Examining the Network Traffic of Facebook Homepage Retrieval: An End User Perspective

Watcharee Wongyai
School of Information Technology
Mae Fah Luang University
Chiang Rai, Thailand, 57100
Email: watcharee.won@mfu.ac.th

Luck Charoenwatana
School of Information Technology
Mae Fah Luang University
Chiang Rai, Thailand, 57100
Email: lcharoen@gmail.com

Abstract—Facebook is one of the most famous social network sites hosting a number of users approaching a billion [1]. Facebook is a cloud-based web site which contains a number of advanced technologies behind the scene. Despite the fact that it is the web site that most users open for all-day and, in some places, all-night long, and its traffic is generally part of all types of networks — wired and wireless, PAN, LAN, MAN, and WAN —, there is virtually no research work or technical paper that report on client’s perspective of what happen when users login onto their Facebook homepages. How many and in what order components loaded are, which and how many servers those components loaded are, or how many TCP streams used, are examples of questions which users and network administrators have a little knowledge of. Our work tries to answer the questions by examining every of over 2,000 packets per a single Facebook homepage retrieval using the Wireshark packet capturing software. From the captured packets, we thoroughly examined all objects that Facebook retrieved and categorized them into groups based on the characteristics. Our investigations found that Facebook retrieves these objects from both international and domestic servers, and over a number of parallel TCP streams. The results also showed that the Facebook traffic exhibited a two-spike bursty pattern, regardless of the loading time.

Keywords-traffic analysis; Facebook forensics; TCP; packet capturing

I. INTRODUCTION

Facebook is arguably the most popular social network site on the Internet. At the end of 2011 it has 845 million users [1]. It is common today to see people around you connect to Facebook all the time. In a LAN network in an office or a computer lab of a university campus, it is quite certain that there will be a number of users being online with Facebook, either in foreground or background processes. Recently Facebook introduces applications, such as the Facebook homepage and the Facebook Messenger, on mobile device of various platforms. This certainly increases connectivity between users and Facebook. Facebook reports that, out of the 845 million users, over 400 of which are access on mobile devices. This means Facebook traffic is persistent in all types of networks – wired and wireless, and from PAN to WAN coverage levels.

Facebook is a cloud-based web application, meaning that its services and storage are distributed in many physical machines in many geographical locations, generally unknown to users. A single web page of Facebook can contain more than one types of information that has different characteristics, for example comment texts and pictures. Facebook deploy various types of cutting edge technologies to support low-latency loading time of such heavily-distributed contents, for instance, Apache Hadoop, Memcached, Haystack, Bigpipe, and Varnish, and improve them constantly in relatively fast pace to ensure the leading position in the market [2].

From network administrators’ standpoint, this hints that the Facebook traffic is not simple. Based on its distributed backend operations, Facebook traffic would certainly not a single HTML file from a single web server. Some fundamental questions arise – How many components are loaded and can they be categorized? Are there certain sequences of what are loaded first and what are loaded last? From how many IP addresses are these components loaded and what are they? Are they UDP or TCP traffic? How many TCP streams are there, and are they accessed in sequence or in parallel? These are questions that lead network operators toward their setups, customizations, and improvements on the network design and planning. For instance, they might realize that the current configuration of a proxy server does not help reducing the loading time of Facebook traffic at all.

This paper is organized as follows. The next section provides literature review of similar research works. In section three, research setups and scenarios are explained. Results and analysis are presented in section four. Section five concludes the paper.

II. RELATED WORKS

Previous studies of Facebook are in numerous areas, socially, such as [3, 4], and technically [5-8]. On the technical side, the paper [5] analyzed popularity and accessibility of Facebook applications through analytic and crawling tools. The paper [6] studies Facebook Chat and investigates on artifacts left on hard-disk drive after the user logs out. The paper [7] studied and compared traffic characteristics of the Bit Torrent, YouTube, and Facebook. The authors captured the applications’ traffic from the backbone network and analyzed it in application level, such as traffic volume and number of packets, and in UDP and TCP flow level.

Authors in [8] conducted experimental works to examine and try to detect web flows of the voice-over-IP application named Skype. The research used the tcpdump program to capture and extract packets for analysis and differentiate the Skype traffic from the normal web traffic. Our work will use
similar methods to capture and analyze the packets but will be different in what to look for and that it is applied to Facebook rather than Skype.

III. METHODOLOGY

Our research work comprises two steps, packet capturing and analysis. The packet capturing is done on a PC that runs two programs, a web browser and the Wireshark packet capturing program. Two Facebook accounts with few and many friends have been used. The account with few friends exhibit a relatively static data updates on the homepage while the other account generally contains more pictures and updates. This is to increase the randomness of the content.

A. Sections of Facebook Homepage

The Facebook homepages in our investigation is the 2012 version. It comprises 6 key sections, as shown in Figure 1. The first section is the top bar that contains notifications, search bar, and control menus. The second section on the left is the navigator area containing shortcuts to other Facebook pages, applications, and group. The third section which dominates the page content is the news feed. This part of the page contains status, pictures, video links, and comments of friends of the user. The fourth part is advertisement and birthday and event notifications. The fifth part is a new section that Facebook recently introduced in late 2011, called Ticker – a real-time updates on status, friendship, photos, videos, and comments. The last section is the Facebook Chat. The chat list window display pictures, names, and status of online friends.

![Figure 1. Sections of Facebook homepage](image)

Apart from their differences in functionality to users, another reason that we categorized the sections is because of that the implementations of these sections are likely to be different. Base on the technical Facebook Engineer’s publication [2], status, photos, videos, and comments can be categorized into two data patterns, a short set of volatile temporal data and a set of data that rarely gets accessed. This results in having these data treated, stored, and accessed differently, and consequently creates different network flow patterns at the client.

B. Packet Capturing and Analysis

All data collections are conducted over an ADSL network of an ISP in Thailand by one user with two Facebook accounts. The operations begin by running a Google Chrome web browser and a Wireshark process. The history of the browser is thoroughly deleted before the capture begins. The HTTPS protocol of Facebook is turned off and the plain HTTP is used so that the payload of the packet is readable. The packet capturing begins when user logins and ends when the homepage is completely loaded. In case of investigations of the Ticker and the Facebook Chat, the capturing continues until some updates of these sections are collected.

Complete packets of two Facebook account homepages traffic have been selectively collected in a period of 2 months. Prior to the real data collection for analysis, a two-week trial of extensive packet capturing had been conducted to validate data and select appropriate time for data collection. It was found that during the peak hours in which the ADSL pipe was highly congested, the Facebook page load was unable to complete regardless of the Facebook account used. The reason behind is that the Facebook page makes requests to a number of servers on multiple IP addresses, some in sequence and some in parallel. When data form an IP address failed to load, the data retrieval from subsequent sources failed too. This hints us that Facebook does not work well under insufficient connection speed.

After the trial period, there are 5 times of a day selected for data collection, in which the network is generally not overloaded and the captured packets are usable for analysis, 01.00, 02.00, 13.00, 14.00, and 23.00 hrs. Since the goal of this research project does not pay much attention on the page’s loading time as it can be greatly vary depending on the connection speed and, in our viewpoint, does not create much impact to users as long as the page load is complete, the completeness of the page load is more important as it would exhibit behaviors, characteristics, and pattern of the Facebook page load. With 5 times a day for several weeks of data collection, the collected data exhibits randomness of a number of texts, profile images, posted images, posted multimedia contents, notifications, online users, Ticker messages, and different advertisement on the Facebook page, as well as sufficient mixture of connection speed for page loading time.

The Wireshark program captures packets in a file. Analyses of these files are conducted using Wireshark tools, such as flow tracer, TCP flow graph, and statistic examination. Various filters, such as IP, TCP, UDP, and HTTP, are applied to the capture file to extract targeted information. Overall, over 60,000 packets are examined by this research project.

IV. RESULTS AND ANALYSIS

A. Types of the Data

Our studies find that the Facebook homepage contains numerous objects of many file types. However, these objects can be categorized into group according to their properties, such as file type, naming pattern, source IP address, and the section or location on the page. Each section of the page
generally contains similar objects, and these objects generally have similar naming patterns and source IP addresses. For example, a friend’s profile image showed in the news feed section has a naming pattern of x_x_x_q.jpg and has been sent from the IP addresses 203.150.x.x.

Table I summarizes types of retrieved objects. The first few types are the core files that belong to the entire page, which are, for example, home.php and ai.php. File types 4 to 9 are controllers of sub components of the homepage, such as the Ticker, the blue bar, and the Chat window. The types 14 and 15 are profile images of friends that appear in front of every post of that user. The file image has 50 x 50 pixels. The images of type 18 are those posted by users in the news feeds section, which usually larger than the profile images. Images and icons of advertisements and game are the file types 19 to 22. The type 23 and 24 are controllers of the Tickers and the ‘More stories” bar. The last file type is called when the user performs actions, such as posting a comment and pressing the Like or Share links.

B. The Number of IP Addresses Accessed

The IP addresses from which the objects have been retrieved have also been illustrated in Table I. Our observations have found that a single Facebook homepage retrieval has made connections to between 15 and 25 different IP addresses, both domestic and abroad. Facebook uses a number of the US-based IP addresses interchangeably to get the core and controls files, most of which are the .php files. These servers seem to be redundant and have a similar set of data that the client can interchangeably access.

The domestic servers seem to be responsible for image storage. Facebook accesses these server for file types 10 to 22, which mostly are images of extensions *.jpg, *.gif, and *.png. Similarly to the international servers, these servers seem to be redundant and are interchangeably accessed.

A special attention has been paid on the file type 18, which are the image files of the naming pattern *n.jpg, as there are many IP addresses attached to them. The *n.jpg are images on the news feed section. It can be either images that the users uploaded themselves, or links of images that they share. Our investigation found 7-10 IP addresses involved. These servers are interchangeably and randomly accessed. There is only 1 IP address that is in Thailand, the rest are not. The abroad IP addresses are from the US, Japan, and South-East Asia countries, which are Singapore and Malaysia. The servers accessed from SEA region has their domain name belonging to the Akamai Technologies [9], a content delivery network provider in which Facebook is one of its customers.

C. Order of Data Retrieval

Regarding the order of object arrival, our study found that the arrival of each homepage section is in fixed order. The object retrieval order is similar to the order of types shown in Table I. Specifically, the page first retrieves some core files of the page, which are the object types 1 – 9. It next pulls a number of other core files, pictures, and scripts of extensions .js, .css, .png, and .gif, which are the objects of types 10 - 13. Thirdly, the page loads the news feed section, which is the largest section of the page. In this part, the objects of types 14 – 18 are loaded, which are profile images, thumbnail of the linked videos, location information, and the images posted by users’ friends. Note that the size of the profile images is 50x50 pixels and that of the friends’ posts is 320 x 320 pixels. Lastly,
the advertisements and the controllers of the Ticker and activities are loaded.

D. TCP Streams

Each time the homepage is completed, there are averagely 2,000 packets retrieved. This is because Facebook limits the contents that will be displayed on screen. As the users scrolls further down the page, more content are gradually loaded. Out of 2000 packets, there are approximately 1,800 TCP packets, and 200 UDP packets. UDP packets are used solely for name resolution. The contents of the page are all transferred over TCP.

As seen from the section B that there are a number of servers that Facebook accesses during the page load. Facebook creates a number of parallel TCP connections to these servers. Facebook does not use one TCP stream per one object. It creates 3-7 parallel streams for data transmissions. On one TCP stream, when an object is completely loaded, the next object is requested over the existing stream.

E. Traffic Pattern of Data Retrieval

This section illustrates the traffic pattern of data retrieval. Figure 2 depicts the amount of data in Kbytes retrieved from the servers since the beginning through the end of the page loading. There are 2 sets of results shown in the figure. The dash line is a Facebook page load of one capturing, and the solid line is of another capturing. They are both a Facebook page load, but on two different occasions with different page load duration. Specifically, the page load of the dash line finished quicker as the capturing was conducted in early morning where utilization of the ADSL link was very low. The solid line output is the result of the slower loading due to greater ADSL link utilization.

Despite the differences in loading durations, it can be seen that the traffic pattern is similar. Specifically, there are two peak periods, approximately at the middle and at the end of the transmission. The first burst is the retrievals of the second set of the core files of types 10 – 15 in Table I. The second peak occurs when Facebook retrieves images, file types 18, for the news feed section. All of our analyses suggest that Facebook data retrieval always exhibits this traffic pattern regardless of the page loading time.

In terms of TCP connection standpoint, it is worth noting that the TCP sessions that used to retrieve data of the first peak did not last until the end of the page loading. In packet analysis, the FIN-flagged packets are exchanged at the middle of the session. Facebook fully terminates the connections of the first data set before making a new set of TCP streams to retrieve images that create the second traffic burst.

Figure 2. Amount of data loaded along the time

V. CONCLUSION

Our key contributions of this paper is based on our motivation that an extensive number of people around the world are using Facebook but only little about the data being retrieved to their computing devices have been known. Many people assume that it might be a little more complicated than a conventional single-HTML file traffic pattern. By investigating what happened when a user load a Facebook homepage and analyzing packets captured by the Wireshark program, we have found that such cloud-based web application is much more complicated than such assumption.

The results firstly showed that there were various kinds of data loaded to the client. These files can be categorized into 24 kinds according to its properties, such as file type, name pattern, source IP address, and location or section on the page. There images retrieved can be categorized further into a number of sub-type. Specifically, there are Facebook icons, profile images, posted images, thumbnails of linked online videos, and advertisement images. The results also showed that there were a number of servers that Facebook had made requests for the homepage objects. These connections are all TCP streams and more than one TCP streams are created to retrieve data simultaneously. Essentially, there are more than 15 servers accessed, both international and domestic. International servers includes those located in US, EU, and Asia. The browser made requests for a certain type of data to a certain group of servers. The servers in the same group are redundant and can offer contents to clients interchangeably. Moreover, the order of data retrieval was also investigated. It was found that the data was not loaded randomly. Each section of the page was loaded in order but the objects in each section were loaded in parallel. The final analysis showed that Facebook data retrieval exhibited the two-spike traffic pattern, at the middle and at the end of the page load. Such pattern occurs regardless of the amount of the page loading time.

In our future works, we aim to investigate further into two directions. First the traffic information of other Facebook applications, such as photos, videos, and games, can be more insightfully investigated. The second direction is to examine the aggregated traffic patterns and characteristics, and seek for enhancement from networking standpoints, such as how we can improve the proxy server that is more suitable for Facebook traffic.
REFERENCES


