Analysis of User Demand Patterns and Locality for YouTube Traffic

Åke Arvidsson
Packet Technologies
Ericsson Research, Sweden
ake.arvidssson@ericsson.com

Manxing Du, Andreas Aurelius
Acreo Swedish ICT AB, Sweden
{manxing.du, andreas.aurelius}@acreo.se

Maria Kihl
Dept. of Electr. and Inform. Technology
Lund University, Sweden
maria.kihl@eit.lth.se

***Abstract*—Video content, of which YouTube is a major part, constitutes a large share of residential Internet traffic. In this paper, we analyse the user demand patterns for YouTube in two metropolitan access networks with more than 1 million requests over three consecutive weeks in the first network and more than 600,000 requests over four consecutive weeks in the second network.**

In particular we examine the existence of "local interest communities", *i.e.* the extent to which users living closer to each other tend to request the same content to a higher degree, and it is found that this applies to (i) the two networks themselves; (ii) regions within these networks (iii) households with regions and (iv) terminals within households. We also find that different types of access devices (PCs and handhelds) tend to form similar interest communities.

It is also found that repeats are (i) "self-generating" in the sense that the more times a clip has been played, the higher the probability of playing it again, (ii) "long-lasting" in the sense that repeats can occur even after several days and (iii) "semi-regular" in the sense that replays have a noticeable tendency to occur with relatively constant intervals.

The implications of these findings are that the benefits from large groups of users in terms of caching gain may be exaggerated, since users are different depending on where they live and what equipment they use, and that high gains can be achieved in relatively small groups or even for individual users thanks to their relatively predictable behaviour.

I. Introduction

The volume of data traffic in cellular networks has been increasing exponentially for the past few years and this trend is predicted to continue over the coming few years, cf. the Ericsson strategic forecast [1] which for the period 2007–2017 gives CAGR of 50% and 65% for mobile PCs and smart phones respectively such that the traffic per month will exceed one EB (10^{18} B) in 2013 and 2015 respectively. Moreover, a large part of the traffic relates to video and this fraction is growing, cf., *e.g.*, the Cisco global visual network index 2011–2016 [2] which reports that mobile video traffic exceeded 50% for the first time in 2011 and predicts that mobile video will increase 25-fold between 2011 and 2016 and account for over 70% of the total mobile data traffic by 2016.

These figures suggest that the largest potential gains from optimised networks relate to video content such as YouTube. Optimisation in this context typically means reducing and moving demand in space and time by on-demand and/or predictive caching in networks and clients. Various forms of caching are being widely studied and deployed to reduce transit traffic and enhance service performance. Web caching was widely used when the world-wide web emerged, but it gradually lost its glory when more advanced HTTP features were exploited [3]. However, more recent works on cachability in the P2P/BitTorrent network community [4], [5] and for user generated content (UGC)/YouTube, [6], [7] suggest that it is time to return to caching. Furthermore, caching has proved to be a vital technique to cope with the bandwidth constraints of the backhaul links to/from the base stations of mobile networks [8], [9], [10], [11].

The gains from caching are highly dependent on user demand patterns. In [12], YouTube user behaviour for PC and mobile users was investigated. In [13] a three month trace of YouTube traffic was collected in a campus network and found a large potential for caching. The work in [14] used the video meta data provided by YouTube to study the global video popularity distribution over a number of years. They also studied how to make the UGC distribution system more efficient by using caching and P2P techniques. In [15], the authors pointed out a small world phenomena and suggested that once a user plays a video clip, the cache should pre-fetch the directly related video clips as they are very likely to be watched (in the near term).

From these papers, we note that some aspects (like network wide caches) have been dealt with in several studies whereas other aspects (such as regional or local caches) are less well known. Therefore, the aim of this study is to investigate the latter aspects when applied to YouTube traffic. The work in this paper is based on detailed traffic measurements in two metropolitan access networks in Sweden. We investigate user characteristics and locality aspects. Further, we analyse the potential gains of caches covering smaller geographic areas. We show that people living in the same town download more similar content than people living in different towns and that the same phenomenon applies to different districts within towns. Further, we show that high gains can be achieved with terminal caches, since users tend to download the same video clip several times.

II. Measurements and Data

The study is based on measurements made by network operators, that are partners of Acreo Swedish ICT AB as a

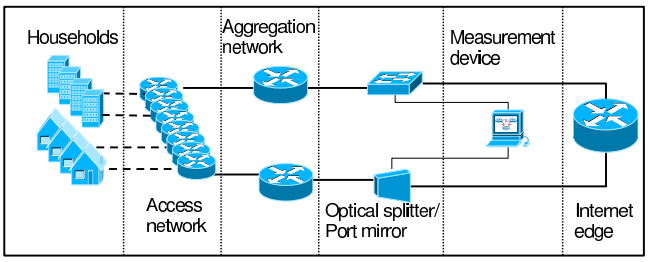

Fig. 1. Network architecture.

part of the IPNQSIS Celtic project. The data originates from two Swedish metropolitan access networks, and the networks and measurements are described below. For privacy reasons, the networks are referred to simply as the north network and the south network respectively in this paper.

A. Networks and Measurements

The data was collected from roughly 5000 households in the north network and 2000 households in the south network. The customers in each network are local residents who can freely choose between different ISPs for access to the Internet. The access speeds range from 1 megabit per second (Mbps) to 100 Mbps depending on the subscription the customer chose. As shown in Fig. 1, the placement of traffic measurement probe in both of the networks is at the edge of the ISP's connection to the Internet.

The measurements were performed with the commercial PacketLogic probe from Procera Networks which performs deep packet and deep flow inspection. The measurements were made in two steps, first, packets passing certain filter rules in the probe were stored in a pcap file and, second, the content of the pcap files was processed to produce request logs. All MAC and IP addresses were anonymised before processing and no data can be traced back to specific users.

B. Data

The data collection was performed during three consecutive weeks, from 00:00 on Monday, January 30, 2012, to 24:00 on Sunday, February 19, 2012, in the north network and during four consecutive weeks, from 00:00 on Monday, January 30, 2012, to 24:00 on Sunday, February 26, 2012, in the south network.

Requests for YouTube clips were identified by "GET video-playback" messages used to request media files, as in [12] which contain unique content identifiers with 16 hexadecimal digits. Some clips were delivered as one large media file (trigged by one request) whereas other clips were delivered as several small media files (trigged by different requests, one for each segment). Users may also alternate the resolution during playback, and each such change will generate a new request with the same identifier. To remove duplicates due to segmentation *etc*., we prevented further counts of content with the same methods as in [12]. Note that such requests are sent also for all previously watched clips except *immediate* replays through the replay button. We remark that the requests for YouTube clips also may be identified "GET watch" messages which are used to download the pages from which the clips are viewed. The different but unique identifiers in these messages enable access to meta data such as content classification, but a severe problem is that YouTube clips may be played in many ways and not all of them include this message. We also remark that YouTube is subject of repeated redesigns hence any attempt to analyse YouTube measurements must include not only data collection but also detailed observations of the current signalling procedures (by, *e.g*., Firebug or similar tools).

TABLE I
REQUEST STATISTICS PER NETWORK.

Network	Content	Overall	Access point	Terminal client
North	Total	1,159,676	203	48.2
	Unique	536,616	94	22.3
South	Total	615,166	294	63.0
	Unique	336,257	161	34.4

The final result contains, for the north network, 1,159,676 requests for 536,616 different clips from 12 geographical districts (identified by operator defined VLAN tags), 5,713 access points (identified by MAC addresses) and 24,059 terminal clients (identified by combinations MAC addresses and web handling agents) and, for the south network, 615,166 requests for 336,257 different clips from 13 geographical districts (identified by manually grouped curbs to which primarily MAC addresses and secondarily DHCP administered IP addresses could be mapped), 2,092 access points (identified by IP addresses) and 9,762 terminal clients (identified by combinations of IP addresses and web handling agents). From the web handling agents we could also differentiate between 26 different browsers (Internet Explorer, Firefox, Chrome, iPhone, iPad, iPod, Android, *etc*.) and deduce 4 different types of hardware (PC, mobile, TV/Playstation and others).

It is noted that the notion of "user" is missing above and the reason is that it is not perfectly clear how to define and detect a user. In this work we will therefore use two different definitions, *viz*. a "terminal client" (distinct combination of MAC or IP address and handling agent) and an "access point" (distinct MAC or IP address).

In the following sections, we will show the results of our analyses. In some sections, only the results for the north network will be shown due the limited space. In these cases, very similar results were obtained for the south network.

III. USER DEMAND CHARACTERISTICS

Distributing the requests over the users, we get the numbers in Table I which suggest that the average terminal client (that requests at least one clip during the measurement period) consumes about two clips per day, one of which is unique.

The numbers in the table do, however, hide a large spread as can be seen in Fig. 2 which is computed by sorting all users in order of their demand and then plotting the accumulated fraction of the demand *vs*. the fraction of users.

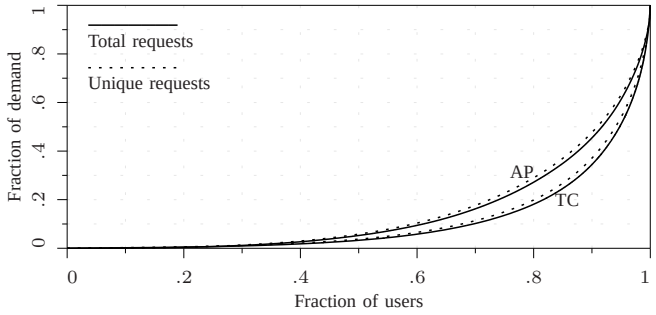

Fig. 2. The distribution of the demand between users in terms of access points (AP, upper curve pair) and terminal clients (TC, lower curve pair) in the north network. Solid curves refer to total requests and dotted curves refer to unique requests.

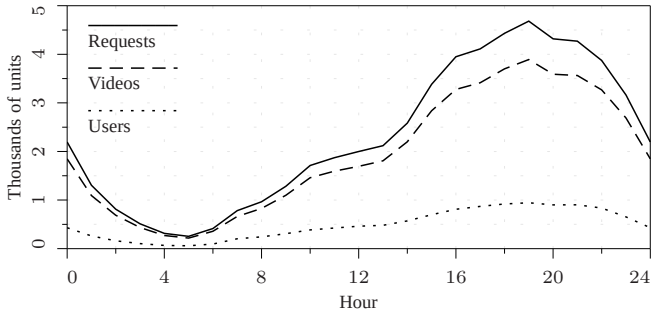

Fig. 3. Average diurnal traffic patterns in the north network.

As can be seen in the figure, 80% of the requests originate from a mere 25% of the access points (20% of the terminal clients), *i.e.* we have a typical heavy tail scenario with a few large consumers and many small consumers. We remark that this spread essentially is the same as the spread between different content items although the latter typically is depicted by plotting demand *vs.* rank in log-log diagrams to obtain Zipf-like charts.

Fig. 3 shows the total number of requests, number of unique videos and number of active users during each hour of a day in the north network averaged over the entire data set. It is seen that the requests steadily increase between 6:00 and 17:00 with a peak between 17:00 and 22:00 after which the number of requests drops rapidly.

IV. Regional Request Characteristics

It is well known that the hit rate of a proxy cache will drop with the number of requests (or, similarly, number of users) the cache is serving. On the other hand, it is also known that there are similarities between requests that originate from users who live close to each other (*e.g.*, in the same country [16]) and/or with common interests (*e.g.*, at a university campus [7]), which means that the efficiency of local caches is a matter not only of how many requests (users) they serve but also of how common or diverse the preferences of these users are.

In this section we examine the extent to which such common interests also occur as a result of geographical proximity. To this end we compare the interests in the two networks and the interests in the different districts of each network.

TABLE II
Hit rates for separated and mixed requests.

Client network	Nominal case	Scaled case	
	Hit rate	Scaling factor	Hit rate
South	0.4370	0.75	0.4056
North	0.4371	0.75	0.4054
Mixed	0.4088	1.50	0.4351

A. Network Request Characteristics

To compare networks we examine the degree of network specific interests by estimating the performance of caches which under comparable circumstances serve users from (i) the north network, (ii) the south and (iii) both networks. Performance is characterised by the probability that an arbitrary clip requested for the first time by an arbitrary terminal client has been requested at least once before by some other terminal client. To make the three cases comparable, we adjust the number of terminal clients in the larger samples (the north network and mixed), first, to the level of the smallest sample (the south network) and, second, scale these values to compensate for different degrees of activity. In formal terms we use two data sets

- $\mathcal{N}$, the set of terminal clients in the north network with cardinality $N = |\mathcal{N}|$ requesting in total C_{N} unique clips, and
- $\mathcal{S}$, the set of terminal clients in the south network with cardinality $S = |\mathcal{S}|$ requesting in total C_{S} unique clips,

and define request rates $\eta_{\mathrm{N}} = C_{\mathrm{N}}/N$ and $\eta_{\mathrm{S}} = C_{\mathrm{S}}/S$ per terminal client, after which we form three sets

- $\mathcal{S}_0$, the entire set $\mathcal{S}$,
- $\mathcal{N}_r$, a randomly selected subset of $\mathcal{N}$ with cardinality $(\eta_{\mathrm{S}}/\eta_{\mathrm{N}})S$,
- $\mathcal{X}_r$, the union of two randomly selected subsets, the first one from $\mathcal{S}$ with cardinality $S/2$ and the second one from $\mathcal{N}$ with cardinality $(\eta_{\mathrm{S}}/\eta_{\mathrm{N}})\, S/2$.

Finally we calculate the user hit rate $h'_{\mathrm{U}}(\Phi)$ for the different sets Φ,

$$h'_{\mathrm{U}}(\Phi) = 1 - \frac{U(\Phi)}{\sum_{\forall u: u \in \Phi} U(u, \Phi)}$$

where $U(\Phi)$ is the number of unique requests in the set Φ $U(u, \Phi)$ is the number of unique requests by user u in the set Φ.

We remark that in this case we use a reduced data set for the south network, where the fourth week has been eliminated, such that the data sets for the two networks overlap completely in time.

The results after 1,000 realisations of the random sets, Table II, show that the comparison appears to be fair (about the same results are obtained for both networks in isolation, cf. the two first rows) and that result of the "mixed network" is not only different but also worse (a lower result is obtained in the comparable, nominal case).

To the last point, note that in the nominal case (with hit rates of 44% for the two separate sets and 41% for the mixed set)

shows that users in the two networks have more in common with users in their own network than with users in the other network. In the scaled case (with hit rates of 41% for the two separate sets and 44% for the mixed set) refers to similar sets except that the cardinalities have been scaled by a factor φ. We thus note that scaling the homogeneous sets by a factor $\varphi = 0.75$ gives hit rates which correspond to the nominal mixed case, and that scaling the mixed set by a factor $\varphi = 1.50$ gives a hit rate which corresponds to the nominal homogeneous case. Noting that the nominal cases correspond to $\varphi = 1.00$ we may say that, in terms the ability to contribute to the hit rate of a cache, a user in the same network brings about 2–3 times as much value as a user in the other network.

(To see these factors, consider the case of one network in isolation with $\varphi = 0.50$ to which we can add either (a) a new set of users from the same network with $\varphi = 0.25$ or (b) a new set of users from the other network with $\varphi = 0.50$. Then note that (a) corresponds to the scaled, homogeneous case, that (b) corresponds to the nominal, mixed case, that two cases perform about the same, and, finally, that the cardinalities of the two added sets differ by a factor of two. Similarly, consider the scaled, nominal cases with $\varphi = 0.75$ to which we can add either (a) a new set of users from the same network with $\varphi = 0.25$ or (b) a new set of users from the other network with $\varphi = 0.75$. Then note that (a) corresponds to the nominal, homogeneous case, that (b) corresponds to the scaled mixed case, that the two cases perform about the same and, finally, that the cardinalities of the two added sets differ by a factor of three.)

B. District Request Characteristics

To compare districts within networks we extract district information for each YouTube request (as outlined in Section II), and compare the request hit rates of unlimited caches serving *particular* districts to those of unlimited caches serving the same number of users but randomly selected from *all* districts. The request hit rate $h_{\mathrm{R}}(d)$ in district d is defined as

$$h_{\mathrm{R}}(d) = 1 - \frac{U(d)}{T(d)}$$

where $U(d)$ and $T(d)$ are the number of unique requests and total requests respectively in district d.

The results for the north network are shown in Fig. 4, and it is seen that the differences between groups of users based on district (solid line) and groups of randomly selected users (dotted line) are small and, in particular, that there is no clear indication of higher hit rates when terminal clients belong to the same district.

To obtain a stronger focus on *users* (as opposed to *requests*) we repeat the above analysis but count hit rate in terms of users rather than requests. That is, request cache hits can be caused by (i) many users requesting the same clip once, (ii) one user requesting the same clip many times or (iii) (more likely) a combination of both of these where a few users request the same clip a few times. User cache hits, on the other hand, can

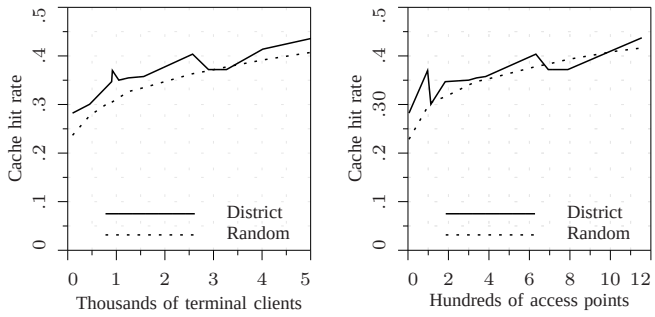

Fig. 4. Request hit rates *vs.* terminal clients (left) and access points (right) in the north network grouped by district and at random respectively.

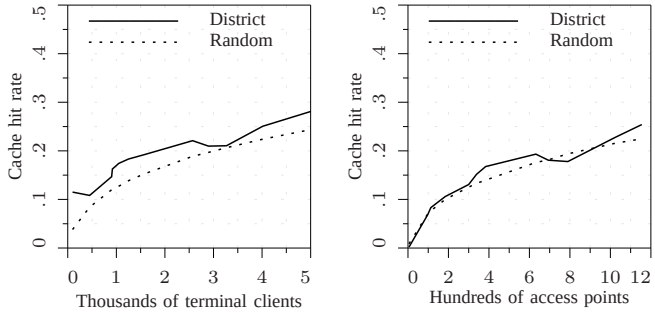

Fig. 5. User hit rates *vs.* terminal clients (left) and access points (right) in the north network grouped by district and at random respectively.

only be caused by many users requesting the clip (at least) once hence this provides a cleaner assessment of possible common interests between users. The user hit rate $h_{\mathrm{U}}(d)$ in district d is thus defined as

$$h_{\mathrm{U}}(d) = 1 - \frac{U(d)}{\sum_{\forall u : u \in d} U(u,d)}$$

where $U(u,d)$ is the number of unique requests by user u in district d.

The results for the north network are shown in Fig. 5, and again it is seen that the differences between users in a district (solid line) and random users (dotted line) are small. To see this, note that the curves for district and random tend to overlap both with respect to access points and terminal clients. Also note that the (subtle) difference between the two cases is explained by the fact that *terminal clients sharing the same access point* have common interests. The difference between the two diagrams is basically that households are split in the left diagram while they are kept intact in the right diagram.

The corresponding results for the south network are shown in Fig. 6, and at a first glance the two sets of results look pretty similar. An important difference, however, is that the differences between users in a district (solid line) and random users (dotted line) in this case are noticeable. To see this, note that the district curves consistently are above the random ones. This means that *terminal clients in the same neighbourhood* have common interests. We believe that the reason for why such commonalities are present in the south network, but not in the north network, is that "regions" are defined differently.

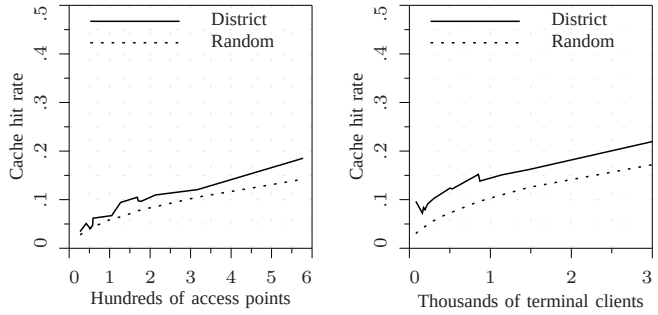

Fig. 6. User hit rates *vs*. access points (left) and terminal clients (right) in the south network grouped by district and at random respectively.

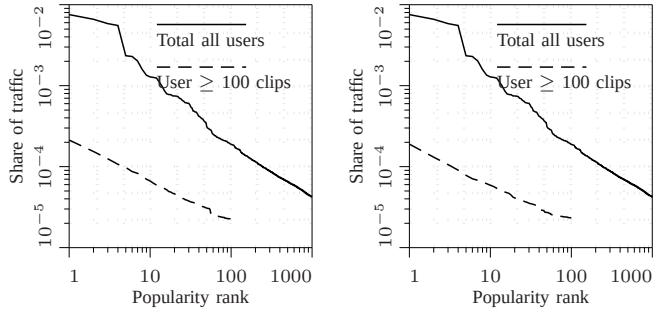

Fig. 7. Traffic share *vs*. popularity rank for the entire group (black solid line) and for individual users with at least 100 distinct requests (dashed line) in the north network. The diagrams refer to access points (left) and terminal clients (right) in the north network.

In the south network, regions are smaller and attempt to capture (manually assessed) areas with common socio-economic factors while, in the north network, regions are larger and (supposedly) reflect network administrative concerns.

Finally it is noted that a significant part of the cache gain is due to repeated requests from the same (set of) user(s). To see this, note the difference between the hit rates in Fig. 4 and Fig. 5 ("intra user gain") and between the right and left diagrams in Fig. 5 and Fig. 6 ("intra household gain").

V. User Request Characteristics

It is well known that the popularity of different objects can be described by Zipf-like distributions. What may be less well known is that this does not only apply to *groups* of users, but also to *individual* users as can be seen in Fig. 7.

The figure shows that the curves for individual users are similar to those of the entire group although we note a flatter slope for individual users than for the entire group.

The hit rates of unlimited user caches, which depend only on the users themselves, are given in terms of averages in Table III. The different hit rates for terminal clients and access points respectively again shows that terminal clients that share the same access point (*e.g.*, persons in a household) tend to have common interests.

A more detailed analysis shows that replay traffic is closely correlated to total traffic both in time and space. This is demonstrated in Table IV which shows the coefficient of correlation between total requests and number of replays per

TABLE III
Common interests between users.

User	North	South
Access point	24.4%	26.5%
Terminal client	21.6%	22.9%

TABLE IV
Correlation between cache gain and YouTube demand.

	Per time		Per user	
User	North	South	North	South
Access point	0.987	0.950	0.987	0.945
Terminal client	0.985	0.942	0.965	0.954

time (measured in five minute periods) and user. The fact that the coefficients are close to unit indicates that times and users associated with many requests in total also are associated with many replays and *vice versa*. Therefore, there may be high gains of using local caching, for example in cellular networks.

To get an idea of the size of a user cache we now examine the characteristics of replays in more detail. To begin we examine how long time clips stay popular, *i.e*. the times after which requests are repeated.

The results are shown in Fig. 8 which depicts the number of observed replays by a terminal client *vs*. time passed since the first observed request from that terminal client.

It is noted that a lot of replays occur soon after the first request (the steep initial slope) and that during the first 24 hours after the initial requests we have small dips after 6 and 18 hours and small peaks after 12 and 24 hours. Over the following days we note a gradually decreasing number of replays, cf. Table V, but with remarkably pronounced peaks every 24 hour after the initial request.

It is, however, important to note that the results are biased in two ways because of the finite observation interval. First, the **interval ends** at some time $t = T$ which means that repeats that occur after, say, one minute can be observed during the entire interval but the first minute, whereas repeats that occur after almost a measurement period only can be observed for a very short time. Second, the **interval begins** at some time $t = 0$ which means that some of the requests seen as the first ones are, in fact, different order replays of initial requests that occurred before the observation interval started.

The *first* effect can be modelled by an "underestimation factor" which, for replays after a time t, amounts to $(T - t)/T$. To see this, note that the numerator is the length of the interval during which the (supposedly) first request must occur while the denominator is the length of the entire interval. The underestimation factor is shown as a dotted line in the diagram, and it is noted that its shape is quite similar to the dropping trend of the observed replays, hence we conclude that the actual drop in popularity may be quite slow.

The *second* effect is illustrated in Fig. 9 which shows the number of unique (left) and total (right) requests over time for content separated by the day it was first seen (1, 2, 5 and 10 respectively). It is seen that that content seen "early" stays more popular than content seen "later". Noting that there is

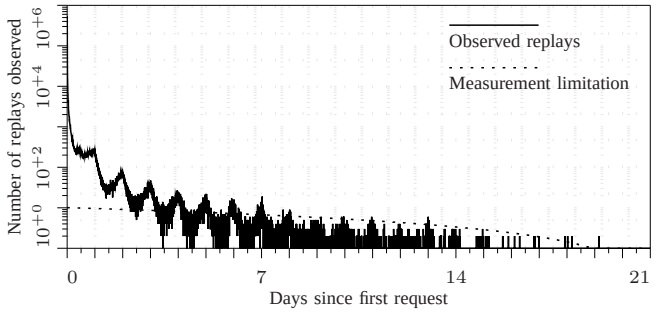

Fig. 8. The number of observed terminal client replays as functions of the time passed since the first observed request for a terminal client (solid line) and the fundamental limits imposed by the observation periods (dotted lines) in the north network.

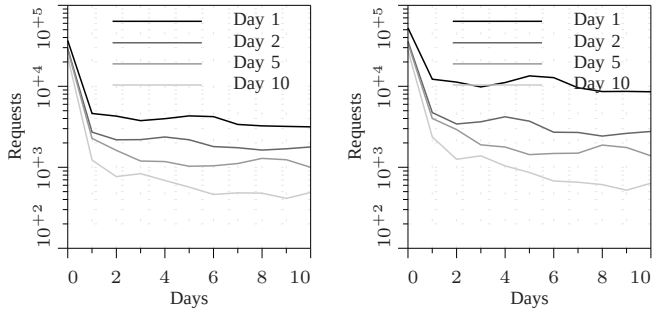

Fig. 9. The number of unique (left) and total (right) requests for clips *vs.* days since they were first requested separated by the day on which the first request was observed in the north network.

TABLE V
PERCENTILES OF TIMES UNTIL REPLAYS.

Fraction	90%	95%	99%	99.5%	99.9%
Time	2h	12h	3d	6d	15d

TABLE VI
TIME LAG PERCENTILES BETWEEN INDIVIDUAL AND COMMON DEMAND FOR THE NORTH NETWORK.

Time lag	All users	Heavy users
0:00	99.8%	88.5%
0:15	99.9%	98.8%
1:00	99.9%	99.8%

no reason to assume that the first day of our measurements is different from the other days, we conclude that there exists a "set" of clips with "long term popularity". Further, we observe many of the most popular clips during the first day, which stay popular for at least the ten days shown in the plots and we observe fewer such clips and/or relatively more less popular clips in this set during the subsequent days. We remark that the members of this set obviously changes over time, but we note that this is hard to see in our measurements.

These conclusions are supported by the results in Fig. 10 where the left diagram depicts the probability that a terminal client will replay a clip as a function of the number of times it has been replayed by that terminal client.

It is noted that the probability that a clip will be replayed one more time tends to grow with the number of times it has been played, until it reaches a saturation value of about 90%. To see this, first note that the probability that a clip viewed for the first time will be replayed is about 15% while the probability that a clip viewed for the second time will be replayed is about 35% *etc*., and then note that the probability that a clip viewed more than ten times will be replayed is about 85%. It is interesting to note that the "converged" replay process thus appears to be memoryless, *i.e*. the probability of further replays becomes independent of the accumulated number of replays.

Next we turn to the pronounced, cyclic replay patterns with peaks every 24 hours. The middle diagram in Fig. 10 depicts for various delays the number of times this delay has been observed between the first time a client requests a clip and the time of the first, second and third time the same client repeats that request. (We remark that, as before, the seemingly fewer observations of replays after longer times can be the effect of that they indeed are fewer, of limited observability due to finite time windows, or both.)

The difference between the left and the middle diagrams in Fig. 10 is thus that in the former we see the general replays whereas in the latter we see specific replays, *viz*. the first, second and third ones. It is immediately seen that the time until repeats exhibit the same pattern as before, with peaks every 24 hours, hence not only replays in general but also each specific replay appears to occur in a relatively regular (and thus potentially predictable) way.

This regularity may be a direct effect of regular requests or an indirect effect of daily traffic variations and regular peak periods. To examine this, we consider the right diagram in Fig. 10 which depicts the total number of requests as a function of time.

Comparing the shapes of the curves in the middle and right diagrams of Fig. 10, it is seen that the former have much sharper peaks than the latter which suggests that at least some of the regularity must be explained by other phenomena than regular peak periods. This observation is further supported by the close match between the aggregated traffic variations and those of individual terminal clients, cf. Table VI which gives the fractions of terminal clients for which certain lags are observed between the aggregated traffic variations and those of the individual terminal clients.

The lags in Table VI are computed by, first, juxtaposing the aggregated 24 hour traffic patterns with those of individual terminal clients and, second, sliding the later in time such that the sum of the absolute differences between the number of observations per five minute period in the two traffic patterns is minimised. It is seen that in most cases the two traffic patterns agree, and that only about 0.1% of all terminal clients (1.0% of heavy terminal clients) exhibit traffic patterns that deviate 15 minutes or more from the aggregate traffic pattern.

VI. TERMINAL CHARACTERISTICS

We now turn to the different kinds of equipment which, as mentioned above, are grouped into four types: PCs, mobiles,

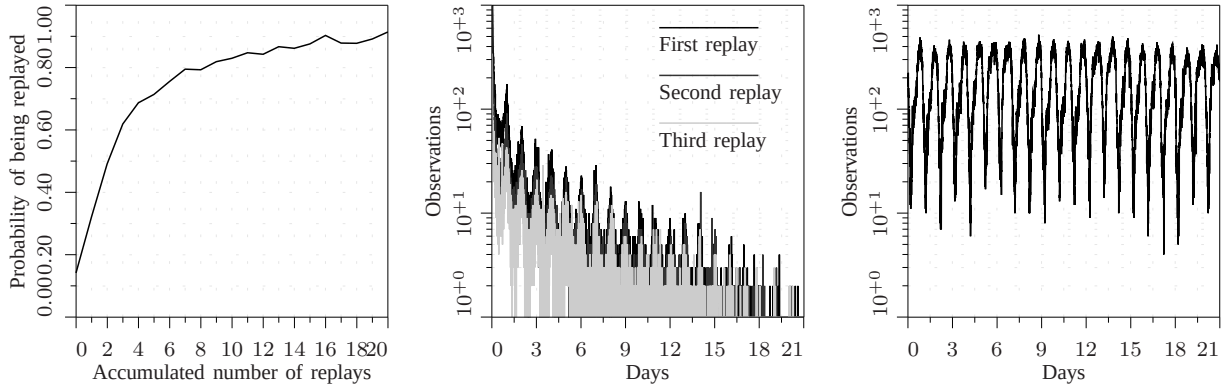

Fig. 10. Left: The probability that a clip will be replayed by a terminal client as a function the number of times it has been played by that terminal client. Middle: The number of replays of clips *vs*. time. Right: The number of clips requested *vs*. time.

TABLE VII
PREVALENCE OF DIFFERENT CLASSES OF EQUIPMENT.

Network	PC	Mobile	TV/Playstation	Other
North	18818	4728	229	284
	78.2%	19.6%	0.95%	1.18%
South	6973	1141	138	1510
	71.4%	11.7%	1.41%	15.5%

TABLE VIII
USAGE OF MOBILE EQUIPMENT IN YOUTUBE DEMAND.

Network	Mobile	No mobile	Mobile only
North	46.3%	53.7%	3.4%
South	29.4%	70.6%	2.7%

TABLE IX
DAILY YOUTUBE DEMAND FOR DIFFERENT TYPES OF EQUIPMENT.

Hardware	North		South	
class	Total	Unique	Total	Unique
PC	2.460	1.962	3.670	2.874
Mobile	1.744	1.252	2.742	1.781
TV/Playstation	1.256	0.936	0.380	0.320
Other	1.428	0.878	0.344	0.201

TABLE X
CACHE HIT RATES PER TYPE OF EQUIPMENT.

Cache	North		South	
arrangement	PC	Mobile	PC	Mobile
Network cache only	0.517	0.448	0.434	0.415
Cache at access point	0.232	0.289	0.253	0.359
Network with above	0.372	0.224	0.242	0.086
Cache at terminal client	0.202	0.282	0.217	0.351
Network with above	0.395	0.231	0.278	0.099

TVs/Playstations and others. (Note that, since the measurements refer to fixed networks, the mobiles we see are are those connected via WiFi.) The prevalence of the different classes is shown in Table VII.

In terms of observed clients, it is seen that the results are about the same for PCs and TVs/Playstations. The most significant type is PCs, with about 75% of the terminal clients, and the least significant type is TV/Playstation, with about 1% of the terminal clients. It is also seen that the results are quite different for mobiles and other. As for mobile terminals, these amount to about 20% in the north network and about 10% in the south network. Finally the remainder amounts to about 1% in the north network and about 15% in the south network. Judging by the numbers in the north network, we believe that at least some of the unknowns in the south network should be classified as mobiles.

A further analysis shows, Table VIII, that mobile terminals seldom are the only means to access YouTube (we note this for about 3% of all access points) while mobiles are relatively common as a complement (we note this for 29–46% of the population).

The amount of content consumed differs between the types as shown in Table IX. It is seen that PCs and mobiles consume more content than other types of devices and we remark that TVs/Playstations are surprisingly unpopular means of accessing to YouTube; not only are they few but the ones that do exist are not used very much. Another interesting observation is that the relationship between unique clips per terminal client and the total clips per terminal client differ; there are relatively more unique clips on PCs than on mobiles.

The last observation suggests that caches in terminal clients would be even more useful in mobiles than in PCs. To examine this, we separate the traffic into four classes, one class per hardware type, and examine the cache hit rates within each such class. The results are given in Table X.

It is seen that the total degree of "content recycling" is slightly higher for PCs than for mobiles and that this applies to both networks. We explain this by the simple facts that (a) there are many more PCs than mobiles and (b) PCs consume more content than mobiles. A more interesting observation is, however, that the potential for recycling in the terminal devices indeed is higher for mobiles than for PCs, and that this is more pronounced in the south network than in the north network. Our analysis shows that users to a higher degree *explore unknown* content on PCs and *repeat known* content on mobiles. Also, further analyses show that this does not imply strong "content migration" since about 96–97% of the items played at all on mobiles were also were played first on

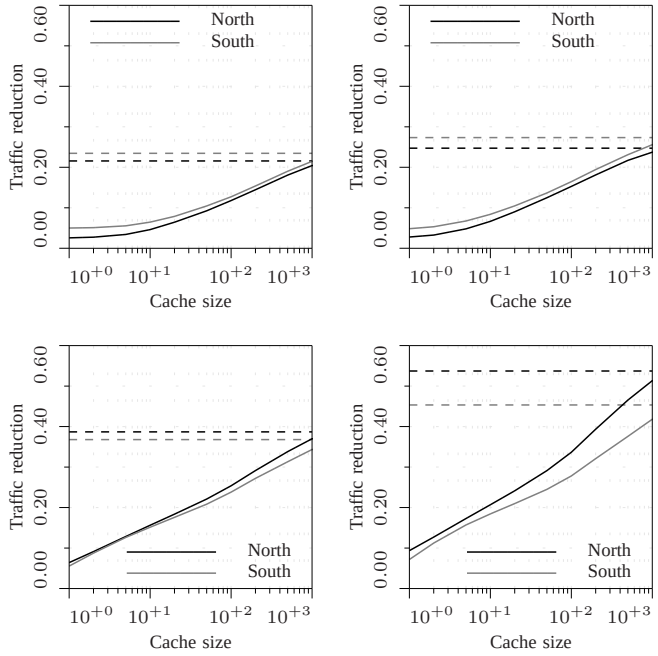

Fig. 11. Traffic reduction for different sizes of limited caches with LRU eviction at terminal clients (upper left), access points (upper right), region head ends (lower left) and network head ends (lower right). Cache sizes are given in percent of unique video clips requested per day and unit (*i.e.* per terminal client, access point, region head end or network head end respectively). Dotted lines indicate reductions obtained from ideal caches.

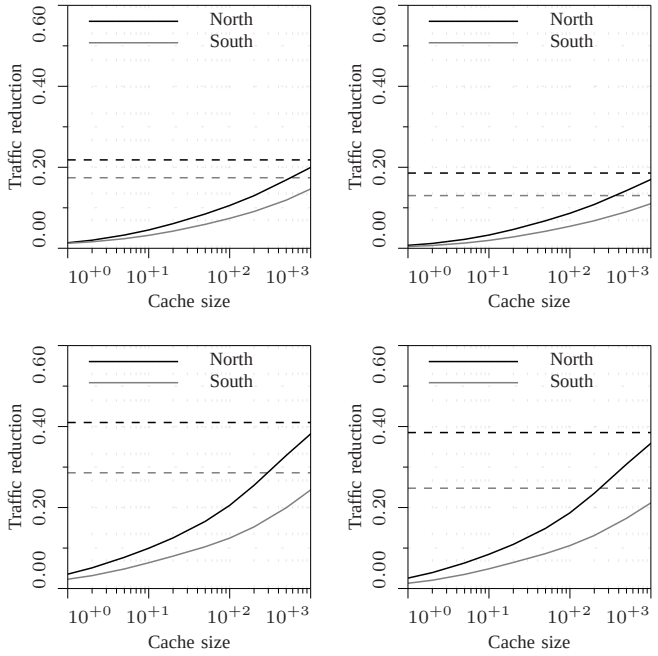

Fig. 12. Traffic reduction for different sizes of limited caches with LRU eviction at region head ends (top) and network head ends (bottom) when combined with (ideal) caches at terminal clients (left) and access points (right). Cache sizes are given in percent of unique video clips requested per day and unit (*i.e.* per region head end or network head end respectively). Dotted lines indicate reductions obtained from ideal caches.

such devices.

VII. Cache Characteristics

This paper deals with different aspects of YouTube traffic. The intention is to reveal any patterns that may be exploited to satisfy user demand in smarter ways. One of the most obvious aspects we have found is the possibility to exploit the fact that many requests are "double repeats" (not only for the same video clip but also from the same terminal agent), and we have found that serving these requests from local caches in the terminal clients can cut YouTube traffic by about 20% over a few weeks (by about 30% for mobile devices). These and other numbers are based on "ideal caches" which store all video clips and on "limited intervals" the beginning and end of which truncate the observations and lead to biased results. In this section we will deal with these aspects.

To examine the impact of cache limitations, we replace the ideal caches by (simplified) realistic ones which (i) can store a limited number of clips, (ii) store all new clips and (iii) when required eject the least recently used clip. To obtain comparable results, we express cache sizes in percent of the unique video clips requested per day and unit at which the cache resides (*i.e*. per terminal client, access point, region or network respectively). These can be rescaled to actual clips through Table I.

Fig. 11 shows the total traffic reduction that would have been obtained by deploying these simple caches during the measurement periods at terminal clients (upper left), access points (upper right), region head ends (lower left) and network head ends (lower right) for the two networks studied. It is seen that cache sizes corresponding to the number of unique videos consumed during ten days yield almost the same reduction as ideal caches.

Fig. 12 shows the marginal traffic reduction that would have been obtained by deploying "cascaded caching" if (ideal) caches at terminal clients (left) and access points (right) were supplemented with (limited) caches at region head ends (top) and network head ends (bottom). Again it is seen that cache sizes corresponding to the number of unique videos consumed during ten days yield almost the same reduction as ideal caches.

Next, to examine the impact of the finite interval, we examine the traffic reduction as a function of time (for an ideal cache).

Fig. 13 shows the total traffic reduction that would have been obtained by deploying ideal caches during the measurement periods at terminal clients (upper left), access points (upper right), region head ends (lower left) and network head ends (lower right) for the two networks studied. As expected, traffic reduction increases over time while the rate at which this happens tends to decrease over time. It is also noted that the finite intervals are noticeable in that there are no signs of final convergence in any of the diagrams.

VIII. Conclusions

In this paper, we have performed detailed statistical analyses of YouTube user demand patterns and locality properties based on data from two municipal networks in Sweden.

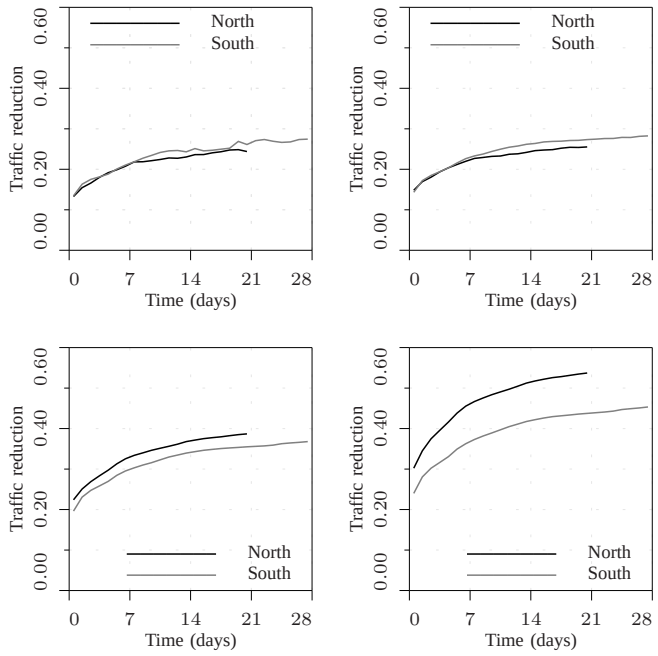

Fig. 13. Traffic reduction *vs*. time by deploying ideal caches at terminal clients (upper left), access points (upper right), region head ends (lower left) and network head ends (lower right).

The paper has demonstrated that a large share (about 80%) of requests for video clips are made for a small number of distinct video clips (about 20%). This phenomenon suggests there is a potential for gains by using caching.

In the continued analysis, we found that caching can be efficient even if the demand is relatively low not only because of (i) similar requests from users living in the same part of the country and (ii) similar requests from users living in the same district but also because of (iii) similar requests from terminal clients sharing the same access point and (iv) similar requests from individual users.

We also found that (i) the probability of further replays grows with the number of previous replays and that (ii) replays exhibit a remarkably regular pattern in time and (iii) can occur after long times.

It was further noted that user caches can provide significant gains and that, such caches can be expected to provide the most hits *when* they are most useful (during peak times) and *where* they are most useful (where the heavy users are).

Finally, PC users and mobile device users showed different content demand patterns and it was seen that user caches may be particularly attractive in mobile devices, since users on mobile devices have a high probability of replays.

Through the IPNQSIS project we will get access to similar measurements made at the same time not only in the two networks above but also in one network in Finland. It would be interesting to expand the locality concept above, which now ranges from terminal client to network head end, to also include different countries. Another interesting way forward is to evaluate the locality measures in [16] and the social aspects in [17] and to compare against their results.

ACKNOWLEDGMENTS

The work in this paper has been partly funded by Vinnova in the CelticPlus project IPNQSIS and the project EFRAIM. Maria Kihl and Andreas Aurelius are members of the Lund Center for Control of Complex Engineering Systems (LCCC). Maria Kihl is a member of the Excellence Center Linköping - Lund in Information Technology (eLLIIT).

REFERENCES

[1] "Ericsson traffic brief january 2012," Ericsson AB, Tech. Rep. EAB 12:006276 Uen, 2012.

[2] "Cisco visual networking index: Global mobile data traffic forecast update, 2011–2016," Cisco Inc., Tech. Rep. FLGD 10459 05/12, 2012.

[3] A. Feldmann, R. Caceres, F. Douglis, G. Glass, and M. Rabinovich, "Performance of web proxy caching in heterogeneous bandwidth environments," in *IEEE INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings*, vol. 1. IEEE, March 1999.

[4] N. Leibowitz, A. Bergman, R. Ben-shaul, and A. Shavit, "Are file swapping networks cacheable? characterizing p2p traffic," in *7th Int. WWW Caching Workshop*, 2002.

[5] F. Lehrieder, G. Dán, T. Hossfeld, S. Oechsner, and V. Singeorzan, "The impact of caching on bittorrent-like peer-to-peer systems," in *Peer-to-Peer Computing (P2P), 2010 IEEE Tenth International Conference on*, Aug. 2010, pp. 1–10.

[6] B. Ager, F. Schneider, J. Kim, and A. Feldmann, "Revisiting cacheability in times of user generated content," in *INFOCOM IEEE Conference on Computer Communications Workshops , 2010*, March 2010, pp. 1–6.

[7] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Watch global, cache local: Youtube network traffic at a campus network — measurements and implications," Department of Computer Science, University of Massachusetts Amherst, Tech. Rep. Technical Report 177, 2008. [Online]. Available: http://scholarworks.umass.edu/cs_faculty_pubs/177

[8] Å. Arvidsson, A. Mihály, and L. Westberg, "Optimised local caching in cellular mobile networks," *Computer Networks*, vol. 55, no. 18, December 2011.

[9] N. Chand, R. C. Joshi, and M. Misra, "Cooperative caching strategy in mobile ad hoc networks based on clusters," *Wireless Personal Communications*, vol. 43, December 2006.

[10] N. Chand, R. Joshi, and M. Misra, "Broadcast based cache invalidation and prefetching in mobile environment," in *High Performance Computing - HiPC 2004*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, January 2005, no. 3296, pp. 410–419.

[11] G. Cao, "A scalable low-latency cache invalidation strategy for mobile environments," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 5, pp. 1251–1265, October 2003.

[12] A. Finamore, M. Mellia, M. Munaz, 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 conference*, ser. IMC '11. ACM, 2011.

[13] 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. ACM, 2007.

[14] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07. New York, NY, USA: ACM, 2007.

[15] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of YouTube videos," in *16th International Workshop on Quality of Service, 2008. IWQoS 2008*, June 2008.

[16] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube around the world: geographic popularity of videos," in *Proceedings of the 21st international conference on World Wide Web*, ser. WWW '12, 2012, pp. 241–250.

[17] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida, "On word-of-mouth based discovery of the web," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, ser. IMC '11, 2011, pp. 381–396.