

Date of acceptance

Grade

Instructor

## **Data Gathering in Digital Homes**

Debarshi Ray

Helsinki September 11, 2012

UNIVERSITY OF HELSINKI

Department of Computer Science

Tiedekunta — Fakultet — Faculty		Laitos — Institution — Department	
Faculty of Science		Department of Computer Science	
Tekijä — Författare — Author			
Debarshi Ray			
Työn nimi — Arbetets titel — Title			
Data Gathering in Digital Homes			
Oppiaine — Läroämne — Subject			
Computer Science			
Työn laji — Arbetets art — Level		Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
		September 11, 2012	55 pages + 14 appendices
Tiivistelmä — Referat — Abstract			
<p>Pervasive longitudinal studies in people's intimate surroundings involve gathering data about how people behave in their various places of presence. It is hard to be fully pervasive as it has traditionally required sophisticated instrumentation that may be difficult to acquire and prohibitively expensive. Moreover, setting up such an experiment is laborious.</p> <p>We present a system, in the form of its requirements, design and implementation, that is primarily aimed at collecting data from people's homes. It aims to be as pervasive as possible, and can collect data about a family in the form of audio and video feed from microphones and cameras, network logs and home appliance (eg., TV) usage patterns. The data is then transported over the Internet to a server placed in the close proximity of the researcher, while protecting it from unauthorised access. Instead of instrumenting the test subjects' existing devices, we build our own integrated appliance which is to be placed inside their houses, and has all the necessary features for data collection and transportation. We build the system using cheap off-the-shelf commodity hardware and free and open source software, and evaluate different hardware and software configurations to see how well they can be integrated and how performant or reliable they are in real life scenarios.</p> <p>Finally, we demonstrate a few simple techniques that can be used to analyze the data to gain some insights into the behaviour of the participants.</p>			
Avainsanat — Nyckelord — Keywords			
layout, summary, list of references			
Säilytyspaikka — Förvaringsställe — Where deposited			
Muita tietoja — Övriga uppgifter — Additional information			

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Methodology . . . . .	4
1.2	Longitudinal Behavioural Studies . . . . .	5
1.3	Digital Homes . . . . .	6
1.4	Prior and Related Work . . . . .	7
<b>2</b>	<b>Requirements</b>	<b>9</b>
2.1	Observational Goals . . . . .	9
2.1.1	Social Interactions . . . . .	9
2.1.2	Personal Habits and Tastes . . . . .	11
2.1.3	Environment . . . . .	12
2.2	Functional Requirements . . . . .	12
2.2.1	Instrumented Home Appliances . . . . .	14
2.2.2	Privacy . . . . .	15
<b>3</b>	<b>Design and Implementation</b>	<b>17</b>
3.1	Overview . . . . .	17
3.2	Key Logger . . . . .	18
3.3	Smartphone Logger . . . . .	19
3.4	Behavioural Observations with BOB . . . . .	19
3.4.1	General Hardware . . . . .	19
3.4.2	Operating System . . . . .	20
3.4.3	Security . . . . .	21
3.4.4	Television and DVD Logger . . . . .	23
3.4.5	Network Logger . . . . .	25
3.4.6	Cameras and Microphones . . . . .	27
3.4.7	Data Uploads and Remote Management . . . . .	28
3.4.8	Installer . . . . .	30

3.5	Alice File Server . . . . .	32
<b>4</b>	<b>Data Analysis and Reliability</b>	<b>35</b>
4.1	Overview . . . . .	35
4.2	Taxonomy for Websites . . . . .	35
4.3	Household Profile . . . . .	40
4.3.1	Distance Calculation and Visualization . . . . .	41
4.3.2	Inferences . . . . .	41
4.4	Reliability of Data Gathering . . . . .	45
<b>5</b>	<b>Conclusions</b>	<b>48</b>
	<b>References</b>	<b>50</b>

## Appendices

### 1 Reference Hardware for BOB

### 2 Categorization of Websites

### 3 Daily Traffic

### 4 Weekly Traffic

## List of Figures

1	A schematic of Alice and BOB. . . . .	17
2	Encrypting the data before storing it on disk. . . . .	22
3	Work flow of the set-top box. . . . .	24
4	Work flow of the television and DVD logger. . . . .	24
5	Work flow of the NAT box, DHCP server and caching DNS server. . .	26
6	Work flow of the network logger. . . . .	26
7	Recording video and audio from cameras and microphones. . . . .	27
8	The positioning of the cameras in one of the apartments. . . . .	28
9	Reverse SSH tunnel between BOB and Alice. . . . .	29
10	Steps to install the BOB software. . . . .	31
11	Data being sent to Alice from different sources. . . . .	33
12	Number of requests to the most popular websites across all test subjects. .	36
13	Similarity of the households based on each component. . . . .	42
14	Overall similarity of the households. . . . .	43
15	Test subject 1: daily traffic in bytes. . . . .	43
16	Test subject 4: distribution of requests made to different categories of websites. . . . .	44
17	Distribution of requests made to different categories of websites. . . .	1
18	Daily traffic in bytes. . . . .	1
19	Weekly traffic in bytes. . . . .	1

## List of Tables

1	Average size of data generated per day by different sensors. . . . .	33
2	Reliability of Sensors . . . . .	46

# 1 Introduction

Information is a very potent tool for intimidating people. If you have the necessary insights about your neighbour's shady past, then it is easier for you to twist their arms into submission. That is why emperors and dictators have their squads of spies and secret police to monitor the activities of the citizens through covert means so that they can retain their control over the people.

In recent times we have seen the dawn of a new information age, where the delivery of information is shifting from newspapers, magazines and books to their digital counterparts. Similar trends can be seen in the way people are communicating today. Citizens are more likely to use electronic mail (email) and short messaging services (SMS) than postal services.

This has affected our lives in different ways. Communication has become cheaper and faster. We can send an electronic mail or an instant message to anybody who is on the Internet, without having to take into consideration when the postal department will be dispatching or delivering mail, or whether some political turmoil or natural calamity will adversely affect the process. With cameras and microphones becoming smaller and cheaper, it has not only become easier to save special moments of our lives for posterity, but enabled crowd-sourcing to democratize reporting and empower the masses [Llo11, Den08]. While things look promising, these are still early days. It remains to be seen whether the overall effect of these developments will be positive or negative. For the moment let us concentrate on the possible drawbacks.

Ubiquitous surveillance of individuals within their intimate surroundings is a serious problem these days [Ara95, PD03]. Online and digital content are easier to monitor. To furtively scan an individual's post, one has to very carefully open the envelope, go through its contents and then seal it again. The whole act should not leave any tell-tale artifacts to alarm the subject. It is hard to do this on a very big scale because it is time and manpower intensive. On the other hand, email can be monitored by instrumenting a network router to capture all the packets originating from the subject's house. This is much more scalable because one can cover the entire population by coercing a handful of Internet service providers (ISPs). It will be difficult for an individual eavesdropper, but governments do have the necessary means for carrying out such a thing.

Moreover, the success of online advertising based revenue models have led to the

proliferation of email, social networking, blogging, microblogging and various other content based services where the user does not have to pay anything to avail themselves of the facilities on offer. This is especially attractive to those who are new to this information age and do not have the knowledge or resources to maintain their own infrastructure. The catch is that the user's personal data are stored on servers owned and operated by the companies offering the services. This way they are able to carry out extensive analyses to create personality profiles of their users. This enables them to figure out usage patterns, behavioural characteristics and personal preferences, which they can leverage to carry out targetted advertising which is their actual source of revenue. The result of the analyses or access to the user's raw digital footprint can be sold to other companies looking for a way to improve the penetration of their products in the market. Prospective beneficiaries include mobile operators, ISPs, banks, insurance companies and departmental stores. Quite often sensitive information like the users' bank or credit card information can be found in their mail boxes, which can be misused by unscrupulous employees of these companies. Not only that, a service provider can simply decide to terminate a user's account, isolating her from her social circle and making it very hard for her to regain access to her own data. This places the user at the mercy of these business houses.

Such asymmetric relationships between the service providers and their subscribers are an inherent back door for rogue elements to wreck havoc on certain sections of society. In 2010, David Barksdale, an engineer at Google was found guilty of breaching the privacy of teenagers using services offered by Google [Hou10]. Earlier, in 2007, two Chinese journalists, Shi Tao and Wang Xiaoning, were jailed after Yahoo supplied the Chinese government with their private information [Mil07].

It is interesting to examine these asymmetric relationships with respect to privacy by carrying out a thorough study to find out to what extent governments and global multi-nationals can infiltrate into our lives, and what aspects of our personality can be deduced from our digital traces. For example, can a totalitarian regime deduce our sexual orientation or figure out who their political adversaries are? Or what can our ISP know about our lives by monitoring our Internet usage? Having a good understanding of how much of our privacy and personal space is under attack will help us in building better defences against possible intrusions.

Carrying out such an experiment would involve setting up a pervasive longitudinal study to monitor the behaviour of a group of volunteers as closely as possible. Based on the collected data their personalities can be reconstructed to understand the

extent to which our privacy can be violated. The better we are able to reconstruct, the higher is the threat.

However, it is hard to be set up such an experiment within people's intimate surroundings, and the success of the experiment depends squarely on the quality of the data that we are able to collect. If the data is not pervasive enough, the reconstructed personalities would be of low fidelity, which would give us a false sense of security. We address this problem by presenting our own data collection and transportation platform that can be used to collect data for such an experiment.

## 1.1 Methodology

In this thesis we work out the requirements for an instrumentation platform that will help us collect vast amounts of data, which only governments and large corporations are capable of doing currently, around a group of volunteer test subjects. We design and implement such a system to collect as much data that can be possibly gathered about the test subjects' behaviour in their various places of presence. We concentrate mainly on studying them within the intimacy of their homes, but also try to cover some of their behaviour and activities when they are on the move or elsewhere. We expect that other research groups who are interested in carrying out similar experiments will be able to use this platform for gathering their own data.

The experiment is run for a year during which the participants are closely, but unobtrusively, monitored. The data obtained in this manner can be analyzed to create household profiles for each family. These profiles can be compared with actual personalities and characteristics of the test subjects to find out the maximum extent and nature of the intrusions that can be carried out. We present a few example data analysis techniques to illustrate this. However, since the amount and diversity of the entire data is huge, we concentrate only on a small subset of the data in this thesis.

Gaps or interruptions in the recordings are undesirable. Therefore, the reliability of the platform is evaluated to see how performant it is in real life scenarios. We discuss the different avoidable and unavoidable scenarios that negatively affect the continuity of the recordings.

Even though the participants are made fully aware of the extent to which they are going to be monitored, steps are taken to protect their privacy. We anonymize all the results in our experiments and in this thesis. It is possible that some visitors



might not be aware that every move inside the apartment is being monitored, or object to being watched. Therefore, the family members are informed about the location of the sensors and they are allowed to temporarily turn them off without our permission if they want. Precautions are taken to prevent unauthorised access to the data in case someone breaks into the test subjects' houses or tries to hijack it while it is getting transported to the researcher.

In social psychology terms we are going to carry out a longitudinal behavioural study. It is hard to be fully pervasive as it has traditionally required sophisticated instrumentation that may be difficult to acquire and prohibitively expensive. However, homes are becoming smarter these days, and we will try to leverage recent advancements in technology to come up with a system that is cheap and easy to assemble, so that social psychology researchers doing similar studies can use it for their own purposes.

## 1.2 Longitudinal Behavioural Studies

Longitudinal behavioural studies is a type of research method used to discover relationships between different psychological variables that are unrelated to background variables. It involves studying the same group of individuals over an extended period of time and can even span multiple decades. The subjects are observed in their natural habitat without any manipulation by the observers, and therefore, great care needs to be taken to avoid interfering with the individuals or disturbing the surrounding environment while still gathering as much data as possible.

This is different from more direct methods like interviews and questionnaires, where people are asked about themselves and their responses noted. There, as Allan Kellehear [Kel93] had put it, the underlying assumption is that important truths about people are best gained through a direct or subtle interrogation of experience attitude and belief. However such direct methods are reactive in nature and tend to intrude as a foreign element into the social setting that is under observation [Lee00]. For example, when subjects are individually interviewed, they have a tendency to project an image of themselves which might not tally with reality in order to maintain their standing in the eyes of the interviewer. Responses might even be influenced by the order and wording of the questions and the rapport between the interviewer and the subjects. For example, if a person has claimed to be interested in politics in response to an earlier question, then she might answer in the affirmative when asked if she had voted even if she had not to hide her self-contradictory behaviour. Such

effects are also seen in other forms of non-naturalistic observation methods as well. The Hawthorne Effect [May45, RD66] is one famous example.

The term 'unobtrusive measures' was coined by Webb [WCSS66] to refer to data gathered by means that do not involve direct elicitation of information from research subjects. Being non-reactive they are presumed to avoid the problems caused by the researcher's presence. The unobtrusive measures suggested by Webb fall under the broad headings of physical traces, archival records and non-participant observation in varying situations. For example, the degree of wear on the floor tiles surrounding a museum exhibit to measure visitor flows, or the size of suits of armour as an indicator of changes in human stature over time.

This kind of naturalistic methods makes our observations more credible because they are occurring in a real and typical scenario as opposed to an artificial simulation within a laboratory. It also facilitates the study of events that are considered unethical to study via other experimental methods. For example, the impact of high school shootings on students attending the high school.

### 1.3 Digital Homes

Recent advancements in technology have caused many personal items of daily use to become smarter. Today we have smarter phones, smarter watches, smarter audio players, smarter heaters, smarter air conditioners, smarter washing machines, smarter refrigerators and smarter television sets. Items that we previously used to think of as electrical appliances have graduated to being programmable computers, and it is this ability to program them that makes them smarter. We can exploit these developments to ameliorate some of the challenges in setting up a longitudinal behaviour study where people are pervasively monitored within their personal space without interfering with their lives.

Since the domestic appliances are programmable, we can instrument them by reprogramming them to keep tabs on their human masters. For example, smartphones can be used to keep track of the users' movements by logging co-ordinates from the Global Positioning System (GPS) unit or the names of the cell phone towers that are encountered. Similarly, a television set-top box can be instrumented to monitor television usage.

## 1.4 Prior and Related Work

The MIT House\_n team have built a highly instrumented apartment-scale research facility called the PlaceLab [ILB<sup>+</sup>05], where new technologies and design concepts can be tested and evaluated in the context of everyday living. The House\_n group is working towards a vision where ever-present computer technology empowers people with "just-in-time" information that helps them make better decisions without losing their sense of control over their environment due to excessive automation. The sensors are also being used to monitor activity in the environment so that researchers can carefully study how people react to new devices, systems, and architectural design strategies in the complex context of the home. The information generated by the PlaceLab is based on unobtrusively observing people within the intimacy of their homes. Therefore this is something that is very closely related to our data collection and transportation system.

However, the PlaceLab is a single-family home, completely built from scratch as a live-in laboratory. Thus it is hard or impossible to retrofit it into the existing houses of the test-subjects, which is what we want to do. We do not want to change the personal habitats of the subjects, but study them within the environment they are used to.

The MIT House\_n team later came up with a portable kit of wireless sensing devices for pervasive computing research in natural settings called the MIT Environmental Sensors (MITes) [TILL06]. It includes six different types of environment sensors, and five wearable sensors. Even though we share many of our design goals with MITes, the kind of data that we intend to collect about the test subjects is different, as we shall see in the next chapter. We have a different set of challenges, which requires a different solution.

With the recent improvements in technology, researchers have started using traces from smartphones [EPL09, ROE09] and behaviour on social networks as sources for their data. The advanced sensing abilities of smartphones allow for automatic gathering of different kinds of behavioural data that is of interest to researchers – location, proximity to other devices, both the data and metadata of mobile communications, users' commands and interaction with the device, calendar, and device state.

Nathan Eagle and Alex Pentland introduced a system for sensing complex social systems by logging the different Bluetooth devices in the proximity of a smartphone

carried by a test subject [ESP06]. The Bluetooth logs are meant to augment the more traditional idea of using cell tower ID to approximate the location of the user, in the absence of cell tower reception. Using this location information they built a predictive classifier that can perceive aspects of a person's life more accurately than a human observer, including the person herself.

The above work involving smartphones is interesting and we try to leverage it while building our instrumentation platform.

## 2 Requirements

The instrumentation platform should be unobtrusive, and it has two main functions – collecting data from different sensors at the test subjects’ homes, and sending it to the researcher. It should ensure that interruptions in the recording are kept to a minimum, and that data gets sent to the researcher at a rate fast enough to keep up with the recordings.

To work out the different types of sensors that are required, we need to decide which aspects of the test subjects’ behaviour we want to observe. We come up with a list of attributes, knowing which we can accurately reconstruct an individual so that we can create personality profiles for them.

### 2.1 Observational Goals

In contemporary personality psychology, the “Big Five” factors or the Five Factor Model [Dig90, Gol93] is used to describe human personality. The Big Five factors are openness, conscientiousness, extraversion, agreeableness and neuroticism. Each of the five factors indicate the presence or absence of certain traits in the test subject’s personality, and is considered to be a comprehensive, empirical and data-driven research finding.

Our objective is to have sufficient data to score the test subjects on the above factors.

#### 2.1.1 Social Interactions

Human beings are social creatures not just because we like company and depend on others, but in a more elemental way where the basic existence as a normal human being requires interacting with other people [GAW11]. Hence, we should start with directing our attention to the social interactions of the test subjects.

Social interactions can be real world interactions, online interactions or a mix of the two. In the real world it consists of friends and relatives with whom the test subjects interact. These interactions can take place either inside the test subjects’ houses, or elsewhere in pubs, offices or shopping malls. Online interactions involve use of forums, chat rooms, email and social networks – basically anything that involves communicating with other individuals. An online interaction can be brought about by a previous real world interaction or the other way round. For example, people

can fix a meeting using email, or agree to keep in touch over email during a meeting. We want to observe the test subjects' friends and social circles very closely. We want to know big they are, whether they meet frequently and go out often, and what sort of roles the test subjects play within their respective circles. If a person has a large number of friends, is the cynosure of all eyes in big gatherings and parties, or is otherwise enthusiastic, talkative, assertive and gregarious by nature, then she has a high extraversion score. On the other hand, if a person is reserved, low-key, less socially involved and tends to focus on a small of number of close friends then she is an introvert [GAL98].

The ideas and issues espoused by the test subjects, their eagerness to help and methods for resolving conflicts should be noted too. If they have divergent ideas which challenge established norms instead of standing by tradition it tells us that they have high levels of openness [MI87]. If they are prejudiced against stigmatized groups then it indicates that they are less agreeable in nature [GBST07]. More agreeable people use constructive tactics, as opposed to coercive ones, to resolve conflicts [JCG01], and are likely to lend a helping hand even the cost of helping is high [GHST07].

We want to know the amount of diversity the test subjects have within their social circle. If a person interacts with foreigners or demonstrates proficiency in multiple languages then it is reasonable to expect that she is either well-read or has travelled broadly. It might also be so that she has a prominent online presence which exposes her to people from different cultures from far away places, and we can verify this possibility from her Internet profile. Preference for variety, an extensive vocabulary and intellectual curiosity are indicators of greater openness [CM92, Jos06].

People have a tendency to express themselves differently on the Internet than they would in the physical world [FSV07]. This can have different manifestations in different individuals, and the degree to which their virtual and real attitudes differ varies as well. The greater anonymity provided by the Internet as compared to the physical world, and the lack of formality of social conduct tends to reveal aspects of an individual's character that would not otherwise be observable [AHWF02]. Ofcourse, people do have lots of reasons to be concerned about their privacy on the Internet [MKA04, GOO11], but the fact that they are sitting behind a computer does provide them with a semblance of anonymity.

Therefore we want to keep track of the test subjects' online interactions over instant messaging networks, social networks, forums and electronic mail. Which means we

should observe the sort of language they use – whether they have a tendency to use profanities or the amount of SMS language versus conventional grammar, the content that they share – do they reveal their true identities, and the subject of their online discourse.

We want to know the entities to which they give out their personal data. It is a good way of estimating how their privacy might be compromised. If they are using a particular service for their email, then information that can be gathered from their mail box can be compromised if the service is attacked by external elements, or if the service provider itself chooses to read her mails and leak the information for profit or due to government coercion [GOO10].

### 2.1.2 Personal Habits and Tastes

The degree of openness of the test subjects can be measured from their intellectual profile [CM92]. People who display an appreciation for the arts and sciences, a readiness to re-examine traditional values, are sensitive to beauty, have a vivid imagination, and an inclination to explore and experiment are the ones with high openness levels.

Therefore we want to know what hobbies the test subjects pursue. If they like to watch movies and plays, it would indicate an appreciation for the arts. If they like to paint, it would indicate aesthetic sense and a vivid imagination. If they indulge in philately and numismatics, it would indicate their inquisitiveness.

We should cover what type of music or clothes the test subjects prefer. More upbeat, conventional and energetic music indicates extraversion [RG03]. Similarly extraverts wear more decorative clothing whereas introverts like practical and comfortable dresses [Sha80].

It is worth noting that the test subjects' hobbies can also shed light on their extraversion score. For example, those who prefer team games are more extraverted than the ones who play individual sports or do not play at all [EMdM07]. Similarly, if a person expresses a liking for reserved activities like tinkering, watching movies and plays, listening to music or fishing then she lies towards the introverted end of the scale.

The general mood of the test subjects should be observed to see how neurotic they are. Frequent bursts of anger or expressions of guilt, long periods of anxiety or depression are the symptoms of neuroticism.

We should observe the daily routine of the test subjects, their food habits, their movements about the house, and their posture while doing different things. Those who stick to a proper time-table involving regular sleeping hours and healthy food, are hard working and prompt in their movements, and maintain proper posture are the ones who are conscientious by nature.

We want to observe the test subjects' browsing habits too. An individual's online profile gives an indication of what sort of person she is. For example, a person who spends a significant amount of time on Internet Relay Chat (IRC) networks or uses Secure Shell (SSH) to remotely log into different computers can be categorized as a computer geek, while those who frequent mainstream news portals and video sharing sites fall into a different category who tend to use the Internet for consuming media.

### 2.1.3 Environment

Human behaviour is influenced by the environment [Bar68]. For example, a person can get irritated or depressed if there are repeated problems with the heating or power supply in her apartment. If the presence of environmental stress was unknown to us, she would have received an undeservingly high neurotic score. Moreover, we can analyse how test subjects react to such stress to calculate their relative neuroticism levels. Those on the higher end of the scale will be more reactive, while those on the lower end will be calmer and less rattled.

People who score high on conscientiousness tend to have tidier homes with better lighting. Similarly, the presence of unconventional decorations, paintings or antiques indicate openness [Gos08].

Therefore we want to observe the ambience within the house – lighting, noise, and the interior decor.

## 2.2 Functional Requirements

In order to meet the above observational goals, the instrumentation platform must meet the following requirements:

- Record the audio inside the house, especially the conversation among different family members.
- Record the video of what is going on inside the house.



- Log television usage of the family members.
- Log details about the Digital Video Discs (DVDs) that are being watched.
- Log Internet usage of the family members.
- Log personal computer usage of the family members.
- Log the presence of wireless access points that can be accessed from within the house.
- Log the presence of Bluetooth devices within the house.
- Record applications used by the test subjects on their smartphones.
- Record Global System for Mobile Communications (GSM) call metadata on the smartphones.
- Record short Message Service (SMS) metadata.
- Log location of the smartphone using GPS samples.
- Log periods of smartphone inactivity.
- Log presence of Bluetooth devices around the smartphones.
- Keep all the collected data secure from malicious attackers.
- Price of each unit should not exceed 500 Euro.

We can meet many of our observational goals if we can see and hear what goes on within the test subjects' houses. It will let us observe their friends circles, things they like to talk about, their opinions, how helpful they are, and their conflict management strategies. We can find out about their hobbies, their tastes in clothes and music, their general mood and their daily routine. It will show us the interior decoration of the apartments too. That is why we want our instrumentation platform to provide us with audio and video coverage of potential hotspots of human activity. For example, living rooms, dining tables, study tables, kitchens and hallways.

Knowing what the test subjects' watch on the television or which DVDs they prefer, we can infer whether they like to watch movies and drama, or the news. With the help of the audio and video data we can analyze the test subjects' reaction to some

important pieces of news – a natural calamity, a terror attack, national election results, or the outcome of an important sports match.

The combination of the network traffic logs and personal computer usage will give us a complete understanding of what the test subjects do on their computer. Using it we can keep track of their online interactions, the websites and services they use, and the entities to which they give out their personal data. It will tell us on which days of the week, or at which times of the day they are more likely to be using the Internet. By looking at the most frequently used applications we can infer whether they use the computer for writing documents, or for preparing presentations, or for listening to music. This is interesting because it is indicative of their routine and habits.

Keeping track of the Bluetooth devices within the house is a way to find out if someone has entered or left the house. Bluetooth enabled phones are getting increasingly popular so we can safely assume that most visitors will be carrying one.

While the other data sources give us an insight into what is going on inside the test subjects' homes, traces from smartphones will let us monitor their behaviour elsewhere.

### **2.2.1 Instrumented Home Appliances**

Looking at the previous section, some of the sensors will be instrumented versions of common home appliances and gadgets. Apart from the personal computers and smartphones, these appliances are special purpose embedded computers with limited or zero mobility requirements. Instead of retrofitting the appliances already owned by the test subjects we want our platform to provide the functionality offered by these gadgets. This is better because we can build and test the instrumentation platform in relative isolation from the environment that we want to place it in. Different households can have widely different varieties of the same appliance, and we would have soon ended up with a combinatorial explosion of setups that need to be supported.

To monitor the personal computers we use a key logger software that can be installed on the computers owned by the families, but for the smartphone traces we want to give them a new phone which will be instrumented to suit our needs. This is based on the fact that an overwhelming percentage of desktops and laptops run some version of the Microsoft Windows operating system [Hol09], which makes it easier

for us to develop a key logger that can be easily retrofitted on existing computers. On the other hand, the smartphone market is way too fragmented [MOB11] for us to safely guess which platforms the test subjects might be using. Hence it is safer for us to give them a pre-instrumented phone.

Therefore we should offer the following functionalities to the test subjects:

- A television set-top box with timeshift functionality.
- A DVD player.
- A wireless access point with enough signal strength to cover the whole apartment. Among all the IEEE standards for implementing wireless local area networks, only 802.11g and 802.11n [Theb] can keep up with the broadband speeds offered by the ISPs in Finland. Hence, they should be supported.
- A network router that can be connected to the Internet uplink provided by the ISP. It should allow wired network devices to join the local network within the house so that they can be interoperable with the wireless devices connected to the access point.
- A firewall to protect the test subjects' network devices from getting attacked by rogue elements in the Internet.
- A smartphone.

### 2.2.2 Privacy

The data collected by our instrumentation platform is going to be very sensitive in nature because it would contain detailed and intrusive logs of the participants' activities over a period of one year. It may contain passwords, information about bank accounts or credit cards, and their presence or absence from their houses during different days of the week. If these fall into the wrong hands it leaves them vulnerable to financial fraud or burglary. Therefore, even though the test subjects are informed of the extent to which they are going to be monitored, we must take steps to protect them from criminal conspiracy.

We must make sure that a malicious attacker who breaks into an apartment and steals the system can not access the collected data, nor should she be able to intercept the data while it gets transported to the researcher. We must also ensure that the

sensors can not be tampered with so that they can not be hijacked away from the system's control.

It is important to anonymize the personality profiles drawn from the data while publishing the results of our experiment. Our goal is not to subject our volunteers to widespread ridicule and scrutiny by exposing their behavioural patterns to the whole world. We merely want to measure the extent to which our personalities can be reconstructed from our digital traces. We do not loose anything by replacing identification information, like names and addresses, with dummy placeholders.

Visitors to the apartments may choose not to participate in the experiment, or there might be situations in which the test subjects feel uncomfortable being watched. Therefore, it should be easy for the participants to turn off the sensors for some time and restart them later.

### 3 Design and Implementation

In this chapter we present the design and implementation of our instrumentation platform as per the functional requirements stated in section 2.2. We start with a broad overview of the whole system and briefly introduce the main components. Once we have an idea of the bigger picture we look at each of the components in greater detail.

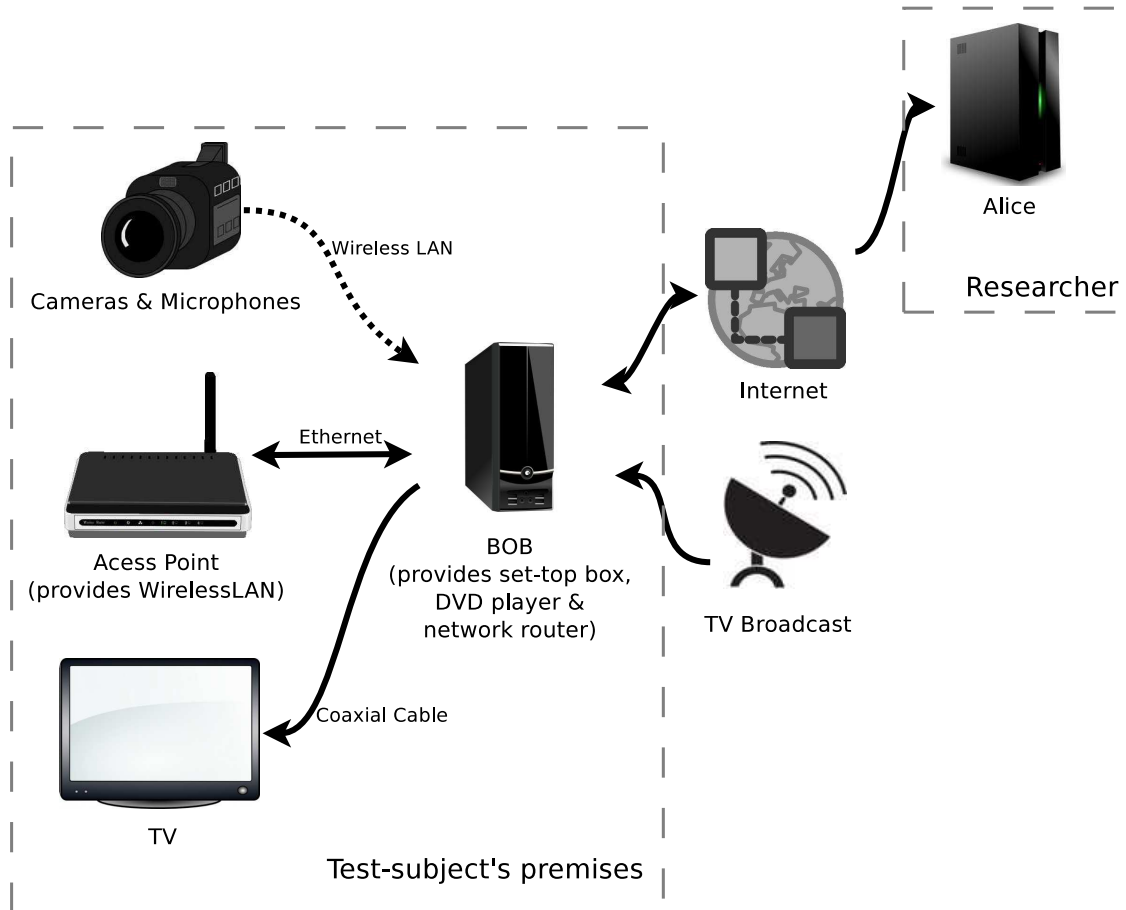


Figure 1: A schematic of Alice and BOB.

#### 3.1 Overview

The instrumentation platform has the following main components:

- A key logger that is installed on the test subjects' personal computers.

- One or more instrumented smartphones that are to be carried by the test subjects.
- An immobile unit called BOB, an acronym for Behavioural Observations, that is to be placed in the test subjects' houses. It co-ordinates all the sensors, except the above two, and caches their data temporarily.
- A file server called Alice at the researcher's end to which the data is sent over the Internet and stored.

It is important to have a cache for each house because we can not guarantee the performance or reliability of the wide area network (WAN) that connects the researcher with the test subjects. The WANs are operated by the Internet Service Providers (ISPs) and can be affected by conditions beyond our control. For example, traffic generated by other network nodes or adverse weather conditions.

## 3.2 Key Logger

Key loggers are installed on the test subjects' personal computers to monitor their computer usage as mentioned in section 2.2. We use PyKeylogger<sup>1</sup>, which has the following logging features:

- Logs all key strokes to a delimited data file.
- Takes a partial screenshot centred at the location of every mouse click.
- Takes a full screenshot at fixed time intervals if the computer is not idle.
- Records the foreground application identifier.

It is hard to know which key presses constitute a password. So we send the logs to Alice over encrypted email using SMTPS<sup>2</sup> every half an hour and delete them from the participants' computers. To prevent malicious changes in the configuration, we use PyKeylogger's feature to protect its control dialog with a password.

---

<sup>1</sup><http://pykeylogger.sourceforge.net/>

<sup>2</sup>It is a method for securing Simple Mail Transfer Protocol (SMTP) with transmission layer security.

### 3.3 Smartphone Logger

All subjects were provided with Symbian-based Nokia N8 smartphones. The phones are instrumented with software that collects all the different kinds of smartphone traces mentioned in section 2.2. We use ContextLogger2 (CL2) [Has10], written by Tero Hasu, to do this. It consists of a logger daemon and a set of supporting programs and libraries, and is configured to observe all the necessary sensors that we need. The data is compressed, and stored locally in SQLite3<sup>3</sup> format until it gets uploaded. It is automatically uploaded to Alice using the Hypertext Transfer Protocol's (HTTP) POST request method over a secure connection every day at 01:00 hr. Uploads are retried at increasing intervals in cases of failure or network inaccessibility. Once uploaded the logs are deleted from the smartphone.

### 3.4 Behavioural Observations with BOB

BOB is basically a headless personal computer with the sensors being either peripherals or software running on it. The computer is installed with software that is customized and configured to fulfill the functional requirements laid out in section 2.2.

#### 3.4.1 General Hardware

To make the system attractive as an instrumentation platform we ensure that it is not a source of constant disturbance or irritation for the test subject. We also minimize the cost and maximize the availability of the individual components to improve the viability of the system. Therefore we decided to use cheap off-the-shelf commodity hardware, pre-existing free software or our own custom software packages. Proprietary technologies come with restricted terms of use and a higher price tag and should be avoided.

We try to reduce the size of the cabinet and the amount of run-time noise. In terms of size, a microATX form factor is the best we can achieve with commodity hardware. The noise is caused by the various cooling fans and the hard disk drive. There is a fan to cool the CPU, one for the switched-mode power supply (SMPS) and another one for the graphics card. We can not get rid of the first two, so the best we can do is to keep their revolutions per minute (RPM) as low as possible.

---

<sup>3</sup><http://www.sqlite.org/>

Select models of modern motherboards make this possible, but we can make things better by trying to reduce the load on the CPU. The lesser the load, the lower the temperature, and we have a quieter fan. To do so, we offload portions of the set-top box's video decoding and post-processing to the GPU, which is a much better vector processor than the CPU [GPU11]. NVIDIA's Video Decode and Presentation API for UNIX (VDPAU)<sup>4</sup> makes it possible to do so. Therefore we use a graphics card which does not have a fan in its cooling unit and supports VDPAU. Regarding the hard disk, our only option is a solid-state drive but we can not use them because they are costlier compared to conventional rotating disk drives [SSD11].

Since the cameras and microphones are special-purpose embedded computers without any moving parts there is not much we have to do about them apart from making sure that they are not too big in size and do not suffer any hardware failure if run continuously for a few years.

### 3.4.2 Operating System

We select an operating system that is stable and meant for long term use. The software components should be well-tested so that they do not have too many bugs, and should continue to receive enough attention against security exploits during the duration of the experiment. Security is important for us, and with limited resources we can not hope to audit the entire code base ourselves. At the same time it should have good hardware support so that drivers for current hardware peripherals are easy to find.

A free (as in freedom) clone of UNIX stands out as the best option. Not only are some of them well-known for their long term stability, but they also have a flexible design, good documentation and community support. We have limited man power and a low budget. Hence the easier it is to tinker with the system and the lesser we have to worry about all the moving parts, the better it is. Documentation and flexibility, coupled with freely available source code and the permission to modify it as needed give us the leeway to modify and integrate the parts as we want to build the whole to fit our needs. A vibrant community reduces the amount of work that we need to do ourselves to build and maintain the system, which is not the case with proprietary alternatives without incurring substantially higher expenses and having to endure lengthy bureaucratic processes. Surely UNIX has had its share of detractors in terms of design, but it is good enough for us and is clearly the winner

---

<sup>4</sup><http://http.download.nvidia.com/XFree86/vdpau/doxygen/html/index.html>



when it comes to the above criteria.

With this mind, our choices are narrowed down to one of the many GNU/Linux distributions and the BSDs – FreeBSD, NetBSD and OpenBSD. A distribution is a collection of pre-packaged system and application software that has been configured and modified to work together as a coherent whole so that the user can do something useful with it. Otherwise one would have to assemble the parts starting from the kernel, C library, linker and loader, and then moving upwards through the init system and all the applications. We decided against the BSDs because the Linux kernel has a wider range of DVB card drivers which are an essential component of our integrated set-top box, and evaluated the Debian, Fedora and Ubuntu GNU/Linux distributions.

Debian is known for its stability and long-term focus, but the software is sometimes a bit old. For example, the absence of the latest drivers from the kernel was conspicuous. Fedora on the other hand is more updated but its thirteen month life cycle does not give us enough time to develop BOB based on a particular version of it and then deploy it. We would prefer something that was supported for atleast a couple of years.

The Long Term Support releases from the Ubuntu project turned out to have the right combination of the bleeding edge and a longer life expectancy. They are supported for a period of three years, which means they keep receiving security fixes and other important updates and new drivers are added as new devices arrive at the market place. Its Debian lineage ensures that some of its ancestor's virtues, like stability, are present in it, and at the same time, a fresh new Ubuntu LTS release is quite as modern as its Fedora counterpart. The presence of yaVDR <sup>5</sup>, a Ubuntu derivative meant for home theatre PCs (or HTPCs), was an additional bonus because it means less work for us while integrating the set-top box into BOB.

### 3.4.3 Security

We need to make sure that if a BOB gets stolen and a person with malicious intent gets physical access to BOB's hard disk, then she should not be able to access the cached data. For that we encrypt the data with the Threefish [THR08] block cipher before it hits the hard disk. Threefish is a symmetric encryption algorithm, so we use RSA [RSA78] to asymmetrically encrypt the symmetric key. Each data file has a

---

<sup>5</sup><http://www.yavdr.org/>

separately generated Threefish key, while there is a single RSA key pair belonging to the researcher. Since there is currently no successful cryptanalysis of Threefish, we regard this scheme to be safe.

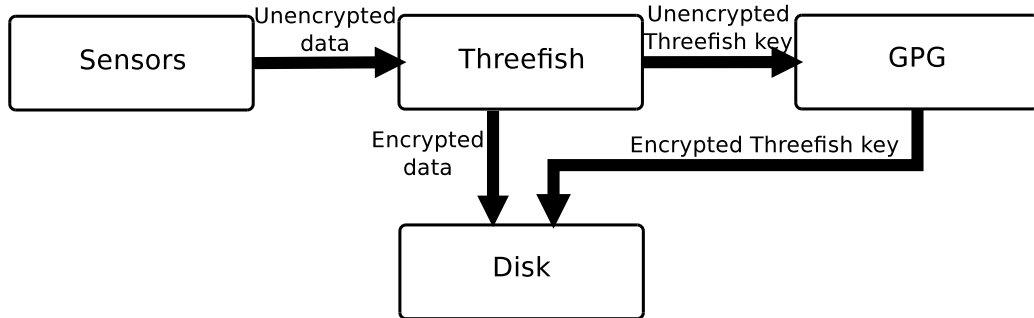


Figure 2: Encrypting the data before storing it on disk.

We use our own implementation of the Threefish algorithm, which is tested against the standard test vectors. For RSA we use GNU Privacy Guard (GPG)<sup>6</sup>, which contains a very widely used and well tested implementation of the algorithm.

To prevent unauthorized intruders from entering the wireless network offered by BOB’s access point, we look at the different options for securing a IEEE 802.11 network – Wired Equivalent Privacy (WEP), Wi-Fi Protected Access (WPA), and Wi-Fi Protected Access II (or WPA2). WEP has been known to be weak for more than ten years now [Wal00], and recently WPA has also been cracked [TB09]. The only viable option is to use WPA2 on our access point. Thus we can be sure that no one will be able to intercept the audio and video being streamed from the cameras and microphones to BOB. Nor will anyone be able to try a remote attack on BOB over the wireless network.

Finally, to ensure the safety of the data while it is being sent from BOB to Alice over the Internet, we use a secure channel offered by the Secure Shell (SSH) version 2 protocol <sup>7</sup>.

---

<sup>6</sup><http://www.gnupg.org/>

<sup>7</sup>SSH is a network protocol specified in RFC 4252 for secure data communication, remote shell services or command execution and other secure network services between two networked computers. The encryption used by SSH is intended to provide confidentiality and integrity of data over an unsecured network, such as the Internet

### 3.4.4 Television and DVD Logger

An integrated television set-top box and DVD player is built into BOB and instrumented to collect details about the TV and DVD usage of the family members. It supports timeshifting and can be controlled with infra-red remote control handsets. Most people have High-Definition Multimedia Interface (HDMI)<sup>8</sup> capable television sets these days [HDM, REU]. Therefore, our set-top box supports HDMI as the primary output for audio and video.

Older TVs might have Digital Visual Interface (DVI) [DVI] or Video Graphics Array (VGA) for video, so we support them too. For people with home theatre systems or pre-HDMI audio setups we provide Sony/Philips Digital Interconnect Format (S/PDIF) and stereo output. In Europe, legacy cathode ray tube (CRT) televisions use Syndicat des Constructeurs d'Appareils Radiorécepteurs et Téléviseurs (SCART) connectors. We offer an adapter to convert VGA and stereo to SCART.

We use Digital Video Broadcasting (DVB) cards for receiving the television broadcast signal. Depending upon whether a particular test subject uses satellite, cable or terrestrial television, we offer DVB-S, DVB-C or DVB-T cards. Ideally we would have liked to provide two cards, so that users can record and view shows from different channel groups simultaneously, but due to the lack of space on the motherboards we use only one.

In recent times Blu-ray Discs (BD) have grown in popularity [BD-a]. However, to support BD we would either need a separate hardware player or a proprietary software application due to legal restrictions on implementing free software players for this new kind of media [BD-b]. Both options are equally expensive, and having a separate hardware player makes it difficult for us to monitor the test subjects' use of Blu-ray. Moreover, the hardware players are big enough to have a negative impact on the footprint of the whole platform. Therefore we chose not to integrate this functionality into the system.

The Video Disk Recorder (VDR) software<sup>9</sup>, written by Klaus Schmidinger, drives the above hardware to provide a functional set-top box. VDR is also used by the yaVDR project, so we benefit from their bug-fixes and updates to the related software packages.

VDR allows the integration of the DVD playback functionality into the set-top box

---

<sup>8</sup><http://www.hdmi.org/>

<sup>9</sup><http://www.tvdr.de/>

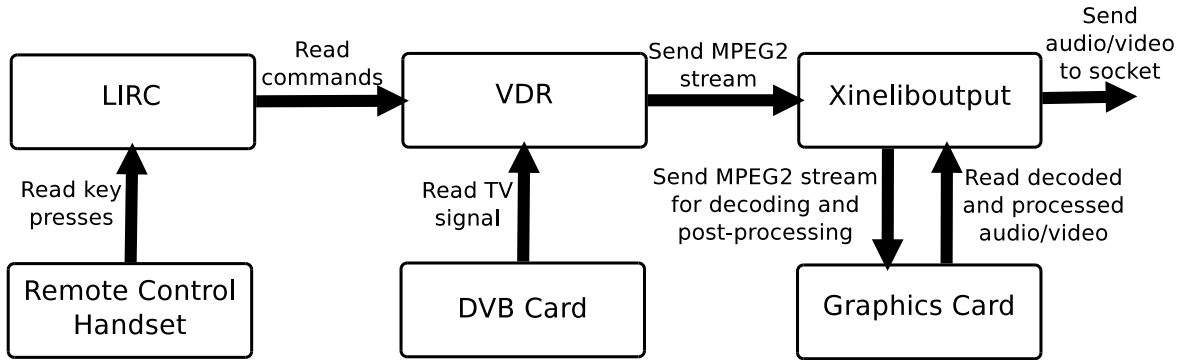


Figure 3: Work flow of the set-top box.

in a way that is very user-friendly. The xineliboutput plugin is used to leverage the graphics card for video decoding and post-processing through VDPAU, and we change the default on-screen display and subtitles to be in Finnish.

Linux Infrared Remote Control (LIRC) is used to decode and send the signal from the remote control handset to the VDR daemon. Since VDR and LIRC are separate processes communicating over a local socket we have the advantage of being able to reset the VDR daemon from the remote control by issuing a signal. For the test subject this is more convenient than having to physically go and hit the reset button on BOB if the set-top box gets stuck while watching television.

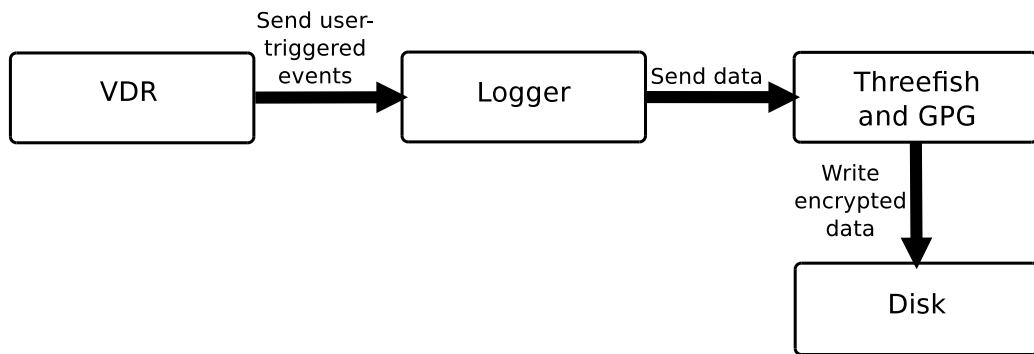


Figure 4: Work flow of the television and DVD logger.

The VDR daemon offers a plugin interface for writing third-party plugins. We use it to write our plugin to log the user-initiated activities within the set-top box into a text file.

### 3.4.5 Network Logger

BOB provides a wireless access point, a network router and an Internet gateway to the test subjects' so that their online behaviour can be monitored. Iptables<sup>10</sup> is used to set it up as a Network Address Translation (NAT) box so that it functions as the gateway and multiple network devices can access the Internet through a single global IP. We provide a Dynamic Host Configuration Protocol (DHCP) server and a caching Domain Name Service (DNS) server using dnsmasq<sup>11</sup>. It makes the home network more enjoyable to use because users would not have to manually configure their network devices, and their DNS queries can be served more quickly.

IEEE 802.11 is a set of standards for implementing wireless local area networks [Theb]. Of these only 802.11g and 802.11n can keep up with the broadband speeds offered by the ISPs in Finland. Therefore we offer them as the wireless network options on our access point.

We use Ethernet [Thea], which is the standard for wired local area network technologies, for connecting BOB and its access point to the Internet uplink, and to allow wired network devices to join the home network. The test subjects are provided with broadband connections with a maximum upstream bandwidth of 10 Mbps, and a maximum downstream bandwidth of 200 or 100 Mbps, depending on the type of the network available near their houses. In practice, the available bandwidth does not go below one-third of the maximum limit.

There are a couple of alternative ways in which we can integrate the home networking functionality in BOB. We can either use a commercially available unit that works as a bridged access point and Ethernet hub for the local network within the house and connect it to the uplink through BOB; or we can use a wireless USB dongle and run the access point on BOB itself.

The former option is simpler in the sense that it comes as a separate unit, which is a small special purpose computer, and we do not need to work on integrating the access point within BOB itself. All we need to do is add an extra Ethernet network interface (NIC) to BOB and connect it to the external unit. On the other hand, the latter option is better aligned towards our overall requirement of being as unobtrusive as possible because we do not need to install a separate unit. However it is tricky because many of the USB-based IEEE 802.11g interfaces that are available

---

<sup>10</sup><http://www.netfilter.org/>

<sup>11</sup><http://thekelleys.org.uk/dnsmasq/doc.html>

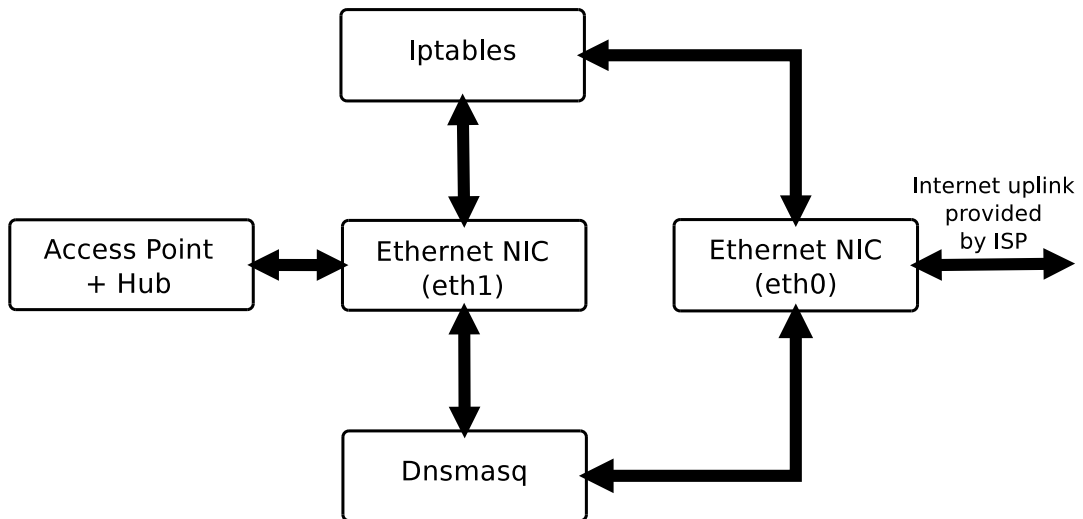


Figure 5: Work flow of the NAT box, DHCP server and caching DNS server.

do not work in the access point or master mode due to lacking drivers, and even if they do they tend to have a weaker signal strength as compared to the commercially available standalone units. Having a weaker signal makes the access point unusable unless the apartment is quite small, and even then it is a problem because stray access points make the wireless channels very noisy within city limits. In practice, having to install a separate unit in the first case is not that much of a problem either because the commercial devices are quite small and neatly built and can be easily placed beside the main BOB box. Moreover, since people will be keeping BOB where they previously kept their set-top box and DVD player, which is below the TV and closer to the floor, it might be convenient to have the flexibility to place the access point slightly higher up than the main unit to have better coverage.

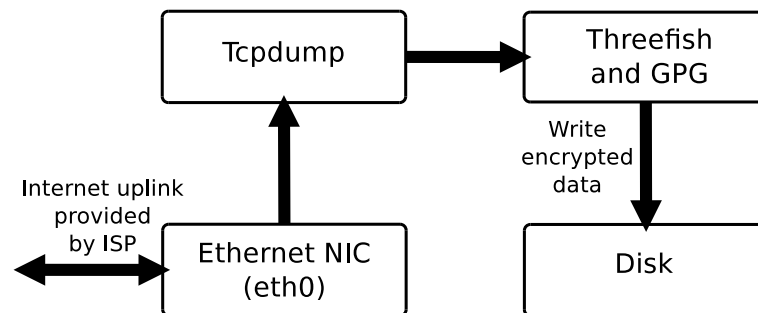


Figure 6: Work flow of the network logger.

Tcpdump<sup>12</sup> is used as a packet sniffer attached to the Ethernet interface of BOB that is connected to the Internet uplink to log the network activity of the test subject. Since the data cached on BOB is uploaded to Alice through the same interface we ignore all packets coming from and going to Alice to avoid pointlessly copying the same data again and again.

### 3.4.6 Cameras and Microphones

To record the audio and video within the test subjects' apartments, we use microphones and cameras that can connect to the wireless network offered by BOB. Thus, they can be placed at different vantage points and the data collected by them can be streamed to BOB over the network.

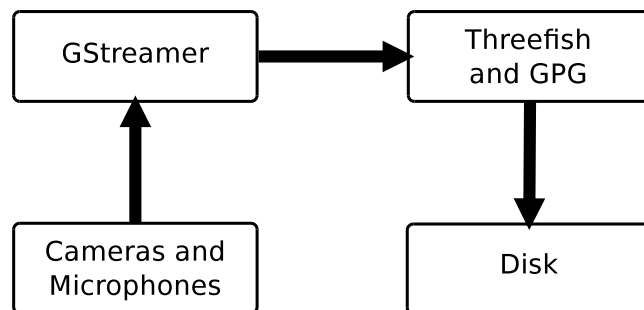


Figure 7: Recording video and audio from cameras and microphones.

While choosing the sensors we ensure that they are of sufficient quality so that the original scene within the house can be reconstructed from the recorded data. Our target environment would have conversations among a small group of people, and voices of different individuals can sometimes overlap in the middle of an exciting discussion. There would also be some amount of movement when a person enters the room, walks by in a passageway, or makes some gesture. These cause artifacts to appear in the collected data, and if the sensors are of poor quality then the amount of artifacts is too high for the data to be of any use.

The microphones and cameras use Real Time Streaming Protocol (RTSP) for streaming the data. So we use a RTSP client to receive the data, and separate the audio and video payloads from the stream. After re-encoding the video as MPEG-4 and the audio as Waveform Audio File Format (WAV), we put them into separate files.

---

<sup>12</sup><http://www.tcpdump.org/>

MPEG-4 and WAV are well known and widely popular, and some cameras have encoders for these formats built into them.

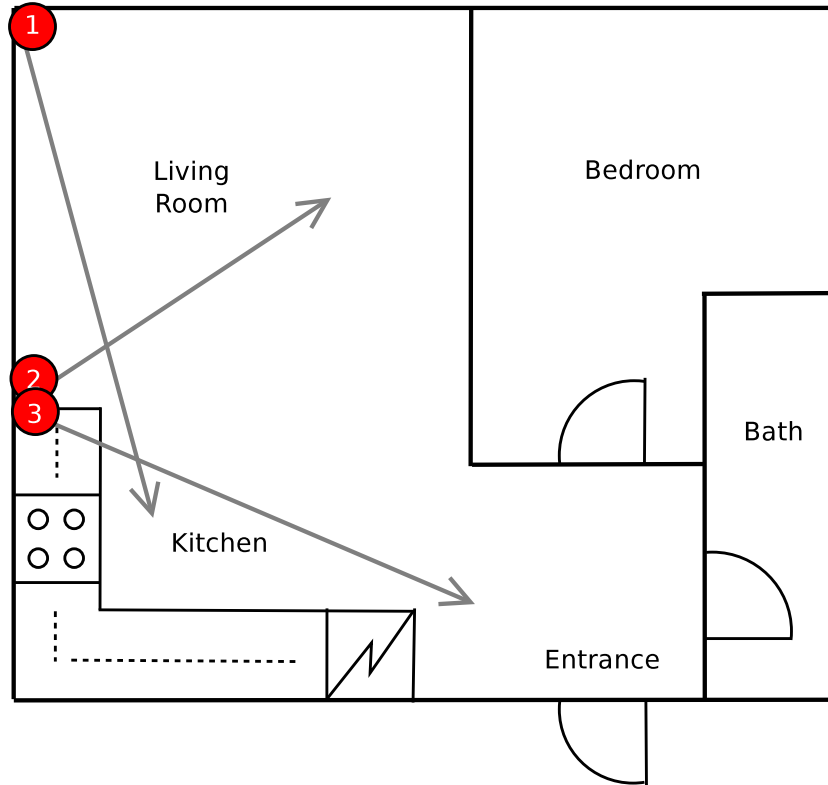


Figure 8: The positioning of the cameras in one of the apartments.

GStreamer<sup>13</sup> is used to record the audio and video feed as described above. It is a comprehensive and well-known multimedia framework for UNIX and provides all the necessary elements for doing so.

### 3.4.7 Data Uploads and Remote Management

Rsync<sup>14</sup> is used to upload the data stored in BOB to Alice over a secure channel as shown in figure 11. Since anomalies in the WAN between the BOBs and Alice can interrupt the transfer, we check whether a restart is needed. Rsync is also capable of limiting the bandwidth consumed during the transfers. This is useful for us to ensure that the transfers consume only a small fraction of the entire Internet bandwidth at the test subjects' end.

<sup>13</sup><http://gstremer.freedesktop.org/>

<sup>14</sup><http://www.samba.org/ftp/rsync/rsync.html>



The BOBs can be remotely managed to enable the researcher, or someone appointed by her, to monitor their status, and diagnose any problems that might have occurred without paying a visit to the test subjects' premises. Common problems are hardware failure, data not getting uploaded due to disturbances in the WAN between the BOBs and Alice, and damaged filesystems caused by sudden loss of power.

The challenge in setting up such a remote management system is that the BOBs might be behind NATs running on the cable modems provided by the test subjects' ISP, and as a result not have a static, globally routable and unicast (ie. public) IP address. Moreover, the ISPs can have firewalls blocking access to certain ports on their networks. Fortunately, it is up to the researcher to decide where Alice is located. Therefore, we can assume that Alice would have a static public IP address without any firewalls interfering with its ports. This means that while it is hard to reach the BOBs from anywhere outside the apartment, the BOBs can connect to any computer, including Alice, on the Internet.

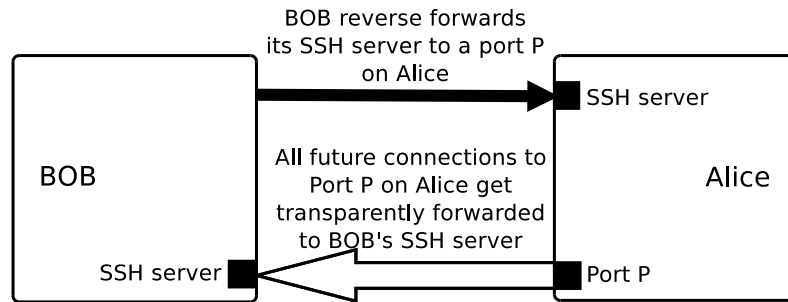


Figure 9: Reverse SSH tunnel between BOB and Alice.

Hence we setup a reverse SSH tunnel from BOB to Alice as shown in figure 9. A connection is initiated from BOB to Alice's SSH server and the port on which BOB's SSH server is listening is reverse forwarded to a local port P on Alice. This establishes a tunnel from port P on Alice to BOB's SSH server. This means that any connections made to that port on Alice will get transparently forwarded to BOB. For each BOB a different value of P is chosen. We get shell access on a specific BOB by first logging on to Alice and then connecting to the port P corresponding to that particular BOB.

### 3.4.8 Installer

We have an installer for quickly setting up the BOB software once the hardware has been assembled and ready. The installation image is a ISO file composed of the data contents of every written sector of an optical disc. This image can either be burnt on to a Compact Disc (CD) or DVD, or written to a USB disk using Ubuntu's `usb-creator`<sup>15</sup> program to create a bootable medium from which to start the installation.

To start, the bootable medium should be inserted and BOB's BIOS should be configured to boot from it. The steps are self explanatory and in many cases default values are provided to make the installation easier. However one should carefully answer the questions shown in figure 10. These are specific to the particular test subject who is going to be observed by this BOB, and hence, the answers will be different for each unit that is installed. Once the installation is finished, restart the computer to begin using it.

Internally, the installer is based on the stock Ubuntu installer called Ubiquity<sup>16</sup>. It is a graphical installer, which is written largely in Python and uses Debian Installer (`d-i`)<sup>17</sup> as a backend. The installation image contains an initial ramdisk (`initrd`)<sup>18</sup> and a kernel. The bootloader loads them into memory and starts the kernel. Once the kernel is ready, the Debian Installer is started. The installation image also contains a Squashfs [Lou11] archive of all the software packages that are to be installed. Once `d-i` is done creating the partitions and file systems on the hard disk, the archive is uncompressed and files are extracted onto the newly created file system hierarchy.

The Debian Installer uses `debconf`<sup>19</sup> for all its user interaction. Ubiquity provides a nice graphical user interface to let the person doing the installation answer all the configuration questions, and these then get written to the `debconf` database, from where `d-i` can read them to complete the installation.

We automate the answers to many of the questions that `d-i` asks by using a preseed<sup>20</sup> file. By doing this we drastically reduce the number of steps that need to be performed by a human during the installation process. Having said that, we do

---

<sup>15</sup><https://launchpad.net/usb-creator>

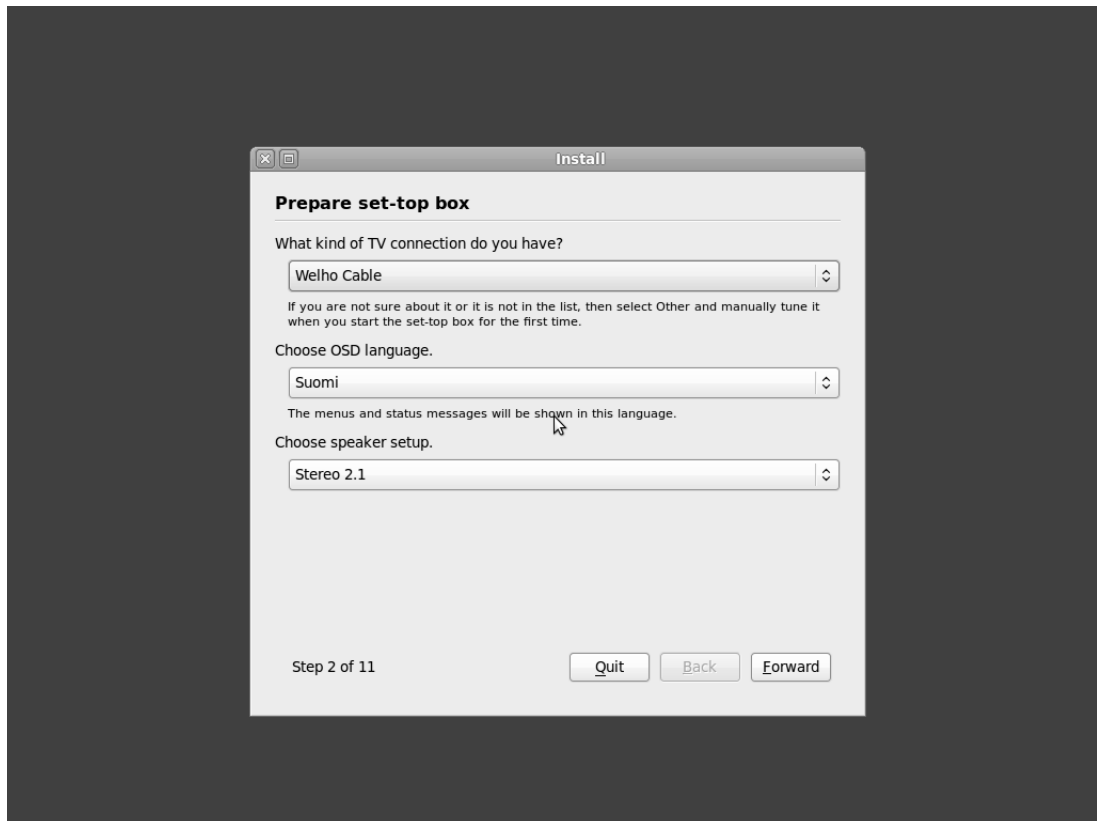
<sup>16</sup><https://wiki.ubuntu.com/Ubiquity>

<sup>17</sup><http://wiki.debian.org/DebianInstaller>

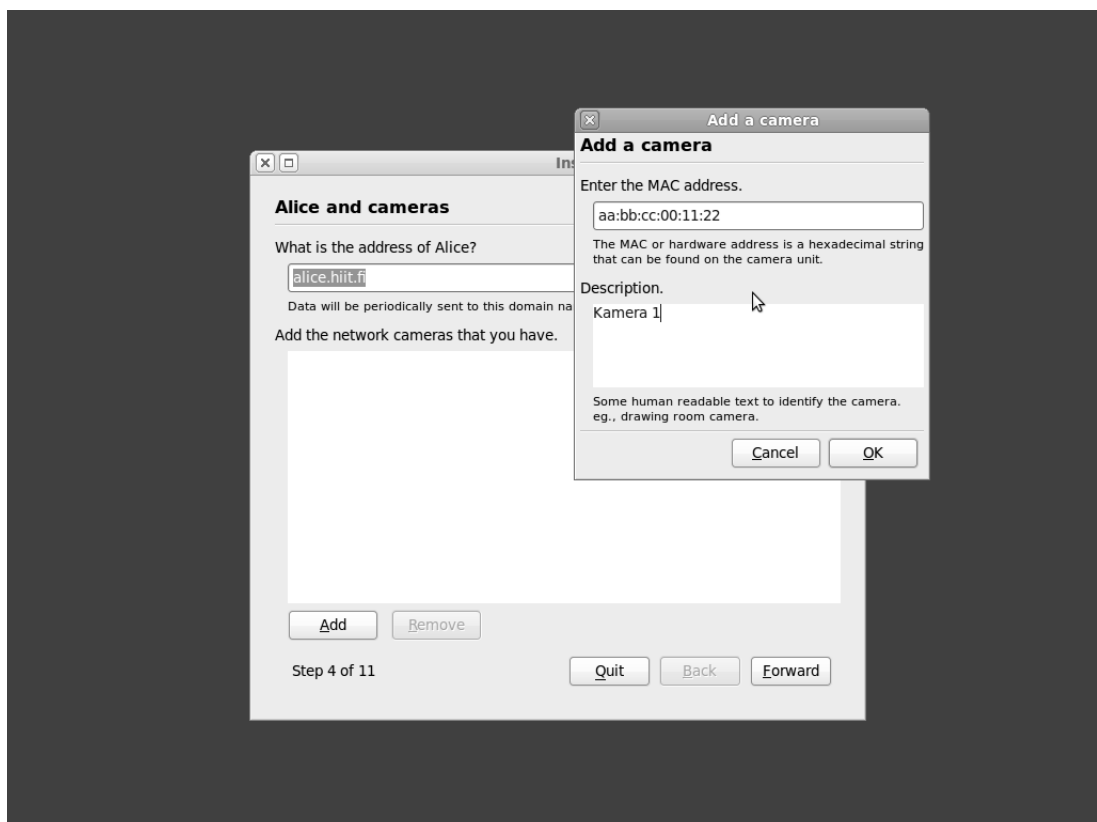
<sup>18</sup>It is a scheme for loading a temporary file system into memory before the real root file system can be mounted during the boot process of a Linux kernel.

<sup>19</sup><http://wiki.debian.org/debconf>

<sup>20</sup><http://wiki.debian.org/DebianInstaller/Preseed>



(a) Prepare the set-top box



(b) Alice and cameras

Figure 10: Steps to install the BOB software.



(c) Who is the test subject

Figure 10: Steps to install the BOB software.

want to present some questions very specific to our use case that are otherwise not asked by d-i. For example, the ones shown in figure 10. We do so by writing our own set of Ubiquity plugins. They present the necessary user interface elements during the initial stages of the installation and record the answers as entries in the debconf database. Later, when d-i has finished setting up the basic operating system, the plugins update the configuration files based on the values in debconf.

### 3.5 Alice File Server

Alice is a file server whose primary objective is to store all the data collected by the key loggers, smartphone loggers and BOBs in the field. To get a rough idea about the size of data generated per day by the sensors we ran a couple of pilot installations of BOB for a month. Although the numbers, shown in table 1, will vary depending on the actual environment in which the sensors were installed and

the behaviour of the test subjects, it serves as a basis for calculating the amount of disk space needed on Alice.

Sensor	Size	Remarks
Camera	7028 MB	Resolution=640x480, FPS=30, MPEG-4
Microphone	1240 MB	WAV
TV & DVD Logger	579 KB	
Network Logger	1589 MB	
Wireless Access Point Scanner	580 KB	
Bluetooth Scanner	580 KB	

Table 1: Average size of data generated per day by different sensors.

Data files stored on Alice are in the same encrypted form as they were on the source devices. As explained in section 3.4.3, the researcher should have the RSA private key to retrieve the Threefish symmetric keys for the specific files that he wants to access. Since this server is accessible over the Internet, we do not store the private key on it. Instead, it is kept on a separate computer disconnected from the network. Whenever a set of files need to be accessed they are copied to the offline node over some portable storage device for decryption. This reduces the chances of some attacker getting access to the private key.

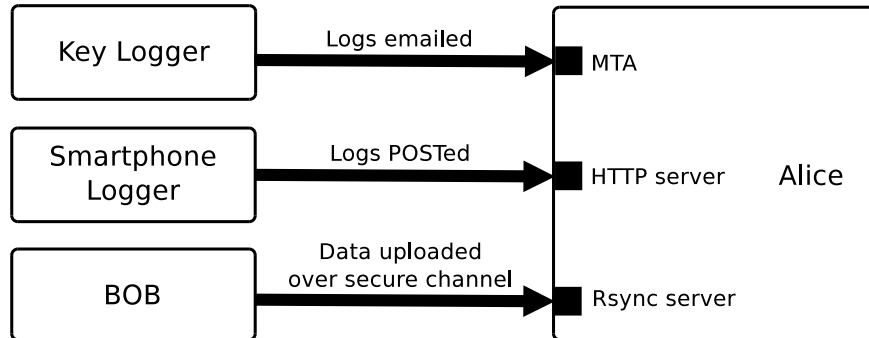


Figure 11: Data being sent to Alice from different sources.

To accept the data sent from the key loggers and smartphone loggers, a Mail Transfer Agent (MTA), and a HTTP server with support for POST requests are run on Alice.

We use Exim<sup>21</sup> as the MTA, and the Apache HTTP Server<sup>22</sup>.

Alice has a publicly accessible SSH server to provide a secure channel for data uploads and the scheme for remotely managing the BOBs as described in section 3.4.7. OpenSSH<sup>23</sup> is the SSH server implementation of our choice. It supports non-interactive passwordless key-based logins so that the clients on the BOBs can connect to Alice without human intervention.

We have some scripts on Alice to automate some of the remote management tasks by leveraging the reverse SSH tunnels from Alice to the BOBs. For example, to check the status of the sensors, how reliable they have been at recording data, and whether the uploads are progressing smoothly or the BOBs are running out of space due to network problems. The output from the scripts is displayed in a human readable format on a password protected website to make it easier for the researcher to keep an eye on what is going on.

---

<sup>21</sup><http://www.exim.org/>

<sup>22</sup><http://httpd.apache.org/>

<sup>23</sup><http://www.openssh.org/>

## 4 Data Analysis and Reliability

Having designed and implemented a data collection and transportation platform, we are going to show a few techniques that can be employed to analyse the data provided by the platform. We also show how reliable the different sensors in the platform are under real world conditions.

### 4.1 Overview

We present a few example data analysis techniques and reliability measures to examine how effectively our instrumentation platform meets the requirements layed out in chapter 2. Since the data that we collect is diverse in nature, we provide techniques to analyze one particular sensor – the network logger.

Based only on the network traffic, the test subjects are profiled based on the following components:

- The frequency of visits to different kinds of websites. (A taxonomy for websites is defined to categorize the sites.)
- The amount of traffic at different hours of the day.
- The amount of traffic on different days of the week.

We provide detailed steps to show how the data was analyzed to score the test subjects on each of the above points. Then, based on these parameters we draw some interesting inferences about them.

When used in a real world experiment outside the laboratory, the reliability of the sensors are adversely affected. Therefore, we present some reliability figures for each sensor to show how good the platform is at monitoring the participants. The reliability of each sensor is calculated, and the causes of imperfection are explored.

### 4.2 Taxonomy for Websites

Based only on the websites visited by the test subjects we draw up a taxonomy for categorizing websites. Since the number of test subjects is very small, and they are all from Southern Finland, our sample set is quite restricted and biased. Therefore

our taxonomy probably can not be generalized and applied on any other subset of websites.

Allotting a website to single category can be tricky because many websites have a wide variety of content. For example, the Finnish national public-broadcasting company, YLE, has news, TV and radio schedules, video streams of current and past programs, and blogs on its website. Therefore, our taxonomy is so designed that before deciding which category a website belongs to, one should determine the primary aim of the website. Sometimes this is made easier by the fact that bigger websites are divided into subdomains, where each has a specific purpose. Each subdomain can then be treated as a separate website and categorized separately.

These days popular websites use content delivery networks (CDNs)<sup>24</sup> to increase access bandwidth and redundancy, and reduce access latency. Therefore the same content can be served by multiple similarly named domains. Since these are logically the same website, we club them together as a single entity.

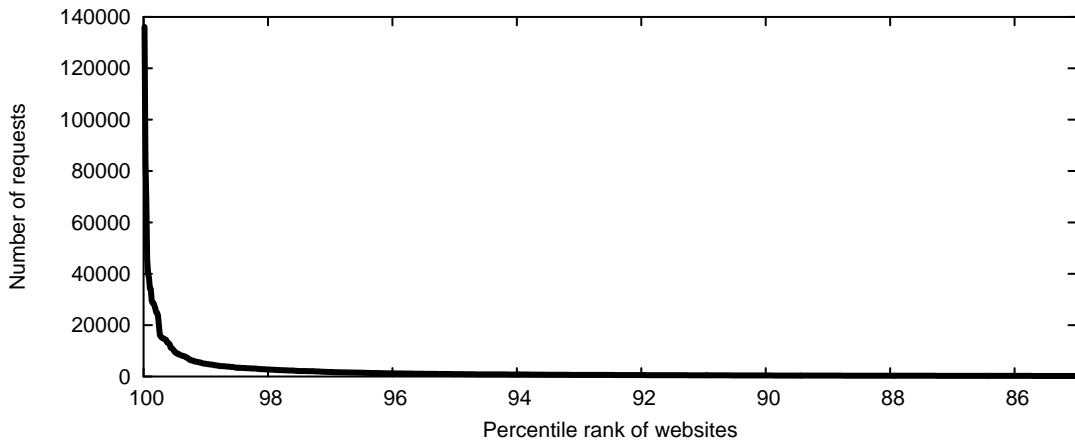


Figure 12: Number of requests to the most popular websites across all test subjects.

We create a list of sites that were accessed by all the test subjects and count the number of HTTP requests made to each of them. The list is sorted in decreasing order of requests, and the percentile rank for each entry is calculated. When plotted on a graph, as shown in figure 12, this looks like a sharply decreasing exponential curve. It is obvious that compared to the most popular websites the others have

---

<sup>24</sup>CDN is a system of computers containing copies of data placed at various nodes of a network. Data content types often cached in CDNs include web objects (text, graphics, URLs and scripts), downloadable objects (media files, software, documents), applications, live streaming media and database queries.



almost negligible number of hits. Even though the websites with lower percentile ranks are often unique to a particular household, the more popular ones are still representative of the general trends. Therefore we consider only those entries that have a percentile rank of 85 or more for our taxonomy.

The following different categories for websites are defined:

- **Adult Entertainment:** Websites that show sexually explicit content for enjoyment and relaxation. They can be similar to a personal website when it is a website of a porn actor or actress, or a media sharing website when users can upload their own sexually explicit material.
- **Advertisement:** Hosts that are part of advertisement serving overlay networks. The advertisements are typically embedded in search engine results, blogs or portals, and are automatically requested and shown to the user as part of the webpage. Some of the domain names that are used for these purposes are track.adform.net, cdn.adnxs.com, adtechus.com, atemda.com, emediate.biz and zedo.com.
- **Dating:** Websites where users can find other single people looking for long range relationships, dating, or just friends.
- **Blog:** Websites that are used to post online diaries and can cover a wide range of topics depending on the author. The content is typically such that it represents the personal opinion of the blogger, or her own experiences or achievements. For example, wordpress.com, blogspot.com.
- **Broadcasting:** Websites that broadcast video or audio content from different publishers. For example, the Video on Demand (VOD) network called ePlayer.
- **Corporate:** Websites providing background information about a business, organization or service. For example, europe.nokia.com, suunto.com, kihlasor-mukset.com.
- **Electronic Commerce:** Websites offering goods and services for online sale, and enabling online transactions for such sales. It also includes those sites whose purpose is to sell products from third parties, or syndicate content from other sellers. For example, ikea.com, sandviks.com, vau.fi, vaukirja.fi, nettiauto.com, pizza-online.fi, smartshopper.com.

- **Encyclopedia:** Websites that are compendiums holding summary information from different branches of knowledge in the form of articles. For example, Wikipedia.
- **Employment:** Websites where employers announce job openings and people can search for jobs based on different criteria. For example, monster.fi.
- **Forums & Groups:** Websites where people can discuss various topics. For example, groups.google.com, groups.yahoo.com.
- **Gambling:** Websites that let users gamble online. For example, bet365.com.
- **Games:** Online gaming sites. For example, hattrick.org, travian.fi, eu.battle.net.
- **Government:** Websites made by the local, state, department or national government of a country. Usually these sites also operate websites that are intended to inform tourists or support tourism. For example, hsl.fi, mol.fi, reit-tiipas.fi.
- **Information:** Websites that give out miscellaneous bits of information. These could be anything from restaurant lunch menus to movie ratings.
- **Maps:** Web mapping services that offers street maps, route planner or urban business locator for numerous countries around the world. For example, maps.google.com, maps.nokia.com.
- **Media Sharing:** Websites that enables users to upload and view media such as pictures, music and videos. For example, Flickr, Picasa, Youtube.
- **Microblog:** A short and simple form of blogging. Microblogs are limited to certain amounts of characters and are similar to status updates. For example, Twitter.
- **News:** Websites dedicated to dispensing news, politics and commentary. For example, bbc.co.uk, iltalehti.fi, iltasanomat.fi, seiska.fi, veikkausliiga.com.
- **Peer-to-Peer:** Websites that index or track BitTorrent files, Manolito music shares.
- **Personal:** Websites about an individual or a small group that are of a non-commercial nature.

- Portal: Websites that are the starting point or gateway to other resources on the Internet. For example [msn.com](http://msn.com), [yahoo.com](http://yahoo.com).
- Search Engine: Websites that index material on the Internet and provides links to information as a response to a query. For example, [www.google.fi](http://www.google.fi), [scholar.google.fi](http://scholar.google.fi), [www.bing.com](http://www.bing.com).
- Software As A Service: Websites that offer web applications hosted centrally on a server, which can be accessed by users using a browser over the Internet. Typical examples of such software are word processors and spreadsheets provided by [docs.google.com](http://docs.google.com).
- Social Networking: Websites where users could communicate with one another, play games, create a network of contacts, share other content from the Internet and rate or comment on the content. For example, Delicious, Facebook, Google Plus, LinkedIn, MySpace, Orkut, Reddit.
- Software Downloads: Websites that allow users to download software programs and provides updates. For example, mirrors of Mozilla and other free and open source software (FOSS) projects, Windows updates, F-Secure or MacAfee updates.
- Translate: Websites that let users translate text written in one language to another. For example, [translate.google.com](http://translate.google.com).
- Weather: Websites that provide weather services. For example, [wwc.instacam.com](http://wwc.instacam.com), [weatherbug.com](http://weatherbug.com), [foreca.fi](http://foreca.fi).
- Webmail: Websites that provide web interface for reading and sending email. For example, [mail.google.com](http://mail.google.com), [mail.yahoo.com](http://mail.yahoo.com).

The advertisement serving overlay networks are ignored while creating the personality profiles. The advertisements are embedded in other websites, and whenever a human user opens such a website requests are automatically issued to show the advertisements. These days online advertising based revenue models are increasingly popular, and having embedded advertisements is hardly a distinguishing feature of a website. Therefore the frequency with which the test subjects visited these networks is not a true indicator of their personality.

These days many Web services and applications provide Application Programming Interfaces (API) to allow people to interface with them from applications other than

traditional Web browsers. For example, microblogging platforms often have APIs for writing clients that are specifically designed for the purpose. To use these APIs one has to access a particular subdomain of the parent website. However, if the parent website offers various different kinds of services then all the APIs are usually offered under the same subdomain. This makes it impossible to know which service was being used through the APIs just by looking at the DNS responses. In such cases we ignore these subdomains.

### 4.3 Household Profile

From the rich collection of data that we have it is possible to reconstruct many different aspects of the test subjects' personalities. However, as mentioned before, we only concentrate on one modality – the network logger. Our aim is to construct profiles to assess the similarities and dissimilarities between the households based on the following components:

- The frequency of visits to different kinds of websites.
- The amount of traffic at different hours of the day.
- The amount of traffic on different days of the week.

Formally, we define a profile as a vector based on the above components. The test subjects can be assessed on the basis of a single component, or the individual component vectors can be combined to get a more wholesome picture of their similarities and dissimilarities.

To create the vector for the categorization of websites, we generate a histogram where the categories correspond to those defined in section 4.2. We look at all the DNS replies logged at the test subject's house and create a mapping from host name to IP address. The number of HTTP requests is calculated for each host name. The host name is allotted to a category based on the taxonomy, and the number of requests for that category is incremented accordingly. Finally, we normalize the resulting histogram so that it sums up to unity.

Similarly for the other two components, the histograms have twenty-four and seven bins respectively. We look at the distribution of network traffic over a single day and a single week, and count the amount of bytes that fall into each slot. The resulting histograms are then normalized.

### 4.3.1 Distance Calculation and Visualization

Given any component  $c$  (as mentioned in section 4.3), and a pair of test subjects  $X_0$  and  $X_1$ , we want to calculate the distance  $d_c(X_0, X_1)$  between them based on it. Kullback-Leibler divergence is used as the distance measure. If the probability distributions of  $X_0$  and  $X_1$  based on component  $c$  are  $X_{0c}$  and  $X_{1c}$  respectively, then the Kullback-Leibler divergence  $D_{KL}(X_{0c}||X_{1c})$  is calculated as given in equation (1).

$$D_{KL}(X_{0c}||X_{1c}) = \sum_i X_{0c}(i) \ln \frac{X_{0c}(i)}{X_{1c}(i)} \quad (1)$$

Since the Kullback-Leibler divergences are not symmetric, we use J-divergence to make them so [JS00].

$$D_J(X_{0c}||X_{1c}) = D_{KL}(X_{0c}||X_{1c}) + D_{KL}(X_{1c}||X_{0c}) \quad (2)$$

We can add some importance to a component by assigning a weight  $\alpha_c$  to it. The exact value of the weights are to be decided by the researcher depending on his domain knowledge and discretion. Here, we consider them to be unity. The overall distance  $d(X_0, X_1)$  between two test subjects  $X_0$  and  $X_1$  can be calculated as given in equation (3).

$$d(X_0, X_1) = \frac{1}{N} \sum_{c=1}^N \alpha_c d_c(X_0, X_1) \quad (3)$$

This way the distances between every pair of test subjects is calculated. Since the distances are originally between vectors in multi-dimensional space, we can not directly visualize them in two dimensions. To do so, we use gradient descent to minimize the errors in the distances when they are projected from their original space on to a two-dimensional plane.

### 4.3.2 Inferences

Data is still coming in from the field, and different test subjects have had BOB running in their apartments for varying periods of time. Therefore we consider only the first three months worth of data in our experiments.

As described earlier in section 4.3.1 we calculate the distances between every pair of test subjects for each of the three different components of the network logger data that we are interested in. Two dimensional visualizations are plotted for each component to get an idea of where each household stands with respect to the others.

This is same as using equation (3) with the value of  $\alpha_c$  as unity for the particular component  $c$  for which we are going to draw the visualization and zero otherwise.

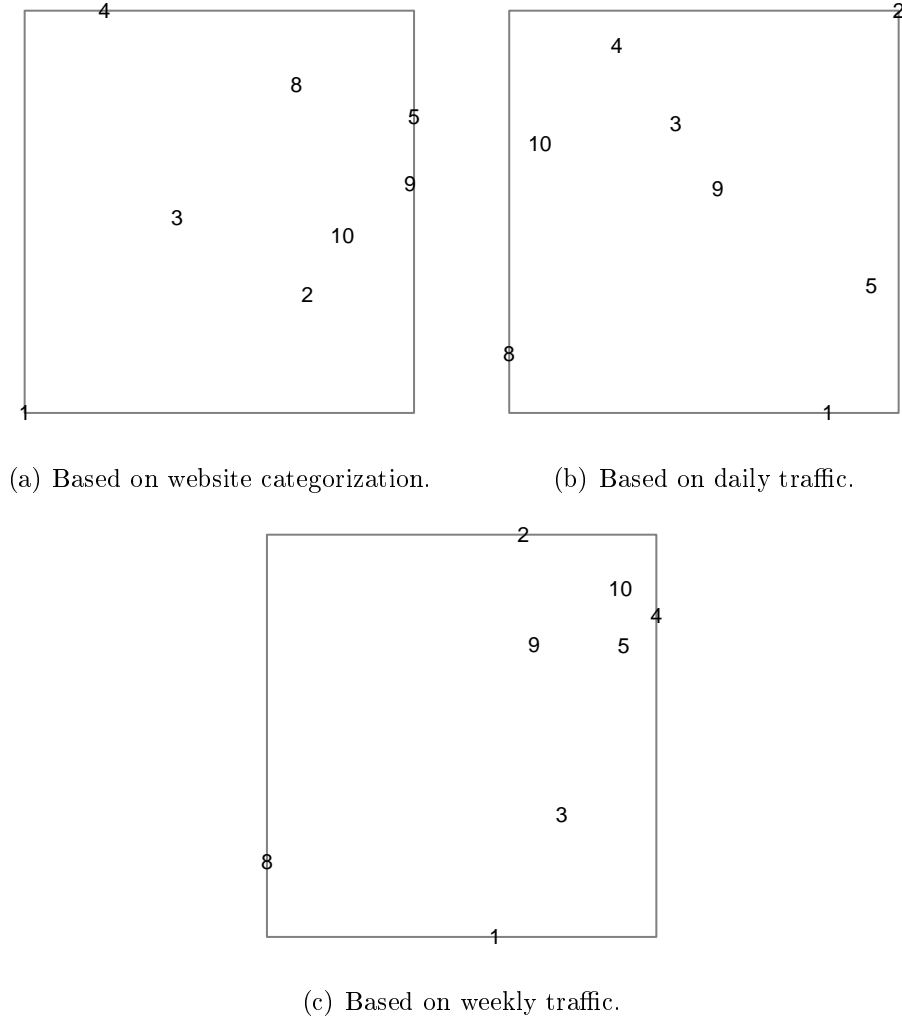


Figure 13: Similarity of the households based on each component.

Similarly, we calculate the overall distances between each pair of test subjects by using unity as the weights for all the components in equation (3). The two dimensional visualization based on these distances indicate the similarity of the households based on all the three components.

Therefore, based on figure 13 and figure 14 we can draw some inferences about the participating households. The histograms for every test subject based on the components mentioned in section 4.3 are provided in the appendices. We only highlight the ones that we found interesting in this section.

From figure 14, participants one, two, four, five and eight look interesting because

they are the outliers. Of them the first participant is the most unusual because she is an outlier in all the similarity diagrams. For the eighth participant this is reflected in figure 13(b) and figure 13(c), while figure 13(a) indicates that the fourth participant visited websites that are different from the rest. The second and fifth households are outliers in figure 13(b).

Let us look at a closer look at these interesting cases to see if we can draw some inferences about their behaviour.

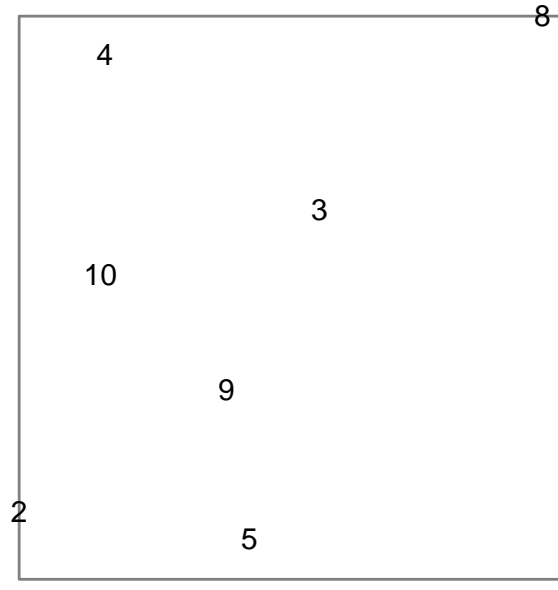


Figure 14: Overall similarity of the households.

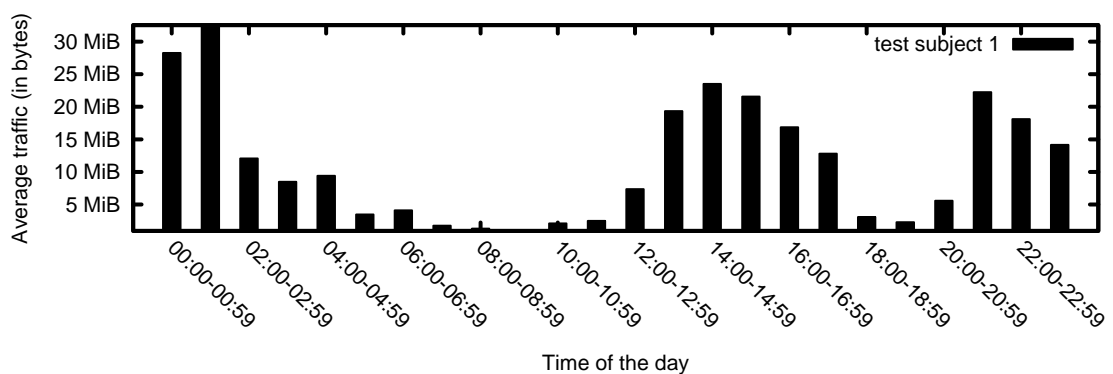


Figure 15: Test subject 1: daily traffic in bytes.

Figure 15 shows that the first test subject is online all round the clock except two

brief periods – from 07:00 hr to 12:00 hr and 18:00 hr to 20:00 hr. Assuming that she has to go to work to earn her livelihood, this indicates that either she does not have fixed working hours, or that she has a very irregular routine, or maybe both. She is a frequent visitor to adult entertainment and social networking websites, with the former having the bigger share.

The second participant is interesting in the sense that there is a single sharp spike in her daily traffic from 21:00 hr to 22:00 hr. For the other participants, either there are multiple spikes or the duration is longer than a single hour.

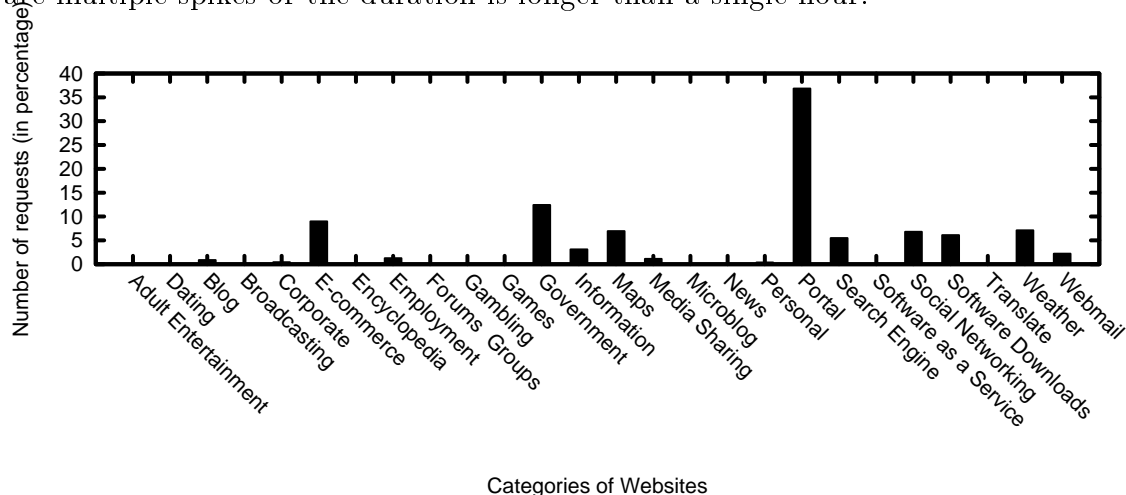


Figure 16: Test subject 4: distribution of requests made to different categories of websites.

The fourth household stands out due to the total absence of websites related to news in figure 16. Every other household has shown some inclination to read the news on the Internet.

The fifth test subject does most of his surfing during the evenings and nights. She usually sleeps after 02:00 hr, sometimes even as late as 04:00 hr, and does not wake up before 10:00 hr. Not a very healthy routine, but predictable because she sticks to it. She is probably a very heavy user of peer-to-peer file-sharing services because there is a disproportionately high amount of traffic at her end, which is unaccounted for by the number of HTTP requests issued by her.

The eighth test subject is unusual because more than 50% of her online activity happens on Mondays and there is none on weekends. Almost 70% of her network traffic occurs between 13:00 hr and 18:00 hr and the activity during the evenings is remarkably low, unlike the other participants.



These inferences can be used to score the participants on the “Big Five” factors as mentioned in section 2.1. Looking at the first test subject’s irregular routine we give her a low conscientiousness score. The fifth subject ranks higher than the first in terms of conscientiousness because even though she does not maintain the healthiest of timetables, she at least sticks to it and does not change her sleeping hours randomly. Going by the heavy use of peer-to-peer file-sharing services at her end, and the presence of popular websites related to movies and TV shows, one can deduce that she consumes a lot of video content from the Internet. This indicates some amount of appreciation for the arts, which translates to a higher openness score.

However, it is hard to draw any further conclusions without looking at data from the other sensors. For example, we know that the first test subject is a frequent visitor to social networking sites, but we do not know anything about her real world interactions. Analyzing the data from the cameras and microphones would give us a better insight into that. Looking at the data from the set-top box would tell us what whether the fifth participant watches any movies or programmes on the television, or he uses only the Internet.

Although the second, fourth and eighth test subjects appear to be interesting from their network logs, we can not conclude anything further about their personality. For example, the absence of visits to news related websites from the fourth household can be due to the fact that they prefer printed material, or it can be due to a complete lack of interest in current affairs. It is impossible to draw a conclusion without more data. Analyzing the audio and video data to check their daily activities can be helpful.

## 4.4 Reliability of Data Gathering

Based on the way they collect data, we have two different types of sensors. The reliability is calculated differently for each type. Sensors like the cameras, microphones, TV and DVD logger, and network logger record continuously round the clock. On the other hand, the wireless access point scanner, and Bluetooth scanner only take snapshots at regular intervals of time.

Reliability for the continuous sensors is calculated as the percentage of the total number of minutes covered during the experiment. A sensor has to work for at least 30 seconds for a particular minute to be counted as covered. For the snapshot

sensors, reliability is the percentage of attempted scans that were successful.

The reliability numbers were estimated by selecting a sample of ten random days from each test subject. So although they are indicative of how the system fared over the course of the experiment, they do not reflect the absolute reliability of the platform.

<b>Sensor</b>	<b>Reliability (in %age)</b>
Cameras	82.645
Microphones	82.645
TV & DVD Logger	73.421
Network Logger	76.078
Wireless Access Point Scanner	90.040
Bluetooth Scanner	90.040

Table 2: Reliability of Sensors

The figures for some of the sensors in table 2 have been adversely affected due to various factors. Some were unavoidable, while others were caused by bugs in the system.

July, August, December, January and February are holiday months in Southern Finland. Many of the test subjects were on vacation during these months and turned off the BOBs in their apartments to save power. On other occasions due to the presence of visitors, who do not want to participate in the experiment, some of the sensors were disabled by the families to protect their privacy.

Sometimes the BOBs become unresponsive due to software bugs. For example, race conditions in the DVB card drivers can cause the whole system to freeze. Situations like these and the ones mentioned above can instigate the test subjects to hard reboot the BOBs. Since the sensors are constantly recording data, the file systems perform a lot of write operations. Hard reboots abruptly interrupt these operations and may break the file system. A broken file system is effectively the same as turning the machine off because the system will fail to boot.

Whenever problems arise in the WAN connecting the BOBs with Alice data accumulates in the BOBs. If connectivity is not restored soon enough they run out of disk space, which prevents all the sensors from recording anything. During the early part of the experiment we discovered that due to a misconfiguration in one of the routers in Alice’s network data uploads were not proceeding as smoothly as they

should have. While we could not track down the faulty router, we made things better by disabling Duplicative Selective Acknowledgement (DSACK)<sup>25</sup> in the BOBs and Alice. Remaining network problems are caused by disturbances in the ISPs' networks and these are beyond our control. If the Internet uplink provided by the ISP goes down, the network logger can not work either.

---

<sup>25</sup>It is an extension of the Selective Acknowledgement (SACK) option for TCP specified in RFC 2883. This extension allows the TCP sender to infer the order of packets received at the receiver, allowing the sender to infer when it has unnecessarily retransmitted a packet.

## 5 Conclusions

We have specified the requirements for an instrumentation platform that can be used to conduct pervasive longitudinal studies within people’s intimate surroundings by gathering data about how people behave in their various places of presence. We have also described how to design and implement such a system, and shown how the collected data can be used to create personality profiles of the test subjects. However, we have only concentrated on a single modality while creating the profiles. More advanced data analysis techniques that take into consideration multiple modalities will reveal much more about the individuals and families as we set out to achieve in section 2.1.

Creating such an intrusive framework and using it to gather data for such a long period of time raises a lot of privacy and ethical issues. To resolve these issues we hired a lawyer and consulted the Data Ombudsman of Finland. Finnish and European Union law allow the collection of data on third persons for scientific research. While we take permission from the test subjects before studying them and notify them about the location of the sensors, it is possible that some visitors might not be aware that every move inside the apartment is being monitored. Therefore we asked them to inform their closest acquaintances about the surveillance and limited ourselves to studying the only the participating subjects. In case it was not feasible for them to explain the study beforehand to their guests, they were allowed to turn off the sensors. Guidelines were established regarding the anonymization of the data. Before releasing any subset of the data or the results of any analyses, it is screened by the members of the research group. In case of doubt, a written disclosure is requested from the particular subject.

Moreover, by allowing us to log their use of the television and Internet the participants might be violating the TV or Internet provider’s terms of service. At the same time we are obliged to co-operate with law enforcement agencies by giving them access to our data if there is suspected criminal activity.

Half yearly interviews with the participants were organized to find out how the instrumentation of their houses affected their daily lives. The subjects were asked about the routines and events that had changed as a result of the continued surveillance. Nakedness, consumption of alcohol, substance use, use of space, sex, and choice of place for conversation were some of the changes that were reported. Most people went through three different phases during the course of the experiment.

The initial confrontation phase starting from the moment their house was wired and lasting for roughly two weeks. During this time the subjects reported “annoyance”, “weirdness” and “novelty” as the dominant feelings. Acceptance and decreasing awareness marked the second phase, which was sometimes interrupted by ad hoc events that reminded them of the presence of the data gathering system. Examples of such events varied from technical problems in BOB itself to moments of personal difficulty.

One of the test subjects left the experiment after six months due to getting increasingly disturbed by the effects of ubiquitous surveillance. She complained that her use of the computer and social interactions were severely affected by the study, and that she was afraid of being harmed in case the data was leaked.

## References

- AHWF02 Amichai-Hamburger, Y., Wainapel, G. and Fox, S., "on the internet no one knows i'm an introvert": Extroversion, neuroticism, and internet interaction. *Cyber Psychology & Behaviour*, 5,2(2002), pages 125–128.
- Ara95 Araya, A. A., Questioning ubiquitous computing. *Proceedings of the 1995 ACM 23rd annual conference on Computer science*, CSC '95, New York, NY, USA, 1995, ACM, pages 230–237.
- Bar68 Barker, R. G., *Ecological Psychology: Concepts and Methods of Studying the Environment of Human Behaviour*. University Press, Stanford, 1968.
- BD-a Blu-ray popularity continues to grow. <http://www.blu-ray.com/news/?id=3653>. [12.09.2011]
- BD-b Don't buy hd-dvd or blu-ray disks. <http://www.fsf.org/news/blu-ray>. [12.09.2011]
- CM92 Costa, P. T. and McCrae, R. R., *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory*. Psychological Assessment Resources, 1992.
- Den08 Dennis, K., Viewpoint: Keeping a close watch - the rise of self-surveillance and the threat of digital exposure. *The Sociological Review*, 56,3(2008), pages 347–357.
- Dig90 Digman, J. M., Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, 41,1(1990), pages 417–440.
- DVI Press release by digital display working group. <http://www.ddwg.org/articles.asp?id=22>. [12.09.2011]
- EMdM07 Eagleton, J., McKelvie, S. and de Man, A., Extraversion and neuroticism in team sport participants, individual sport participants, and non-participants. *Percept Mot Skills*, 105,1(2007), pages 265–75.
- EPL09 Eagle, N., Pentland, A. S. and Lazer, D., Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106,36(2009), pages 15274–15278.

- ESP06 Eagle, N. and (Sandy) Pentland, A., Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10, pages 255–268.
- FSV07 Fritsch, T., Schiller, J. and Voigt, B., Personal behavior and virtual fragmentation. *Proceedings of the international conference on Advances in computer entertainment technology*, ACE '07, New York, NY, USA, 2007, ACM, pages 60–63.
- GAL98 Introversion, 1998. [http://findarticles.com/p/articles/mi\\_g2602/is\\_0003/ai\\_2602000328/](http://findarticles.com/p/articles/mi_g2602/is_0003/ai_2602000328/). [28.09.2011]
- GAW11 Hell hole, 2011. [http://www.newyorker.com/reporting/2009/03/30/090330fa\\_fact\\_gawande](http://www.newyorker.com/reporting/2009/03/30/090330fa_fact_gawande). [27.09.2011]
- GBST07 Graziano, W. G., Bruce, J., Sheese, B. E. and Tobin, R. M., Attraction, personality, and prejudice: liking none of the people most of the time. *Journal of Personality and Social Psychology*, 93,4(2007), pages 565–582.
- GHST07 Graziano, W. G., Habashi, M. M., Sheese, B. E. and Tobin, R. M., Agreeableness, empathy, and helping: a person x situation perspective. *Journal of Personality and Social Psychology*, 93,4(2007), pages 583–599.
- Gol93 Goldberg, L. M., The structure of phenotypic personality traits. *American Psychologist*, 48,1(1993), pages 26–34.
- GOO10 Google transparency report – government requests, 2010. <http://www.google.com/transparencyreport/governmentrequests/>. [21.09.2011]
- GOO11 Google agrees to biennial privacy reviews, 2011. <http://www.thinq.co.uk/2011/3/30/google-agrees-biennial-privacy-reviews/>. [20.09.2011]
- Gos08 Gosling, S., *What Your Stuff Says About You*. Basic Books, New York, 2008.
- GPU11 Folding@home distributed computing, 2011. <http://fah-web.stanford.edu/cgi-bin/main.py?ctype=osstats>. [05.10.2011]

- Has10 Hasu, T., Contextlogger2 - a tool for smartphone data gathering. Technical Report, Helsinki Institute for Information Technology, 2010.
- HDM High-definition multimedia interface (hdmi) adopters and affiliates. [http://www.hdmi.org/learningcenter/adopters\\_founders.aspx](http://www.hdmi.org/learningcenter/adopters_founders.aspx). [12.09.2011]
- Hol09 Holwerda, T., Ballmer: Linux bigger competitor than apple, 2009. [http://www.osnews.com/story/21035/Ballmer\\_Linux\\_Bigger\\_Competitor\\_than\\_Apple](http://www.osnews.com/story/21035/Ballmer_Linux_Bigger_Competitor_than_Apple). [21.01.2012]
- Hou10 Hough, A., Google engineer fired for privacy breach after 'stalking and harrasing teenagers', 2010. <http://www.telegraph.co.uk/technology/google/8003925/Google-engineer-fired-for-privacy-breach-after-stalking-and-harrasing.html>. [01.01.2012]
- ILB<sup>+</sup>05 Intille, S. S., Larson, K., Beaudin, J. S., Nawyn, J., Tapia, E. M. and Kaushik, P., A living laboratory for the design and evaluation of ubiquitous computing technologies. *CHI '05 extended abstracts on Human factors in computing systems*, CHI EA '05, New York, NY, USA, 2005, ACM, pages 1941–1944.
- Int02 Intille, S. S., Designing a home of the future. *IEEE Pervasive Computing*, 1, pages 76–82.
- Int06 Intille, S. S., The goal: Smart people, not smart homes. 2006.
- JCG01 Jensen-Campbell, L. and Graziano, W., Agreeableness as a moderator of interpersonal conflict. *Journal of Personality and Social Psychology*, 69,2(2001), pages 323–61.
- Jos06 Jost, J. T., The end of the end of ideology. *American Psychologist*, 61,7(2006), pages 651–70.
- JS00 Johnson, D. H. and Sinanovic, S., Symmetrizing the kullback-leibler distance. Technical Report, IEEE Transactions on Information Theory, 2000.
- Kel93 Kellehear, A., *The unobtrusive researcher: a guide to methods*. Allen and Unwin, St. Leonards, NSW, Australia, 1993.



- Lee00 Lee, R. M., *Unobtrusive Methods in Social Research*. Open University Press, Buckingham, 2000.
- Llo11 Lloyd, R., Crowd-sourced data hold potential for positive change and human rights abuses, 2011. <http://blogs.scientificamerican.com/observations/2011/02/18/crowd-sourced-data-hold-potential-for-positive-change-and-human-rights> [20.09.2011]
- Lou11 Lougher, P., Squashfs - a squashed read-only filesystem for linux, 2011. <http://squashfs.sourceforge.net/>. [05.01.2012]
- May45 Mayo, E., *The Social Problems of an Industrial Civilization*. Division of Research, Harvard Business School, Boston, 1945.
- MI87 McCrae, R. R. and Ingraham, L. J., Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology*, 52,6(1987), pages 1258 – 1265.
- Mil07 Mills, E., Yahoo settles lawsuit with jailed chinese journalists, 2007. [http://news.cnet.com/8301-10784\\_3-9815950-7.html](http://news.cnet.com/8301-10784_3-9815950-7.html). [01.01.2012]
- MKA04 Malhotra, N. K., Kim, S. S. and Agarwal, J., Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Info. Sys. Research*, 15, pages 336–355.
- MOB11 Gartner newsroom, 2011. <http://www.gartner.com/it/page.jsp?id=1764714>. [21.01.2012]
- Neg95 Negroponte, N., *Being Digital*. Vintage Books, New York, 1995.
- PD03 Palen, L. and Dourish, P., Unpacking "privacy" for a networked world. *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '03, New York, NY, USA, 2003, ACM, pages 129–136.
- RD66 Roethlisberger, F. and Dickson, W., *Management and the worker. An account of a Research Program conducted by the Western Electric Company, Hawthorne Works, Chicago*. Harvard University Press, Cambridge Massachusetts, 14th edition, 1966.

- REU        Dvi on the decline as hdmi and displayport grow. <http://www.reuters.com/article/2008/01/28/idUS142983+28-Jan-2008+BW20080128>. [12.09.2011]
- RG03        Rentfrow, P. J. and Gosling, S. D., The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84,6(2003), pages 1236–1256.
- ROE09        Raento, M., Oulasvirta, A. and Eagle, N., Smartphones: An emerging tool for social scientists. *Sociological Methods and Research*, 37,3(2009), pages 426–454.
- RSA78        Rivest, R. L., Shamir, A. and Adleman, L., A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM*, 21, pages 120–126.
- Sha80        Sharma, R. S., Clothing behaviour, personality and values: A correlational study. *Psychological Studies*, 25,2(1980), pages 137–142.
- SSD11        Sandisk 2011 financial analysts day presentation, 2011. <http://url.ca/5b0q6>. [05.10.2011]
- TB09        Tews, E. and Beck, M., Practical attacks against wep and wpa. *Proceedings of the second ACM conference on Wireless network security, WiSec '09*, New York, NY, USA, 2009, ACM, pages 79–86.
- Thea        The Institute of Electrical and Electronic Engineers, Inc., IEEE 802.11: Ethernet. <http://standards.ieee.org/about/get/802/802.3.html>. [12.09.2011]
- Theb        The Institute of Electrical and Electronic Engineers, Inc., IEEE 802.11: Wireless local area networks. <http://standards.ieee.org/about/get/802/802.11.html>. [12.09.2011]
- THR08        Threefish, 2008. <http://www.schneier.com/threefish.html>. [01.01.2012]
- TILL06        Tapia, E. M., Intille, S. S., Lopez, L. and Larson, K., The design of a portable kit of wireless sensors for naturalistic data collection. *in Proceedings of PERVASIVE 2006*. Springer-Verlag, 2006, pages 117–134.

- Wal00 Walker, J., Unsafe at any key size; an analysis of the wep encapsulation. Technical Report, IEEE P802.11, 2000.
- WCSS66 Webb, E. J., Campbell, D. T., Schwartz, R. D. and Sechrest, L., *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Rand McNally, 1966.

## Appendix 1. Reference Hardware for BOB

Here is the list of hardware components that we used to assemble the BOBs for our experiment:

- MicroATX cabinet
- AMD Athlon II X2 240e 2.8 GHz dual-core processor, with 2.0 MB cache
- ASUS M4N68T-M LE V2 motherboard, with NVIDIA GeForce 7025 / nForce 630a chipset
- 1 GB 1333 MHz DDR3 memory
- ASUS EN 210 Silent graphics card, with NVIDIA GeForce 210 GPU, 512 MB DDR3 video memory, and DVI, VGA and HDMI connectors
- 1.5 TB 3.5 inch 5400 RPM serial ATA (SATA) hard disk
- DVD disc drive
- DVB-C PCI card based on the Mantis VP-2040 chipset
- Remote control handset
- A serial port adapter for the remote control receiver<sup>26</sup>
- Ethernet PCI card
- Belkin N150 Surf wireless router
- D-Link DCS 2121 wireless network cameras
- Bluetooth USB dongle
- IEEE 802.11bg USB dongle

---

<sup>26</sup><http://www.lirc.org/receivers.html>

## Appendix 2. Categorization of Websites

Figure 17 shows the percentage of HTTP requests made to different categories of websites by each test subject.

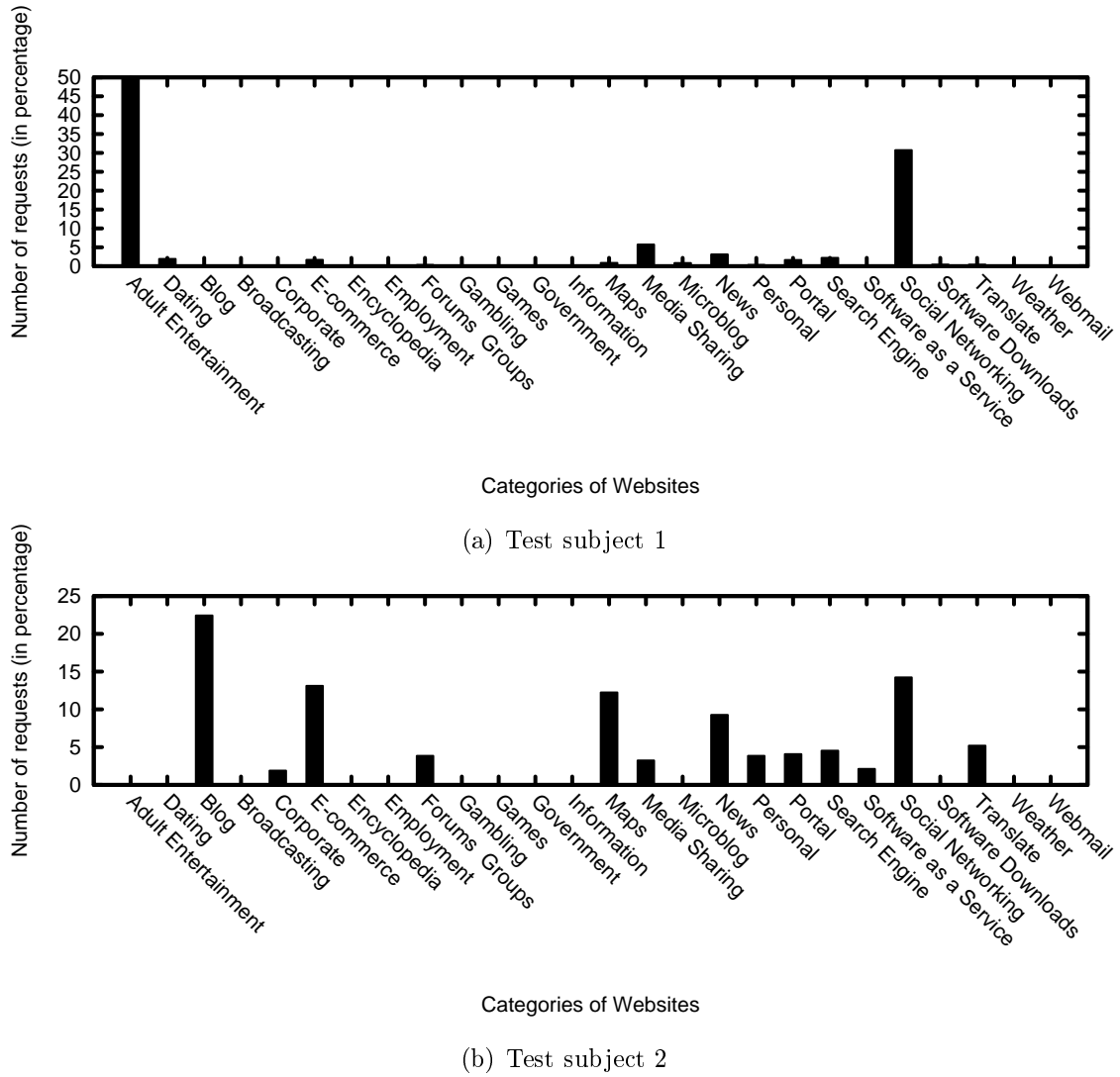
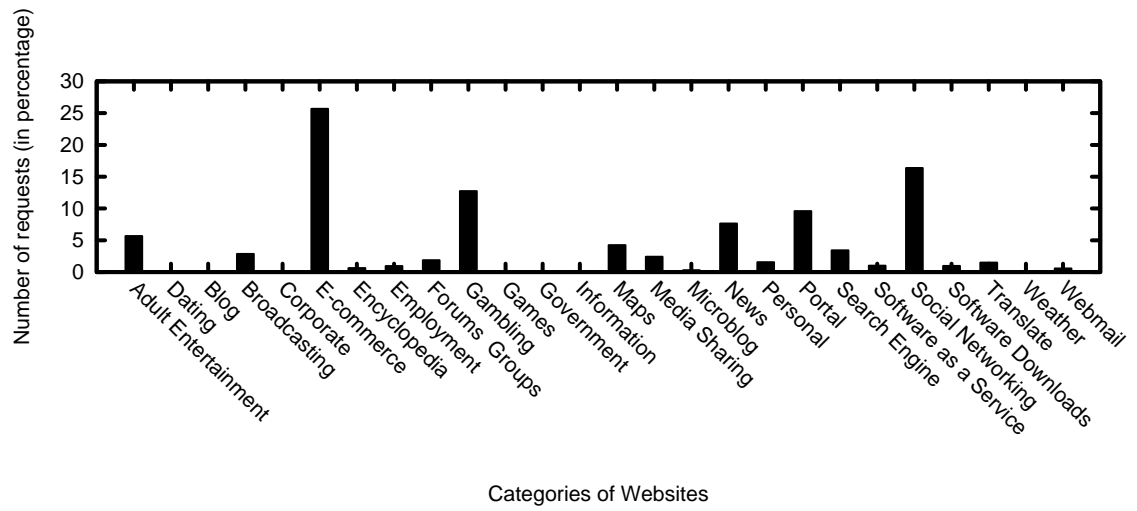
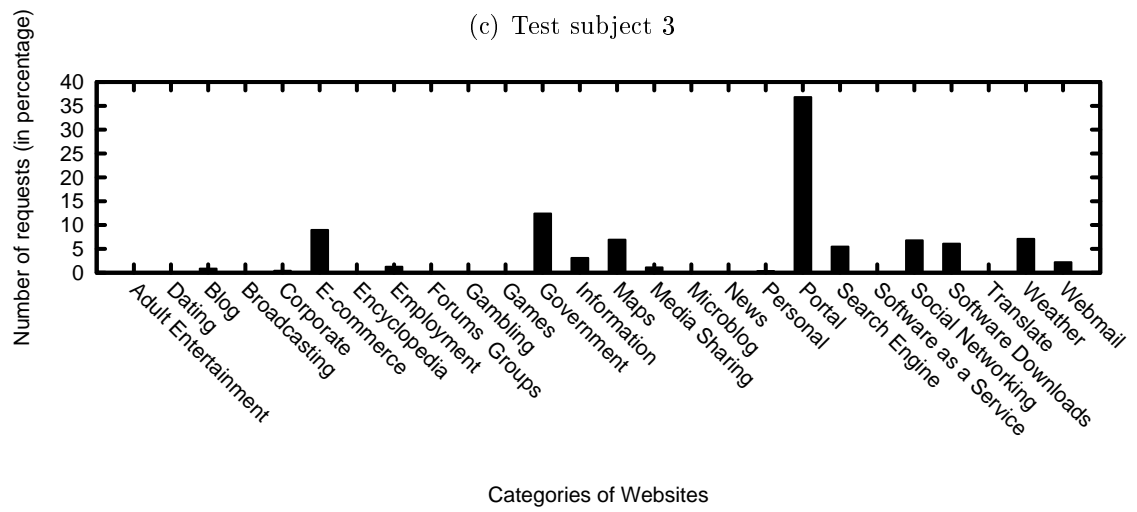


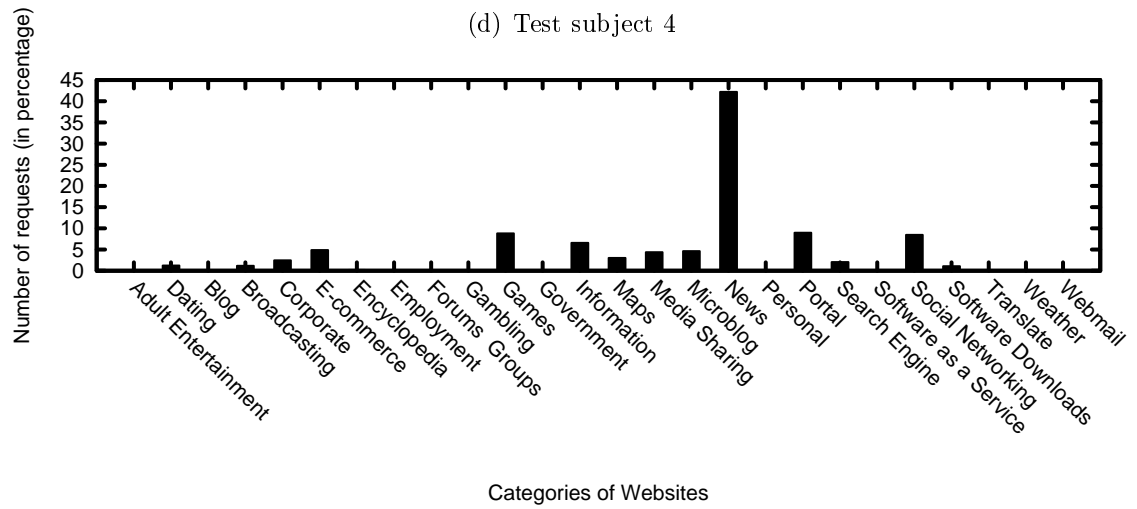
Figure 17: Distribution of requests made to different categories of websites.



(c) Test subject 3

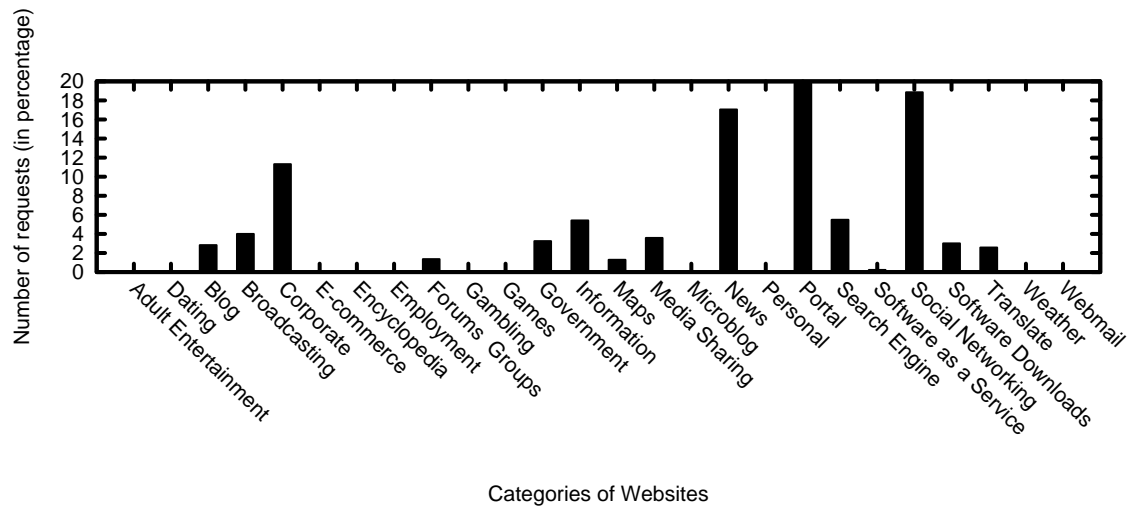


(d) Test subject 4

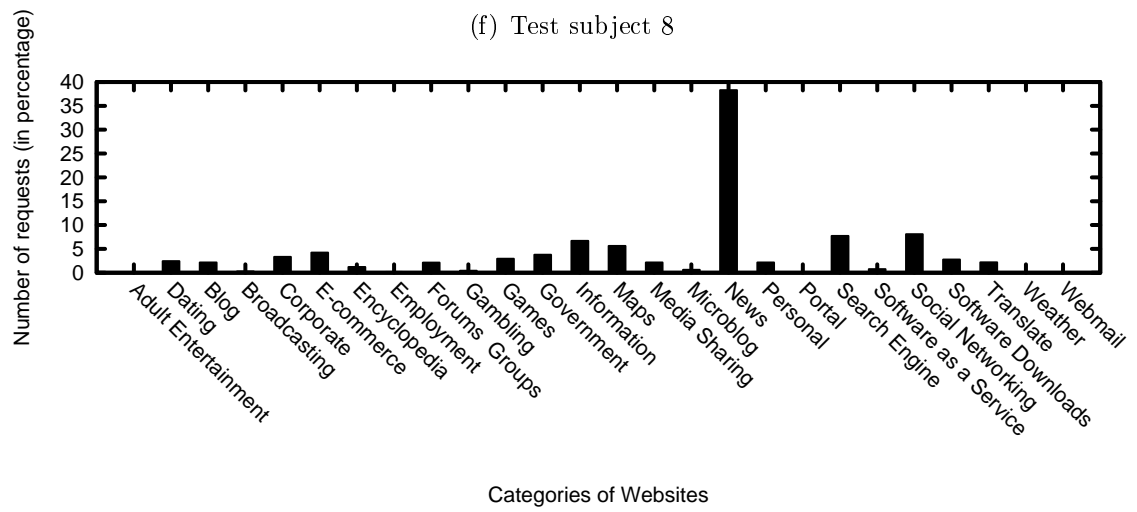


(e) Test subject 5

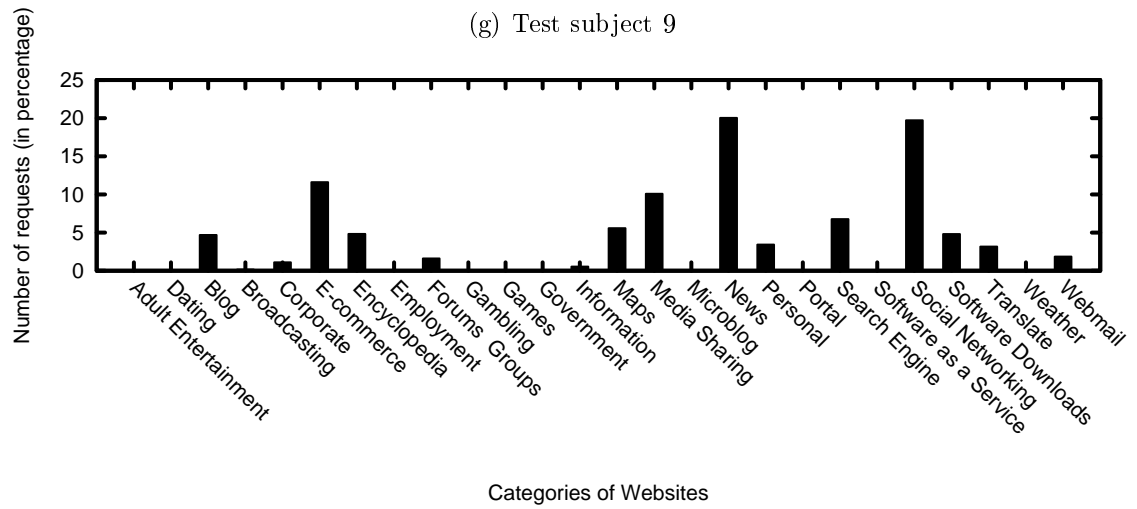
Figure 17: Distribution of requests made to different categories of websites.



(f) Test subject 8



(g) Test subject 9

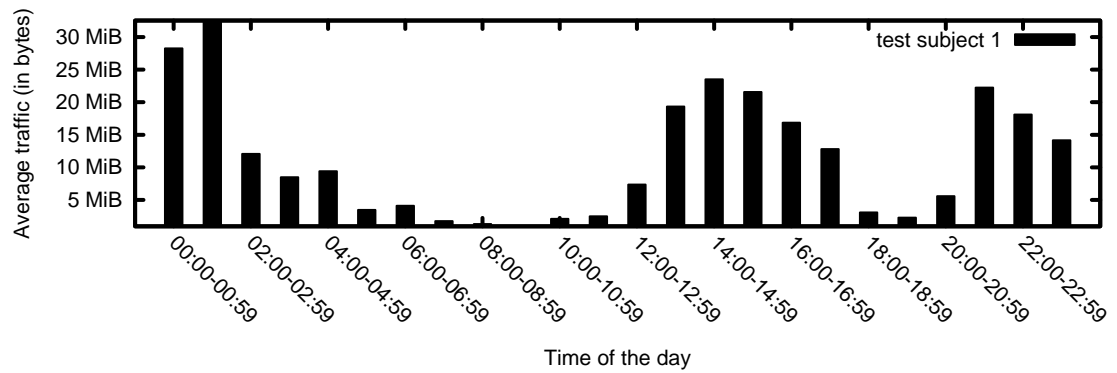


(h) Test subject 10

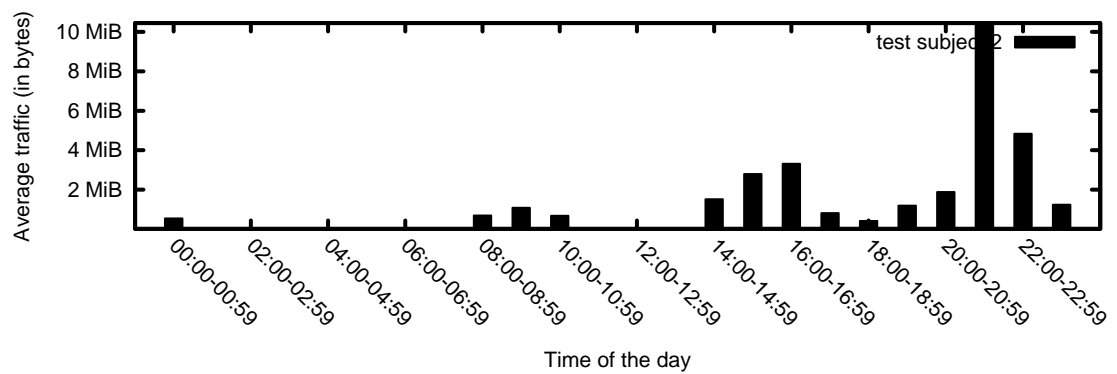
Figure 17: Distribution of requests made to different categories of websites.

## Appendix 3. Daily Traffic

Figure 18 shows the amount of Internet traffic during different intervals in a day for each test subject.



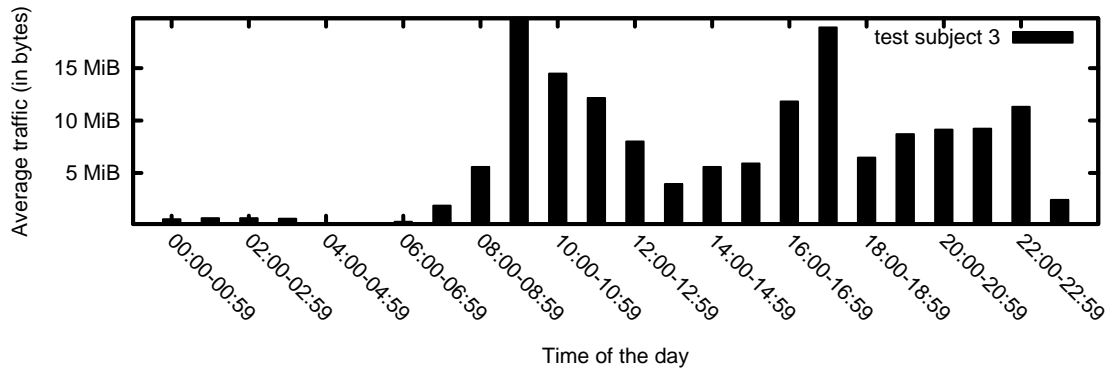
(a) Test subject 1



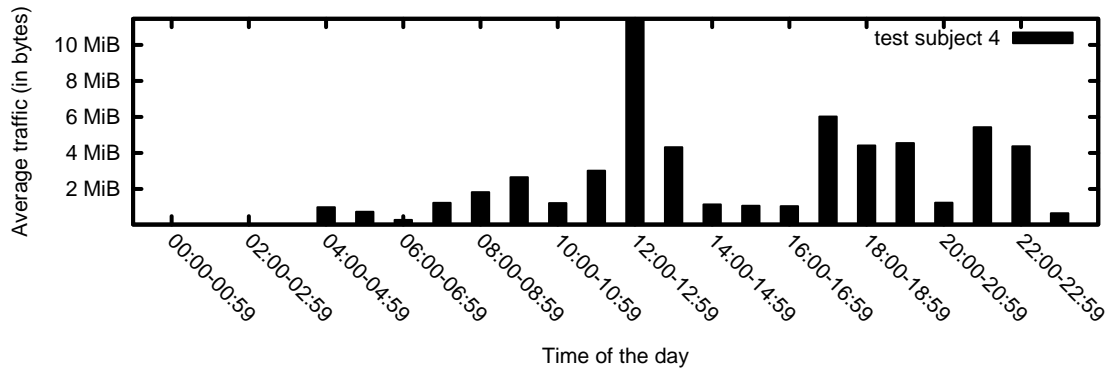
(b) Test subject 2

Figure 18: Daily traffic in bytes.

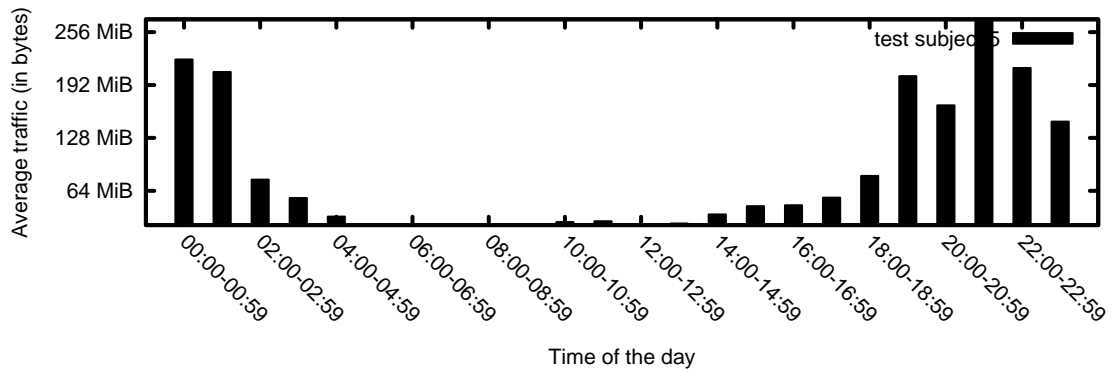




(c) Test subject 3

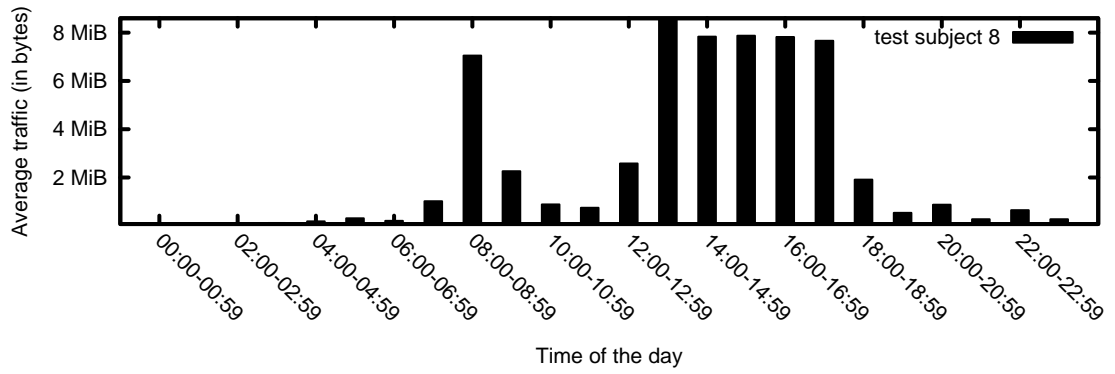


(d) Test subject 4

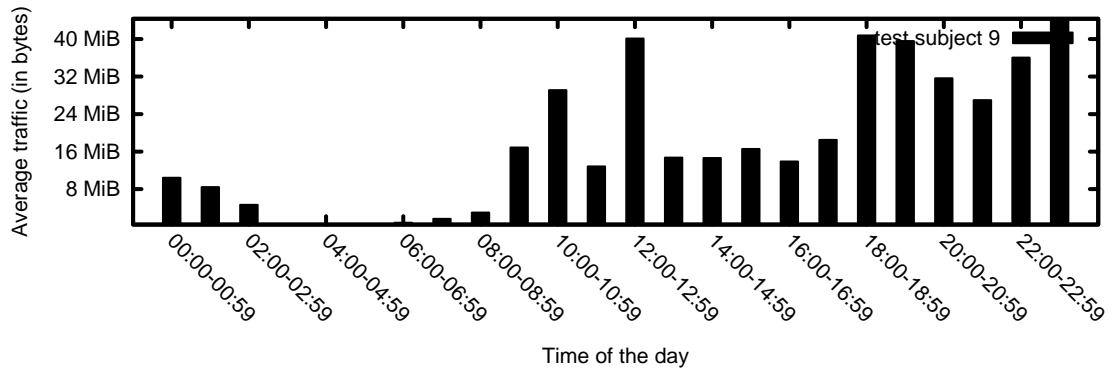


(e) Test subject 5

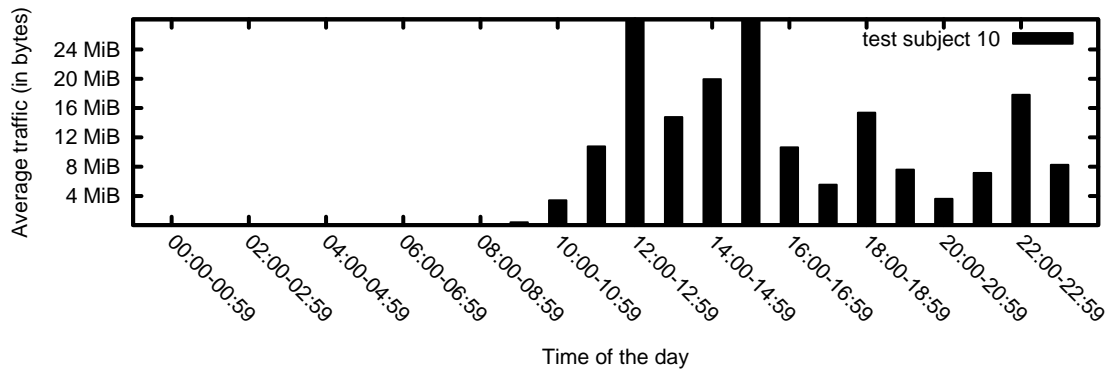
Figure 18: Daily traffic in bytes.



(f) Test subject 8



(g) Test subject 9

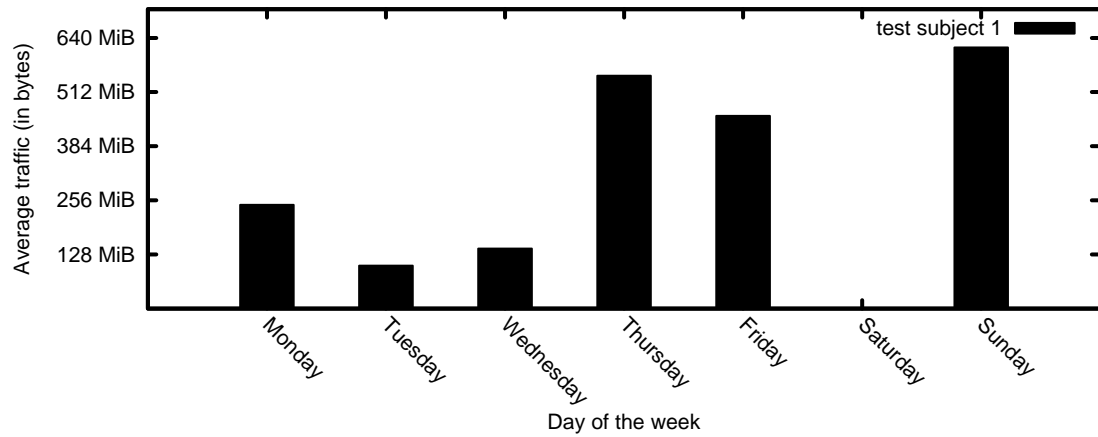


(h) Test subject 10

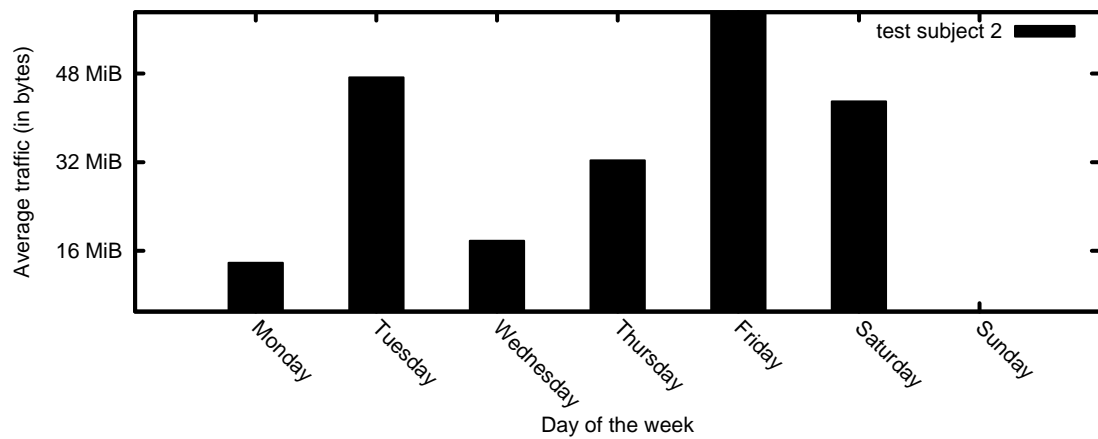
Figure 18: Daily traffic in bytes.

## Appendix 4. Weekly Traffic

Figure 19 shows the amount of Internet traffic during different days of the week for each test subject.

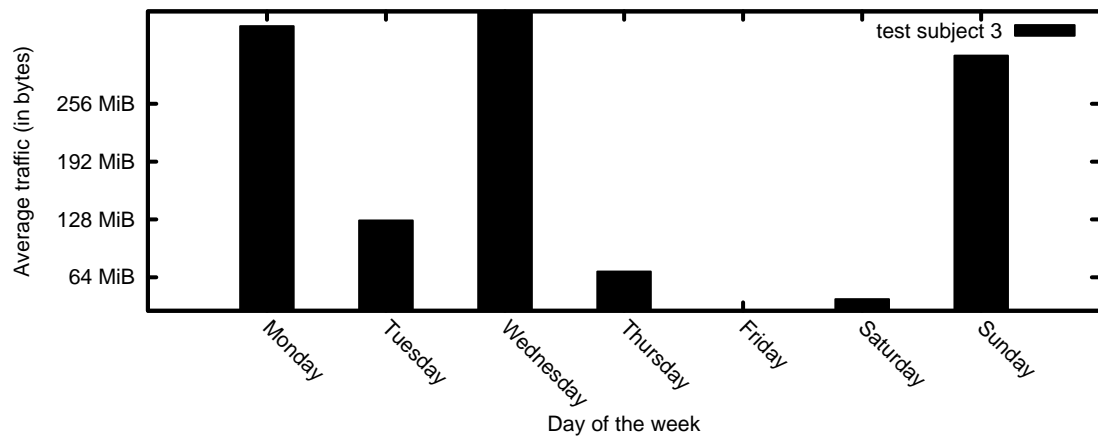


(a) Test subject 1

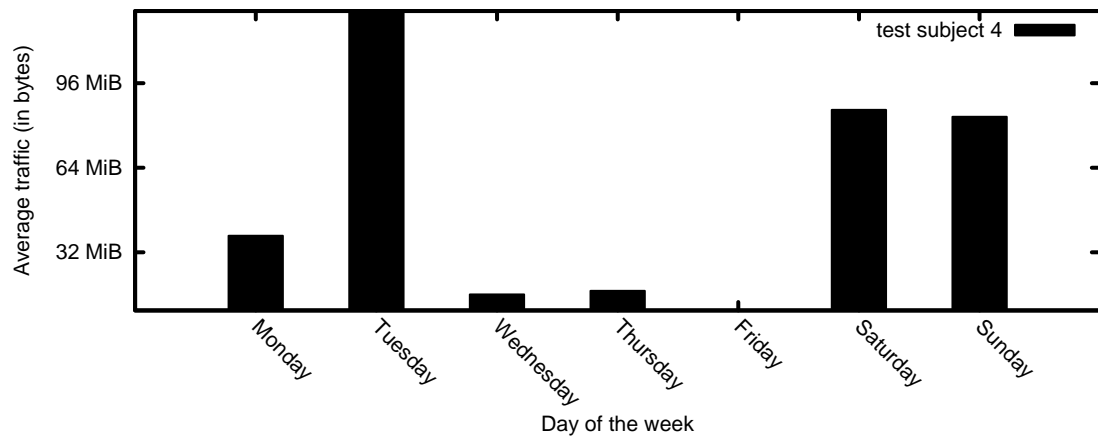


(b) Test subject 2

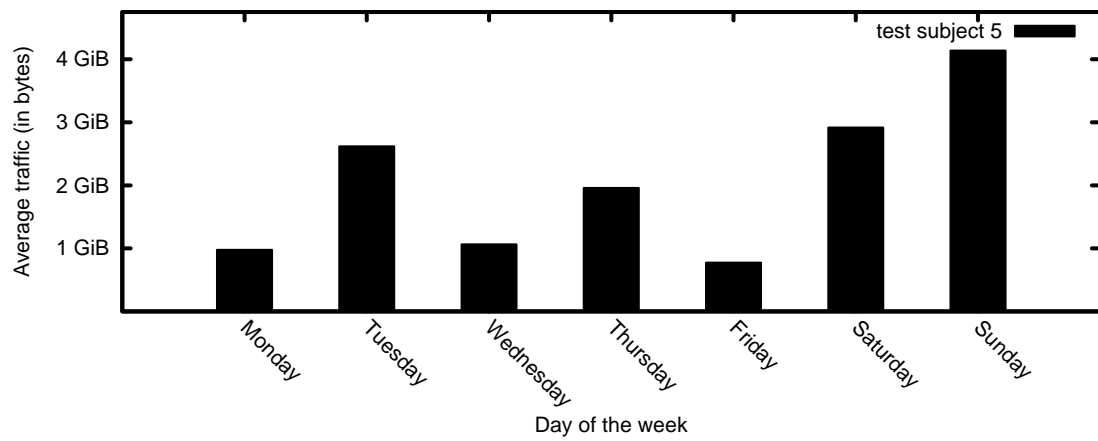
Figure 19: Weekly traffic in bytes.



(c) Test subject 3

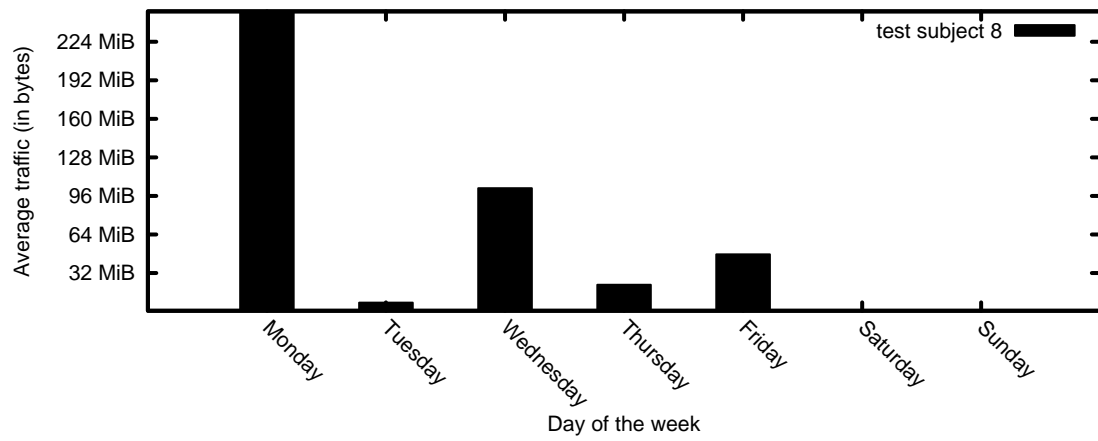


(d) Test subject 4

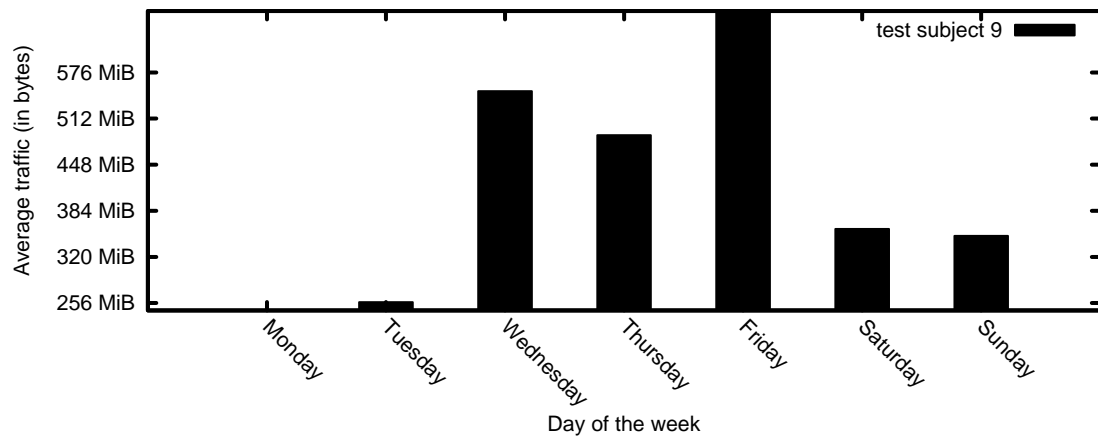


(e) Test subject 5

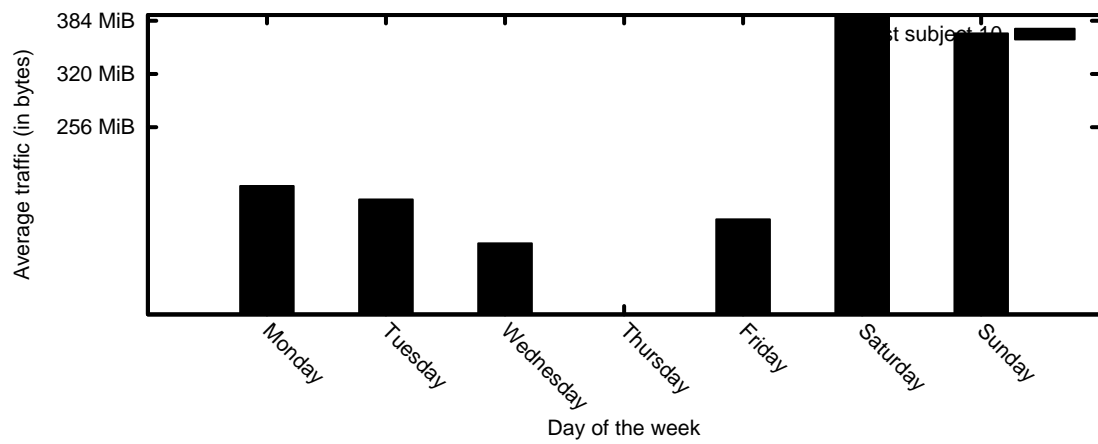
Figure 19: Weekly traffic in bytes.



(f) Test subject 8



(g) Test subject 9



(h) Test subject 10

Figure 19: Weekly traffic in bytes.