Understanding Internet User Behavior: Towards a Unified Methodology

Christina Lagerstedt, Andreas Aurelius,
Hemamali Pathirana, Claus Popp Larsen
Netlab
Acreo AB
Kista, Sweden
christina.lagerstedt@acreo.se
andreas.aurelius@acreo.se
hemamali.pathirana@acreo.se
claus.popp.larsen@acreo.se

Olle Findahl
World Internet Institute
Gävle, Sweden
olle.findahl@wii.se

***Abstract*—Understanding user behavior and Internet usage patterns is fundamental in developing future access networks and services that meet technical as well as end user needs. User behavior is routinely studied and measured, but with different methods depending on the research discipline of the investigator, and these disciplines rarely cross. We tackle this challenge by combining measurement methods from different research disciplines, with the purpose of finding a complete picture of the user behavior and working towards a unified analysis methodology. In this paper, we provide results, based on measurements performed with three different methods: Internet traffic measurements, web questionnaires and diaries. All measurements were performed in the Acreo National Testbed, where we have access to both the network infrastructure and the end users. By comparing the different measurement methods we find that it is difficult for the end users to estimate the time they spend on different Internet activities although they are fairly well able to estimate the frequency of usage. We see that it is more difficult for end users to estimate their usage of streaming media applications than their HTTP traffic. We also find that though the diaries are quite accurate, the traffic measurements give us a much more detailed picture of the end user activity. We conclude that having a testbed with real end users is invaluable to this kind of study and we emphasize the importance of having access to Internet traffic to gain detailed knowledge of end user behavior.**

Keywords-testbed; traffic measurements; user behavior; FTTH

I. Introduction

In a previous paper [1], we have described our work on the development of a unified methodology for Internet behavior usage measurements. In this paper we will expand the work concerning the methodology description and include additional results.

As the Internet continues to evolve and offer new services, it takes up a larger part of our lives. We find new ways to communicate, interact and entertain ourselves. This puts new demands on access networks [2] and requires new insights into the behavior of those who use them. We believe that understanding user behavior and needs is the key to developing future networks and services that are accessible, reliable and that address the needs of real end users.

There are several ways to study user behavior. From the technical side, the data traffic can be measured and analyzed. Other common ways are to use surveys or diaries. Traffic measurements are routinely performed by operators, but results are rarely published because the operators are unwilling to share this information with competitors. On the other hand, published behavior studies are almost always based on surveys with individuals (e.g., telephone interviews with a large population). Such surveys often attract considerable interest in the public debate, and far-reaching conclusions may be drawn.

The question that we deal with in this paper is: To what extent are such surveys reliable? People can forget about their Internet activity, they may not know what they did, they may not know what their children did, they may lie about sensitive subjects, etc. We believe that in order to get a comprehensive picture of the evolution of Internet behavior, a combination of methods from different disciplines is needed. This will lead to a more detailed knowledge of the user behavior, and, by evaluating results from different methods we will obtain a better knowledge of their respective limitations. Furthermore, this way it can be verified whether surveys are correct or whether there perhaps is a systematic bias in survey answers that may lead to misleading results.

The purpose of this study is to develop a unified methodology where different kinds of surveys are combined with actual traffic measurements. We compare three different methods of looking at Internet user behavior;

1) Internet protocol (IP) traffic measurements,
2) web questionnaire and
3) diary.

Combining technical measurements with surveys requires test subjects in order to get statistical data. At Acreo we put a lot of effort into developing the Acreo National Testbed, ANT, which enables us to perform in-depth measurements and test new technology and equipment as well as to interact with end users. We have signed agreements with the end users where they agree to give feedback and participate in surveys and investigations. This means that the network conditions such as network topology, link speeds, service setup, etc as well as user metadata such as the number of people in a household, age, etc are known. This gives us a unique opportunity to perform measurements of user behavior, and to compare the results from different measurement techniques in order to evaluate the validity of the results since we have access to real user traffic measured on a household level.

The rest of the paper is organized in the following way. Section II describes the state of the art. Section III outlines the details and limitations of the specific methods. The test environment and the test population are described in Section IV, while data on the specific measurements is detailed in Section V. The measurement results are compared in a systematic way in Section VI, where we make some general observations and go through the results from the respective methods one-to-one. In this systematic review, we also discuss the limitations of the methods and the currently performed measurements in Section VII. Section VIII is dedicated to conclusions and future work.

II. State of the art

There are traffic measurement studies in the literature. Most of these are based on traffic from campus areas [3] [4] or measure highly aggregated traffic [5] [6] [7] [8] [9] [10] [11], which makes it very difficult to draw conclusions about detailed user statistics. IP network traffic studies on a household level can be found in [12] [13] but as far as we know traffic measurements have not been used together with questionnaires and diaries, which are well known and often used methods [14] [15] [16] [17] [18] [19] [20] [21] [22].

Surveys, administered to a sample of the population, are the most common method to get information about Internet behavior [16]. An interviewer can ask questions by phone or at home. The questionnaire can also be sent by post or be available online and be administered by the individuals themselves. Questions can be asked about attitudes, beliefs and past behaviour, i.e., what people generally think and do.

Another way to get to know how people use the Internet is to let them fill in a diary, where they make notes about what they do during the 24 hours of the day. The diary method has a long tradition in social sciences and is normally used in field studies where the data is collected in situ, from real people in real situations [23]. Notes about what the person is doing during the 24 hours of the day are written down. The method has been used to measure different aspects of Internet behavior such as time displacement [24] or usage patterns [25] although not as frequently as surveys [26].

Both diary studies and surveys are prone to human error in answering questions and filling in diaries. We know that answers gathered from these types of surveys have systematic errors due to the difficulty of estimating time and the unwillingness to answer questions regarding certain areas of Internet usage (e.g, illegal such as file sharing and personal, such as sites with adult content) [27].

A third method, without these limitations, is Internet traffic measurements. This method measures all traffic going in and out from a household or user to the Internet. It can be used to register for example traffic volumes, applications used and web sites visited as well as background traffic such as automatic updates.

III. Pros and Cons of the Methods

In this section, we will discuss the strengths and weaknesses of the different measurement methods used in this study: web questionnaires, diaries and traffic measurements.

A. Questionnaire

A questionnaire is an efficient way of collecting information from a large number of respondents. Questions are administered to individuals, by phone or at home, by post or online. Questions can be standardized; the sample can be representative of the population, which is a major strength. This method also gives the possibility of covering many areas from attitudes and reflections to general behavior and very detailed information.

A weakness of this method is that the outcome is strongly dependent on the ability of the respondents to answer the questions. This depends partly on the ability of the researcher to formulate relevant and understandable questions and partly on the willingness of the respondents to answer sensitive questions in a truthful way. Also their ability to estimate how often they perform different activities on the Internet and how much time they spend online during an average week is crucial for the outcome of the method. Triangulation with other methods of measuring Internet behavior can help resolve these issues.

B. Diary

If a questionnaire measures past behavior, a diary focuses on the present. It is about real people in their everyday life, at a specific time and in a specific place. This is also the strength of the diary. Diaries also have an exploring quality that surveys lack, which can lead to unexpected discoveries. To keep a diary during a few days requires a high level of commitment from the participants, which is a weakness. There is therefore a need for detailed instructions and frequent administrative contacts [28]. To what extent the behavior noted down in the diaries is exhaustive and reflects the simultaneous measurements of Internet traffic as well as their regular Internet behavior will become obvious during the triangulation.

C. Traffic Measurements

Traffic measurements record actual network traffic capturing all user activity without bias or human error. Depending on the equipment and methods used, a deep level of detail can be achieved. It is also possible to measure activities that are not induced by active end users such as automatic updates or applications that are left running with no user present such as file sharing applications. This may be both a possibility and a disadvantage, since it may be difficult to distinguish between user induced activities and computer induced activities.

One disadvantage of this method is the enormous amount of data generated. It requires large storage space and efficient analysis tools. The important question is whether the analysis and classification of traffic measurements can make these data comparable with data from the questionnaires and the diaries. There is also the problem of coupling measurements of traffic from an IP-number to a specific end user for example with shared computers or in family households. Having the possibility of addressing a test population with a given profile is also a limiting factor of the method. In general, traffic measurements are collected in geographically specified populations, whereas for a specific test one may require a population which is representative of the country, i.e., defined by parameters such as age, sex, occupation, etc.

IV. Test Environment and Population

As mentioned in the introduction, the measurements in this study were performed in the Acreo National Testbed (ANT), which has previously been described in [29] and [30]. The purpose of ANT is to provide an environment for testing new technology and equipment as well as a way to interact with end users. Contrary to lab based testbeds, this is a live network with real end users or test pilots. In return for being test pilots the end user households are given free access to services like Internet and IPTV. Fiber to the home (FTTH) is the main access technology in the testbed, and a schematic picture of the network is shown in Figure 1.

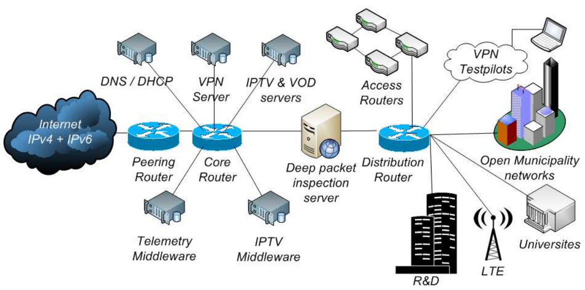

Figure 1: Schematic picture of the Acreo National Testbed, ANT.

The FTTH installation at the test households is active Ethernet providing 100 Mbit/s symmetrical connections to each household. The number of households in the testbed changes over time according to the current tests that are being performed. At the time of this study, there were approximately 40 active households in the testbed. Approximately 20 were apartments in a building centrally located in a town in Sweden. The rest of the households, approximately 20 single dwelling units, were connected to the testbed via a fixed wireless access network, depicted as HSPA (High Speed Packet Access) in Figure 1.

A letter of invitation was sent out to all users in the testbed describing the measurements to be performed and asking for volunteers. A form was attached to the invitation letter, where detailed information about the household was to be filled in, such as the number of people in the household, number of computers, ages, etc. The test pilots were also assured that their answers would be treated anonymously.

Based on the response from the test users, 5 households were selected to take part in this study. The selection criteria for the 5 households were that different kinds of household constellations should be represented. Thus 2 single households and 3 family households were chosen. Of the family households, one had preschool children, one had teenaged children and one had both younger and older children. The household details are summarized in Table 1. Household number 2 had two IP-numbers, one used by the parents that will be denoted 2.1 and one mainly used by the teenagers of the family that will be denoted 2.2.

All the participants, except those twelve years or younger, were later classified into different user groups according to usage patterns revealed by their answers to the web questionnaires. The typology has been accomplished through two separate studies which encompass 2000 telephone interviews, based on a random sample of the Swedish population from 16 years of age and older [31] [32].

Characteristic for the groups of Internet users that turned up in the cluster process is that there were two extremes. One group, the advanced enthusiasts, has access to everything and uses Internets potential fully. The other group, its opposite, the cautious, have limited access to new technology and use it very restrictedly. Between these two extremes there are a number of groups whose usage patterns differ in various ways. But they are essentially variants of two basic patterns: the traditional and the modern. The traditional usage pattern is built around the Internets informative qualities and the modern usage pattern, supported by the young Internet generation, rests on the interactive, communicative characteristics of the Internet.

The participants in the study were sent gratifications in the form of movie tickets when their responses had been received. They were not informed that they would receive any gratifications beforehand and the gratification was thus not an incentive for taking part in the study.

Table 1: Characteristics of the members of the 5 households.

Household	Household members	Age	Education	Usage pattern
1	Man	30	University education	Enthusiastic traditionalist
2	Woman	41	University education	Traditionalist
	Man	46	Higher certificate	Traditionalist
	Boy	15	At school	Advanced enthusiast
	Girl	17	At school	Enthusiastic modernist
	Girl	15	At school	Enthusiastic modernist
3	Woman	34	9-year compulsory	Modernist
	Man	32	University education	Traditionalist
	Pre-school child	4		Video and gaming
	Pre-school child	6		Video and gaming
4	Man	58	Higher certificate	Traditionalist
5	Woman	39	Higher certificate	Traditionalist
	Man	35	More than compulsory, no higher certificate	Traditionalist
	Boy	17	At school	Advanced enthusiast
	Boy	12	At school	Moderate use; gaming, file sharing

The measurements in this paper are based on data from 5 testbed households. The reason for using a small population in the study is that we wanted to perform a qualitative analysis of their Internet usage, i.e., the main point is not to gather data for statistical analysis, but rather to study the user behavior in depth, as we have not found any similar studies in the literature. As this is a novel comparative methodology it was also important to develop methods and analysis tools that can be scaled up to considerably larger populations [33] where the amount of data to be handled will be much more extensive. That is, we do not claim to be able to make statistical conclusions in this study, but interesting trends for further study will be pointed out.

We are not able to include all types of users detailed in Table 1 with this small population. We have however made an effort to include households with different profiles to be as inclusive as possible. Nevertheless, before extending our population, we would like to establish a solid foundation for our work on developing a unified methodology that can be used for more substantial conclusions at a later stage.

V. Measurements

The combined methodology consists of three parts: traffic measurements, web questionnaires and diaries. The details of each type of measurement are given below.

A. *Traffic Measurements*

The traffic measurements have been performed using PacketLogic (PL) [34], a commercial traffic management device used in many commercial broadband access networks all over the world. Traffic is identified based on packet content (deep packet inspection and deep flow inspection) instead of port definitions. The device can identify more than 1000 Internet application protocols, and the signature database is continuously updated. The identification process is connection-oriented, which means that each established connection between two hosts is matched to a certain application protocol. When a new connection is established, the identification of this connection begins. The identification algorithm searches for specific patterns, signatures, in the connection. The patterns are found in the IP header and application payload. The PL uses the traffic in both directions in the identification process. The measurement point is depicted as deep packet inspection in Figure 1.

The PL can track and identify several hundred thousand simultaneous connections, storing statistics in large databases. The statistics database records the short-time average amount of traffic in inbound and outbound directions as well as the total traffic for all nodes in the network. Data may be stored aggregated over 5 minute periods or detailed on a per connection basis.

The measurement setup, although giving detailed measurements, has certain constraints. First, the traffic is measured per household and not per person and the analysis in this paper is therefore done on a per household basis. There is also a 5 minute resolution in the measurements, which may have an impact on measurements of applications that are used in short time periods such as instant messaging. The data cut-off is 1 kbps, which may influence the measurements of certain applications such as gaming where the amount of data is generally very low. The signature database in this study was not up-to-date due to old hardware, which may result in a larger amount of unknown traffic. An upgrade will be performed before follow-up studies are performed. The Internet traffic of each household was measured both during the days when the household recorded their diaries (17-18 May 2009) and for a complete month (May 2009) to get enough statistics to compare with the web questionnaire. Statistics on what web sites were visited by the different households were monitored for two weeks.

B. *Web Questionnaires*

Each household member was asked to answer a web questionnaire regarding their long-term media and Internet activity/behavior in general during the last weeks and months. In the case of preschool children, the parents were asked to answer for them. The questionnaire used was basically the same as the one used by the World Internet

Institute in their yearly study of the Internet behavior of 2000 Swedes [35]. It contained basic questions concerning family situation, education and occupation as well as questions about attitudes, computer knowledge and Internet activity. The Internet activity questions included questions about the frequency of use of different Internet applications and how often different types of web sites are visited such as banks or newspapers, etc. Finally, the respondents were asked to estimate the total time they usually spend online at home, at work, at school and in other places.

C. Diaries

The members of each household were asked to complete a 24-hour diary logging their activities during two consecutive days, Sunday and Monday 17-18 May 2009. In the diary the test pilots were asked to fill out four columns:

- Daily activity (sleep, work/school, leisure time activities, meals etc.)
- Media usage (TV, newspaper, radio, book, etc.)
- Internet activity when at home (web browsing, playing games online, visiting community, downloading material from the internet, etc.)
- Web address or service/application used

Each day was divided into 15-minute intervals. In the case of preschool children in the household, the parents were asked to note down their Internet activities.

VI. Results

In this section, the results of the measurements are presented. First, some general results are given and then the results from the different methods are compared.

A. General Results

Comparing data from the questionnaires with the results from a representative study of the Swedish population using the same questionnaire, we can classify the Internet usage patterns of the test persons according to the four typical usage patterns that have earlier been found in [31].

The most striking usage pattern is that of the advanced enthusiasts. Two teenage boys in households 2 and 5 belong to this group. They use the Internet for everything to a much larger extent than most people. They share files, use social networks and read blogs. The Internet is very important to them. Their opposites are the cautious. They do not spend much time on the Internet and when they do, its in order to look for facts or find information. No one in this study belongs to this group.

Between these two extremes we find the majority of Internet users, the traditionalists and the modernists. The modernists are mostly interested in communication and entertainment. But they also use the Internet for information and fact-finding. There are three modernists among the test persons. They are all women and two of them, the teenage girls (household 2), are so called enthusiastic modernists as they spend a lot of time online.

The traditionalists are the largest group in the population as well as among the test persons. They do not spend as much time on the Internet as the preceding groups and are mostly interested in its traditional role as a source of information, for checking facts and for practical matters. They are generally positive to the Internet but regard other media as more important. Two women (household 2 and 5) and five men (household 1, 2, 3, 4, and 5) can be classified according to this usage pattern. The Internet usage patterns among the test persons are shown in Table 1.

The traffic measurements show that the households participating in the study are active mainly during afternoons and evenings with shorter bursts of traffic during the morning and lunch hours. This corresponds to the traffic patterns established both in the ANT-testbed, see Figure 2, and in municipal networks of similar characteristics [36].

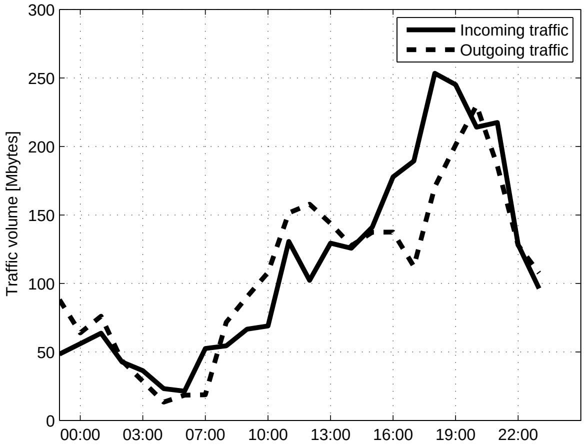

Figure 2: Schematic picture of the Acreo National Testbed, ANT.

We also note that the average time spent online calculated from the traffic measurements is greater during the weekend than the weekdays for the family households while the opposite is true for the households without children, see Table 2. The assumption here is that for the family households, the time of day when household members use Internet applications will be more spread out during the weekend and of course there are more people at home with leisure time. However, this assumption should be confirmed for a larger population.

B. Traffic Measurements (Long Term Behavior) vs. Web Questionnaires (Long Term Behavior)

The over-all estimation in the questionnaires (Q) of the frequency of use of the Internet corresponds to what is shown in the traffic measurements (TM). All households use the Internet daily. However, one older man, a traditionalist in

Table 2: Average time per day spent on Internet applications per household as registered by the traffic measurements performed during May 2009.

Household	All days [min/day]	Weekends [min/day]
1	253	111
2.1 / 2.2	432 / 75	614 / 66
3	588	496
4	196	154
5	1047	1162

Household 4, underestimates his general frequency online. Traffic measurements show that he uses the Internet daily but in his answer to the questionnaire he states that he uses Internet a few times a week. At the same time he says that he visits certain web sites daily, which corresponds with the results from the traffic measurements. So in the main, the questionnaire answers as to how often the test pilots were online, were confirmed by the traffic measurements, see Table 3.

Table 3: Frequency of general Internet usage. A comparison between the results from traffic measurements (TM) and a questionnaire (Q) given to the members of 5 households about their Internet behavior at home. The numbers represent the number of days where activity was registered by the measurement equipment. D=Daily usage, W=Weekly usage, Y=Yes.

Household	Frequency		HTTP (days/month)	
	TM	Q	TM	Q
1	D	D	23	D
2.1	D	D	16	D
2.2	D	D	28	D
3	Y	Y	27	D
4	D	W	28	D
5	D	D	26	D

In the families with teenagers, the questionnaires show that the children are the main users of the Internet. The traffic measurements cannot couple a specific person to a specific Internet activity. However the measurements show a traffic mix corresponding to the expectations from the age group with parents visiting banks and web shops and teenagers using instant messaging, gaming, blogging and visiting social communities [35].

The user penetration of a number of applications seen in the traffic measurements is found in Table 4. This is in good agreement with the answers from the web questionnaires. As is expected, HTTP is used by all households as well as the SSL protocol, which is used by for example Internet shops and banks. All of the households also use HTTP media stream as well as flash video, which means that they look at streaming material on the Internet.

The frequencies of use reported in the questionnaires were seen, on a general level, to correspond well with the measured data. The questionnaire also posed questions about the frequency of use of more specific Internet activities such as using e-mail, visiting banks online, reading newspapers online, visiting social networks, using instant messaging, phoning online, listening to music through Spotify or using file sharing applications to download music or film. The answers from the questionnaires are for the most part confirmed by the traffic measurements, see Table 5. This includes activities such as file sharing and listening to streaming music online through Spotify. The test pilots were able to give estimations of how often they used for example Spotify or file sharing applications and how often they visited certain types of web sites such as newspapers or banks.

Table 4: User penetration of specific applications seen in the traffic measurements.Specific use of the Internet. A comparison between traffic measurements (TM) during 28 days and web questionnaires (Q) concerning Internet behaviour at home. Y=Yes, N=No, D=Daily. User penetration of specific applications seen in the traffic measurements.

Application Household	HTTP	HTTP Media Stream	BitTorrent	Spotify	SSL	Flash video	MSN Messenger	Skype
1	x	x	x		x	x	x	
2	x	x	x	x	x	x	x	x
3	x	x		x	x	x		
4	x	x			x	x		
5	x	x	x		x	x	x	

Table 5: User penetration of specific applications seen in the traffic measurements.Specific use of the Internet. A comparison between traffic measurements (TM) during 28 days and web questionnaires (Q) concerning Internet behaviour at home. Y=Yes, N=No, D=Daily.

Household	e-mail		Banking		Newspapers	
	TM	Q	TM	Q	TM	Q
1	D	D	Y	Y	Y	D
2.1	D	D	Y	Y	Y	D
2.2	D	D	N	N	Y	D
3	Y	Y	Y	Y	Y	D
4	D	D	Y	Y	Y	D
5	Y	Y	Y	Y	Y	D

One test pilot reported a high online gaming activity in the web questionnaire, which was not recorded by the traffic measurements. A factor influencing the traffic measurements is the cut-off of 1 kbps for the data collection. The transfer rates are usually low for gaming applications and the traffic may thus not register in the statistics. However, the discrepancy between the measurements and the questionnaire is in this case so large that at least a great part of it is deemed to be due to other circumstances for example that he used a computer in some other place for gaming. It has been discussed whether questionnaire answers give a true picture of activities that may be illegal or not socially acceptable.

In this study, the majority of the households answered that they used file sharing applications occasionally, which is in agreement with the measurement results. We thus conclude that the households are comfortable with answering questions about file sharing activities and that eventual errors are not a result of unwillingness to divulge this information. It should be noted here that the test pilots are used to answer questionnaires and may not be representative in the sense that they may be less shy than other users when sharing sensitive information.

Considering streaming media, the estimations of the test pilots were seen to be reliable when answering questions regarding specific streaming media applications, such as Spotify, or specific web sites with streaming media content, such as YouTube. But more general questions concerning how many hours a week they listened to streaming music or watched streaming video, were more problematic for the test pilots to answer and the answers were in some cases incomplete. The result is shown in Table 6. The reason for this difficulty is probably that the question was difficult to understand. The test pilots may not be aware of what using streaming media means although they are familiar with Spotify and Youtube.

Table 6: User penetration of specific applications seen in the traffic measurementsSpecific use of the Internet. A comparison between the results from traffic measurements (TM) during 28 days and a questionnaire (Q) given to the members of 5 households about their Internet behavior at home. The category Streaming audio includes applications such as Spotify, SHOUTcast, Flash audio over HTTP, etc. while streaming video includes for example Flash video over HTTP, HTTP media stream and Joost. The numbers represent the number of days where the measurement equipment registered activity. D=Daily, W=Weekly.

Household	Streaming Audio		Streaming Video	
	TM	Q	TM	Q
1	3	W	19	—
2.1	1	Seldom	0	0
2.2	15	D	21	D
3	15	—	6	W
4	18	0	25	W
5	27	—	28	D

We have up to now looked at frequency of use. Now we will compare Internet usage time, in hours and minutes. The question asked in the web questionnaire was: How many hours and minutes per week do you usually use Internet at home? As can be seen in Table 7, the answers correspond well with measurements of http traffic. But the traffic measurement method measures all Internet traffic such as file sharing, streaming media, IP telephony, automatic updates, etc as well as http traffic. As long as the computer is connected to the Internet there will be traffic going to and from the computer. This total Internet time is for every household more than 100 percent longer than their own estimations of their Internet activities. It seems that the Internet users are not able to estimate this total Internet time, at least not without much more detailed and comprehensible instructions.

Table 7: Time online. A comparison between the results from traffic measurements (TM) and a questionnaire (Q) given to the members of 5 households about their Internet behavior at home.

Household	Hours/week HTTP		Hours/week Total	
	TM	Q	TM	Q
1	16	14	30	—
2.1	9	12	—	—
2.2	50	56	120	132
3	10	8	69	—
4	14	10	23	—
5	49	48	122	—

C. Traffic Measurements (Long Term Behavior) vs. Diaries(Short Term Behavior)

All test persons were asked to fill in a diary and to make notes about what they were doing during two days (48 hours) chosen to be a Sunday and a Monday, 17 and 18 May 2009. During the same time, all Internet traffic going in and out from their IP-number was measured and the web sites visited were logged. On a general level, there was a total correspondence between the diaries and the traffic measurements as to when the test persons were using the Internet, which is shown in Table 8.

Table 8: A comparison between the results from traffic measurements and a two days diary given to the members of 5 households about their Internet behavior at home. D=Daily, day 1=day 1 of the recorded diary.

Household	Frequency		HTTP	
	TM	Q	TM	Q
1	D	D	D	D
2.1	day 1	day 1	day 1	day 1
2.2	D	D	D	D
3	D	D	D	D
4	D	D	D	D
5	D	D	D	D

The activities noted down in the diaries were compared to the traffic measurements and we can see that the same activities were registered in both, see Table 9. Although the diaries generally correspond well with the measured Internet activity, there are also some exceptions. For household 5 the measurements show Internet activity that is not noted down in the diary. The traffic consists of HTTP, HTTP media stream and HTTP download traffic. Also for household 4 one household member has at several times noted down that Spotify (music streaming application) was used in conjunction with either at work or outside. The measurements however show no traffic from this application. Either the usage of this application has taken place away from home or on a device that does not send traffic via the IP monitored by the traffic measurements.

Another noteworthy discrepancy is a number of visits by one of the households to websites containing adult content,

Table 9: Specific Internet activities. A comparison between the results from traffic measurements and a two days diary given to the members of 5 households about their Internet behavior at home. The numbers represent the number of hours during the two days where activity was registered by the measurement equipment. Y=Yes, D=Daily.

Household	E-mail		Banking		Newspapers		Social Network		IM		Spotify [Hours]	
	TM	Diary	TM	Diary	TM	Diary	TM	Diary	TM	Diary	TM	Diary
1	Y	Y			Y	Y	Y	Y				
2.1	Y	Y	Y	Y								
2.2							D	D	D	D		16
3	Y	Y	Y	Y			Y	Y			15	15
4	Y	Y	Y	Y	Y	Y						
5	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	

which was registered by the measurement instrument. Compared with what is detailed in the diaries of the household members it is seen that HTTP activity has been written down without any indication of the specific web sites visited. In the case of these types of web sites we would expect both diary and questionnaire material to show less activity or to be less specific than what is actually the case.

One of the households also shows activity from the filesharing application Bit Torrent KRPC which is a protocol keeping track of seeders and content for the Bit Torrent application. In the diaries, no such activity has been noted down. The conclusion here is that there is a computer in the household causing the activity without the presence of an active user. This type of activity that is not user induced is in fact difficult to measure using questionnaires or diaries and adds important details to the results concerning user behavior.

A more general observation is that the Internet usage registered by the traffic measurements seems to be, in several cases, more versatile and complex than what can be found in the diaries. We exemplify this by looking at some of the teenagers participating in the study. They use the Internet in several ways at the same time. They watch TV online while at the same time communicating with friends and visiting a social network site. This combination of activities is not seen in the diary. In the diary they may have noted down only one activity like watching TV online or just surfing. There is a need to define more carefully what is meant by certain concepts and to be aware that several Internet activities can go on at the same time as the main activity.

D. Diaries (Short Term Behavior) vs. Web Questionnaires (Long Term Behavior)

The general usage pattern that can be extracted from the questionnaire also shows up in the diary. The advanced enthusiasts and the enthusiastic modernists are very active according to the diaries, and traditionalists with very restricted use of the Internet are the same both in the questionnaire and the diary, see Table 10. Most of the activities the test persons say they perform daily in the questionnaire can also be found in the diary.

Table 10: Time online. A comparison between the results of a web questionnaire and a two days diary given to the members of 5 households about their Internet behavior at home.

Household	Hours/day HTTP	
	Q	Diary
1	2.0	1.25
2.1	1.7	0.75
2.2	3.0	3.25
3	1.2	2.0
4	1.4	1.1
5	7.0	5.5

There are exceptions such as when the test pilot is away from home during one of the two days when the diary was scheduled, or is busy with something completely different from what he/she usually does. But in most cases there is a high correspondence between the daily activities marked in the questionnaire and what is noted in the diary, even if two days are not enough for the whole repertoire of activities to show up. This is especially evident among the frequent users who use the Internet several times a day, as an important part of their everyday life.

VII. Discussion

Three different methods have been used to measure the way people use Internet in their homes: A questionnaire, a diary and traffic measurements. There is a high agreement between the methods. People who are active users of the Internet according to the traffic measurements are also active users according the questionnaire and the diary. Those who say in the questionnaire that they use the Internet rarely also show low Internet activity in the diary and in the traffic measurements.

There is also an agreement on a more specific level. The different activities that the respondents say that they usually do when going online can be found in the technical measurements of the Internet activities, and these activities also show up in the diaries. In most cases there is also a rather good fit between the time of Internet use measured from the traffic measurement and the users own estimate of the hours and minutes online. We found one exception of an underestimation. We also found a boy of 15 years who reported a lot of gaming that was not found in the traffic measurements.

This high level of agreement applies to the use of http websites. There seems to be more problems when it comes to streaming audio and video applications. The reason for this can be that the more passive use of streaming media as a second hand activity is more difficult to estimate, but also that the questions were not formulated in a comprehensive way. Not everyone knows what streaming audio is. Supporting this conclusion is the very good fit between the traffic measurements and the estimates of those Internet users who listen to the music service Spotify. When the question is specific and tied to a single activity it is also easier to give a good estimate. However, the questionnaire and especially the diary do not give the full picture of the Internet usage. In many cases, a lot of different activities go on at the same time. This is typical of the Internet usage of teenagers and younger people. They visit social websites, communicate with instant messaging, at the same time as they listen to music and watch TV. All these activities do not show up in the diary. Behind the term surfing a lot of activities can be hidden. Only traffic measurements can uncover this more complicated interplay between different activities although the diary form can be further elaborated to cover more complex situations.

At last, there is a more general problem. A direct comparison between traffic measurements and answers from a questionnaire and a diary can be problematic, as they do not measure exactly the same things. There is lot of network traffic that is not directly induced by the user, such as automatic updates and file sharing activities that happen in the background. For the most part, this is something that the users are unaware of. It is therefore necessary to develop a way to filter out those activities and further develop the traffic measurements, before a fair comparison can be made.

VIII. Conclusion and Future Work

In this paper, we have proposed and applied a unified methodology using three different methods to study Internet user behavior; traffic measurements, web questionnaires and diaries with the purpose of verifying and comparing the different methods as well as gaining more insight into user behavior.

From the measurements, we conclude that the test pilots are well able to describe some of their short term behavior seen in the diaries, although some activities were not noted in the diaries. The long term behavior seen in the web questionnaires are fairly accurate in describing frequencies of use specific applications and visits to specific web sites. The estimation of the amount of time spent on different activities was seen to differ from that of the traffic measurements, with a slight tendency to underestimate the time spent. An even more powerful conclusion is the complex and rich picture of user behavior, which is obtained via traffic measurements. Here, details and behaviors that are not exposed in diaries or questionnaires are visible. This gives new insights into user behavior as well as valuable feedback for better construction of question based investigations in the future.

The study was conducted in 2009 and since then, the amount of time people spend on the Internet has increased, especially when it comes to the use of smart phones. This will make it more difficult in the future to estimate the total time of Internet use, as the number of ways you can connect to the Internet increases: computer, mobile phone, TV, game console, etc. It will be more difficult but not impossible.

Another major result is the importance of the testbed to the study. Here, we have the possibility of making measurements in a controlled environment with real end users. From the traffic measurements we gain much more insight into the behavior of the end users than can be obtained from only questionnaires or diaries. However, from the questionnaires and diaries, we also gain a better understanding of how the end user perceives specific services giving valuable information when interpreting the results from the traffic measurements. From the network side, this can be used to improve the quality of service both from the technical and the end user perspective.

We will continue to develop our testbed under the Central Baltic Testbed project. In this project, we will work towards creating a distributed measurement platform. In this way we will have data from a large population spread out across the country. Our future work will continue with a wider study to follow up on the results presented here.

Acknowledgment

The work in this project has been partly financed by the VINNOVA project Broadband Behavior and the European Regional Development Fund project Acreo National Testbed. Also, it has been partly funded by the Celtic IPNQSIS project and Swedish National Strategic Research Area (SRA) within the program TNG (The Next Generation) and the project eWIN.

References

[1] C. Lagerstedt, A. Aurelius, H. Pathirana, C. P. Larsen, and O. Findahl, "Unified methodology for broadband behavior measurements in the acreo national testbed," *The Seventh International Conference on Digital Telecommunications, ICDT*, 2012.

[2] M. Chesire, A. Wolman, G. M. Voelker, and H. M. Levy, "Measurement and analysis of a streaming-media workload," *Proceedings of the 3rd conference on USENIX Symposium on Internet Technologies and Systems*, vol. 3, 2001.

[3] M. Alvarez-Campana, A. Azcorra, J. Berrocal, D. Larrabeiti, J. I. Moreno, and J. R. Prez, "Castba: Internet traffic measurements over the spanish r&d atm network," *5th HP Openview University Association Workshop*, 1998.

[4] Y. Bhole and A. Popescu, "Measurement and analysis of http traffic," *Journal of Network and Systems Management*, vol. 13, no. 4, pp. 357–371, 2005.

[5] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot, "Packet-level traffic measurements from the sprint ip backbone," *IEEE Network*, vol. 17, no. 6, pp. 6–16, 2003.

[6] J. Zhang, J. Yang, C. An, and J. Wang, "Traffic measurement and analysis of tunet," *2005 International Conference on Cyberworlds (CW'05)*, 2005.

[7] K. Fukuda, K. Cho, and H. Esaki, "The impact of residential broadband traffic on japanese isp backbones," *SIGCOMM Comput. Commun. Rev., ACM*, vol. 35, p. 1522, 2005.

[8] G. Maier, A. Feldmann, V. Paxson, and M. Allman, "On dominant characteristics of residential broadband internet traffic," *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, ACM*, p. 90102, 2009.

[9] (2012, 12) Akamai. [Online]. Available: http://www.akamai.com/stateoftheinternet/

[10] (2012, 12) Sandvine. [Online]. Available: http://www.sandvine.com/downloads/documents/Phenomena_1H_2012/Sandvine_Global_Internet_Phenomena_Report_1H_2012.pdf

[11] (2012, 12) Cisco. [Online]. Available: http://www.cisco.com/en/US/netsol/ns827/networking_solutions_sub_solution.html\#\~overview

[12] M. Kihl, P. Odling, C. Lagerstedt, and A. Aurelius, "Traffic analysis and characterization of internet user behavior," *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2010.

[13] A. Aurelius, C. Lagerstedt, I. Sedano, S. Molnar, M. Kihl, and F. Mata, "Tramms: Monitoring the evolution of residential broadband internet traffic," *Future Network & Mobile Summit 2010 Conference Proceedings*, 2010.

[14] J. George, "Researching life in e-society with diary studies," in *Proceedings of the 2006 European Conference on Information Systems*, no. 67, June 2006.

[15] K. Vermaas and L. van de Wijngaert, "Measuring internet behavior: Total time diary and activity diary as research methods," *Journal of Information Technology Theory and Application*, vol. 7, no. 11, 2005.

[16] O. Findahl, "Vad säger internetstatistiken? (what does internet statistics tell us)," Stiftelsen för internetinfrastruktur .SE, Tech. Rep., 2008.

[17] (2012, 12) Nordiskt informationscenter fr medie- och kommunikationsforskning, nordicom. [Online]. Available: http://www.nordicom.gu.se

[18] (2012, 12) Nielsen. [Online]. Available: http://se.nielsen.com/site/index.shtml

[19] (2012, 12) Post och telestyrelsen, pts. [Online]. Available: http://www.pts.se/sv/

[20] (2012, 12) Gallup. [Online]. Available: http://www.gallup.com/home.aspx

[21] (2012, 12) Mediamatning i skandinavien, mms. [Online]. Available: http://www.mms.se/

[22] (2012, 12) Statistiska centralbyrn, scb. [Online]. Available: http://www.scb.se

[23] N. Bolger, A. Davis, and E. Rafaeli, "Diary methods: capturing life as it is lived," *Annual Review of Psychology*, vol. 54, p. 579616, 2003.

[24] J. P. Robinson, "Introduction to issue 2: It, mass media and other daily activity," *IT & Society*, vol. 1, no. 2, pp. 1–8, 2002.

[25] K. Ishii, "Internet use in japan a time diary method," *Paper presented at the APIRA conference in HongKong and Macao*, 8 2004.

[26] S. Sieber, "The integration of field work and survey methods," *American Journal of Sociology*, vol. 6, no. 78, pp. 1335–1359, 1973.

[27] H. Selg and O. Findahl, "Nya anvandarmonster. jmforande analys av tva anvandarstudier." Uppsala universitet. Nationellt IT-anvndarcentrum NITA., Tech. Rep., 10 2008.

[28] T. Crosbie, "Using activity diaries: Some methodological lessons," *Journal of Research Practice*, vol. 2, no. D1, 2006.

[29] C. P. Larsen, C. Lindqvist, H. Pathirana, R. Lindstrm, E. Modin, and A. Aurelius, "Ant: The acreo national testbed status and plans," in *Proc. NOC 2007*, 6 2007.

[30] (2012, 12) Acreo national testbed, ant. [Online]. Available: http://www.acreo.se/en/Technology-Areas/Broadband-Technology/Projects/Current-Projects/Acreo-National-Testbed

[31] O. Findahl, "The swedes and the internet 2007," World Internet Institute, Gävle, Tech. Rep., 2007.

[32] ——, "The swedes and the internet 2011," World Internet Institute, Gävle, Tech. Rep., 2011.

[33] W. Cleveland and D. X. Sun, "Internet traffic data," *Journal of the American Statistical Association*, pp. 79–985, 1995.

[34] (2012, 12) Procera networks. [Online]. Available: http://proceranetworks.com

[35] O. Findahl, "The swedes and the internet 2009," World Internet Institute, Gävle, Tech. Rep., 2009.

[36] M. Kihl, C. Lagerstedt, A. Aurelius, and P. Odling, "Traffic analysis and characterization of internet user behavior," in *Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems*, 10 2010, pp. 224–231.