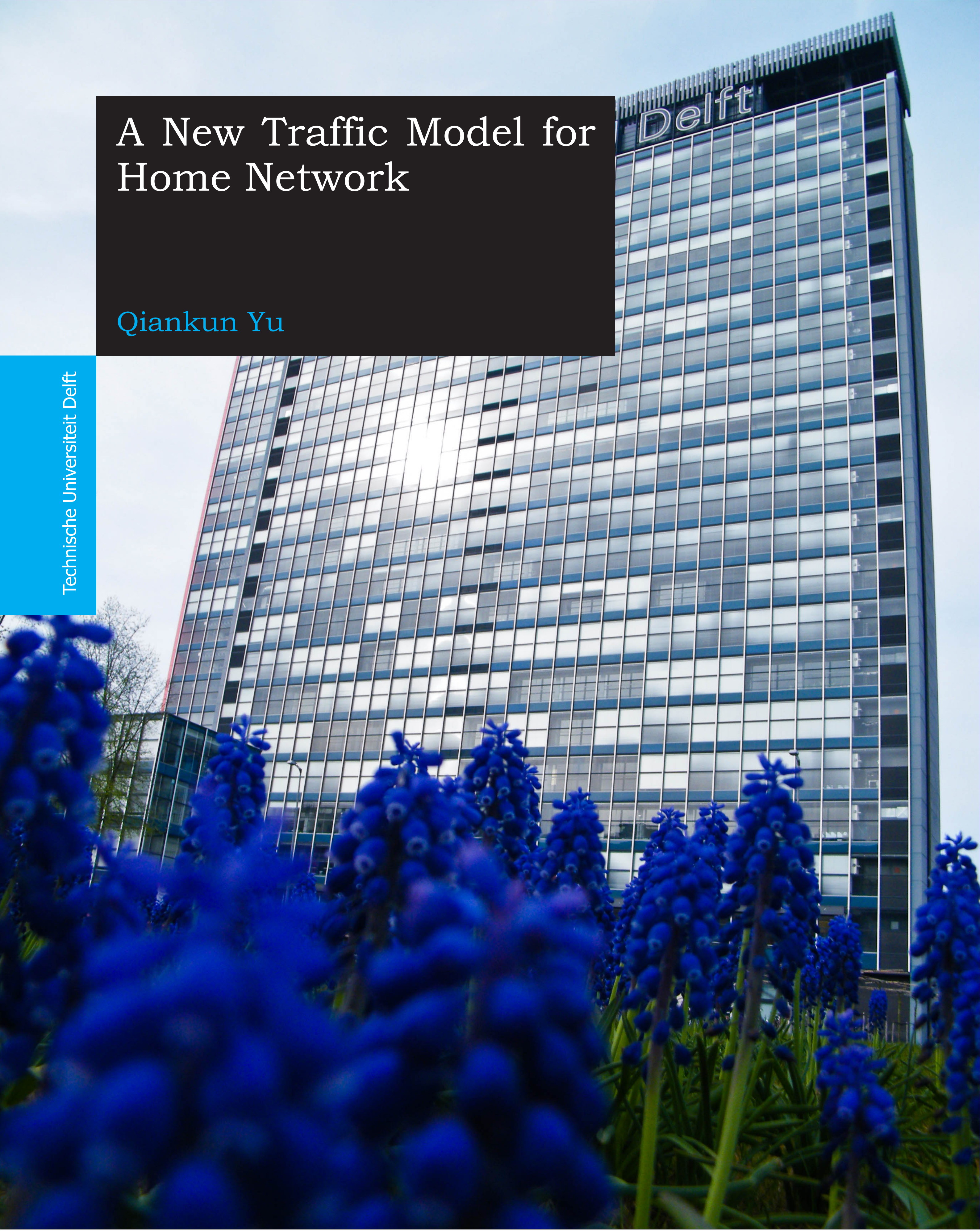
A New Traffic Model for Home Network

Qiankun Yu

Technische Universiteit Delft



TU Delft
Delft
University of
Technology

Challenge the future

A New Traffic Model for Home Network

by

Qiankun Yu

in partial fulfillment of the requirements for the degree of

Master of Science

in Electrical Engineering
Telecommnunication & Sensing System

at the Delft University of Technology,

to be defended publicly on Thursday October 27th, 2016 at 14:30 PM.

Student number:	4402774	
Supervisors:	Dr. Ir. Huijuan Wang	
	Dr. Sylvie Dijkstra-Soudarissanane	
	Ir. Shuang Zhang	
Thesis committee:	Prof. dr. A. Hanjalic,	TU Delft
	Dr. ir. Huijuan Wang,	TU Delft
	Dr. Sylvie Dijkstra-Soudarissanane,	TNO Corporation
	Dr. ir. Sicco Verwer,	TU Delft

TU Delft
Delft University of Technology

Preface

This thesis would not have been possible without the help and support of many people. I want to thank Dr. Huijuan Wang for supporting and guidance in this procedure; and Dr.Dijkstra-Soudarissanane for giving me the chance to do my graduation thesis with TNO and guiding my research by providing valuable inputs during our meetings. I would like to thank Ir. Shuang Zhang for providing an industry insight and being constantly involved in my research.

I want to thank my dear parents, who give me a lot of love and care during this thesis. I would like to thank my friends, who give me a lot of encouragement in the this period. This includes particular Ang, Yaweg, Aaron, Zhenyu ,Yanbo, Quan, Boning.

Qiankun Yu
Delft, October 2016

Summary

Home network is heterogeneous and dynamic, which consists of an indefinite number of devices of various kind, running applications of distinct kind, such as IPTV, online gaming. With an increasing demand of these devices, home network provide a crucial ”last mile” to the Internet. To ensure better Quality of Service of different applications, providing a foreknowledege of home network traffic behavior is essential to help solving this issue.

In this thesis, we firstly characterise home network traffic by calculating the probability density and discover small traffic rate($\leq$ 40kbit/s) has a long-tail characteristic. Moreover, according to all traffic measurements in 15 households, 92.85% are smaller than 1 Mbit/s, which shows home network has long period of low bandwidth consumption and short period of high bandwidth consumption.

Secondly, considering the given incoming traffic are the sum traffic of various applications and background traffic, which used to self-contain program update and daemon running. Ignoring the background traffic in each home network, we record the new traffic data larger than background traffic and notate as nonzero traffic. Taking advantage of Gaussian mixture model, we could learn the probability density at each sampling points(index) of nonzero traffic data. Then we defined the index with the highest probability density in a Gaussian component as the peak point, the 1-sigma area around this point as a burst. And the 95th percentile bandwidth is taken as the height in this burst. Then we can characterise all traffic measurements with different bursts.

Thirdly, based on the idea of burst, we develop a new traffic model based on Markov chain methodology to predict where will the next peak point of a burst happen, the width of the next burst and the how much bandwidth will be needed. At the end of this model, leave-out-one cross validation(LOOCV) are used to validate our predict model.

Last but not least, due to the evolution of new services and applications, we establish an experimental platform to measure new traffic data in a controlled home environment in TNO lab. We register all activities details happened during the experiment and analyze the effect of several traffic activity scenarios on Internet traffic and in-home traffic. Later we discover devices with wireless connects consumption more bandwidth than the wired.

Key words: Home network Burst Gaussian Mixture Model(GMM) Markov Chain LOOCV Incoming Traffic in-home traffic

Qiankun Yu

Delft, Octobor 2016

Contents

1

Introduction & Contributions

1.1. Problem Description

With the fast development of technology, networking has been rapidly adopted by many households. Nowadays, a household comprises a variety of connected devices not only including personal devices like laptops, smartphones, but also game consoles, printers, IPTV. These devices connect among themselves and share a broadband connection to the Internet via a local area network(LAN), which is called in this thesis a home network. The demand for new broadband application and services, such as online gaming, video streaming, email and file transfer, etc., has opened a new era for telecommunication network across the world. When home users perform different applications simultaneously, for example, one user watches a video on Youtube, another watches IPTV and another person plays online gaming at the same time, then it's possible that one user will receive some delay to continue the activity. The quality of these new services, however, depends critically on the performance of the home network, known as the crucial last mile [1].

In telecom networks, normally different home networks are connected to one telecom cabinet, which owns a fixed bandwidth capacity. Since each home owns different kinds and number of digital devices and the connections of these digital devices differ from different households, home networks are known to be dynamic and technologically heterogeneous. Usually, a household is provided with Internet coming from the central office, or via a street cabinet.The operator deploys its network with a certain risk margin, where it assumes that not all households are making use of the full bandwidth. The worst case scenario would then happen when all households connected to one cabinet is suddenly asking for the full allocated bandwidth. To avoid this problem, if the worst case scenario happens too often, the operator needs to split the cabinet or assign households to other boards.

To solve this problem, we need a better resource management within the resource-scarce home network, which could benefit of both home network users and operators. To achieve this target, a new home network traffic model is needed to help customers better understand their network usage and let operator could dynamically and optimally allocate bandwidth to gain more profit as well. This thesis is part of TNO home network research.

1.2. Scope and Limitations

The scope of this thesis focuses on analysing home network behavior through the aggregated services running on various home devices. By characterising the measured traffic data, we aim to propose and develop a traffic prediction model to give service providers and vendors an insight of the network usage in different households so that Internet service providers could better manage bandwidth resource to safeguard the QoS of home network.

However, one limitation of this project is that characteristics of home network traffic are largely unknown owning to the difficulty to record traffic behavior beyond the residential gateway. The second limitation is the complexity of conducting traffic measurement inside home. Besides, from the view of privacy and the regulatory constraints, it is extremely difficult to obtain the necessary permission for both operators and home network users. Thus, collecting useful data to build a prediction model is a hard task. The third one is that the data set of measured home network traffic is relatively small comparing with the number of home networks, and has district restriction (only in the Netherlands), which may not be sufficient enough to validate the prediction model.

1.3. Objectives and Research Questions

1.3.1. Objectives

The main objective is to propose and develop a new traffic model to predict the traffic behaviour in the future and provide the Internet service provides(ISP) a rough idea that what is happening inside a home network.

1.3.2. Research Questions

1. Unlike other kind of network traffic, one significant feature of home network traffic is the small traffic bit rates. Apart from this, what kind of characteristics does this kind of traffic has?

2. Home network traffic strongly depends on Internet habits of home users. So to develop a prediction, with which aspects can we start to solve this problem? And how to do that?

3. Whether is it possible to implement a controlled environment in the lab to simulate some

ofter used applications and services to let Operators roughly know what is happening inside a home network?

1.4. Methodology and Contributions

To achieve the objectives in this thesis, a new traffic prediction model is proposed and developed using Gaussian mixture model(GMM) and Markov chain methodology. The main contribution of this thesis can be identified as:

- Combining with Gaussian mixture model, we define a new concept of *burst* to characterise home network traffic.
- Based on the idea of burst, propose a new traffic model, which can systematically predict when traffic is most likely to happen, how long it will last and how much bandwidth will be required .
- Leave-out-one cross validation method is used to validate the prediction model.
- Establish a experimental platform of home network traffic in the lab and perform some often-used application and services to check the effect bringing to the home network.

1.5. Thesis Outline

This thesis includes three main parts: characterise home network traffic measurements, develop a new traffic prediction model and traffic measurement in a controlled home network environment. The detailed of the rest chapters are explained as blow:

Chapter 2 Background: We introduce the background of home network and current researches of home network traffic.

Chapter 3 Analysis of Home Network Traffic: We characterise the raw home traffic measurements and apply the concept of "burst" to model traffic measurements based on Gaussian mixture models(GMM).

Chapter 4 A New traffic prediction model: By means of discrete-time Markov chains and *burst*, a new traffic prediction model is developed to predict when traffic is most likely to happen, how long it will last and how much bandwidth a home network will use during this period. At the end of this chapter, leave-out-one cross validation(LOOCV) is applied to validate the prediction model.

Chapter 5 Controlled Home Environment Research: we establish an experimental platform for home network to monitoring traffic in a controlled environment. By register all activities, we compare two kinds of traffic inside home network in different scenarios and conclude that different connections of home devices has different requirements of traffic bandwidth.

Chapter 6 Conclusion & Future Work: This chapter summarizes the work we have done in this thesis and addresses some other aspects of home network traffic could be improved in the future.

2

Background

In this chapter, we introduce three characteristics of home network and study the current researches of home network traffic from the aspect of its features and traffic monitoring methods. In section 2.4.2, we discuss various session definitions and implement different session time calculation methods of home network based on a relevant research. Finally, an overview of home network traffic prediction model is presented in section 2.5.

2.1. What Is Home Network?

Home networking is a supportive system to share information and services among different network users and devices as well. As networking and telecommunication systems are gradually being used in many different environments, there is an increasing demand for an automation of processes and monitoring of activity. Most users seek for versatile devices that enable easy control and easy access to/from the outside world. Moreover, they are interested in how much information they retrieve and analyze from their home network. To ensure a flexible yet controllable and stable network, a user very often needs to have a home network that is comprehensive and well understood. The home network is the ”last mile” of a network, but also the most complex node of a network. Therefore, a thorough study of the behavior of a home network is needed.

Different from other kind of Internet networks, such as enterprise network, public network and ISPs, home networks has three prominent characteristics [2]:

Local administrators. Public networks or data centers usually have one or more professional network administrators to manage, secure and monitor networks. Home networks usually are not locally managed by an administrator. Instead, the Internet provider usually manages the Internet connection delivered to the home from a central office or street cabinet. Any problem occurring in a home network is most often managed by the home user, who is often not certified.

Heterogeneous topology. For different homes, diverse digital devices are interconnected using a combination of wired and wireless communication technologies, such as Ethernet, UltraWide Band(UWB), MoCA, HomePlugAV over coax cables or powerlines, HPNA over coax cables or phonelines, etc.. For this reason, it's hard to identify how many devices are connected to the Internet and how many networks are existing in a home. Due to the peculiar characteristics and distinctions of wired and wireless communication technologies, home networks are anticipated to be heterogeneous.

Privacy consideration. If a user is in an enterprise network, his activity is mostly likely monitored, therefore there is no privacy. Home users have higher expectations to their privacy. Most of home users are sophisticated with the network routines of their own home. Thus, it's possible that they don't have a centralized management or control of their devices and information.

These three factors increase the number of home network faults and difficulty for services provides and vendors to isolate the problems. Moreover, most householders are deficient in settling network problems, which makes them frustrated.

In this thesis, we define a home network as the network of different home devices connected to the Residential Gateway (RG). These devices usually have different types of interfaces, such as LAN (Ethernet, Telephone cables) and WLAN (WiFi, Bluetooth). Figure 2.1 depicts an example of a simple home network.

2.2. Home Network Traffic

Home network traffic has similarities with other traffic of other network, like stochastic and dynamical changes with time and space, also self-similarity of network traffic and long range dependence. Apart from these common points, home network has several peculiar properties. The first one is the total traffic volume is relatively small comparing with other networks. The second feature of home network traffic is home devices seldom send large volume of traffic at the same time. It means home networks have long period of low network activities, which are dominated by small sets of home equipments. This feature results in that fact that network traffic are with high randomness and distributed far from the average traffic value.

A relevant home traffic studies done by Reggani A. et al. [3]is to compare the local and wide-area traffic based on the traffic data from multiple kind networks (also including home networks). They point out that most of traffic in home network is wide-area traffic, while the majority of users has more local traffic at work. Meanwhile, for example, from the view of the application mix, P2P traffic are existed in home networks but not at work network. K. Cho et

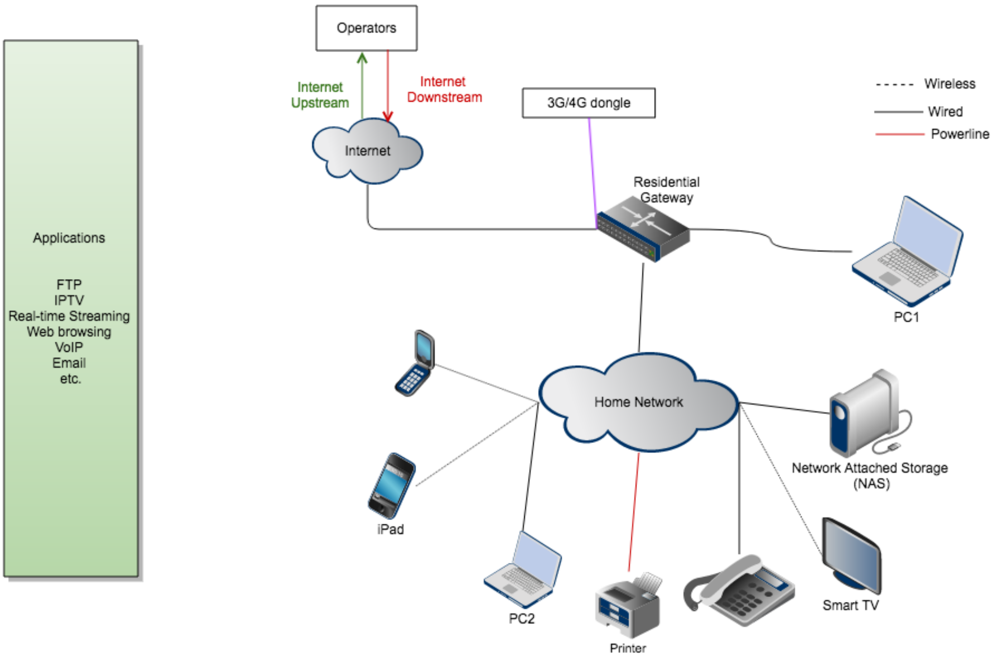


Figure 2.1: *A simple Home Network Overview*

al. [4] examines the growth of residential user-to-user traffic in Japan, a country with a high penetration rate of residential broadband access, and studies the impact of these traffic on usage patterns and traffic engineering of commercial backbone networks.

Karagiannis T. et al. [5] studies home network traffic at the packet-level by correlating with diaries capturing user experiences with three home networks, and they show that the lack of cooperation across applications in a home network leads to observable performance problems and associated user frustration. Similar to home network, G. Maier et al. [6] passively collects packet-level traffic data of residential networks at aggregated routers of a large Internet service provider, and analyzes the performance of dominant characteristics of residential traffic from different applications and their impacts on the overall traffic. These studies are conducted from outside home networks and focus more on packet-level from the view beyond the Residential Gateway.

Many studies have been done in the view of the inside of home network as well. One representative studies deriving from inside home network is done by K. Xu et al. [7]. They study the characteristics of home network mainly from two aspects: IP address and application ports at different measurements. By statistics analysis of source and destination traffic, they know whom home users contact mostly or which website home users visit frequently. Taking advantage of Principal Component Analysis (PCA), they analyze the traffic pattern among network applications and find the cluster of application ports exhibit explicit temporal correlations of traffic.

2.3. Home Traffic Monitoring

Network traffic monitoring is a network management process to inspect, administer and analyze the performances of network traffic [8]. The aim of network traffic monitoring is to provide accessible and smooth operations inside a network. In this procedure, network sniffing and packet capturing techniques might be also applied. Moreover, traffic monitoring enables better network management, admission control, adaptive QoS policies, SLA agreement, adaptation to improve monitization of network consumption and so on. Many traffic parameters, such as capacity, Internet throughput, latency, traffic rate, etc. could be conducted for future analysis, including traffic prediction.

Home network performance depends on a comprehensive overview of all home network traffic. Various kinds of network monitoring tools and techniques are proposed to study network performance. For home network, traffic monitor tools can be divided into two categories: end-host based monitors and gateway-based monitors. HostView is an end-host based monitoring tool used to collect traffic data directly operated on users' computers [9]. In spite of measuring specific performance indicators, like RTTs, data rate, TCP resets, TCP retransmission, etc., a markable feature of this tool is to prompt the feedback from users on network performance. HomeMaestro [5] is also a host-based traffic monitoring system that ensure home network monitoring at application level without changing the network configuration and protocols. It could detect whether applications competing traffic flow would cause network problems by tracking traffic rate and latency. This kind traffic monitors, however, are not preferred for low resource consumption but rather for high accuracy with by investing more resources [10]. Besides, for high heterogenous equipments inside homes, it's difficult for end-host based monitors to deploy on all possible devices. Comparing with previous monitors, gateway-based traffic monitoring tool are optimized when tracking network performance from an overall view.

Srikanth S. et al. [11] conduct a research about home network traffic monitoring among United States home users and measure Internet access link performance from the router by taking advantage of SamKnows and BISmark deployments . First, they respectively record the throughput details of different access links with different users and same Internet Services Provider (ISP) via these two softwares, then drew the conclusion that different measurement methods focus on different aspects of throughput. Moreover, according to the upload or download throughput with different ISPs, they discover that the upload throughput is more consistent due to the low bit rate comparing with download throughput. Second, they measure two latency metrics of networks: last-mile latency (the latency to the first hop inside a ISP's network) and latency under load (the latency of user experiencing during a upload and download). SanKnows is used to measure end-to-end latency under load, while last-mile latency under load was measured on BISmark platform. It is concluded that last-mile latency and jitter were affected by access technologies and buffer size has an effect on latency under load. Despite of these, they emphasized that home network equipments could apparently affect traffic performance, but didn't tell the specific impacts.

Another traffic monitoring tools is a bandwidth measure tool PRTG developed by Paessler, which represents data using straightforward graphs and tables with the assumption of supporting SNMP [12]. One application of PRTG is in home networks done by A.Delphinanto et al.[13]. The tool is placed at the home gateway to monitor traffic between different interfaces and offers the possibilities to monitor traffic with different type of UDP, HTTP, etc.. It recorded the byte numbers going through WAN, WLAN and Ethernet port in seven days and extracted incoming and outgoing traffic of each interface from the router. In this project, the author also pointed out that PRTG could only monitor IP traffic. Another attractive traffic monitoring method is achieved via a real-time behavior monitoring system HomeTPS [7, 14]. This system, which operates on a separate server, collects and analyzes network traffic among digital infrastructures in home networks through programable home routers. They utilized the platform to monitor the overhead of traffic data, incoming traffic and outgoing traffic of one home network. And they showed the possibility that using this traffic monitoring platform could help analyze traffic data exchanged not only between home devices and Internet end hosts, also among home network devices, which assists home network users to deeply understand their network usage and figure out the anomalous of network traffic.

2.4. Session and Duration

2.4.1. Duration and Application Sessions

Two important terminology related to traffic measurements are introduced in this section: duration and session. Duration is the length of time it will take to complete an activity or interaction. It could be short to milliseconds or long enough to hours, days. It depends on the specific situation. For the future research on traffic modeling, if the traffic at the session level, then it should involve start time and duration of each session.

There are fruitful definitions of a session. For instance, a user session begins when the user logs in to Secure Global Desktop (SGD) and ends when he logs out. For an application session, it begins when a user start an application and ends when the application exists. And each application session corresponds to an application running on SGD. Thus, the time a session lasts differ according to different session perceptions. Some connections and sessions last only long enough to send a message in one direction. For a connection-oriented network like ATM, a session can be a call established and terminated by a signaling command. However, other sessions may last longer, usually with one or both of the communicating parties being able to terminate it. For Internet applications, each session is related to a particular port, a number that is associated with a particular upper layer application. For instance, a client builds a TCP connection to port 21 on a server to establish a FTP session and the connection will be closed when the session is completed.

In telecommunication, a session is a range of information exchanges between two communication points that happen during the span of a single connection [15]. Take a simple scenario for example, one end point A requests for a connection with another end point B, when end point B replies an agreement to end point A, A will take turns to exchange signaling and data with B. The session begins since the connection is established at both ends and stops when the connection is interrupted, which is associated with the human activity. In a computing world, a session is a limited time to communication between two systems [16]. In OSI model, the session layer controls the connections between two end points, including establish, manage and terminate the connections [17]. In a parallel and distributed computing environment, a session is an abstraction of diverse forms of ”structured communication” [18].

A common client-server session type is a HTTP session on application layer. A HTTP session is set up by a Web browser every time a user visits a website. Browsing each page constitutes an independent session. A session is the entire time a user spend on the website, that is the period from the user’s first arrival at the web page until the time he leaves. Another simple example is an email or SMTP session. When a user uses an email client, like Apple Mail or Outlook, the user initiates an SMTP session. In this procedure, many interactions are processed, for example, sending the account information to the mail server, receiving and downloading new emails or messages from the mail server. Once the downloading is finished, the session is complete. Online chat or instant messaging session is another case between two individual computer, called peer-to-peer communication (P2P). However, the roles of both computers are neither server or client. One application of P2P is BitTorrent file sharing. In the BitTorrent network, file downloads consist of more session with other computers. A P2P session ends when the connection between two systems are terminated.

2.4.2. Home Network Session

The above mentioned sessions are respectively corresponding to different applications or services, which could be monitored and measured in public. However, for home networks, due to the privacy constraint, it’s impossible to monitor different sessions of user’s applications. To remedy this problem, we physically define a session in a home network as a related progression of events devoted to a particular activity, which intuitively relates to home user’s actions and our purposes.

A relevant home network research done by A.Delphinanto [13] propose an empirical method to calculate session period, which can be described by firstly calculating the changes between two consecutive traffic samples. They suppose an average precision of existing bandwidth tools for home network equalling to 250 Kbit/s, then a session period is regarded as the distance between two absolute traffic changes larger than 250 Kbit/s. One advantage of this method is to remove any over- and under-estimation to accurately characterise traffic rates.

At last, they discover the distribution of application session time in home networks follows generalized Pareto distribution and the average session time is 500s, which shows the similar results as in a special public wireless network done by Balachandran et al. [19]. Balachandran point out that most users in a public wireless LAB process very short session period (less than 10 minutes) and the distribution of session time in this kind network also follows generalize Pareto distribution.

On basis of A.Delphinanto's work, three session time calculation algorithms are proposed by TNO colleagues. In these algorithms, the idea of traffic regions[1] is considered to weaken the heterogeneous of home network traffic. Taking advantage of connected-component labeling algorithm, all traffic measurements in each households are classified into different traffic regions.Positive and negative traffic changes between two neighbouring traffic samples stand for the up and down traffic,respectively. Then, three algorithm of session time calculation can be concluded as follow.

- A session time refers to the sample interval times the distance between the i-th time stamp from the beginning of up traffic and the i-th time stamp from the beginning of down traffic, called First Up First Down(FUFD).
- A session time refers to the sample interval times the distance between the i-th time stamp from the beginning of up traffic and the i-th time stamp from the end of down traffic, called First Up Last Down(FULD).
- A session time refers to the sample interval times the distance between the i-th time stamp from the beginning of down traffic and the i-th time stamps from the end of up traffic, called First Down Last Up(FDLU).

We implement the three session time calculation algorithms together with A.Delphinanto's methods and find when increasing the average precision of bandwidth, the average session time has the worst performance in A.Delphinanto's methods, while the rest three algorithms has the similar performance.

2.5. Home Traffic Prediction

As network infrastructure provides limited resources and users have dramatically increasing requirement of network, it's much better to balance the contradiction and forecast traffic flows and information volume in networks before arriving than adjusting for them later. From users' views, prediction helps them to have an optimum management and control of the networks in advance. And for operators, based on predictions, they can effectively allocate network resource to analyze or forecast the network state to timely adjust real-time polices for max or

[1] Regions refers to a group of IP endpoints that share common characteristics and resources [20].

optimal network resource consumption, such as increasing network utilization, safeguarding user's perceived network QoS, etc. Considering detailed functions, traffic rate prediction includes many kinds of aspects, for example, predict traffic state or traffic distribution in the next period, or the range of traffic rates volatility. From prediction types, it can be partitioned into long and short period prediction. Long period prediction provides more traffic details, making more accurate planning and management possible, while short period prediction helps to improve the QoS, network congestion and resource control [21]. Additionally, from specific techniques, traffic prediction includes training-based technique and non-training-based technique. Obviously, the difference between them is whether it has training phase which is used to recognize the model parameters.

Choosing appropriate traffic parameters to build an accurate prediction model is a key point for traffic prediction. For dynamical network traffic, direct and indirect traffic prediction approaches are considered to achieve the target [22] . The indirect traffic prediction approach indicated a stochastic model of network traffic. Forecasters pick up some related prediction models, and calculate parameter values of each model. Then test how well the estimation model could suitable for the observed data. By contrast, direct traffic prediction approach is more accurate and challenging due to the derivation from sufficient traffic data.By considering the self-similarity and long-rang dependence of network traffic, these characteristics increase the difficulty to build a traffic prediction model.

H. Yin et al. proposed a new time series model, named adaptive autoregressive(AAR), based on network bandwidth usage in dynamical networks. The model shorten the shorten the memory transformations of traffic data and fit ARMA model to the transformed series, which is useful to predict self-similar network traffic and increase the prediction accuracy. Additionally, this model decreases the complexity of computation and provides a more precise approach to forecast the dynamic network traffic.

Another traffic prediction approach is an extended fractional Brownian traffic (efBt) model based on multifractal modeling [23]. From the whole properties of traffic processes, multifractal can observe different traffic behaviors in diverse time scales, and better describe the irregularities of traffic processes. The fractional Brownian traffic (fBt) model is a self-similar model to depict irregular signal with bursts and long range via a fixed Hurst parameter H (H indicates a certain autocorrelated relation [24]), and describe the accumulated traffic volume using the fBm process. The efBt model overcomes the limitation of capacity in fBt model and use *Hőlder function* $H(t)$ to represent Hurst parameter H. Through simple calculation with these parameters (*Hőlder function*, traffic mean and traffic standard derivation), they received the volume of the efBt traffic. Based on the extended fractional Brownian traffic model, they designed a novel predictor via FIR Wiener Filter.

The most relevant study is done by A.Delphinanto [13] to predict home network traffic fluctuation in the next certain period based on entropy and discrete-time Markov chain methodology. The average session time mentioned in section 2.4.2 gives an insight of the duration of different network applications. Then all traffic measurements are evenly divided into different period by this parameter. In each entropy period, traffic data is transformed to entropy. Taking normalized entropy as the input state of Markov chain, the transition probability matrix is obtained to estimate the probability of future traffic states. One advantage of this model is the home gateway will not be required to allocate resources to store any historical traffic data. Based on this model, we verify the state division of Markov chain using dynamical time warping technique and find that higher similarity exists in traffic measurements with higher fluctuation, detailedly explained in Appendix A. One shortage in this model is the prediction result just tell the fluctuation of future traffic, not the traffic bit range of the fluctuation in the next entropy period. Another shortage is the session time follows a generalized Pareto distribution as mention in section 2.4.2, so that the average of all application session time is not enough to indicate the duration of network application usage.

2.6. Conclusion

In this chapter, we give a brief introduction to home network and detailedly present the current researches on three aspects of home network traffic: 1)current techniques to monitor traffic; 2)different application sessions and our definition of session in home network;3) traffic prediction models for dynamical network and time series traffic data. Based on these researches, we build our new traffic model in the following chapters.

3

Analysis of Home Network Traffic

In this chapter, we explain the home network traffic measurements and analyze the measured data which will be used in our prediction model. Considering the very nature of a home network bearing low network activities in long period of time, we use the idea of ”burst” to model this kind of traffic based on Gaussian Mixture Model(GMM). To the best of our knowledge, this method have never been applied to home network traffic.

3.1. Home Traffic Measurements

We start from introducing the home network traffic measurements used in this thesis. Take a simple residential gateway for example(Fig.3.1). Port 1,2,3,4,5 are Ethernet port and port 6 is a wireless access port. Port 1 is connected to Internet and port 2 is configured as access-point mode, connecting to a PC which runs traffic monitoring tool. The PC should be kept separated from the rest of network to avoid the inference. And PRTG is used to measure the bandwidth of the incoming and outgoing traffic of each port except port 2 with the minimum time interval 10s. For different home networks, home devices are conncted through Ethernet cable and powerline in port 3,4,5 and Wi-Fi in port 6.

Notate the incoming traffic at port i as $Port_{incoming}(i)$, $(i = 1,3,4,5,6)$, then the total aggregated traffic $B_{incoming}$ is the sum of incoming traffic from all ports(except port 2), formulated in Eq.3.1:

$$B_{incoming} = port_{incoming}(1) + \sum_{i=3}^{6} port_{incoming}(i) \tag{3.1}$$

The same configuration of the residential gateway was implemented in 15 households and

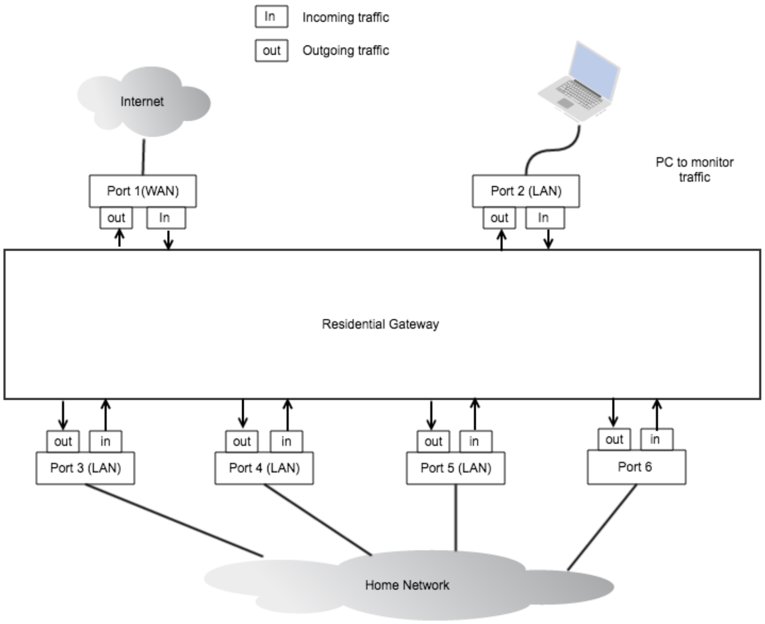


Figure 3.1: Configuration of the residential gateway.

all traffic was measured for one week. Though Eq.3.1, the incoming traffic in each household is calculated, which is the raw traffic bit rates(the unit is Kbit/s) we use in this thesis.

According to all traffic measurements, the maximum traffic bit rate is around 92 Mbit/s and 92.85% of all traffic measurements are smaller than 1 Mbit/s. We calculate the probability mass function(PMF) of all incoming traffic bit rate in 15 households with bin size of 1 Kbit/s shown in Figure 3.2. It represents the histogram result of traffic bit rates smaller than 40Kbit/s. We use power law and generalized Pareto distribution(GPD) to fitting this kind of data. The cyan line shows the power law fitting results, explained in Eq.3.2.

$$pmf(x) = a \cdot x^{b}, \quad a = 0.24, b = -1.113 \tag{3.2}$$

where x is the aggregated incoming traffic. The fitting results of generalized Pareto distribution in red dotted line we obtained is same as the result in [13], which is formulated in Eq. 3.3

$$pmf(x) = \text{gpf}(x|k,\sigma,\mu) = \left(\frac{1}{\sigma}\right)\left(1 + k\frac{x-\mu}{\sigma}\right)^{-1-\frac{1}{k}} \\ (k = 0.56, \sigma = 3.1\text{kbps}, \mu = 0\text{kbps}) \tag{3.3}$$

where, k is the shape parameter, σ is the scale parameter, μ is the threshold or location. Both two fittings reach good fitting results with Root mean squared error(RMSE) of 0.0473(GPD) and 0.022(power law). These results prove that smaller home network incoming traffic has long-tail characteristic and home network traffic are long period with low network activities and short period with high traffic bit rates.

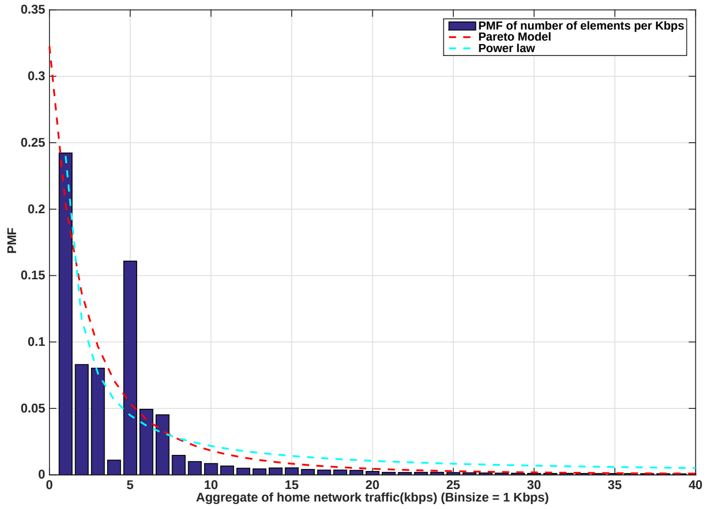


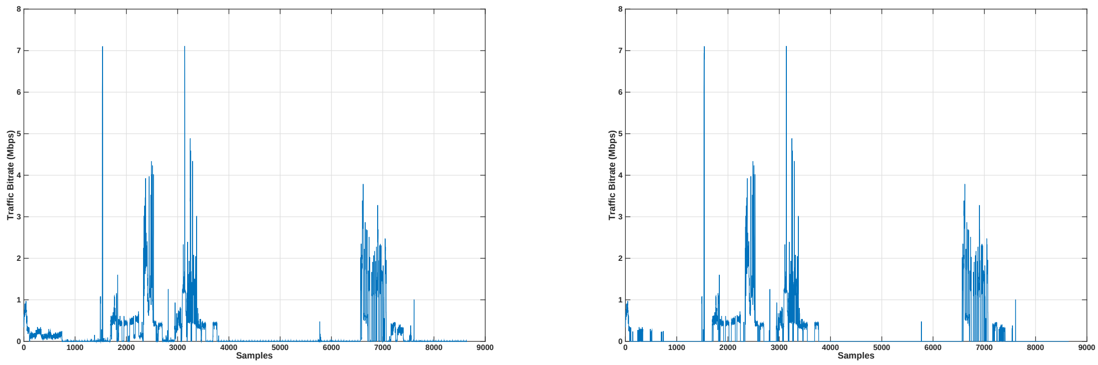
Figure 3.2: *Probability mass function of the aggregated incoming home network traffic among 15 households.x axis is the aggregated incoming traffic, with the unit of Kbit/s, y axis is the probability of a possible traffic bit rate.*

3.2. Motivation of Data Modeling

As mentioned in section 3.2,the given home network traffic data has the same number of samples, but the measurements started at different time during a day. We evenly divide all traffic samples of each home network into seven equal parts as they were acquired over 7 days precisely. Each day is defined as one cycle. In each cycle, the period is 24 hours and the number of traffic samples is $60 \times 60 \times 24 \div 10 = 8640$. Hence, the same ordinal of the sample in different cycle in each home corresponds to the the same time on different days. From the whole view of given home network traffic data, we find traffic bit rates have different magnitudes and little similarities among 15 households. Vertically comparing, traffic bit rates of different cycles within one home network, the same conclusions are reached.

How to extract useful information from the set of heterogeneous traffic data to conduct a practical prediction seems to be an impossible assignment. Back to the root of the question, in the view of particular feature, we are thinking of whether it's possible to pick up relative high traffic in a home network and predict their behaviors.

The answer is yes. Thanks to a related study about the cascades of information-resharing on Facebook done by J.Cheng et al. in [25]. They suppose a recurrence occurs when a peak of image resharing number is observed in the time series. Additionally, this peak should last a certain number of days and the height of each peak must be larger than certainer thresholds. Then, the idea of using *burst* to characterise a recurrence is proposed. By only characteristics of a cascade' s initial burst, strong performance in predicting whether and when it will recur in the future is demonstrated. Motivated by this research, we intend to use "burst" to model our traffic measurements.



(a)

(b)

Figure 3.3: *Raw traffic measurements(a) and nonzero traffic measurement of cycle 4 in home network 1. x axis is the sample index, y is the traffic bit rates.*

3.3. Traffic Data Processing

To model our traffic measurements with the idea of *burst*, The first step is data processing. The aggregated incoming traffic through the Resident Gateway (RG), are the sum traffic of various applications and background traffic. The background traffic as defined in this project, refers to the traffic used to self-contain program update and daemon running. This part of traffic is regarded as basic network bandwidth demand to support normal operation of a home network. Due to the distinction of bandwidth usage in different home networks, for simplification, we suppose that the background traffic is equal to the mean traffic rate of all traffic measurements in each home network and traffic measurements smaller than this threshold can be ignored, converting to zero. Then, a new set of traffic bit rates is obtained with same size of raw traffic measurements, notated as nonzero traffic, which is the basis to model traffic behavior.

Take one period of traffic measurements for example. Fig.3.3a depicts the raw traffic bit rates of cycle 4 in home network 1. The background traffic in home network 1 is equal to 157.07*Kbps*. Ignoring traffic bit rates smaller than this threshold, the nonzero traffic in this cycle is plotted in Fig.3.3b. We find in this period when a peak traffic happens, some relatively small traffic are around this sample. And for some small traffic bit rates, it lasts a relatively long period. This phenomenon also appears in other households. When applying the burst definition as presented in [25] to our nonzero traffic, it turns out hard to scientifically set reasonable thresholds of a burst, such as how larger the traffic bit rates can be regards as the peak traffic, the lowest bound of burst width,etc., to model the heterogeneous traffic data.

However, an alternative view of the matter is from time stamps of nonzero traffic data. Each index of nonzero traffic represent one specific sampling time stamp. Ignoring the traffic magnitudes, we extract indices of all nonzero traffic data. By learning the relation among these indices, the index, which has the highest probabilities in a particular period, can be taken as a peak point. Thus, the concept of a burst in this thesis can be described as a certain area around this peak point, representing traffic is more likely to happen comparing with other periods.

To achieve this aim, modeling nonzero traffic data based on *burst* has turn out to be a probability density estimation problem. Based on all indices of nonzero traffic bit rates in 15 households, Gaussian mixture models(GMM) provide an effective way to extract all peak points in each household.

3.4. Model Traffic data using Gaussian Mixture Model(GMM)

Gaussian mixture models are widely used in data mining, pattern recognition, machine learning, and statistical analysis. It is often applied to find clusters in a set of data points. Additionally, from the view of statistic, a Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities [26]. And latent variables of mixture models are determined by finding maximum likelihood solutions, typically using the expectation-maximization algorithm, or EM algorithm. In this section, we introduce the Gaussian Mixture Model (GMM) and explain how to represent the peak point of a burst with GMM.

3.4.1. *Gaussian Mixture Models(GMM) using EM algorithm*

Since all nonzero traffic are 1-D time series data, we explain GMM and EM algorithm with 1-D Gaussian distribution. Suppose x is the input 1-D data, N is the number of these data, the probability density of a Gaussian is expressed as

$$g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\text{exp}\left(-\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{3.4}$$

where, μ and σ are mean and variance of the distribution, respectively. Suppose there are K components in a Gaussian mixture model, the GMM is defined by a weighted sum of K Gaussian distributions:

$$p(x;\theta) = \sum_{i=1}^{K} w_i g(x;\mu_i,\sigma_i) \tag{3.5}$$

where w_i's are mixture weights, satisfying the constraint

$$w_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{K} w_i = 1 \tag{3.6}$$

and each mixture component is identified by an index i, $i = 1,2,...,K$.For mixture of Gaussian function, the likelihood function can be defined as

$$\Lambda(x;\theta) = \prod_{n=1}^{N}\sum_{i=1}^{K} w_i g(x_n;\mu_i,\sigma_i) \tag{3.7}$$

The logarithm of the likelihood function $\Lambda(X;\theta)$ in equation. 3.7 can be written as

$$\lambda(x;\theta) = \sum_{n=1}^{N} \log \sum_{i=1}^{K} w_i g(x_n;\mu_i,\sigma_i) \tag{3.8}$$

Thus, a Gaussian mixture model is parameterized by three values: the mean, variance and mixture weights from all component densities, donated as $\theta = \{w_i, \mu_i, \sigma_i\}_{i=1}^{K}$. And the parametric density estimation problem can be defined as to find the parameter vector θ with the maximum likelihood estimate, expressed as

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\ \Lambda(x;\theta) = \underset{\theta}{\text{argmax}}\ \lambda(x;\theta) \tag{3.9}$$

EM algorithm is a general and powerful method for finding the maximum likelihood estimate of the parameters in statistical models, in which unobserved variables exist. EM starts with initial values $w_i^{(0)}$, $\mu_i^{(0)}$, $\sigma_i^{(0)}$ and iteratively with two steps(E step and M step) until convergence to a local maximum of the likelihood function [27].

E Step is to construct the bound of the logarithm expectation. Suppose M is the iterate number. Given the old parameters $w_i^{(m)}$, $\mu_i^{(m)}$, $\sigma_i^{(m)}$, the prior probability, also called membership probability, $w^{(i)}(m|i)$ is computed by

$$w^{(m)}(i|n) = \frac{w_i^{(m)} g\left(x_n;\mu_i^{(m)},\sigma_i^{(m)}\right)}{\sum_{i=1}^{K} w_i^{(m)} g\left(x_n;\mu_i^{(m)},\sigma_i^{(m)}\right)} \tag{3.10}$$

Then, the bound $b_m(\theta)$ can be formulated as

$$\begin{aligned} b_m(\theta) &= \sum_{n=1}^{N}\sum_{i=k}^{K} w^{(m)}(i|n)\log p(i,n) - \sum_{n=1}^{N}\sum_{i=k}^{K} w^{(m)}(i|n)\log w^{(m)}(i|n) \\ &= \underbrace{\sum_{n=1}^{N}\sum_{i=k}^{K} w^{(m)}(i|n)\log w_i g(x_n;\mu_i,\sigma_i)}_{\text{part 1}} - \underbrace{\sum_{n=1}^{N}\sum_{i=k}^{K} w^{(m)}(i|n)\log w^{(m)}(i|n)}_{\text{part 2}} \end{aligned} \tag{3.11}$$

The second part in equation 3.11 is known, the minimizing the bound $b_m(\theta)$ is equal to mimimize the first part in this equation.

M step is to generate new estimate of θ: $w_i^{(m+1)}, \mu_i^{(m+1)}, \sigma_i^{(m+1)}$ based on the maximization of $b_m(\theta)$. GMM parameter set θ iterates as following:

$$\mu_i^{(m+1)} = \frac{\sum_{n=1}^{N} w^{(m)}(i|n) x_n}{\sum_{n=1}^{N} w^{(m)}(i|n)} \tag{3.12}$$

$$\sigma_i^{(m+1)} = \sqrt{\frac{\sum_{n=1}^{N} w^{(m)}(i|n)(x_n - \mu_i^{n+1})^2}{\sum_{n=1}^{N} w^{(m)}(i|n)}} \tag{3.13}$$

$$w_i^{(m+1)} = \frac{1}{N} \sum_{n=1}^{N} w^{(m)}(i|n) \tag{3.14}$$

Overall, when the initial values of $w_i^{(0)}, \sigma_i^{(0)}, \mu_i^{(0)}$ are known, a local maximum of the likelihood function 3.7 can be computed by equation 3.10 to 3.14 until convergence.

From a general perspective, in this thesis, we set the initial estimates of $w_i^{(0)}, \sigma_i^{(0)}, \mu_i^{(0)}$ like this:

- $\mu_i^{(0)}$: a $1 \times K$ vector with the mean of all input data the uniform distribution with K components.
- $\sigma_i^{(0)}$: a $1 \times K$ vector with equal values of the standard deviation of the input data
- $w_i^{(0)}$: a $1 \times K$ vector with equal values of $\frac{1}{K}$

Following the above study of Gaussian mixture models, how to determine a suitable component number in a Gaussian mixture model to better fit the input data has been a crucial problem.

A simple solution to this problem is to use *Bayesian Information Criteria* (BIC) to penalize the complexity of the GMM. BIC is a criterion for model selection among a finite set of models, defining as [28]

$$\text{BIC} = -2 \times \log \hat{L} + k \times \log(n) \tag{3.15}$$

Where, $\hat{L}$ is the maximized values of likelihood function of a Gaussian mixture model. Combining with the loglikelihood function in equation 3.8, BIC can be rewritten as

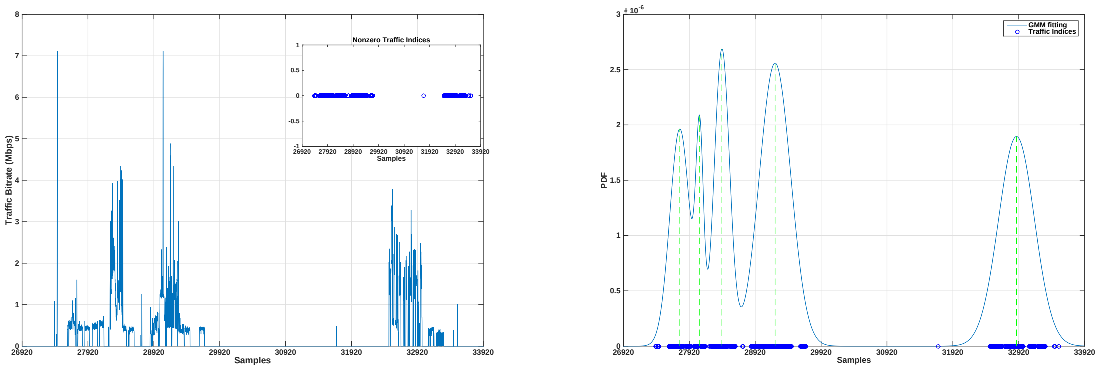
$$\text{BIC} = -2 \sum_{n=1}^{N} \log \sum_{i=1}^{K} w_i g(x_n; \mu_i, \sigma_i) + k \times \log(n) \tag{3.16}$$

where, k is the parameters used in GMMs. The Gaussian mixture model with the lowest BIC is preferred.

3.4.2. *Implementation of Gaussian Mixture Models*

For each household, we tend to build a Gaussian mixture model to characterize nonzero traffic in this home network. However, due to the larger number of traffic measurements in each household, it's hard to figure out how many mixture numbers will be needed in each GMM, even the range of mixture number.

We first apply Gaussian mixture model to each cycle in different home networks. Considering the computation complexity, we set the range of mixture number from 1 to 15 to build different Gaussian mixture models and choose the best Gaussian mixture model with the lowest BIC. The results show that the mixture number of Gaussian components in each cycle is smaller than 15.Hence, the mixture number in each home network is smaller than 105.



(a)

(b)

Figure 3.4: *(a) The nonzero traffic from the 2690th sample to 33920th sample in home network 1. The zoom shows the nonzero traffic indices. (b) The GMM result during this period. The blue circles on X-axis represent the nonzero traffic indices. The blue line shows the probability density of GMM fitting. The green dotted lines represent the mean value of each Gaussian component.*

Based on this result, a wide range of mixture numbers to model all indices in each home network is set from 1 to 105. The GMM results with the lowest BIC under this constraint are reached and will be used as the input of prediction in chapter 4.

3.4.3. *Burst Definition*

Fig.3.4b captures part of Gaussian mixture model results in home network 1. Obviously, five Gaussian components are used to model this part of data. For each Gaussian component, the mean points owns the highest probability density, which indicate the traffic is most likely to happen at this point.

Based on the idea and the symmetry of Gaussian distribution, we can define a burst according to nonzero traffic indices and Gaussian mixture results: in a Gaussian component, the index corresponds to the mean value is notated as a peak point; the 1-sigma area around this index is called a burst. In this area, traffic are more likely to happen. Hence, one Gaussian component corresponds to one burst, and one burst corresponds to one peak. A burst can be described in Fig.3.5.

Differing from the burst definition in [25], our burst is considered based on the probability density of each nonzero traffic indices and gives an insight of when traffic is most like to happen and the duration that traffic is more like to happen along the time. It doesn't tell how much traffic enters into the network in this burst. In telecommunication, 95th percentile industry standard is a widely used mathematical calculation to evaluate the regular and sustained use of a network connection [29]. It provides a method to closely indicate how much bandwidth capacity needed in the network than using average traffic rate or maximum bandwidth [30]. In this thesis, we choose 95th percentile bandwidth to represent how much bandwidth a home

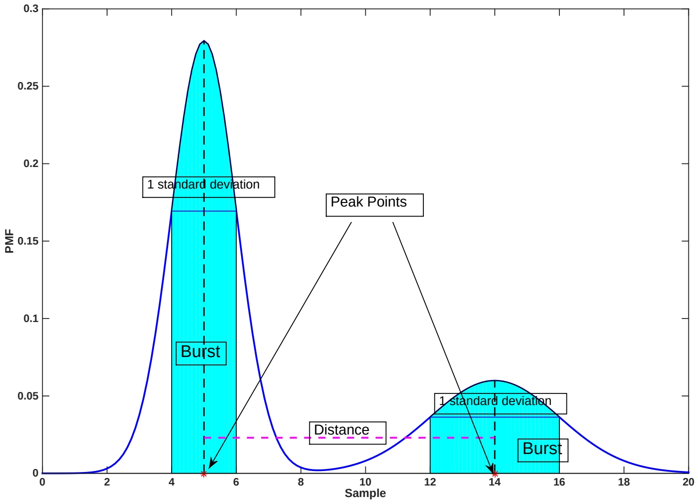


Figure 3.5: *A Gaussian mixture model with* $\mu_1 = 5, \sigma_1 = 1, w_1 = 0.7$ *and* $\mu_2 = 14, \sigma_2 = 2, w_2 = 0.3$. *The black dotted lines represent the symmetry axes, which correspond to red stars on x-axis. The cyan area is a defined burst with the width of one standard deviation. The magenta dotted line shows the difference between two peak points.*

network used in the burst. It is calculated based on the raw traffic measurements in this burst.

3.5. Conclusion

In this chapter, we characterise the incoming traffic bit rates using power law and generalize Pareto distribution and find that small traffic data has long-tail features, which indicate a significant characteristic of home network traffic that long period with low bandwidth usage and short period with high bandwidth usage. Considering this feature, we extract traffic bit rates higher than background traffic and record the corresponding sample indices. On the strength of Gaussian mixture model, ”burst” is defined based on these indices.

4

A new traffic prediction model

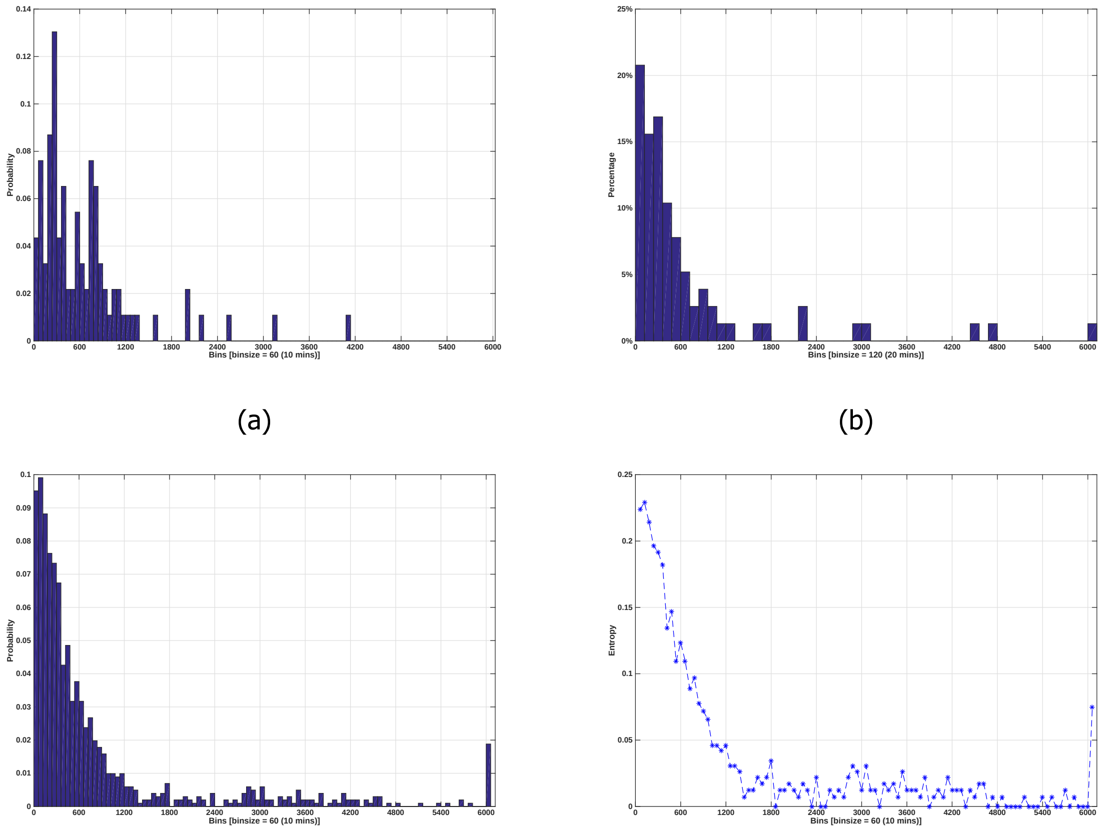
In the last chapter, we characterize home network traffic based on the idea of burst. Hence, in this chapter, we propose and develop a method to predict when traffic is mostly to happen, how long the traffic will last and the bandwidth usage of the next burst. At the end of this chapter, cross validation is applied to analyze the performance of this prediction.

4.1. Prediction Model

4.1.1. *Prediction of Peak Point in the Future Burst*

In order to predict the peak of the next burst, we first define the possible states of Markov chain and calculate state transition probabilities based on the dataset in order to obtain the Markov transition.

Based on results of Gaussian mixture models(GMM), we pick up all mean values of each Gaussian component in 15 household. As defined in section 3.4.3, these values represent all peaks in different bursts. We calculate all distances between each two neighbouring peaks and compute the probabilitiese of these distances in each household. Fig.4.1a and Fig.4.1b respectively depict the PMF results in home network 2 and home network 5. The bin-size of distances in the figure is set to 60 samples, which equals to 10 mins, and each bin corresponds to one distance range. We find that in these two households, small distances take high probability and this phenomenon also appears in the rest 13 households. In light of this feature, we compute the probability of distances according to all measurements in 15 households in Fig.4.1c and discovery the distribution of distances follows power law, described as: $y = 8.826 \cdot x^{-0.962}$, where y is the probability density function and x is the distance between each two consecutive peak points.



(a)

(b)

(c)

(d)

Figure 4.1: *(a),(b),(c) the probability mass function of distances in home network 2, 5 and among all households. (d)Entropy of distances. x axis is the distance between two consecutive peak points.y axis is the probability (entropy) that a burst has a certain peak.*

Information entropy is a measurement to quantize the amount information [31], which has been applied to data classification, such as Internet backbone traffic behaviors cluster by K.Xu et al.[32]. We take the intuitive meaning of entropy to measure the information amount of distance ranges between two peak sample stamps in this project.

Donate $p(q)$ is the probability of qth distance range in Fig.4.1c, we calculate the entropy of each distance range through Eq.4.1, shown in Fig.4.1d.

$$h(q) = -p(q)\log_2 p(q) \tag{4.1}$$

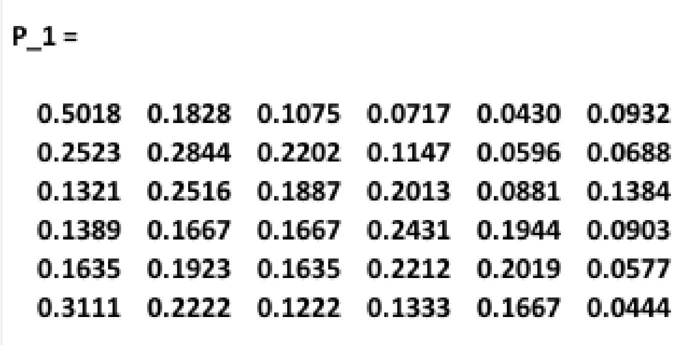
The total distance entropy is the sum of the entropy in each distance range, equalling to 4.855bits. We divide the distance states of Markov chain with the principle that the entropy of a distance range to belong to each state is the same, which means the amount information each state provides is the same. Assume all distances can be sufficiently represented into six states, thus the entropy in each state is around 0.8 bits. Then the distance state of Markov chain are explained in Table 4.1. The transition between each two states is considered to follow a discrete-time Markovian process with the property that the next state only depends

Table 4.1: *Possible states of Markov chain. Distances refer to the difference between two neighbouring peak sample stamp.Intervals refers to the period of the corresponding sample ranges,calculated by distance multiplying 10s*

States	Distances	Time Intervals [1]
x_1	(0, 180]	(0, 30min]
x_2	(180, 360]	(30min, 1h]
x_3	(360, 600]	(1h, 100min]
x_4	(600, 1020]	(100min,170min]
x_5	(1020,2880]	(170min, 8h]
x_6	>2880	>8h

on the current state.

Assume the distance between two peak points is in state x_i,$i \in [1, 6]$, then the transition probability to another state x_j can be written as $P_1(x_j|x_i)$, which corresponds to the i-th row and j-th column of the transition matrix. After calculating the transition between each two states among 15 households, the transition matrix of this Markov chain P_1 is obtained :



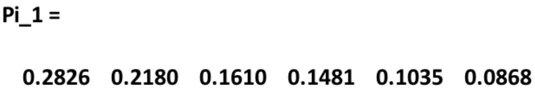
```
P_1 =

   0.5018   0.1828   0.1075   0.0717   0.0430   0.0932
   0.2523   0.2844   0.2202   0.1147   0.0596   0.0688
   0.1321   0.2516   0.1887   0.2013   0.0881   0.1384
   0.1389   0.1667   0.1667   0.2431   0.1944   0.0903
   0.1635   0.1923   0.1635   0.2212   0.2019   0.0577
   0.3111   0.2222   0.1222   0.1333   0.1667   0.0444
```

Hence, given an initial state, donated as C_k(a 1×6 vector), then the state probability vector after $m(m \in \mathbb{Z}^+)$ steps can be calculated using the formula 4.2:

$$C_{k+m} = C_k \cdot P_1^m \tag{4.2}$$

Since the transition matrix P_1 is irreducible and has finite states, when m goes to infinite, the Markov chain will go into a steady state and the steady state π_1 is reached by Eq. 4.3:

$$\pi_1 = \underset{m\to\infty}{\text{limit}} \; C_{k+m} = \underset{m\to\infty}{\text{limit}} \; C_k \cdot P_1^m \tag{4.3}$$



```
Pi_1 =

   0.2826   0.2180   0.1610   0.1481   0.1035   0.0868
```

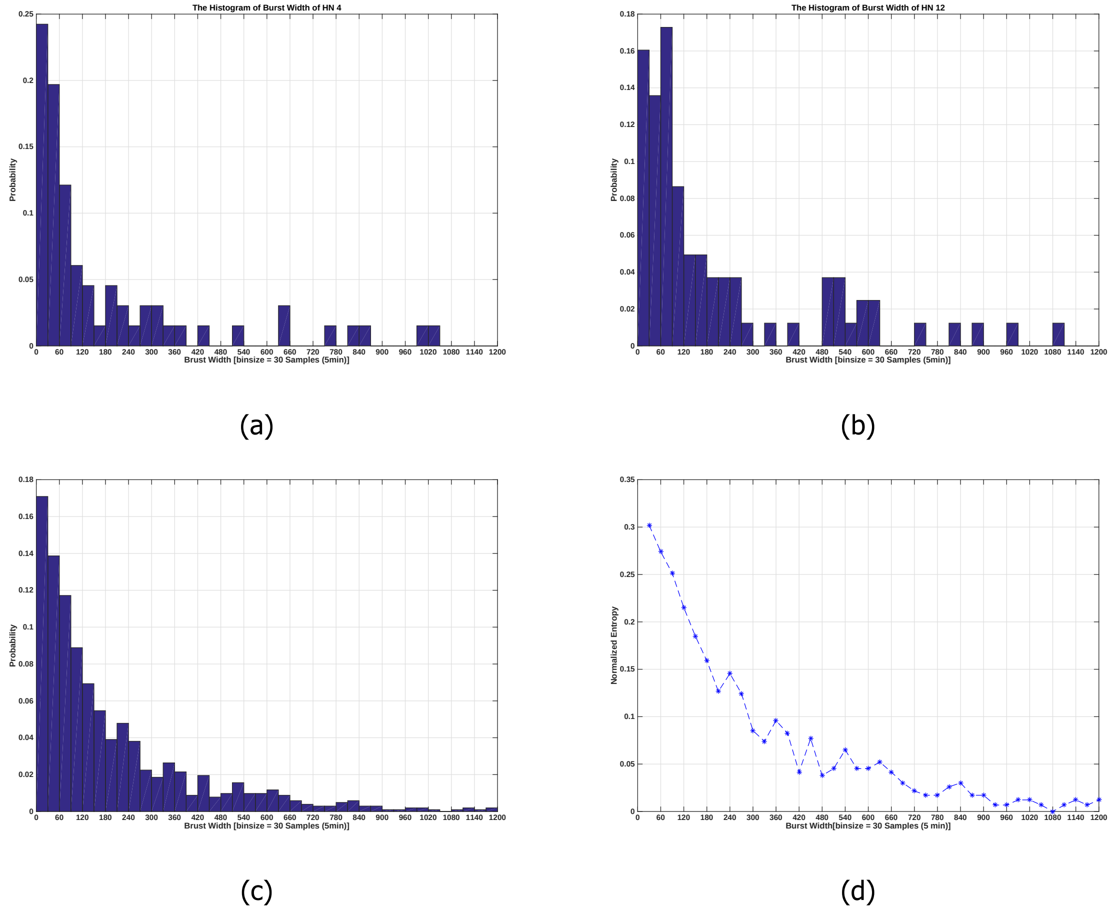


Figure 4.2: *(a),(b),(c) the distribution of burst width in home network 8, 14 and among all households. (d) Entropy of burst width, x axis is the width of a burst and y axis is the probability (entropy) that a burst has a given width.*

4.1.2. *Prediction of Future Burst Width*

According to the definition of burst width in section 3.4.3, we calculated the widths of each burst among all households. Burst width stands for the duration of the traffic. Markov chain prediction method is also used to predict the probability of possible traffic range in the future. Firstly, we define the states of this Markov chain based on the distribution of existing burst widths.The maximum burst widths among 15 households includes 1199 traffic measurements. Then we calculate the probability mass function(PMF) of all burst widths in each household from 0 to 1200 with the bin size of 30 traffic measurements, equaling 5min. And a similar characteristic appears in each household that is small burst widths account for a larger proportion of the bursts, such as in home network 4 (Fig.4.2a) and 12 (Fig.4.2b).

Upon this characteristic, we depict burst widths among 15 households expressed as a probability mass function (PMF) in Fig.4.2c and the distribution of burst width also follows power law.

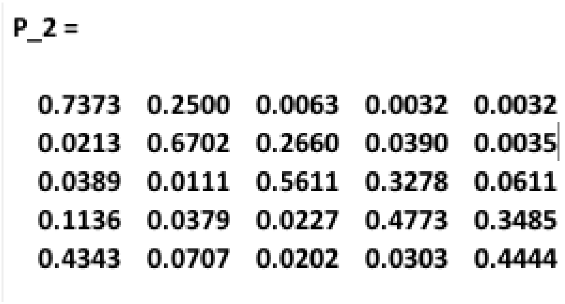
Similar to the state definition methods in section 4.1.1, entropy of burst width is calculated

Table 4.2: *Possible states of burst width based on entropy. Burst Width Ranges how how many traffic measurements in one burst. Time Intervals indicate how long one burst lasts.*

States	Burst Width Ranges	Duration
y_1	(0, 60]	≤ 10min
y_2	(60, 150]	(10min, 1500s]
y_3	(150, 270]	(1500s, 2700s]
y_4	(270, 510]	(1000s,5100s]
y_5	(510, 1200]	(5100s, 200min]

and shown in Fig.4.2d and the state selections of burst widths also follow the principle that the entropy of each state is the same, thus the states of burst width are explained in Table 4.2.

The transition from one burst width range to another range is supposed to follow a discrete-time Markovian process, where the next state only depends on the current one. Then we computed the transition matrix in P_2. The data of the i-th row and j-th column in transition matrix P_2 represents the probability from current state y_i jumping to another state y_j in the next time step, $i, j \in [1, 5]$.



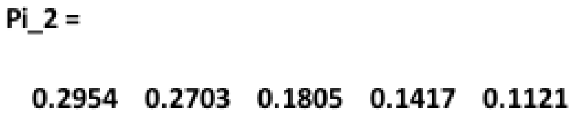
```
P_2 =

    0.7373    0.2500    0.0063    0.0032    0.0032
    0.0213    0.6702    0.2660    0.0390    0.0035
    0.0389    0.0111    0.5611    0.3278    0.0611
    0.1136    0.0379    0.0227    0.4773    0.3485
    0.4343    0.0707    0.0202    0.0303    0.4444
```

Given the probability that the initial burst width is being in each state R_k, the future state of burst width after $m (m \in \mathbb{Z}^+)$ steps is calculated in Eq. 4.4:

$$R_{k+m} = R_k \cdot P_2^m \tag{4.4}$$

Since the transition matrix P_2 is irreducible and has finite states, when m goes to infinite, this Markov chain will go into a steady state after m steps and the steady state π_2 is reached by Eq. 4.5:

$$\pi_2 = \operatorname*{limit}_{m \to \infty} R_{k+m} = \operatorname*{limit}_{m \to \infty} R_k \cdot P_2^m \tag{4.5}$$



```
Pi_2 =

    0.2954    0.2703    0.1805    0.1417    0.1121
```

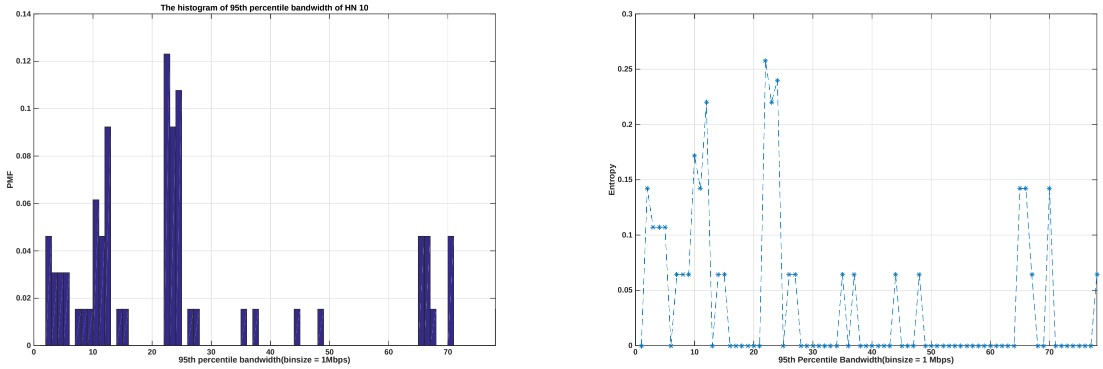
4.1.3. *Prediction of Bandwidth Usage in the Future Burst*

After predicting the peak point and burst width of a future burst, we forecast how much traffic will enter the network for the next burst. We calculate the 95th percentile bandwidth based on traffic bitrates in each burst to represent the bandwidth usage during this period. In view of the heterogeneous of bandwidth usage in different bursts, we discover, except home network 10, the bandwidth usage in the rest 14 households has one common characteristic that is more than 90% of the 95th percentile bandwidths are smaller than 10Mbps. For home network 10, around 75% of bandwidth consuption fo each burst is spread across the range from 10Mbps to 80Mbps.

Thus, to give a close view of the bandwidth usage of the next burst for different home networks, based on the 95th percentile bandwidth, we tend to classify 15 households into two groups: one group(Group 1) is households 10, which occupies very higher bandwidth usage; the second group (Group 2) includes the rest 14 households that most of bandwidth usage are smaller than 10 Mbps. Then, the prediction of bandwidth usage in the next burst are conducted in these two groups.

Prediction of bandwidth usage in Group 1:

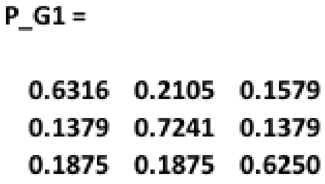
The example in the group is home network 10, which has very high bandwidth usage explained in Fig.4.3a. The bin size of 95th percentile bandwidth is 1 Mbps. Similar to the prediction method in section 4.1.2, entropy is calculated as the basis for the state division, shown in Fig.4.3b According to results of the Gaussian mixture models, there are 65 burst in this home network, which means the number of 95th percentile bandwidths is also 65. Considering the data set under this situation is too small, we assume only three states could sufficiently represent all bandwidth usage in this household.



(a) (b)

Figure 4.3: *(a) 95th percentile bandwidth histogram in home network 10, (b) CDF of 95th percentile bandwidth in the same network. x axis is the 95th percentile bandwidth, y axis is the probability (entropy) that a burst has a known bandwidth usage.*

With the idea that each state of bandwidth usage has the same amount of information, the three states based on the traffic measurements in home network 10 are defined as: e_1 = [0,11Mbps], e_2 = [11Mbps, 25Mbps] e_3 = [25Mbps, 78Mbps]. Suppose the transition between two states follow the a discrete-time Markovian process. After calculating all probabilities of all transitions, the Markov chain can be presented in the matrix P_{G1}, in which the data of the i-th row and j-th column represent the probability from current state e_i jumping to state e_j in the next time step, $i, j \in [1, 3]$.



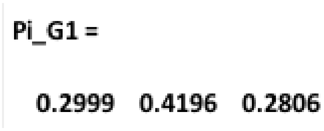
```
P_G1 =

    0.6316    0.2105    0.1579
    0.1379    0.7241    0.1379
    0.1875    0.1875    0.6250
```

Hence, given the probability that the intial bandwidth usage is being in each states H_k, the future state of bandwidth usage after $m (m \in \mathbb{Z}^+)$ steps is calculated in Eq.4.6:

$$H_{k+m} = H_k \cdot P_{G1}^m \tag{4.6}$$

Since the transition matrix P_{G1} is irreducible and has finite states, when m goes to infinity, this Markov chain will go into a steady state after m steps and the steady state π_{G1} is reached by Eq.4.7:

$$\pi_{G1} = \operatorname*{limit}_{m \to \infty} H_{k+m} = \operatorname*{limit}_{m \to \infty} H_k \cdot P_{G1}^m \tag{4.7}$$

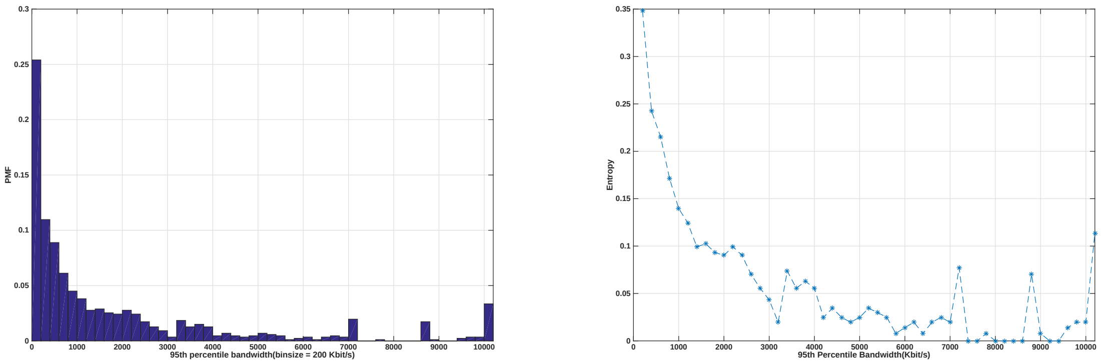


```
Pi_G1 =

    0.2999    0.4196    0.2806
```

Prediction of bandwidth usage in Group 2:

This group includes 14 home networks and we use Markov chain methodology to develop our model of bandwidth usage of the next burst in a home network. We calculate the probability of 95th percentile bandwidth among these 14 households in Fig.4.4a. The bin size of 95th percentile bandwidth is set to 200kbit/s. Based on the probability distribution, we computed the entropy of the corresponding 95th percentile bandwidth, depicted in Fig.4.4b.

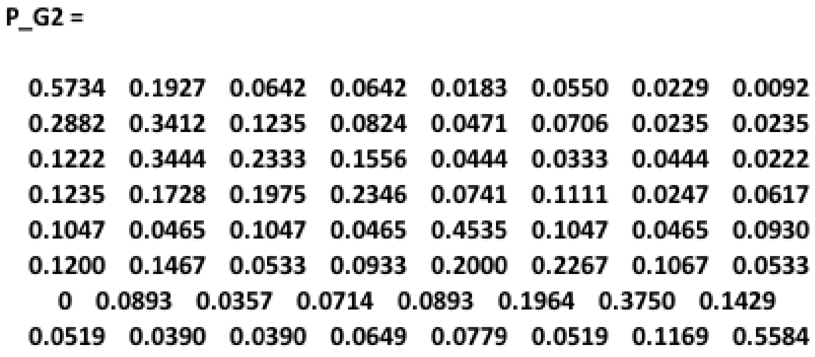
For the state definition, we assume the 95th percentile bandwidth can be well represented into eight sates, which follows the law that the information each state contains is the same. Then the states are defined by the corresponding entropy, namely z_1 = [0, 200Kbit/s], z_2 = [200Kbit/s, 600Kbit/s], z_3 = [600Kbit/s, 1Mbps], z_4 = [1Mbps, 1.6Mbps], z_5 = [1.6Mbps,



(a) (b)

Figure 4.4: *(a) Histogram of 95th percentile bandwidth. (b) Entropy of 95th percentile bandwidth. x axis is 95th percentile bandwidth.x axis is the range of bandwidth usage (95th percentile bandwidth), y axis is the probability (entropy) that a burst has a known bandwidth usage.*

2.4Mbps], z_6 = [2.4Mbps, 3.8Mbps], z_7 = [3.8Mbps, 6.6Mbps], z_8 > 6.6Mbps. Suppose the transition between two states follow a discrete-time Markovian process, in which the next state depends only on the current state and is independent from previous states. After calculating the transition between different states, the transition matrix is expressed in P_{G2}, in which the data of i-th row and j-th column stands for the probability form state z_i jumping to state z_j in the next time step, $i, j \in [1, 8]$.



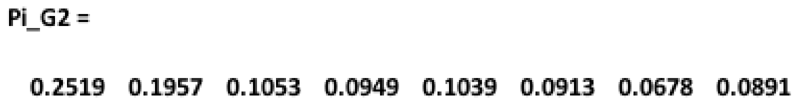
```
P_G2 =

    0.5734    0.1927    0.0642    0.0642    0.0183    0.0550    0.0229    0.0092
    0.2882    0.3412    0.1235    0.0824    0.0471    0.0706    0.0235    0.0235
    0.1222    0.3444    0.2333    0.1556    0.0444    0.0333    0.0444    0.0222
    0.1235    0.1728    0.1975    0.2346    0.0741    0.1111    0.0247    0.0617
    0.1047    0.0465    0.1047    0.0465    0.4535    0.1047    0.0465    0.0930
    0.1200    0.1467    0.0533    0.0933    0.2000    0.2267    0.1067    0.0533
         0    0.0893    0.0357    0.0714    0.0893    0.1964    0.3750    0.1429
    0.0519    0.0390    0.0390    0.0649    0.0779    0.0519    0.1169    0.5584
```

Hence, given the initial states of bandwidth usage G_k, the future state of 95the percentile bandwidth after $m (m \in \mathbb{Z}^+)$ steps is calculated in Eq.4.8:

$$G_{k+m} = G_k \cdot P_{G2}^m \tag{4.8}$$

Since the transition matrix P_{G2} is irreducible and has finite states, when m goes to infinities, this Markov chain will go into a steady state after m steps and the steady state π_{G2} is reached by Eq. 4.9:

$$\pi_{G2} = \text{limit}_{m \to \infty} G_{k+m} = \text{limit}_{m \to \infty} G_k \cdot P_{G2}^m \tag{4.9}$$



```
Pi_G2 =

  0.2519   0.1957   0.1053   0.0949   0.1039   0.0913   0.0678   0.0891
```

4.2. Performance Analysis of a New Traffic Model

Owing to the memoryless of Markov chain, one advantage of our prediction model is the prediction only depends one the current state of each Markov chain, which only require a small set of data samples. In this section, we analyse the performance of our model.

4.2.1. *Model Validation*

Cross validation is a model validation method widely used to measure the performance of a prediction model [33]. Basically, cross validation classifies all data samples into two sets: (1) training set, used to train or develop data set; (2) validate sets used to test the model. Typically, there are two kinds of cross validation: exhaustive cross-validation and non-exhaustive cross-validation. Considering traffic measurements in this project is quite small, we use Leave-one-out cross-validation((LOOCV)) to validate our model.

As discussed in section 4.1, three aspects of a burst are respectively predicted: peak point, burst width and bandwidth usage. We apply LOOCV to each of our prediction model separately and the specific validation steps are explained as follow.

1. All burst widths (distances, bandwidth usages) in one households are considered as an independent observation. Thus, 15 households represent 15 independent observations.
2. In one home network, all burst widths (distances, bandwidth usages) are divided into two data sets: a testing set which only contains one burst width(distance, bandwidth usage), and the rest burst widths (distances, bandwidth usages) belong to the training set.
3. All burst width (distances, bandwidth usages) in these two sets of data follows the same state divisions in Table 4.2 (Table4.1,z_1 to z_8).
4. A Markov transition matrix is built on the training set using the same methods in section 4.1.2 (section 4.1.1, section 4.1.3).
5. Then the state probability vector can be obtained through Eq.4.4 (Eq.4.2, Eq.4.8), where the initial states were set as the corresponding state of the burst width before the testing set.
6. The performance of the model is measured by the True Positive Rate (*TPR*) parameter, which gives the number of correct predictions for what will be the next state. Two prediction methods are employed as following:

   1) *M1:* The burst prediction model has five states, we suppose the next burst width at each state with the same probability to appear (all transition probabilities equal to 0.2). (For

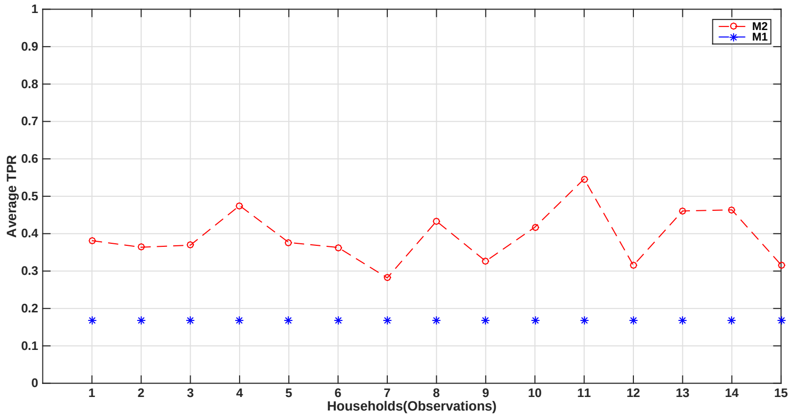


Figure 4.5: For the prediction of peak point of the next burst, the average TPR of *M1* and *M2* used in LOOCV among 15 households

prediction model, six sates are defined, the next distance at each state is supposed with the same probability, equalling to 1/6; The bandwidth prediction model in G2 has eight states, the next bandwidth usage at each state with the same probability to appear, equalling to 0.125)

2) *M2:* According to the steady state vector, we suppose the next burst width (distance, bandwidth usage) is at the state with the highest probability.

7. Counting the TPR for all validation set, we donated the average as the final TPR.
8. Repeated step 2 to 7 to other 14 households.

4.2.2. *Performance Analysis*

The result of average TPR of three prediction models with two prediction methods $M1$ and $M2$ in 15 households are depicted in Fig.4.5, Fig. 4.6, Fig.4.7. Due to the probabilistic character of our prediction model, normally the average TPR will not achieve 100%. And it's obviously from these three figure, the performance of prediction method in M2 is better than the random state distribution in M1. Since difference households behave in different ways, the average *TPR* of three prediction models among 15 households differ from each other.

For the second prediction result(M2) in LOOCV, the average *TPR* of 15 households in peak point prediction model is in the range of [0.3 0.6] and only one household obtains the *TPR* larger than 0.5. While the performance of burst width prediction model is slightly better that the average *TPR* of 10 households are larger than 0.5 and the whole average TPR is in the range from 0.35 to 0.75. And for the prediction of bandwidth usage, four households has very small average TPR and the distinction of average TPR between different households are larger than the other two prediction models. An explanation to this is the heterogeneous of bandwidth usage of different burst in 15 households.

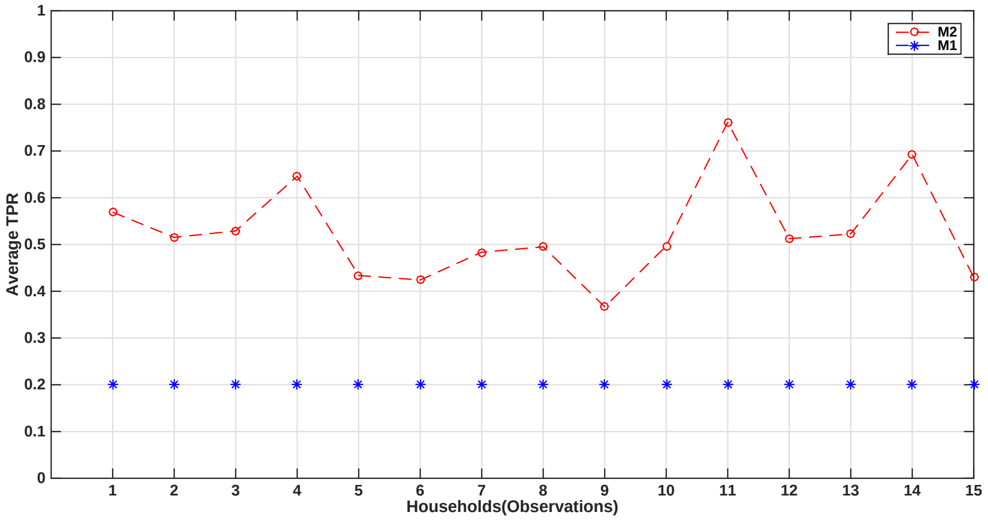


Figure 4.6: For bursts width prediction, the average TPR of *M1* and *M2* used in LOOCV among 15 households

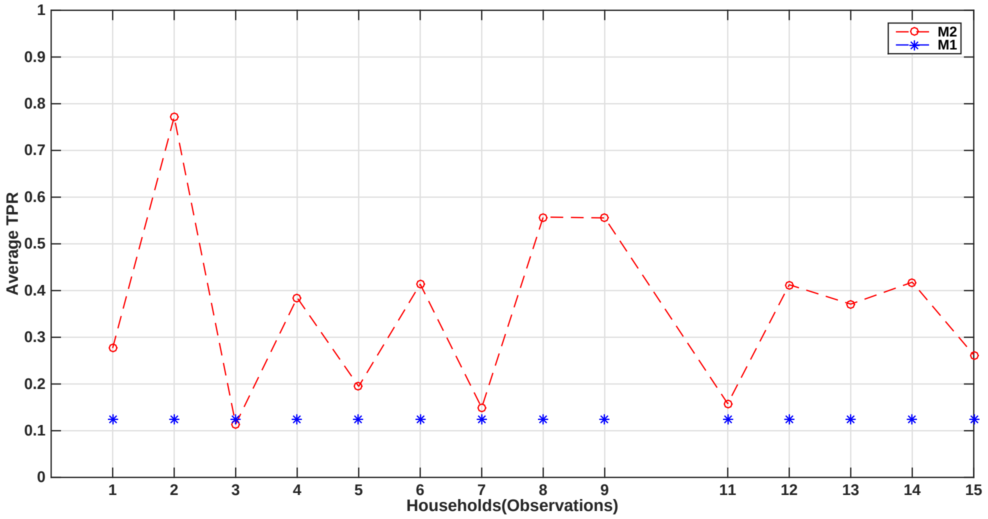


Figure 4.7: For the prediction of bandwidth usage in Group 2, the average TPR of *M1* and *M2* used in LOOCV among 15 households

On the whole, the average TPR results in three prediction models don't obtain very high values as expected. This is because the data set of bursts extracted from our traffic measurement are relative small. In order to increase the accuracy, more traffic measurments are needed.

4.3. Conclusion

In this chapter, we proposed a new traffic model based on the idea of *burst* to predict when traffic is mostly likely to happen in the future, the width of the next burst and the bandwidth usage in the next burst. At the end, leave-one-out cross validation are used to analyze the performance of our prediction model.

5

Controlled Home Environment Research

In the previous chapter, we have characterised the behaviours of heterogeneous home networks and built a new traffic prediction model to systematically forecast the future traffic behavior. The traffic bit rates in the databased are relatively small, compared to the current network bandwidth and the bandwidth consumption of online applications and services. Hence, in this chapter, we establish an experimental platform to collect a new set of traffic data set and provide an insight of the effects of home activities to traffic bandwidth under a controlled home environment.

5.1. Experiment Platform

The experimental platform is created in the TNO lab to simulate a home network under a controlled home environment, where all activities can be monitored and controlled. And network users behave as they would be using the network at their own home. The physical map of the controlled home network shown in Fig.5.1 follows the structure diagram in Fig. 2.1. Different home devices are connected to the Residential Gateway(RG) through Ethernet cable, Wi-Fi and powerline: two iPads, mobile devices and one computer connect to the HG through Wi-Fi, while the printer and a network attached storage(NAS) connect to the RG using Ethernet cable.

To monitor home network traffic, one prerequisite is to run Simple Network Management Protocol(SNMP) on RG. Due to the firewall of WRT1900AC doesn't support SNMP, thus, OpenWrt [34] is used to reboot the firmware on the router. Then one computer (as PC1 in Fig.2.1) connecting to the router could run PRTG software via SNMP to collect traffic data through

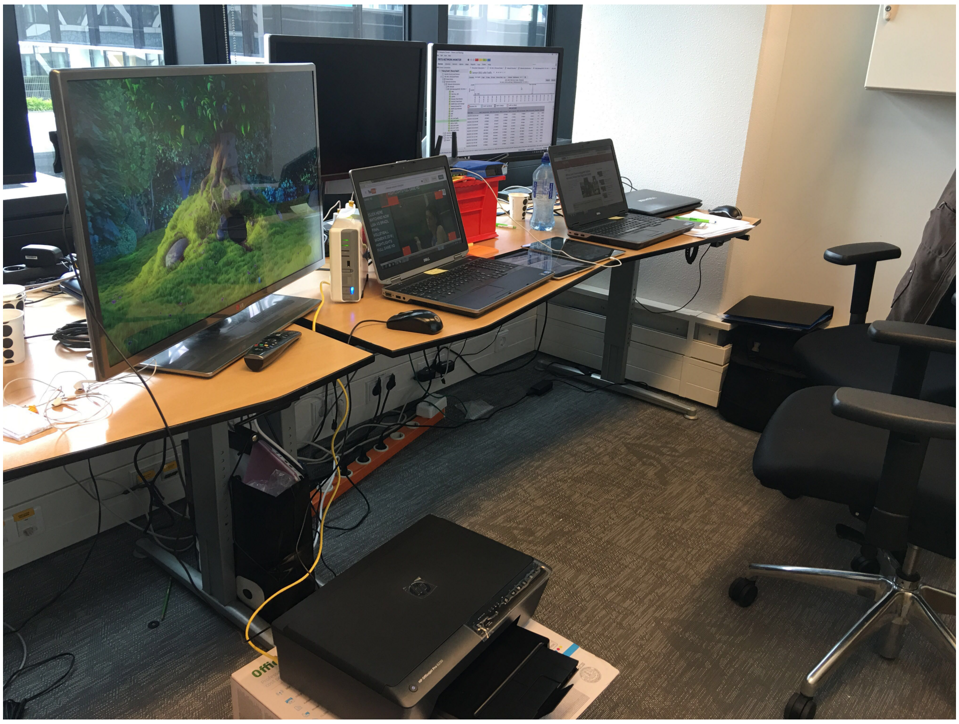


Figure 5.1: *The controlled home network configuration in the lab*

each port of the router with the time interval of 10 seconds. The controlled home network is measured for seven working days from Sep.1st to Sept.9th. During this period, we register all activities during office hours(9:00 to 17:00) running on different home devices.

5.2. Analysis of Home Activities

The incoming traffic through the router, used to build the new traffic model in section 3.2, is calculated using the Eq.3.1, which includes two parts: one is Internet traffic and the other part is traffic passing through different devices inside the home though the Router. The second part traffic is called in-home traffic, referring to traffic going through sources and servers, both of which are inside the home network.

From the prospective of privacy, Operators only know how much bandwidth each home network requires, but don't have any idea on what kind of applications running inside a home. Accounting to the registered activities in this experiment, we briefly provide an insight of the bandwidth usage of some often used applications inside a home network based on some main scenarios from our experiment.

5.2.1. *Scenario 1: Different Connectivities of a Printer*

In this scenario, we check with the bandwidth usage of a inkjet printer with wired and wireless connections to print a black and colourful page respectively, plotted in Fig.5.2. We start the

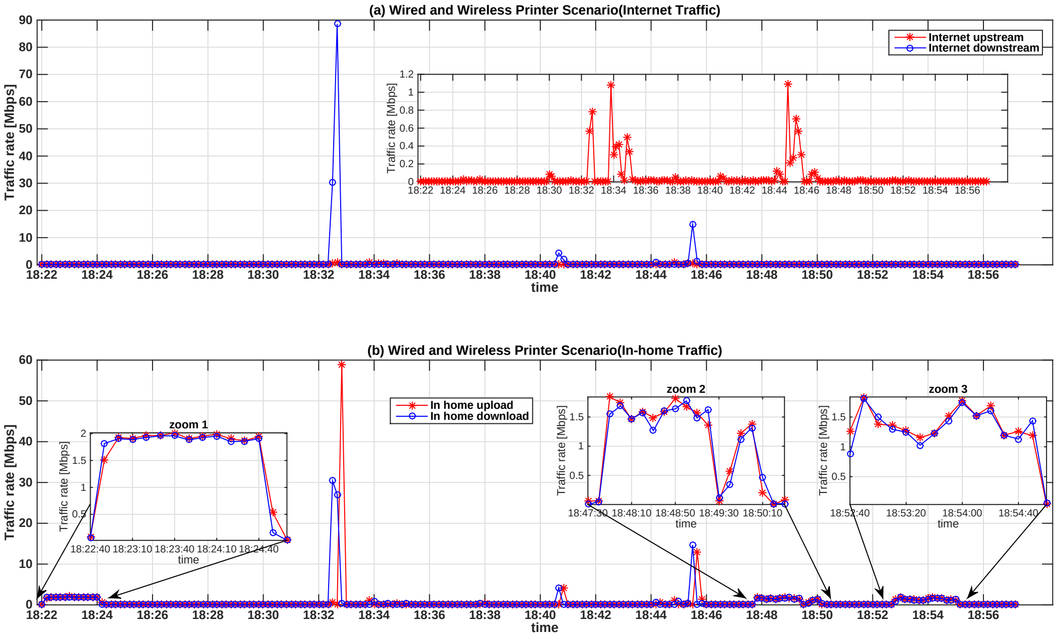


Figure 5.2: *Bandwidth usage of a printer in different scenarios*

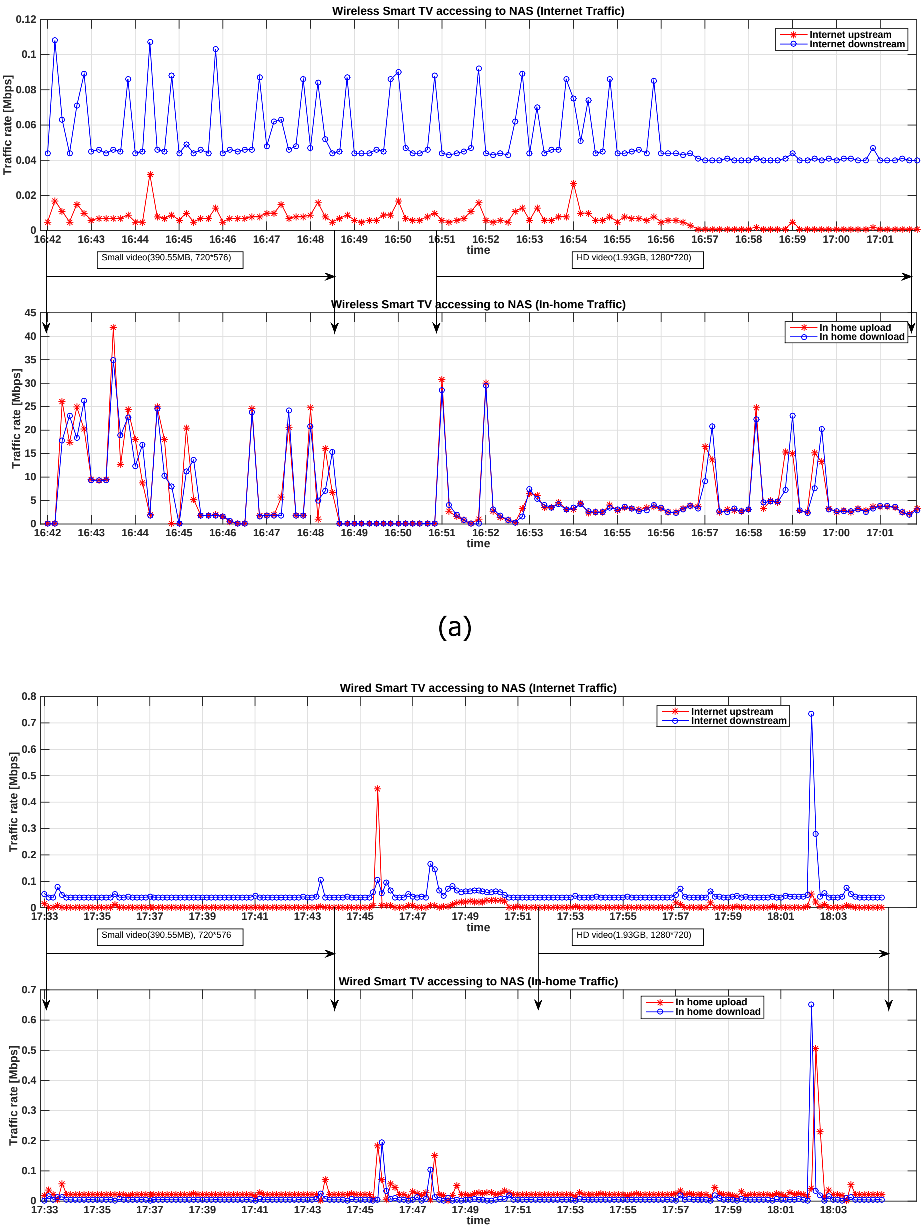
analysis from the in-home traffic in Fig.5.2 (b): zoom 1 presents that the wired printer print a colour page of 13M; zoom 2 shows the printer with wireless connection prints the same colour page; zoom 3 shows the wireless printer print a black page. The wired printer also print the same black and white page at time 18:26 and this activity only last 20s. Due the bandwidth usage is too small (100kbps) and the duration is short, it doesn't show in the graph. For the very peak traffic point in both two graphs in Fig.5.2, the high bandwidth consumption is caused by the printer unplugging from the router and a PC plugging to the router.

Thus, several conclusions can be drawn from this scenario:

1) We can say printing inside a home network mainly cause in-home traffic.
2) With a wired printer, printing a small file requires less bandwidth than printing a larger file, while for the printer with wireless connection, the bandwidth requirement under these two conditions are the same.
3) The connection change of home devices will cause very high bandwidth consumption.

5.2.2. *Scenario 2: Different Connectivities of Smart TV*

Without Internet, we implement a scenario that wired and wireless smartTV gets access to different resources on NAS and discover that smart TV with wireless connection consumes around 40Mbps, which is much more than the smart TV with wired connection(the required bandwidth is around 670Kbps), shown in Fig.5.3. In Fig.5.3a, from the time 16:42 to 16:48:40, the smart TV is playing a small video with resolution of 720 × 576. When we forwards the movie, the bandwidth will increase in the next 10s occupying a 4s delay, while when the wireless smart TV plays a HD TV with the resolution of 1280 × 720, the latency decreases to 2s. However, similar phenomenons doesn't appear with a wired smart TV. The required bandwidth for both small video and HD video are the same, and when forwarding or rewinding a small video or a HD video, the video can be continuously played.



(a)

(b)

Figure 5.3: *Internet traffic and in-home traffic with different connections of smartTV accessing to NAS*

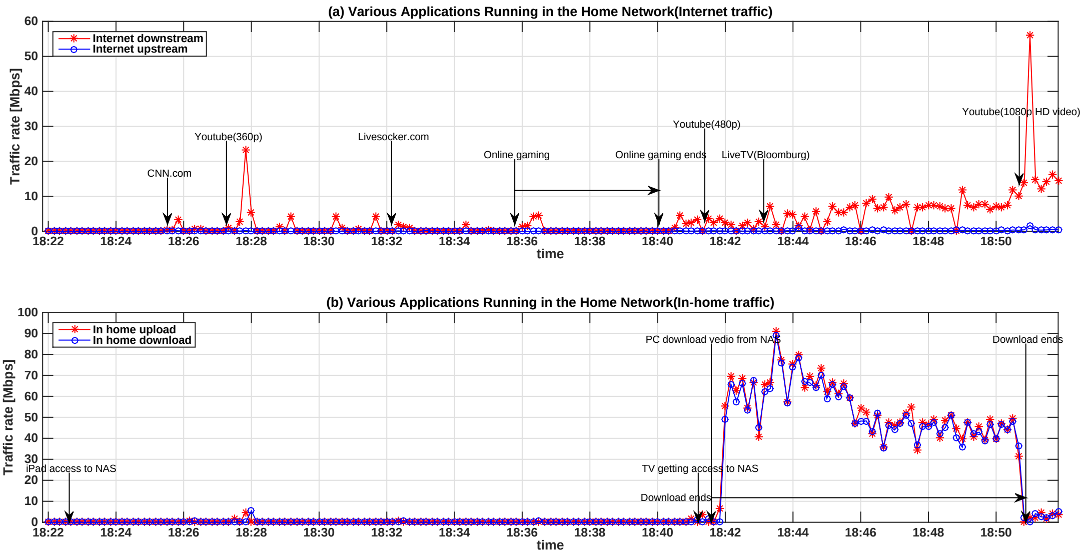


Figure 5.4: *Bandwidth usage of various applications getting accessed*

5.2.3. *Scenario 3: Various Applications Running in the Home Network*

In this scenario, various applications performance in the home network. We plot the internet traffic and the in-home traffic with the corresponding activities, shown in Fig.5.4.

For internet traffic Fig.5.4 (a), the consumption of most applications are smaller than 10Mbps, except watching a HD video on Youtube requires larger bandwidth. And we also find the bandwidth requirement of watching a live video is smaller than watch a video already existed on the website, because there is no buffer for live streaming. And for example, watching video on the youtube, it will automatically buffer when you open it, which requires higher bandwidth.

5.2.4. *Other Scenarios*

Apart from the aforementioned three scenarios, we also checks the bandwidth usage of different devices for same application, for example, youtube. And we discover, the mobile phone connecting to the network using Wi-Fi has shorter buffer time and higher bandwidth requirements than watching the same video on a PC. Whereas, after short buffering, in both situation, traffic bitrates drop dramatically.

5.3. *Conclusion*

In the chapter, we propose an experimental platform to measure home network traffic. According to this platform, we measure the traffic going through each port of the residential gateway and record the corresponding network activities. Afterwards, we analyse some scenarios on different applications and connections.

6

Conclusion & Future Work

6.1. Conclusion

With the increasing network requirement of home users and new services getting accessed, how to manage network resource and safeguard the QoS of new applications has been a tough problem for service provides. In this section, we discuss the answer to our research questions.

Research Question 1: *Unlike other kind of network traffic, one significant feature of home network traffic is the small traffic bit rates. Apart from this, what kind of characteristics does this kind of traffic has?*

According to the given traffic measurements, the maximum traffic bit rate is around 92 Mbit/s and 92.85% of all traffic measurements are smaller than 1 Mbit/s. We depict all traffic bit rates in 15 households through probability mass function(pmf) and find the pmf follows a power law or a generalized Pareto distribution, which indicates a significant feature of home network that it has long period of low network activities and short period of high bandwidth usage.

The third characteristic of home network is heterogeneous. Based on our traffic measurements, we divide all traffic samples into 7 cycles for 15 households and discover, for each home network, traffic rates have different scales and very high traffic are randomly distributed.

Research Question 2: *Home network traffic strongly depends on Internet habits of home users. So to develop a prediction, with which aspects can we start to solve this problem? And how to do that?*

Due to the heterogeneous of home network, it is really hard to determine which aspects we can do to achieve this target. As mentioned in the last question, most traffic bit rates are relatively small. Thus, we intend to pick up relative high traffic in a home network and predict the traffic behaviors. Our contribution in this thesis is to define a new burst concept combining our traffic measurement and to systematically predict future burst behavior using Markov chain methodology. The detailed steps are described.

The given incoming traffic in 15 households is the sum traffic of various applications and the background traffic. The background traffic as defined in this project,refers to the traffic used to self-contain program update and daemon running. This part of traffic is regarded as basic network bandwidth demand to support normal operation of a home network. Suppose in each home network, the background traffic is equal to the mean traffic bit rate. So traffic bit rates, smaller than this threshold can be ignored and the remaining traffic will be used to characterize the behavior of home network, notated as nonzero traffic. By taking all indices of nonzero traffic, we apply Gaussian mixture model(GMM) to learn the relations among traffic points. According to the character of Gaussian distribution, the mean value in each Gaussian component has the highest probability density, which means at this point traffic is most likely to happen. Then we define the 1-sigma area around this index is a burst and the 95th percentile bandwidth in this area is used to present the burst height.

Based on the idea of *burst* and results of Gaussian mixture model in each home network, we apply Markov chain methodology to predict when the next peak point happen, the width of the burst and how much traffic will enter into the network during this burst. At the end of the model, we use leave-one-our cross validation(LOOCV) method to analyze the performance of our model.

One advantage of our model is to give service provides and vendors an insight of which period traffic are mostly likely to happen and how much they need to allocate to the home network. Due to the memoryless of Markov Chain, the home gateway will not be required to measure a very long time traffic .

Research Question 3: *Whether is it possible to implement a controlled environment in the lab to simulate some ofter used applications and services to let Operators roughly know what is happening inside a home network?*

Apart from traffic model part, this thesis also establishes an experimental platform to measure new traffic data in a controlled environment. During this experiment, we registered every details of all activities happened in the network for seven working day. By analysis the effect of different applications on Internet traffic and in-home traffic, it's possible to provide operators a rough knowledge of traffic activities inside home.

6.2. Future Work

- In our Gaussian mixture, we use an indirect way to determine the range of mixture model. According to our results of GMM, numbers of Gaussian components are much smaller than the upper bound. In order to reduce computational complexity, how to define reasonable numbers for different households needs to be improved.
- The traffic measurement used in our thesis are measured in 2010. During to quick evolvement of applications and services, we are not sure whether this prediction model can will predict the current home traffic. Thus, new real home network traffic needs to be measured. And then apply our prediction methodology to new data set.

A

Appendix for State Validation of Markov Chain in the Previous Traffic Model

In the traffic prediction model proposed by A.Delphinanto [13], the normalized entropy are evenly divided into 5 states, each of which corresponding to one states with concrete theory evidence. To make up this shortage, considering traffic measurements are based on time series, we choose dynamic time warping (DTW) algorithm to compare the similarity between two sequences to prove the rationality of Markov chain states.

A.1. Introduction of Dynamic Time Warping

Dynamic time warping (DTW) is a method to find an optimal alignment between two sequences under some constraints [35] by ignoring differences in time dimension. It measures a distance-like values between two certain sequences. Suppose two sequences $X = (x_1, x_2, ..., x_n)$ for $n \in [1, 2, ..., N]$ and $Y = (y_1, y_2, ..., y_m)$ for $m \in [1, 2, ..., M]$. Denote a feature space by $\mathcal{F}$ (that is, for all elements in X and Y belongs to $\mathcal{F}$) and the local distance (cost) measure is chosen as the parameter to depict this feature by the function

$$c : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}_{\geqslant} 0 \tag{A.1}$$

Generally, if X and Y are similar to each other, $c(x, y)$ is small, otherwise $c(x, y)$ is larger with less similarity between X and Y. By estimating local distance between each pair of sequence X and Y, cost matrix C can be obtained by $C(n, m) := c(x_n, y_m)$. According to the local cost measurement, the total cost $c_p(X, Y)$ of a warping path between sequence X and Y can be written as

$$c_p(X,Y) := \sum_{l=1}^{L} c(x_{n_l}, y_{m_l}) \tag{A.2}$$

An optimal warping path $p*$ between X and Y has the minimal local cost total cost among all possible warping paths. And this is the idea that DTW algorithm intends to achieve. Based on cost matrix C, a related concept accumulated cost matrix D is defined with size $N \times M$

$$D(n,m) := DTW(x_{1,2,...,n}, y_{1,2,...,m}). \tag{A.3}$$

Meanwhile, the accumulated cost matrix M has the following identities:

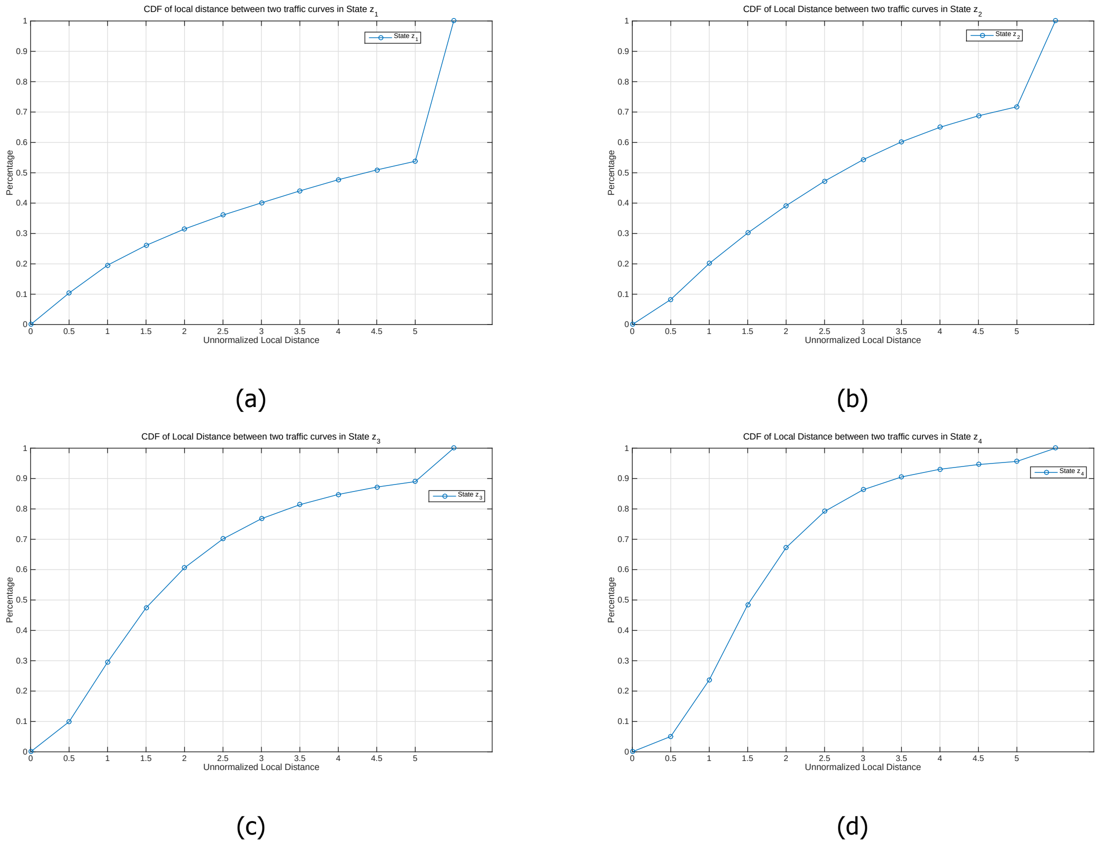
(1) $D(n,1) = \sum_{k=1}^{n} c(x_k, y_1)$, for $n \in [1,2,...,N]$;
(2) $D(1,m) = \sum_{k=1}^{m} c(x_1, y_k)$, for $m \in [1,2,...,M]$;
(3) $D(N,M) = \min\{D(n-1,m-1), D(n-1,m), D(n,m-1)\} + c(x_n, y_m)$, for $1 < n \leq N$ and $1 < m \leq M$.

A.2. Implement and Results

Obviously, DTW$(X,Y) = D(N,M)$ with operation complexity $O(MN)$. In our project, each two traffic curves in one states with same length 50 samples. In one comparison, the optimal warping path has obtained after computing with $O(2500)$. For our traffic measurement, there are 1209 traffic states with four possible states (z_1, z_2, z_3, z_4) in 15 households and we find that for each network around 1000 state samples are in state z_1. Thus, the calculation in states z_1 will be calculated more than 10^8 times, combining with the calculation of DTW algorithm, operation complexity increases to 2×10^{11} times. The detailed steps of local distances calculations are as following :

(1) According to four Markov states z_1, z_2, z_3, z_4, traffic curves are separately recorded;
(2) Normalize each traffic curve by dividing the maximum traffic rate in this curve. Thus, all traffic rates are in the range of 0 to 1;
(3) In state z_1, horizontally calculate local distance between each two traffic curves;
(4) Considering computing speed, partly compare distances among all home networks;
(5) In state z_2, z_3, z_4, compare each two traffic curves among all households;
(6) Setting the bin size to 0.5, we calculate the number of elements per bin and cumulative distribution of DTW distances.

We plot the CDF of local distances between every two traffic curves in state z_1, z_2, z_3, z_4 (Fig. A.1) from above steps. The last value of X-axis corresponds to the local distance



(a)

(b)

(c)

(d)

Figure A.1: *(a), (b), (c), (d) show the local distance distribution with four states z_1, z_2, z_3, z_4 respectively among all home networks. X-axis stands for local distance obtained by DTW algorithm with the range from 0 to infinity. The last point with y=1 corresponds to x value, which is larger than 5.*

which is larger than 5. In state z_1 (Fig. A.1a), almost half of local distances are larger than 5, which means every two traffic curves in this state has very little similarities. This phenomenon improved a little bit in state z_2 with more than 70% of distance are smaller than 5. Futhermore, state z_3 and z_4, indicating higher uncertainty of home network traffic, obtain smaller local distance of warped paths from the histograms shown in Fig. A.1c and Fig. A.1d (In state z_3, about 60% local distance are smaller than 2, as well as 67.25% in state z_4).

In home network traffic prediction, Operators intend to know more about traffic uncertainty to better allocate network resources. Obviously, the kind of state divisions is propitious to predict states with higher uncertainty, which satisfies the aim of traffic prediction. Besides, local distance histograms of each home network in state z_1 are plotted shown in Fig. A.2. In each home network, about 50% local distance are smaller than 2, and more than 70% distance are smaller than 5. Thus, we can say there exists a certain level similarity between each states. And with the increasing of traffic uncertainty, it appears higher similarities in related states. This character satisfies the prediction aim, which is reasonable to be used in Markov prediction methodology as well.

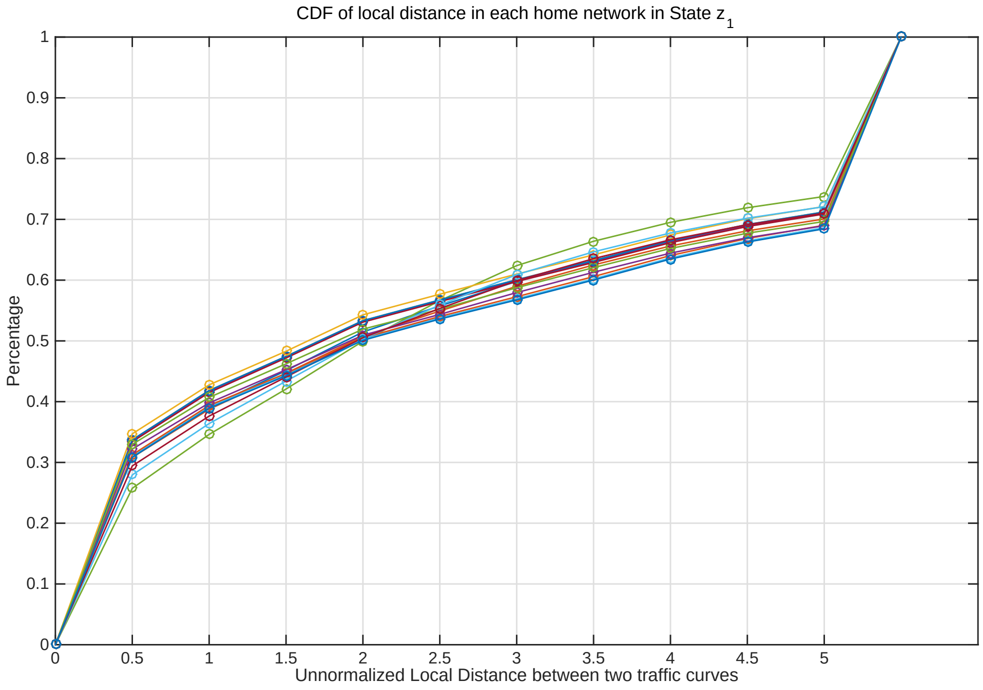


Figure A.2: *Local distance between each two traffic curves in each home network using DTW algorithm. X-axis represents the unnormalized distances, with the range from 0 to infinity. The last point with y=1 in the graphs corresponds to x value, which is larger than 5.*

Bibliography

[1] N. Jayant, Broadband Last Mile: Access Technologies for Multimedia Communications (CRC Press, 2005) chapter 6.

[2] W. K. Edwards, R. E. Grinter, R. Mahajan, and D. Wetherall, *Advancing the state of home networking,* Communications of the ACM **54**, 62 (2011).

[3] R. T. Ahlem Reggani, Fabian Schneider, *An end-host view on local traffic at home and work,* in Passive and Active Measurement (Springer, 2012) pp. 21–31.

[4] K. Cho, K. Fukuda, H. Esaki, and A. Kato, *The impact and implications of the growth in residential user-to-user traffic,* in ACM SIGCOMM Computer Communication Review, Vol. 36 (ACM, 2006) pp. 207–218.

[5] T. Karagiannis, C. Gkantsidis, P. Key, E. Athanasopoulos, and E. Raftopoulos, Homemaestro: Distributed monitoring and diagnosis of performance anomalies in home networks, Tech. Rep. (Tech. Rep. MSR-TR-2008-161, Microsoft Research, 2008).

[6] G. Maier, A. Feldmann, V. Paxson, and M. Allman, *On dominant characteristics of residential broadband internet traffic,* in Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference (ACM, 2009) pp. 90–102.

[7] K. Xu, F. Wang, L. Gu, J. Gao, and Y. Jin, *Characterizing home network traffic: an inside view,* in Wireless Algorithms, Systems, and Applications (Springer, 2012) pp. 60–71.

[8] https://www.techopedia.com/definition/29977/network-traffic-monitoring, accessed Dec. 10, 2015.

[9] D. Joumblatt, R. Teixeira, J. Chandrashekar, and N. Taft, *Performance of networked applications: the challenges in capturing the user's perception,* in Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack (ACM, 2011) pp. 37–42.

[10] R. T. Ahlem Reggani, Fabian Schneider, *Tracking application network performance in home gateways,* in Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International (IEEE, 2013) pp. 1150–1155.

[11] S. Sundaresan, W. De Donato, N. Feamster, R. Teixeira, S. Crawford, and A. Pescapè, *Measuring home broadband performance,* Communications of the ACM **55**, 100 (2012).

[12] https://www.paessler.com/info/network_traffic_monitor, accessed Nov 28, 2015.

[13] A. Delphinanto, Network and Service Monitoring in Heterogeneous Home Networks (Technische Universiteit Eindhoven, 2012).

[14] K. Xu, F. Wang, and M. Lee, *Hometps: Uncovering what is happening in home networks,* in Consumer Communications and Networking Conference (CCNC), 2012 IEEE (IEEE, 2012) pp. 40–41.

[15] http://searchsoa.techtarget.com/definition/session (), access Jan.15,2016.

[16] http://techterms.com/definition/session (), accessed Jan 15, 2016.

[17] https://en.wikipedia.org/wiki/OSI_model, accessed Jan 16, 2016.

[18] S. Capecchi, I. Castellani, M. Dezani-Ciancaglini, and T. Rezk, *Session types for access and information flow control,* in CONCUR 2010-Concurrency Theory (Springer, 2010) pp. 237–252.

[19] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, *Characterizing user behavior and network performance in a public wireless lan,* in ACM SIGMETRICS Performance Evaluation Review, Vol. 30 (ACM, 2002) pp. 195–205.

[20] Network Region Configuration Guide (Communication Manager 3.0), Tech. Rep. (Avaya Labs, October 2005).

[21] M. F. Zhani, H. Elbiaze, and F. Kamoun, *Analysis and prediction of real network traffic,* Journal of networks **4**, 855 (2009).

[22] M. Dashevskiy and Z. Luo, *Network traffic classification and demand prediction,* http://www.ntu.edu.sg/home/SSHo/CPBook/Chapter12.pdf, accessed Dec 18, 2015.

[23] F. H. T. Vieira, G. R. Bianchi, and L. L. Lee, *A network traffic prediction approach based on multifractal modeling,* Journal of High Speed Networks **17**, 83 (2010).

[24] M. W. Garrett and W. Willinger, *Analysis, modeling and generation of self-similar vbr video traffic,* in ACM SIGCOMM Computer Communication Review, Vol. 24 (ACM, 1994) pp. 269–280.

[25] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec, *Do cascades recur?* in Proceedings of the 25th International Conference on World Wide Web (International World Wide Web Conferences Steering Committee, 2016) pp. 671–681.

[26] C. M. Bishop, *Mixture models and the em algorithm,* Microsoft Research, Cambridge (2006).

[27] C. Tomasi, *Estimating gaussian mixture densities with em – a tutorial,* http://www.cse.psu.edu/~rtc12/CSE586/papers/emTomasiTutorial.pdf.

[28] https://en.wikipedia.org/wiki/Bayesian_information_criterion, accessed in Sep. 10, 2016.

[29] https://en.wikipedia.org/wiki/Burstable_billing (), access Sep.20, 2016.

[30] *95th percentile bandwidth metering explained and analyzed,* ().

[31] T. M. Cover and J. A. Thomas, Elements of information theory (John Wiley & Sons, 2012).

[32] K. Xu, Z.-L. Zhang, and S. Bhattacharyya, *Profiling internet backbone traffic: behavior models and applications,* in ACM SIGCOMM Computer Communication Review, Vol. 35 (ACM, 2005) pp. 169–180.

[33] *Cross-validation (statistics),* https://en.wikipedia.org/wiki/Cross-validation_(statistics), access on Oct. 15,2016.

[34] https://wiki.openwrt.org/toh/hwdata/linksys/linksys_wrt1900ac_v1, access in Aug.20, 2016.

[35] M. Müller, Information Retrieval for Music and Motion, Vol. Chapter 4 (Springer, 2007).