(Not) Just Another Measurement Study of Online Social Networks

Rohit Koul and Raja Bala
Department of Computer Sciences,
University of Wisconsin Madison, Madison, WI, USA
{rkoul,bala}@wisc.edu

ABSTRACT

Online Social Networking sites like Facebook, MySpace and Orkut have become a popular way to share and disseminate content. A few studies have characterized how information spreads over these networks, but none of them have focused on the prevalence of fat content in them. In this paper, we collect and analyze data from the largest online social network in terms of the number of users and the amount of fat content generated, Facebook and compare that with a popular non-OSN, CNN. We analyze the data gathered and argue that the although Facebook is very effective in serving as much as 50 times more fat content than CNN, some of the pre-fetching policies employed by it can be improved. We also look at how the two sites behave differently for mobile users. We posit that a better understanding of some of the user interactions and the content delivery systems employed by these sites, coupled with the knowledge of the user geographical locations could improve the user experience manifold.

General Terms

social networks, measurement, performance, Facebook, CNN

Keywords

online social networks, prefetching, Facebook, caching, redundancy, information dissemination

1. INTRODUCTION

The Internet has spawned different types of information sharing systems including the web. Over the past five years, Online Social Networks (OSNs) have gained significant popularity and are now among the most popular sites on the web. Facebook (over 400 million users), MySpace (over 200 million users), Orkut (over 100 million users) and LinkedIn (over 50 million users) are a few examples that portray the scale of such networks.

Off late, Facebook has emerged as the most dominant platform of choice. Facebook is a social utility that helps people communicate more efficiently with their friends, family and coworkers, using technologies that facilitate the sharing of information through the social graph, the digital mapping of people's real-world social connections. Photos and videos, which we term as 'fat' content for the rest of this paper, comprise a huge chunk of the information shared. Adding to this are over 100,000 third-party applications, including games, written using the Facebook developer API.

[10] suggests that 50% of the active Facebook user base log on to Facebook on any given day. The amount of time people spend on Facebook per month clocks over 500 billion minutes!

A recent study [8] explores some of the properties of OSNs, the present methodologies available, and discusses various challenges associated with measuring them. Earlier work studied the graph properties of online communities, high level properties based on snapshots of individual OSNs, and issues related to anonymization and privacy. This was generally accomplished through crawling measurement techniques. To the best of our knowledge, there are no known studies which document the nature of macro-OSN traffic, such as Facebook, based on the content being exchanged, from both a mobile and non-mobile perspective.

Furthermore, related studies on the macro level properties of OSNs used Twitter [1] and Flickr [7] as the platforms of choice. Twitter, is a micro-blogging service that relies on short 140 character long text messages and has no fat content whatsoever. It also comes under the category of Micro Online Social Networks, on the basis of the brevity of content exchanged. Flickr, on the other hand, is an online photo and video sharing service, but does not cater to as broad an audience as Facebook.

This paper attempts to address the following questions.

How much of the Facebook traffic can be termed as 'fat'?

[11] indicates that Facebook users have uploaded over 15 billion photos and at the peak, there are 550,000 images served per second. Hence it is an interesting research problem to figure out how much bandwidth intensive fat traffic is.

How different are the network level traces for mobile users when compared to non-mobile ones?

[10] indicates that there are more than 100 million active users currently accessing Facebook through their mobile devices. Given the resource limitations of screen space

and energy in mobile devices, we need to compare mobile and normal Facebook interactions.

Facebook uses Akamai as its content delivery network. What insights can be gained into its content distribution mechanism?

More than 25 billion pieces of content (web links, news stories, blog posts, notes, photo albums, etc.) are shared each month via Facebook [10]. Any insights into the content distribution and dissemination are vital for improving the end-user experience.

Is pre-fetching used to improve the user experience? If so, how much of content is pre-fetched? Can intelligent caching techniques be used?

There is a high probability that the avid Facebook user clicks on content that is liked by a lot of people. User interaction in OSNs tends to be bursty in nature and recognizing the increase in activity to trigger the pre-fetching of content he/she is most likely to view is an interesting problem in itself.

How does the traffic characteristics of an OSN like Facebook compare with a popular non-OSN like CNN?

Rate of change of contents in an OSN is different from content owner controlled web sites. A popular news site like cnn.com, that is centrally administered and deals with timely information dissemination, has a higher rate of change than individually updated pages on an OSN. Thus, one would expect the traffic characteristics to be starkly different as well. Photos form a large chunk of the data elements present in a news web page. Frequent access to the same page might allow for different content distribution strategies when compared to the home page (also known as the 'Wall') of a Facebook user, which is heavily customized on his/her interactions and that of friends.

The remainder of this paper is structured as follows: In Section 2, we give a description of our approach and present our macro analysis with regards to the Facebook's content distribution and dissemination and follow it up with the details regarding our trace analysis. Section 3 deals with the results and observations of our approach, which we follow up in Section 4 by presenting some other interesting properties of the Facebook and CNN worlds. In Section 5, we describe some of the related work done in this field and their limitations. Finally we conclude in Section 6 by describing the work to be done in future.

2. MEASUREMENT METHODOLOGY

We had access to the packet traces for the ingress and egress traffic of a large American University (University of Wisconsin –Madison). These packet traces provided a detailed view of bi-directional traffic, with attributes like timestamps, source/destination addresses, requested URIs and even payloads. Assuming that all the ingress and egress traffic went through a single link monitored without loss, we can make concrete statements about the usage patterns. We had access to roughly 200 hours of egress data but not much (~30 minutes) of ingress data. Hence, we focused more on the statistics that could be derived from the egress data and its co-relation with the corresponding ingress data.

2.1 Using Browser Plugins

We started by doing a static macro analysis of the approximate content distribution patterns for Facebook using some of the Firefox browser plugins like Tamper Data [12]. Live HTTP Header [13], Firebug[14] and Web-Developer [15]. These tools allowed us to inspect individual elements on a Facebook page and figure out the servers from which the browser was loading the content. A typical Facebook page contains user-interactions in the form of text wall posts, messages and feeds. It also loads the re-sized photos of the users friends, any other video-thumbnails and photos posted on the wall (either by the user or as feeds of friends) and various advertisements. Clicking a video or an album takes the user to another set of pages with another re-sized version of the original photos. Table 2.1 captures our findings about various servers and the type of content they serve

	A	B	C	D	E
1	**Domain Name**	**Content Served**	**IP**	**Location**	**Network**
2	profile.ak.fbcdn.net	ProfileThumbnails	209.18.42.162	Herndon, VA	Road Runner
3	photos-g.ak.fbcdn.net	PPhotos/Ads	72.247.219.75	Cambridge, MA	Akamai
4	photos-d.ak.fbcdn.net	PPhotos/Ads	72.247.219.76	Cambridge, MA	Akamai
5	photos-b.ak.fbcdn.net	PPhotos/Ads	72.247.219.50	Cambridge, MA	Akamai
6	photos-f.ak.fbcdn.net	PPhotos/Ads	72.247.219.40	Cambridge, MA	Akamai
7	photos-h.ak.fbcdn.net	PPhotos/Ads	72.247.219.27	Cambridge, MA	Akamai
8	photos-e.ak.fbcdn.net	PPhotos/Ads	72.247.219.89	Cambridge, MA	Akamai
9	photos-c.ak.fbcdn.net	PPhotos/Ads	72.247.219.66	Cambridge, MA	Akamai
10	photos-a.ak.fbcdn.net	PPhotos/Ads	72.247.219.34	Cambridge, MA	Akamai
11	sphotos.ak.fbcdn.net	Photos/Fat	72.247.219.49	Cambridge, MA	Akamai
12	static.ak.fbcdn.net	Ads	205.213.110.14	Madison, WI	WiscNet
13	hphotos-snc3.fbcdn.net	Photos/Fat	69.63.183.3	Palo Alto, CA	Facebook
14	hphotos-snc1.fbcdn.net	Photos/Fat	69.63.180.186	Palo Alto, CA	Facebook
15	b.static.ak.fbcdn.net	Ads	205.213.110.8 /	Madison,WI	WiscNet
16	platform.ak.fbcdn.net	Apps	205.213.110.8 /	Madison, WI	WiscNet
17	creative.ak.fbcdn.net	Ads	205.213.110.7 /	Madison, WI	WiscNet
18	external.ak.fbcdn.net	External Links	205.213.110.8 /	Madison, WI	WiscNet
19	vthumb.ak.fbcdn.net	Video Thumbnails	72.247.219.97	Cambridge, MA	Akamai
20	hphotos-sjc1.fbcdn.net	Photos/Fat	69.63.183.35	Palo Alto, CA	Facebook, Inc
21	photos-snc1.fbcdn.net	Photos/Fat	69.63.178.42	Palo Alto, CA	Facebook, Inc
22	video.ak.facebook.com	Videos	72.247.219.74	Cambridge, MA	Akamai

Table 2.1

Facebook also uses a unique naming scheme for the images. Table 2.2 describe the scheme

Content Type	Nomenclature
Thumbnails (photos and videos)	End with _t
Album pics (Entire album view)	End with _s
Album pics (Individual)	End with _n
Most Profile pics (on wall)	End with _q

Table 2.2

From our static analysis, one can conclude that almost all the image and video content is delivered to the end-users via Akamai CDN.
The primary domain for delivering the profile photos is profile.ak.fbcdn.net, whereas all the thumbnails and album view pics are served via photos-*.fbcdn.net. The full-size pictures are served by hphotos* and sphotos*.fbcdn.net. In addition Facebook pushes all information to its users via the servers channel=[a-z0-9]*facebook.com

A similar static analysis of CNN revealed all the image content to come from the main .cnn.com server and the CDN nodes i.cdn.turner.com and i2.cdn.turner.com.

To further test our approach and to get a measurement of the amount of fat vs non-fat content accesses on Facebook and CNN, we analysed the traces (mostly in .pcap format) using tShark[16]

We initially isolated the requests and responses for the fat content based on the end-users view of the Facebook servers distributing the content and hence did not take into account the servers serving small profile images and ads. we later verified this by looking at the content lengths of all the images requested and served, the details of which are presented in the next section. We further fine-grained the methodology to analyse the traffic per 10-minute traces since as per [9], one could conclude that an average Facebook user does not generally interact with an OSN continuously for more than 10 minutes.

To identify IP addresses we used reverse DNS lookup mechanisms and public databases and used that to analyse the traffic characteristics at the Facebook nodes over a 41 hour period.

2.2 Mobile Traffic

In order to check whether Facebook and CNN behaved differently for the users who used the mobile versions of the sites in question, we forged the *User-Agent* header sent by a PC browser to that of a mobile phone browser. We tested for approximately 130 different mobile models across 10 providers.
As already discussed, one of our goals was to see if users from a certain geographical location have any similarities in the access patterns that could be exploited.

For this, we looked at the Facebook Fan-pages. A Facebook Page is a public profile that enables one to share a business and product. with Facebook users. There are thousands of fan pages on Facebook pertaining to local businesses, artists, bands, television shows, celebrities, sports, movies etc with hundreds of users interacting with each other everyday. A typical fan page is rich in fat content and has a very high volume of user-interactions.for example, the largest fan-page has approx 6 million fans with thousands of interactions everyday. Facebook provides a way for the admin of such pages to get insights into the users interacting on their page. We acquired the administrative rights to one of the popular Indian undergraduate college - fan-page (BITS Pilani) which has about 8000 active users and interactions per day, and a popular football club (Arsenal). We also created a fan-page for a popular American TV character (Adam Baldwin, who plays Agent John Casey in Chuck on NBC.com) and used media like twitter and forums to publicize it - thereby gaining ~400 members within a month. We then used the Facebook 'page insights' feature to gather statistics about the user base.

3. RESULTS

We analyzed ~43 hours worth of egress Facebook and CNN traffic to estimate the percentage of 'fat' requests in them .

3.1 Overall Statistics

Table 3.1(a) depicts the results

	Facebook	CNN	Facebook access via Mobile
Total Requests	15491107	349099	8968
Fat Requests	2734516	231704	7428
%age	17.65	66.37	82.82

Table 3.1 (a)

It seems that most of the requests via mobile phones is for checking fat content (images,videos). We further analyzed another 96 hours for Facebook and got similar ratios.

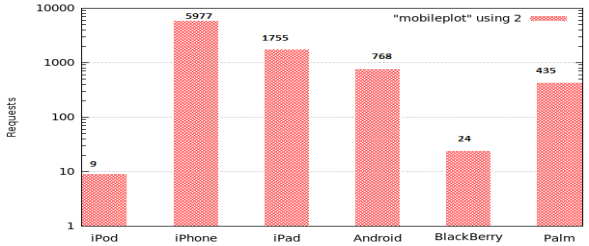
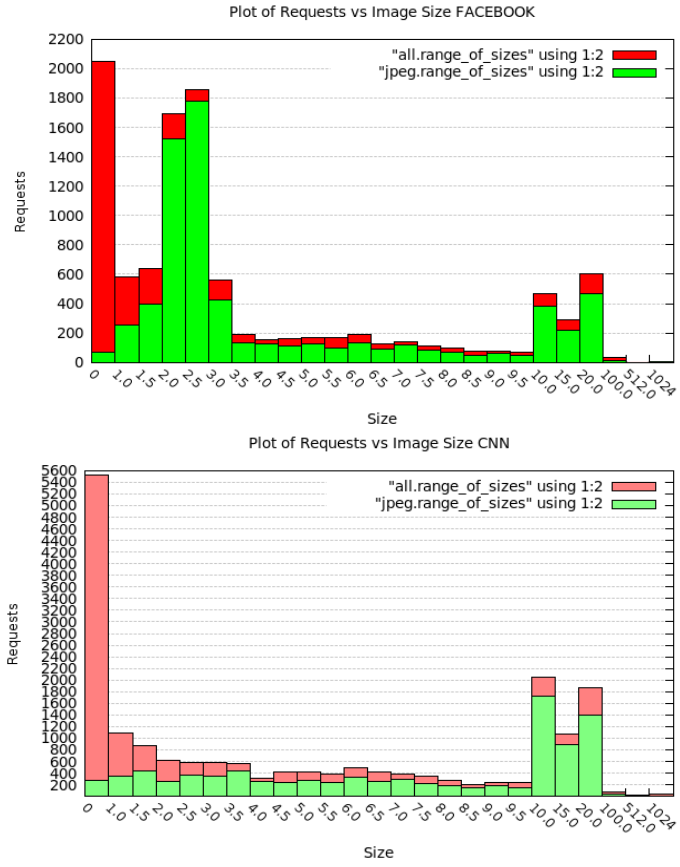


Figure 3.1(b)

For the mobile phones, we also analyzed the devices being used. Apple products are clearly favored by students to access Facebook. Figure 3.1(b) shows a relative histogram for the same.

3.2 Relative Image Sizes



Figures 3.2 (a) and (b)

Figures 3.2 (a) and (b) show the distribution in image sizes for Facebook and CNN responses respectively. From the graphs it is clear that CNN has a higher ratio of images with smaller and larger sizes. JPEG is not the dominant format for smaller sizes (< 1 KB) with GIF and PNG ruling the roost in both the cases.

3.3 Host and Server Statistics

To get an idea of the number of active Facebook users at a given time and the number of 'fat requests' (excluding the small profile and advertisement images) sent by them, we analyzed a continuous subset of the egress traffic spanning 40 hours, using a one hour interval. We find that the number of Facebook users follows the expected diurnal pattern, with peaks at 22:00 hours. The number of users is quite high upto midnight, after which it quickly dies down. The number of fat requests per user is generally between 32 and 70, which can be attributed to browsing a few photo albums. Figure 3.3 (a) depicts this information. Note that the y-axis uses logscale.

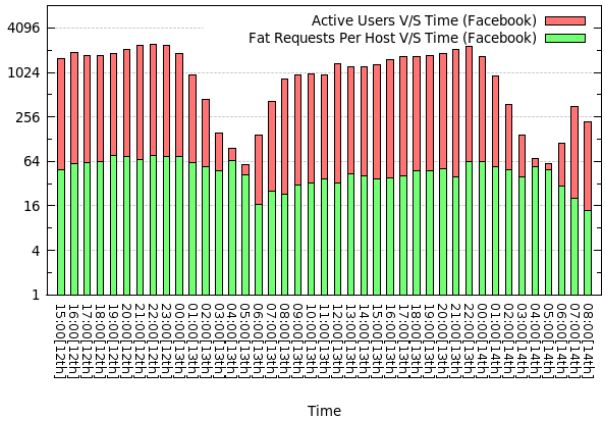


Figure 3.3 (a)

We then estimated the number of servers that dish out the fat content over the 40 hour trace.
We also recorded the average number of requests sent to each server, for every one hour interval of the trace. As expected, the number of active servers follows a diurnal pattern, with a peak around 22:00 hours. The number of fat requests per server however has a lot of variation, as seen in Figure 3.3(b) At midnight, we find around 750 fat requests being processed per server. The pattern is repetitive as expected.

We also found that requests for the same object from the same host (in the same session) sometimes go to different servers. This seems counter-intuitive as one would expect Akamai to route the request to the same server that handled the request the first time, at least in the same session.

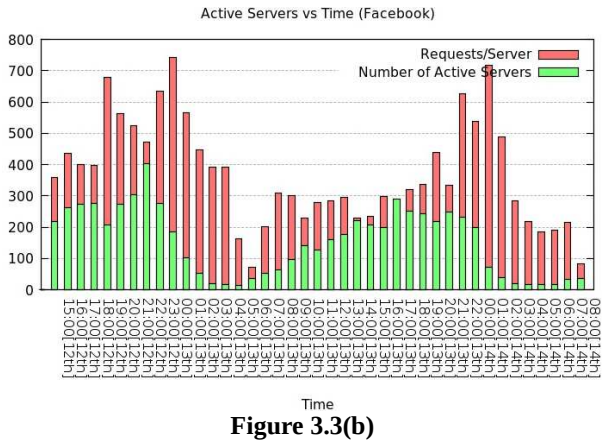


Figure 3.3(b)

To understand the magnitude of fat requests finding their way to the Akamai content delivery servers, we again used the 40 hour trace, broken into one hour intervals. There seem to be day to day variations (12th being a Monday sees higher fat requests than 13th). This cannot

be used as a generalization as it has a number of variables in question. Nevertheless, the numbers are quite astounding! Figure 3.3© depicts this

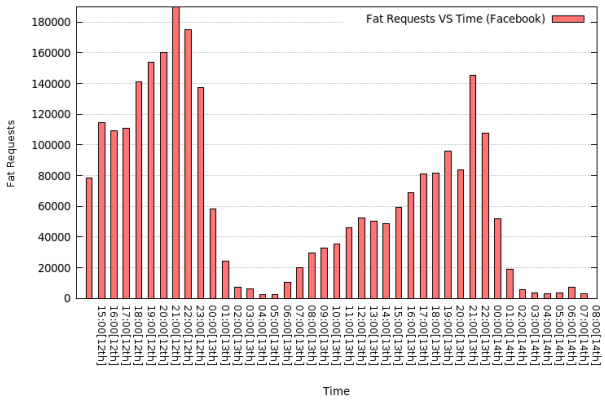
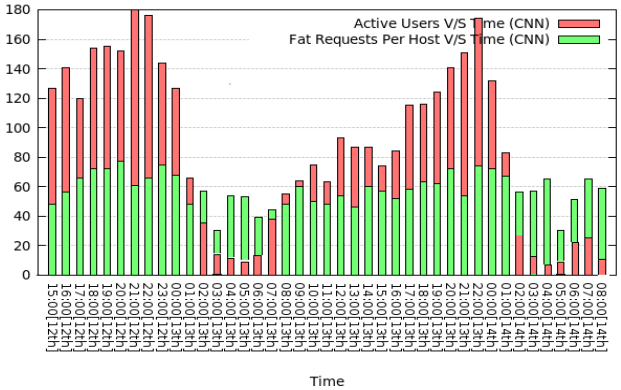
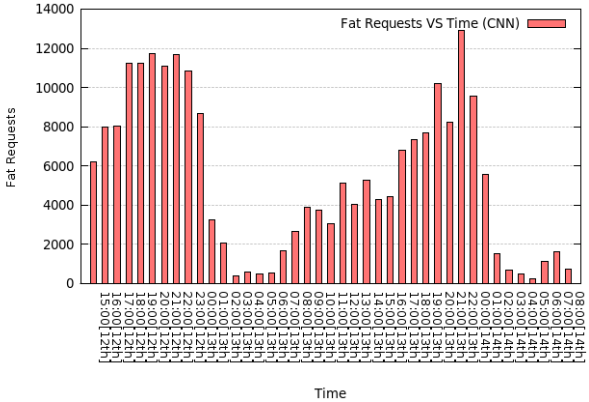
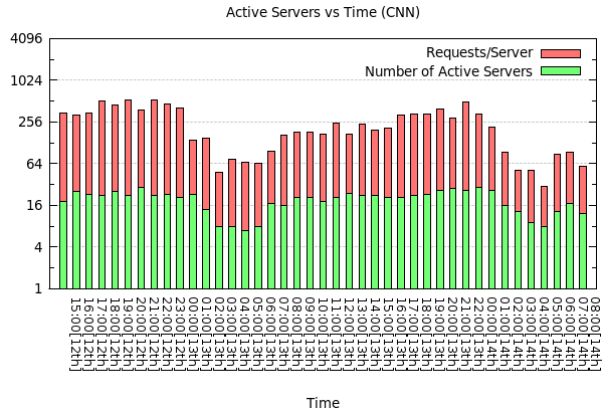


Figure 3.3 (c)

We extracted the same information for fat requests destined to CNNs content delivery servers as well. We see the expected diurnal pattern in all the three figures below (Fig 3.3(d), 3.3(e), 3.3(f)), but at a significantly lower magnitude, be it the number of active CNN users, the magnitude of fat requests or the number of CNN servers dishing out the fat content.
The load on the servers (in terms of fat requests per server) is similar to that of Facebook though.







3.4 Fan Page Insights

Figure 3.4 (a) and (b) show screenshots of Facebook analytics for the fan page corresponding to one of the popular TV shows in USA. It shows the media consumption from a spatial and temporal perspective. In certain fan pages, city level details can also be mined. Content in these fan pages have a high geographic interest locality. For example, the Arsenal football fan club, which is a London based club, has more than half of its active users from London. Such detail can be used to improve data distribution in the Akamai CDN for Facebook.

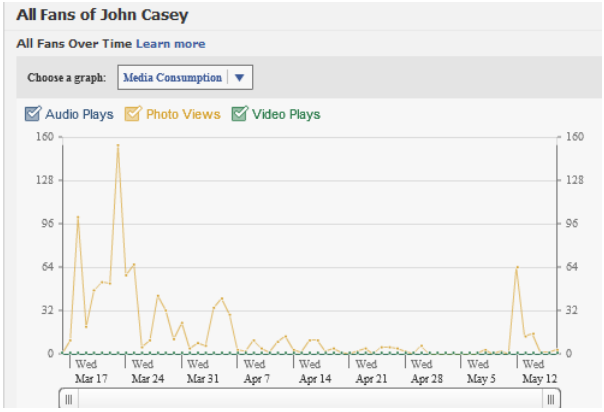


Figure 3.4(a)

Figure 3.4 (b)

4. OTHER OBSERVATIONS

Our study, especially the static analysis of the data brought the following interesting observations to light as well:

1. Facebook AJAX scripts prefetch atleast 1 extra photo during an album access. We observed that when a user clicks on an album and views photo N, the AJAX calls also prefetch the N+1 photo. Hitting "Next" to see the N+1th photo, then triggers a prefetch of N+2 photo. This is in sync with other photo sharing services which use this technique to improve the user experience

2. In certain cases, if a user views a photo in an album and remains inactive for some time, a subsequent request for next photo in the album, triggers the prefetch of at least 3 more images from the start of the album. For example, assuming the album cover for an album with 60 photos is photo # 10, a user click on the album cover will display the cover photo i.e photo #10. At this point, if the user hits "Next" after some time of inactivity, one can see photos #1,2, and 3 getting prefetched. Some arbitrary time later Facebook returns to the normal mode of prefetching an extra image as described in 1, only this time it does not start from the N+1th image but N-kth image. In our runs, we found the k to be arbitrary. So far, we haven't figured out a rational explanation to this behavior and we believe it is a crude attempt to do over-optimization and needs to be investigated further.

3. The prefetching behavior is not seen in the case of CNN. This is also not seen in the case of Facebook accesses via mobile devices.

4. For most touch phones (like iPhone, HTC Android), facebook redirects the user to touch,Facebook.com. For all other mobile phones (like Symbian OS Nokia N Series), the default mobile Facebook server is m.facebook.com. In our tests, we found touch.facebook.com to be displaying more images and m.facebook.com being more restrictive in fat content display. There is also no rational explanation to this. We believe a better approach would be to figure the b/w of the user out and then decide on the content delivery.

5. The fat content for the Facebook access via a mobile device is also delivered via the CDN nodes. However , that for CNN appears to come from .cnn.com domain.

6. CNN shortens the URLs in the case of mobile accesses unlike Facebook.

7. Facebook is still not used extensively for video sharing. We found approximately ~0.0017% requests for videos uploads/views.

8. The requests are evenly load balanced across the Akamai facebook CDN domain names (photos*.fbcdn.net). This, however is not the case with CNN wherein there is a 1:8 ratio of request loads.

5. RELATED WORK.

There have been attempts to do measurement studies for OSN's before, we however argue that a large scale study of the Facebook site especially with respect to fat-content dissemination and its implications on the network has not been attempted before.

Balachander et al [1] identify distinct collections of the user content from the publicly available data and try to characterize their behaviors and geographic growth. However their work is limited to Twitter, a highly popular micro-blogging application which allows the users to post no more than 140 character messages. However our focus is more on the fat-content like videos and images.

Nazir et al. [2] monitor and characterize the usage of third party Facebook applications. By further studying the interactions between these applications and the OSN users, the same authors identify some potential performance bottlenecks within the Facebook server infrastructure [3]. Prefetching can be used to enhance the user experience here.

Wilson et al [4] have attempted to do a large scale study of the Facebook network by crawling Facebook regional networks (which have been phased out by Facebook since last year), However their analysis is more focused on whether social links are real indications of the real life interactions and hence is significantly different from our approach.

Mislove et al [5] gather datasets for multiple social networks (namely Hi5, Orkut, LiveJournal and Youtube) and try to identify various characteristics of an OSN. However, their study, though relevant in understanding how to measure an OSN, does not focus on the amount of the fat-content generated or accessed.

6. CONCLUSION AND FUTURE WORK

Our study has helped gain more insights into the content distribution of Facebook and how it compares with an popular non-OSN site CNN wrt the fat content accesses. However, there are still some rough ends to be sorted out that will strengthen our results.

For starters, we would want to analyze more ingress traffic and try to draw some correlations between the egress and ingress traffic wrt user sessions and the activities per session.
We have also written a Facebook Application that mines the user data during their active session on Facebook. We will add more functionality to gather data regarding user friends, album and videos liked.

Finally, we would attempt to analyze more Facebook Fan pages since the level of user interaction on such pages is very high and we do see a lot of spatial locality in the active users.

6. REFERENCES

[1] Balachander Krishnamurthy , Phillipa Gill , Martin Arlitt, A few chirps about twitter, Proceedings of the first workshop on Online social networks, August 18-18, 2008, Seattle, WA, USA

[2] Nazir,A., Raza,S., Gupta,D., Chuah,C.-N., and Krishnamurthy, B. Network-level footprints of Facebook applications. In Proc. ACM IMC (2009).

[3] Nazir A., Raza,S., Chuah,C.-N.,. Unveiling Facebook: A measurement study of social network based applications. In Proc. ACM IMC (2008).

[4] C. Wilson, B. Boe, A. Sala, K.P. Puttaswamy, and B. Y. Zhao, User interactions in social networks and their implications. In Eurosys '09: Proceedings of the fourth ACM European conference on Computer systems, pages 205-218, New York, NY, USA, 2009 ACM.

[5] Mislove, A., Marcon, M., Gummadi, K. P.,Druschel, P., and Bhattacharjee, B. Measurement and analysis of online social networks. In Proc. ACM IMC (2007)

[6] Sarita Yardi, Amy Bruckman, Modelling the flow of information in a social network

[7] Cha, M., Mislove, A., Gummadi, K. O., A measurement-driven analysis of information propagation in the Flickr social network. WWW2009

[8] B. Krishnamurthy. A measure of online social networks. In COMSNETS, 2009.

[9] Fabian Schneider, Anja Feldmann, Balachander Krishnamurthy, Walter Willinger Understanding Online Social Networks from a networks perspective. Proceedings of IMC'09, November 2009

[10] Facebook Statistics: http://www.facebook.com/press/info.php?statistics

[11]Facebook: Storing billions of photos http://www.facebook.com/note.php?note_id=76191543919

[12] Tamper Data https://addons.mozilla.org/en-US/firefox/addon/966/

[13] LiveHTTP header *https://addons.mozilla.org/en-US/firefox/addon/3829/*

[14] Firebug http://getfirebug.com

[15] Webdeveloper *https://addons.mozilla.org/en-US/firefox/addon/60/*

[16] tshark : *www.wireshark.org/docs/man-pages/tshark.html*