UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

Analysis of Web Protocols Evolution on Internet Traffic

Karikari Abina Mary

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática

(2º ciclo de estudos))

Orientador: **Prof. Nuno M. Garcia**

Covilhã, Junho de 2014

Acknowledgements

First and foremost, I want to express my sincere gratitude to the Almighty God, for his protection, strength and support. Him alone has been the brain behind my success during the period of this study.

I am greatly indebted to my supervisor Professor NUNO. M. GARCIA for his advice, guidance, corrections and useful suggestion during the various stages of this research work. His scholarly criticism and interest in this research has been very rewarding, his promptness and sense of duty during my consultation with him has been highly commendable and which has made great impact in my life. My profound gratitude goes to him.

My special appreciation goes to Mr Oluwafemi Olawale, Mr William Okorukwu Okey, Mr Diallo Ousmane and all my colleage in ALLAB (Assistance Living and Telecommunication Laboratory) for their assistance and support in one way or the other towards the success of this research work. My profound gratitude also goes to my Pastor Mrs Funmi Olanrewaju and her husband Mr Tola Olanrewaju for their prayer and word of encouragement towards the success of this research work.

My profound gratitude also goes to my beloved parent Mr Kingsley Karikari and Mrs Comfort Karikari, Mr Nosa Osasere and my beloved uncles Mr Clement Conno, Mr Kingsley Oridomo for their prayer, moral and financial support towards the success of this project.
This acknowledgement would not be complete without my appreciation to my lovely daughters Blessing Osasere and Joy Osasere for their understanding and cooperation throughout the period of this course.

Resumo

Esta pesquisa foca-se na análise de dez anos de tráfego de Internet, a partir de 2004 até 2013, capturado e medido pelo Mawi Lab numa ligação de fibra óptica entre o Japão e os Estados Unidos da América. O tráfego recolhido foi analisado para cada um dos dias nesse período, e também conjuntamente nesse período. As questões de pesquisa iniciais incluíram testar a hipótese de ser observável no tráfego gerado, a alteração das aplicações em uso na Internet e a alteração dos padrões de uso da Internet. Vários protocolos foram analisados exaustivamente, incluindo HTTP, HTTPS, TCP, UDP, IPv4, IPv6, SMTP e DNS. O efeito da transição do IPv4 para o IPv6 também foi analisado. As conclusões foram tiradas, as questões de pesquisa foram respondidas e a hipótese de pesquisa foi confirmada.

Palavras-chave

Tráfego da Internet, medição de tráfego, análise de tráfego, transição do IPv4 para o IPv6, segurança na Internet, histórico de tráfego, uso da Internet, evolução da Internet.

Abstract

This research focus on the analysis of ten years of Internet traffic, from 2004 until 2013, captured and measured by Mawi Lab at a link connecting Japan to the United States of America. The collected traffic was analysed for each of the days in that period, and conjointly in that timeframe. Initial research questions included the test of the hypothesis of weather the change in Internet applications and Internet usage patterns were observable in the generated traffic or not. Several protocols were thoroughly analysed, including HTTP, HTTPS, TCP, UDP, IPv4, IPv6, SMTP, DNS. The effect of the transition from IPv4 to IPv6 was also analysed. Conclusions were drawn and the research questions were answered and the research hypothesis was confirmed.

Keywords

Internet traffic, traffic measurement, traffic analysis, IPv4 transition to IPv6, Internet security, traffic history, Internet usage, Internet evolution.

Contents

List of Figures

List of Tables

Acronyms

TCP Transmission Control Protocol
UDP User Datagram Protocol
FTP File Transfer Protocol
HTTP Hypertext Transfer Protocol
HTTPS Hypertext Transfer Protocol Secure
IP Internet Protocol
DNS Domain Name System
SMTP Simple Mail Transfer Protocol
ISP Internet Service Provider
MSN Microsoft Network
OSI Open System Interconnected Model
NCP Network Control Protocol
ARPA Advanced Research Projects Agency
IPv4 Internet protocol version 4
IPv6 Internet protocol version 6
POP3 Post Office Protocol
ACK Acknowledgement
NACK Negative acknowledgement
SSL Secure Sockets Layer
SSH Secure shell
IMAP Internet Message Access Protocol
SNMP Simple Network Management Protocol
TFTP Trivial File Transfer Protocol
NFS Network File System
TLS Transport Layer Security

Chapter 1

Introduction

1.1 Background

Over the past decades, there have been several technological applications that leverage its performance on the Internet. However the drastic increase in research and technological advancement and application development in recent time has popularized the term Internet.
The Internet which refers to the global information system that is logically linked together by a globally unique address space based on the Internet Protocol (IP). Its subsequent extensions which is capable of supporting communications using the Transmission Control Protocol, User Datagram Protocol (TCP|UDP) suite and other IP-compatible protocols; and also provides, uses or makes accessible, either publicly or privately, high level services.
Traditionally, UDP which is a much simpler protocol when compared to TCP, because it doesn't require connection setup delays, flow control, or retransmission, and has been used as a transport layer protocol for real-time applications in recent time. Currently, more than 80 percent of the Internet's bandwidth [1] is consumed by TCP-based applications, such as the Hypertext Transfer Protocol (HTTP) and Hypertext Transfer Protocol Secure (HTTPS). Under the TCP flow control, this uses a sliding window flow mechanism. Network traffic is recognized by detection of packet loss. However, when this occurs, the packet is retransmitted. It is still an accepted assumption that Internet traffic is dominated by TCP [2,3].
However, the rise of new streaming multimedia applications [4] such as Microsoft Network (MSN), Skype, Facebook, YouTube, *etc.* and new P2P protocols that try to avoid traffic shaping techniques (such as RST packet injection) will increase the use of UDP as a transport protocol in future. This substantial increase in UDP usage has raises serious concerns about fairness and stability in the Internet because; currently UDP lacks functionality to adapt to network traffic congestion.

The tremendous increase in the use of these protocols in several applications and the need to meet user needs has necessitated research on Network traffic analysis. Network traffic analysis is a process of capturing network traffic and inspecting it closely to determine what is happening in the network [5]. It is also known by several other names: network analysis, protocol analysis, and packet sniffing and packet analysis to name a few.
Network traffic analysis which is a scientific approach that involves collection and gathering of network traffic data over a period of years, processing, analyzing and interpreting them into a useful information for effective and efficient decision making and tries to highlight the

use and performance of these protocols over a period of time and how they can be improved to meet future use.

This desire to conceptualize network traffic in a prevailing communication network has helped to tackle vast range of problems, including security, attacks and monitoring general health of the network. This project is aimed at evaluating the Analysis of web protocols evolution on Internet traffic.

This research focuses on traffic statistics of the traces that have already been collected every day, for 15 minutes, starting from 2004 January -2013 December. The traffic traces that were utilized in this research was obtained from the online traffic data repository maintained by the Measurement and Analysis on the Wide Internet (MAWI) working group of the WIDE Project and from TMA portal- European research portal on traffic monitoring at www.tma-portal.eu [6].

The longitudinal study on the analysis of web protocols on Internet traffic were investigated at (packet connection level) and its application usages. The research examined and analyzed the Internet traffic over these periods (2004-2013). Also an attempt was made to highlight the trend on Internet applications on Social Network such as Microsoft Network (MSN), Skype and Facebook with the view to analyze their trend on Internet Network Traffic and security issues within the study period.

1.2 Objective

The main aim of this research is to evaluate the changes in Internet Protocol (IP) traffic, and to analyze and visualize the network traffic data at packet connection levels. We try to address various concerns related to the percentage of TCP/UDP traffic, monitor the relevance of the web protocols traffic over the Wide Area Network (WAN) and the Internet. However, an attempt was made on the history, traffic growth and usage of Internet Applications such as Skype, Facebook and MSN.

The achievements of these aims were realized by the following objectives:

(I) analyzing the previously recorded TCP and UDP connection traffic, such as the amount of traffic being transmitted from or received by the host machine, based on different applications.

(II) Examining the details of the data transmission including the size of messages transmitted or received during a given period of time, the source IP address, the destination IP address , the source and destination port number, the type of protocol that have been used.

(III) Evaluating the traffic growth and usage of the Internet Applications.

1.3 Research Questions

Given the current changes in technology and increasing popularity on the numerous online services and applications such as VoIP, VIDEO, and WEB that obtain its performance mostly on

TCP and UDP protocols, how has this change affect the statistics for Network Traffic on the Internet from 2004 -2013?
The following research questions were established so as to provide concrete objectives for the thesis.
(I) what is the evolution of the ratios of use for the different transport protocols over the studied 10 years? Here we addressed various concerns related to the percentage of TCP|UDP traffic, HTTP traffic, IPv4 |IPv6 traffic and their statistics.
(II) What is the effect of these protocols on the Internet Traffic for this period (2004 - 2013)?
(III) What is the evolution on the amount of traffic being transmitted from or received by the hosts in this link? Here we obtained previously recorded flows of IP traffic on the Internet for a particular link, evaluate and analyzed the records of the link traffic, *i.e*. The amount of traffic being transmitted from or receive by the hosts.

1.4 Hypothesis

The change in the Internet application ecosystem in the last 10 years is perceivable through the analysis of the Internet traffic at a random yet ergodic collection point.

1.5 Thesis organization

This thesis is organized in five chapters.

Chapter 1 introduces the scope of this work; define the object, research questions and the hypothesis, the remaining chapters are organized as follows:

Chapter 2 discussed the State of the Art, it also covers the evolution TCP/IP and the Internet, including a description on the Open System Interconnected Model (OSI) layers and details of headers for several important protocols, network traffic, Internet traffic Measurement, the evolution of Internet traffic and COST Action IC0703.

Chapter 3 describes our data trace selection point and the methodology involved in analyzing the traffic data. It also discusses about the approach utilize for analyzing the traffic data, Limitation on the recorded data and the information collected from the WIDE project and MAWI.

Chapter 4 presents the results for our data trace analysis for each protocol that we use for these research and analyze our results at different levels of granularity.

Chapter 5 summarizes the conclusion, followed by the future work for further study of our work presented in this thesis.

Chapter 2

State of the Art

Before discussing the evaluation on the usage of different web protocols on Network Traffic, it is useful to talk about the evolution of TCP|UDP/IP and Internet Applications on the Network Traffic. This chapter is divided into six sections, each relating to a important aspect of the research. The first section explains the evolution of TCP/IP and the Internet, including a description on the Open System Interconnected Model (OSI) layers and the details of the headers for several important protocols. Section two discusses the network Traffic including network monitoring, Traffic flow, Traffic profiling and Traffic analysis. Section three focuses on Internet Traffic Measurement, discussing the evolution of Internet Traffic and COST Action IC0703. As to allow us to make comparisons and understand the nature of the collected network Traffic, section four discusses some of the most used network applications for the studied period. This includes MSN, Skype and Facebook. Section five shows recent trends on network traffic growth, including Web browsing and Web based email. Finally section six presents data on the expected growth of social network traffic.
These sections will allow us to make a thorough analysis on the collected data trace.

2.1 The Evolution of TCP/IP and the Internet

Today the Internet is known as a network of networks that is basically changing social, political, and economic structures, and in many ways obviating geographic boundaries. This prospective is merely the recognition of predictions that go back nearly forty years ago. In a series of memos dating back to August 1962, J.C.R. Licklider of MIT discussed his "Galactic Network" and how social interactions could be enabled through Networking.

TCP/IP is a standard suite of protocols that is designed for huge networks consisting of network segments that are linked by routers. It is the protocol that is used on the Internet, which comprises of thousands of Networks worldwide that connect research facilities, universities, libraries, government agencies, private companies, and individuals [7]. TCP/IP is a set of network standards that specify the details of how computers communicate, as well as a set of conventions for interconnecting Networks and routing traffic. The Internet certainly provides such a national and global infrastructure and, in fact, interplanetary Internet communication has already been seriously discussed. TCP/IP was initially designed to meet the data communication needs of the U.S. Department of Defence (DOD).

In the late 1960s the Advanced Research Projects Agency (ARPA, now called DARPA) of the U.S. Department of Defence began a partnership with U.S. Universities and the corporate research community to design open, standard protocols and build multi-vendor networks.

Together, the participants planned The Advanced Research Projects Agency Network (ARPANET), this was the first packet switching Network.

The first experiment for the four-node version of ARPANET went into operation in 1969. These four nodes were connected together at three different sites via 56 Kbit/s circuits, using the Network Control Protocol (NCP). The experiment was a success, and the trial Network finally evolved into a useful operational Network, the "ARPA Internet".

In the year 1974, Vinton G. and Robert E. proposed in a paper the design for a new set of core protocols, for the ARPANET. The official name for the set of protocols was called TCP/IP Internet Protocol Suite, which is generally referred to as TCP/IP, which is taken from the names of the Network layer protocol (IP) and one of the transport layer protocols (TCP).

The Institute of Information Sciences at University of Southern California presented a reference document in January 1980 [8] describing the values of the Internet Protocol, designed to be used in an environment of computer communication networks positioned to packet switched systems interconnected between them.

In 1985 ARPANET was faced a problem of congestion and the National Science Foundation's decide to developed NSFNET to support the prior net which was finally closed in 1989. The NSFNET was built on several regional networks and peer networks such as NASA Science Network. There was a network architecture connecting campuses and research organizations connected also to super computer facilities in 1986. Due to increase in speed of transmissions over the past years, the backbone was moved to a private company in 1991, this innovation make them to start charging for connection for their services, and companies like IBM developed ANSNET in parallel which was now aimed to enrich these companies.

Below is the Summary of TCP/IP milestones:
The history of TCP/IP can be traced back to research conducted by the United States Department of Defence (DOD) Advanced Research Projects Agency (DARPA) in the late 1960s and early 1970s [7].
The following list highlights some important TCP/IP milestones:

- In 1970, ARPANET hosts started to use Network Control Protocol (NCP), a preliminary form of what would become the Transmission Control Protocol (TCP).
- In 1972, the Telnet protocol was introduced. Telnet is used for terminal emulation to connect dissimilar systems. In the early 1970s, these systems were different types of mainframe computers.
- In 1973, the File Transfer Protocol (FTP) was introduced. FTP is used to exchange files between dissimilar systems.
- In 1974, the Transmission Control Protocol (TCP) was specified in detail. TCP replaced NCP and provided enhanced reliable communication services.

- In 1981, the Internet Protocol (IP) (also known as IP version 4 [IPv4]) was specified in detail. IP provides addressing and routing functions for end-to-end delivery.
- In 1982, the Defense Communications Agency (DCA) and ARPA established the Transmission Control Protocol (TCP) and Internet Protocol (IP) as the TCP/IP protocol suite.
- In 1983, ARPANET change from the name NCP to TCP/IP.
- In 1984, the Domain Name System (DNS) was introduced. DNS resolves domain names (such as www.example.com) to IP addresses (such as 192.168.5.18).
- In 1995, Internet service providers (ISPs) began to offer Internet access to businesses and individuals.
- In 1996, the Hypertext Transfer Protocol (HTTP) was introduced. The World Wide Web uses HTTP.
- In 1996, the first set of IP version 6 (IPv6) standards were published.

2.1.1 The Open System Interconnected Model (OSI)

In 1984, the International Standard Organization (ISO) designed a stand for the communication framework for heterogeneous systems in networks, this system is called Open System Interconnection Model (OSI). The OSI reference model provides a framework to break down complex inter-networks into such components that can be more easily understood and utilized.

The purpose of the OSI is to have an easy communication with other computer anywhere in the world, as long as they follow the OSI standard [9].

This OSI reference model is divided into seven levels, and each of these levels in OSI Model has its own working functionality; these levels are remote but on the other hand cascaded to each other and have communication functionality in a proper flow between them. With reference to above standard communication framework, this set of layers known as OSI layers and their functionalities are presented as shown in figure 1 below.

OSI Model			
	Data unit	Layer	Function
Host layers	Data	Application	Network process to application
		Presentation	Data representation, encryption and decryption
		Session	Interhost communication
	segment	Transport	End -end connections and reliability, Flow control
Media layers	packet	Network	Path determination and logical addressing
	frame	Data link	Physical addressing
	Bit	Physical	Media , signal and binary transmission

Figure 2.1 - The OSI Reference Model.

The functionality of each layer (or group of layers) is described in a bit more detail below.

2.1.1.1 Physical Layer

The first layer of a network is the Physical layer, the Physical layer is exactly what its name implies: the physical infrastructure of a network.
This includes the cabling or other transmission medium and the network interface hardware placed inside of computers and other devices which enable them to connect to the transmission medium. The purpose of the physical layer is to take binary information from higher layers, translate it into transmission signal or frequency, transmit the information across the transmission medium, receive this information at the destination, and finally translate it back into binary before passing it up to the higher layers.

2.1.1.2 Data Link Layer

In OSI Reference Model the Data Link Layer is the second layer, it is the layer that is responsible for control methods which provides proper format of data and it can access data flow errors in physical layer. The data format in data link layer is in the form of frames, therefore we say that the data link layer is responsible for defining data formats to include the entity by which information is transported. Error control procedures and other link control procedures may occur in physical layer [10]. Like cyclic redundancy check (CRC); the error checking mechanism that run at the time of transmission of a frame from source side.
The same mechanism will run at the destination side if they found any difference after comparison then receiver makes a request to source to send that frame again. The data link layer is further subdivided into two layers, Logical link Control (LLC) and Media Access

Control. The logical link control is responsible for flow control and error detection in data. Whereas media access control is responsible for controlling the traffic congestion and physical address reorganization.

2.1.1.3 Network Layer

The third layer in OSI Reference Model is the Network Layer, this layer is responsible to make a logical connection between source and destination. The data at this layer is in the form of packets. The network layer protocols provide the following services which is the connection mode and IP Addressing.
Connection mode: The network layer has two types of connection between source and destination, first one is known as connectionless communication which does not provide connection acknowledgement. The example of connectionless communication is Internet Protocol (IP). The second type of connection is connection-oriented which provides connection acknowledgement. TCP is an example of this connection.

IP Addressing: In computer networks every node has its own unique ID. By this unique ID sender and receiver always make right connection. This is because of the functionality of network layer protocol, which has source address and destination address in their header fields. So there is less chance of packet loss, traffic congestion and broadcasting.

2.1.2 Internet Protocol

The IP is in the third network layer of the OSI model that contains addressing information and some control information that enables packets to be routed. IPv4 is documented in RFC [11] as the primary network-layer protocol in the Internet protocol suite alongside with TCP. IP represent the heart of the Internet protocols because it provides a connectionless best- effort delivery of datagram's through an Internetwork service, which means (No logical connection between the user and the network is established prior to data transmission).

The data units are transmitted as independent units. "and providing fragmentation and reassembly of datagram's to support data links with different maximum-transmission unit (MTU) sizes [12]. Because of this feature, IP is robust, however unreliable. An IP packet can be lost, duplicated or arrive out of order. IP was not designed to deal with these problems. It does not provide error recovery or flow control. These functions can be provided by an upper layer (transport layer) connection-oriented protocol, e.g. TCP. Currently there are two versions of the IP protocol, IP version 4 (IPv4), and IP version 6 (IPv6).

2.1.2.1 Internet Protocol Version 4 (IPv4)

IPv4 is one of the protocols that is widely used in the Internet. All communication across the Internet currently relies on IPv4 protocol. In order to understand this protocol in more detail, first we need to look at the address scheme. IPv4 addressing contains four octets and each octet represents 8 bits of a binary number. The entire address space of IPv4 contains 32 bits of binary number, which mean IPv4 has 2^32 addresses that are equivalent to 4,294,967,296 different addresses. According to [13]. IPv4 contains four classes of address, as shown in Table 2.1 below.

Table 2.1 - IPv4 Address Range

Class	High order	Start	End
Class A	0	0.0.0.0	127.255.255.255
Class B	10	128.0.0.0	191.255.255.255
Class C	110	192.0.0.0	233.255.255.255
Class D (Multicast)	1110	224.0.0.0	239.255.255.255
Class E	1111	240.0.0.0	255.255.255.255

In [14], IPv4 address is written in dot decimal notation and it contains three types of address, which include unicast, broadcast, and multicast address.

2.1.2.2 Internet Protocol Version 6 (IPv6)

Ipv6 is the new version of the Internet protocol which was designed to overcome the shortcoming of IPv4. Authors in [15] stated that "*IPv6 was designed to incorporate all of the patches, changes, and best practices developed from over twenty years of IPv4 Internet engineering, into a new next-generation protocol to support the expansive growth of Internet communications and applications*". The development of IPv6 is not just resolving the address space but also provide better performance and improvement over IPv4 [16].

It has been almost two decades that IETF NGtrans had proposed IPv6. According to the Authors in [17]. "*IPv6 had been proposed at IETF as the next generation of Internet Protocol at early in the 1990s and it is now ready for practical use after trial phase*". Both IPv4 and IPv6 have different addressing format, as IPv4 addressing format is written in decimal notation and IPv6 addressing format is written in hexadecimal notation [14].

IPv6 supports unicast, anycast, and multicast address. On the other hand, IPv4 support unicast, anycast, and broadcast address. IPv6 has more efficient forwarding mechanism than

IPv4 due to the 40 bytes fixed header size that allows routers to make faster decisions in forwarding IPv6 packets [18]. There are number of advantages that IPv6 has over IPv4. Next section will discuss in detail the differences between these two protocols.

2.1.2.3 Comparison of IPv4 and IPv6 features

IPv6 packet header has fewer fields when compare to IPv4 header. IPv4 contains fourteen header fields whileIPv6 has eight header fields. However, the size of IPv6 header is double the size of IPv4 header, which means the difference between these two protocols' headers is 20 bytes. This is due to the length of source and destination IPv6 address in IPv6 header field [19].

There are changes in IPv6 header as compared to IPv4 header:

- The Header Length field in IPv4 header is not present in IPv6.
- The type of Service field in IPv4 header changed to Traffic Class and Flow Label field in IPv6.
- The Source address and destination address of IPv4 contains 32 bit long for each field whereas IPv6 contains 128 bit long for each field.
- Time to Live field in IPv4 header changed to Hop Limit field in IPv6.
- The Protocol field in IPv4 header changed to Next Header field in IPv6.
- IPv6 header does not contain Options and Padding fields.

<table>
<tr><th>+</th><th>Bits 0-3</th><th>4-7</th><th>8-15</th><th>16-18</th><th>19-31</th></tr>
<tr><td>0</td><td>Version</td><td>Header length</td><td>Types of Service (now Diffserv and ECN)</td><td colspan="2">Total Length</td></tr>
<tr><td>32</td><td colspan="3">Identification</td><td>Flags</td><td>Fragment Offset</td></tr>
<tr><td>64</td><td colspan="2">Time to live</td><td>Protocol</td><td colspan="2">Header Checksum</td></tr>
<tr><td>96</td><td colspan="5">Source Address (32 bits)</td></tr>
<tr><td>128</td><td colspan="5">Destination Address (32 bits)</td></tr>
<tr><td rowspan="2">160 or 192</td><td colspan="5">Option</td></tr>
<tr><td colspan="5">Data</td></tr>
</table>

Figure 2.2 - Internet Protocol version 4 header.

The IPv4 header fields are described in the following list [20]:

- Version: The first header field in an IP packet is the 4-bit version field. For IPv4, this has a value of 4 (hence the name IPv4).
- Internet header length: The second field is a 4-bit Internet Header Length (IHL) telling the number of 32-bit words in the header. Since an IPv4 header may contain a variable number of options, this field specifies the size of the header (this also coincides with the offset to the data). The minimum value for this field is 5 (rfc791), which is a length of 5×32 = 160 bits. Being a 4-bit field the maximum length is 15 words or 480 bits.
- Type of service: This 8 bit field specifies the datagram's precedence (importance), delay, throughput and reliability.
- Total length: Specifies the total length of the datagram (in octets), including the header. Since this field is 16 bits in length, a datagram length of up to 65536 octets can be specified.
- Identification: Each datagram assembled receives a unique identification number. If the datagram becomes fragmented, this identification number is used to reassemble the datagram when it is received.
- Flags: The flags are the next 3 bits in the datagram. The first is unused. The next is the DF (Don't Fragment) flag. If this is set to 1, then the datagram cannot be fragmented. If the IP layer cannot send datagrams across the network without fragmenting, and the DF flag is set, then no datagrams can be sent. The next flag is MF (More Fragments), and specifies that the current datagram is part of a fragmented message and that more fragments are to follow. If MF is set to 0, then this is the last fragment in the message.
- Fragment offset: When MF is set, this field indicates the position of the current fragment relative to the starting fragment, and thereby allows reassembly.
- Time to live: The TTL specifies how long a datagram can remain on the network. It is usually set to 15 or 30. Whenever a datagram passes through a host/router, the TTL is decreased by 1. If a datagram reaches 0, the current node discards it and sends a message back to the originator so that it can resend. This process ensures that gateways do not become bottlenecked, and ensures that datagrams do not travel forever if a network path contains a loop.
- Protocol: This field contains a code representing the transport protocol of the segment passed to the IP layer. In turn, at the receiving end, this field indicates which upper layer protocol is to receive the data portion of the IP datagram. Common values are 1 for ICMP, 6 for TCP and 17 for UDP.
- Header checksum: A form of CRC, the checksum is calculated using a quick algorithm, using data in the IP header only. Because the TTL value is decreased at every node

the datagram passes through, the checksum must also be recalculated at each stage. This checksum gives some protection against corruption.

- IP addresses: The 32 bit source and destination IP addresses. Durr.
- Options: Additional header fields (called options) may follow the destination address field, but these are not often used. Note that the value in the IHL field must include enough extra 32-bit words to hold all the options (plus any padding needed to ensure that the header contains an integral number of 32-bit words).

<table>
<tr><th></th><th colspan="2">0-4</th><th>8</th><th>12</th><th>16</th><th>20</th><th>24</th><th>28</th><th>32</th></tr>
<tr><td>0</td><td colspan="2">Version (4)</td><td colspan="2">Traffic Class (8)</td><td colspan="5">Flow label (20)</td></tr>
<tr><td>64</td><td colspan="4">Pay load length (16)</td><td colspan="2">Next header (8)</td><td colspan="3">Hop limit (8)</td></tr>
<tr><td>128</td><td colspan="9">Source Address (128 bits)</td></tr>
<tr><td>192
256</td><td colspan="9">Destination Address (128 bits)</td></tr>
</table>

Figure 2.3 - Internet Protocol version 6 header.

The IPv6 header fields are described in the following list [21]:

- Version – 4-bit Version number of Internet Protocol = 6.
- Traffic Class – 8-bit traffic class field. See Traffic Class.
- Flow Label – 20-bit field. See IPv6 Quality-of-Service Capabilities.
- Payload Length – 16-bit unsigned integer, which is the rest of the packet that follows the IPv6 header, in octets.
- Next Header – 8-bit selector. Identifies the type of header that immediately follows the IPv6 header. Uses the same values as the IPv4 protocol field. See Extension Headers.
- Hop Limit – 8-bit unsigned integer. Decremented by one by each node that forwards the packet. The packet is discarded if Hop Limit is decremented to zero.
- Source Address – 128 bits. The address of the initial sender of the packet. See IPv6 Addressing.
- Destination Address – 128 bits. The address of the intended recipient of the packet. The intended recipient is not necessarily the recipient if an optional Routing Header is present.

2.1.3 Transport Layer

The fourth layer in OSI reference model is Transport Layer. It contains two types of protocols, first is Transport Control Protocol (TCP) which is connection oriented protocol and supports some upper layer protocols like Hypertext Transfer Protocol (HTTP) and Simple Mail Transfer Protocol (SMTP).
The second is User Datagram Protocol (UDP) which is a connection less protocol. Like TCP it also supports some upper layer protocols such as Domain Name System (DNS) and file transfer protocol (FTP). The Transport layers is responsible for the reliability of the link between two end users and for dividing the data that is being transmitted by assigning port numbers to its layer 4 packages, known as segments. The main thing in transport layer protocols is that they have port addresses in their header fields.

2.1.3.1 Transmission Control Protocol (TCP)

TCP is designed to provide a connection oriented ordered reliable byte stream on top of the connectionless unreliable IP [22]. It was also designed to run above IP, providing reliable data transmission with flow control. TCP is a connection-oriented protocol, which means "*A user and network set up a logical connection before transfer of data occurs. Usually, some type of relationship is maintained between the successive data units being transferred through the user/network connection.*" [12]. TCP uses sequence numbers and checksum facilities to ensure that a segment of data is not damaged during the transmission. It also allows retransmission by sending acknowledgement message back to the sender, when the segment is received correctly, a positive acknowledgement (ACK) is returned to the sender, otherwise, a negative acknowledgement (NACK) is returned; in this case, the sender would retransmit the data. In addition, TCP also uses the sequence numbers to deliver the segments in order even if the segments arrive over the network out of order, TCP also checks for the duplication.

Another useful feature provided by TCP is flow-control. It is based on the "sliding-window" technique. A window size value is assigned to the transmitter. The transmitter is only allowed to transmit a specified number of bytes within this window. On receiving of the correct ACKs, the window slides forward. The transmitter must stop the transmission when the window is closed. Another point to mention is the port number. Each application process needs to identity itself by a port number, which is used to identity which application program should receive the incoming traffic. Since the port number allows several programs to communicate concurrently, it can be used to support multiplexing capabilities.

+	Bits 0-3	4-7	8-15	16-31
0	Source Port			Destination Port
32	Sequence Number			
64	Acknowledgment			
96	Data Offset	Reserve	Flags	Window
128	Checksum			Urgent Pointer
160	Potions (optional)			
160/192	Data			

Figure 2.4 - Transmission Control Protocol Header.

The following descriptions summarize the TCP header fields illustrated in Figure 4 [20]:

- Source Port and Destination Port: Identifies points at which upper-layer source and destination processes receive TCP services.
- Sequence Number: Specifies the number assigned to the first byte of data in the current message. In the connection-establishment phase, this field also can be used to identify an initial sequence number to be used in an upcoming transmission.
- Acknowledgment Number: Contains the sequence number of the next byte of data the sender of the packet expects to receive.
- Data Offset: This 4-bit field specifies the size of the TCP header in 32-bit words. The minimum size header is 5 words and the maximum is 15 words thus giving the minimum size of 20 bytes and maximum of 60 bytes. This field gets its name from the fact that it is also the offset from the start of the TCP packet to the data.
- Reserved: Remains reserved for future use.
- Flags: Carries a variety of control information, including the SYN and ACK bits used for connection establishment, and the FIN bit used for connection termination.
- Window: Specifies the size of the sender's receive window (that is, the buffer space available for incoming data).
- Checksum: Indicates whether the header was damaged in transit.
- Urgent Pointer: Points to the first urgent data byte in the packet.
- Options: Specifies various TCP options.
- Data: Contains upper-layer information.

2.1.3.2 User Datagram Protocol (UDP)

UDP is a simple datagram- oriented transport layer protocol. Each output operation by a process produces exactly one UDP datagram, which causes one IP datagram to be sent. This is different from a stream-oriented protocol such as TCP where the amount of data written by an application may have little relationship to what actually gets sent in a single IP datagram. RFC 768 [23] is the official specification of UDP.

In addition, UDP is functionally at transport layer protocol. It is connectionless, and does not provide a reliable transport. On the other hand, it gives an application a direct access to the datagram service of the IP layer. The multicast and broadcast services are available by using UDP. The UDP header contains the source port number, destination port number, total length and checksum. (Cisco Networking Academy Program, 2th edition. Cisco Press.2001) [24]. Both UDP and TCP have checksums in their headers to cover their header and their data. Unlike the TCP, UDP adds no reliability, flow-control, or error-recovery functions to IP.

UDP headers contain fewer bytes and consume less network overhead than TCP Because of its simplicity. UDP is useful in situations where the reliability mechanisms of TCP are not necessary, such as in cases where a higher-layer protocol might provide error and flow control. UDP is the transport protocol for several well-known application-layer protocols, including Network File System (NFS), Simple Network Management Protocol (SNMP), Domain Name System (DNS),and Trivial File Transfer Protocol (TFTP).

+	Bits 0-15	16-31
0	Source Port	Destination Port
32	Length	Checksum
64	Data	

Figure 2.5 - User Datagram Protocol Header.

The UDP header format contains four fields, as shown in Figure 5 These include source and destination ports, length, and checksum fields [20]:

- Source port is the field that identifies the sending port when important and should be assumed to be the port to reply to if needed. If not used, then it should be zero.
- Destination port identifies the destination port and is required.

- Length contains 16-bit field that specifies the length in bytes of the entire datagram: header and data. The minimum length is 8 bytes since that's the length of the header. The field size sets a theoretical limit of 65,527 bytes for the data carried by a single UDP datagram.
- Checksum is the 16-bit checksum field that is used for error-checking of the header and data.

2.1.3.3 UDP and TCP Performance

The User Datagram Protocol (UDP) and Transmission Control Protocol (TCP) are the “siblings” of the transport layer in the TCP/IP protocol suite. They perform the same role, providing an interface between applications and the data-moving capabilities of the Internet Protocol (IP), but they do it in very different ways. The two protocols thus provide choice to higher-layer protocols, allowing each to select the appropriate one depending on its needs [25].
Below is the table which helps illustrate the most important basic attributes of both protocols and how they contrast with each other:

Table 2.2 - UDP and TCP comparative features.

Characteristic/ Description	UDP	TCP
General Description	Simple high speed low functionality 1"wrapper" that Interface application to the network layer and does little else	Full-featured Protocol that allows application to send data reliably without worrying about network layer issues
Protocol Connection setup	Connectionless data is sent without setup	Connection- Oriented; connection must be Established prior to transmission
Data Interface to Application	Message base- based is sent in discrete package by the application	Stream-based, data is sent by the application with no particular structure.
Reliability	There is no guarantee that the packet or message sent would reach at all	There is absolute guarantee that the data transfer remain intact and arrived at the same order in which it was sent
Retransmissions	Not performed. Application detect lost data and retransmit if need.	Delivery of all data is managed and lost data is retransmitted automatically
Features Provided to Manage Flow of Data	None	The flow control is using sliding windows; windows size adjustment heuristics; congestion avoidance algorithms
Overhead	Very low	Low, but higher than UDP
Transmission Speed	Very high because there is no error checking of packets	The speed of TCP is slower than UDP
Data Quantity Suitability	Small to moderate amounts	Small to very large amount of data

2.1.4 Session Layer

The fifth layer in OSI Reference Model is Session Layer. The Session Layer is responsible for session management i.e. start and end of sessions between end-user applications [26]. It is used in applications like live TV, video conferencing, VoIP etc, in which sender establishes multiple sessions with receiver before sending the data. Session Initiation protocols (SIP) is an example.

2.1.4.1 Presentation Layer

The sixth layer in OSI Reference Model is Presentation Layer. This layer is responsible for presentation of transmitted/received data in graphical mode. Data compression and decompression is the main functionality of this layer. The data encryption is done before transmission in presentation layer.

2.1.4.2 Application Layer

The last layer of OSI Reference Model is Application Layer. This layer organizes all system level applications like DNS, HTTP, Post Office Protocol (POP3), SMTP, Secure shell (SSH), Telnet, E-mail services etc.
The World Wide Web supports two well-known transport protocols which is, HTTP [27] and HTTPS [28]. These two protocols have different costs and provide different security guarantees for the web applications deployed on top of them. At one end, HTTP is inexpensive to use but provides no security guarantees for any web application deployed on top of it. While at the other end, HTTPS is expensive to use but provides three important security guarantees for any web application deployed on top of it. These three security guarantees are (1) server authentication (2) message integrity (3) message confidentiality [29].

In most cases, HTTPS is also augmented with a password protocol in order to provide the added guarantee of client authentication. (Note that HTTP cannot be easily augmented with a password protocol to provide client authentication).
There are two primary differences between an HTTPS and an HTTP connection
HTTPS connects on port 443, while HTTP is on port 80, HTTPS encrypts the data sent and received with SSL, while HTTP sends it all as plain text.

2.1.4.3 HTTP Protocol

The Hypertext Transfer Protocol (HTTP) is a protocol to transmit data at the application layer between hosts. The protocol was designed in 1989 by Tim Berners-Lee at CERN in combination with the Uniform Resource Locator (URL) and the Hypertext Markup Language (HTML). It is a communication scheme to transmit data units which are parts of websites in the WWW and defined in the RFC 2616 [29]. HTTP is a stateless protocol allowing asynchronous connections between client and server. It needs a reliable connection to transmit data. Mostly TCP is used for this purpose although it can run on other reliable protocols too. Up to now there exist two different versions of HTTP: 1.0 and 1.1, the later is a down-compatible extension of the previous version. In HTTP 1.0 the client/server relation can only set up separate connections for every request. In this case a client creates a new TCP connection for each object request separately. The TCP protocol is not optimized for this kind of data transfer.

In fact the slow start mechanism will harm the performance of such protocols. As one of the main advantages HTTP 1.1 offers so called persistent connections. In such a connection the client can request multiple objects at once. Hence, there is only one open TCP connection for the whole webpage. The drawback to this method is the fact that there are more open connections that must be handled by the servers. The second performance improvement in HTTP 1.1. is pipelining. This feature enables the client to request multiple objects without waiting for the response from the server. In combination with the persistent connection this feature fills the available resources much more efficiently. In general the HTTP header may hold optional information not standardized, which allows special applications to implement modified data communications.

2.1.4.4 HTTPS Protocol

Hypertext Transfer Protocol Secure (HTTPS) is a widely used communications protocol for secure communication over a computer network, with especially wide deployment on the Internet. Technically, it is not a protocol in itself rather, it is the result of simply layering the Hypertext Transfer Protocol (HTTP) on top of the Secure Sockets Layer/ Transport Layer Security SSL/TLS protocol, thus adding the security capabilities of SSL/TLS to standard HTTP communications [30].
In its popular deployment on the Internet, HTTPS provides authentication of the web site and associated web server that one is communicating with, which protects against Man-in-the-middle attacks. Additionally, it provides bidirectional encryption of communications between a client and server, which protects against eavesdropping and tampering with and/or forging the contents of the communication [31]. In practice, this provides a reasonable guarantee that one is communicating with precisely the web site that one intended to communicate

with (as opposed to an imposter), as well as ensuring that the contents of communications between the user and site cannot be read or forged by any third party.
In the past, HTTPS connections were primarily used for payment transactions on the World Wide Web, e-mail and for sensitive transactions in corporate information systems. In the late 2000s and early 2010s, HTTPS began to see widespread use for protecting page authenticity on all types of websites, securing accounts and keeping user communications, identity and web browsing private.

2.1.5 Network security

Network security refers to any activities designed to protect your network. It consist of the technologies and processes that are deployed to protect network from internal and external threats.
Network security involves all activities that organizations, enterprises, and institutions undertake to protect the valve and ongoing usability of assets and the integrity and continuity of operations. Effective network security targets a variety of threats and stops them from entering or spreading on the network.

System and network technology is a means technology for a wide variety of applications. Security is essential to networks and applications. Although, network security is a vital requirement in emerging networks, there is an important lack of security methods that can be easily implemented. There exists a “communication gap” between the developers of security technology and developers of networks.
Network design is a well-developed process that is based on the Open Systems Interface (OSI) model. The OSI model has several advantages when designing networks. It offers modularity, flexibility, ease-of-use, and standardization of protocols. The protocols of different layers can be easily combined to create stacks which allow modular development.
The implementation of individual layers can be changed later without making other adjustments, allowing flexibility in development. In contrast to network design, secure network design is not a well-developed process. There isn’t a methodology to manage the complexity of security requirements. Secure network design does not contain the same advantages as network design.

When considering network security, it must be emphasized that the whole network is secure. Network security does not only concern the security in the computers at each end of the communication chain. When transmitting data the communication channel should not be vulnerable to attack. A possible hacker could target the communication channel, obtain the data, decrypt it and re-insert a false message. Securing the network is just as important as securing the computers and encrypting the message.

When developing a secure network, the following need to be considered [57]:

- Access – authorized users are provided the means to communicate to and from a particular network
- Confidentiality – Information in the network remains private
- Authentication – Ensure the users of the network are who they say they are
- Integrity – Ensure the message has not been modified in transit
- Non-repudiation – Ensure the user does not refute that he used the network

2.1.5.1 Importance of Network Security

Network security is important for a variety of reasons. First of all, it is important to ensure the company's reputation will not be marked by a security breach leaking customers' information. Large, small, known and unknown companies are all at risk to an attack led by a hacker. One security breach and the reputation of the company can immediately take a turn for the worse.

Once a company is educated about their network's strengths and weaknesses, they will gain a better understanding of areas they may be at risk to an attack and be able to take appropriate measures to pinpoint areas where security needs to be reinforced. Network security helps to protect the networks and the network-accessible resources from unauthorized access, and consistent and continuous monitoring and measurement of its effectiveness combined together. The primary goal of network security is to provide controls at all points along the network perimeter which allows access to the network and only let traffic pass if that is authorized, valid and of acceptable risk.

The purpose of network security is to protect networks, network devices and network messages from unauthorized access, usually by outsiders:

- To provide control at all points along the network perimeter in order to block network traffic that is malicious, Unauthorized or that otherwise presents risk to the network.
- To detect and respond to attempted and actual intrusions through the network.
- To prevent network messages that is sent across networks from being intercepted or modified.

Network security controls cannot completely eliminate risk. The goal is to minimize risk as much as possible and to avoid unnecessary or excessive risk. The goal of network security is really to 'enable' network connectivity. Without network security, the risks/costs of network connectivity would be very expensive.

2.1.5.2 Network Security Threats

Security threat is a condition or event with potential to harm network resources in the form of destruction, disclosure, fraud etc. Network security threats include impersonation, eavesdropping, denial-of-service, packet replay and packet modification. Security threat can be categorized into four parts and these categories are the ways or forms through which threats can be carried out on a network. These threats can be categorized as unstructured versus structured, and external versus internal:

- **Unstructured Threats**: Unstructured security threat is the kind of threat created by an inexperienced person trying to gain access to a network. They commonly use common hacking tools, like shell scripts, and password crackers. A good security solution should easily thwart this kind of attack. In other words, these kinds of hackers could not be underestimated because they can cause serious damage to network.
- **Structured Threats**: Unlike unstructured threats, structured threat hackers are well experienced and highly sophisticated. They use sophisticated hacking tools to penetrate networks and they can break into government or business computers to extract information. On certain occasions, structured threats are carried out by organized criminal gangs or industry competitors.

2.1.5.3 External Threats

External threats can arise from individual or organization working outside of a company who do not have authorized access to the computer systems or network. They work their way into a network mainly from the Internet or dialup access servers. External threats can vary in severity depending on the expertise of the attacker, Both experienced and inexperienced hackers could pose external threats.

2.1.5.4 Internal Threats

An internal security threats occurs when someone from inside your network creates a security threats to your network. Interestingly, the CSI (Computer Security Institute) study has found that, of the 70 percent of the companies that had security breaches, 60 percent of these breaches come from internal sources. Like external threats, the damage that could be caused by such a hacker depends on the expertise of the hacker. (Orbit-Computer Solutions 2012).

2.2 Network Traffic

With the increasing knowledge in Internet applications and the need to transmit and receive information in a timely, secured and accurate manner, the need to study the network traffic and analysis for effective and efficient decision making by its users, has become an interesting research area.

This scientific approach involves collection and gathering of network traffic data over a period of years, processing, analysing and interpreting them into a useful information for effective and efficient decision making. The moment the data are collected from a particular point on your network for a period of time, the real fun begins that is performing traffic analysis on the data.

The methodology adopted varies from place to place. Different researchers tend to adapt different approaches depending on the environments and the polices governing the place. However, the general guide is that if you permit everything that isn't explicitly denied, then you should look for those items that are explicitly denied. If you deny everything that isn't explicitly permitted, then you'll need to look for those items that aren't explicitly permitted.

It is pertinent to point out that in many environments, no single person will know what activity is really unauthorized, particularly on a server-by-server or host-by-host basis. In which cases, there is need to consult the network polices governing the environment. This stage of traffic polices and data pre-processing lead us to the statistical analysis of these data which is known as network traffic monitoring and analysis.

2.2.1 Network Monitoring

This is the process to monitor a computer network to prevent that the network goes too slow or to look out for failing systems, including notifying the network administrator via email, pager or other alarms [58].

Network monitoring is the use of grouping and analysis tools to accurately determine traffic flows, utilization, and other performance indicators on a network [59]. A good monitoring tools gives you both hard numbers and graphical aggregate representations of the state of the network. This helps the network administrator to visualize precisely what is happening in the network, so as to know where adjustments may be needed.

These tools can help answer critical questions, such as:

- What are the most popular services used on the network?
- Who are the heaviest network users?
- At what time of the day is the network most utilized?
- What sites do your users frequent?
- Is the amount of inbound or outbound traffic close to our available network capacity?

- Are there indications of an unusual network situation that is consuming bandwidth or causing other problems?
- Is our Internet Service Provider (ISP) providing the level of service that we are paying for?
- This should be answered in terms of available bandwidth, packet loss, latency, and overall availability.

And perhaps the most important question of all:

- Do the observed traffic patterns fit our network planning and expectations?

There are two types of Network monitoring techniques, active monitoring and passive monitoring. Active monitoring reduce system overhead by using small number of probe packets that have smaller sizes compare to real data packet so that performance measures may not be accurate. While Passive monitoring monitors a lot of data packets, it has system overhead problem so that its performance is more accurate and reliable than active monitoring [60].

2.2.2 Traffic Flow

The environment of Internet traffic can better be understood by knowing the concept of the flow. Traffic flow or network flow is a sequence of packets from a source computer to a destination, which may be another host, a multicast group, or a broadcast domain. RFC 2722 [61] defines traffic flow as "an unreal logical equivalent to a call or connection. RFC 3697 [62] defines traffic flow as "a sequence of packets sent from a particular source to a particular uncast, any cast, or multicast destination that the source desires to label as a flow. A flow could consist of all packets in a particular transport connection or a media stream. Flow is also defined in RFC 3917 [63] as a set of IP packets passing an observation point in the network during a certain time interval.

Alternatively, the definition of flow may also be coined as, a series of packets that share the same *source IP*, *destination IP*, *source port*, *destination port* and the *protocol*. This is most commonly known as *five-tuple* IP flow, which is an aggregation of individual flows.
Network flow data symbolizes a summary of sessions between two end hosts. It further aids in network analysis and security issues. Flow data is autonomous of packet payloads. The flow tool or analyzer is dependent on the amount of information collected from packet headers and its important metrics. In addition, the network flow data deeply enhances the visualization of discrete network events such as protocol analysis or length distribution without the need for payload analysis.

The knowledge of flow data aids in understanding how different flows compete in a network to acquire network resources. Packets having similar *five-tuple* information belong to the same flow. The most significant thing to remember is that a network flow can be considered either as unidirectional flow or bidirectional flow. In unidirectional flow, the flow attribute is categorized in one direction i.e. from source to destination or vice versa. Whereas, in a bidirectional flow, the attributes are categorized considering both directions.

2.2.3 Traffic Profiling

As the Internet continues to grow in size and complexity, the challenge of effectively provisioning, managing and securing it has become inextricably linked to a deep understanding of Internet traffic. Although there has been significant progress in instrumenting data collection systems for high speed networks at the core of the Internet, developing a comprehensive understanding of the collected data remains a daunting task. This is due to the vast quantities of data, and the wide diversity of end-hosts, applications and services found in Internet traffic. Because of these challenges that will encounter every day in the Internet, there is need for us to use traffic profiling in our network plan.
Traffic profiling is the ability to look at the network traffic and identify potential security risks. Today network profiling should include not only the local area network traffic but also wireless traffic as well as any traffic that flowing through the routers and firewalls.

2.2.4 Traffic Analysis

Traffic analysis is the science of extracting information from metadata, or otherwise known as traffic data, produced by a communication. These include the routing data, length and timing of the communication stream. Recent work in this area includes using timing information to reduce the entropy of passwords sent using SSH [64] and guessing if a particular web page is already locally cached by a user [65]. Research into anonymous communication has also provided some insights about how the shape of traffic contained in a channel can be used to trace the communication. The onion routing project [66] presented strong evidence for the need to use dummy cover traffic in order to hide these patterns. Traffic analysis of HTTP transactions through anonymizing proxies has been mentioned in [67] and [68]. In [67] Hintz analyzes traffic packet lengths at the TCP level, in order to attack the SafeWeb [69] service.

Traffic analysis can be used to extract a variety of information. It can be used for identification, when the information extracted is used to find out who the sender of some data is, or which particular network card is active. It can also be used for profiling when the aim of the analysis is to extract some information about the target, such as their type or

status. Finally traffic analysis can be used for information extraction when the objective of the analysis is to extract some of the information contained in a particular conversation.

2.3 Internet traffic measurement

Researches have been ongoing for the past decade on the field of web and Internet traffic protocols and related topics. In recent years, several research works have been proposed due to the dramatic increase in applications that leverage its performance on the Internet. These section presents the evolution of Internet traffic measurements to one of today's most recent and relevant proposals in this topic.

2.3.1 The Evolution of Internet Traffic Measurement

The Internet did not initially employ any native comprehensive measurement mechanism, mainly due to its own decentralized and layered design which facilitated transmission of data between end-points without needing any visibility into the details of the underlying network.
This lack of detailed measurement capabilities was also reinforced by the Internet best-effort service model that offers no hard performance guarantees to which conformance needs to be measured [32]. However, the need to gain visibility into the Internet's internal behaviour has become increasingly imperative for a number of different beneficiaries, including network operators and administrators, researchers and service providers. The people who actually run the network initially needed to be able to detect traffic anomalies and infrastructure failures. Hence some inspired diagnostic tools started being developed as the Internet was growing larger.
The Internet has been continually evolves in scope and complexity, hence creating a complex task to characterize, understand, control, or predict the network behaviour. The field of Internet traffic analysis research includes many research works in the literature, representing various attempts to classify whatever traffic samples a given researcher has to look at, with no systematic integration of results. Due to these reasons researchers have started investigating the behaviour and usage patterns of computer networks in order to create realistic models of the traffic sources. These efforts have led to the emergence of new research themes dealing with measurement methodologies, inferences and statistical analyses of the Internet traffic characteristics with the aims of improving its performance. More recently, Internet service providers have started considering the provision of quality of services (QoS) through quality of experience (QoE) in order to increase revenue by implementing non-flat-rate usage pricing within the research community. [33], [34].
The research work of [35] played an important role in the advancement of Internet, not only to the empirical characterization of end-to-end Internet routing behaviour and packet dynamics, but also to the actual birth and subsequent tremendous popularity of inter-network measurements. The research work deployed large number of Internet sites and used TCP and

route information, to assess the traffic dynamics of the dominant transport protocol. In addition, the routing behaviour across a representative number of geographically-spread end-to-end Internet paths were also investigated. Using a significant number of traces, the work empirically examined among other routing pathologies such as packet delay and loss, as well as bandwidth bottlenecks across the Internet. The paper concluded that, with due diligence to remove packet filter errors and TCP effects, TCP-based measurement would provide a viable means for assessing end-to-end packet dynamics.

Sporadic studies on local and wide area network traffic measurements can be traced back to the beginning of 1980's, yet it was the second half of the same decade when a considerable number of highly-cited studies focused on monitoring operational network traffic and characterizing several aspects of its aggregate behaviour. The effects of LAN traffic was examined by [36]. The research analyzed an eight-week LAN traffic monitoring and concluded that packet arrivals on the Ethernet are not adequately described by the often-assumed Poisson model. The low bit error rate experienced, the bursty nature of the network load, and the strong locality properties of the LAN traffic were also observed. However, in order to address these issues, [37] proposed a new model for packet arrival processes based on the concept of packet trains due to the observation that packet inter-arrival times on a ring LAN topology were not exponentially distributed.

Wide-area traffic monitoring study on the 56 Kb/s link that connected the Bell Labs corporate network to the Internet was carried out in [38], the research presented packet and byte count statistics, protocol decomposition, and length frequencies for TCP and UDP wide-area traffic. A later more comprehensive yet similar study by [39] examined characterized bulk transfer and interactive wide-area network traffic. The paper concluded that the characterized bulk transfer approach was dominance in traffic. The calculation of packet size and connection duration distributions, inter-packet latencies, as well as sizes of packet bursts among other statistics were carried out in [40] through the analysis of the characteristics of operational traffic captured during a 5-hour interval on the UK-US academic 384 Kb/s network link. The paper concluded that appropriate mechanisms to continuously assess and monitor the network's traffic-perceived performance, is vital for such a global communications medium to be able to offer consistently predictable performance characteristics.

Leland *et al*. used long traces of captured Ethernet LAN traffic to characterize its nature as statistically self-similar, and hence very different from conventional telephone traffic and from commonly considered formal models for packet traffic, such as Poisson-related, packet train, and fluid flow models [41].
Observations on the patterns and characteristics of wide-area Internet traffic was presented in [42]. They discovered that Internet is rapidly growing in number of users, traffic levels, and topological complexity. This discovery was as a result of increasingly driven economic

competition. The Authors in [43] investigated more than 4000 traces from 1998 to 2003 to find the relations between bit rates and traffic statistics. Traffic pattern is a very clear and typical way to display traffic variation within the recording period as observed in [42]. They made experiments to reveal the traffic characteristics in terms of packet sizes, flow duration and volume over the two time scales, one day and seven days. The researchers in [44] reviewed 10 years development of Lang-Range Dependence (LPD) theory and used LPD to model the complex traffic of the Internet. However, more researchers believed that, instead of bit and packet rate approach, flow level traffic would be better used for explaining the intrinsic characteristics of Internet.

Analysis of IP traffic workload at a single measurement site at NASA Ames Internet exchange point (AIX) from May 1999 through March 2000 was examined in [45]. They observed that there is no significant change in the overall packet size distribution, nor in the ratio of TCP to UDP traffic during this period, but the proportion of fragmented traffic was on the rise.

Borgnat *et al* researched on sketching the evolution of Internet Traffic in [46], by collecting internet traffic data for seven years in order to analyse the evolution and trend of the internet traffic, both at the TCP/IP layers (packet and flow attributes) and application usage. The paper proposed a random projection (sketch) based analysis procedures, which provide practitioners with an efficient and robust tool to disentangle actual long term evolutions, from time localized events such as anomalies and link congestions.

Barakat *et al*. analyzed TCP flow by means of Markovian model in a differentiated service network [47]. They also established a Poisson Shot-noise model in flow level. As a matter of fact, modeling the traffic at the packet level has proven to be very difficult [48], because traffic on a link is the result of a high level of multiplexing of numerous flows whose behaviour is strongly influenced by the transport protocol and by the application. It is not easy to judge which model is more ideal for the Internet, it all depends on which application the model is used. For example, detection of anomalies (e.g. denial of service, link failure) require an accurate traffic model. While in a protocol and application agnostic environment, a more general model is needed.

Fraleigh *et al*. [49] describe the IPMON traffic monitoring system and report observations of traffic on OC-12 links in the Sprint E-Solutions backbone network over a 24-hour period in the middle of 2001. This is the first published traffic study from OC-12 links in a commercial backbone network. They found that on some links over 60% of the traffic is generated by new applications such as distributed file sharing and streaming media, while only 30% is web traffic. The results of this paper provide a snapshot of traffic characteristics typical for the Sprint IP backbone, but it is unclear if they are representative of other networks.

The WAND network research group of the University of Waikato conducts bidirectional measurements on the OC3 access link that connects the University of Auckland to the public Internet [50]. Since July 1999 they have collected several comprehensive data sets spanning

periods from one week up to seven months. The data are publicly available and have been used in a number of studies (see review in [50]).

Some researchers would like to analyze the Internet from the view of application. Back in the nineties, when FTP and Mail accounted for half of the traffic volume, until HTTP becomes the majority [51]. And the invention of Peer-to-Peer (P2P) nearly toppled the pattern of Internet traffic, it could be considered as killer Internet application [52]. The Internet service providers, on the other hand, are reluctant to see this change as P2P consumes huge amount of bandwidth resource. And they react, inclining to interfere their customers" file sharing [53]. But the technologies seem to keep in pace, while modern P2P application uses random port numbers, making itself hard to be detected from authorities and Internet service providers who have the illegal P2P file-sharing concern [54], which may cause inaccurate P2P traffic measurement.

Fukuda and his team collected month-long aggregated traffic logs for seven major ISPs in Japan, in order to analyze the macro-level impact of residential broadband traffic [55]. They have an advantage of keeping a large dataset which covers 41% of the total customers in Japan. The collecting method for traffic logs is to use MRTG (a tool to monitor the traffic load on network-links) or RRD tool (an open source tool for storage and retrieval of time series data) which are usually providing aggregated traffic information. And they have reached several conclusions in their report. For example, about 30% of the daily traffic volume is promised, while the rest 70% is a fluctuation pattern with peak in the evening hours, which is much larger than that in campus or office networks. The residential traffic accounts two-thirds of the ISP backbone traffic, which means that backbone traffic is dominated by the residence behaviour.

2.3.2 Cost Action IC0703

The issue of Internet Traffic Monitoring research has been increasing, due to the number of technologies and new applications that have been developed in the past few years, to increase our understanding of the behaviour of network traffic [56]. This is can be done by the accessibility of hardware and large storage solutions at accessible cost. Presently, there are numerous research groups that is involved in the development monitoring tools and methods to acquire, analyze and interpret traffic data from the live networks. These developments have spread to different network environments such as wired access, broadband backbone, campus WLAN, 3G cellular WAN, etc. The term used to describe such research activities is called Traffic Monitoring and Analysis (TMA).

Traffic Monitoring and Analysis (TMA), is a research group that is working in the field of Traffic Monitoring and Analysis. It was developed by COST Action IC0703, its functions are

to serve the research community at large, in the areas of large-scale performance monitoring, network validation and troubleshooting, detection of weaknesses and failures inside complex infrastructures and ultimately ensuring a higher level of network robustness.
Furthermore, it is an essential research area within the field of Communication Networks, that involves several research groups around the globe that are collectively advancing our understanding of network traffic monitoring, real packet networks and their users. Since modern packet networks are highly complex and ever-evolving objects. Understanding, developing and managing such systems is difficult and expensive in practice. This is the reason why TMA techniques play an important role in the operation of the network environment.

Apart from its practical importance, it is an intellectually attractive research field due to the following reasons. First, the inherent complexity of the Internet has attracted many researchers to face traffic measurements since the pioneering times. Second, TMA offers a fertile ground for theoretical and cross-disciplinary research such as the various analysis techniques being imported into TMA from other fields while at the same time providing a clear perspective for the exploitation of the results in a real environment.
In other words, TMA research has an intrinsic potential to reconcile theoretical investigations with practical applications, and to realign curiosity-driven with problem-driven research.

The COST Action IC0703 began on March 2008, with a duration plan for 4 years. It is part of the COST program (ICT domain) which is now run by the European Science Foundation. It is an intergovernmental framework for European Cooperation in Science and Technology, It has involved more than 50 research groups from 26 different countries, that is promoting the coordination of nationally funded research on an European level.
Each COST Action aims at improving the coordination and exchange between European researchers involved in a particular field or cross-disciplinary topic, and helps to open European Research to cooperation worldwide. Although the COST TMA Action program has finished since March 2012, the portal will still remain active for the years to come.
The primary goal of the TMA Action was to establish of a recognizable community out of a set of research groups and individuals who are working across Europe in the field of Internet traffic monitoring and network measurements. This goal has been largely achieved, and the TMA community that has formed around the Action is now a recognized entity, inside and outside Europe, even after the formal termination of the Action.

Objectives of the Cost Action IC0703
The COST Action have two main objectives which is the primary and secondary objectives:

- **Primary objectives of IC0703**

The primary objectives of the COST Action (CA) is to boost the quality and the impact of European research in the field of Traffic Monitoring and Analysis (TMA).

The first goal is to give grounds to the European research agenda in the field, by promoting realistic coordination among the different research groups and sharing of operational know-how (lessons-learned, problems found during practical deployment, ideas for real-world exploitation of TMA techniques, etc.).

The second action is foster the ability of making systems and organizations to work together in the area of monitoring tools, data formats and analysis modules developed by the researchers.

This action will promote the adoption and improvement of existing shared databases of information sources and traffic traces (e.g. MOME), that could be used to compare the dataset when developing and testing novel algorithms/tools/modules.

- **Secondary objectives of IC0703**

  The secondary objectives of the COST Action is to become the central reference point for the European research in the field of TMA. This will help to the launch of collaborations and joint activities with non- European entities in the field, e.g. CAIDA in US, and the research communities of other fields.

These objectives will be achieved by implement the following instruments:

- To create a TMA portal and associated electronic collaborations tools for targeting external audience.
- To organize regular management meetings every year, and addressing particular importance to technical discussions and presentations, to achieve this goal they need invite different representatives from EU projects and externals experts to participate regularly and contribute to the technical discussion.
- By organize workshops yearly where they will invite guest speakers and lecturers from other scientific fields.
- By releasing a series of Annual Reports on the survey of state-of-art of TMA research and level of real-world applications.
- To organize seminars and summer schools for young researchers.
- To organize short visits for senior participants that will last for a month and long-term visits for young researchers for 3 months to be due on a competitive basis.

2.4 Internet Applications

With rapid increase in research and development of applications and devices that run on Internet network platform, there is an urgent need to investigate these Internet applications for assurance to developers and end users. Internet application refers to the entirety of all interactive services that run on Internet network platform which are used to perform several tasks over the Internet.

These applications operate either as a server-based in which case it uses Internet protocol to receive requests from a client or as a client-base which is typically the web browsers that are requesting pages from the Server [70]; finally, there are applications that act both as servers and as clients, known as Peer-to-peer.

Over the years and in recent time, several Internet applications have emerged and with the globalization trend with its accompanying research and technological advancement, more and more applications may emerge in future. There are several Internet applications presently in use and each has almost the same or slightly difference services rendered to the clients. These Internet applications range from MSN, Skype, Dropbox, YouTube, Yahoo messenger, email clients and others. to web-browser, Facebook, Linkedin, Myspace,Twitter, Orkut, Flikr and others. However, for the purpose of this research work we are going to look at three major Internet applications that are mostly in use all over the world, which are MSN, Skype and Facebook.

2.4.1 MSN

The Microsoft Network known as (MSN) is a collection of Internet sites and services provided by Microsoft. MSN was created by the Advanced Technology Group at Microsoft, headed by Nathan Myhrvold. It was originally conceived as a dial-up online content provider like America Online, supplying proprietary content through an artificial folder-like interface integrated into Windows 95'sWindows Explorer file management program.

Categories on MSN appeared like folders in the file system [75] MSN was officially known as 'The Microsoft Network,' in August 24, 1995 when the service was launched with Windows 95, and was included with Windows 95 installations and promoted through Windows and other Microsoft software released at the time. Product support and discussion was offered through the MSN service, information such as news and weather, basic e-mail capabilities, chat rooms, and message boards similar to news groups.

The range of services offered by MSN has changed since its initial release in 1995. MSN was once a simple online service for Windows 95, an early experiment at interactive multimedia content on the Internet, and one of the most popular dial-up Internet service providers that use port 80,443 and any port above 1025. MSN was primarily a popular Internet portal. Microsoft used the MSN brand name to promote numerous popular web-based services in the late 1990s. Most notably Hotmail and Microsoft Messenger service, before reorganizing many of them in 2005 under another brand name, Windows Live. MSN.com was the 17th most visited domain name on the Internet [76].

2.4.2 SKYPE

Skype was founded by Janus Friis from Denmark and Niklas Zennström from Sweden in the year 2003 [77]. The Skype software was developed by the Estonians Ahti Heinla, Priit

Kasesalu, and Jaan Tallinn [78]. In April 2003, Skype.com and Skype.net domain names were registered and in August 2003, the first public beta version was released [79]. One of the initial names for the project was “Sky peer-to-peer”, which was then abbreviated to “Skyper”. Skype is a software application that was developed by KaZaa [80] which allows its users to talk to each other using the Internet. In that respect, Skype is a VoIP (Voice over IP) provider that uses all destination port above 1024 or port 80 and 443 that allows anyone with Internet access and the Skype software to contact other Skype users.

In essence, it is very similar to the MSN and Yahoo IM applications, as it has capabilities for voice calls, instant messaging, audio conferencing, and buddy lists. However, the underlying protocols and techniques it employs are quite different. Initially, Skype only allowed Skype users to talk to each other, i.e. non-Skype users could not make or receive phone calls. Calls to other users within the Skype service are free, while calls to both traditional landline telephones and mobile phones can be made for a fee using a debit-based user account system.

Skype has also become popular for its additional features which include instant messaging, file transfer, and video conferencing. Skype has 663 million registered users as of 2010 [81]. The average number of users connected each month was 145 million in the fourth quarter of 2010, versus 105 million a year earlier, while paying customers rose over the same period to an average 8.8 million per month, from 7.3 million. Skype reached a record with 30 million simultaneous online users on 28 March 2011 [82].

The network is operated by Microsoft Skype Division, which has its headquarters in Luxembourg. Most of the development team and 44% of the overall employees of Skype are situated in the offices of Tallinn and Tartu, Estonia [78]. eBay acquired Skype Limited in September 2005 and in April 2009 announced plans to spin it off through an initial public offering in 2010, it was acquired by Silver Lake Partners in 2009. Microsoft agreed to purchase Skype for $8.5 billion on May 2011 and the company is to be incorporated as a division of Microsoft called Microsoft Skype Division. Some network administrators have banned Skype on corporate, government, home, and education networks, citing reasons such as inappropriate usage of resources, excessive bandwidth usage and security concerns.

2.4.3 Facebook

Facebook is a social networking service that was launched in February 4, 2004 by Mark Zuckerberg while studying Computer Science at Harvard University [46]. It was originally known as The Facebook and the name was taken from sheets of paper that Zuckerberg distributed among freshmen students and staff to profile them. With the introduction of Facebook, Zuckerberg‘s initial crude profiling network experienced enormous popularity, and

within a month, half of Harvard‘s population had a profile of their own. The network started to spread among other universities in the Boston area, the Ivy League schools and finally to other US Universities.

In August 2005, Zuckerberg purchased the domain name Facebook.com, as it is known today, for $200,000 [83]. In September 2005, it was opened to US high schools and in a short period of time the networking site was being used by students in other countries as well .

Facebook quickly established itself as a social utility that helps people communicate more efficiently with their friends, family and co-workers [84]. It focused all its efforts into creating technologies that enable users from all walks of life share information more efficiently, thereby creating a digital map of people's real-world social connections [84]. According to the Statistics on Facebook's website, it uses HTTP and HTTPS protocol, and is the second most-trafficked PHP *"hypertext pre-processor"* site in the world, and one of the largest MySQL installations anywhere, running thousands of databases [84]. While empowering its members with sharing tools, Facebook has also pioneered in providing its users with a set of privacy controls that can be used to efficiently control the amount of information being shared.

Facebook has over one billion active users, [85] of which more than half of them are using Facebook on their mobile device. Users must register before using the site, after which they may create a personal profile, add other users as friends, and exchange messages, including automatic notifications when they update their profile. Additionally, users may join common-interest user groups, organized by workplace, school or college, or other characteristics, and categorize their friends into lists such as "People From Work" or "Close Friends". Among the many reasons of joining Facebook, the most important one is to get in touch with old and current friends and acquaintances.

Facebook seems to be gradually moving the trend of communications away from the more traditional ways of telephone and email toward the Internet and cyberspace. It is the sixth-most trafficked site in the United States and the number one photo-sharing site on the Internet (Facebook 1, 2007, online). According to the company, Facebook is a social utility that helps people understand the world around them. The company develops technologies that facilitate the spread of information through social networks, allowing people to share information online the same way they do in the real world.

In May 2005, Accel partners invested $12.7 million in Facebook, and Jim Breyer added $1 million of his own money to support the investment [86]. Facebook was ranked the most used social networking service by Compete.comstudy in January 2009, because of the monthly active users worldwide [87]. Facebook finally filed for an initial public offering on February

1, 2012; it is headquartered in Menlo Park, California [88], and on May 18, 2012 they began selling stock to the public and trading on the NASDAQ [87]. Based on the income Facebook generated in 2012 which is about US$5 billion, they joined the Fortune 500 list for the first time on the list published in May 2013, being placed at position 462 [89].

Facebook is considered the 5th most successful startup company of all time, by market capitalization, revenue, and growth [90]. In 2012, Facebook was valued at $104 billion, and by January 2014 its market capitalization had risen to over $134 billion [91,92]. At the end of January 2014, 1.23 billion users were active on the website every month, while on December 31, 2013, 945 million of these total were identified by the company as mobile users. The company celebrates its tenth anniversary in the week beginning February 3, 2014.

2.4.4 Web browsing and Web based email access

A Web browser is an application used to access and view websites. While Web browsing is the process of accessing the web browser in order for us to get information provided by web servers in private networks or files in file systems. Common web browsers include Microsoft Internet Explorer, Google Chrome, Mozilla Firefox, and others. The primary function of a web browser is to render HTML, the code used to design or "markup" webpages. Each time a browser loads a web page, it processes the HTML, which may include text, links, and references to images and other items, such ascascading style sheets and JavaScript functions. The browsing processes these items, then renders them in the browser window.
Web-based email is any email a client implemented as a web application accessed via a web browser. Examples of webmail software are Roundcube or SquirrelMail, examples of webmail providers include AOL Mail, Gmail, Outlook.com and Yahoo! Mail. Practically every webmail provider offers email access using a webmail client, and many of them also offer email access by a desktop email client using standard email protocols, while several Internet service providers provide a webmail client as part of the email service included in their internet service package.
It is widely assumed that the traffic on the Internet is mostly due to web browsing and to the access to emails using a web platform. This is an issue that we will address in the research.

2.5 Network Traffic Growth

The evolution of the Internet over the last ten years has been accompanied by the development, growth, and use of a wide variety of network applications. These applications range from text-based utilities such as file transfer, remote login, electronic mail, and network news from the early days of the Internet, to the advent of desktop videoconferencing, multimedia streaming, the World-Wide Web, and electronic commerce on today's Internet. The conventional wisdom with respect to the Internet is that size matters.

Some scholars have suggested that the Internet's size is governed by a form of Moore's Law, growing at a constant rate [93].

Size is also generally regarded to be an important determinant of value. It has long been recognized that the number of possible connections increase quadratically with the number of endpoints [94]. Metcalfe's Law posits that if the value of a network goes up in proportion to the number of connections and the costs of increasing network size increase linearly, increasing a network's size necessarily increases its value [95]. This logic was used to justify the enormous investments that fueled the dot-com bubble [96]. The most popular and extremely misleading myths of the dot-com and telecom bubbles was that Internet traffic doubles every 100 days (3 months, or 4 months). Based on this fact the result of the Network traffic growth on Internet Applications Analysis finding that we carried out confirmed the above theories Purported by Moore's law and Metcalfe's law.

In the past, it used to be enough to have an online presence on the Internet for the one-way broadcasting and dissemination of information. But today, social networks such as Facebook, Skype and MSN are driving new forms of social interaction, dialogue, exchange and collaboration. Social networking site facilitate users to exchange ideas, to post updates and comments, or to participate in activities and events, while sharing their wider interests. From general chit-chat to propagating breaking news, from scheduling a date to following election results or coordinating disaster response, from gentle humour to serious research, due to the fact that social networks are now used for different reasons by various user communities, it has given rise to the increase in the usage of the network traffic. *i.e* Morgan Stanley estimates that there were about 830 million "unique" users of social networks worldwide at the end of 2009. Based on a total Internet user population of 1.7 billion at the end of 2009, according to ITU's World Telecommunication/ICT Development Report 2010, this suggests that around half of all Internet users could currently be using social media applications.

Many social network users access these services over their mobile phones. According to ITU's report *Measuring the Information Society 2010*, mobile broadband subscriptions reached an estimated 640 million at the end of 2009, driven by growing demand for smart phones, new applications and social networking services, and are set to exceed 1 billion this year. The market research firm eMarketer projects that just over 600 million people will use their phones to tap into social networks by 2013, compared with 140 million in 2009. Facebook passed the historic milestone of 500 million users on 21 July 2010.

The diagram below shows how many users are drawn to some popular social networks in early 2010.

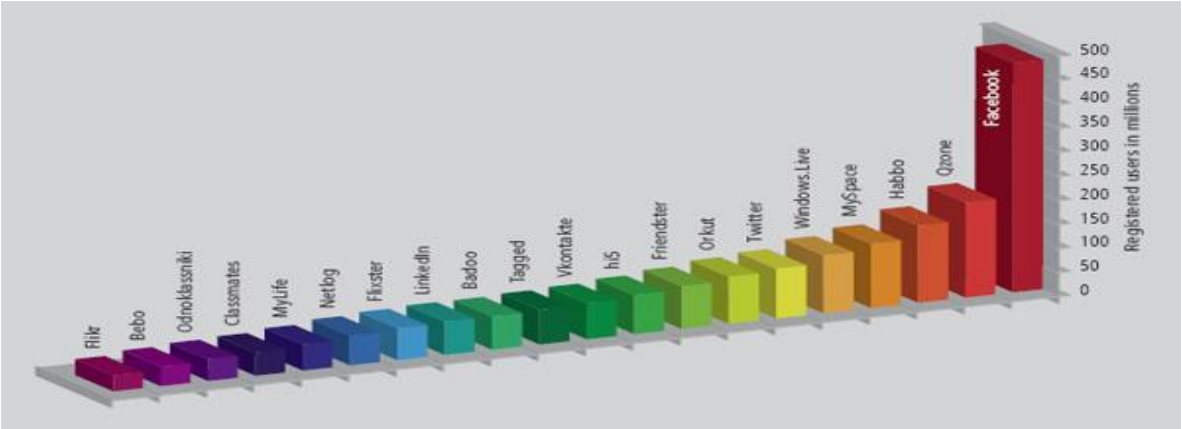

Figure 2.6 - Twenty most popular social network website May 2010.

2.6 The growth in Social network Traffic

The Internet transports roughly 10 billion gigabytes of data in a month a figure that some observers expected to quadruple by 2012, although we could not confirm this prediction. The market and advertising research company Nielsen estimates that the average time spent on social networks grew from three hours in December 2008 to five and a half hours in December 2009, based on a survey of social media use in ten countries.

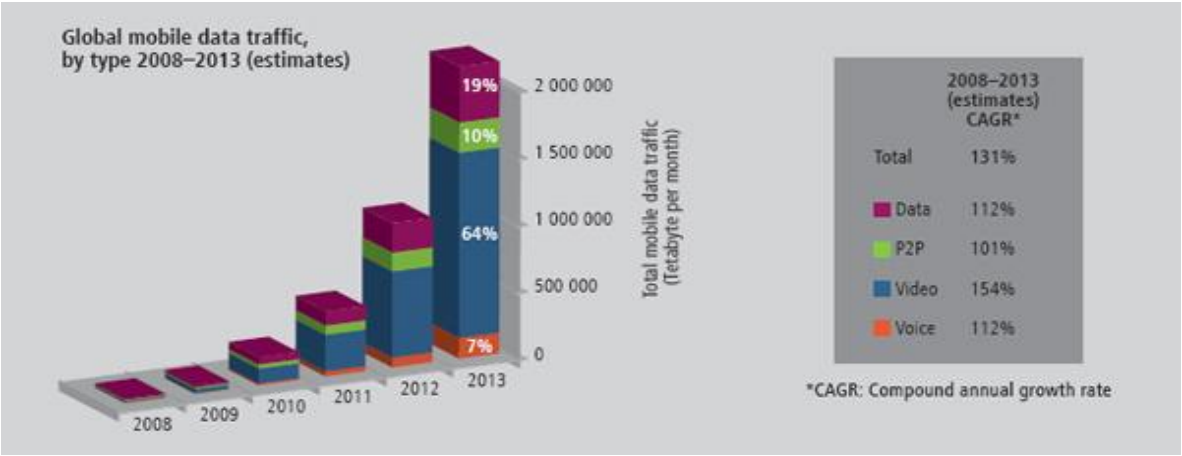

Figure 2.7 - Expansion in the demand for video 2008-2014.

Half of all mobile data use in the United Kingdom is accounted for by Facebook, so social media look set to continue driving future growth in traffic, with video-streaming applications (such as YouTube) expected to account for a large proportion of that traffic. Over the past years, Internet data traffic has grown by a factor of 56, driven partly by people uploading more data. On average, people uploaded fifteen times more data in 2009 than they did just three years previously. Cisco projects that global mobile data traffic will grow by sixty-six times from 2008 to 2013, with video forecast to account for around 64 per cent of all global mobile data traffic by 2013 [93].

Chapter 3

Data Trace Selection

In this chapter, all the traffic analysis is based on the information collected from the WIDE project and Measurement and Analysis on the Wide Internet (MAWI) traffic measurement. A thorough description of our network will be given in this chapter, and we will explain why we selected these data trace, where and how the data is collected.

3.1 The WIDE Project and MAWI

The Widely Integrated Distributed Environment (WIDE) project [71] was launched in 1988 in Japan and is made up of more than 100 loosely bound organizations from all over the world. The visionary goal of WIDE is to construct a dependable Internet (that can used by people from all walks of life in any situation with a sense of security). WIDE research activities cover all different layers of the Internet, including activities such as flow measurements with sFlow/NetFlow and analysis of IPv6, and DNS.
The Measurement and Analysis on the Wide Internet (MAWI) is a research group that assists researchers provides a traffic repository of data captured on the WIDE backbone and to evaluate traffic anomaly detection in the area of Internet traffic analysis. It consists of a set of labels locating traffic anomalies in the MAWI archive (sample points B and F). The MAWI traffic repository is the one responsible for archiving the traffic data that is collected from the WIDE backbone networks. The WIDE network (AS2500) is a Japanese academic networking group connecting universities and research institutes that has been responsible for providing anonymized packet traces to the public since 1999, total volume of available data exceeds 1TB as of 2008. The data used here are all publicly available on the website (http://www.fukuda-lab.org/mawilab/[72]).

3.2 MAWILab

The MAWI Lab duties are to locate anomalies in the MAWI archive with a simple traffic taxonomy that consists of four different labels: anomalous, suspicious, notice and benign.
These labels are acquired using an advanced graph-based methodology that compares and combines different and independent anomaly detectors. The data set is updated daily and it includes new traffic from upcoming applications and anomalies which are explain below:

- The label anomalous that is assigned to all abnormal traffic are identified by an efficient anomaly detector.

- The anomalous label suspicious is assigned to all traffic that is probably anomalous but not clearly identified by the MAWILab method.
- The anomalous label notice is assigned to all traffic that is not identified as anomalous by the MAWILab method but can be identify by at least one anomaly detector.
- The remaining traffic is labeled benign because none of the anomaly detectors identified them.

3.3 Description of MAWI dataset and collection point

The traffic dataset collection point for this thesis is from a backbone network with 100Mbps link called the WIDE network (AS2500) in Japanese academic network that is connecting Universities and research institutes in a link between Japan and the USA. As it was not clear the nature of the users that generate the tributary traffic, we have email two representatives for MAWI Lab, and to the question "How do you classify the collected traffic? Residential, Academia, Enterprises?" we received the following answer "(...) one is connected to academic Japanese traffic, and the other is to the rest of the Internet.(...)" (*sic*). Regarding if the traffic was either inbound or outbound, we were answered that the traffic is both inbound and outbound. Therefore we can safely assume that this traffic sample is ergodic as it will include communications with companies and with individual users, but it will also include communications with other academic institutions.

The dataset consists of a set of labels locating traffic anomalies in the MAWI archive [6] (sample points B and F). The labels available in the dataset are sample points A,B,C,D,E and F, respectively. The MAWI traffic repository is in charge of archiving the traffic data collected from the WIDE backbone networks.

The datasets used in our research analysis are daily 15- minutes packet traces data captured at Sample point-B from 2004/01 to 2006/06, then at Sample point-F from 2006/07 to 2013/12 (total of 3653 days analysis) which is connected between a link Japan and US.
These are transit links of the WIDE network, and the link of B was replaced in July 2006 by the link F. (although, some traces were missing just after the upgrade until 2006/10). At sample point B, congestions were often observed, the link was a 18Mb/s, with 100Mb/s Committed Access Rate. The link for F is over-provisioned, it started as a full 100Mb/s link and upgraded to a bandwidth of 150Mb/s on June 1 2007.
The MAWI Lab measured the traffic every day from 14:00 to 14:15 JST (Japanese Standard Time, UTC+9), these corresponding traces with IP addresses anonymized and payloads removed are made available to the public along with a summary information web page about the traffic. The traffic of the WIDE transit link is mostly trans- Pacific commodity traffic between Japanese research institutions and non-Japanese commercial networks, as WIDE

peers with all the major domestic ASes at the Internet Exchange Points it operates, and international traffic between academic networks goes through other international research networks. The traffic of the transit link is also asymmetric as WIDE has other trans- Pacific links, meaning that many flows can be observed only in one direction. This makes us to study the traffic separately for each direction, being labeled US2Jp, for traffic going to Japan, and Jp2US, for outgoing traffic, as most traffic is between Japan and the USA. The traffic is highly aggregated. A 15-minute-long trace usually contains 300k-500k unique IP addresses, and contains various kinds of anomalies. We examine the evolution of the traffic for 10 years using these dataset, under both congested and over provisioned conditions.

3.4 Limitations on the recorded data

During the collection of the traffic dataset trace on the MAWI backbone link, We make several key observations from our study. Over the 10 years of measurement periods, we can make several remarks:

- We notice that, 159 days of traces are missing due to the scheduled network maintenance. Except for the three month gap (link update) in June–Aug. 2006, the concern for data continuity is minimal.
- We also discovered that traffic measurement was recorded twice on these days (28/04/2006 between 19:15- 19:45hrs and on 30/06/2006 between 18:15- 18:29hrs.
- There were two type of traffic trace measurement done in the MAWILab, one is connected to the Japanese academics traffic and the other is the rest of the Internet *i.e* residential and enterprise traffic.
- According to section 3.2 each trace is tagged with a summary of anomalies from MAWILab. In a nutshell, MAWILab identifies anomalies using a combination of four anomaly detectors and takes advantage of a community mining algorithm to aggregate the detectors results.

3.5 Reason for choosing this collection point

The evaluation of Internet traffic monitoring analysis on networking environments can be very complex since measuring Internet traffic is a laborious and expensive task, measurement projects typically want to archive not only their analysis results, but also the raw data, such as packet level traces or flow data. Moreover, the data needs to be anonymized, because of security and privacy issues. Archiving raw data is furthermore important to keep scientific results reproduceable, to allow comparisons between historical and current data, to make additional analysis regarding different aspects possible, and finally to share datasets with the research community. Archiving of network traces is not always a trivial task, especially for longitudinal, continuous measurement activities. This is the reason why we decide to choose TMA portal (European research portal on traffic monitoring and analysis) as a site for our

collection point, because it is a research group that is working in the field of Traffic Monitoring and Analysis, is an essential research area within the field of Communication Networks, that connects several research groups around the globe that are collectively advancing our understanding of Internet traffic measurement on networks environment and their users.

3.6 The scope of the selected data trace

In this thesis, the selection of data for our analysis was divided into the following categories:

- we analyze the traffic measurements taken from two sample point (B and F) from the MAWI repository backbone network link [6];
- we provide a longitudinal study of traffic analysis measurement from 2004 to 2013 (10 years) over a trans-Pacific backbone link, in terms of IP protocols IPv4 and IPv6, the transport layer TCP and UDP, the applications layer, Hypertext Transfer Protocol HTTP, the average packet sizes and Byte size for IPV4 and IPv6.

We were mostly interested in the statistics of the collected data, and therefore we did not download and analyzed the raw data traces. The relevant statistics were summarized in an application over a spreadsheet, and comparison and analysis was performed in the results and its evolution in time.

Chapter 4

Result for Data Trace Analysis

In this chapter we present and discuss the results for our data trace analysis. For each Protocol, we trail the analysis of the data from three different layers, which are the network, transport and application layer. We analyze the results at various levels of granularity. After each step, we try to provide an insight on the rationale for the behaviour of the data.
The presented charts cover 10 years of Internet traffic at the before mentioned collection point.
We have selected a number of features from the recorded traffic, including analysis of IPv4 and IPv6 that would allow us to draw conclusions on our initials research questions which are:
(I) What is the evolution of the ratios of use for the different transport protocols over the studied 10 years? Here we addressed various concerns related to the percentage of TCP|UDP traffic, HTTP traffic, IPv4 |IPv6 traffic and their statistics.
(II) What is the effect of these protocols on the Internet Traffic for this period (2004 - 2013)?
(III) What is the evolution on the amount of traffic being transmitted from or received by the hosts in this link? Here we obtained previously recorded flows of IP traffic on the Internet for a particular link, evaluate and analyzed the records of the link traffic, *i.e.* the amount of traffic being transmitted from or receive by the hosts.
Taking these research question into consideration, we have selected the following data features from the selected data traces: percentage of TCP, UDP, HTTP, HTTPS, SMTP SSH, IMAP, Telnet, DNS, POP3, Other protocols that are inside TCP and UDP, number of packet of IPv4 and IPv6 and finally, the average packet size of IPv4 and IPv6.

4.1 Analysis of our Results

The charts below show an overview of result for traffic trace analysis presented in different parameters, such as, IPv4, IPV6, HTTP vs. HTTPS, TCP vs. UDP, SMTP, IPv4 vs. IPv6 and DNS.

Figure 4.1 show the recorded number of packets for IPv4 protocol collected for a period of ten years. From the our chart below we can see that the over these 10 years the number of transmitted packets has increased by 10 times because we go from 10 to the 7 power to 10 to the 8 power, we also have some peaks in here but the numbers of the packet have increase constantly. This is an expected result.

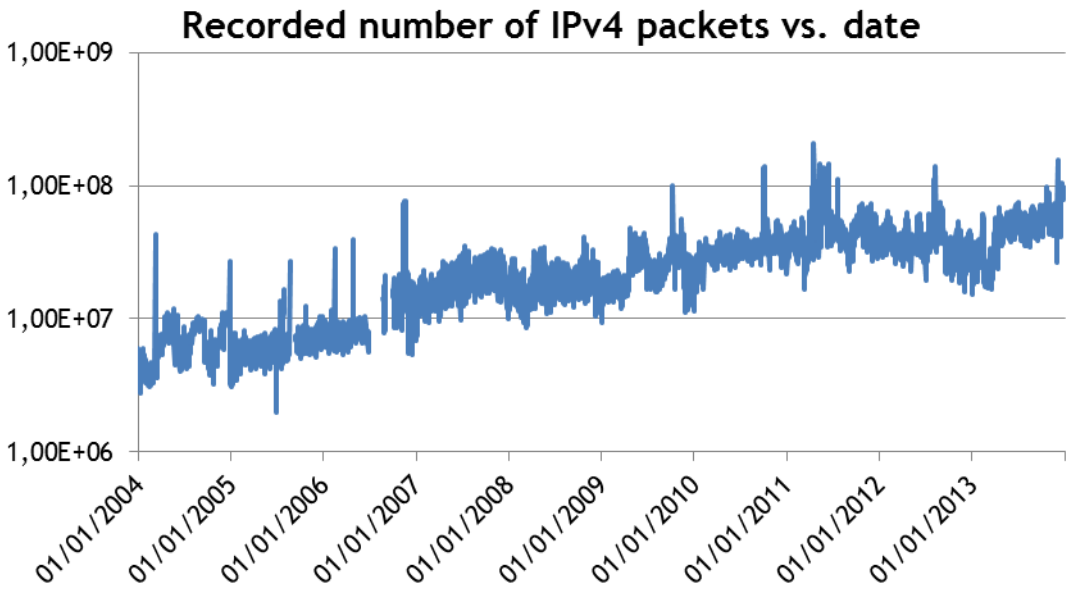

Figure 4.1 - IPv4 number of packets.

Due to the characteristics of the collection point, and due to the fact that we do not know how the evolution on the number of users for the analyzed networks has progressed, we cannot confirm the "Nielsen Law" that states that the number of (high-speed) Internet connections doubles every 21 months [73]. Nevertheless, Nielsen's Law results would agree with the measurements and the data plotted in Figure 4.1.

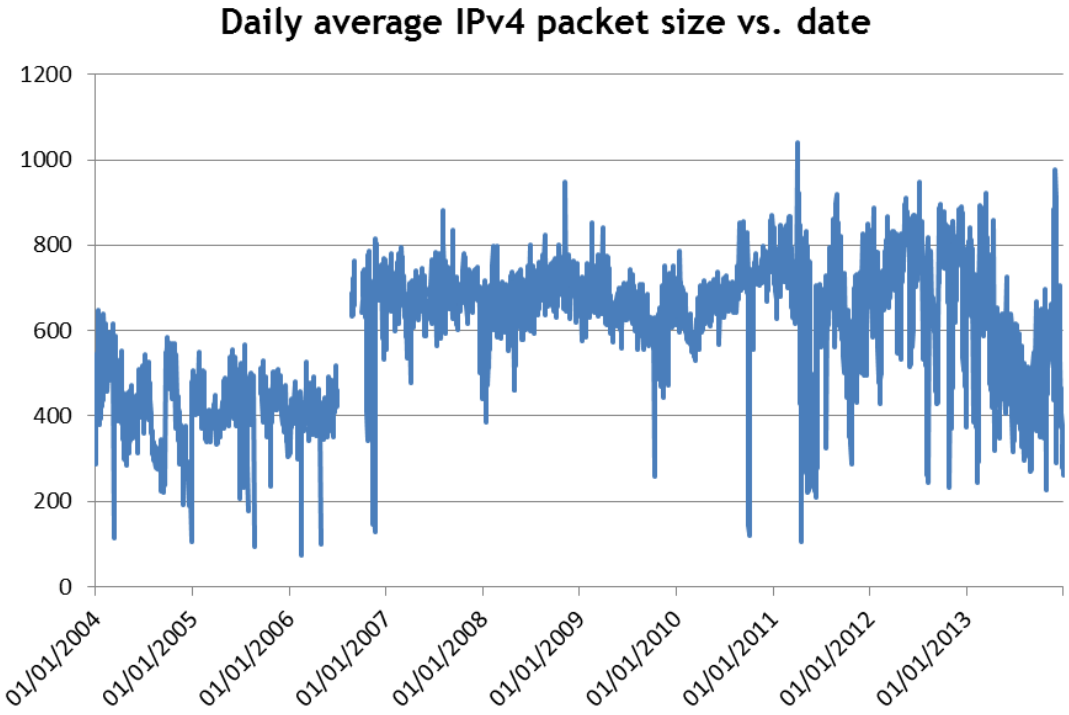

Figure 4.2 - IPv4 average packet size.

Figure 4.2 show the daily average packet size results for IPv4 protocol collected from the MAWI Lab for a decade. Mainly looking at our chart we observed that before the link upgrade, the average packet size was the smaller when compared to the rest of the years, which may have to do with congestion from the backbone link. Although we cannot be prove this hypothesis, we cannot find other viable explanation since a change in the used applications that generate the traffic is not expected to occur in such a short period of time.

Another observation from our chart is that from 2011 onward we notice a higher variance on the average packet size of IPv4. This higher variability suggests that some new applications may have started to be used thus creating the change in the traffic profile.

Figure 4.3 shows an overview of the recorded number of packets for IPv6 protocol collected for the same number of years as for Figure 4.1. In this chart we can see that IPv6 has generated over a million packets per day (in average) since middle 2012. Between 2004 - 2007, in average, there was less than 1000 packets per day and finally in 2007 - 2011 the number of the packet was around 100,000 packets.
This increase in the number of IPv6 packets is consistent with the growth of IPv6 usage, as observed in Figure 4.5, in middle 2012 and in 2013 that account for 30% of the overall traffic ratio of IPv4 versus IPv6.

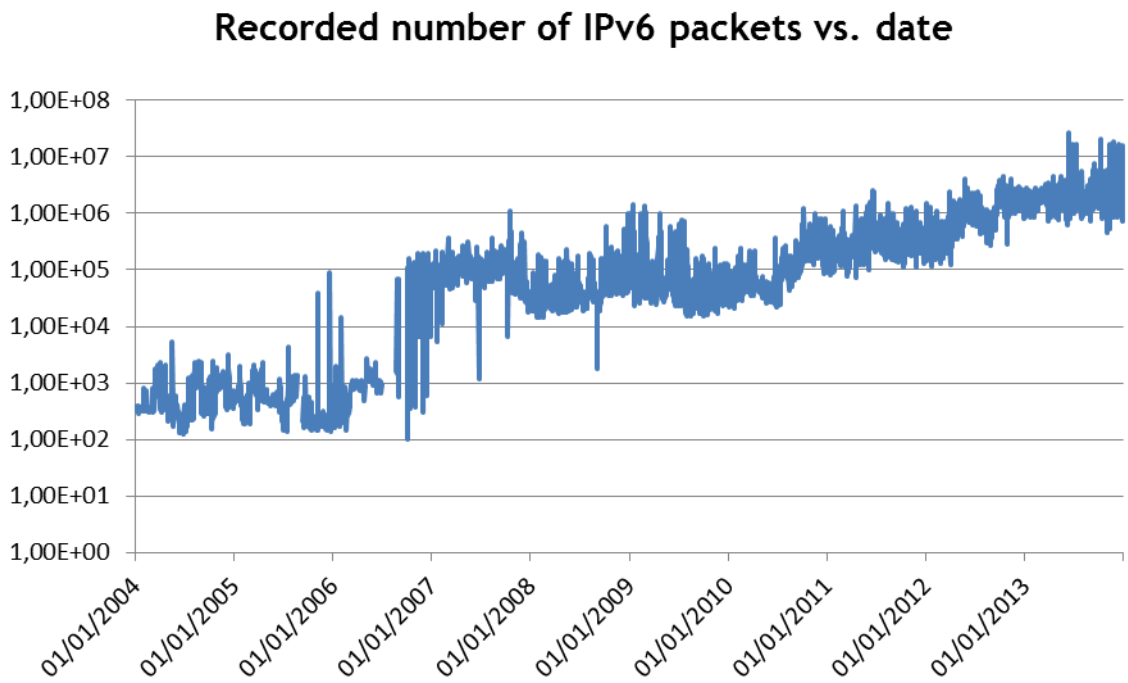

Figure 4.3 - An overview IPv6 number of packets.

In figure 4.4 the IPv6 daily average packet size was analyzed, from our findings it seems that lately the average packet size in IPv6 has been increasing. This increase appears to be related to the increase of IPv6 traffic with reference to our chart IPv4 vs. IPv6 in figure 4.5 below.

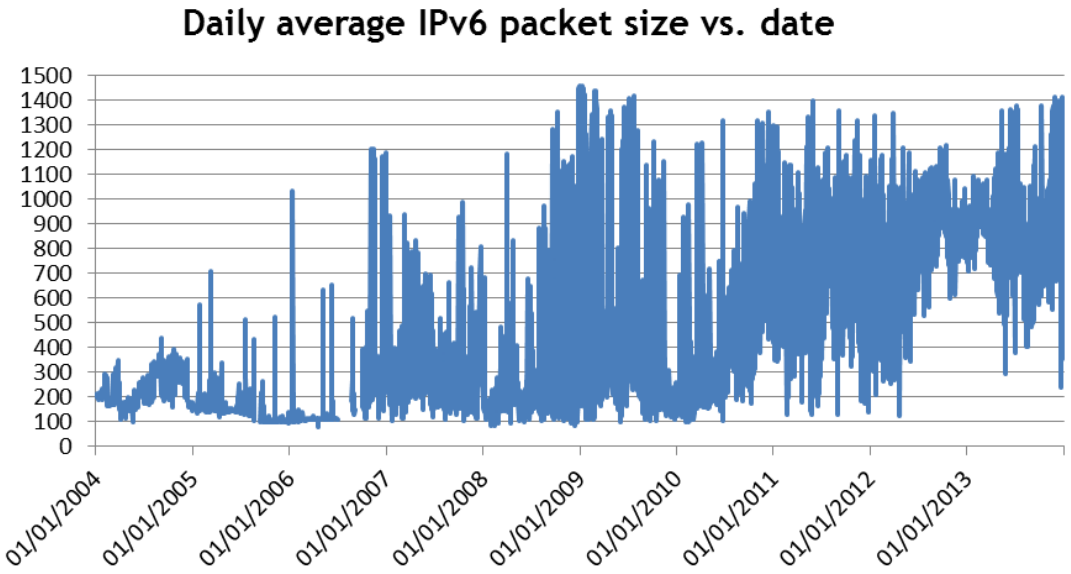

Figure 4.4 - IPv6 average packet size.

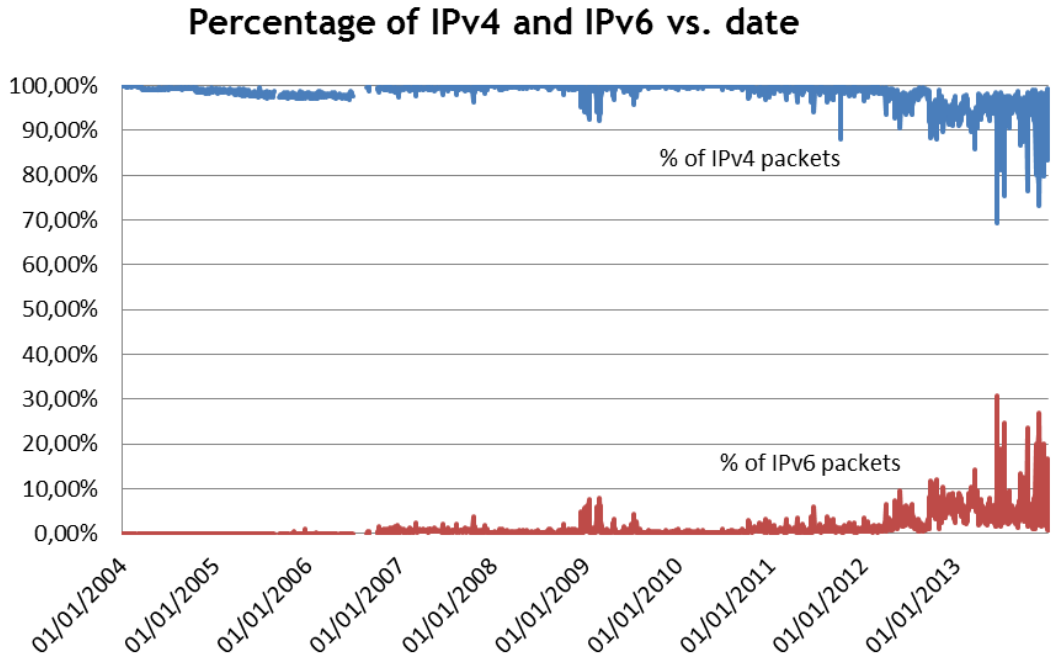

Figure 4.5 - IPv4 and IPv6 packets.

In Figure 4.5 we compare the evolution of ratios for the number of IPv4 and IPv6 packets in the link. It seems that IPv6 has been gaining traffic since middle 2012 and in 2013 at times it reaches some 30% of the overall traffic. Secondly, from 2004 to 2007 the volume of IPv6

traffic was residual, so in 2007 it started gaining some volume, while still residual when compared to the volume of IPv4. In 2009 we couldn't tell what happen but we observed that in 2010 the traffic was like before in 2004 - 2006. Finally, as observed previously, in middle 2012 the ratio of IPv6 versus IPv4 number of packets has taken off, and this may explain the increase in the average size of IPv6 packets as seen in Figure 4.4.

The results of the analysis in our chart in figure 4.6 shows the recorded number of average packet size in months from 2004 -2013 for TCP and UDP protocols. From our chart we can see that TCP and UDP make almost 100% of the traffic initially. In late 2011 we can see a decrease in TCP, but we cannot see an increase in UDP. Other layer 4 protocols that account for the missing volume of traffic percentage are ICMP, some IPSec, some IPv6 in IPv4 and some fragmented traffic. This means that there are some other protocols that have been used but we don't know what they are because the collection application at the collection link could not identify these protocols, referring them as "other protocols".

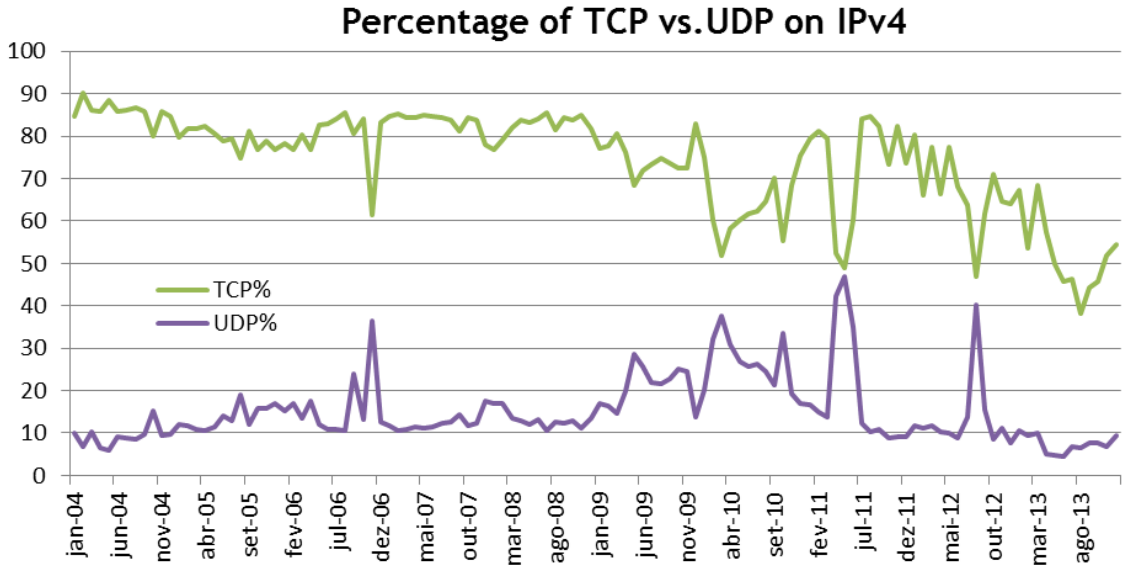

Figure 4.6 - TCP vs. UDP on IPv4 protocol.

While this could probably mean that there had been an increase in the number of applications that are using ports outside the "Well-known" port range as defined by RFC 6335, we will see in Figure 4.7 that this is not the case.

The chart in figure 4.7 shows the percentage of other protocols that are inside IPv4 and IPv6 (and, as mentioned previously, the values for IPv6 are residual). We observed that some time the unknown protocol were 30% - 40% or at most 60% of the TCP traffic, but we do not use these analysis for IPv6 because of low percentage of IPv6 therefore the numbers were not meaningful. As discussed previously, well known protocols such as ICMP, IPSec and other are not charted here, and account for the missing part as seen in Figure 4.6.

So the possible conclusion is that the number of applications that are using ports that are outside the range of well-known ports has been decreasing since late 2011.

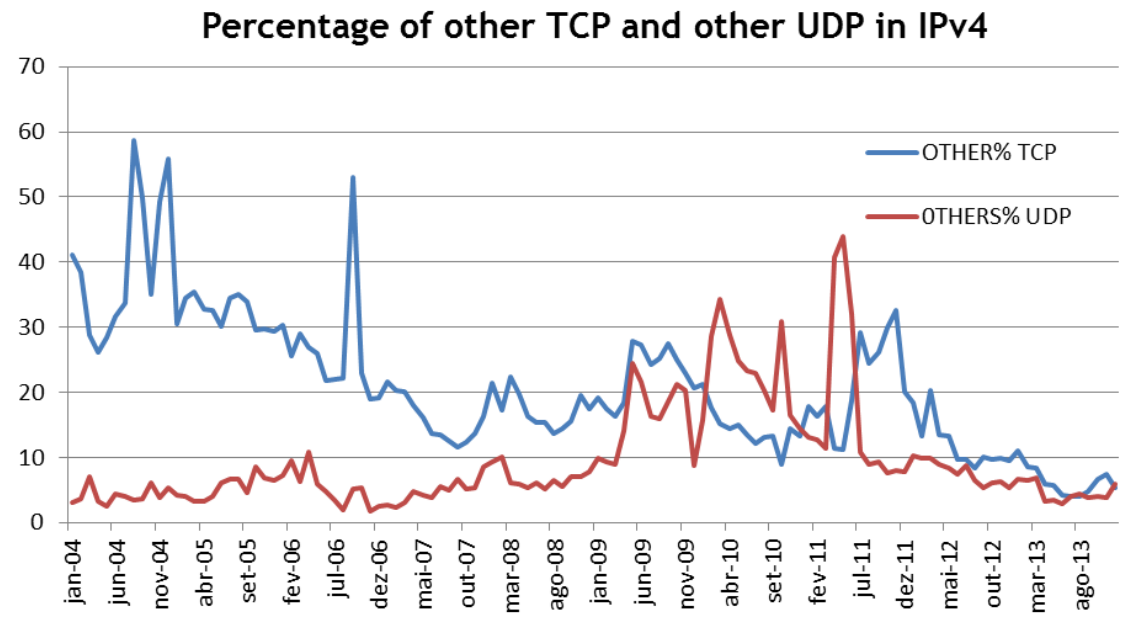

Figure 4.7 - other% TCP and other% UDP on IPv4 protocol.

Figure 4.8 shows that around month 26 (which is February 2006), HTTPS started growing and since then it is more frequent than HTTP (in average). This is probably related to the joint effort to make the web a more secure place for data, and to the fact that many popular websites have started to offer their contents over HTTPS. We can also see that spikes in usage are common for HTTP and HTTPS thus allowing us to infer that the relation between

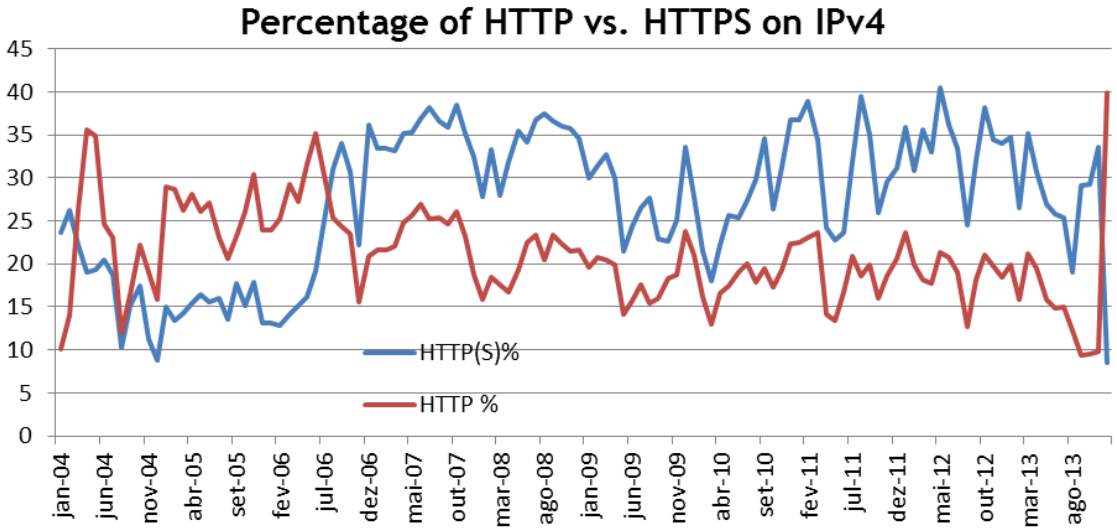

Figure 4.8 - HTTP vs. HTTPS on IPv4 protocol.

these two protocols has very much stayed unaltered since late 2006, and even before, with the exception of the switching period occurred in 2006.

In Figure 4.9 below we can witness the decrease in SMTP. This is probably due to the fact that increasingly users have switched to web based email management platforms such as Gmail or Hotmail, since we know that the number of emails has not decreased [74].
We can also see an increase in SSH usage at around 2005 and 2008, while for the period from 2010 to 2013 the SSH traffic appears to be stable at around 1%. We cannot provide an explanation to the increase of SSH traffic in the periods mentioned before.

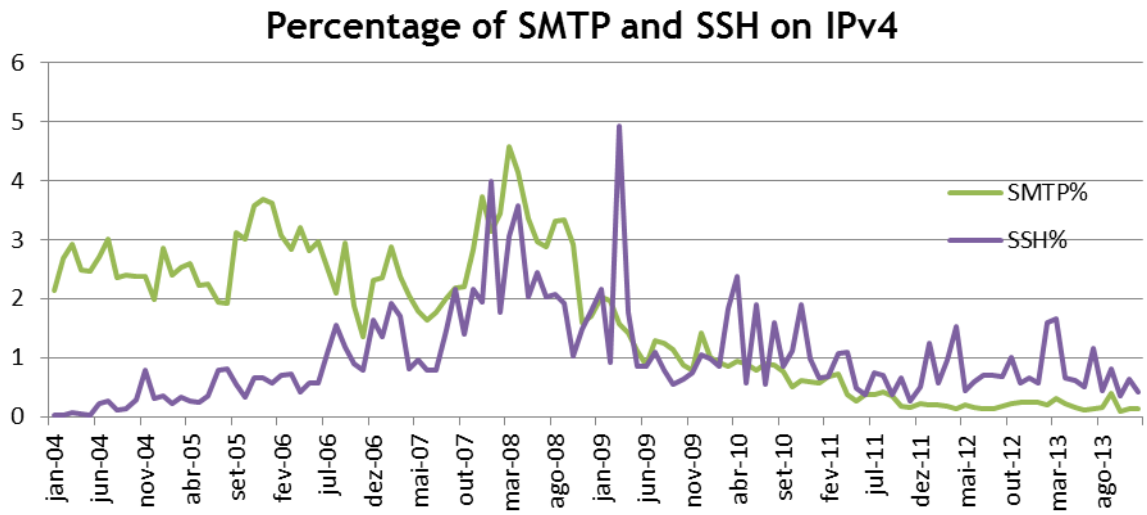

Figure 4.9 - Percentage of SMTP and SSH on IPv4 protocol.

Finally we analyze DNS traffic. Figure 4.10 shows the result of the percentage of DNS packets in IPv4. From our findings it seem that despite the increase on the overall traffic as seen on the chart IPv4 in Figure 4.1, there has not been a change in the ratio of DNS as shown in the chart, showing that an increase in 10x over the period of 10 years did not made a dramatic change in the ratio that the DNS service was used.

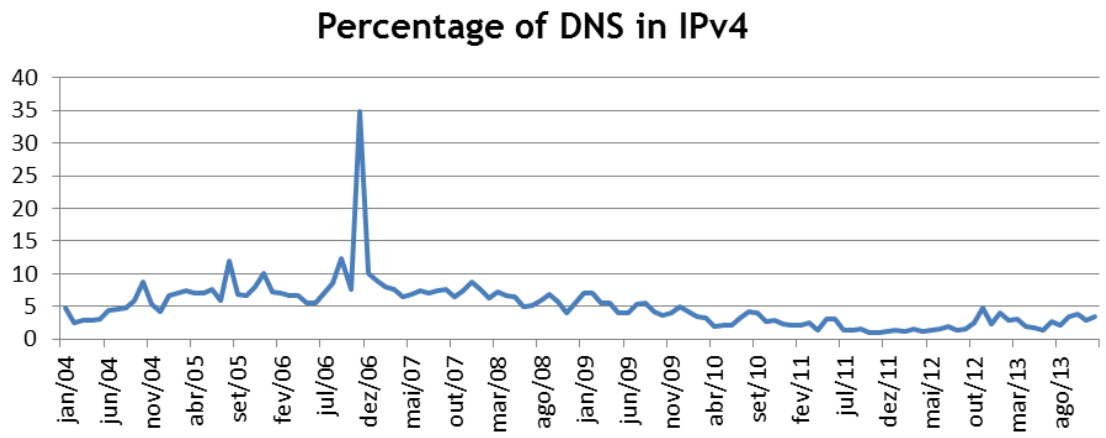

Figure 4.10 - DNS in IPv4 protocol.

In fact, if any, the ratio has dropped. For the last 4 years the ratio of DNS has been in average below 5% and for the initial 6 years below 10% in average.

We have detected a spike in November 2006, and proceeded to verify the data related to that period. Although the data is confirmed, we cannot explain why it happens.

This activity was recorded between 11-21 November 2006. Analyzing the data with more detail, we were able to plot the chart in Figure 4.11. We can see that there was a set of specific days where there was an unusual activity for DNS traffic, while for the other days, the values averaged 10% in line with the remaining days of the period as seen in Figure 4.10.

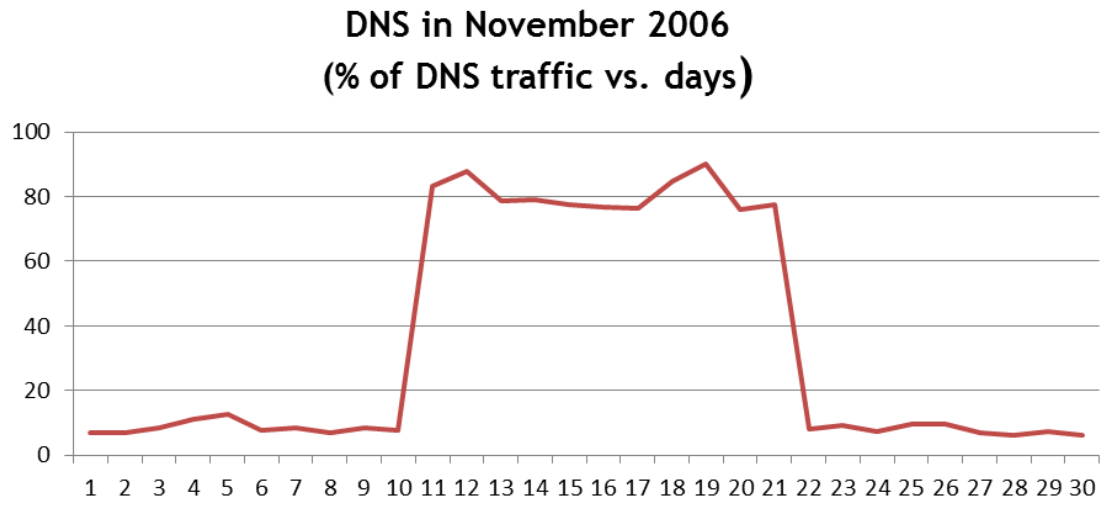

Figure 4.11 - DNS result November 2006.

4.2 Conclusion

In this chapter we discussed the results for our data trace analysis. Although the trace files are in the dump format and contain all the recorded packets, we chose to make use of the statistics for each of the data files made available on the MAWI archive. Then, we described

how the usage of these data have increased the volume of Internet traffic for the evolution of these period. Further, we have seen the structure of the data and its variability according to the approach we adopted. Finally, we were able to analyze our result and draw a conclusion that there has been increase in the usage some of these protocols in the last 10 years.

Chapter 5

Conclusion

In this dissertation, the analysis of Web Protocols evolution on Internet traffic, collected over Trans-Pacific backbone links (MAWI) was analyzed, evaluated and optimized to study the underlying data trace analysis for a period of 10 years. This chapter concludes the dissertation. It summarizes our contributions, points out the applicability of our results and suggests areas for future work.

5.1 Summary and contribution

During course for this research, we learned the functionality of various protocols and how they can be incorporated together to analyze network traffic effectively. We have produced results in three main layers, which are: the network, transport and application layer. We summarize these contributions below.

First, we presented a chart of Internet traffic data trace for 10 years of continuous wide-area network activity at the above mentioned collection point. These traces contain a record of every packet flowing through the wide-area network link between Japan and US at their respective sites every day for 15 minutes. In addition we selected a number of features from the recorded traffic, each records includes analysis of all the protocols: percentage of TCP, UDP, HTTP, HTTPS, SMTP SSH, IMAP, Telnet, DNS, POP3, Other protocols that are inside TCP and UDP, number of packet of IPv4 and IPv6 and finally, the average packet size of IPv4 and IPv6 as mention in chapter four to investigate the volume of Internet traffic.

Secondly our evaluation shows a unique day by day longitudinal study of long data trace for the above mentioned year which shows that, the estimations of traffic statistics exhibit a huge variability, largely due to traffic condition variations (congestions, restrictions,. . .) and anomalies constantly but randomly occurring. This impairs the possibility of drawing long term evolution conclusions on our research questions.

Third, going through all these findings we were able to answer the following research questions on our thesis and draw a conclusion that from 2004 to 2010, TCP and UDP make use of almost 100% the traffic, but there has been a decrease in TCP and no increase in UDP since late 2011 to 2013. We also observed that since February 2006 HTTPS has been growing and is more frequently use than HTTP, another observation is that from 2011 we notice a higher variance on IPv4 packets and a consistent increase in the usage of IPv6 in the middle of 2012 and 2013 that account for 30% of the overall traffic ratio of IPv4 versus IPv6.

Fourth, to answer our second research question, going through our chart we notice that there has been increase in the usage some of these protocols and that some other new protocol/application are been used also on the Internet traffic in the last 10 years.

So, in conclusion, we asked three questions to be addressed by this research. We have selected a link which has been collecting and registering traffic for a consecutive and coherent period of time, and, because of the nature of traffic in this link, we can assume that the samples are ergodic, *i.e.*, they are representative not only of the whole period the traffic has been transmitted, but also to other links who aggregate traffic from residential, enterprise and educational users. Finally we have selected the relevant features of this traffic and drawn feasible answers to our research questions.

Question 1 was “What is the evolution of the ratios of use for the different transport protocols over the studied 10 years?" After analysis, we have found that the ratios for the different protocols have in fact changed in the last years, having found, in our opinion, two outstanding results, and one comment. The first outstanding result is connected to the inversion of rations between HTTP and HTTPS, in middle 2006. The second outstanding result is the loss of prevalence of TCP and UDP in the last two years. The comment has to do with common claims that email and in particular spam email is responsible for a large percentage of Internet traffic. We have not been able to confirm that, moreover, we have witnessed a decrease in the amount of SMTP traffic in the last five years, being now confined to under 2% of all link traffic.

Question 2 was "What is the effect of these protocols on the Internet Traffic for this period (2004 - 2013)?" We have concluded that there has been a timid increase of IPv6 versus IPv4 on the last two years of the study. The prevalence of HTTPS also means that communication flows have been "burdened" with security features. We have also been able to confirm that there is a change in the balance of other protocols (*e.g.* TCP and UDP), and that this change probably means that other applications have been deployed and used and these are not using regular TCP or UDP segments.

Question 3 was “What is the evolution on the amount of traffic being transmitted from or received by the hosts in this link?" Finally, we have concluded that the increase in the number of transmitted packets follows closely Neilsen's Law.

Our initial research hypothesis was that the change in the Internet application ecosystem in the last 10 years is perceivable through the analysis of the Internet traffic at a random yet ergodic collection point. Taking into consideration all of the above conclusions we are able to state that the hypothesis is confirmed.

5.2 Future Work

With the work that has been done so far, the evolution of web protocols on Internet traffic has been evaluated and optimized. The result shows there is room for further improvement in this area, since the traffic trace on the MAWI backbone link is a continuous process, therefore we suggest that there is need to monitor recent records and also other collection sites to access if the our conclusions are confirmed and the observed trends are maintained.

Bibliography

[1] http://www.nlanr.net/Flowsresearch/fixstats.21.6.html.

[2] M. Fomenkov, K. Keys, D. Moore, and K. Claffy, "Longitudinal study of Internet traffic in 1998-2003," in *WISICT*, 2004.

[3] W. John and S. Tafvelin, "Analysis of Internet backbone traffic and header anomalies observed," in *ACM IMC*, 2007.

[4] P. Pan, Y. Cui, and B. Liu, "A measurement study on video acceleration service," in *IEEE CCNC*, 2009.

[5] B. R. Chang, and H. F. Tsai, "Improving network traffic analysis by foreseeing data packet-flow with hybrid fuzzy-based model prediction," *Expert Systems with Applications,* vol. 36, no. 3, Part 2, pp. 6960-6965, 2009.

[6] http://www.tma-portal.eu/topics/resources/traces-collections.

[7] http://technet.microsoft.com/en-us/library/bb726991.aspx.

[8] University of South California (1980), DOD Standard Internet Protocol, RFC 760.

[9] http://en.wikipedia.org/wiki/OSI_model.

[10] [36] Charles M. Kozierok, *The TCP/IP Guide: a comprehensive, illustrated Internet protocols reference* .

[11] RFC791 J. Postel. *Internet Protocol*. RFC 791, IETF, September 1981.

[12] Uyless Black, "TCP/IP & Related Protocols (Second Edition)", 1994, McGraw-Hill Ryerson, Limited.

[13] Cisco (2007-2009). *Classful IP addressing*. Retrieved May 20, 2009, from http://curriculum.netacad.net/virtuoso/servlet/org.cli.delivery.rendering.servlet.

[14] Govil, J. (2007). On the investigation of transactional and interoperability issues between IPv4 and IPv6.*Proceedings of the 2007 IEEE International Conference on Electro/Information Technology*(pp. 604-609). Washington: IEEE Computer Society.

[15] Green, D., Fiuczynski, M. E., & Marc, E. (2006). IPv6 translation for IPv4 embedded systems. *Proceedings of the 2006 IEEE Sarnoff Symposium*. Washington: IEEE Computer Society.

[16] Wang, Y., Ye, S., & Li, X. (2005). Understanding current IPv6 performance: A measurement study. *Proceedings of the 10th IEEE Symposium on Computers and Communication*(71-76).Washington: IEEE Computer Society.

[17] Hiromi, R. & Yoshifuji, H. (2006). Problems on IPv4-IPv6 network transition. *Proceedings of the International Symposium on Applications and Internet Workshops* (pp. 38-42). Washington: IEEE Computer Society.

[18] Govil, J., Govil, J., Kaur, N., & Kaur, H. (2008). An examination of IPv4 and IPv6 Networks: Constraints and Various Transition Mechanisms. *Proceedings of the 2008 IEEE Southeast on*(pp. 178-185). Washington: IEEE Computer Society.

[19] Davies, J. (2008a). *Understanding IPv6* (2nd ed.). Washington: Microsoft Press.

[20] http://www.just2good.co.uk.
[21] http://docs.oracle.com/cd/E19683-01/817-0573/6mgc65bb3/index.html.
[22] RFC793 J. Postel. Transmission Control Protocol. RFC 793, IETF, September 1981.
[23] RFC768 J. Postel. User Datagram Protocol. RFC 768, IETF, August 1980.
[24] RFC768 J. Postel. User Datagram Protocol. RFC 768, IETF, August 1980.
[25] Communication Networks/TCP and UDP Protocols - Wikibooks... en.wikibooks.org/wiki/...Networks/TCP_and_UDP.
[26] http://www.fujitsu.com/downloads/TEL/fnc/pdfservices/TCPIPTutorial.pdf.
[27] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext Transfer Protocol – HTTP/1.1," RFC 2616 (Draft Standard), Jun. 1999, updated by RFC 2817. [Online].Available: http://www.ietf.org/rfc/rfc2616.txt
[28] E. Rescorla, "HTTP Over TLS," RFC 2818 (Informational), May 2000. [Online]. Available: http://www.ietf.org/rfc/rfc2818.txt.
[29] R. Fielding, J. Gettys, and J.C. Mogul. RFC 2616: Hypertext Transfer Protocol – HTTP/1.1. IETF, 1997.
[30] SSL: Intercepted today, decrypted tomorrow, Netcraft, 2013-06-25.
[31] HTTP Secure - Wikipedia, the free encyclopedia.
[32] Duffield, N., Sampling for Passive Internet Measurement: a Review, Statistical Science, Volume 19, Issue 3, Pages 472-498, Institute of Mathematical Statistics, 2004 .
[33] Romaniak, P., Mu, M., Mauthe, A., D'Antonio, S., Leszczuk, M.: Framework for the Integrated Video Quality Assessment. In: 18th ITC Specialist Seminar on Quality of Experience, Blekinge Institute of Technology, Karlskrona, Sweden (May 2008).
[34] Serral-Graci`a, R., Barlet-Ros, P., Domingo-Pascual, J.: Coping with Distributed Monitoring of QoS-enabled Heterogeneous Networks. In: 4th International Telecommunication Networking Workshop on QoS in Multiservice IP Networks, Venice, Italy, February 2008.
[35] Paxson, V., E., Measurements and Analysis of End-to-End Internet Dynamics, PhD Thesis, University of California, Berkeley, CA, USA, April 1997.
[36] Amer, P., D., Kumar, R., N., Kao, R., Phillips, J., T., Cassel, L., N., Local Area Broadcast Network Measurement: Traffic Characterisation, IEEE Computer Society International Conference (Compcon'87), San Francisco, February 1987.
[37] Jain, R., Routhier, S., Packet Trains-Measurements and a New Model for Computer Network Traffic, IEEE Journal of Selected Areas in Communications (JSAC), Volume 4, No. 6, September 1986, pp. 986-995.
[38] Cáceres, R., Measurements of Wide Area Internet Traffic, Technical Report (CSD-89-550), University of California at Berkeley, USA, 1989.
[39] Cáceres, R., Danzig, P., B., Jamin, S., Mitzel, D., J., Characteristics of Wide- Area TCP/IP Conversations, ACM SIGCOMM Computer Communication Review, Volume 21, Issue 4, Pages 101-112, September 1991.

[40] Crowcroft, J., Wakeman, I., Traffic Analysis of some UK-US Academic Network Data, in Proceedings of The First Annual Conference of the Internet Society (INET'91), Copenhagen, June 1991.

[41] Leland, W., E., Taqqu, M., S., Willinger . W., On the Self-Similar Nature of Ethernet Traffic, ACM SIGCOMM'93, San Francisco, California, USA, September 13-14, 1993.

[42] Kevin .Thompson, G.J.Miller, R.;Wilder, Wide-area Internet traffic patterns and characteristics, Network, IEEE , Volume 11, Issue 6, December 1997, pp.10-23.

[43] M.Fomenkov, K.Keys, D.Morre, and K Claffy, Longitudinal Study of Internet Traffic in 1998-2003, on Proceedings of the Winter International Symposium on Information and Communications Technologies WISICT'04, January 2004.

[44] T.Karagiannis, M.Molle, and M.Faloutsos, Long-Range Dependence Ten Years of Internet Traffic Modeling, IEEE Internet Computing Sep-Oct 2004, pp.57-64.

[45] S. McCreary and kc claffy, "Trends in wide area IP traffic patterns: a view from Ames Internet eXchange," in *13th ITC specialist seminar: IP Traffic measurement, modeling and management*, Sept. 2000.

[46] P.Borgnat, G.Dewaele, K.Fukuda, P.Abry, and K.Cho, Seven Years and One Day: Sketching the Evolution of Internet Traffic, IEEE Infocom 2009 proceedings, April 2009, pp.711-719.

[47] C.Barakat, P.Thiran, G.Iannaccone, C.Diot, and P.Owezarki, A Flow-based Model for Internet Backbone Traffic, ACM SIGCOMM Internet Measurement Workshop, November 2002.

[48] C.Barakat and E.Altman, A Markovian Model for TCP Analysis in a Differentiated Services Network, INRIA, France, 2003.

[49] C. Fraleigh, S. Moon, C. Diot, B. Lyles, F. Tobagi, "Packet-level traffic measurements from a tier-1 IP backbone," Sprint technical report TR-01-110101.

[50] J. Micheel, I. Graham, N. Brownlee, "The Auckland data set: an access link observed," in *Proceedings of the 14th ITC specialists seminar on access networks and systems*, 2001.

[51] KC Claffy, G.Polyzos, and H.Braun, Tracking Long-Term Growth of the NSFNET, Communications of the ACM, vol.37, no.8, August 1994, pp.34-45.

[52] G.Fox, Peer-to-peer Networks, Computing in Science & Engineering, Volume: 3, pp.75-77, 2001.

[53] T.Karagiannis, A.Broido, N.Brownlee, KC Claffy, and M.Faloutsos, Is P2P Dying or Just Hiding, IEEE COMM Globecom, November 2004.

[54] A. Smith, Comcast prevails over FCC in Web traffic fight, CNNMoney.com, April 6th 2010, http://money.cnn.com/2010/04/06/technology/net_neutrality_fcc_comcast/, retrieved April 9th, 2010.

[55] K.Fukuda, K.Cho, H.Esaki, The Impact of Residential Broadband Traffic on Japanese ISP Backbones, ACM SIGCOMM, Volume 35, pp.15-22, January 2005.

[56] http://www.eg-climet.org/ES0702-e.pdf.

[57] Dowd, P.W.; McHenry, J.T., "Network security: it's time to take it seriously," Computer, vol.31, no.9, pp.24-28, Sep 1998.

[58] Introduction to Traffic Analysis by George Danezis University of Cambridge, Computer Laboratory.

[59] Network monitoring implementation guides and tutorials, http://wiki.debian.org/Network_Monitoring.

[60] N. Brownlee, C. Mills, and G. Ruth (1999-10). "RFC 2722 - Traffic Flow Measurement: Architecture". IETF. Retrieved 2010-02-11.

[61] J. Rajahalme, A. Conta, B. Carpenter and S. Deering (2004-03). "RFC 3697 - IPv6 Flow Label Specification". IETF. Retrieved 2010-02-11.

[62] J. Quittek, JT. Zseby, B. Claise, and S. Zander (2004-10). "RFC 3917 - IPFIX Requirements". IETF. Retrieved 2010-02-11.

[63] K. C. Claffy, H. W. Braun, and G. C. Polyzos, "A parameterizable methodology for Internet traffic flow profiling," *Selected Areas in Communications, IEEE Journal on,* vol.13, no. 8, pp. 1481-1494, 1995.

[64] D. Song, D. Wagner, and X. Tian. Timing Analysis of Keystrokes and SSH Timing Attacks In 10th USENIX Security Symposium, 2001.

[65] Edward W. Felten and Michael A. Schneider. Timing Attacks on Web Privacy. Proc. of 7th ACM Conference on Computer and Communications Security, Nov. 2000.

[66] Paul Syverson, Gene Tsudik, Michael Reed, and Carl Landwehr. Towards an Analysis of Onion Routing Security. LNCS 2009.

[67] Andrew Hintz, Fingerprinting Websites Using Traffic Analysis, Privacy Enhancing Technologies Workshop 2002, San Francisco.

[68] Qixiang Sun, Daniel R. Simon, Yi-Min Wang, Wilf Russell, Venkat Padmanabhan, Lili Qiu, Statistical Identification of Encrypted Web Browsing Traffic, 2002 IEEE Symposium on Security and Privacy, Oakland.

[69] SafeWeb, One Click Privacy Anywhere, Anytime http://www.safeweb.com.

[70] http://www.ask.com/question/example-of-Internet-application.

[71] Packet traces from wide backbone," http://tracer.csl.sony.co.jp/mawi/.

[72] K. Cho, K. Mitsuya, and A. Kato. Traffic data repository at the WIDE project. In USENIX 2000 Annual Technical Conference: FREENIX Track, pages 263–270, June 2000.

[73]http://www.nngroup.com/articles/law-of-bandwidth.

[74]http://www.radicati.com/wp/wp-content/uploads/2011/05/Email-Statistics-Report-2011-2015-Executive-Summary.pdf.

[75] "Microsoft timeline and profile". About.com Web Trends. Retrieved 2012-05-01. visited 16-03-2014.

[76] "First Look: The Microsoft Network, by Robert J. Ambrogi". Retrieved 2009-08-07. visited 16-03-2014.

[77] "About Skype: What is Skype?". Retrieved 28 July 2010.

[78] "Skype — A Baltic Success Story". credit-suisse.com. Retrieved 24 February 2008.

[79] "Happy Birthday Skype: Even monkeys use it now".*Emirates 24/7*. 28 August 2013. Retrieved 28 August 2013. visited 08.03.2014.

[80] Kazan. http://www.kazaa.com.
[81]Skype grows FY revenues 20% reaches 663 mln users.
http://www.telecompaper.com/news/ March 2011. Visited 08. 3. 2014.
[82] Skype – The Big Blog – 30 million people online on Skype.
http://blogs.skype.com/en/2011/03/30_million_people_online.html.
Visited 08. 03. 2014.
[83] http://www.msba.org/sec_comm/sections/solo/docs/Obtaining Records from Social Networking Websites.pdf.
[84] "Facebook Tops Billion-User Mark". The Wall Street Journal (New York). October 4, 2012. Retrieved October 4, 2012. Visited 10-03-2014.
[85] "Jim Breyer (via Accel Partners)". *CNBC*. May 22, 2012. Visited 10-03-2014.
[86] Kazeniac, Andy (February 9, 2009). "Social Networks: Facebook Takes Over Top Spot, Twitter Climbs". *Compete Pulse blog*. Retrieved February 17, 2009. . Visited 10-03-2014.
[87] "Facebook, Inc. Financial Statements". Securities and Exchange Commission. February 1, 2013. Retrieved February 1, 2013. Visited 10-03-2014.
[88] "Birthday boy Mark Zuckerberg to get $100bn gift". *The Times of India*. Associated Press. May 14, 2012. Archived from the original on May 14, 2012. Visited 10-03-2014.
[89] "Facebook squeaks onto the Fortune 500". *USA Today*. May 6, 2013.
Retrieved May 19, 2013. Visited 10-03-2014.
[90] Mark Milian and Marcus Chan (May 18, 2012)."Facebook's Valuation: What $104 Billion Is Worth".*Bloomberg Technology*. Retrieved January 11, 2014. Visited 10-03-2014.
[91] "Facebook Inc. Overview". Marketwatch. Retrieved January 30, 2014.
Dominic Rushe (January 29, 2014). Visited 10-03-2014.
[92] "Facebook posts record quarterly results and reports $1.5bn profit for 2013". *The Guardian*. Retrieved January 30, 2014. Visited 10-03-2014.

[93] Guo-Qing Zhang et al., *Evolution of the Internet and its Cores*, 10 NEW J. PHYSICS 123027, at 3 (2008).
http://iopscience.iop.org/1367-2630/10/12/123027 (concluding that the size of the Internet follows a form of Moore's Law, doubling every 5.32 years).
[94] Daniel F. Spulber & Christopher S. Yoo, Networks in Telecommunication:
Economics and Law 121 (2009).
[95] Bob Metcalfe, *Metcalfe's Law (the Blog) Has a POV: Point of View*, COCKRELL School of Engineering.
http://www.engr.utexas.edu/metcalfe/blog/blogpov (reproducing the original 1980 slide that led to Metcalfe's Law).
[96] Bob Briscoe et al., *Metcalfe's Law Is Wrong: Communications Networks Increase in Value as They Add Members—But by How Much? The Devil Is in the Details*, IEEE SPECTRUM, July 2006, at 35, 37.