Residential Network Traffic and User Behavior Analysis

YICHI ZHANG

KTH Information and
Communication Technology

Master of Science Thesis
Stockholm, Sweden 2010

TRITA-ICT-EX-2010:276

Residential Network Traffic and User Behavior Analysis

Yichi Zhang

yichi@kth.se
yichi.zhang@acreo.se

Supervisor in Industry: Andreas Aurelius

Acreo AB

Supervisor and Examiner: Prof. Gerald Q. Maguire Jr.

Royal Institute of Technology

Abstract

Internet usage is changing and the demands on the broadband networks are ever increasing. So it is still crucial to understand today's network traffic and the usage patterns of the end users, which will lead to more efficient network design, energy and costs savings, and improvement of the service offered to end users. This thesis aims at finding hidden patterns of traffic and user behavior in a residential fiber based access network. To address the problem, a systematic framework of traffic measurement and analysis is developed. It involves PacketLogic traffic data collecting, MySQL database storing, and traffic and user behavior analysis by using Python scripts. Our approach provides new insights on residential network traffic properties and Internet user habits of households, covering topics of aggregated traffic pattern, household traffic modeling, traffic and user penetration for applications, grouping analysis by cluster and subscriber, and concurrent application analysis. The analysis solutions we provide are based on open source tools without proprietary, giving the most flexibility for codes modification and distribution.

Sammanfattning

Internetanvändningen förändras ständigt och detta medför ökande och förändrade krav på bredbandsnäten. Detta medför att det är viktigt att förstå dagens nätverkstrafik och slutanvändarnas användningsmönster. Denna kunskap kan leda till mer effektiv nätverksdesign, energi-och kostnadsbesparingar, och förbättring av tjänster som levereras till slutanvändare. Denna rapport syftar till att hitta dolda mönster av trafik-och användarbeteende i ett fiberbaserat accessnät. För detta syfte har mätningar av IP-trafik samlats in Med hjälp av en kommersiell trafikhanterare, PacketLogic. Data har lagrats i en MySQL-databas. Analysen har utförts med MySQL och Python-script. Resultaten ger nya insikter om nätverkstrafik och hushållens Internetvanor hos, med avseende påaggregerat trafikmönstret, hushållsbaserad trafikmodellering, penetration av applikationer, klusteranalys och användaranalys. Lösningarna är baserade på open source-verktyg, vilket ger störst flexibilitet för modifiering och distribution av kod.

Terminology and Abbreviation

API	Application Programming Interface is an interface implemented by a software program to enable interaction with other software
CDF	Cumulative Distribution Function, describes the probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x
DHCP	Dynamic Host Configuration Protocol, an autoconfiguration protocol used on IP networks
DSL	Digital Subscriber Line, a network access technology using the telephone infrastructure
DNS	Domain Name System is a hierarchical naming system for computers, services, or any resource connected to the Internet or a private network
DB	Database
DW	Data Warehouse
FTP	File Transfer Protocol is a standard network protocol used to exchange and manipulate files over a TCP/IP-based network, such as the Internet
FTTH	Fiber-To-The-Home, a network access technology
GUI	Graphical User Interface, a user interface based on graphics
HTTP	HyperText Transfer Protocol is a communication protocol for the transfer of information on the intranet and the World Wide Web
IM	Instant Messaging, a form of real-time direct text-based communication between two or more people using personal computers or other devices, along with shared software clients
IP	Internet Protocol. An IP address is a numerical label that is assigned to devices participating in a computer network that uses the Internet Protocol for communication between nodes
ISP	Internet Service Provider is a company or business that provides access to the Internet
MAC	A Media Access Control address uniquely identifies a network adapter or an interface card
Module	A Python module is a file containing Python definitions and statements
MySQL	A relational database management system
Python	A programming language
P2PTV	Peer-to-peer applications designed to redistribute video streams in real time on a P2P network

P2P	A peer-to-peer network is a network of connected nodes that don't communicate according to the client-server model. All computers can act as a server and a client
QoS	Quality of Service refers to resource reservation control mechanisms rather than the achieved service quality
SQL	Structured Query Language, is a database computer language designed for managing data in relational database management systems
TCP/IP	The Internet Protocol Suite is the set of communication protocols used for the Internet and similar networks
TTL	Abbreviation for Time to live, is a limit on the period of time that a unit of data (e.g. a packet) can experience before it should be discarded
UDP	User Datagram Protocol is a protocol to send data over an IP-network
VoIP	Voice over IP, telephony over the Internet
VPN	A virtual private network is form of communication over networks that are public in ownership, but emulate a private network in terms of security

Table of Contents

Chapter 1

Introduction

Advances in computer technology and the Internet have changed the way America works, learns, and communicates. The Internet has become an integral part of America's economic, political, and social life.

– Former US president, Bill Clinton

1.1 Background

The Internet has become an essential part of many persons' daily lives, from ordering a song online to sharing files and documents at work, from chatting with an old friend to playing an online game in a given role. Residents pay to Internet service providers(ISP) for broadband service, and it is ISP's duty to ensure network resources can adequately match their customers' ever-increasing demand for Internet applications such as web browsing, Skype[1], WoW[2], etc. Service providers and researchers are always interested in seeking out the variety of underlying traffic demands, finding the way to fully exploit current network resources and maximize the profits. In this thesis, we will discuss these issues from an empirical aspect, delve into patterns of the residential network traffic and user behavior as well.

Residential network, as one part of the Internet, is a mature service in many countries.

[1] Skype is a software application that allows users to make voice calls over the Internet.
[2] World of Warcraft, often referred to as WoW, is a massively multiplayer online role-playing game (MMORPG)

However, the users in residential networks usually have different goals than those in other environments. There are no strict regulations for them to download and upload, compared with campus network and enterprise network users. In most situations, the only limitation for them is the bandwidth they have, according to their subscription. Research on residential network is not easy to find as researchers rarely have large-scale access to residential network. Since users have the right to choose available ISPs, it may be difficult to coordinate among ISPs to get necessary data.

On the other hand, measuring and analyzing the Internet traffic in itself is not an easy job. Many obstacles exist as listed below.

a. The Internet is a huge system. It has been expanded from one link in 1969, as a part of the ARPANET research network[1], to 1.7 billion of users in 2009[2] (one third of the population of the world). Besides its enormous scale, complex physical structure adds the difficulty of measuring the Internet traffic. Although OSI-7 layer model and TCP/IP protocols are widely adopted in today's Internet, the underlying infrastructures are deployed and maintained by independent organizations, which means it is difficult to coordinate among those responsible companies to collect necessary data.

b. Privacy issues arise. People worry about the privacy abuse since all consumers' data go through their ISP's network, which means the service provider has the capability to acquire detailed data from their consumers. Usually ISPs do make a commitment like what BT[22] and Verizon[23] have done, they collect information just for billing, marketing, or improving their connectivity service. And they commit that they do not identify personal details. However, sometimes ISPs state that they will make their information available for third parties such as the government because of law enforcement. For example, according to USA Patriot Act, ISPs should give up more user information[24]. In Sweden, three Internet Service Providers, among them Tele2, one of the country's three major broadband operators, have stated that they will erase traffic data to protect their customers' privacy. This move is regarded as reaction to the IPRED[3] law[25], which gives copyright owners the right to ask for customers' identity from ISPs, if a court agrees. As a conclusion, it is difficult to solicit ISPs to provide traffic information. The data set should not contain sensitive information but be meaningful enough so that traffic could be analyzed.

c. Ever-growing encrypted data adds more challenge to the traffic analysis. Although encryption means more operation steps and less-efficient traffic transmission to the user, there still exist plenty of reasons for using encryption. The Internet is not secure. It is possible to tap into the network, and capture or maybe even change the information which is sensitive or confidential. Companies obviously do not expect this would happen, they want all their employees' emails and documents remaining intact and uneavesdropped. E-commence also required confidential network to have safe transaction. Besides, some P2P users attempt to encrypt their file-sharing traffic as they do not want their P2P traffic identified by ISPs, which may lead to ISPs throttling this traffic.

[3] Intellectual Property Rights Enforcement Directive, which is a highly contested anti-piracy law. IPRED took effect in Sweden on Apr 1st ,2009.

d. Captured data may be not what it is supposed to be. It is not always possible to directly measure quantities of interest, if captured data is only components in aggregate measurements[26]. For example, troubleshooting packet loss requires knowing network performance on individual links, while in practice it may only be feasible to measure performance between two hosts, that is, the composite performance along a path that comprises several links. Another example is measuring the household's traffic may exclude the traffic between neighboring households, as the traffic collector may stand inside the ISP's network instead of household's access point. If households are using P2P or direct file transfer program, some direct traffic between them is not detected by the measurement equipment.

e. Too much data problem. The volume which is collected could be enormous. A single high speed network interface could in principle generate hundreds of gigabytes of (unsampled) flow statistics per day if fully utilized[26]. So it leaves a question whether to keep all of the data or use certain ways to restore featured data, which is adequate for analysis.

When we analyze network traffic, we also bring user behavior into concern. As a matter fact, Internet user behavior is constantly changing. It is easily influenced by individual's habits, income, or other personal reasons. Moreover, people may change their behavior in response to some new services' emerging, like Facebook[4] and YouTube[5]. Young people today may spend more hours on solely Facebook than all web-surfing time three years ago. However, research on the aggregated level of Internet user behavior is of great importance for service providers, network operators and , and social scientists. Internet user behavior will reflect the social customs of an epoch. And vice versa, the society can influence one's behavior on the Internet.

1.2 Motivation

The importance of residential traffic analysis affects many fields. Commercial service provider can optimize their network based on their own measurement result, for example, in order to give QoS guarantees for home users. Researchers can utilize the data to design more efficient application or protocols for residential network. In social science field, statistics can be learned to know about network users' behavior and insights into social issues, e.g. the effect of the IPRED Law.

1.3 Overview of thesis

The thesis is performed in VINNOVA[6] Broadband User Behavior Project, while some parts of the thesis is performed in the project TRAMMS[3] (Traffic Measurements and Models in

[4] Facebook is the most popular social networking website today. http://www.facebook.com
[5] YouTube is the most popular video-sharing website today. http://www.youtube.com
[6] http://www.vinnova.se/

Multi-Service Networks), which is coordinated by Acreo AB[4] with 11 partners from Sweden, Spain, and Hungary. In this thesis project, data from a Swedish commercial municipal network and a testbed network is analyzed. This thesis focuses on developing desirable methodology of collecting data and data analysis. With advanced traffic measurements on the application level, more features of the residential traffic and user behavior are expected to be disclosed.

1.4 The reader

The reader is assumed to be familiar with the Internet and TCP/IP suite. It is also assumed that the reader has basic knowledge of telecom network and devices. And it is helpful to understand the procedure how our results are generated if the reader knows fundamentals of SQL and programming language.

Chapter 2

Previous Work

Here we divide them into three sub-sections. Nevertheless, they may interleave in different researches.

2.1 Internet traffic measurement and analysis

Measurement and analysis of Internet traffic is important to get knowledge about the characteristics of the traffic, thus it has drawn a lot of researchers' attention. Some researchers made long-term investigation of Internet. Borgnat and his team collected internet traffic data for 7 years in order to sketch the evolution trend of the internet traffic[5]. Fomenkov and Keys have investigated more than 4000 traces from 1998 to 2003, to find relations between bit rates and traffic statistics[6]. Traffic pattern is a very clear and typical way to display traffic variation within the recording period, like what K.Thompson and G.J.Miller did[17]. They had 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.

It is of great interest to find analytical models for describing Internet traffic. Karagiannis, Molle, and Faloutsos in their paper[7], reviewed 10 years development of Lang-Range Dependence(LPD) theory and use LPD to model the complex traffic of the Internet. However, more researchers believed that, instead of bit and packet rate, flow level traffic would be better used for explaining the intrinsic characteristics of Internet. Barakat et al. analyzed TCP flow by means of Markovian model in a differentiated service network[9]. 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[8], because traffic on a link is the result of a high level of multiplexing

of numerous flows whose behavior 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.

More people would like to analyze the Internet from the view of application. Back in the nineties, FTP and Mail accounted for half of the traffic volume, until HTTP becomes the majority[11]. And the invention of Peer-to-Peer(P2P) nearly toppled the pattern of Internet traffic, it could be considered as "killer Internet application"[12]. 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[14]. 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[13], which may cause inaccurate P2P traffic measurement.

2.2 Residential network

Residential network characteristics have been studied in some countries.

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[15]. 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[7] or RRDtool[8], 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. And, the residential traffic accounts two-thirds of the ISP backbone traffic, which means that backbone traffic is dominated by the residence behavior.

Gregor Maier et al. [27] have described passive, anonymized packet-level observations from monitoring the network activity for more than 20,000 residential DSL customers in a German urban area. They have deployed Endace DAG[9] network monitoring cards at aggregation points for traffic capture, and identify packets by using Bro system[10]'s Dynamic Protocol Detection (DPD). DPD tries to parse each byte stream with parsers for numerous protocols, deferring determination of the corresponding application until only that application's parser recognizes the traffic. Gregor and his team study the traffic in 2008 and 2009, and discover that HTTP traffic dominates instead of P2P. They infer that HTTP's domination may be due to ever-increasing usage of video portals such as youtube.com and news sites, which contributes 25% of all HTTP traffic.

[7] A tool to monitor the traffic load on network-links. http://oss.oetiker.ch/mrtg/
[8] An open source tool for storage and retrieval of time series data. http://oss.oetiker.ch/rrdtool/
[9] Data Acquisition and Generation Technology, which is developed by developed by New Zealand company Endance. http://www.endace.com
[10] http://www.bro-ids.org

The shortage of using Bro's DPD is not well capable of identifying P2P traffic, thus causing a large portion of unclassified application in which some are using well-known P2P ports. Gregor also find that for many TCP connections the access bandwidth-delay product exceeds the advertised window, which makes it impossible for the connection to saturate the access link.

2.2 User behavior

It is very common that people use survey and questionnaire to learn user behavior. Liu and Day[16], in 2002, has made an island-wide telephone survey in Taiwan, to find out the "globalness" of Taiwan users' Internet behavior, and the factors contributing to these patterns of use. World Internet Institute[11] is the home of study "Svenskarna och Internet", which is an annual quantitative survey interviewing 2000 people as a representative sample of the Swedish population.

Making different categories is a common approach for user behavior study. Clustering method is one way which could be effective identifying those subgroups. Yamakami has made a research [18] based on k-means clustering[12]. In order to analyze the user behavior of the mobile Internet, he first performs a preparatory survey to label three different kinds of mobile user behavior (a) always active (b) irregular (c) prime time use. Then he develops a certain way to establish a data set by using content access logs from a mobile service provider. After that, Yamakami applies the k-means(k=3) to have 3 user clusters, which represent the assumed three different user behavior. Thus he has obtained main behavior models as groups, and makes further research to summarize each user groups' profiles with the attributes.

D.Yinan et al. [20] also use k-means method, which helps to find clustering subgroups in a Chinese dial-up broadband network. They have employed an elegant method to specify k value. They have defined C_{intra} and C_{inter} as coefficient of variation of intra cluster distance and coefficient of variation of inter cluster distance respectively. And β_{cv} is defined as ratio between the intra and inter cluster variance C_{intra} / C_{inter}. We can easily infer that β_{cv} turns smaller when k becomes larger. However, k cannot be too large as for the purpose of clustering is to select appropriate user classification. Most ideal k can be obtained when β_{cv} becomes relatively stable. By using this method, Yinan et al. have got the k value as 10, and found each group's attributes and made comparison among them.

2.4 Previous work within Acreo

Several master students have contributed to the research on traffic measurement and analysis. Tomas Sisohore from KTH examined various traffic parameters (flow, bandwidth, packet size and

[11] http://www.wii.se

[12] K-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

inter-arrival-time) for traffic measurements. And he also took a lot of efforts on HTTP media stream statistics, drawing graphs of intensity, activity and volume. He has given a broad and comprehensive insight for HTTP stream traffic. There's one limitation for Sisohore's work, that is, he used IP address instead of household to go through his research. As the network he investigated was running DHCP, he had no way to learn from a user-oriented aspect. But another master student, Tomas Bonnedahl, in his report has fixed this problem. Bonnedahl has looked at user penetration in regard to applications and traffic volume distributions. He also made comparison between two networks, DSL households and FTTH households. After doing data collection and analysis, he has confirmed the assumption that symmetric connections with high capacity (FTTH) are preferred peers in P2P network. Conclusion that upper 10% of high bandwidth users transferred large amounts of data was also drawn in Bonnedal's report. It is worth pointing out that the DSL network consisted of just about 100 households. This figure is comparatively small so that it makes the results less reliable. Therefore this study should be compared to other studies conducted on similar networks in case of taking as a whole conclusion. Since Acreo use PacketLogic[13] which has deep packet inspection function for traffic measurements, much research has covered application level. Phat Hoang and David Herrero from Lund University delved into different application-level traffic, such as P2P, gaming, streaming multimedia, online music, etc. And they have presented very interesting view of those traffic patterns. Hoang has also displayed IPRED Law's impact on P2P and streaming video traffic.

[13] Further description can be found in next chapter.

Chapter 3

Network Traffic Measurement

In this thesis project, all the traffic and user behavior analysis is based on the information collected from traffic measurement. A thorough description of our network will be given in this chapter, and we will explain where and how the data is collected.

3.1 Overview of the network architecture

The network we study during this thesis is a municipal IP access network in Sweden. The network offers broadband Internet access as well as some other services (such as IPTV) to its customers who are the local residents. This residential network is open-fiber based, which means that the customers can independently and freely choose among the services offered by the different ISPs who connect to the network. DHCP is deployed to allocated IP addresses for all the households. Usually the lease time for IP addresses varies with the service provider with a minimum value of 20 minutes[], which means during 20 minutes period one active IP represents one active household.

Most of the household covered by our research are using FTTH (Fiber To The Home) broadband access. Approximately 2600 FTTH households are included in our measurements. Those FTTH households are spread all over the town, which are consisted of different social groups with Swedish or foreign background. The maximum downloading speeds for them range from 1 Mbps to 100 Mbps, depending on which subscription customers have chosen. The broadband speed could be asymmetric, i.e. the uplink bandwidth is lower than the downlink one. It is mostly common that customer has a 10/100 Mbps upload/download speed.

The network also has about 200 households connected with DSL (Digital Subscriber Line). The DSL households live in two traditional Swedish residential areas with private owned houses. Therefore the DSL households are a rather homogenous group with mostly ethnic Swedish background and culture. All households in this group have a 24Mbps downstream rate and 3Mbps upstream.

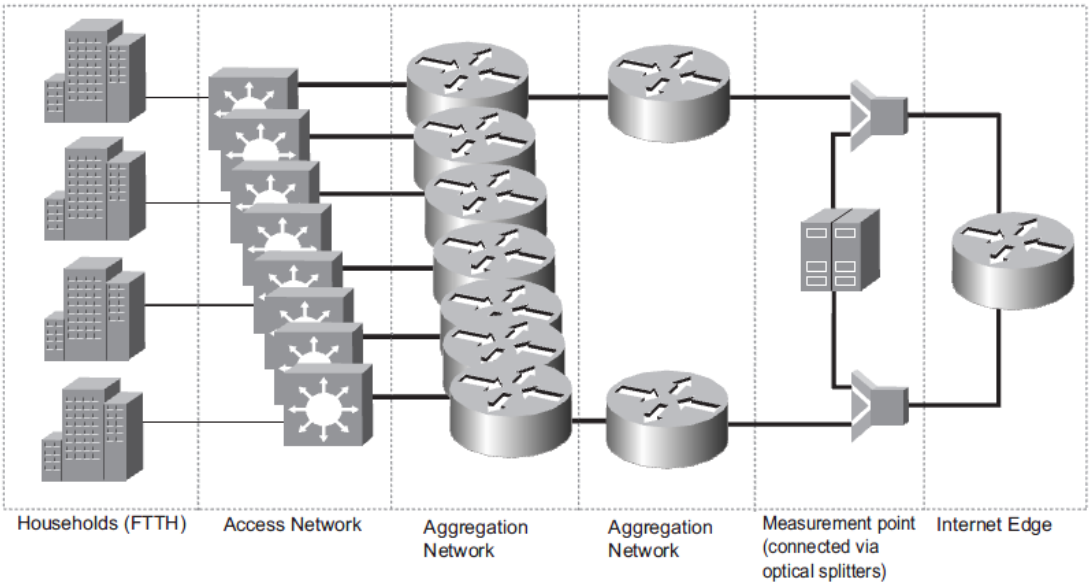


Figure 3.1 Municipal network architecture[28]

Figure 3.1 has shown the architecture and the measurement points for the municipal network. The thin lines denote 1 Gbps links while the bold lines denote 10 Gbps links. FTTH households' access networks are aggregated to connect the Internet Edge Point, where the service providers are connected to the network. The measurement equipment is deployed in the middle of aggregation network and the Internet Edge. It is using optical 50/50 splitters, which merely split the optical signal into two exact copies, so that the traffic in the network will not be affected by the measurement device. Since there are two redundant links to this node, measurement hardware with two physical GB Ethernet channels has been used. Details of the measurement equipment will be given in the next section.

3.2 Measurement equipment

The measurement is performed by using PacketLogic (PL), a commercial real-time hardware/software solution which is used mainly for traffic surveillance, traffic shaping or as a firewall. But PacketLogic can also be used for traffic analysis and for gathering statistics about the traffic. It is a rack-mounted system which can be equipped with different interfaces (The standard con-figuration is delivered with Fast Ethernet and Gigabit Ethernet interfaces).

PacketLogic system can be regarded as a transparent Layer 2 appliance. And it acts on Layer 2 to Layer 7 to manage today's complex traffic. Its self has a statistics database that collects and stores data. It records the average amount of traffic in the inbound and outbound directions as well as the total traffic for all nodes in the network.

Deep packet inspection (DPI) is the feature function for PacketLogic. It uses the Datastream Recognition Definition Language (DRDL) to identify different application protocols, which is based on a broad range of criteria instead of merely port definitions. The identification relies on bidirectional information like the packet sequence in a handshake, header information, protocol, actual payload, and other distinguishing characteristics of an application. This way DRDL can properly identify even encrypted applications. Some applications create multiple flows. DRDL interconnects control and data sessions of protocols like FTP. During the identification process DRDL aggregates detailed traffic properties like MIME-type, filename, chat channel and SIP caller ID. This granularity enables the administrator not only to see the Xbox Live traffic, but rather the Xbox Live users who are playing Halo 3. Till Feb 2010, the DRDL database consists of more than 1600 signatures[29].

PacketLogic also has a Python API that can enable programmatic configuration and retrieval of data from the system. This allows everything from small scripts retrieving statistics data to large programs integrating the PacketLogic with other systems to be written. In this thesis, Python API for PacketLogic is utilized for extracting specified data from PL database system and then exporting to external data warehouse.

3.3 Data warehouse

Data warehouse is a certain type of database. As our database is mainly designed for facilitating querying and analysis, it would be more accurate to call it data warehouse (DW) rather than a regular database. The purpose of a database is to record, while that of the data warehouse is to respond to analytical questions. The establishment of a data warehouse is based on regular database, however, DW usually contains read-only data, which is specially picked and never changed since the day it is stored.

Since data warehouse is a certain type of database, in this report, these two terminologies is interchangeable.

3.3.1 Database facilities

We choose MySQL as the database system to store and organize all the data, because MySQL is a full-featured database system and it is open sourced at extremely low costs as well. Our

MySQL database is established on a Ubuntu Linux distribution server, which has 4 CPUs to support all kinds of queries and data manipulation.

There are several ways to connect the MySQL database.

1. Use third-party proprietary or free graphical administration applications (such as MySQL Workbench[14]). MySQL Workbench frontend is an official integrated environment that enables users to graphically administer MySQL databases and visually design database structure. The GUI for MySQL Workbench is very user-friendly, which can automatically check the grammar error of queries. The shortage of Workbench is comparatively slow performance due to excessive memory usage of client machine.
2. Use telnet to log into the host and use the MySQL client programs (mysql, mysqladmin, mysqldump, etc). It is a more direct way to interact with the database server.
3. Use MySQL API to access the database. As for the prevalence of MySQL, API for different programming languages is well-developed and available. With the help of MySQL API, user can customized his own programs and manipulate the database batchly.

During the thesis, all above the methods are utilized for different purposes. MySQL Workbench is used for checking correctness of SQL queries; Telnet and remote login is usually made to get preliminary results; and we write Python scripts (which will import MySQL API module) to further analyze the data set.

Since June 2007, we have recorded and stored a huge scale of traffic data of about 2800 households in one MySQL database. The total disk usage of the database is more than 100 GB so far. Database establishment is the groundwork for all further analysis.

3.3.2 Data structure and content

This section is divided into two parts as we have use two databases of different structure for performing analysis. The latter one is an upgrade for the previous one with better organized structure, although the two databases contain the same amount of information.

Here we first start the description of the old data warehouse.

3.3.2.1 Old data warehouse

With the help of PacketLogic Python API, Internet traffic data is obtained and stored to construct one table which is called PL. Table 3.1 is an extract from PL table. It contains information about date and time, ip address, application and protocol (which is categorized and recognized by the PacketLogic signature database), and incoming and outgoing traffic volume.

Every row of PL table is a record for a certain service for one IP. The granularity of the record is 5 minutes. If one application or protocols are detected by PacketLogic during this 5

[14] MySQL Workbench is a visual database design tool that integrates SQL development, administration, database design, creation and maintenance into a single, seamless environment for the MySQL database system. http://wb.mysql.com/

minutes, it will be recorded in PL table. The traffic volumes for trafficIn and trafficOut are calculated in the unit kbps, which is an average value for the 5 minutes duration. Even if the value is too small to count as zero, the entry will be kept in record. Nevertheless, if there is no traffic of certain application in that 5-min interval, nothing will be added into the database. IP addresses are hashed in PL table to comply with privacy laws. The hashed digest is anonymized and irreversible, but still unique to distinguish with each other for analysis.

Table 3.1 An extract from PL table

datetime	Ip	service	trafficIn	trafficOut
2009-04-01 19:00:00	2bf2a5eb9ec0f2d51bcdfe2154702630	HTTP	253	8
2009-04-18 14:55:00	011107069e800aa678b959cc75bd8682	BitTorrent encrypted transfer	290	1691
2009-04-18 14:55:00	011107069e800aa678b959cc75bd8682	MSN messenger	0	0

As DHCP is used for our municipal network, IP addresses are shared with the households. So another table BECS which contains logs from DHCP servers is created. Table 3.2 is an extract from BECS table. It contains fields of date and time, ip address (hashed for privacy reason), households MAC address, the switch id and port number which directly connect to the household, and subscription information which implies customers' bandwidth.

Table 3.2 An extract from BECS table

datetime	ip	mac	aswitch	aport	iservice
2009-04-01 19:00:00	2bf2a5eb9ec0f2d51bcdfe2154702630	0050.22a2.3897	fogd4a-01	fastethernet12	ISP1-subsription1
2009-04-18 14:55:00	011107069e800aa678b959cc75bd8682	0018.4d2f.c6ed	salt1-01	fastethernet13	ISP2-subscription2

We can see from Table 3.2, fields of datetime and ip are the same as that in PL table. MAC address uniquely identifies a network adapter or an interface card which is used by the customer. And the switch id, together with port number, can indicate one specific household regardless of changing of IPs. The last field of BECS is customers' subscription information, which brings a large advantage for our research. We can study one user's behavior bearing in mind that which subscription he or she is using.

There is one flaw for BECS table. As BECS is extracted from DHCP log and converted into 5 minutes interval to match PL table's interval, it remains the possibility that one households' data is missing during the 5 minutes interval while the IP is released and allocated to another household by DHCP. Although it is usual that DHCP lease time is more than 20 minutes, as mentioned in 3.1, we find a few entries in BECS that IPs are relocated to different households within 20 minutes. The occurrence is insignificant. Comparing with large amounts of data we have, we consider it statistically negligible.

Two tables can be linked together by tying rows from both BECS and PL. Queries matching the datetime field and IP address field at the two tables (in SQL, that is 'bes.datetime =

pl.datetime and becs.ip = pl.ip' in WHERE clause) will result in a per household basis approach. Figure 3.2 has shown this relation between BECS and PL tables.

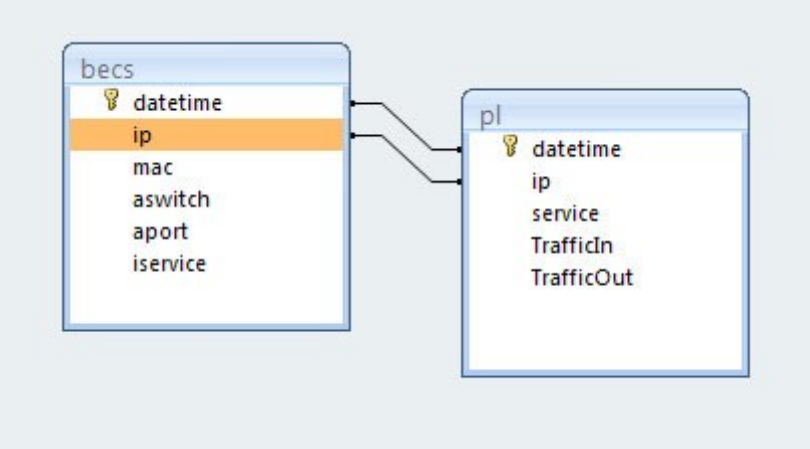


Figure 3.2 Link between BECS and PL

3.3.2.2 New data warehouse

In order to speed up the queries' time and save disk space, one new data warehouse for traffic analysis is established in June 2010. The super large table BECS and PL are disparted into relatively small one, as shown in Figure 3.3.

BECS is renamed as ip. Columns (aswitch, iservice and aport) are stripped off and replaced by id number (as_id, is_id and ap_id). Two columns (begin and end) are added, which will indicate the internet using time (ip allocating and leasing time) for current ip user. An column (id) gives every row in this table an identifying number.

PL, on the other hand, is renamed as acreo. Column (service) is also taken out and substituted by id number (s_id). Column (id) also functions as identifying each row in acreo. In addition, one column (ip_id) is mapping the column (id) in ip table, which builds relationship with the two tables. By using SQL query ('acreo.ip_id = ip.id' in WHERE clause), results can be gained in per household approach.

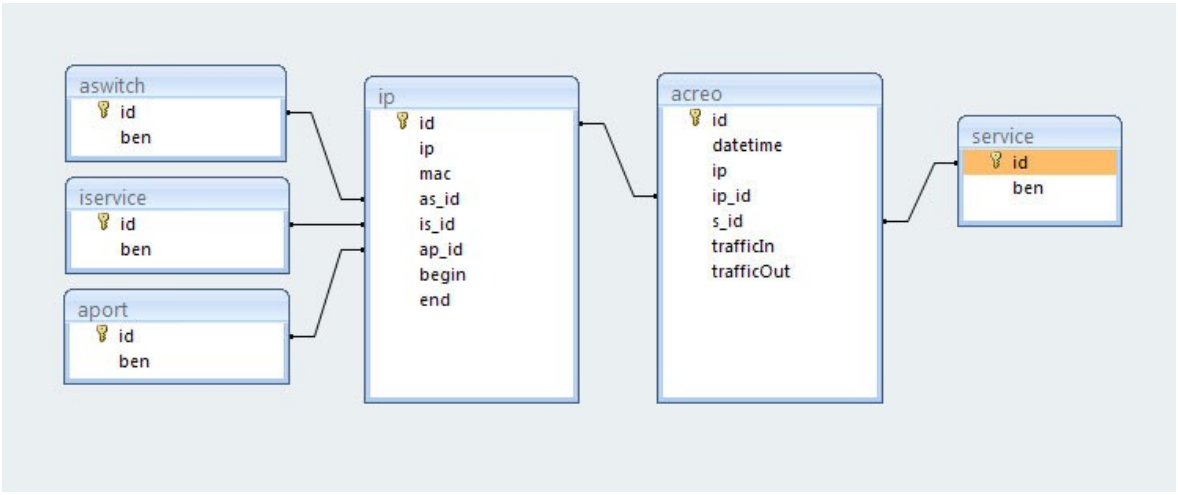


Figure 3.3 New database structure

Chapter 4

Method for Traffic and User Behavior Analysis

The main content for this chapter is method designed for traffic and user behavior analysis. A brief introduction of tools will be given in the first part, which is followed by a series of experiment methods.

4.1 Analysis tools

Based on all information which is stored in our data warehouse, data mining method is implemented to extract patterns and to obtain features for the huge residential traffic. Several analysis tools are involved in our research. Script language Python is majorly utilized throughout the research to realize the data mining procedure. With the help of powerful Python API (MySQL and Matplotlib), all procedures in Python programs, from collecting data to analyzing and drawing diagrams, are integrated to generate results for analysis.

A. Python

Python is a remarkably powerful dynamic script programming language which is used in a variety of application fields. It is chosen as our major analyzing programming language for the following reasons.

• Python is Open. Unlike MATLAB[15], Python and most of its APIs are under an open source license which makes them freely usable and distributable. For Acreo AB, it is better to use an open source analyzing tool so that all methods and results could be regenerated and distributed without any proprietary limitation.

• The language itself is powerful. Python uses a clear and readable syntax. It also supports high level dynamic data types, which makes the work easier in our data analysis.

• Python language is easy to learn as for its complete documentation and friendly user community.

• Python embraces a variety of well-support API (Application Programming Interface), which is also called modules in Python. Despite the extensive standard libraries developed by Python team, third party modules are also mature for virtually every task. As mentioned in previous chapter, PacketLogic API helps to dump all traffic data to our databases. Furthermore, in our analysis process, we have used MySQL API to communicate with the database, Matplotlib to handle scientific calculation and analysis.

• It is easy to port Python programs between platforms. All codes could possibly remain unchanged to be executed in Windows or in Linux. In our research, codes during development are usually written and tested in Windows for friendly user interface, and mature codes are executed in Linux distribution server to generate analysis results for high performance.

In order to make programming efficient, we use open-source Eclipse[16] IDE (integrated design environment) for authoring, modifying, compiling, deploying and debugging. Eclipse is originally designed for Java[17] assistance, but we can use PyDev[18] which is a third-party plug-in for Eclipse. There are also some other IDEs which support Python programming, such as IDLE[19], Netbeans[20], Eric Python and so on. Eclipse with PyDev is chosen for the reason that it has powerful debugging capability, friendly user interface, and it is with open source license.

B. Python MySQL API

Python MySQL API —— MySQLdb[21], is the interface we use in Python to communicate with the popular MySQL database. In Python codes, simply using the following statement can realize the module-import procedure.

```
import MySQLdb
```

The simplest possible database connection is:

[15] MATLAB is a numerical computing environment and fourth-generation programming language. MATLAB allows matrix manipulations, plotting of functions and data and other function. It is developed by The MathWorks, Inc. http://www.mathworks.com/products/matlab/

[16] http://www.eclipse.org/.

[17] Java is one of the most popular programming languages. http://www.java.com/en/.

[18] Pydev is a Python IDE for Eclipse. http://pydev.org/

[19] IDLE is a Python IDE built with the tkinter GUI toolkit, which is supported by Python Software Foundation. http://docs.python.org/library/idle.html

[20] http://netbeans.org/

[21] http://sourceforge.net/projects/mysql-python/

```
conn = MySQLdb.connect (host = host_name,
                        user = user_name,
                        passwd = password,
                        db = db_name)
```

This creates a connection to the MySQL server running on the remote or local machine using the standard UNIX socket (or named pipe on Windows), login name, password, and use a database which specifies a database name as a variable.

To perform an SQL query, a cursor is needed within MySQLdb. The cursor object contains the execute method, which is used for running SQL languages, such as select, add entry and modify entry. One example of selecting query is demonstrated as below.

```
cursor = conn.cursor()
query = "select * from pl limit 5"
cursor.execute(query)
result = cursor.fetchall()
for row in result:
   print row[0],row[1],row[2],row[3],row[4]
```

It can be seen that the example uses a simple 'select' query, to find out 10 rows of content of pl table. 'cursor.execute(query)' is the step to execute the query while 'result' records the position for a sequence of data rows. A 'for' loop is used to extract data from the query result, and print them out. The program's output shows below.

```
2007-06-01 00:00:00 000f6cbafcd16e74e47f22caf2b22c75 Undetermined 0 0
2007-06-01 00:00:00 0012b19d979ad792d80757691df7021a Undetermined 0 0
2007-06-01 00:00:00 0013ceb139d9458e3b768ac9081ee766 Undetermined 0 0
2007-06-01 00:00:00 001f16b23d3965b1f6aa28621a9ae7cf Undetermined 0 0
2007-06-01 00:00:00 0025c63a90fdca03c77bb14fef1b62f2 Undetermined 0 0
```

We compare the result with the output by using MySQL client, which execute the command 'select * from pl limit 5' directly.

```
mysql> select * from pl limit 5;
+---------------------+----------------------------------+-------------+----------+-----------+
| datetime            | ip                               | service     | trafficIn | trafficOut |
+---------------------+----------------------------------+-------------+----------+-----------+
| 2007-06-01 00:00:00 | 000f6cbafcd16e74e47f22caf2b22c75 | Undetermined |        0 |         0 |
| 2007-06-01 00:00:00 | 0012b19d979ad792d80757691df7021a | Undetermined |        0 |         0 |
| 2007-06-01 00:00:00 | 0013ceb139d9458e3b768ac9081ee766 | Undetermined |        0 |         0 |
| 2007-06-01 00:00:00 | 001f16b23d3965b1f6aa28621a9ae7cf | Undetermined |        0 |         0 |
| 2007-06-01 00:00:00 | 0025c63a90fdca03c77bb14fef1b62f2 | Undetermined |        0 |         0 |
```

```
+-------------------+---------------------------------+-------------+----------+-----------+
5 rows in set (0.00 sec)
```

Apparently the two approaches return the same results. Thus ensures we may utilize the convenience of Python script to make further traffic data analysis.

C. Matplotlib

Matplotlib is a module for making 2D or 3D plots of arrays in Python. Although it has its origins in emulating the MATLAB graphics commands, it is independent of MATLAB, and can be used in a Pythonic, object oriented way. Furthermore, unlike MATLAB, Matplotlib is open and could be distributed freely.

Matplotlib makes a heavy use of NumPy[22], which is a scientific-computing Python extension module featuring in N-dimensional array and matrix type. NumPy benefits our research a lot because much of our implemented analysis is based on large arrays and matrix. We can directly use built-in functions of Numpy while making calculating.

Except for powerful calculating, Matplotlib also embraces high performance of figure plotting. By using Matplotlib, one can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc. The most convenient thing is that, we can combine MySQL API and Matplotlib together, and write merely one script to realize the task of extracting data from SQL server, processing and analyzing data, and plotting them in a figure. There is no need to do the numerical work in one program, save the data, and plot it with another program. Compared with MATLAB, it is more suitable for large scale of data processing and fast scripting.

4.2 Data analysis method

Most of our analysis work uses captured data stored in the data warehouse. The captured data includes data from 2007 until the present year 2010. To better organize the experiments operated, here we divided them into six categories: aggregated traffic pattern, household traffic model, application traffic pattern, user penetration rate for applications, clustering, and current application analysis.

[22] http://www.numpy.org/

In this chapter, we give the models and methods associated with these categories. Results and discussion are in Chapter 5.

4.2.1 Aggregated traffic pattern

Aggregated traffic pattern analysis mainly focuses on an overview of network traffic characteristics. From an ISP's point of view, it is important to have an idea of how traffic varies during a day, what is the trend in overall traffic, etc. We have divided our investigation of aggregated network traffic into three lengths of periods. They are long-term, one-week and one-day, each of which will be discussed separately.

4.2.1.1 Long term traffic pattern

Thanks to the long term record over which data has been collected and stored in our data warehouse, it is possible to gain some insights into a long-term insight of network traffic patterns. For example, all households' traffic data can be summed and plotted to see how the Internet is developing over a long period. Below is the SQL script which is used for calculating traffic over a given interval.

```
mysql> select datetime, sum(trafficIn)/8/1024/1024*300 as TrafficInGB, sum(trafficOut)/8/1024/1024*300 as TrafficOutGB from pl2009 where datetime >= '2009-04-01 00:00' AND datetime <= '2009-12-31 23:55' group by date(datetime);
```

In the above query, all applications' traffic is accumulated together to have the total traffic volume. It can be seen from the query that traffic amount is divided by 8 and 1024^2 and then multiplied by 300. This converts units from mbps to GB, as the original data is sampled every 5 minutes. Both incoming traffic and outgoing traffic (to the client, respectively downloading and uploading traffic) will be selected, thus we can have a rough idea of the amount of bidirectional traffic and total bandwidth consumption.

4.2.1.2 Single week traffic pattern

First we select a week's worth of data to examine the daily traffic pattern, to see difference between weekdays and weekends.

```
mysql> select datetime, sum(trafficIn)/8/1024/1024*300 as TrafficInGB, sum(trafficOut)/8/1024/1024*300 as TrafficOutGB from pl2009 where datetime >= '2009-09-07 00:00' AND datetime < '2009-09-13 23:55' group by dayname (datetime);
```

This is a similar query to the long term one, but it only considers a period of one week. This could be repeated for several different weeks to see that whether there is a certain weekly pattern based upon a series of single-week traffic experiments.

An average one-week traffic pattern can also be calculated in order to compare to the traffic of individual single-week patterns. The SQL query is shown below. The standard deviation is also introduced to give a thorough idea of how much each weekday varies.

```
mysql> select datetime, std(TrafficInGB), avg(TrafficInGB), std(TrafficOutGB), avg(TrafficOutGB), dayname(datetime) from (select datetime, sum(trafficIn)/8/1024/1024*300 as TrafficInGB, sum(trafficOut)/8/1024/1024*300 as TrafficOutGB, EXTRACT(YEAR_MONTH from datetime) as yearmonth, EXTRACT(DAY from datetime) as day from pl2009 where datetime >= '2009-04-03 00:00' AND datetime <= '2009-12-31 23:55' group by yearmonth,day) as A group by dayofweek(datetime) order by dayofweek (datetime);
```

4.2.1.3 Single day traffic pattern

Single-day traffic pattern is computed as hourly records for one day. An experiment can be performed on randomly chosen days, to see whether there exists a general pattern or not. An average single-day's hourly traffic pattern will be computed. The SQL query to extract data for a single-day's experiment is:

```
mysql> select datetime, sum(trafficIn) from pl2009 where datetime >= '2009-09-01 00:00' AND datetime < '2009-09-02 00:00' group by hour(datetime) ORDER BY datetime;
```

Like the previous section, we calculate the an average one-day hourly traffic pattern with standard deviation.

```
select datetime, std(TrafficInGB), avg(TrafficInGB), std(TrafficOutGB),avg(TrafficOutGB) from (select datetime, sum(trafficIn)/8/1024/1024*300 as TrafficInGB, sum(trafficOut)/8/1024/1024*300 as TrafficOutGB, EXTRACT(YEAR_MONTH from datetime) as yearmonth, EXTRACT(DAY_HOUR from datetime) as dayhour from pl2009 where datetime >= '2009-04-01 00:00' AND datetime <= '2009-12-31 23:55' group by yearmonth,dayhour) as A group by hour(datetime);
```

4.2.2 Household traffic modeling

A traffic model can be used to predict the future behavior of a real traffic stream. Ideally, the traffic model should accurately represent all of the relevant statistical properties of the original traffic, but such a model may become overly complex. In this thesis we consider a model of a household traffic volume. We expect the model will help predict the behavior of traffic as it passes through the network.

Our approach begins by plotting a Cumulative Distribution Function (CDF) associated with single household's contribution. The CDF plotting will represents the probability, at each proportion of households, of reaching that volume or a lower one. We have the hypothesis that CDF of both the incoming traffic and outgoing traffic belong to a long tail distribution, since the long tail feature exists in some Internet traffic characteristics which are well studied, such as packet inter-arrival time[7], holding time of a modem[21], HTTP file size, HTTP request and reply length[31], etc.

4.2.2.1 Least square curve fitting

A question is: how can we find the proper model? Based upon early description of Internet traffic[5][7] and by computing the CDF of the observed traffic (see Figure 5.8 in Section 5.2), several probability distributions are potential candidate. The best fit will be chosen by using

curve-fitting. We use least squares to find the parameters of a curve. Mathematically, the least (sum of) squares criterion that is minimized to obtain the parameter estimates is [30]

$$Q = \sum_{i=1}^{n} [y_i - f(x_i, \beta)]^2$$

Where (x_i, y_i), $i = 1, \ldots, n$, is a simple data set consists of n points, $f(x_i, \beta)$ is the model function, and vector β is the parameter. In our case, we perform iterative algorithm to find best β which will generate minimum value of Q.

4.2.2.2 Long-tailed distribution

In long-tailed distributions a high-frequency or high-amplitude population is followed by a low-frequency or low-amplitude population which gradually "tails off" asymptotically. The events at the far end of the tail have a very low probability of occurrence. A long-tailed distribution is an important subclass of a heavy-tailed distribution. In mathematical terms, a function F is heavy-tailed if its tail decreases slower than any exponential distribution.

The Weibull distribution and Pareto distribution are the famous examples of long-tailed models. Their mathematical expressions, along with exponential distribution are listed in Table 4.1.

Table 4.1 Probability distributions

Distribution	Probability Density p(x)	Cumulative probability F(x)
Exponential	$\frac{1}{a} e^{-\frac{x}{a}}$	$1 - e^{-\frac{x}{a}}$
Weibull	$\frac{1}{a}\left(-\frac{x}{a}\right)^{c-1} e^{-\left(\frac{x}{a}\right)^c}$	$1 - e^{-\left(\frac{x}{a}\right)^c}$
Pareto ($k>0, a>0, x \geq \boldsymbol{k}$)	$\frac{ak^a}{x^{a+1}}$	$1 - \left(\frac{k}{x}\right)^a$

4.2.2.3 Experiment SQL query

In order not to be biased by single observation, we randomly picked several days for analysis.

The SQL query used for the experiment of one of these days is:

```
mysql> SELECT SUM(trafficIn)/8/1024*300 AS dataInMB, SUM(trafficOut)/8/1024*300 AS dataUtMB,
aswitch, aport FROM pl2009 as pl, becs2009 as becs WHERE becs.ip = pl.ip AND becs.datetime = pl.datetime AND
pl.datetime >= '2009-05-01 00:00'  AND pl.datetime <= '2009-05-01 23:55' GROUP BY aswitch, aport ORDER BY dataInMB
DESC;
```

4.2.3 Application traffic patterns

In this section our attention focuses on traffic at the application level. As the application name of the traffic can be identified, we may gain additional insight into the composition of residential traffic.

4.2.3.1 Categorized application traffic patterns

We categorized traffic by dividing it into categories based on the nature and usage of a set of applications. More specifically, we divided traffic into six categories: web browsing, Instant Messaging(IM), media streaming, online gaming, peer-to-peer(P2P) file sharing, and others. The web browsing category mainly contains HTTP traffic. IM was popular among the network users during these years, and we have included MSN Messenger, Yahoo! Messenger, and Skype in this category. Media streaming includes all types of online stream, regardless whether it is P2P based or not. Flash[23] embedded in a web page is classified in this category. Online gaming traffic is generated by user's gaming client. P2P file sharing only includes the P2P traffic used for file sharing, while excluding P2P streaming and P2P gaming. Finally we have classify as "others" all of the traffic which does not fit in any of the above category, such as DNS, FTP, IPSec, VPN, routing protocols, etc. The SQL query used for this experiment is:

```
mysql> select service, sum(trafficIn),sum(trafficOut) from pl2009 where datetime>='2009-05-01 00:00' and datetime <= '2009-05-01 01:00' group by service;
```

After extracting data we use python scripts to classify service into different application categories, and draw them by using pie chart.

4.2.3.2 Top traffic applications

After categorizing all applications, we can compute which specific application is prevalent and which ones are not that popular. Since we have long time period, we expect to find some interesting results.

Although the SQL queries used in this type of experiment are similar to the previous case, we analyze the resulting data to search for heaviest traffic in each category. The amount of traffic for different categories will be shown in a vertical bar chart.

4.2.4 User penetration

Apart from looking at the traffic volumes, the penetration of applications gives a detailed understanding of application usage. Following the trends in an application's penetration is a powerful instrument for analyzing trends. By penetration in this section, we mean the percentage of households who have used the particular application during the measurement time. This analysis is particularly suitable for describing those applications which do not consume large amounts of bandwidth, such as Instant Messaging traffic, online game traffic, etc. On the other

[23] Flash is a vector animation (read about vector animation software) software, originally designed to create animations for display on web pages.

hand, it helps the researchers to get an idea of user behavior for all categories of applications.

By executing SQL query, we can extract the data necessary to count the number of distinct households for each application. Given this data we may compare an application's prevalence with other applications in the same category. An example of such query is:

```
mysql> select service,count(distinct aswitch,aport) from pl2009 as pl, becs2009 as becs where pl.ip=becs.ip and pl.datetime=becs.datetime and pl.datetime>='2009-05-01 00:00' and pl.datetime<='2009-05-01 01:00' group by service;
```

4.2.5 Grouping analysis

Grouping analysis is applied in order to discern more features of Internet user behavior. In this section, two approaches will be introduced. The traffic data will be divided into sub-groups by using k-means clustering and by subscriber.

4.2.5.1 Group by k-means cluster

Clustering analysis, which was mentioned in Chapter 2, is a common and effective technique to determine the intrinsic grouping in a set of unlabelled data, and to further investigate each groups' behavior. Clustering by traffic volume is an intuitive way of thinking about the traffic. But judging from the results which are presented in section 5.2, it is difficult using this method to determine the bound for each cluster as the household traffic volume follows a smooth distribution. Hence we use the k-means algorithm to cluster the households, and then attempt to describe the characteristics of each cluster.

The k-means algorithm is a commonly used algorithm for data analysis. Based on the input parameter k, it partitions a set of objects into k clusters so that each object belongs to the cluster with the nearest mean. K-means is widely deployed due to its simplicity and efficiency.

In our project, a hypothesis was that three clusters could be found: heavy traffic households, medium traffic households, and light traffic households. As P2P users usually generate a large portion of traffic, especially for outgoing traffic (uplink from the users), the cluster will be computed using k-means on the outgoing traffic volume. It is expected that the three clusters have their own distinguished Internet using behavior. Later in this thesis project user behavior in terms of HTTP active time will be studied on a k-means cluster basis. From the results in section 5.3, HTTP has a penetration rate of about 98%. That is the reason why HTTP was chosen as indicator of user activity.

The SQL script shown below extracts the data for each household's HTTP active times in minutes, as well as their bidirectional traffic.

```
mysql> SELECT IFNULL( ActiveMinute, 0 ) , B . * FROM (
```

```
SELECT count( DISTINCT ( pl.datetime) ) *5 /14 AS ActiveMinute, aswitch, aport FROM pl2009 AS pl, becs2009 AS becs
WHERE becs.ip = pl.ip AND becs.datetime = pl.datetime AND pl.datetime >= '2009-05-01 00:00' AND pl.datetime <=
'2009-05-14 23:55' AND pl.service = 'HTTP' GROUP BY aswitch, aport )A
RIGHT JOIN (
SELECT SUM( trafficIn ) /8 /1024 *300 /14, SUM( trafficOut ) /8 /1024 *300 /14, aswitch, aport FROM pl2009 AS pl,
becs2009 AS becs WHERE becs.ip = pl.ip AND becs.datetime = pl.datetime AND pl.datetime >= '2009-05-01 00:00'
AND pl.datetime <= '2009-05-14 23:55' GROUP BY aswitch, aport )B
ON ( A.aswitch = B.aswitch AND A.aport = B.aport ) ORDER BY ActiveMinute DESC
```

As all entries stored in the database have a 5-minutes interval, pl.datetime*5 is the approximation for duration in minutes. The experiment is based on two weeks data, so the query will divided by 14 to get an average daily statistic.

4.2.5.2 Subscriber grouping

The dataset used in our project was collected from a real residential network. Several ISPs were contributing to the whole dataset. Because it is a users' right to choose the subscription they favor, in this section, we group users based upon the subscription bandwidth they have chosen, to determine whether there is a relationship between subscription bandwidth and the users' actual traffic volume.

The scanning interval is also taken as two weeks, and the traffic will be computed as an average daily volume. The corresponding SQL query is:

```
mysql>SELECT iservice, SUM( trafficIn ) /8 /1024 *300 AS dataInGB, SUM( trafficOut ) /8 /1024 *300 AS dataOutGB,
aswitch, aport FROM pl2009 AS pl, becs2009 AS becs WHERE becs.ip = pl.ip AND becs.datetime = pl.datetime AND
pl.datetime >= '2009-05-01 00:00' AND pl.datetime <= '2009-05-14 23:55' GROUP BY aswitch, aport ORDER BY iservice
```

4.2.6 Concurrent application analysis

Concurrent application analysis is another approach for studying user behavior. It investigates the hidden relationship between different categories of applications, i.e. to see what two applications are more likely being used simultaneously. It is a common sense that Internet user will not solely browse the webpage or solely download the movie via P2P without opening other applications. This thesis will try to unveil the statistical patterns by utilizing the dataset.

We are inspired by association rule learning method[36], which is used to discover interesting relations between variables in large databases. For our case, we set a certain interval of time and look for hidden association rules between different application categories.

Figure 4.1 and Figure 4.2 demonstrate the flow chart of the programs which are used for concurrent application analysis. The main program concurrent.py is shown in Figure 4.1. It first constructs 2 matrixes. One $5 \times NUMOFIP \times NUMPOFTIME$ matrix (denoted data matrix) is to

help store the occurrence of 5 application categories for each IP within each time slot. The other matrix (denoted probability matrix) has 5×5 elements, to stores the final result, showing the probability of two concurrent categories' application. After construction of these matrices, a series of nested loops invoke queries, which combined with category matching, are executed to get the values of data matrix. When $5 \times NUMOFIP \times NUMPOFTIME$ of values are obtained, another program manmatrix.py is called. The flow chart for manmatrix.py is displayed in Figure 4.2. It describes how to compute 5×5 statistical results from the $5 \times NUMOFIP \times NUMPOFTIME$ data matrix. Note that the concurrent probability here is calculated by dividing occurrence with numbers of users of a given application within the certain interval. The final results will present in Chapter 5.

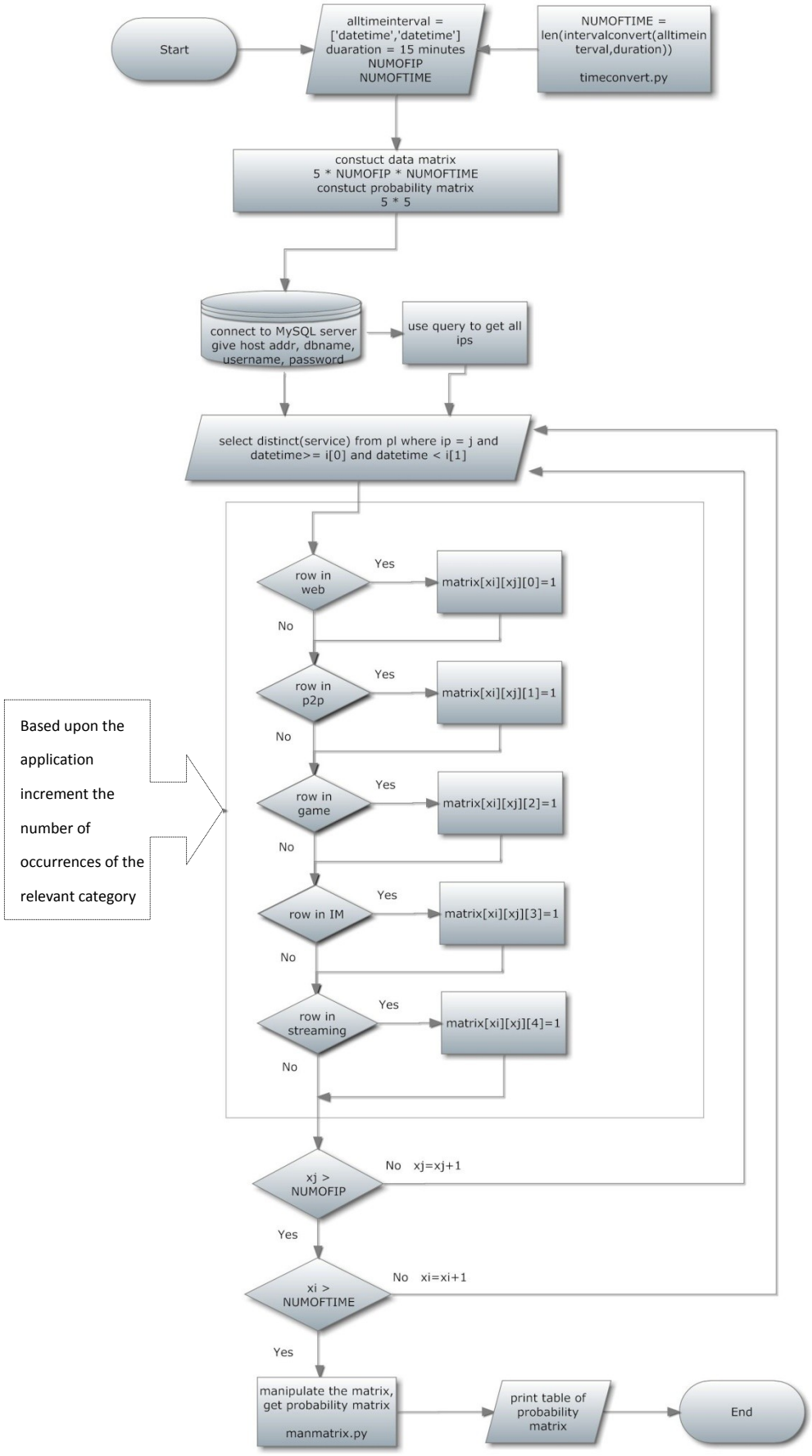


Figure 4.1 Flow chart for main program of current application analysis

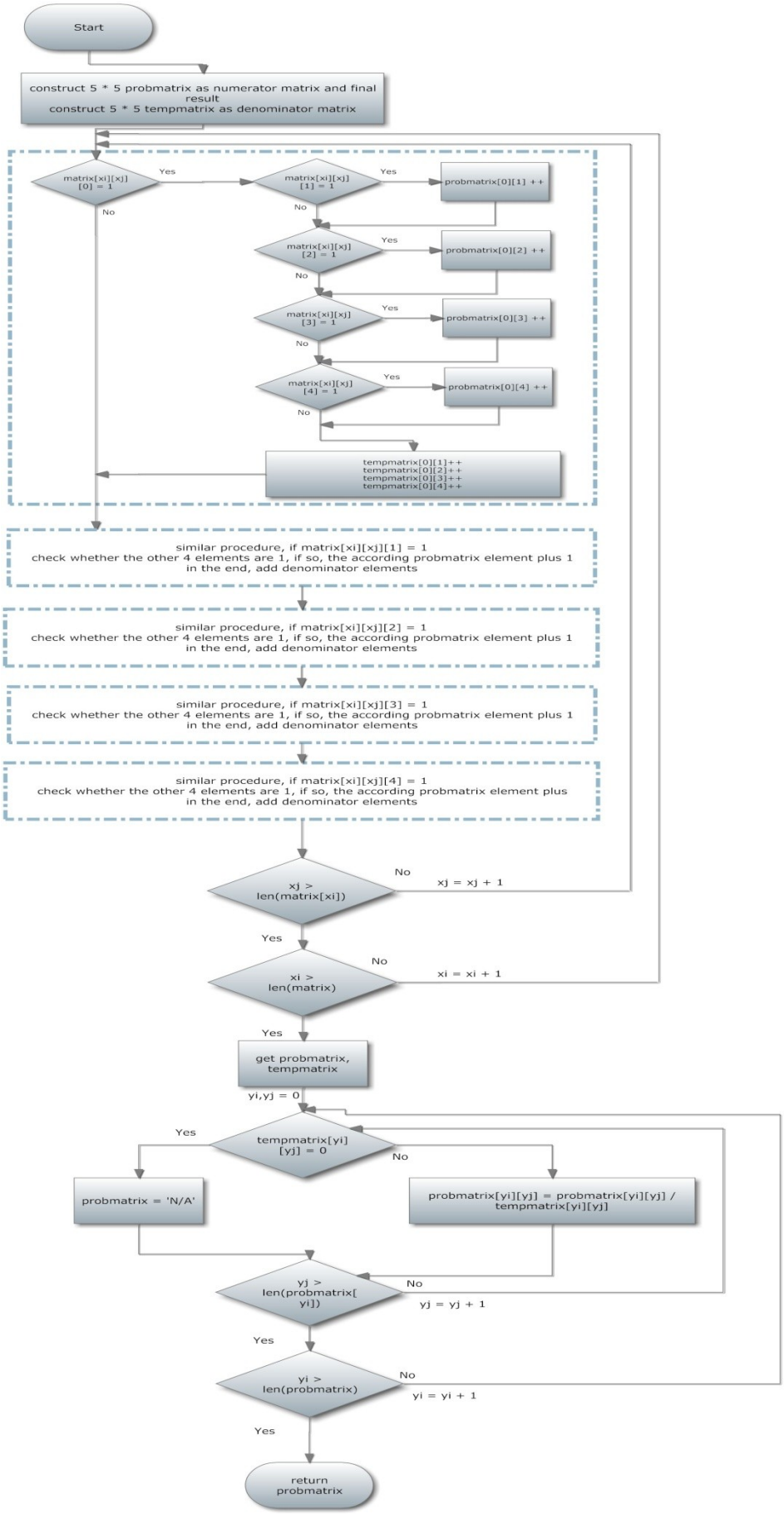


Figure 4.2 Flow chart for manmantrix.py (manipulate the data matrix)

Chapter 5

Results and Analysis

In the previous chapter, a series of experimental methods were introduced. Their results will be presented and discussed in this chapter following the same order as the methods were presented. The analysis will be given along with suitable figures and tables.

5.1 Aggregated analysis of network traffic

5.1.1 Long term traffic trends

We look for traffic trends in a long period, from March 27th, 2009 to December 31st, 2009. In Figure 5.1, the blue curve depicts the incoming traffic (traffic to the customer), the red curve shows the outgoing traffic, and the green curve shows the total traffic volume (simply the sum of the blue and the red curves).

As shown in Figure 5.1, the outgoing traffic volume usually surpasses the incoming traffic volume. That is a result of file-sharing applications. It is important to note that in this residential network the access lines in the network are generally symmetric and of high speed. This means that the users in the network can act as servers for file sharing application, thus they are providing the server function to users who have high downlink data rates but low uplink data rates, such as DSL or wide area mobile networks.

Both direction of traffic usually exhibit the same trend, hence if the incoming traffic increases then the outgoing traffic will probably increase too. Over this period there are two turning points. One is April 1st, 2009, the network traffic volume suddenly drops to roughly half of the traffic volume of the previous day. This result was reported by some news agencies[32]. A possible

reason given in Phat's master thesis report[34] is that the IPRED Law came into force on that specific day, hence the file-sharing traffic decreased dramatically. Another turning point appears around middle of August, when the traffic gradually increases until it reaches the level just before IPRED Law's enforcement. This increase in traffic is seen every year at the end of Swedish people's nation-wide summer vacation. The major part of the yearly traffic growth occurs in the fall, and this year the growth rate is likely to be higher, due to the bounce back after the IPRED Law's effect.

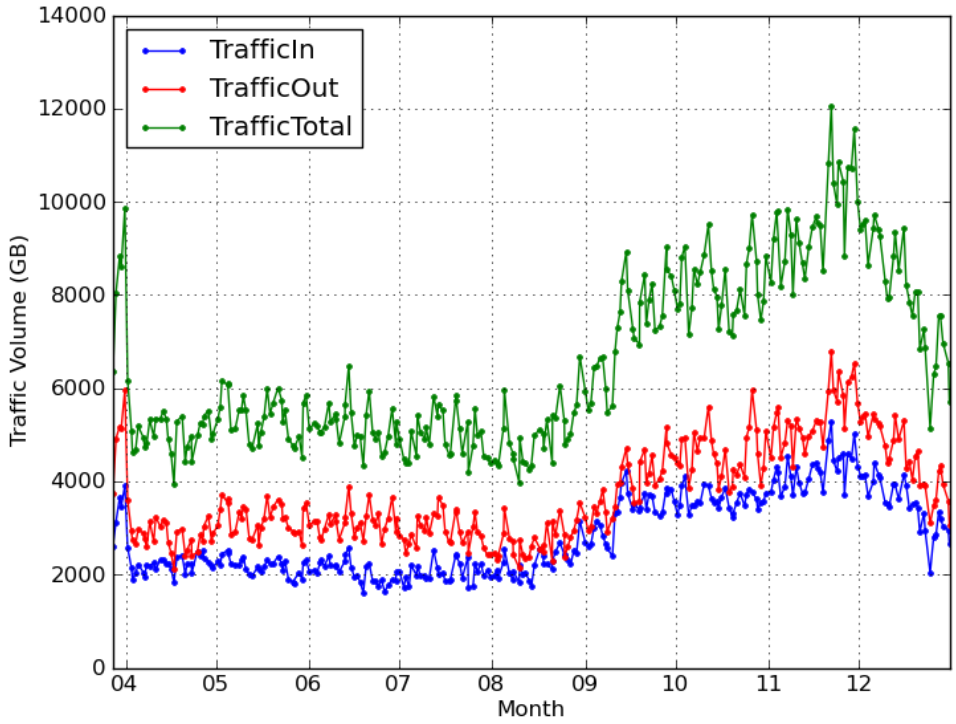


Figure 5.1 Aggregated daily traffic from 2009-04-01 to 2009-12-31

5.1.2 Single week traffic pattern

An example of weekly traffic pattern is shown below, in order to analyze the difference between weekdays and weekends. Figure 5.2 shows the traffic for a specific week, September 07th to 13th. During this week, we see little remarkable difference between the single days except for a slightly increased traffic volume (both the incoming and outgoing traffic) during the weekends. However, looking at another week August 3rd to 9th, as shown in Figure 5.3, there is more traffic on weekdays than at weekends. We infer that it is due to the summer vacation period.

To have a more general idea of the daily traffic during each week, we calculate the average (arithmetic mean) and standard deviation of each week day during the whole captured period which starts from April 1st to the end of that year. The result is shown in Figure 5.4.

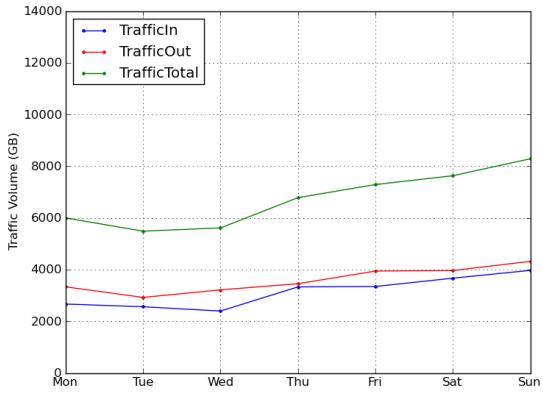


Figure 5.2 Single week traffic pattern from 2009-09-07 to 2009-09-13

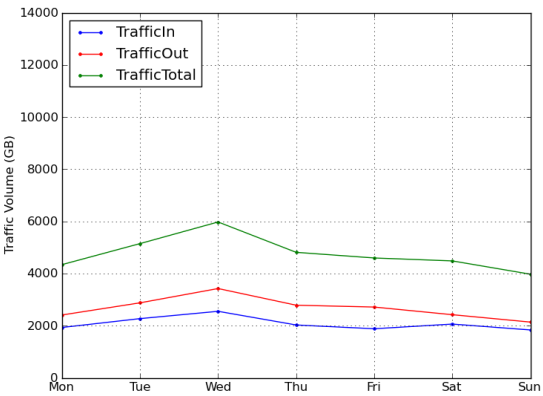


Figure 5.3 Single week traffic pattern from 2009-08-03 to 2009-08-09

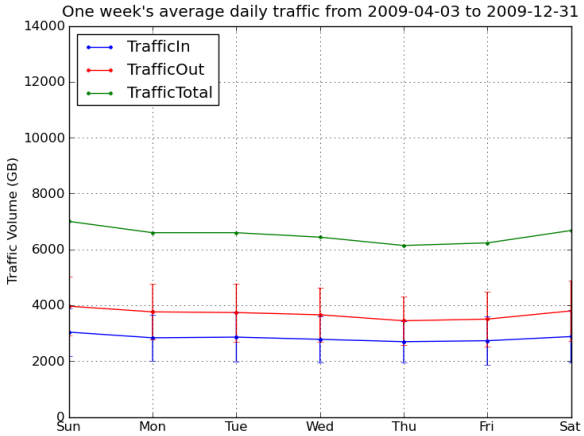


Figure 5.4 average of single week traffic pattern during 2009-03-27 to 2009-12-31

We can judge from the figure, there is no obvious difference of traffic between each day. Weekends' traffic has only a slight increase comparing with the workdays. Outgoing traffic is approximately 25% higher than incoming traffic every day of the week.

5.1.3 Single day traffic pattern

Traffic volume varies strongly during the course of a day. Figure 5.5 shows the aggregated traffic volume associated by the hour of the day. Figure 5.6 uses the same method to display another day's traffic data. Since one day's traffic is easily influenced by single heavy users, we calculate the average traffic volume with its standard deviation for each hour from March 27th to December 31st, shown in Figure 5.7.

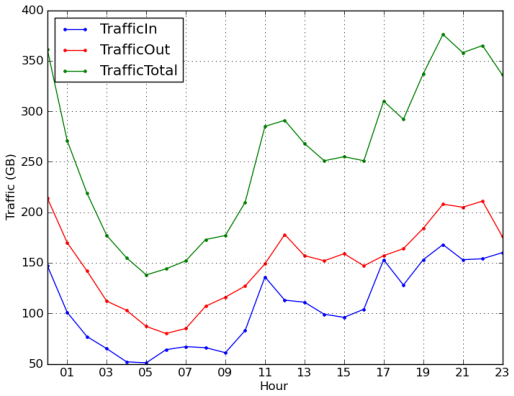


Figure 5.5 2009-04-01 day traffic

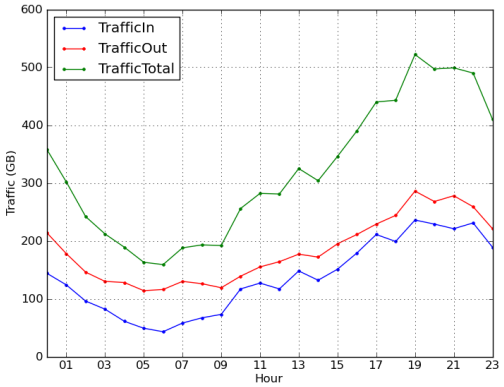


Figure 5.6 2009-10-01 day traffic

We can see that during the period 17:00-24:00 is the network traffic is at its peak time of the day. These 7 hours are accounting for roughly 50% of the whole day's traffic. While at 6 a.m. the network experiences the lowest load. This hourly traffic pattern reveals the aggregated residential network users' living behavior, i.e. when people stay at home, when most people go to sleep, etc.

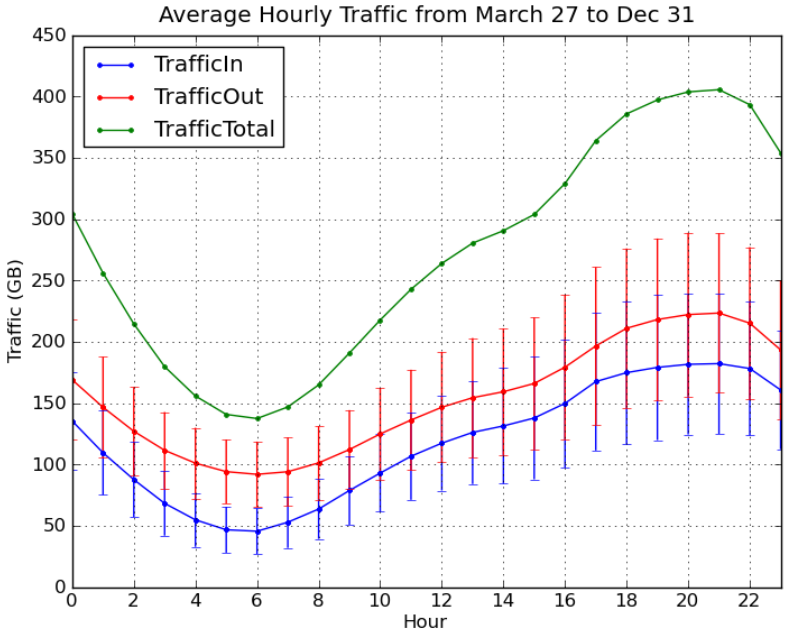


Figure 5.7 Average of daily traffic during Mar 27th to Dec 31st

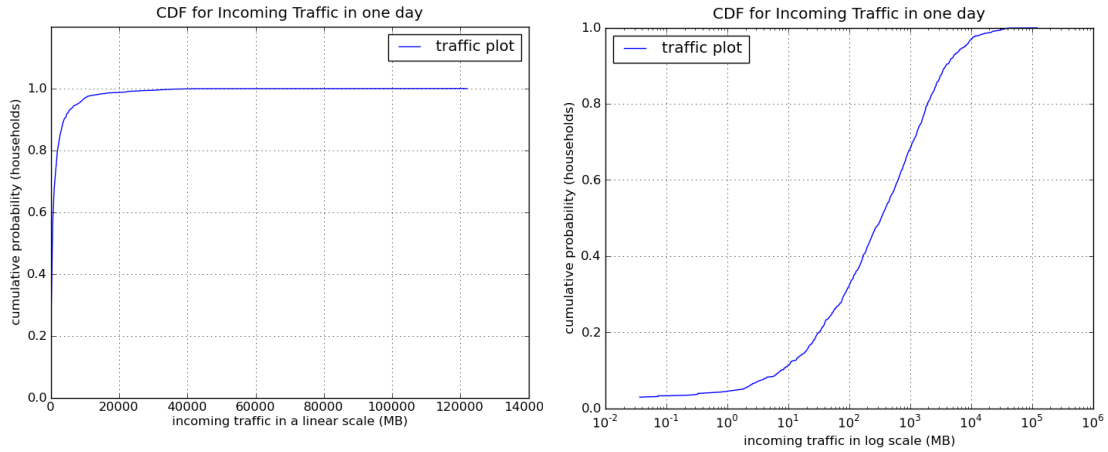
5.2 Household traffic modeling

Household traffic is of interest, since the household is the entity that subscribe for service from the Internet service provider. This is also a unique feature of the data analyzed in this report,

since access to household level data is very scarce.

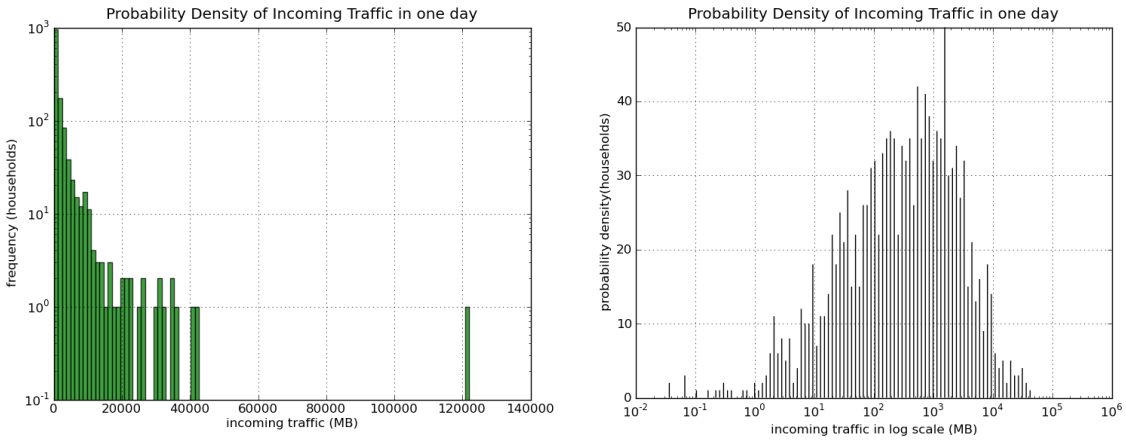
In order to model household traffic, a random day was chosen for analysis. The dataset consists of households who generated incoming or outgoing traffic within the 24 hours. The parameter for comparison in this analysis is the traffic volume generated by the household during the selected measurement time. A household is here defined as a physical port on the access switch in the access network.

A comparison of traffic volume between households is done with a Cumulative Distribution Function (CDF), shown in Figure 5.8(a) and 5.8(b) (in which the X axis is plotted using a logarithmic scale). It can be seen that more than 95% of the households consume less than 10GB on 2009-05-01, while the remaining portion shapes into a long tail.



(a) x axis in a linear scale *(b) x axis in a logarithmic scale*

Figure 5.8 CDF for incoming traffic on 2009-05-01



(a)y axis in logarithmic scale *(b) x axis in logarithmic scale*

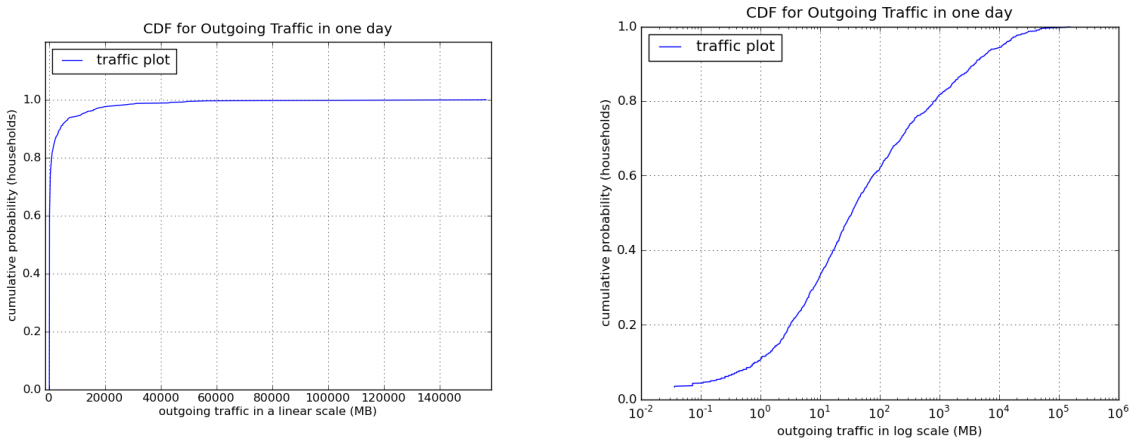
Figure 5.9 PDF for incoming traffic on 2009-05-01

Figure 5.9(a) presents the histogram for the traffic probability density distribution, which further demonstrates the long tail character. There is a very high probability that a household

consumes no more than 10GB in the 24 hours, but there still exists a few households who generate large incoming traffic. For example, on 2009-05-01, one household downloaded more than 120GB from the Internet. It can be assumed from this phenomenon that the distribution is heavy-tailed (because the tail decreases slower than any exponential distribution).

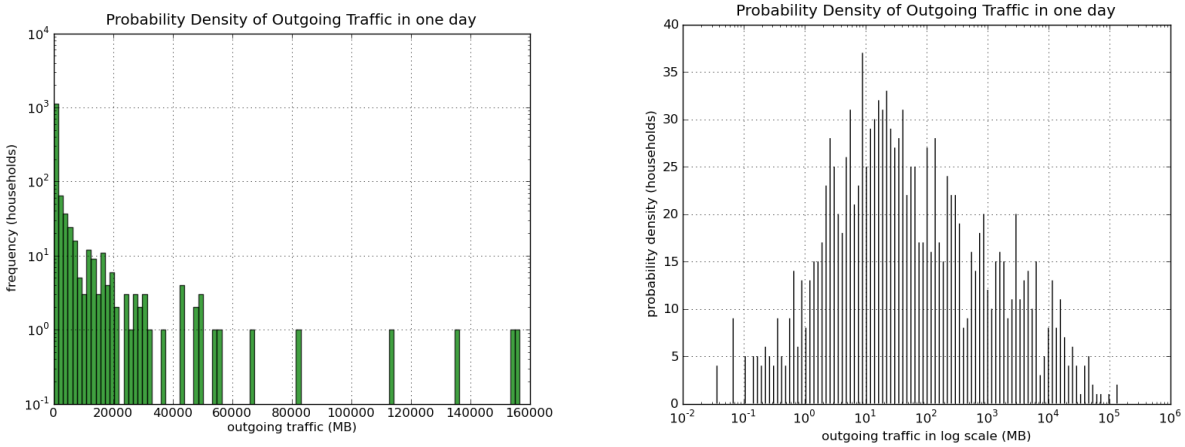
Figure 5.9(b) shows the probability density distribution with a logarithmic X axis. We can see that most households download traffic from 10MB to 10GB (accounting for 89% of total traffic) on that day, which explains the steep slope in Figure 5.8(b) within that range. From the same figure we can see that the amount of households which consume less than 1 MB is rather small.

We also take the outgoing traffic (uplink) into account. For the same day 2009-05-01, we have the accordingly Figures 5.10 and Figure 5.11. A long-tailed distribution is displayed as well. It can be seen that, the number of 'heavy user' (who consumes large network resources) is larger than that of incoming traffic. The slope of the CDF is not as steep as the incoming traffic CDF, which means that the probability density embraces a comparatively flatter distribution.



(a) x axis in a linear scale *(b) x axis in a logarithm scale*

Figure 5.10 CDF for outgoing traffic on 2009-05-01

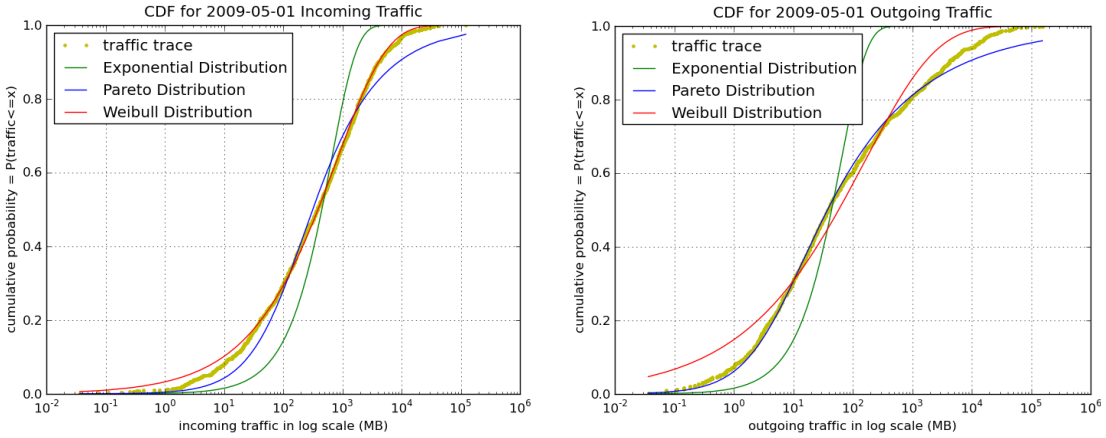


(a) y axis in logarithm scale *(b) x axis in logarithm scale*

Figure 5.11 PDF for outgoing traffic on 2009-05-01

To better understand the characteristic of these households, we have considered several long-tailed mathematical models to fit our dataset to. These models are exponential, Pareto, and Weibull distributions. We use least square fitting to find proper parameters for each of these possible distributions. The fitted curves for the CDF is shown in Figure 5.12, (a) for incoming traffic and (b) for outgoing traffic.

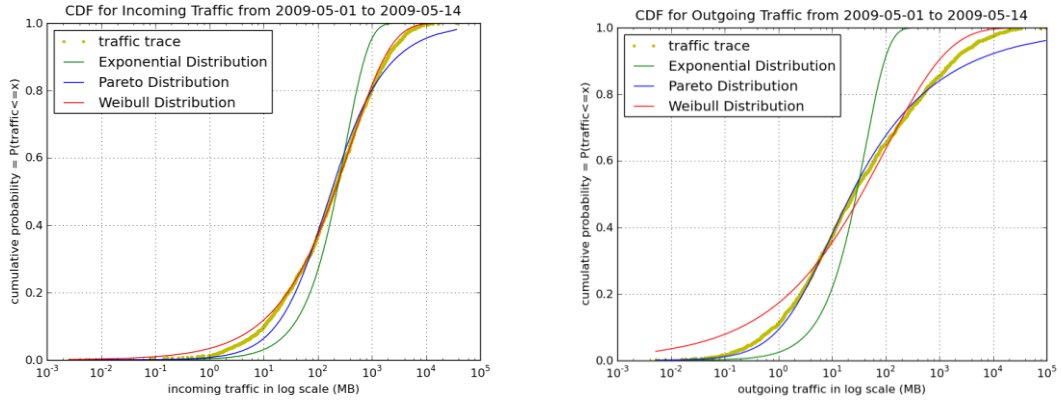
In Figure 5.12(a), the Weibull distribution provides the best fit, especially in the region of high incoming traffic volume. When the household has less than 100MB traffic, the Weibull model slightly overestimates the cumulative probability. Exponential distribution apparently doesn't fit the CDFs well, nevertheless, it shows that the incoming traffic belongs to the heavy tailed family, as it is not bounded by the exponential function. From Figure 5.12(b), we can see that Pareto Distribution is the best fit for outgoing traffic. It has a very good fit for households with less than 1GB uplink traffic, but it underestimates the traffic when households have more uplink traffic.



(a) incoming traffic *(b) outgoing traffic*

Figure 5.12 CDF curve fitting on 2009-05-01

We also fit these distributions to the curves of an average day's traffic, see Figure 5.13. The traffic data was collected from 2009-05-01 to 2009-05-14, lasting 2 weeks. We calculated the average traffic for each household and plotted the results. The Weibull and Pareto distribution are still the best fit for incoming and outgoing traffic respectively.



(a) incoming traffic *(b) outgoing traffic*

Figure 5.13 CDFs and curve fits for average daily traffic from 2009-05-01 to 2009-05-14

In order for model evaluation, we performed a series of experiments and computed CDF curves and fit distributions to them on various days by utilizing a Python batch script. Datasets are selected from several days late in September 2009. The results are shown in Figure 5.15 and Figure 5.16.

Judging from those figures, Weibull is always the better model for incoming traffic. The disadvantage of applying the Weibull distribution is that it will overestimate the portion of 'light user' (who do not have much incoming traffic). While for outgoing traffic, Figure 5.16 shows that Pareto fits best, except for periods dominated by 'heavy users' when there is some mismatch with the Pareto distribution. In the future work, one should consider a composite model which has a threshold, i.e., below the threshold we use Weibull to model the incoming traffic while another model will be utilized for traffic above this threshold. Some economics models have been studied by using such methods[33].

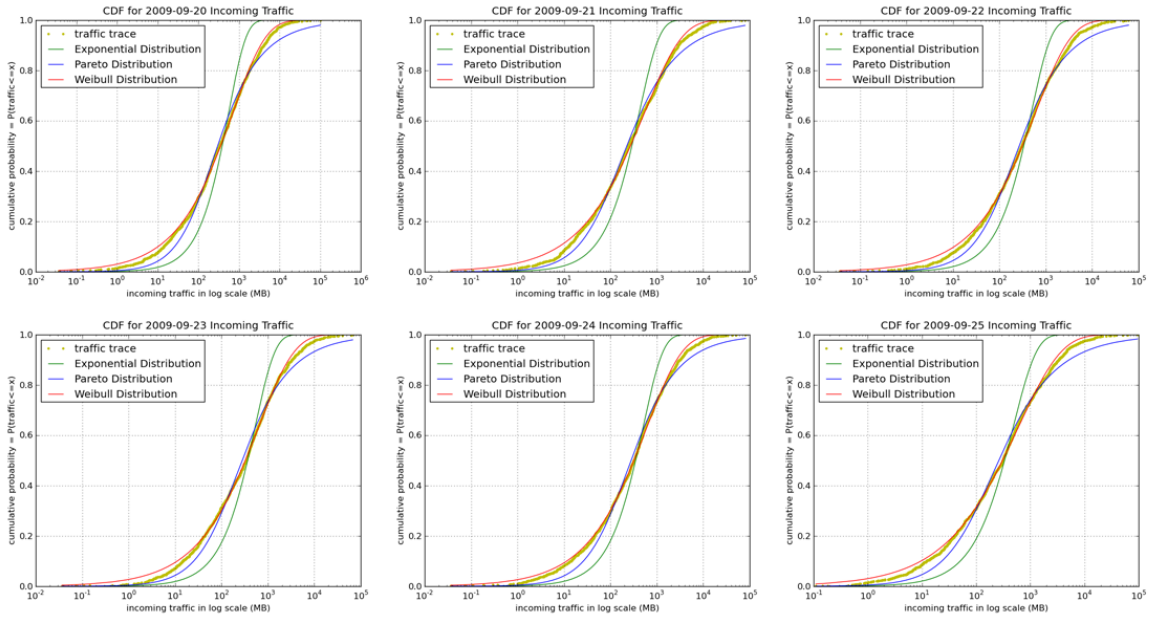


Figure 5.14 curve fitting for incoming traffic CDF, from 2009-09-20 to 2009-09-25

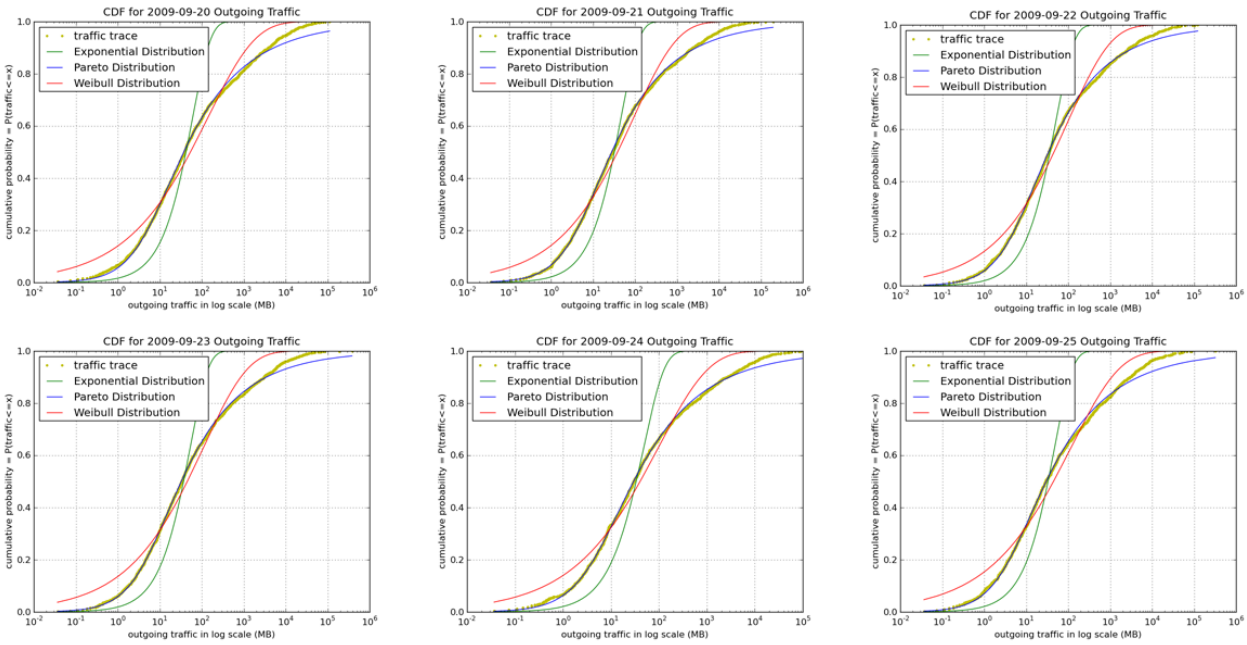


Figure 5.15 curve fitting for outgoing traffic CDF, from 2009-09-20 to 2009-09-25

Based upon multiple experiments, we have come to an empirical conclusion that incoming traffic per household is well described by a Weibull distribution, while the outgoing traffic per household has a more Pareto like distribution. Both of these distributions mean that most households consume a small portion of the overall network traffic, while the majority of traffic is due to a few 'heavy users'. Now that we have fit distributions for both incoming and outgoing household-based traffic, we can use to estimating the parameters of a distribution based on observed data, or test hypotheses about them.

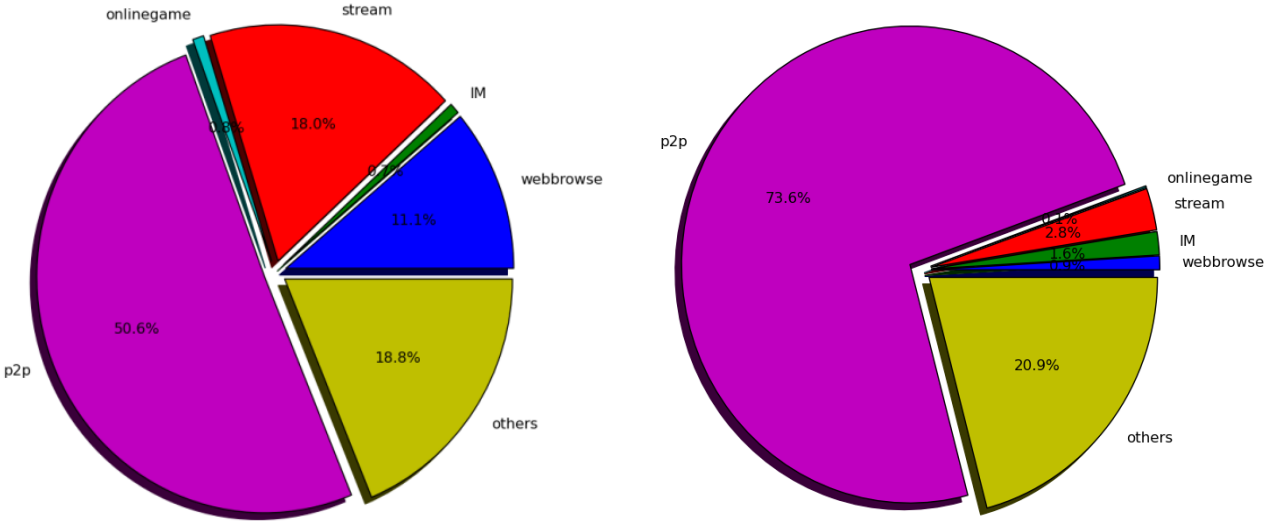
5.3 Application traffic patterns

5.3.1 Categorized application traffic patterns

Figure 5.16 and 5.17show pie charts displaying categorized application traffic. Different colors correspond to different categories. Blue is for web browsing; green is for Instant Messaging; red is for media stream generated in the traffic; online gaming traffic is depicted in cyan; P2P is in purple; and other applications which are not included in any other category are depicted in yellow.

Figure 5.16 shows the traffic volumes during two weeks, from 2009-05-01 to 2009-05-14. In Figure 5.16(a), it can be seen that P2P counts for roughly half of the incoming traffic. Media stream accounted for 18% of the incoming traffic, which is reasonable because video and audio content usually requires high bandwidth. The 'others' category mainly consists of FTP transfers,

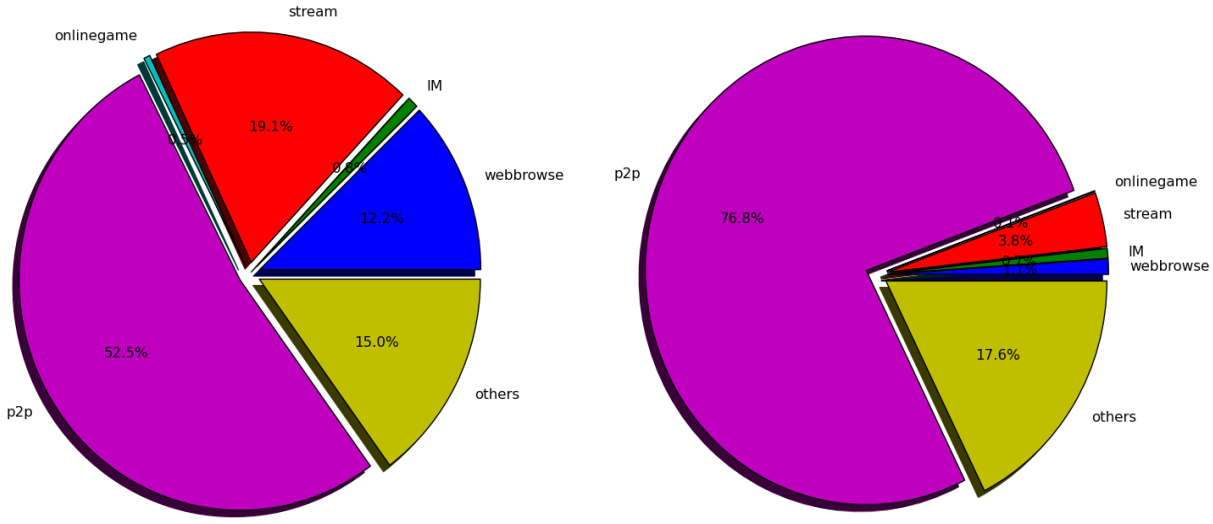
encrypted data and unrecognized traffic, which are not the focus of this thesis. Web browsing comprises 11.1% of all traffic, while Instant Messaging and online game comprise 0.7% and 0.8% of all of the traffic respectively. In Figure 5.16(b), P2P accounts for 73.6% of the uplink(outgoing), which dominates the overall application traffic.



(a) incoming traffic for two weeks *(b) outgoing traffic for two weeks*

Figure 5.16 application traffic by category, data from 2009-05-01 to 2009-05-14

Figure 5.17 shows the data five months later, i.e. October 2009. Proportion of P2P traffic has increased on both incoming and outgoing direction.



(a) incoming traffic in two weeks *(b) outgoing traffic in two weeks*

Figure 5.17 application traffic by category, data from 2009-10-01 to 2009-10-14

One Internet research[35] made by Henrik and Klaus in 2009 also reveals the patterns for different categories' application. Their research objects covered ISPs and universities in 8 countries and regions, including Europe, Middle East, and Africa. The results show that P2P generates most traffic in all regions, and web traffic accounts for a high percentage. Those results match the findings we got for our network, which means the global Internet traffic pattern has similar characteristics.

5.3.2 Top traffic applications

After learning the relative composition of different categories' traffic, it is interesting to see which specific application generated the traffic. Figure 5.18 and Figure 5.19 displays the top 5 application protocols of the P2P category for both traffic directions. It should be noted that the traffic data is identified by the protocols the application uses and several protocols may be associated one application.

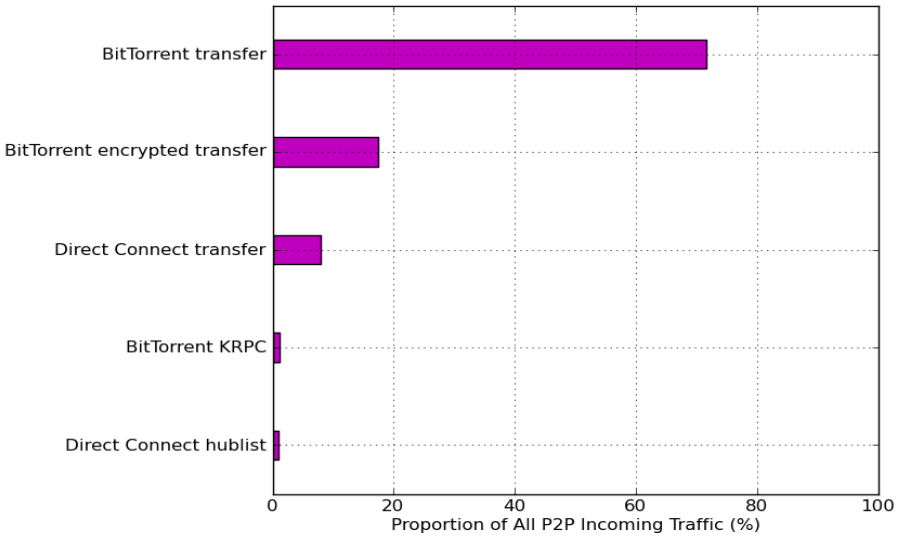


Figure 5.18 Top 5 application protocol of P2P incoming traffic data from 2009-05-01 to 2009-05-14

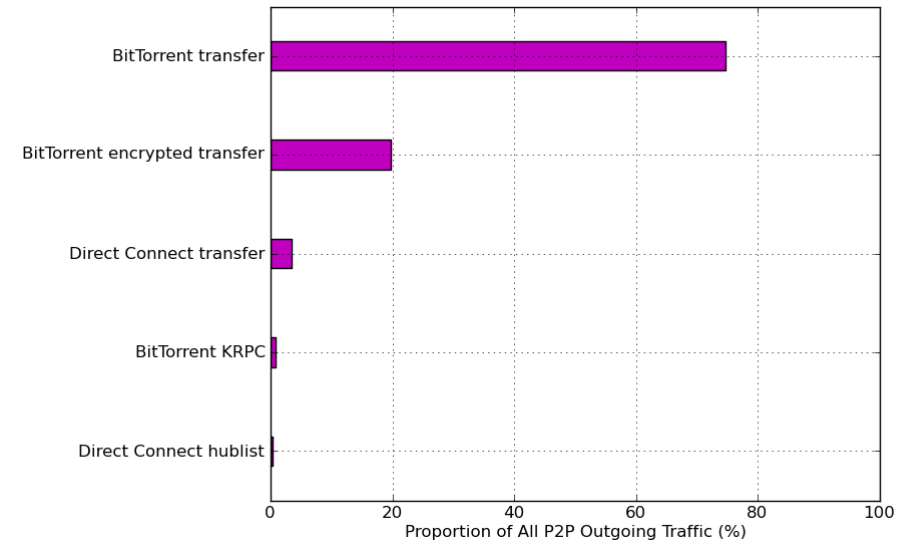


Figure 5.19 Top 5 application protocol of P2P outgoing traffic data from 2009-05-01 to 2009-05-14

Clearly, BitTorrent[24] is the dominate application in terms of generation of P2P traffic. Because BitTorrent clients also act as servers to re-distributed the content that they have received, we expect to see that the traffic is rather symmetric.

[24] BitTorrent is a peer-to-peer file sharing protocol used for distributing large amounts of data. It is also the name of company which develops and maintains the BitTorrent protocol. http://www.bittorrent.com/

Figure 5.20 and Figure 5.21 shows the top 5 streaming applications in both directions. It can be seen that Flash video over HTTP[25] accounts about 50% of the downlink traffic. However, in the uplink direction, Spotify[26] accounts for the largest amount of traffic. This first is expected as Flash video over HTTP is mostly used as a unidirectional application, while Spotify is P2P based. Some other P2P based applications can be found on the list, such as PPStream[27] and Sopcast[28]. And they each contribute a large share of outgoing traffic.

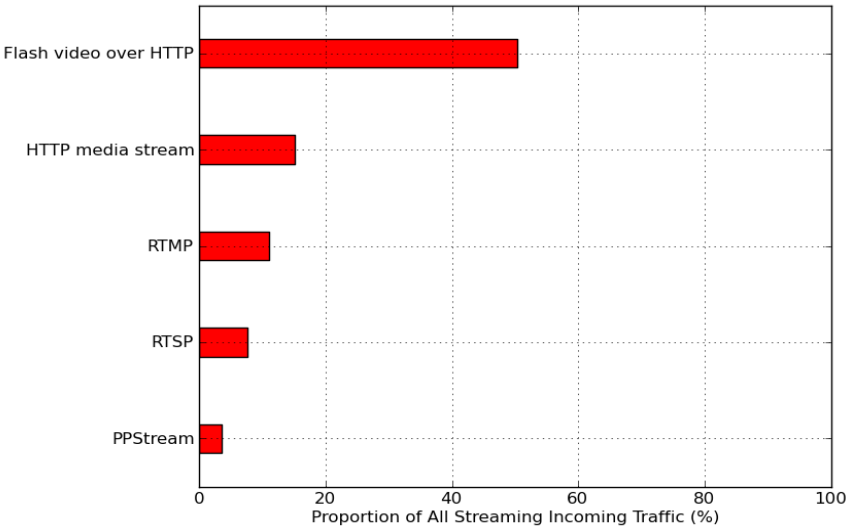


Figure 5.20 Top 5 application protocol of stream incoming traffic data from 2009-05-01 to 2009-05-14

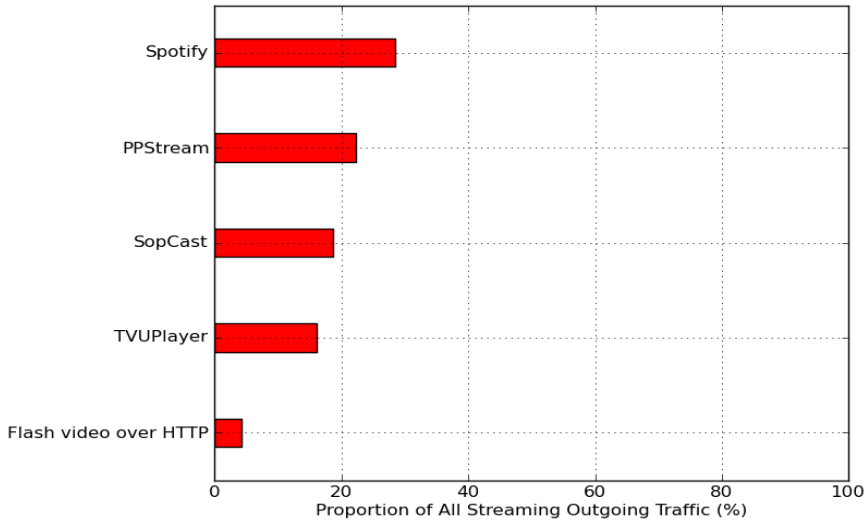


Figure 5.21 Top 5 application protocol of stream outgoing traffic data from 2009-05-01 to 2009-05-14

[25] Adobe's Flash protocol is widely used to add animation, video, and interactivity to Web pages. One famous example is Youtube.

[26] Spotify is a music streaming service with desktop applications. http://www.spotify.com/

[27] PPStream is a Chinese peer-to-peer streaming video network software. http://www.ppstream.com/

[28] Sopcast is another Streaming Direct Broadcast System based on P2P. http://www.sopcast.com/

5.4 User penetration

The results in this section examine penetration of applications, in relation to households. For example, if only one person in the household has used P2P applications and P2P traffic has been captured, then the whole household will be considered to be using P2P. Table 5.1 displays the most popular application protocols classified by categories. The dataset that was used for this was captured during two weeks, and all application protocols are listed in a descending order.

For the category of web browsing, HTTP has a penetration rate of 98.1%. It is obvious that most users will browse web pages while they are using the Internet.

In the P2P category, a list of application protocols can be found. BitTorrent KRPC achieves 85.5% penetration rate, making it the most popular P2P protocol. BitTorrent uses a type of "distributed sloppy hash table" (DHT) for storing peer contact information for "trackerless" torrents, and KRPC is the DHT designed for BitTorrent. Note that the second, third, and fourth most common protocols (in terms of penetration into households) are also BitTorrent protocols, showing the undoubted dominance of BitTorrent in P2P category. In the fifth position is Gnutella discovery. This protocol together with the 12th ranked Gnutella transfer protocol, belongs to the first decentralized P2P application Gnutella[29]. The sixth ranked, eDonkey[30] is another well-known P2P application, with 25.3% penetration of encrypted data and 4.8% of unencrypted data. Judging by the list of P2P penetration, it can be inferred that peer discovery and maintenance or tracking protocol usually have a higher penetration rate, e.g. BitTorrent KRPC has a larger penetration than BitTorrent encrypted or unencrypted transfer (BitTorrent transfer is independent protocol. A household can generate encrypted or unencrypted simultaneously), Gnutella discovery is 8 times more prevalent than the Gnutella transfer protocol. Furthermore, it is obvious from the results of previous section, that the applications with a large penetration are not necessarily the ones generating the most traffic, specifically while Gnutella and eDonkey are widely present these applications account for only a small portion of the traffic volume.

In the media stream category, Flash video over HTTP has the highest penetration, which matches the fact that it also accounts for the most incoming streaming traffic. While Spotify clients generate the most outgoing streaming traffic (referring to Figure 5.21), Spotify only has 21% penetration rate. The second and third heaviest outgoing traffic generators, PPStream and SopCast, have 12.6% and 3.7% penetration rate respectively.

The penetration rate for online game and Instant Messaging are also displayed in Table 5.1, which can be used for further study of user behavior.

[29] Gnutella used to be the most popular P2P file sharing network on the Internet. More information can be found in http://rfc-gnutella.sourceforge.net/index.html

[30] eDonkey network is a is a decentralized P2P file sharing network. http://www.emule-project.net/home/perl/general.cgi?l=1

Table 5.1 User Penetration Rate from 2009-05-01 to 2009-05-14

Category	Application Protocols	Penetration Rate (%)
Web Browsing	HTTP	98.1
	HTTP download	95.6
	Flash	34.5
	Java Web Start	22.2
Peer to Peer	BitTorrent KRPC	85.5
	BitTorrent encrypted transfer	53.5
	BitTorrent transfer	39.8
	BitTorrent tracker	36.4
	Gnutella discovery	31.5
	eDonkey encrypted	25.3
	Ares	24.5
	Direct Connect hublist	13.4
	Thunder UDP	8.3
	uTP	8.1
	eDonkey	4.8
	Gnutella transfer	3.6
Online Game	Source engine server	8.7
	World of Warcraft	8.0
	World of Warcraft login	7.8
	KartRider	7.1
	Source engine map transfer	4.1
	Project Entropia	3.9
	QQGame	3.5
	Playstation.net	3.1
	Massive Ad Client	3.0
	GameSpy server query	2.7
	PlayStation 3 firmware update	2.6
	Xbox Live login	2.4
	Xbox Live server browser	2.4
	Xbox Live update	2.4
	Battle.net	2.0
	PunkBuster	1.8
	id Tech 3	1.7
	Half-Life	1.6
	GameSpy	1.6
	Unreal keepalive	1.6
	Half-Life engine server	1.4
	Warcraft 3	1.4
	Call of Duty 4	1.2
	Call of Duty 2	1.1
	Counter-Strike: Source	1.0

Category	Application Protocols	Penetration Rate (%)
Instant Messaging	MSN messenger	76.7
	Windows Messenger Service	72.7
	MSN messenger echo	70.9
	MSN messenger chat	60.9
	Skype-P2P	46.7
	Skype-TCP	41.6
	Skype-UDP	37.3
	Skype version check	34.3
	RTP	30.1
	Skype login	28.9
	Skype discovery	27.0
	MSN messenger echo tcp	23.9
	SIP	22.7
	Yahoo! messenger	19.7
	Skype-SSL	19.5
	Skype-Hub2Hub	16.9
	Yahoo! messenger udp	13.3
	MSN messenger video	12.2
	XMPP	12.2
	Vivox	12.1
	Windows Live Messenger over HTTP	10.4
	MSN messenger video over udp	9.7
	QQ	8.8
	MSN messenger over HTTP	7.3
	OSCAR	6.5
Media stream	Flash video over HTTP	87.4
	HTTP media stream	77.2
	RTMP	56.3
	PPLive	40.2
	SHOUTcast	34.5
	RTSP	27.8
	Joost	27.2
	RTMPT	22.1
	Spotify	21.0
	RTSP media stream	14.1
	PPStream	12.6
	iTunes Store	10.6
	Flash audio over HTTP	6.0
	RTCP	5.2
	TVUPlayer	4.3
	SopCast	3.7

5.5 Grouping analysis

5.5.1 Grouping by k-means clustering

Figure 5.22 shows a scatter plot of HTTP active time versus the household's outgoing traffic. HTTP active time is computed as the duration of successive periods with HTTP traffic. Clusters are differentiated by colors. Red denotes heavy-traffic households, purple denotes medium, and cyan denotes light-traffic households. We computed each household's average traffic volume for a period of 2 weeks and plotted this volume on a logarithmic scale. To divide the household in terms of their traffic into clusters we used a k-means clustering. Three clusters were identified by using k-means, the two boundaries between these three clusters are 224MB and 0.843MB. And the three clusters respectively have 540, 706 and 634 households respectively, from a total of 1880 households. The 540-households in the heavy traffic cluster accounts for 98.5% of the total outgoing traffic. The medium traffic cluster accounts for 1.43% and the light traffic cluster accounts for 0.656% of the total outgoing traffic.

With respect to HTTP active time, the heavy traffic cluster exhibits a much more even distribution along the x axis. Among the 540 members of this cluster, there are about 50 who have a more than 15 hours daily HTTP active time. We assume that those people have the habit of leaving the computer on and are using a web application that automatically refreshes the web page(for example, an web e-mail interface, newspages, etc). Additionally, they have left their P2P software running, hence they generated a large traffic. Compared to the heavy traffic household cluster, the medium traffic cluster's users mostly exhibit a less-than-10-hours daily HTTP active time. For the 634 light-traffic households, their HTTP active time per day is mostly no more than 5 hours.

We have also tried to cluster the group by incoming traffic, but the HTTP active time plot does not display a distinct set of clusters. This may result from the fact that a user running a P2P application can consume a lot of incoming traffic even if they are only online for a short period of time.

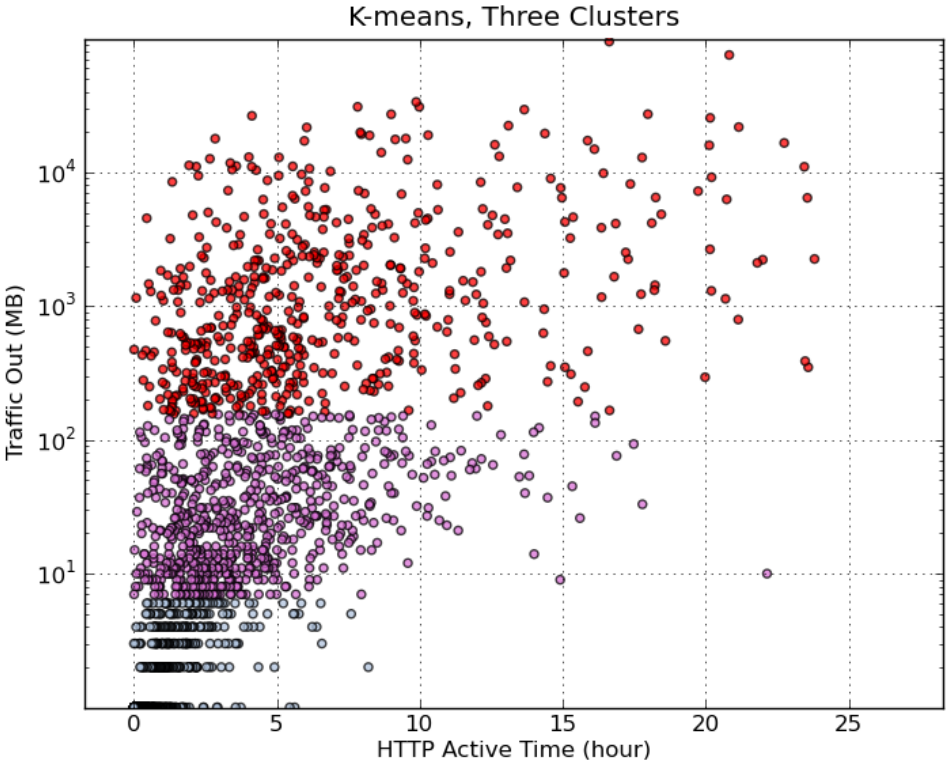


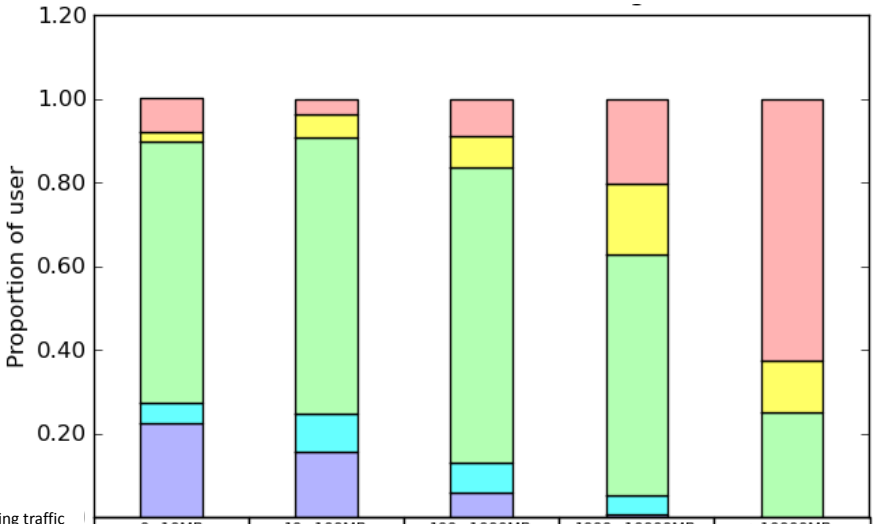
Figure 5.22 K-means cluster analysis, dataset from 2009-05-01 to 2009-05-14

5.5.2 Grouping by subscriber

Figure 5.23 displays a series of bar charts for households classified by different total incoming traffic volumes during a two week period of time. Traffic volume is divided into decimal order of magnitude bins. The proportion of traffic in this bin is grouped by type of subscription. Each type of subscription is denoted by a different color. Each bar illustrates the proportion of subscribers by type of subscription within the traffic volume bin. The first bar represents households who have less than 10 MB of daily incoming traffic (as averaged over 2 weeks). In this bin, households with 100Mbps subscription comprise 8% of users, while those with 30Mbps, 10Mbps, 5 Mbps and 1Mbps subscription have 2%, 62%, 5%, and 22% proportion respectively. The second to fifth bars show the other bins of incoming traffic volume.

As the user number of subscribers with each type of subscription varies, it seems more meaningful to make horizontal comparison (i.e., comparing users with the same type of subscription). In Figure 5.23 we see that households with 100Mbps subscriptions comprise an increasing proportion of users as the traffic volume increases (specifically 8%, 4%, 9%, 20%, and 63% for the five respective bins). The 30Mbps subscribers reach their peak (as proportion of users) in 1000~10000MB bin. While the 10Mbps subscribers, which have the majority of households, follow a trend of 'first rising, then descending'. Their shares are 62%, 66%, 71%, 57%, and 25%. Finally, as expected 5Mbps and 1Mbps subscribers are virtually extinct in '>10000MB' bin (this is

expected as 1 Mbps subscriber could maximally download 10800 MB per day).



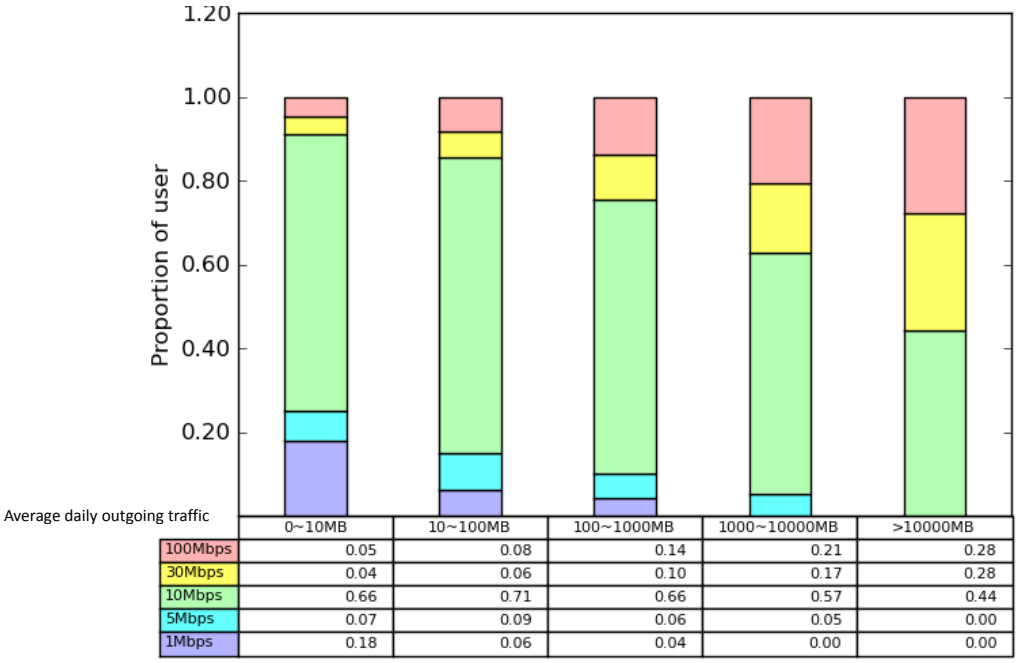
Average daily incoming traffic

	0~10MB	10~100MB	100~1000MB	1000~10000MB	>10000MB
100Mbps	0.08	0.04	0.09	0.20	0.63
30Mbps	0.02	0.06	0.07	0.17	0.13
10Mbps	0.62	0.66	0.71	0.57	0.25
5Mbps	0.05	0.09	0.07	0.05	0.00
1Mbps	0.22	0.16	0.06	0.01	0.00

Figure 5.23 Incoming Traffic as a proportion of users clustered by type of subscription
Data set from 2009-05-01 to 2009-05-14

When household proportion by subscription is classified by outgoing traffic volume, as shown in Figure 5.24, a much smoother trend could be found. The proportion of 100Mbps and 30Mbps subscribers increases monotonically with traffic volume. Both 10Mbps and 5Mbps subscribers have a 'first rising, then descending' trend, and a peak in 10~100MB bin. 1Mbps subscribers only have a significant presence in the bins for less than 1000MB traffic, with a monotonically descending trend of '22%,16%,6%,1%, and zero'.

We can see that some high-bandwidth households do not fully utilize their resources, but the low-bandwidth households are very much limited by their subscription (which sets a maximum data rate). In generally speaking, network users generated an outgoing traffic which corresponding to their subscription.



Average daily outgoing traffic	0~10MB	10~100MB	100~1000MB	1000~10000MB	>10000MB
100Mbps	0.05	0.08	0.14	0.21	0.28
30Mbps	0.04	0.06	0.10	0.17	0.28
10Mbps	0.66	0.71	0.66	0.57	0.44
5Mbps	0.07	0.09	0.06	0.05	0.00
1Mbps	0.18	0.06	0.04	0.00	0.00

Figure 5.24 Outgoing Traffic as a proportion of users clustered by type of subscription Data set from 2009-05-01 to 2009-05-14

5.6 Concurrent application analysis

Table 5.2 displays a 5×5 matrix showing the likelihood of usage of concurrent applications from each pair of categories. Every cell in the body of the table represents the probability of the occurrence of column's header applications while people are using the row header's applications. For example, the third cell of first column containing the value of 0.72 infers that: when people are playing online games, there is a 72% of the possibility that they are generating web browsing traffic at the same time. But on the other hand, there is only 7% of chance that the web browsing users are having an active online game client meanwhile. The asymmetry here is due to the weighting by numbers of users of a given application.

Some interesting conclusion can be drawn from this concurrent application table.

(a) P2P is exclusive. It always has the lowest concurrent probability when people are generating other categories of traffic. This might be explained by users only running P2P applications when they are not engaged in other activities, since P2P tends to consume large amounts of bandwidth and would have a negative effect on their experience of other applications. The phenomena can also be explained by the fact that the P2P downloads and uploads may take a long time, and are often performed with the user absent from the computer, e.g. at night, when the PC is left on but is unattended. The other, and probably the main reason, is that many P2P applications get content from the most responsive peers and if the user is actively engaged in an

application that utilizes a large fraction of their bandwidth, then they will be a less desirable peer.

(b) Both web browsing and instant messaging are popular among users and users tend to use them concurrently.

(c) Media streaming applications are strongly tied to web browsing. Media streaming occurs 87% of the time when web browsing traffic is detected. This is partially due to the definition of media streaming because Flash is classified into this category, and Flash is usually embedded in a web page. It may also reason from the fact that media streaming commonly requires a buffering time which causes people to do some short term activity, such as browsing websites or chatting online.

(d) Online gaming has some interesting aspects. It appears that when people are playing online games, 72% of them have web browsing traffic (only slightly less than the 87% of streaming, which was explained in (c)), 48% of these users have some P2P traffic(larger than other categories), 74% of them have an Instant Messaging client active(larger than other categories), 32% of them generate media stream(larger than other categories). The conclusion could be made that when users are gaming they are not focused on accomplishing other tasks, so they do not focus their attention - hence they multitask - by doing several things at a time.

Table 5.2 Concurrent category of applications
(in a 15 min interval, data set from 2009-05-01 to 2009-05-14)

	Web browsing	P2P file sharing	Online game	Instant Messaging	Media streaming
Web browsing	N/A	0.46	0.07	0.67	0.32
P2P file sharing	0.46	N/A	0.04	0.46	0.17
Online game	0.72	0.48	N/A	0.74	0.32
Instant Messaging	0.56	0.39	0.06	N/A	0.22
Media streaming	0.87	0.46	0.08	0.72	N/A

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The goal of this thesis project was achieved. I successfully uncovered the hidden characteristics of residential network traffic as well as observed some patterns of user behavior. A systematic method of traffic measurement and analysis has been developed. The procedure starts by collecting data from PacketLogic, and stores the collected data a data warehouse. This data can be used subsequently for analysis. In order to extract patterns from this large amount of data, Python scripts combined with MySQL queries have been written for each type of analysis. All of my analysis (involving data extraction, programming, graphs and table creation) are built upon open source software.

Thanks PacketLogic's deep packet inspection, we can identify different protocols at the application layer. Additionally, an individual household's traffic information could be recognized by using information from the DHCP logs. Moreover, coordination among ISPs benefitted our project as we were able to cover a large residential area. These factors together provided great opportunity to conduct research on Internet traffic and user behavior.

Several topics concerning Internet traffic analysis have been studied and discussed in this thesis. They are listed below.

• We have investigated aggregated traffic over different temporal periods. It was found that outgoing traffic (uplink for households) always exceeds the amount of incoming traffic (downlink for households). Some fluctuations have been discovered in long-term traffic, these are believed to be the effect of the IPRED Law and the Swedish summer vacation period. Weekly traffic patterns

were examined, but it turns out there is not a big difference between traffic on weekdays and weekends. As these observations were made over a period of 9 months, we considered this result to be a general conclusion. Within a single-day's traffic, we have observed that the peak hours are 17:00-24:00 and that this period accounts for half of the whole day's traffic.

• Models have been established for traffic distributions of the observed households. These traffic distributions follow a long tailed distribution. Several heavy tail model hypotheses have been raised and parameters we fit are gained by using a least-square method. Of the models considered, the Weibull distribution appears to be the better model for incoming traffic, while the Pareto provides the best fit for outgoing traffic. In both directions we observed that a small portion of the households are responsible for most of the network traffic. These two models also applied to data which was captured several months later. Despite the fact that part of the curves have some mismatch (in terms of 'lighter user' for Weibull and 'heavy user' for Pareto), they are the acceptable models for expressing traffic distribution by households.

• By utilizing PacketLogic deep packet inspection function, we identified application level protocols, and thus could identify specific application traffic. Six categories of applications were defined: web browsing, Instant Messaging, media streaming, online game, P2P file sharing, and other applications. The shares for both direction of traffic were shown in Figure 5.17 and Figure 5.18 on page 36. It can be seen that P2P accounts for a little more traffic in September 2009 than in May 2009, which may be due to an IPRED Law bounce back effect. The top five specific application protocols have also been listed in Figure 5.18, Figure 5.19, Figure 5.20, and Figure 5.21 on page 38 and 39, for each category of traffic. BitTorrent is the dominant application in the P2P file sharing category, accounting for 90% of all application traffic classified in this category. Flash over HTTP accounts for roughly 50% of incoming media streaming traffic. But for outgoing traffic, P2P based media streaming such as Spotify, PPStream, and SopCast account for the most network traffic.

• User penetration rate is another interesting topic. HTTP has a penetration rate of 98.1%, indicating its popularity among these users. In the P2P category, BitTorrent is the mostly used P2P protocol, with 85.5% penetration of BitTorrent KRPC protocol, and 53.5% Encrypted Transfer and 39.8% normal transfer. For P2P file sharing applications, tracking protocols generally have a higher penetration than transfer protocols. In category of media stream, Flash over HTTP has 87.4% penetration. Although P2P based media streaming applications accounts for a large share of total traffic volume, their penetration rate is not that high (i.e. the few users that use these protocols account for a large share of the total traffic).

• We have used group analysis methods to delve into user behavior patterns. K-means clustering was used successfully to cluster the outgoing traffic volumes into three subsets. A conclusion has been reached that 'heavy traffic users' are inclined to have longer HTTP active time. Another approach for grouping is to group by type of subscription. These results showed that not all high maximum data rate subscriber households fully utilize their resources, and traffic generated by low bandwidth subscriber households is very limited by their subscription speed. We also observed that outgoing traffic volumes will follow the subscriber's maximum data rates.

• In the section on concurrent application, a table of probabilities has been computed which indicate the likelihood of concurrent applications for each category. From this table we can observe P2P's usage tends to exclude all other applications, web browsing and IM's popularity, a close relationship between media streaming and web browsing, and online game users' behavior. Undoubtedly features of concurrent are not limited to these and additional conclusions can be drawn by changing the parameters for our experiments, such as adjusting the sample interval, adjusting the overall time period, looking for specific applications instead of application categories, and so on. The method we have provided for the concurrent applications explores a new way for Internet traffic user behavior analysis.

6.2 Future work

There are two fields of future work for our project's successor. One is to explore and refine traffic measurements to have more valuable and up to date data for research. The other is to perform more analysis in order to discern more hidden patterns among the available data.

In the field of traffic measurement, keeping track of data is the highest priority. By doing so, temporal consistency of the data warehouse can be maintained, thus providing opportunities for long-term or year-by-year analysis. On the other hand, our data warehouse only stores part of the traffic information which has been captured by PacketLogic. Additional features (such as flow and session information, the two parties in an application protocol, etc) could be extracted by using PacketLogic's Python API. This requires both changes in the data warehouse design and additional Python programming.

For traffic and user behavior analysis field, several tasks remain for future efforts.

• More results can be extracted from the additional data which has been collected, as this thesis was limited to the data stored during a period from March 2009 until the end of 2009. Once the data from year 2010 has been incorporated into data warehouse, contrasts should be made among the same objects at different periods of time. For example, a comparison between categorized application patterns in 2009 with that in 2010. Moreover, an updated long-term aggregated traffic pattern could be presented.

• The model for traffic distribution among households can be refined. Mismatches exist between the two best models we have. In future work, some other long tailed models should be considered. Or a piecewise function is introduced to better describe the distribution.

• Further investigate the relation between P2P protocols' penetration with P2P application penetration. Although we have captured different P2P protocols including tracking and transfer protocols, the real user penetration cannot be estimated. The actual user penetration is not simply the sum of tracker penetration and transfer penetration, or the sum of unencrypted and encrypted

transfer penetration. Protocol analysis is necessary to solve the problem of estimating the actual user penetration for P2P file sharing.

• Explore more hidden patterns among concurrent applications. This thesis analyzed and presented the probability of concurrent applications as grouped in categories. Future work could further examine the relationship between occurrence of different protocols. Moreover, to have concurrent probability as a function of time would be helpful to understand users' behavior. The concurrent probability is likely to vary from morning to evening, but we have not yet examined this. We propose the hypothesis that there is a stronger correlation between P2P file sharing and Instant Messaging during daytime than in the middle of the night, as P2P file sharing applications are usually left and continue to operate without the user's presence throughout the night.

References

[1] B.Stewart, Internet History, http://www.livinginternet.com/i/ii_summary.htm, retrieved April 1st, 2010
[2] Internet World Stats, http://www.internetworldstats.com/stats.htm, retrieved April 1st, 2010
[3] A.Aurelius, TRAMMS Project Information, January 1st, 2007. www.celtic-initiative.org/~pub/Project-leaflets/Webquality/tramms-lq.pdf, retrieved Apr 1st, 2010
[4] Acreo AB, www.acreo.se, retrieved April 1st, 2010
[5] 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
[6] 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
[7] 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
[8] 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
[9] C.Barakat and E.Altman, A Markovian Model for TCP Analysis in a Differentiated Services Network, INRIA, France, 2003
[10] L.Rodrigues and P.Guardieiro, A Spatial and Temporal Analysis of Internet Aggregate Traffic at the Flow Level, IEEE COMM Society, Vol.2, November 2004, pp. 685-691
[11] 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
[12] G.Fox, Peer-to-peer Networks, Computing in Science & Engineering, Volume: 3, pp.75-77, 2001
[13] T.Karagiannis, A.Broido, N.Brownlee, KC Claffy, and M.Faloutsos, Is P2P Dying or Just Hiding, IEEE COMM Globecom, November 2004
[14] 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
[15] 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
[16] C.Liu, W.Day, S.Sun,and G.Wang, User Behavior and the "Globalness" of Internet: From a Taiwan Users' Perspective, Journal of Computer-Mediated Communication, Volume 7, No.2, January 2002
[17] K.Thompson, G.J.Miller, R.;Wilder, Wide-area Internet traffic patterns and characteristics, Network, IEEE , Volume 11, Issue 6, December 1997, pp.10-23
[18] T.Yamakami, A User-perceived Freshness Clustering Method to Identify Three Subgroups in Mobile Internet Users, on proceeding of International Conference on Multimedia and Ubiquitous Engineering, April 2008
[19] F.Molina-Castillo,C.López-Nicolás, H.Bouwman, Explaining mobile commerce services adoption by different type of customers, Journal of Systemics, Cybernetics and Informatics, Volume 6, No. 6, December 2008, pp. 73-79

[20] D.Yinan, Y.Hao, L.Zhenming, Broadband dial-up user behavior identification and analysis, on proceeding of IC-BNMT '09. 2nd IEEE International Conference, October 2009
[21] J.Färber, S.Bodamer, J.Charzinski, Measurement and Modelling of a Internet Traffic at Access Networks, EUNICE Summer School, September 1998
[22] British Telecom, Privacy and Policy, http://www2.bt.com/btPortal/application?pageid=pan_privacy_policy&siteArea=pan, retrieved August 21st, 2010
[23] Verizon, Customer Proprietary Network Information (CPNI) for Telecom Consumers, http://www22.verizon.com/about/privacy/cpnitelecom/, retrieved August 21st, 2010
[24] Emily E. Terrell, Introduction to Module V: The USA Patriot Act, Foreign Intelligence Surveillance and Cyberspace Privacy, Harvard Law School Open Education, http://cyber.law.harvard.edu/privacy/Introduction%20to%20Module%20V.htm, retrieved August 21st, 2010
[25] Mats Lewan, Swedish ISPs vow to erase users' traffic data, CNET News, April 28, 2009, http://news.cnet.com/8301-1023_3-10229618-93.html, retrieved August 21st, 2010
[26] Nick Duffield, Sampling for Passive Internet Measurement: A Review, Statistical Science 2004, Vol. 19, No. 3, pp. 472–498
[27] Gregor Maier, Anja Feldmann, Vern Paxson, Mark Allman, On Dominant Characteristics of Residential Broadband Internet Traffic, on proceeding of the 9th ACM SIGCOMM conference on Internet measurement conference, November 2009, pp. 90-102
[28] C.Lagerstedt, M.Kihl, A.Andreas, A.Berntson, and T.Westholm, Measuring and Modeling HTTP Streaming in IP Access Networks, May 2010, http://www.acreo.se/upload/Partners/TRAMMS-Measuring-Modeling-HTTP-Streaming-IP-Access-Networks.pdf, retrieved August 21st, 2010
[29] Procera Networks, DRDL Technology,http://www.proceranetworks.com/drdl-technology.html, retrieved August 21st, 2010
[30] U.S. Commerce Department NIST, Engineering Statistics Handbook, June 2003, http://itl.nist.gov/div898/handbook/pmd/section4/pmd431.htm, retrieved August 21st, 2010
[31] B.A. Mah, An Empirical Model of HTTP Network Traffic, INFOCOM '97. On proceeding of Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies, Vol.2, April 1997, pp.592-600
[32] BBC NEWS, Piracy law cuts internet traffic, April 2009, http://news.bbc.co.uk/2/hi/7978853.stm, retrieved August 21st, 2010
[33] P.Vasile and C.Roxana, On Composite Models: Weibull-Pareto and Lognormal-Pareto, A comparative study, Romanian Journal of Economic Forecasting, Volume 3, June 2006, pp.32-46
[34] P.Hoang, Internet Traffic Analysis, Master thesis, April 2010
[35] H.Schulze and K.Mochalski, Internet Study 2008/2009, Ipoque, http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009, retrieved October 28th, 2010
[36] R. Agrawal, T. Imielinski, and A. Swami, Mining Association Rules Between Sets of Items in Large Databases, Proceedings of the 1993 ACM SIGMOD international conference on Management of data, June 1993, pp.207-216

TRITA-ICT-EX-2010:276

www.kth.se