LUND UNIVERSITY

Analysis of Facebook content demand patterns

Kihl, Maria; Larsson, Robin; Unnervik, Niclas; Haberkamm, Jolina; Arvidsson, Åke; Aurelius, Andreas

Published in:
[Host publication title missing]

DOI:
10.1109/SaCoNeT.2014.6867760

2014

Link to publication

Citation for published version (APA):
Kihl, M., Larsson, R., Unnervik, N., Haberkamm, J., Arvidsson, Å., & Aurelius, A. (2014). Analysis of Facebook content demand patterns. In *[Host publication title missing]* IEEE - Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/SaCoNeT.2014.6867760

Total number of authors:
6

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Analysis of Facebook content demand patterns

Maria Kihl, Robin Larsson,
Niclas Unnervik, Jolina Haberkamm
Dept. of Electr. and Inform. Technology
Lund University, Sweden
{maria.kihl, robin.larsson}@eit.lth.se

Åke Arvidsson
Ericsson Research; Packet Technologies,
Ericsson AB, Sweden
ake.arvidsson@ericsson.com

Andreas Aurelius
Acreo Swedish ICT AB
andreas.aurelius@acreo.se

***Abstract*—Data volumes in communication networks increase rapidly. Further, usage of social network applications is very wide spread among users, and among these applications, Facebook is the most popular. In this paper, we analyse user demands patterns and content popularity of Facebook generated traffic. The data comes from residential users in two metropolitan access networks in Sweden, and we analyse more than 17 million images downloaded by almost 16,000 Facebook users. We show that the distributions of image popularity and user activity may be described by Zipf distributions which is favourable for many types of caching. We also show that Facebook activity is more evenly spread over the day, compared to more defined peak hours of general Internet usage. Looking at content life time, we show that profile pictures have a relatively constant popularity while for other images there is an initial, short peak of demand, followed by a longer period of significantly lower and quite stable demand. These findings are useful for designing network and QoE optimisation solutions, such as predictive pre-fetching, proxy caching and delay tolerant networking.**

I. Introduction

The volume of data traffic in cellular networks has been increasing exponentially for the past few years and it is predicted that this increase will continue over the coming few years as well, cf. the Ericsson Traffic and Market Report [1] which for the period 2011–2017 predicts a mobile data compound annual growth rate (CAGR) of 60% which results in a total of more than 8 exabytes (1 exabyte $= 10^{18}$ bytes) per month by 2017 and the Cisco Virtual Networking Index [2] which for the period 2011–2016 predicts a global data CAGR of 29% which results in a total of more than 110 exabytes per month.

The rapid growth in traffic combined with relatively slow growth in revenues lead to a continuous need for operators to reduce their costs. At the same time, however, operators must stay competitive and focus on performance to avoid churn as quality rankings are becoming public [3], [4] and device unlocking is or will be protected by law [5], [6]). Optimising network utilisation without negatively impacting performance requires understanding of user behaviour and preferences.

An important factor in this context is social networking which is one of the most important applications today. In this category Facebook is by far the largest one; in October 2012 Facebook had one billion active users of which 81% were users outside the U.S. and Canada, and in Sweden more than 50% of the population use Facebook. Moreover, Facebook is used both in mobile and fixed networks; in September 2012 600 million monthly active users were seen using Facebook mobile products [7]. Noting that Facebook users not only post and read status updates, but also upload and download many images and videos, we conclude that Facebook traffic will include a large amount of content objects the handling of which may be improved by various network and QoE optimisation schemes, such as caching and prediction-based pre-fetching.

Surprisingly, very few papers have analysed Facebook user behaviour and examine demand patterns from actual traffic. Some general traffic statistics for Facebook are shown in [8] and a trend detection system for Facebook posts is proposed in [9]. Moreover [10] considers Facebook posts and analyses interaction activities and friendships while user interaction also was investigated in [11]. Finally, usage patterns of different Facebook applications were analysed in [12]. However, to the best of our knowledge, there are no papers that have analysed Facebook images with regards to download patterns and popularity.

Therefore, in this paper we present a detailed analysis of user demand patterns and content popularity for Facebook, *i.e.*, without evaluating particular solutions such as, *e.g.*, prediction-based pre-fetching, proxy caching or delay tolerant networking.

The study is focused on images and based on data from traffic measurements in two metropolitan access networks in Sweden. We find that the distributions of both content popularity and user activities may be described as Zipfian and that demand for images can be viewed as a first, short phase of high but declining demand followed by a second, longer phase of lower but seemingly stable demand. We also find that, compared to other applications, Facebook usage is relatively evenly spread throughout the day and that the weekly peaks can occur in the middle of the week.

The paper is organized as follows. In Section II we describe the networks and data collection procedures. Further, in Section III we present results on content demand patterns, image popularity and potential caching gains. Finally, in Section IV we present some conclusions.

II. Data collection and processing

In the following section, we will describe the networks, measurements and data collection procedures.

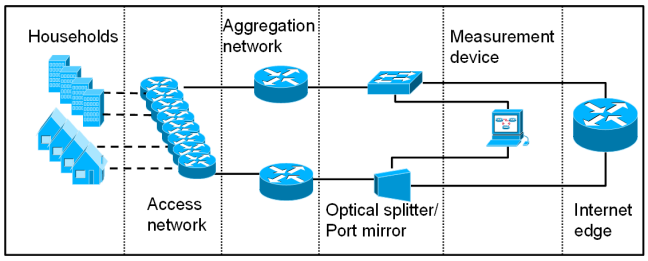

Fig. 1. Network architecture.

A. Networks and measurements

The study is based on measurements made by two Swedish network access network operators that are partners of Acreo Swedish ICT AB and as a part of the IPNQSIS and NOTTS Celtic projects. For privacy reasons, the networks are referred to as the north and south networks.

The data sets include roughly 5000 households in the north network and 2000 households in the south network respectively. The customers in each network are local residents who can freely choose between different ISPs for access to the Internet. As shown in Figure 1, traffic measurement probes were placed at the network head end edge routers which bridge the metro networks to the Internet. This means that all Facebook traffic from all subscribers can be captured by the probes.

The measurements were performed with a commercial PacketLogic (PL) probe from Procera Networks, which performs deep packet and deep flow inspection. The measurements were made in three steps. First, all traffic to *facebook.com* and *fbcdn.net* was store in PCAP files, second, MAC and IP addresses in the PCAP files were scrambled such that no data can be traced back to specific users and, third the analysis scripts were run on-site after which the anonymous meta data was transferred for detailed analysis.

B. Data collection and processing

The measurements lasted for 13 days in the north network, from Wednesday, October 17$^{\text{th}}$ to Monday, October 29$^{\text{th}}$, 2012 and for 18 days in the south network, from Friday, September 21$^{\text{st}}$ to Monday, October 8$^{\text{th}}$, 2012.

Facebook images may be downloaded "automatically" or "manually". The first group is formed by pictures that are downloaded without explicit user requests and includes profile pictures, pictures related to advertisements and pictures on initial log in pages (such as small album pictures, icons, and pictures connected with posts *etc.*). The second group is formed by pictures that are downloaded as a result of explicit user requests and includes, *e.g.*, larger versions of pictures which users have clicked on. In this paper, we consider *profile pictures*, *thumbnail* images, which are automatically downloaded, user-generated images shown in the news feed, and *large images*, which are manually downloaded as users click on or point at thumbnail images. Other images (such as advertisements) are thus not part of the investigation in this paper. Images related to advertisements seem less relevant from the perspective of understanding user behaviour.

All images are requested by GET requests which means that picture URLs are part of the HTTP headers. Such URLs can be found by filtering on the supported picture formats; jpg, gif, png and tif. Different pictures can be identified by the names of the pictures which also are given in the HTTP headers.

To identify users we keep track of Facebook identification numbers (FBIDs), which are unique static numbers assigned to each user. FBIDs are distributed through the unsigned 32-bit integer space except newly assigned FBIDs which are prefixed by the number 10000 [13]. FBIDs are, however, not included in the requests but can be found by examining Facebook cookies and it was discovered that client browsers repeatedly send batches of cookies to Facebook during the active sessions. Below is a truncated example of what cookie-batches can look like when they are extracted from the data packet:

```
"Cookie": "datr=[...]; fr=[...]; lu=[...];
c_user= 0123456789;"
```

As can be seen, the `c_user` Cookie ID contains the FBID. To map GET requests to FBIDs, timestamps and MAC addresses were used; a request at time t sent from a MAC address a was thus matched to the FBID at time t' on MAC address a for which $|t - t'|$ was minimised. A few GET requests, on the order of 1%, could not be mapped to unique FBIDs since the time granularity was one second and, on a few occasions, different FBIDs were sent from the same MAC address within one second hence unique mapping was impossible. We believe that this is a minor problem, however, since a unique mapping could be performed for the remaining 99% of all GET requests.

In the following sections, we will present various results showing user content demand patterns, image popularities and life-times. The results of identifying requests, images as well as FBIDs are summarised in Table I.

We remark that traffic from Facebook users who have enabled secure browsing is not part of this study, since their corresponding traffic is encrypted. However, our data indicates that during this time period, only a small subset of the users had enabled secure browsing, hence our assumption is that encrypted sessions only would have a minor impact on the results. Later, Facebook changed its system, and now all users have secure browsing enabled by default.

III. Results

In the following section we will present the results of the data analysis. The analysis has been focused on user activity, image characteristics and traffic patterns.

A. User activity

User activity levels may of course be measured in several ways. A typical metric is the durations of individual user

TABLE I
REQUESTS, IMAGES AND FACEBOOK IDENTIFIERS IN THE TWO DATA SETS.

	North				South			
	Profile picture	Thumbnail	Large image	Total	Profile picture	Thumbnail	Large image	Total
Requests	23,475,766	5,778,277	2,045,900	31,299,943	12,763,755	3,109,786	1,240,241	17,113,782
Images	5,804,311	2,898,408	1,213,792	9,916,511	4,905,812	1,980,046	944,477	7,830,335
FBID	11,562	10,289	9,187	12,175	5,086	4,405	3,926	5,473

sessions but, unlike the Internet applications analysed in [14], sessions may be more difficult to identify as many Facebook users in many cases are (semi-)permanently logged in. Another possibility particular to social networks is to measure user activity by the number of messages posted, the number of "like clicks", or the number of downloaded images.

In this paper we take the latter approach; users who download many images are considered to be more active than users who download few images. Figure 2 shows a ranking of users in the two networks in terms of total (upper, solid curve) and unique (lower, dashed curve) downloaded images. It is seen that the average activity, which in the north network amounts to about a total of 198 requests for images (about 63 unique images) per user and day, is unevenly distributed between the users such that on the order of 10% of the users have a relatively high activity with about 10,000 requests per day while about 50% of the users have a much lower activity with less than 1,000 requests per day. The south network shows a similar user behaviour. We can compare this result to the patterns for YouTube users who in the same networks tend to watch on the average about two clips per day with a qualitatively similar spread between users [15].

B. Image popularity and life time

Having seen the significant demand for Facebook images we now assess the possible gains from different optimisations. To this end we note that images with (many requests from) few users during a short time may need different treatment (and give different gains) compared to those with (many requests from) many users and/or during a long time hence we examine how demand is distributed between users and over time.

Figure 3 shows the popularity characteristics for the downloaded images in the north network (top) and the south network (bottom) for thumbnails (left), large images (middle) and profile pictures (right). The rankings refer to total number of request for downloads (top solid curves) and number of unique users that requested downloads (bottom dashed curves) respectively. We can conclude that the curves in all cases have significant heavy tails; most downloads relate to a limited set of pictures whereas the majority of images only are downloaded a few times and by a few users. Such concentration of demand in general is favourable when it comes to caching (and many other forms of optimisation), but we note that the present slope of about -0.5 is relatively small in this context.

Another important aspect is image life times. If images only are popular during very short times after being uploaded to Facebook, the content in a network cache will need to be renewed often and makes prediction more difficult. We define the life time of an image as the time between the first and last visible downloads.

Figure 4 depicts the cumulative distribution function, $F(t)$, for the life time distribution of the different types of images in both networks that were seen during the first day of measurements. The cumulative distribution function is derived as

$$F(t) = \frac{1}{N} \sum_{\tau=1}^{t} n(\tau)$$

where N is the total number of downloads, and $n(\tau)$ is the number of downloads at time τ seconds from the first download; note that the first download at time $\tau = 0$ is not counted. This calculation gives a simple estimate of an image's life time. Of course, an image may have a longer life time, since the measurement period is limited. Similar results were found for the south network.

It is seen that the popularity for thumbnails and large images can be described in two phases; a *first* one of high but decreasing demand during about one day and a *second* one of low but relatively stable demand for the rest of the measurement period. However, for profile pictures, the popularity is rather evenly distributed over the days, which is compliant with the expected user demand patterns for these images. Profile pictures have a more extended popularity than thumbnails and large images.

Figure 5 depicts the same phenomenon but in terms of requests per time since the first observation. The graphs for the north network are shown, and the data from the south network shows similar behaviour. As before we see there is a first phase of strong but declining initial demand and a second phase of seemingly relatively stable demand, where the second phase is more pronounced for profile pictures. The initial "sub peaks" as well as the regularly occurring daily "spikes" may be explained by weekly and daily traffic variations which we will examine in more detail below. Again we can compare with YouTube for which the same general drop can be seen but with much more pronounced daily spikes [15].

C. Traffic variations

Another important aspect of traffic optimisation is time; in a strict sense only peak hour traffic matters since well engineered networks would be under-utilised at other times.

To see how Facebook activities vary over time we consider the upper graphs in Figure 6 that shows the rate of requests for images over the measurement period in the north and south networks respectively. As can be seen, there are long peaks

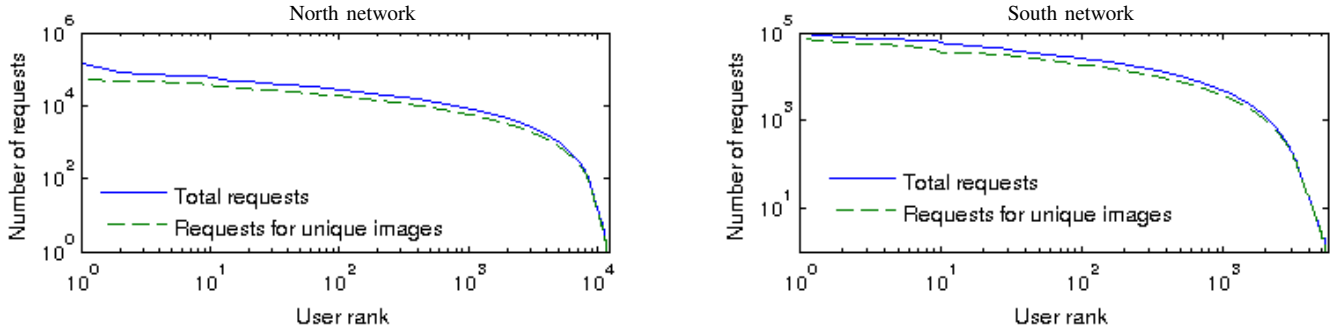

Fig. 2. Ranking of users in terms of number of downloaded images in the north network (left) and the south network (right). The user with the highest amount of downloads is ranked as number 1. The top curves (solid) show user ranking in terms of total requests. The bottom curves (dashed) show user ranking in terms of unique requests. The ranking in the two curves are not matched, that is, user number 1 with respect to totals may not be user number 1 with respect to unique requests *etc*.

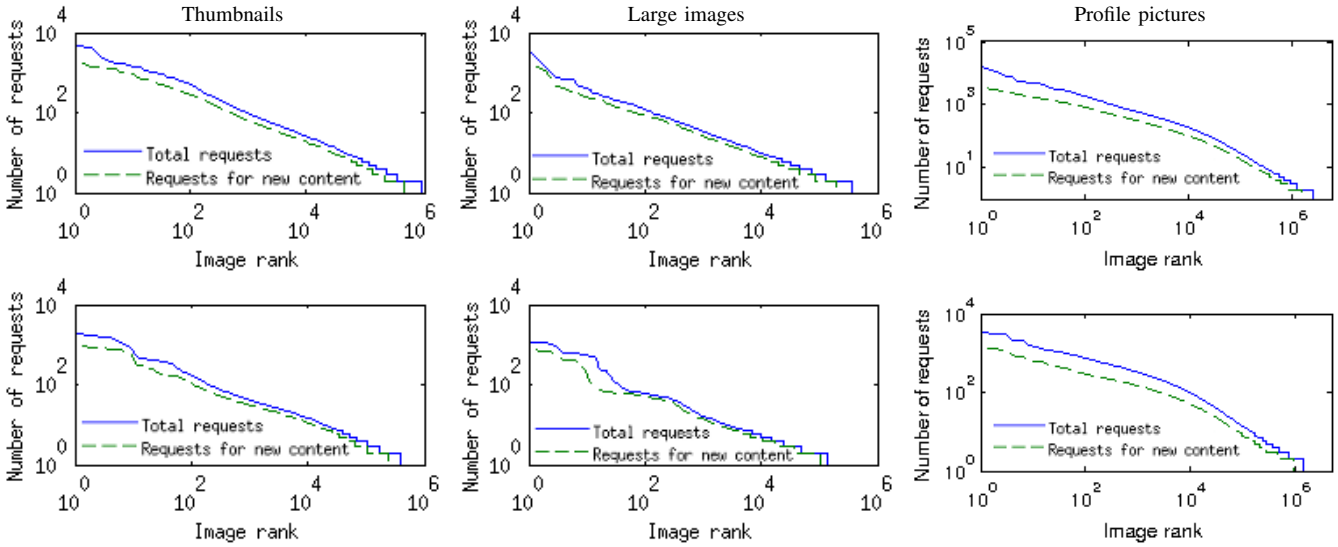

Fig. 3. Ranking of images in the north network (upper) and south network (lower) with respect to total and unique requests respectively; thumbnails (left), large images (middle) and profile pictures (right). The upper curves (solid) show the ranking with respect to total number of requests and the lower curves (dashed) show the ranking with respect to unique FBID requests.

during each day. There is most of the time no particular difference between weekdays and weekends. However, in the north network there are two days in the middle of the measurement period (Oct. 23 and 24) during which a much higher rate of requests was generated than usual. When examining the lower curve with the rate of requests for new images, it can be seen that most of these requests are for "old" images. Obviously, something happened during these two days that triggered this spike of requests.

The more evenly spread Facebook activities are also illustrated in lower part of Figure 6, where the average diurnal traffic pattern are shown for the two networks. Again it is seen that there are no clear peak hours in the evenings, but the activity is increasing from the morning and it is not decreasing until late at night. It is interesting to note that this is different from other Internet applications, such as web browsing and streaming [14], for which distinct peaks were noted in the evenings between 8 *p.m.* and 10 *p.m.*

D. Potential caching gains

In the previous sections, we have investigated Facebook user activity, image popularity and traffic variations, all of which will have an impact on various network optimisation schemes, for example network caches. In order to illustrate the potential caching gains for Facebook images, Figure 7 shows the potential gain from caches (*i.e.*, the gain that would be achieved in caches without deletions) at the network edges as functions of time. As can be seen, the two networks have rather different cache hit dynamics, partly due to their different sizes (12,175 FBIDs in the north network and 5,473 FBIDs in the south network), and partly due to the "odd" request spikes in the north network. It can be seen that the cache hit ratios reach almost 70% in the north network and almost 55% in the south network although none of the values seem to have converged.

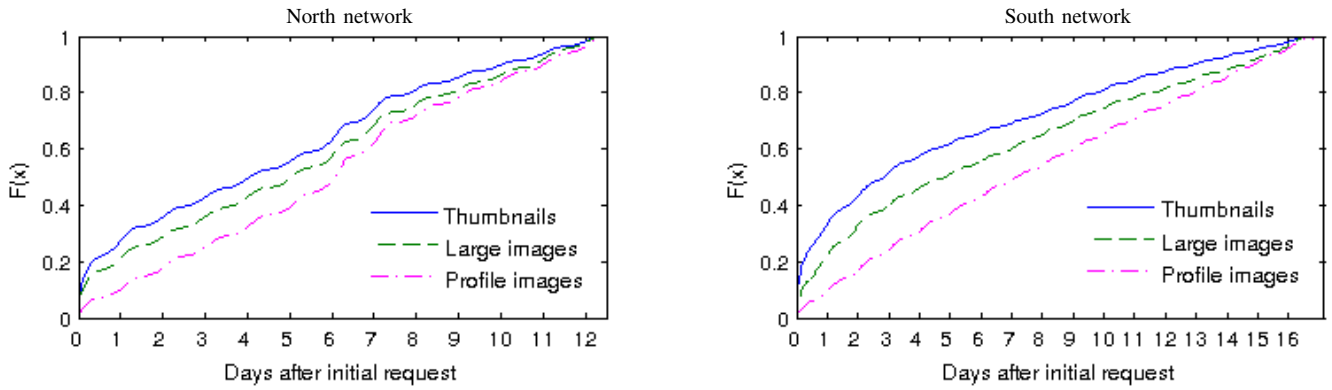

Fig. 4. The cumulative distribution function for the life time of thumbnail images, large images, and profile pictures seen during the first day in the north network (left) and the south network (right).

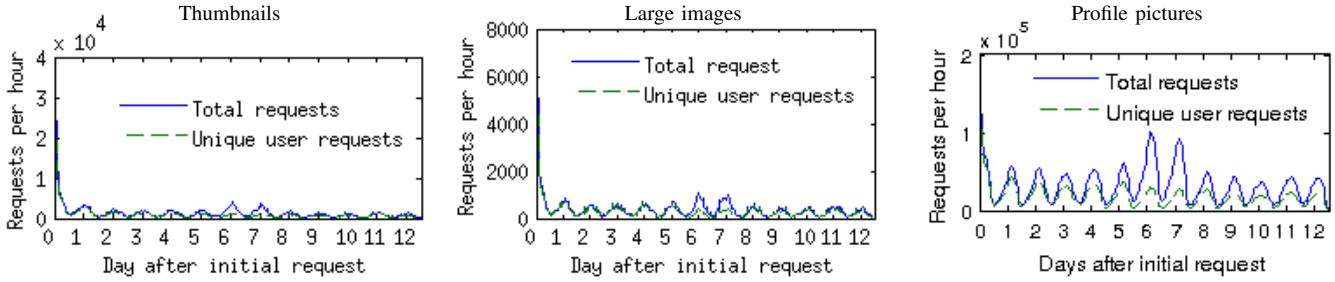

Fig. 5. Requests over time for images first seen on day one in north network; thumbnails (left), large images (middle) and profile pictures (right). The upper solid curves show the total number of requests and the lower dashed curves show the number of unique FBID requests.

IV. Conclusions

In this paper, we have analysed Facebook data from two weeks measurements in two residential metropolitan access networks, with in total more than 7 million images downloaded by more than 16,000 Facebook users. The purpose of the analysis has been to investigate user demand patterns and content popularity. We have shown that there is a large number of pictures downloaded each day, and that the demand is unevenly distributed with a small number of heavy users. The top 10% of the users request on the order of 10,000 pictures per day, whereas 50% of the users request less than 1,000. The majority of these requested images are profile pictures. The image popularity distributions exhibit a heavy tail, which is favourable in, *e.g.*, caching solutions.

Further, we showed that thumbnails seem to have relatively constant popularity while the lifetime of other images can be described by two phases; a first phase with a high but decreasing demand over roughly one day; a second phase with a low but relatively stable demand throughout the measurement period. Regarding the potential for caching, we have shown that ideal caches would achieve hit ratios of 70% and 55% in the north and south networks respectively, although none of the values seem to have converged during the measurement period.

Acknowledgements

The work in this paper has been partly funded by Vinnova in the CelticPlus projects IPNQSIS and NOTTS, and the national 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). Andreas Aurelius is partly financed by the Swedish National Strategic Research Area (SRA) within the program TNG (The Next Generation) and the project eWIN.

References

[1] "Traffic and Market Report — On the Pulse of the Networked Society," Ericsson AB, Tech. Rep. 198/287 01-FGB 101 220, 2012.

[2] "Cisco Visual Networking Index: Forecast and Methodology, 2011–2016," Cisco Inc., Tech. Rep. FLGD 10584 08/12, 2012.

[3] "November ISP Rankings for the USA," http://blog.netflix.com/2012/12/november-isp-rankings-for-usa.html, accessed: 12/03/2013.

[4] "J.D. Power Telecom Studies," http://www.jdpower.com/consumer-ratings/telecom/index.html, accessed: 12/03/2013.

[5] "Så säger lagen om att låsa upp din iPhone," ("This is what the law says about unlocking your iPhone"), http://www.idg.se/2.1085/1.118062l, accessed: 12/03/2013 (in Swedish).

[6] "Obama Administration: Mobile Phone 'Unlocking' Should Be Legal," http://business.time.com/2013/03/05/obama-administration-mobile-phone-unlocking-should-be-legal, accessed: 12/03/2013.

[7] "The New York Times: Mark Zuckerberg." [Online]. Available: http://topics.nytimes.com/topics/reference/timestopics/people/z/mark_e_zuckerberg/index.html

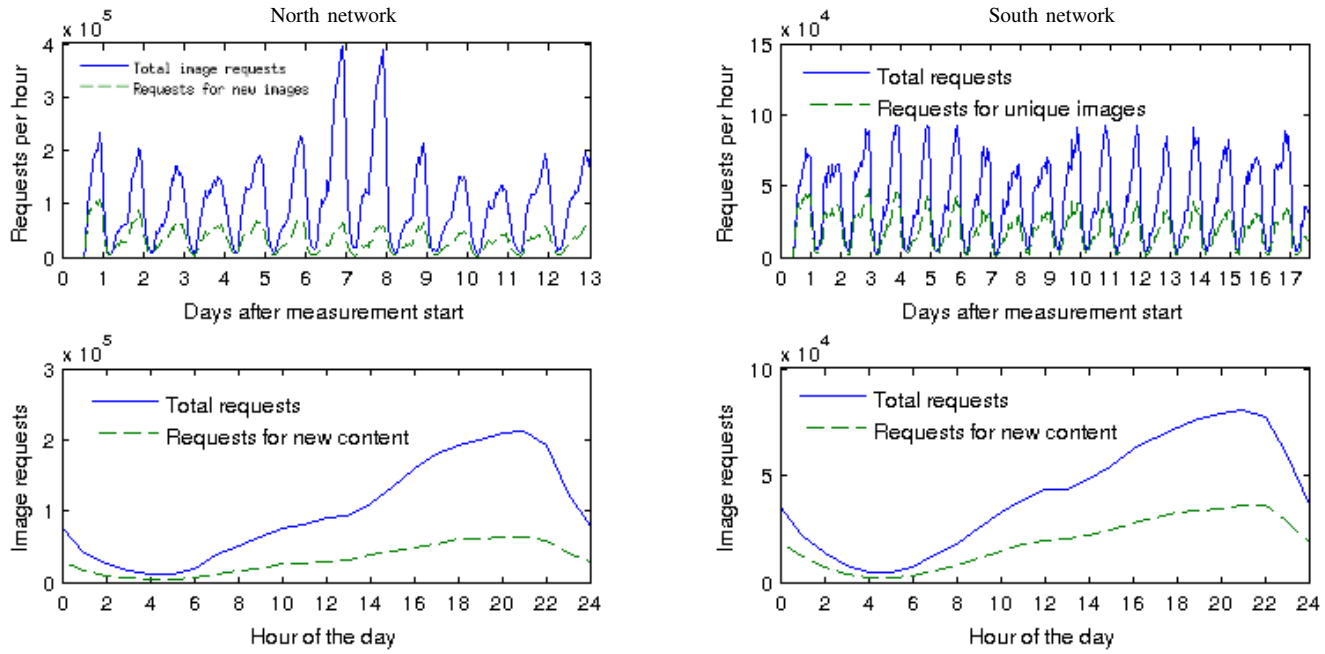

Fig. 6. Traffic profile for the north network (left) and south network (right). Top: Entire periods Wednesday, October 17 to Monday, October 29, 2012 and Friday, September 21 to Monday, October 8, 2012 respectively. Bottom: Averages over all days. The upper (solid) curves refer to the total number of requests and the lower (dashed) curve to the requests for new images (not previously downloaded)

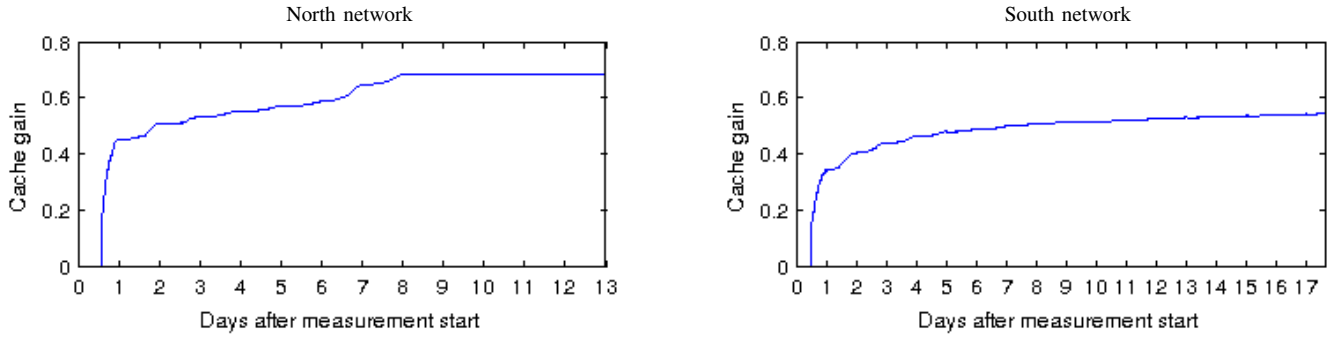

Fig. 7. Potential cache gains over time in the north network (left) and south network (right).

[8] Z. Moczar and S. Molnar, "Comparative Traffic Analysis Study of Popular Applications," in *Proc. of the 17th international conference on Energy-aware communications*, 2011.

[9] I. P. Cvijikj and F. Michahelles, "Monitoring Trends on Facebook," in *Proc. of the IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*, 2011.

[10] K. Nguyen and D. A. Tran, "An Analysis of Activities in Facebook," in *Proc. of the IEEE Consumer Communications and Networking Conference*, 2011.

[11] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the Evolution of User Interaction in Facebook," in *Proc. of the 2nd ACM SIGCOMM Workshop on Social Networks*, 2009.

[12] M. Gjoka, M. Sirivianos, A. Markopoulou, and X. Yang, "Poking Facebook: Characterization of OSN Applications," in *Proc. of the first ACM workshop on Online social networks*, 2008.

[13] S. Catanese, P. D. Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Crawling Facebook for Social Network Analysis Purposes," in *Proc. of the ACM International Conference on Web Intelligence, Mining and Semantics*, 2011.

[14] M. Kihl, A. Aurelius, C. Lagerstedt, and P. Ödling, "Traffic Analysis and Characterization of Internet User Behavior," in *International Congress on Ultra Modern Telecommunications and Control Systems*, 2010.

[15] Å. Arvidsson, M. Du, A. Aurelius, and M. Kihl, "Analysis of User Demand Patterns and Locality for YouTube traffic," in Proc. of the 25th International Teletraffic Congress, 2013.