3.58%	23.87%	308.19
Comedy	3.83%	6.51%	25.48%	381.72
News & Politics	2.44%	2.34%	24.48%	141.64
Shows	0.30%	0.41%	25.12%	125.96
Other	23.56%	21.74%	25.22%	382.76

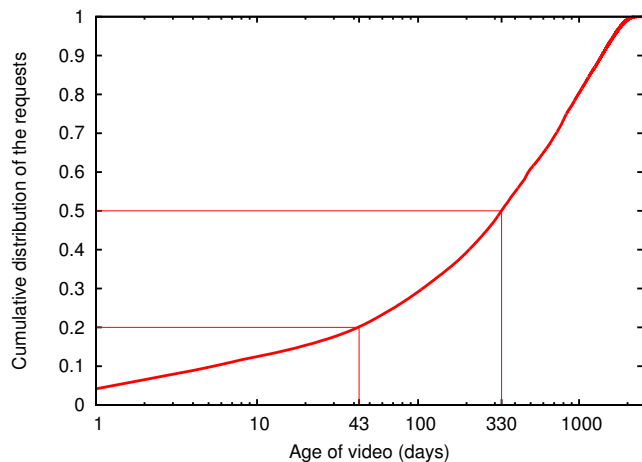


Fig. 12. Cumulative distribution of the requests, depending on the age of the requested video. 20% of requests are for videos that are less than 43 days old, while half of the requests are for videos that are more than 11 months old.

An interesting measure to better understand the characteristics of video categories and the way users watch that type of video is the replay rate. This time, we define the replay rate for a given video as the proportion of requests that has been already sent by a same user. A high replay rate shows that a category contains videos that are often replayed by users. Again, music videos show a high popularity: almost 35% of the requests are targeting a music video already requested by the same user. This is likely due to users employing YouTube as a free music streaming jukebox in order to listen to their favourite music many times. However, users are more likely to request sport and news videos only once: more than 75% of the requests for both categories are not sent again by users. We note that the overall replay rate is intuitively higher than we expected. A requested video is not necessarily played by a user, but can also be requested again to load more comments, or more related videos, for example. Figure 11 shows the distribution of replay rates for different categories. We observe a difference in the proportion of videos played only once (when the replay rate is 0) depending on the category. Whereas 66% of music videos are played once by users, 74% of sport videos are played once. However, for all categories, 10% of the videos have a replay rate greater than 50%. This suggests that videos are prone to be replayed by users, whatever the category they belong to.

Finally, we observe a large variation of the video age, depending on their category. The metadata retrieved from the YouTube API also contained the video creation date, allowing

us to calculate the age of the video at the time it was requested by a user. In Figure 12, the distribution of the age of the video for the different requests in our dataset is shown. It shows that about 20% of the requests are for videos that are less than 43 days old at the time of request, while half of the requests are targeting videos older than 11 months. Table I also shows that while the median age for videos in the Music category and other categories are more than one year old, the median age of a video for a show or a piece of news is only around 20 weeks, as that type of content is usually only interesting for a short period after its original publication.

D. Trends

One of the interesting differences between videos on YouTube in this era of social media as compared to traditional video distribution channels is the possibility for a piece of content to *go viral* and rapidly increase in popularity due to sharing on social media platforms or due to publicity through some other channel. In this subsection we are interested in observing such trends and epidemic effects in popular videos and try to identify the videos in our dataset that have experienced such effects.

Requests for videos can be triggered by different factors. For example, a link can be included on a web page, or in an email, or shared on a social network. The user can also hear or think of a video and search for it directly. Johansen [13] models such human activities by defining the response time of internet users. This is the time between the event causing the video request (for example, the publication of the link on a website) and the actual video request. It can be modelled by a power law time distribution described by:

$$\phi(t) \sim \frac{1}{t^{1+\theta}}, \text{ with } 0 < \theta < 1 \quad (1)$$

Here, the exponent θ can be determined empirically from the data. After this response time, the action of viewing a video can be the cause itself to another similar action by another user. Typically, when a video becomes viral, users watching it share and talk about that video, which will be then requested by other users after a response time. This epidemic behaviour can be modelled by the self-excited Hawkes condition Poisson process [14], using the response time function $\phi(t)$ in (1) as a component:

$$\lambda(t) = V(t) + \sum_{i, t_i \leq t} \mu_i \phi(t - t_i) \quad (2)$$

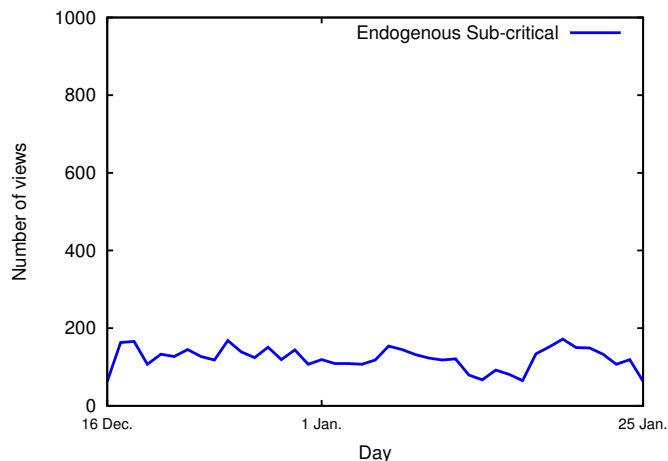


Fig. 13. Example of endogenous sub-critical video: there is no particular peak of popularity. The distribution is closed to the one of a Poisson process.

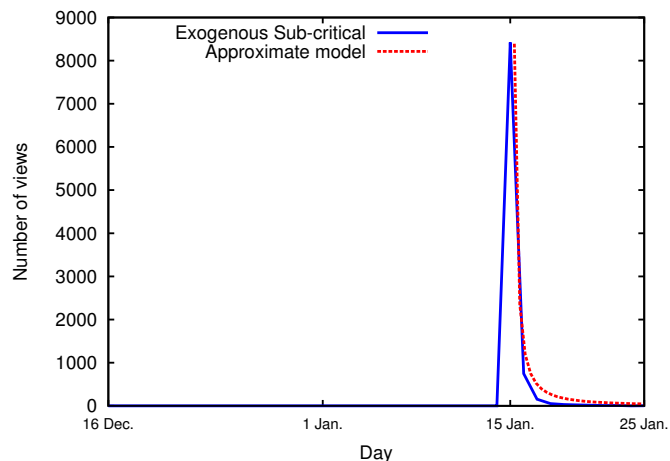


Fig. 15. Example of exogenous sub-critical video: a sudden burst of popularity is observed around 15 January, followed by a quick loss of popularity.

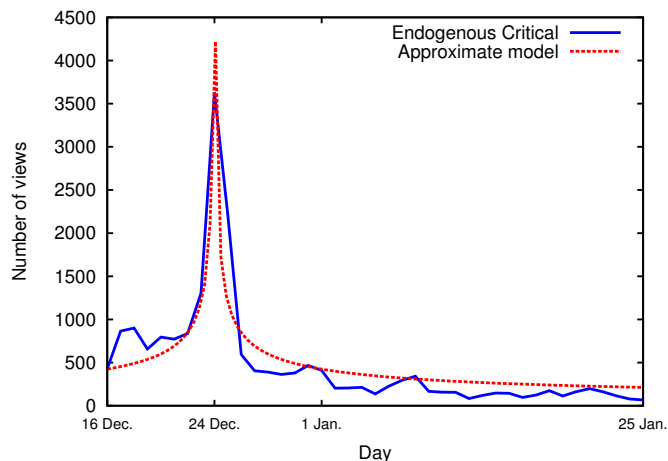


Fig. 14. Example of endogenous critical video: a Christmas video. The number of requests increases until reaching a peak on Christmas eve, and then decreases following a power law.

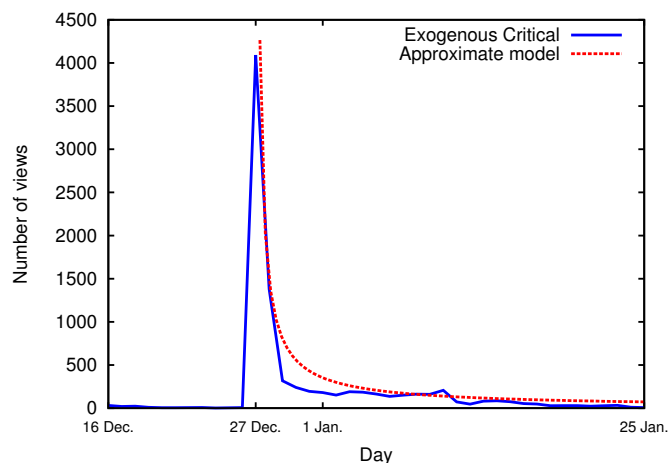


Fig. 16. Example of exogenous critical video: a sudden burst of popularity is observed around 27 December, followed by a slow decrease of popularity.

where μ_i is the number of potential viewers who will be influenced after t_i , which is the time when the user i shared the video. But the instantaneous rate of requests $\lambda(t)$ is not only described by the epidemic process, as the event triggering the requests initially might be still happening: for instance, the link to the video is still available on the web page. Hence, other users that are not taking part of the epidemic phenomenon might still request the video. For this reason $V(t)$ is added as a component to the model to capture the views that are not triggered by the epidemic effect.

From the rate of requests described by (2), Crane and Sornette [15] define two classes of videos, subdivided into 2 sub-classes:

Endogenous, when the requests are mainly driven by the video itself without external factors (eg, without publication on a website). Endogenous videos can be:

- 1) **Sub-critical (Figure 13)**, when there is no strong epidemic effect (μ_i is small). In that case the requests will obey a simple stochastic process, and we will not

observe any trend in popularity.

- 2) **Critical (Figure 14)**, when the views of users influence and trigger more requests from other users. We will then observe a growing number of views until a maximum after which the number of views will slowly decrease. Sornette and Helmstetter [16] show that such activity is approximated by $\frac{1}{|t-t_c|^{1-2\theta}}$.

Exogenous, when the requests are mainly driven by external factors such as a publication on a website, an appearance in a popular TV show, or a coverage on the news. Exogenous videos typically experience a sudden burst of views, followed by a gradual decrease that depends on the sub-category:

- 1) **Sub-critical (Figure 15)**, when there is no strong viral effect. In that case the video will quickly cease to be popular and such activity is approximated by $\frac{1}{(t-t_c)^{1+\theta}}$ [16].
- 2) **Critical (Figure 16)**, when the viral effect makes the decrease of popularity much slower. Such activity is approximated by $\frac{1}{(t-t_c)^{1-\theta}}$ [16].

The three last sub-classes can be modelled with a power law function of different exponent. The number of requests of exogenous videos typically show a sudden burst, before decreasing:

- 1) quickly when there is no viral effect, following a power law with a higher exponent $1 + \theta$
- 2) slowly due to the viral effect, following a power law with a smaller exponent $1 - \theta$

In both cases, because there was usually a very small number of views before the sudden burst, the peak day present a high proportion of views compared to the following days, which is however lower when the following days experience an epidemic phenomenon. Crane and Sornette [15] show that this proportion F can reliably be used to determine the class. F is defined as:

$$F = \frac{\max(v_i)}{\sum_{i=1}^N v_i} \quad (3)$$

where v_i is the number of views for day i , and N is the number of days observed.

As described in Table II, a rule of thumb used in previous works ([15], [17]) is that exogenous sub-critical videos (that quickly lose popularity) present a proportion F greater than 0.8 and exogenous critical videos (that slowly lose popularity) show a proportion F between 0.8 and 0.2. Videos for which $F < 0.2$ are endogenous critical videos. Crane and Sornette show that the classification is insensitive to slight changes to the boundaries 0.8 and 0.2 [15].

To apply these models on our dataset, we only consider popular videos, that have been requested at least 1,000 times in total by all users during the data collection period. This sample of popular videos represents only 1,318 videos (0.014% of the videos), but these videos are still responsible for 22.62% of the requests and are targetted by more than half of the users.

Table III shows the distribution of videos per class. The endogenous sub-critical videos were identified by running a Chi-square goodness-of-fit test on the videos to identify distributions following a Poisson process. We used a p-value threshold of $p < 0.01$ to determine the endogenous sub-critical videos, which constitute the largest group, as found in previous studies. For the three other classes, that all follow a power law, we computed F for each video. We observe that more than half of the popular videos are endogenous critical. Our results are consistent with the results obtained by analysing data from the YouTube website [17]. Note that for all our calculations, we normalised the number of views with the number of total requests collected for each day, to avoid the missing data in some days of our dataset to bias the popularity of videos.

The trends observed in our dataset, although restricted to one European country, can be confirmed in a global scale from the YouTube website. For instance, for a given video, we extract the global view count statistics from YouTube, and

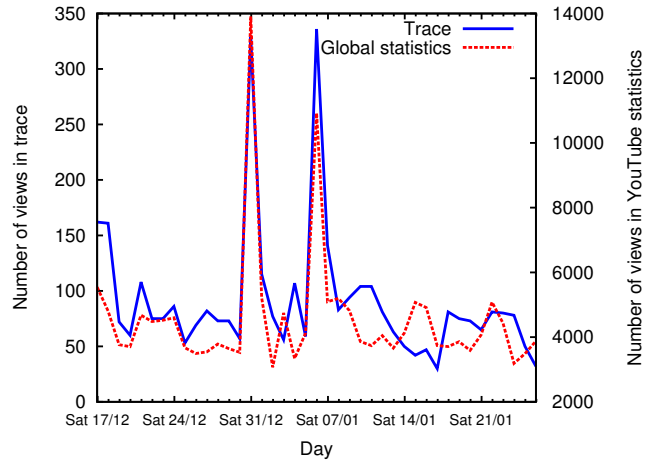


Fig. 17. Viewing statistics for one selected video both from our dataset and from the global YouTube viewing statistics. The popularity of the video in our dataset can clearly be seen to follow the same trends as for the global access statistics for YouTube, in particular with the two very noticeable peaks in popularity occurring at the same time for both data sources.

Figure 17 shows that the two sudden peaks of popularity we observe in our dataset are also visible in a global scale. This gives us increased confidence that our results, albeit from a limited period of time and geographic area, still can give us some insight into YouTube usage on a more global level.

V. CONCLUSIONS

This paper has presented the first detailed measurement study of YouTube demand patterns on a nationwide cellular network. Four key aspects have been analysed on approximately 75 million requests: the users, the distribution of the videos, the video themselves, and the trends of popularity.

Around 30% of the users in our dataset are also seen accessing YouTube over the cellular network. This confirms that the global popularity of YouTube as observed in the Internet in general is also valid in cellular networks. However, the usage of the service varies across users, with only 20% of the users generating 78% of the requests, and 90% of the users accessing the service less than 10 days in a collection period of 41 days.

As for many other types of online content, the Pareto principle is also verified, with 20% of the videos being targetted by 84% of the requests. The different categories of videos present different characteristics. For example, music videos are more popular than other categories, and the videos in the News and Politics category are more recent than entertainment videos at the time of the requests. Finally, the trends of popularity observed in the cellular network are similar to the ones observed globally in previous works.

Our results suggest a high potential for cacheability in the cellular network. For instance, by keeping 20% of the videos in a proxy cache server, 84% of the requests could have been replied locally. For a better efficiency, proxy servers should be kept closer to the users and distribute content to small geographical areas. Future work includes a study of the cacheability of YouTube videos in local proxy servers, by analysing the video demand patterns for each cell.

TABLE II. RULE OF THUMB TO SIMPLY DETERMINE A VIDEO TREND CLASS.

Endogenous critical	Exogenous sub-critical	Exogenous critical
$F < 0.2$	$0.2 < F < 0.8$	$F > 0.8$

TABLE III. DISTRIBUTION OF VIDEOS PER CLASS.

	Endogenous sub-critical	Endogenous critical	Exogenous sub-critical	Exogenous critical
Videos	458	721	4	135
Proportion	34.7%	54.7%	0.3%	10.2%

Building on our results, pre-fetching algorithms for cellular network can also leverage the meta information such as the category in order to improve the caching schemes: for instance by favouring music videos when choosing videos to keep in the cache memory, as they are more popular and more likely to be replayed.

VI. ACKNOWLEDGMENTS

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 246016.

REFERENCES

- [1] Cisco Systems, Inc. (2013, May) Cisco Visual Networking Index: Forecast and Methodology, 2012-2017.
- [2] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," in *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006*, ser. EuroSys '06. New York, NY, USA: ACM, 2006, pp. 333–344. [Online]. Available: <http://doi.acm.org/10.1145/1217935.1217968>
- [3] G. Nencioni, N. Sastry, J. Chandaria, and J. Crowcroft, "Understanding and decreasing the network footprint of catch-up tv," in *Proceedings of the 22nd international conference on World Wide Web*, ser. WWW '13. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013, pp. 965–976. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2488388.2488472>
- [4] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *Networking, IEEE/ACM Transactions on*, vol. 17, no. 5, pp. 1357–1370, 2009.
- [5] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network - measurements, models, and implications," *Comput. Netw.*, vol. 53, no. 4, pp. 501–514, Mar. 2009. [Online]. Available: <http://dx.doi.org/10.1016/j.comnet.2008.09.022>
- [6] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07. New York, NY, USA: ACM, 2007, pp. 15–28. [Online]. Available: <http://doi.acm.org/10.1145/1298306.1298310>
- [7] A. Arvidsson, M. Du, A. Aurelius, and M. Kihl, "Analysis of user demand patterns and locality for youtube traffic," in *Teletraffic Congress (ITC), 2013 25th International*, 2013, pp. 1–9.
- [8] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "Youtube everywhere: impact of device and infrastructure synergies on user experience," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 345–360. [Online]. Available: <http://doi.acm.org/10.1145/2068816.2068849>
- [9] L. A. Adamic and B. A. Huberman, "Zipfs law and the internet," *Glottometrics*, vol. 3, no. 1, pp. 143–150, 2002.
- [10] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos," in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*. IEEE, 2008, pp. 229–238.
- [11] G. Tyson, Y. Elkhatib, N. Sastry, and S. Uhlig, "Demystifying porn 2.0: a look into a major adult video streaming website," in *Proceedings of the 2013 conference on Internet measurement conference*, ser. IMC '13. New York, NY, USA: ACM, 2013, pp. 417–426. [Online]. Available: <http://doi.acm.org/10.1145/2504730.2504739>
- [12] J. Juran, "Universals in management planning and control," *Management Review*, pp. 748–761, November 1954.
- [13] A. Johansen, "Response time of interuants," *Physica A: Statistical Mechanics and its Applications*, vol. 296, no. 34, pp. 539 – 546, 2001. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0378437101002023>
- [14] A. G. Hawkes and D. Oakes, "A Cluster Process Representation of a Self-Exciting Process," *Journal of Applied Probability*, vol. 11, no. 3, pp. 493–503, Sep. 1974. [Online]. Available: <http://dx.doi.org/10.2307/3212693>
- [15] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proceedings of the National Academy of Sciences*, vol. 105, no. 41, pp. 15 649–15 653, 2008. [Online]. Available: <http://www.pnas.org/content/105/41/15649.abstract>
- [16] D. Sornette and A. Helmstetter, "Endogenous versus exogenous shocks in systems with memory," *Physica A: Statistical Mechanics and its Applications*, vol. 318, no. 3, pp. 577–591, 2003.
- [17] F. Figueiredo, F. Benevenuto, and J. M. Almeida, "The tube over time: characterizing popularity growth of youtube videos," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 745–754.