HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Industrial Engineering and Management

Antero Kivi

MOBILE INTERNET USAGE MEASUREMENTS – CASE FINLAND

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Engineering

Espoo, 24.4.2006

Supervisor: Professor Heikki Hämmäinen
Instructor: Professor Heikki Hämmäinen

HELSINKI UNIVERSITY OF TECHNOLOGY   ABSTRACT OF MASTER’S THESIS
Department of Industrial Engineering and Management

Author:	Antero Kivi		
Title:	Mobile Internet Usage Measurements – Case Finland		
Number of pages:	8+128	**Date:**	24.4.2006
Library location:	TU	**Publishing language:**	English
Professorship:	S-38 Networking Technology		
Supervisor:	Professor Heikki Hämmäinen		
Instructor:	Professor Heikki Hämmäinen		

Mobile Internet is the outcome of two intense and global trends of the recent years: mobile/wireless and the Internet. Despite the potential the hundreds of millions of new mobile devices sold globally each year present, little information on mobile data service usage apart from mobile operator portals is currently available.

In this research, mobile Internet usage was measured in fall 2005 using three fundamentally different methods to answer the question *What are the characteristics of consumer mobile Internet usage in Finland?* First, data on 80-90% of all Finnish mobile subscribers and terminals was collected with mobile operators’ charging-oriented reporting systems. The observed Finnish mobile terminal installed base was old and did not widely support key features for data usage as e.g. packet data capability was in 48% and WCDMA capability in less than 1% of terminals. Nokia’s market share was a remarkable 87%. Smartphones constituted 6% of all terminals, over 99% of which were Nokia’s Symbian handsets and one third of these Nokia communicators. The terminal base was fairly concentrated as the 50 most common models made up 88% of all terminals. Some 92-94% of all mobile subscribers were postpaid subscribers, 75% of them consumers. While 99% of consumers had operators’ default usage-based packet data tariff plan, the remaining 1% created 82% of all consumer subscriber packet data traffic. Second, 50% of all Finnish mobile network packet data traffic was captured in TCP/IP header collection –based measurements. Strikingly, the Windows operating system originated 65% of all packet data traffic in mobile networks. Moreover, VPN usage created 46% of traffic volume leading to a very high 85% share of UDP traffic. The Internet APN accounted for 90% of all packet data traffic. Third, a panel of 500 Finnish Symbian S60 handsets was monitored with software installed in the handsets. Panelists with higher radio capability handsets used packet data more frequently and in higher volumes. Data usage volumes were also higher for users with relatively cheaper fixed fee packet data plans. Operator sites and infotainment dominated web/wap site visits with 32% and 33% shares of all visits. Using handset as a modem formed a 21-25% part of all smartphone data traffic. The most active 20% of data users created 80% of traffic, even when modem traffic was excluded. Browsing was the most important data application area with a 72% share of non-modem traffic, and its relative share increased with data usage volume.

In conclusion, Finnish mobile data usage is currently business driven. Traffic to non-operator controlled sites appears to be important. The usage of 3G terminals and effectively flat-rate packet data tariffs seems to increase data usage considerably, and browser is a central application also in mobiles. Mobile operators are recommended to include items on off portal traffic to their regular reporting. Similar measurements enabling evaluation of the development of Finnish mobile data usage should be repeated. The measurement methods could also be productized or sold to operators as a service by a 3rd party. Potential ways to utilize the handset-based data are numerous.

Keywords: Mobile Internet, mobile data service, off portal, mobile consumer behavior, smartphone, traffic measurement, mobile terminal base

TEKNILLINEN KORKEAKOULU
Tuotantotalouden osasto

DIPLOMITYÖN TIIVISTELMÄ

Tekijä: **Työn nimi:**	Antero Kivi Mobiilin Internetin käytön mittaukset Suomessa		
Sivumäärä:	8+128	**Päiväys:**	24.4.2006
Työn sijainti	TU	**Julkaisukieli:**	Englanti
Professuuri: **Työn valvoja:** **Työn ohjaaja:**	S-38 Tietoverkkotekniikka Professori Heikki Hämmäinen Professori Heikki Hämmäinen		

Mobiili Internet on seurausta kahden viimeaikaisen voimakkaan ja globaalin trendin, liikkuvuuden ja Internetin yhdistymisestä. Huolimatta satojen miljoonien vuosittain globaalisti myytyjen mobiililaitteiden edustamasta potentiaalista, operaattoreiden portaalien ulkopuolista mobiilien datapalveluiden käyttöä ei tunneta hyvin.

Tässä tutkimuksessa mitattiin mobiilin Internetin käyttöä syksyllä 2005 käyttäen kolmea erilaista menetelmää vastattaessa kysymykseen *Mitkä ovat mobiilin Internetin kuluttajakäytön ominaispiirteet Suomessa?* Ensimmäiseksi, mobiilioperaattoreiden laskutukseen perustuvien raportointijärjestelmien avulla kerättiin aineistoa 80-90% Suomen mobiilitilaajista ja –päätelaitteista. Suomen päätelaitekanta havaittiin vanhaksi ja datakäytölle keskeiset ominaisuudet rajallisesti levinneiksi, sillä pakettidata-kyvykkyys oli 48% ja WCDMA kyvykkyys vain 1% päätelaitteista. Nokian markkina-osuus oli huomattava 87%. Kaikista päätelaitteista älypuhelimia oli 6%, joista 99% oli Nokian Symbian puhelimia ja näistä kolmannes Nokian kommunikaattoreita. Päätelaitekanta oli varsin keskittynyt 50 yleisimmän mallin vastatessa 88% kaikista päätelaitteista. Mobiilitilaajista 92-94% oli postpaid-tilaajia, 75% heistä kuluttajia. Vaikka 99% kuluttajista oli oletusarvoisen käyttöperusteisen pakettidatahinnoittelun piirissä, loput 1% loi 82% kuluttajatilaajien pakettidataliikenteestä. Toiseksi, 50% suomalaisten mobiiliverkkojen pakettidataliikenteestä mitattiin keräämällä TCP/IP otsakkeita. Silmäänpistävin tulos oli Windows-käyttöjärjestelmän 65% osuus kaikesta pakettidataliikenteestä. VPN-käyttö loi 46% liikennevolyymistä johtaen UDP-liikenteen hyvin korkeaan 85% osuuteen. Internet APN vastasi 90% kaikesta pakettidataliikenteestä. Kolmanneksi, 500 suomalaisen Symbian S60 puhelimen paneelia monitoroitiin puhelimiin asennetulla sovelluksella. Radioltaan kyvykkäämpiä puhelinten käyttäjät käyttivät pakettidataa useammin ja enemmän. Datakäytön volyymit olivat korkeampia suhteellisesti halvemman kiinteän hinnoittelun käyttäjillä. Operaattoreiden sivustot ja tietoviihde hallitsisivat web/wap käyttöä 32% ja 33% osuuksilla kaikista vierailuista. Modeemikäyttö muodosti 21-25% osuuden kaikesta älypuhelinten dataliikenteestä. Aktiivisin 20% datakäyttäjistä loi 80% liikenteestä, myös kun modeemikäyttöä ei huomioitu. Selainkäyttö oli tärkein datasovellusalue 72% osuudella ei-modeemi-liikenteestä, sen suhteellinen osuus kasvoi datakäytön volyymin myötä.

Suomalainen mobiilidatakäyttö on tällä hetkellä yrityskäyttäjävetoista. Operaattoreiden portaalien ulkopuolinen liikenne vaikuttaa tärkeältä. Kolmannen sukupolven päätelaitteiden ja kiinteän hinnoittelun käyttö näyttää kasvattavan datakäyttöä merkittävästi, ja selain on keskeinen sovellus myös mobiilissa. Mobiilioperaattoreiden tulisi liittää portaaliliikenteen ulkopuolinen käyttö sisäiseen raportointiinsa. Vastaavat mittaukset tulisi toistaa mobiilidatakäytön kehittymisen arvioimiseksi. Käytettyjen mittausmenetelmien tuotteistaminen tai myynti palveluna operaattoreille kolmannen osapuolen toimesta on mahdollista. Päätelaitepohjaisella aineistolla on lukuisia käyttökohteita.

Avainsanat: Mobiili Internet, mobiili datapalvelu, off portal, mobiilikuluttaja-käyttäytyminen, älypuhelin, liikennemittaus, päätelaitekanta

Preface

This work was carried out at the Networking laboratory at the Helsinki University of Technology as a part of the national LEAD project. Without the contribution of the project's industrial partners this research would not have been possible.

I wish to express my gratitude to the people that have supported me in this work. First and foremost, I want to thank Professor Heikki Hämmäinen for the guidance and insights he has provided throughout the research process. In addition, I am especially grateful to Markus Peuhkuri for his extensive assistance with the traffic measurements. I would also like to thank Hannu Verkasalo for his help with the handset usage measurements, and the rest of the networking business team for their general feedback. I also wish to thank Juho Simpura for reviewing my text.

Finally, I want to thank my family for their support during my studies.

Espoo, April 24th 2006

Antero Kivi

Table of Contents

List of Figures

List of Tables

Acronyms and Terms

2G	Second generation mobile telephony systems. Characterized by the use of digital transmission enabling much higher capacity than first-generation analog systems. Main 2G systems are GSM, cdmaOne, D-AMPS, and PDC.
2.5G	Generation 2.5. Broadly includes all advanced upgrades for 2G networks. In the GSM context refers to one or all of the following technologies: HSCSD, GPRS, and EDGE.
3G	Third generation mobile telephony systems. Systems specified by the IMT-2000 family of specifications, e.g. UMTS and cdma2000.
3GPP	3rd Generation Partnership Project. A standards organization promoting and developing the UMTS specifications.
APN	Access Point Name
ARPU	Average Revenue Per User. A basic measure used by mobile operators to evaluate the revenue stream from a single subscriber.
BD	Billing Domain
BG	Border Gateway
Bluetooth	An industrial specification for wireless personal area networks. Bluetooth provides short range wireless communications between various devices, such as mobile phones and laptops.
CDR	Charging Data Record (formerly Call Detail Record)
CSD	Circuit Switched Data. The original circuit-switched form of data transmission developed for GSM.
DNS	Domain Name System/Server/Service
EDGE	Enhanced Data Rates for GSM/Global Evolution. Essentially a radio access network upgrade increasing data transmission rates of GPRS (→ Enhanced GPRS / EGPRS) and HSCSD (→ Enhanced CSD / ECSD) by up to threefold. Sometimes called “2.75G”.
FICORA	Finnish Communications Regulatory Authority
GERAN	GSM/EDGE Radio Access Network
GGSN	Gateway GPRS Support Node

GPRS	General Packet Radio System. Packet-switched form of data transmission (bearer service) developed for GSM. Also refers to the packet-switched core network used in both GSM and UMTS mobile networks. GPRS sometimes also refers to packet-switched data transmission over GSM network while EGPRS implies EDGE is supported. As “GSM data” typically refers to circuit-switched data, the terms GSM/GPRS, EDGE and W-CDMA are used to imply packet-switched data transmission using the respective bearers in this study.
GSM	Global System for Mobile communication. The most successful and widely used 2G system. Operates at the 900, 1800 and 1900 MHz frequency bands.
GSM bandwidth	GSM terminals are referred to as single band, dualband and triband terminals depending on which GSM frequency bands they support.
HSCSD	High-Speed Circuit-Switched Data. A method for increasing GSM’s circuit-switched data transmission rates.
IMEI	Mobile station Equipment Identity. A unique code identifying each mobile equipment in the GSM/UMTS system
IMS	IP Multimedia Subsystem / IM Subsystem
IMSI	International Mobile Subscriber/Station Identity. A unique code identifying each mobile subscriber in the GSM/UMTS system.
IMT-2000	International Mobile Telecommunications at 2000 MHz
IP	Internet Protocol
MMS	Multimedia Messaging Service. A service evolved from SMS for transmitting not only text but various types of multimedia content (e.g. images, audio, video).
MNO	Mobile Network Operator. A company owning a radio spectrum license and operating a complete mobile network infrastructure. MNOs lease out network capacity to MSOs.
MSO	Mobile Service Operator. A company selling mobile network access to end users. Responsible for e.g. subscriber billing and customer care.
MVNO	Mobile Virtual Network Operator. A term generally referring to all mobile operators not owning a radio spectrum license and, thus, a radio access network. MVNOs can be divided into subcategories based on the range of core network components they own and operate.
NAT	Network Address Translation
OS	Operating System
OTA	Over The Air
PDA	Personal Digital Assistant

PLMN	Public Land Mobile Network
PS	Packet Switched
PSTN	Public Switched Telephone Network
SGSN	Serving GPRS Support Node
SIM	Subscriber Identity Module
SMS	Short Message Service. A service for sending short messages of text between mobile phones.
SPSS	Statistical Package for Social Sciences
TAC	Type Allocation Code. The first 8 digits of the IMEI code identifying the model and origin of the mobile equipment.
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System. 3G mobile network evolved from GSM. Uses the existing GSM core network while introducing a completely new air interface based on W-CDMA.
UTRAN	Universal/UMTS Terrestrial Radio Access Network.
VPN	Virtual Private Network. A private communications network used within an organization or by several different organizations to communicate over a public network.
W-CDMA	Wideband-CDMA (Code Division Multiple Access). Solution for the air interface in 3G systems. W-CDMA is frequently used as a general term to refer to all 3G standards using W-CDMA as its air interface.
W-LAN	Wireless LAN (Local Area Network). Provides a limited-range wireless network connection to users in proximity of an access point using unlicensed frequency bands.
WAP	Wireless Application Protocol

1 Introduction

This chapter provides a background for the research and describes the research questions, objectives and scope of the research. The primary research methods are then introduced. Finally, the structure of the research report is outlined.

1.1 Background of the Research

Mobile Internet is an outcome of the merging of two intense and global trends of the recent years: mobile/wireless and the Internet. Moving the Internet to mobile devices presents plenty of potential for consumers and businesses alike, as the current Internet applications will probably be accompanied by brand-new and yet unimaginable application areas in the mobile environment. For operators and equipment manufacturers, the spread of mobile Internet presents a solution to falling ARPUs and a possible way to avoid or prolong commoditization, not to mention the opportunities it presents to other mobile industry actors such as content and service providers or new entrants from the Internet world.

While the mobile industry and the very concept of mobile Internet suffered some drawbacks in the turn of the millennium in Europe, breakthrough has been achieved elsewhere such as Japan, where the i-mode mobile Internet service has been hugely popular (e.g. Funk 2001). The importance of mobile Internet is inevitably going to rise worldwide, in one form of another, as a growing proportion of the hundreds of millions of mobile handsets sold globally each year will be mobile Internet capable. This belief is also shared by the mobile industry, as several major mobile industry actors have jointly lobbied for the approval of a new top level domain name ".mobi" intended for the easier addressing of mobile Internet services. (mTLD 2006)

In spite of the potential mobile Internet presents, little information on the adoption of mobile data services is currently available. Even less is known about the content and characteristics of mobile Internet traffic. Existing knowledge is primarily extracted from secondary sources, such as mobile sector companies' annual reports, retail statistics, consumer surveys, or industry experts' educated guesses. Naturally commercial actors do have exact information on their proper sales figures or the traffic flowing through their networks, but have little incentive to share that data due to its strategic nature. This situation is prevalent, at least in the case of Finland, which can still be considered to be a somewhat special mobile market despite the recent rise of many Asian markets.

This research is a part of a bigger entity, namely the national LEAD project (Optimal Rules for a Leading Mobile Data Market), which is aimed at supporting the renewal of Finnish telecommunications cluster by improving the understanding on market dynamics in the mobile data business. The participants of the project represent broadly the Finnish mobile industry landscape including the National Technology Agency of Finland (TEKES), Finnish Communications Regulatory Authority (FICORA), Helsinki University of Technology (TKK), Nokia, and three major Finnish mobile operators: TeliaSonera, Elisa and DNA. (LEAD 2006) The cooperative participation of the influential industry actors gives access to a very representative set of data for the Finnish market. Thus, by publicizing market-wide knowledge on the current usage and drivers of mobile Internet this research contributes both to the body of knowledge and lays the ground for new value creation opportunities for businesses.

1.2 Research Questions

The Finnish mobile data service market has not developed in pace with the leading markets, such as Japan and South Korea (Minges 2004). In order to be able to develop the market, current data service usage patterns must be better understood. While the operators know the usage levels of data services provided in operator portals, general usage patterns of "off portal" mobile Internet traffic to outside services on a market-wide scope are relatively unknown to both operators and academia. This research problem is narrowed down to a specific research question, which is further divided into four sub-questions.

The **research question** and the associated sub-questions are:

Q: What are the characteristics of consumer mobile Internet usage in Finland?

- *a) What is the profile of the Finnish mobile terminal installed base and consumer mobile packet data subscriber population?*
- *b) What is the profile of mobile Internet traffic generated by Finnish consumer mobile subscribers?*
- *c) What are the characteristics of mobile Internet usage by Finnish smartphone using consumer mobile subscribers?*
- *d) What are the current drivers of consumer mobile Internet usage in Finland?*

1.3 Objectives of the Research

The research questions will be answered by achieving the corresponding objectives of the research. The **objectives of the research** are:

O_1: *Provide a description and an evaluation of alternative methods to measure mobile Internet usage, and document the measurement processes used in adequate detail.*

O_2: *Provide statistics on the active mobile terminal base observed in Finnish mobile networks, and of the mobile subscriber population of major Finnish mobile operators.*

O_3: *Provide statistics on the volumes and content of mobile Internet traffic generated by Finnish consumer mobile subscribers.*

O_4: *Provide statistics on packet data traffic –related handset usage patterns of Finnish smartphone using consumer mobile subscribers.*

O_5: *Identify the current drivers of mobile Internet usage in Finland based on the obtained statistics.*

In order to better tackle the fifth objective (and the associated research sub-question *d*) four **hypotheses** on possible drivers of mobile Internet usage are used:

Having a handset enabling higher data transmission rates (e.g. 3G phones) facilitates the use of data services over the network. Does the increased ease of use increase usage? The first hypothesis is:

H_1: Users of handsets with higher data transmission capability use more data services.

It has been suggested that the data traffic profile of mobile subscribers with fixed or flat fee pricing on data transmission would start to resemble Internet traffic profile instead of the WAP traffic profile observed with fees based on purely usage volume. While this suggestion is hard to verify with the current means, a more general level hypothesis founded on e.g. quantity discounts can be made. The second hypothesis is:

H_2: Having (effectively) flat rate data pricing increases data service usage volume.

There are some behavioral aspects regarding technology and demographics that can practically be considered as general knowledge. First, the fact that men are more prone to using new technology than women is verified by e.g. gender distributions of engineering school students (e.g. Salokangas 2002). Similarly, it is fairly reasonable to state that generally young people try out new technologies more than old people. The validity of these two basic hypotheses concerning the effect of gender and age on smartphone usage will be tested in this study. Thus, hypotheses 3 and 4 are as follows:

H_3: Men are more active in using their smartphone handsets than women.

H_4: Young people are more active in using their smartphone handsets than old people.

1.4 Scope of the Research

While everybody seems to know what mobile Internet means, the very term “mobile Internet” is rather vaguely defined. For this reason, research on mobile Internet is very diverse (see chapter 2.4 Previous Mobile Internet Usage Research). Thus, a **definition of mobile Internet** must be provided for the purposes of this study.

According to Merriam-Webster’s Collegiate Dictionary:

- *Mobile* refers to something “capable of moving or being moved”.
- *Wireless* is something “having no wire or wires” or more specifically “operating by means of transmitted electromagnetic waves”, which can be related to either “radiotelephony, radiotelegraphy, or radio” or ”data communications using radio waves”.
- *Cordless* refers to “having no cord” and more specifically “powered by a battery”.
- *Cellular* is generally something “relating to, or consisting of cells”. More specifically it refers to something “relating to, or being a radiotelephone system in which a geographical area (as a city) is divided into small sections each served by a transmitter of limited range”.

Thus, for a communications device intended to be mobile, it has to be both wireless and cordless. However, all mobile devices, such as satellite phones or devices with WLAN

(Wireless LAN / Wireless Local Area Network) cards are not necessarily based on cellular technology.

The term *Internet* is not comprehensively defined. The US Federal Networking Council defined the term "Internet" in 1995 as "the global information system that (i) is logically linked together by a globally unique address space based on the Internet Protocol (IP) …; (ii) is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite …, and/or other IP-compatible protocols; and (iii) provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure …" (FNC 1995) The Internet Security Glossary in RFC 2828 defines "Internet" as the single, interconnected, worldwide system of commercial, government, educational, and other computer networks that share the set of protocols specified by the IAB (Internet Architecture Board) and the name and address spaces managed by the ICANN (Internet Corporation for Assigned Names and Numbers). (Shirley 2000)

Thus, a broad definition for mobile Internet would include the use of private, public, or commercial services in the global TCP/IP suite based information system accessed using packet switched data transmission with any mobile communications device. However, in this study the term *mobile Internet* is further narrowed down to refer to the services enabled by the data transmission capabilities of cellular mobile telephone networks used by associated mobile terminal devices to access the services provided outside the domain of mobile network or service operators by 3rd party service providers.

Definitions for mobile terminal devices must also be provided. In general, a mobile terminal refers to all devices capable of connecting to mobile telephone networks. Mobile terminals primarily consist of mobile phones henceforth called *handsets* to distinguish them from other mobile terminal types, data cards, and other modems/modules capable of accessing mobile networks. Furthermore, a smartphone is generally considered any handheld device that integrates personal information management and mobile phone capabilities in the same device. Often this includes adding phone functions to already capable personal digital assistants (PDAs) or putting "smart" capabilities, such as PDA functions, into a mobile phone (handset). Another key feature of a smartphone is that additional applications developed by device manufacturers, operators, or 3rd parties can be installed on the device. However, in this study the term *smartphone* refers only to PDAs and handsets functioning in mobile telephone networks, identified by the operating system (e.g. Symbian, Windows Mobile, Linux Mobile, Palm) used in the terminal in question.

While some parts of the research do include broader takes of the mobile market, the **primary scope** of this study is:

<u>Mobile Internet traffic</u> as defined above, with the following limitations:

- Mobile cellular network technologies are limited to the GSM/UMTS family network technologies used in Finland (see chapter 2.2 GSM/UMTS Mobile Networks).
- Data transmission in these networks is limited to packet data traffic only, thus excluding all circuit-switched data transmission.

- Internet traffic is limited to the “off portal” or “by-pass” traffic going through mobile networks to outside Internet, thus excluding operator-provided services such as SMS, MMS and WAP.

Private consumers, meaning:

- The observed mobile subscribers are private persons, identified by their type of mobile subscription.

Finnish mobile market, meaning one of the following:

- Observations are made in Finnish mobile networks.
- Observations are made from the handsets of Finnish mobile subscribers.

1.5 Research Methods

A **literature study** will be conducted to form an understanding of the underlying technologies and to establish the methods used to measure mobile Internet usage. Moreover, a review of previous research on the research subject will be conducted. In addition, the technical complexity involved in the empiric part of the study necessitates several **expert interviews** for the proper understanding of the measurement conditions.

A **measurement on operator reporting systems**, i.e. data collection based on the participating mobile operators’ charging-oriented reporting systems provides information on the mobile terminal installed base, the packet data subscriber population, and the volumes of packet data and mobile Internet traffic generated by mobile subscribers.

A **mobile network packet data traffic measurement** based on the collection of TCP/IP headers from the GSM/UMTS core network gives access to the content of the traffic, most notably to the used application protocols and the visited web sites. Moreover, additional analysis enables studies on the underlying mobile terminal operating systems.

Handset –based usage measurements are realized with a research platform including a monitoring software installed on voluntary mobile subscribers’ Symbian S60 handsets. The research software observes the usage of applications and handset features, writes usage events to a log file and sends the encrypted files regularly over GPRS/WCDMA to a server for further analysis. Moreover, some background information on the participating users is collected with questionnaires filled out during the registration process and after the study.

1.6 Structure of the Report

The top level structure of the master’s thesis is as follows: After the first, introductory chapter, chapters 2 and 3 describe the background for this study. Chapters from 4 and 5 have an empiric focus whereas chapters 6 and 7 combine the previously known and empirically obtained information. The three fundamentally different measurement methods used impose themselves on the general structure of report, which is presented in the figure below.

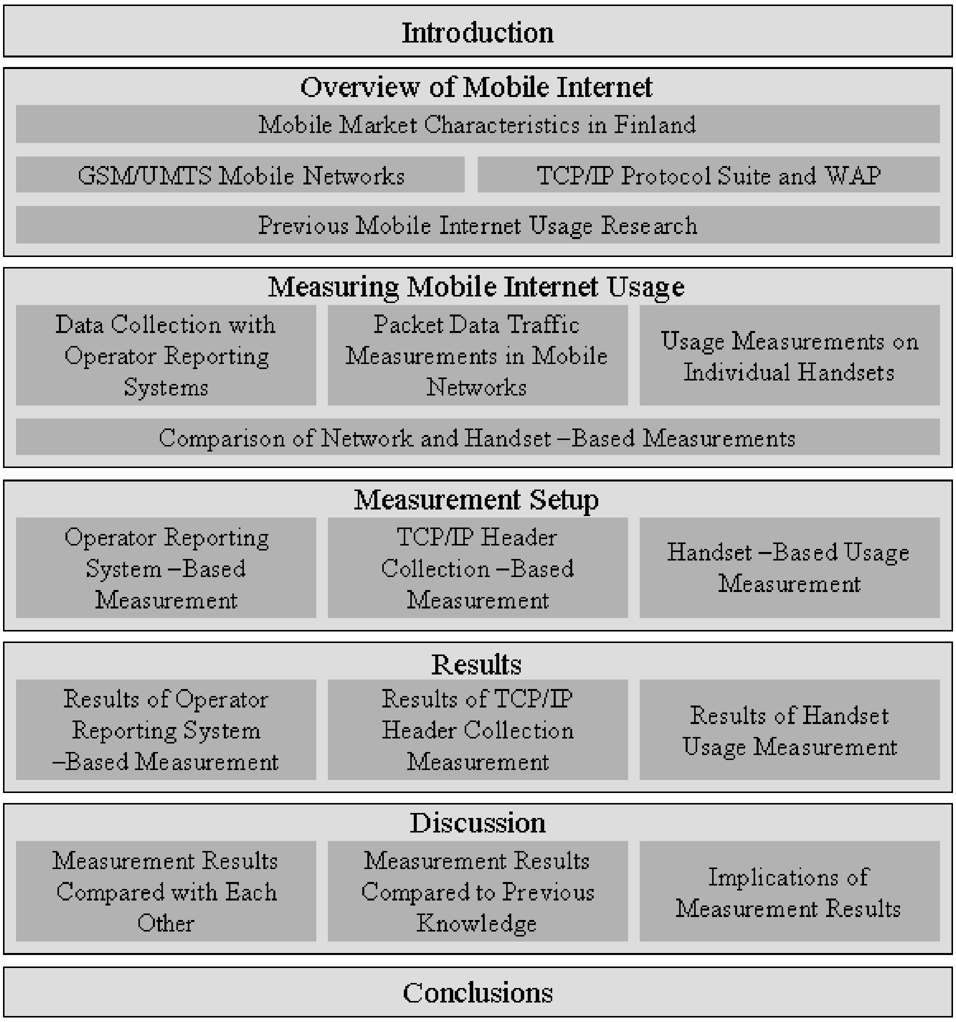


Figure 1 Structure of the master's thesis

Chapter 2 gives an overview on mobile Internet. Some general mobile market characteristics are first treated. Then, a general presentation on GSM/UMTS networks, TCP/IP protocol suite, and WAP is given. Finally, the relevant previous research on the subject of mobile Internet is presented.

Chapter 3 lays out the relevant background issues related to the measurement methods. The chapter describes what kind of usage data can be obtained with different methods and how they should be implemented in theory. A comparison of the measurement methods is also provided.

Chapter 4 describes the setup of the operator reporting system, TCP/IP-header collection, and handset –based measurements. Specifications detailing the specific data items sought in each of measurement are presented before treating the required steps in the actual implementation of the measurements and associated analyses. Finally, a comprehensive analysis on the potential sources of measurement and processing error is presented.

Chapter 5 presents the results from all three measurements separately from each other. Results of operator reporting system –based measurements concentrate on Finnish mobile terminal installed base and Finnish mobile subscriber population. Results of TCP/IP header collection measurements describe some general packet data traffic patterns before treating application protocol traffic and web protocol traffic in more detail. Results of the handset usage measurements describe the sample of people included in the measurement and some general handset usage patterns. Then packet data usage focus is resumed with a description of general packet data usage patterns and packet data generating application usage. Finally, smartphone browsing patterns are studied.

Chapter 6 discusses the measurements by comparing the results with each other and with previous knowledge, before evaluating the implications of the results.

Chapter 7 concludes the research, with a summary of the major findings and some notions on the research reliability and validity. Finally, some tangible recommendations and suggestions for further research are provided.

2 Overview of Mobile Internet

An overview on mobile Internet is presented in this chapter. First, a brief introduction on the mobile market characteristics in Finland is given. Second, a general overview of GSM/UMTS mobile networks is presented. Third, the underlying networking protocols are described. Fourth, the range of previous studies on the research subject is discussed.

2.1 Mobile Market Characteristics in Finland

According to a research of eBird Scandinavia Oy (Snellman 2005) for the Finnish Ministry of Transport and Communications based on interviews of several industry actors, the **size of the Finnish mobile service market** is estimated to be 253 million € in 2005. The size is expected to grow slightly due to growth in content services and data services, whereas the size of person-to-person communications is decreasing. The 29 million € worth of mobile data services represented about 10% of the total market size, with the share of packet-switched services grabbing a dominant 98% of this. The size of mobile data service markets is in rapid 36% growth. According to the annual reports of TeliaSonera and Elisa, the mobile subscriber ARPU (Average monthly Revenue Per User) in Finland has dropped from about 38€ at the end of 2004 to under 29€ at the end of 2005. More detailed data service ARPUs are not publicly available.

One particularity of the Finnish mobile market is the very strong presence of **Nokia Corporation** and 40% of its employees in Finland, due to Nokia's Finnish origins. Regarding **mobile operators**, the market has a few major actors. TeliaSonera, Elisa, and Finnet (or DNA Finland) operate the three Finnish GSM/UMTS networks, each of which providing nation-wide GSM/GPRS, and limited EDGE and UMTS coverage. Sonera, Elisa, and DNA are also the major service operators in the market, in addition to the MVNOs TeleFinland (part of TeliaSonera) and Saunalahti (part of Elisa). The operators also have brand operators Kolumbus (Elisa) and ZeroForty (TeliaSonera) with visibility on the market. Moreover, Ålands Mobiltelefon Ab is a MNO owning regional licenses to operate both GSM- and UMTS-networks on the Åland islands. According to a survey-based research made by Otantatutkimus Oy for Ficora (Otantatutkimus 2005) in November – December 2005, 98% of the roughly five million Finns have a mobile phone. Three major operators dominate the scene, as the combined market share of TeliaSonera is 50% (Sonera 37%, TeleFinland 13%),while Elisa's (Elisa 23%, Saunalahti 10%, Kolumbus 5%) share is 38%, and DNA's 11%. The remaining operators have a 1% market share.

Packet data usage pricing is managed according to three general schemes in Finland. Purely *usage-based* plan means the subscriber is billed a certain amount of euros for each megabyte of transferred packet data. A *block-based* scheme refers to the situation, where the subscriber can use packet data with a fixed fee up to a certain monthly megabyte limit, i.e. the block size, and pay according to the usage-based plan for the surpassing traffic. A *flat-rate* (or flat fee) plan means that a fixed monthly fee is paid regardless of the amount of traffic generated. The usage-based packet data plan is currently the default plan in all postpaid subscriptions. Prepaid subscribers are typically offered usage-based plans only. Some operators have introduced combinations of the above general tariff categories, and also differentiated the offering by controlling the transmission rates of certain flat-rate plans according to the size of the monthly fee. Ficora has provided a comprehensive table detailing the packet data usage prices of all Finnish operators for December 2005 (Ficora 2005).

2.2 GSM/UMTS Mobile Networks

A general overview of GSM/UMTS networks is given in this chapter. Considering the scope of this study, particular interest is given to the packet switched domain and the GPRS core network. Similarly, basics of charging and billing systems and GPRS roaming are also presented, as are the typical mobile network connections to external packet data networks and mobile devices from the network point of view.

2.2.1 General GSM/UMTS Network Architecture

Generally, a Public Land Mobile Network (PLMN) provides land mobile telecommunications services to the public. A very high-level organization of the GSM/UMTS network is presented in the figure below (3GPP TS 23.002).

Figure 2 A high-level organization of the UMTS network

An UMTS network has one or more **access networks**, using different types of access techniques (3GPP TS 32.101). The network supports two types of cellular radio access networks (RANs), Base Station Subsystem (BSS) for GSM and the Radio Network System (RNS) for UMTS. The BSS is the GSM RAN, which has incorporated EDGE in later standard releases as GERAN (GSM/EDGE Radio Access Network). The RNS is based on Wideband-CDMA radio link, and is also called UTRAN (Universal / UMTS Terrestrial Radio Access Network). (3GPP TS 23.002)

The **core network** is logically divided into a Circuit Switched (CS) domain, a Packet Switched (PS) domain and an IP Multimedia Subsystem (IM subsystem / IMS). The CS domain (respectively PS domain) refers to the set of all the core network entities either offering circuit switched (respectively packet switched) connection for user traffic or supporting the related signaling. These two domains are overlapping, i.e. they contain some common entities. A PLMN can implement only one domain or both domains. The IM subsystem comprises all core network elements for provision of IP multimedia services delivered over the PS domain. (ibid)

There are several **service specific entities**, dedicated to the provisioning of a given set of services. The implementation of these entities in a given PLMN should have limited impact on all the other entities of the PLMN. The service specific entities are mostly located in the core network. (ibid)

The **User Equipment** (UE), or Mobile Station (MS) in GSM vocabulary, consists of the variety of the physical equipment with different levels of functionality used by a PLMN subscriber. (ibid, 3GPP TS 23.101)

2.2.2 Packet Switched Domain and GPRS Core Network

A simplified figure on the GSM/UMTS packet switched domain and GPRS core network is presented below.

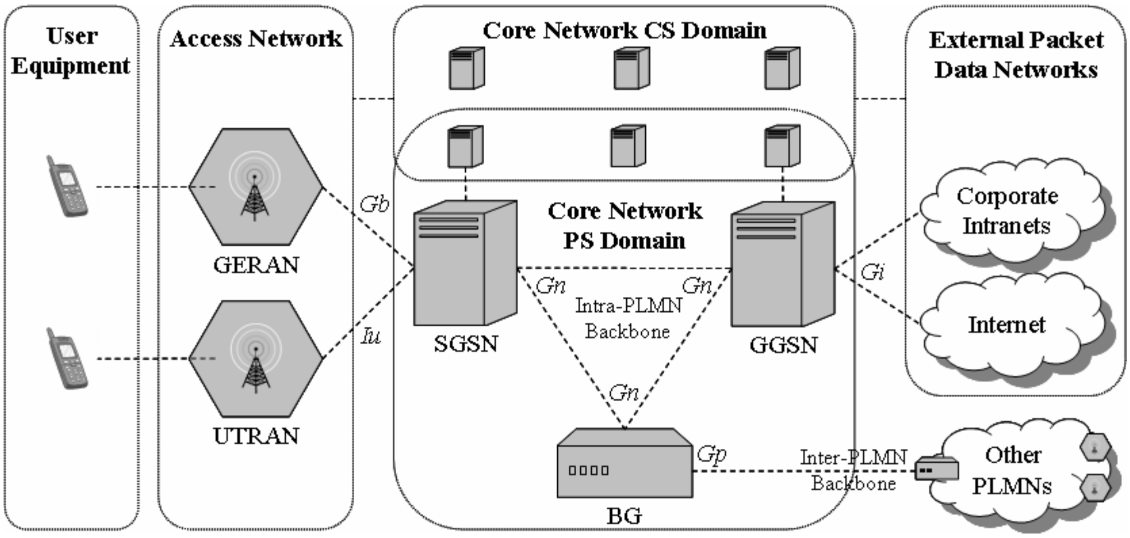


Figure 3 GSM/UMTS packet switched domain and GPRS core network

The **GPRS Support Nodes** (GSN) constitute the interface between the radio system (GERAN and UTRAN) and the fixed networks for packet switched services. The GSNs perform all necessary functions to handle the packet transmission to and from the mobile stations. The **Serving GPRS Support Node** (SGSN) is the core element of a GPRS network responsible for mobility management, session management, tunneling and routing, compression, authentication, encryption, and the generation of charging data records. The **Gateway GPRS Support Node** (GGSN) is responsible for providing gateway access to external packet data networks (PDNs) and also generates charging data records. The GGSN terminates the GPRS Tunneling Protocol (GTP) tunnels and forwards the data packets to destination networks. GGSNs also often include firewall capabilities. (3GPP TS 23.002)

The **Border Gateway** (BG) is a gateway between a PLMN supporting GPRS and an external inter-PLMN backbone network used to interconnect with other PLMNs also supporting GPRS. The role of the BG is to provide the appropriate level of security to protect the PLMN and its subscribers. (ibid)

In order to use GPRS services, the MS must first make its presence known to the network by performing a “GPRS attach” to the SGSN. This makes the MS available for SMS over GPRS, paging via the SGSN, and notification of incoming packet data. In order to send and receive packet data by means of GPRS services, the MS must activate the **Packet Data Protocol context** (PDP context) that it wants to use. This operation makes the MS known in the corresponding GGSN and enables interworking with data networks.

(23.060) The **GPRS Tunneling Protocol** (GTP) is the protocol used between GSNs, and between SGSN and UTRAN. There are three forms of GTP, one for transfer of user data in separated tunnels for each PDP context, one for control reasons (e.g. setting up and deleting PDP contexts), and one for transferring charging data. (3GPP TS 29.060)

Some of the **interfaces** (Gi, Gn, Gp, Gb, Iu) between the various networks and network elements in GSM/UMTS mobile networks are also presented in the figure.

2.2.3 Charging and Billing Systems

Charging is a very important function for PLMN operators. As GSM/UMTS is a global standard with global roaming functionality being in a central role, charging is specified by the 3GPP with considerable detail. For subscriber charging and billing to work properly, appropriate charging information needs to be generated and collected by the network elements of the PLMN and forwarded to appropriate charging and billing systems. Charging data is also used as evidence of customer actions and for statistical analysis of service usage.

The functional requirements for charging are always the same across all domains, services and subsystems. These requirements are defined in the common charging architecture presented in 3GPP TS 32.240. However, this is only a logical view of the actual domain / service / subsystem specific charging architecture. The physical mapping of the logical architecture onto each domain / service / subsystem depends on the case in question. A simplified version of the UMTS logical **charging architecture** is presented in the figure below. (3GPP TS 32.240).

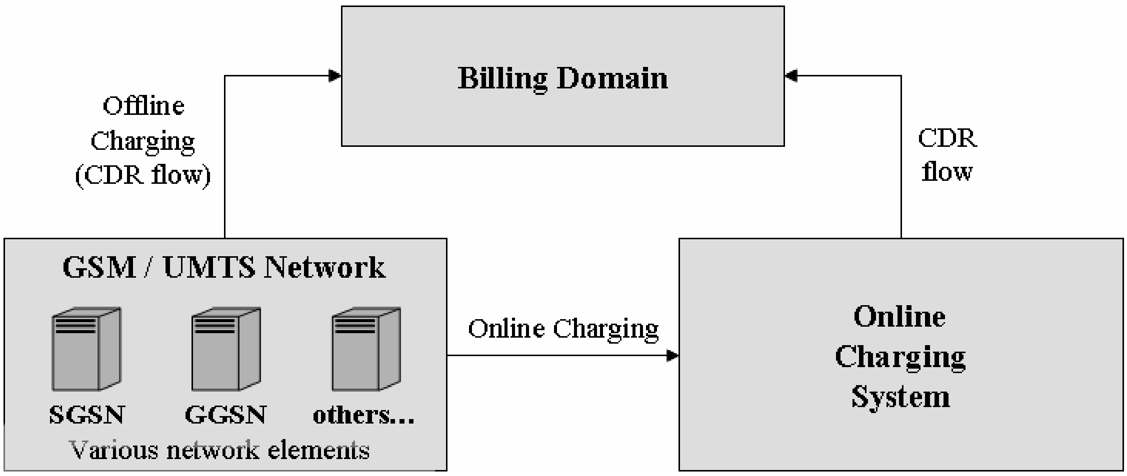


Figure 4 Simplified UMTS logical charging architecture

GSM/UMTS networks provide functions that implement offline and/or online **charging mechanisms** on the bearer (e.g. GPRS), subsystem (e.g. IMS) and service (e.g. MMS) levels. In order to support these charging mechanisms, the network performs real-time monitoring of resource usage, on the above three levels in order to detect the relevant chargeable events. **Offline charging** is a process where charging information for network resource usage is collected concurrently with the resource usage. The charging information is then passed through a chain of logical charging functions. At the end of this process, CDR (Charging Data Record, formerly Call Detail Record) files are generated by the network, and are then transferred to the network operator's Billing Domain (BD) for the purpose of subscriber billing, inter-operator accounting, and/or other applications (e.g. statistics). **Online charging** is a process where charging information for network resource usage is collected concurrently with that resource usage similarly to offline charging. However, authorization for the network resource usage must be obtained by the network prior to the actual resource usage. This authorization is granted by the Online Charging System (OCS) upon request from the network. When receiving a network resource usage request, the network assembles the relevant charging information and generates a charging event towards the OCS in real-time. The OCS then returns an appropriate resource usage authorization based on a query on a prepaid subscriber account in the OCS. The resource usage authorization may be limited in its scope (e.g. volume of data or duration) and therefore the authorization may have to be renewed from time to time as long as the user's network resource usage persists. **Prepaid and postpaid** billing arrangements are usually related to online and offline charging methods, respectively. However, they are not rigorously the same thing as e.g. an operator can have postpaid subscribers on credit control by using online charging mechanisms. (ibid)

Charging Data Record (or "ticket"), is a formatted collection of information about a chargeable event (e.g. time of call set-up, duration of the call, amount of data transferred, user identity, terminal identity …) used in billing and accounting. For each party to be charged of a chargeable event a separate CDR is generated, i.e. more than one CDR may be generated for a single chargeable event. Charging (online and offline) can be categorized into two distinct **charging classes**, namely event based charging and session based charging. Event based charging implies that a chargeable event is defined as a single end-user-to-network transaction, e.g. the sending of a multimedia message. This chargeable event is then mapped to an appropriate charging event, resulting in a single CDR in offline charging or in a single credit control and resource usage authorization procedure in online charging. In contrast, session based charging is characterized by the existence of a user session, such as a circuit call, or a GPRS PDP context. This user session is then matched by a charging session, resulting in the generation of multiple charging events and the creation of one or more CDRs in offline charging or the performance of a credit control session in online charging. (ibid)

Most of the internal functions of the BD and the OCS are outside the scope of 3GPP standardization with only the reference points for charging information transfer from the network to BD and OCS being parts of the 3GPP standards. (32.240) Thus, **billing systems** are typically not provided by major telecommunications infrastructure vendors, but by special billing system manufacturers. Nevertheless, billing is a very important issue for operators and service providers as it is often the most important contact point with the customer. In principle, the billing process is simple. The billing system gets the CDRs from the network, calculates the cost of service used based on current tariffs, and dispatches the bill statements to customers. (Korhonen 2001)

2.2.4 GPRS Roaming

GPRS networks can support **two roaming scenarios** called “home GGSN roaming” and “visited GGSN roaming”. While roaming, the MS always performs a GPRS attach to the SGSN in the visited network (VSGSN) to register to the network. If this is the initial attachment in the new PLMN, then the VSGSN communicates with the Home PLMN (HPLMN) to validate whether the user is allowed to roam or not, and relays this to the MS. For GPRS data transfer and access to external packet data networks the user must then perform the PDP context activation with a GGSN. This procedure, together with the data sent by the MS and the subscribed data stored in the VSGSN, will determine which roaming scenario the user will activate. The two roaming scenarios are illustrated in the figure below.

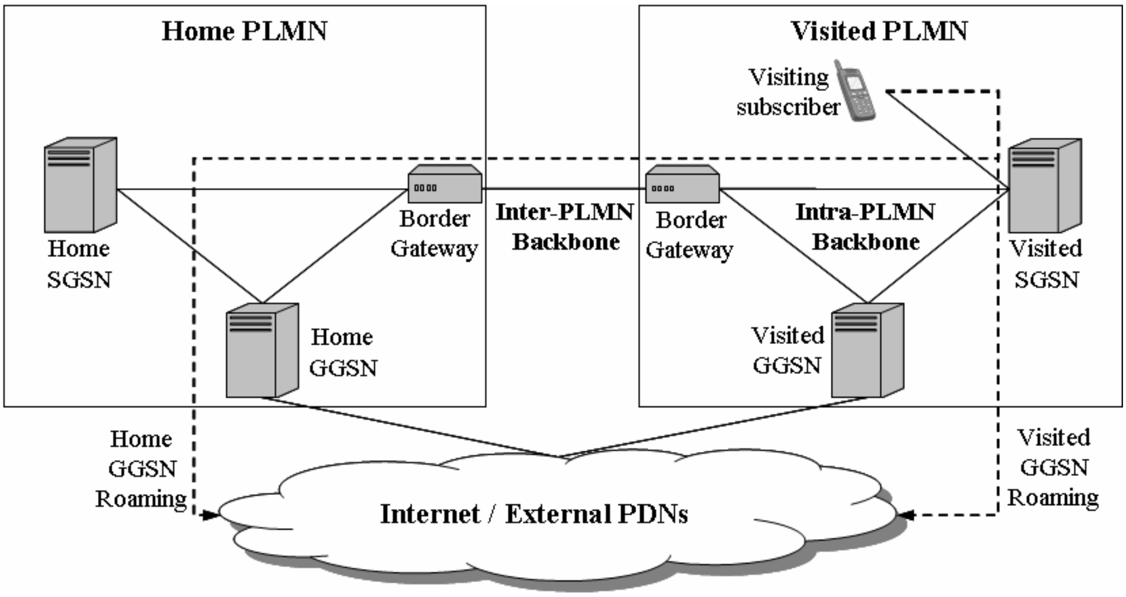


Figure 5 GPRS roaming scenarios

In **home GGSN roaming** (or PLMN roaming) the roaming user registers to the VPLMN using the VSGSN and then activates a context using the HGGSN. There will be data and signaling exchanges via the border gateways across the Inter PLMN Backbone in order to establish the context. Once the context is active, user traffic is transported to the HPLMN first and then routed to external networks via the HGGSN. Thus, both the HGGSN and the VSGSN create CDRs.

In **visited GGSN roaming** (or Internet roaming) the roaming user registers to the VPLMN using the VSGSN, similar to the previous scenario. However, for context activation the VGGSN is used. There will be no data and signaling exchanges across the Inter PLMN Backbone, as the context is established within the VPLMN and user traffic is routed directly to external networks through the VGGSN. Thus, the VPLMN has complete control over the call and generates all CDRs, similarly to the circuit switched GSM roaming. (GSMA 2005a)

Transferred Account Procedure (TAP) is the process that allows the VPLMN to send billing records of roaming subscribers to their HPLMNs. The CDRs generated by the

VPLMN are transferred on a regular basis to the billing system of the VPLMN for pricing or rating. The CDRs produced on behalf of roaming subscribers are converted and grouped in files under the TAP format. The transfer of TAP records between the VPLMN and the HPLMN may be performed directly, or more commonly, via a Clearinghouse. This might include some delay for the roaming information to reach the HPLMN. Invoicing between the operators typically happens once per month. At the HPLMN the TAP record is converted into an internal format and added together with any locally produced CDRs. (GSMA 2005b)

2.2.5 External Packet Data Network Connections

The PLMN is connected to external packet data networks through the GGSN. In the GPRS backbone, an **Access Point Name** (APN) is a reference to a GGSN. To support inter-PLMN roaming, the internal GPRS DNS functionality is used to translate the APN into the IP address of the GGSN. The APN is composed of two parts: the mandatory APN Network Identifier defining the external network to which the GGSN is connected, and the optional APN Operator Identifier defining in which PLMN GPRS backbone the GGSN is located. An APN consisting of both the Network Identifier and Operator Identifier corresponds to a DNS name of a GGSN. (3GPP TS 23.003)

The network operators usually have separate APNs at least for WAP and Internet, and often also for MMS. In addition, some corporations have dedicated APNs (corporate APNs) for added security when connecting to corporate intranets while others resort to the use of virtual private networks (VPNs) through the Internet APN for enabling intranet connections. The domain for operator servers includes among others gateways, proxy servers, web and wap servers, application servers, mobile middleware components, and wireless / GPRS accelerators. A basic scenario for external packet data network connections is presented in the figure below.

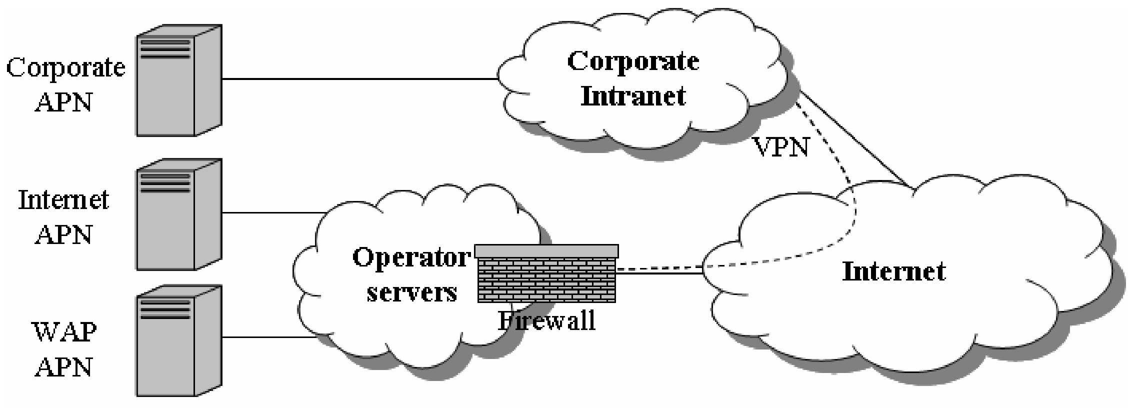


Figure 6 External packet data network connections

2.2.6 User Equipment and Mobile Station

The **User Equipment** (or Mobile Station) refers to the physical equipment used by a PLMN subscriber. It comprises the Mobile Equipment (ME) and possibly a removable smartcard called UMTS Subscriber Identity Module (USIM), or Subscriber Identity Module (SIM). (3GPP TS 23.002, 3GPP TS 23.101)

The USIM (or SIM) is a logical application running on a smartcard. These removable smart cards are commonly referred to as SIM cards. Each mobile subscriber in the GSM/UMTS system can be identified by a unique IMSI code (International Mobile Subscriber/Station Identity) of a maximum length of 15 decimal digits, which is contained in the USIM. The IMSI is composed of three parts: the three-digit Mobile Country Code (MCC) uniquely identifying the country of domicile of the mobile subscriber, the two to three digit Mobile Network Code (MNC) identifying the home PLMN of the mobile subscriber, and the Mobile Subscriber Identification Number (MSIN) identifying the mobile subscriber within a PLMN.

Similarly, the **Mobile Equipment** is uniquely identified by the IMEI (International Mobile station Equipment Identity) or the IMEISV (International Mobile station Equipment Identity and Software Version number). The IMEI is composed of an 8-digit Type Allocation Code (TAC) issued to equipment manufacturers, an individual six-digit Serial Number (SNR) uniquely identifying each device within the TAC, and a spare digit. The IMEISV is identical to the IMEI concerning the TAC and SNR but has instead of the spare digit a two-digit Software Version Number (SVN) allocated by the manufacturer identifying the software version number of the mobile equipment. (3GPP TS 23.003) The GSM Association (GSMA) is responsible for the allocation of IMEI code (or TAC code) ranges to equipment manufacturers. The GSMA has appointed the United Kingdom – based BABT (British Approvals Board of Telecommunications) to carry out the function of IMEI allocations on behalf of the industry. (BABT 2005)

Generally, **GPRS devices** are divided into classes based on terminal equipment functionality. GPRS classes (A, B and C) indicate device's capability to use GPRS and GSM simultaneously. Class A devices can be connected to both GPRS and GSM services simultaneously, whereas Class B devices mobile phones can be attached to GPRS and GSM services, using one service at a time. Class C devices are also attached to either GPRS or GSM service, while switching between the two has to be made manually. GPRS multislot classes determine the maximum achievable data rates in both the uplink and downlink directions. The classes range from 1 to 10 indicating the amount of timeslots the device is capable of using. (GSMA 2005)

2.3 TCP/IP Protocol Suite and WAP

A brief introduction to the TCP/IP protocol suite and the Wireless Application Protocol (WAP) is given in this chapter.

2.3.1 TCP/IP and Application Protocols

The functioning of the Internet is based on the **TCP/IP protocol suite**. TCP/IP software is organized into five conceptual layers, four software layers that build on a fifth layer of hardware. A simplified model of the TCP/IP protocol suite with the dependencies among the major higher level protocols is presented in the figure below. In the figure, each square represents one protocol residing directly above the other protocols it uses.

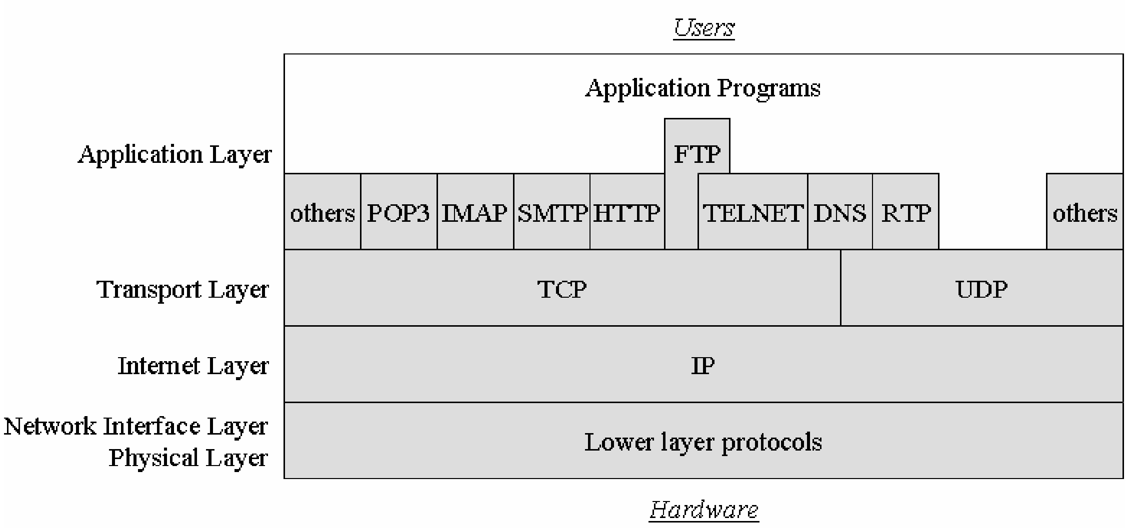


Figure 7 Major higher level TCP/IP protocols

At the highest level, the **application layer**, users invoke application programs that access services available across a TCP/IP internet. In general, each TCP/IP application defines its own application protocol and interacts with either TCP or UDP to achieve end-to-end transport. Each application program chooses the style of transport needed, which can be either a sequence of individual messages (UDP) or a continuous stream of bytes (TCP). Major application protocols related to e.g. email (SMTP, POP3, IMAP) and the web (HTTP) are presented in the figure above.

The **transport layer** provides end-to-end communication from one application program to another. The User Datagram Protocol (UDP) provides an unreliable connectionless delivery service, whereas the Transmission Control Protocol (TCP) is a connection-oriented, reliable-delivery byte-stream protocol. Using TCP, applications on networked hosts can create reliable pipe-like connections to one another, over which they can exchange data. UDP and TCP use IP to transport messages between machines, but add the ability to distinguish among multiple destinations within a given host computer.

The **Internet layer** handles communication from one machine to another. It accepts requests to send a packet from the transport layer along with an identification of the machine (IP address) to which the packet should be sent. It encapsulates the packet in an IP datagram, uses the routing algorithm to determine where the datagram should be delivered, and passes it to appropriate network interface. Correspondingly, the Internet layer also handles incoming datagrams, checks their validity, runs the routing algorithm, and chooses the transport protocol that will handle the packet if the datagram was meant for this host. As seen in the figure, Internet Protocol (IP) is the only protocol common to all applications.

The **Network Interface Layer** accepts IP datagrams and transmits them over a specific network hardware, which forms the **Physical Layer**. (Comer 2000)

2.3.2 Wireless Application Protocol

Most of the original Internet technology has been designed for desktop and larger computers and reliable medium to high bandwidth data networks. Mass-market, hand-held wireless devices present a more constrained computing environment. The **Wireless Application Protocol (WAP)** was designed as a solution to this constraint, to insulate the Internet-based application from the wireless mobile network infrastructure. WAP is actually a set of technical standards that enables users of wireless devices to access, and receive content from the Internet similarly to using a web browser.

There are two major **WAP releases**, versions 1 and 2. Version 1.x features a custom WAP protocol stack with WAP variants of transport (TCP and UDP) and application (HTTP) layer protocols. A WAP protocol gateway function is needed to make the protocol conversion between WAP and TCP/IP stack protocols. Version 2.x no longer uses the custom protocol stack and the WAP gateway. Wireless profiles of TCP (see WAP Forum 2001b) and HTTP (see WAP Forum 2001c) are used instead. There are more than one possible stack configurations, some of which are presented in the WAP Architecture specifications (WAP Forum 2001a). Moreover, as WAP version 1.x systems coexist with WAP version 2.x systems, some phones feature dual stack support, i.e. support for both WAP 1.x and 2.x protocol stacks. (WAP Forum 2001a) The web and wap standards are likely to converge further in the future.

At the moment, a separate **WAP APN** is used by mobile operators to direct WAP traffic to WAP gateways and operator-controlled portals. Some services of the outside Internet can also be accessed via the WAP APN.

2.4 Previous Mobile Internet Usage Research

There is a wide range of studies on “mobile Internet”, each focusing on a certain aspect of the research area encompassed by the term. Mobile Internet research can be roughly categorized at least in the following ways:

- Data collection methods
- Data collection conditions
- Focus/aim of the research
- Research organization / researcher

There are several **methods to collect data** for mobile Internet research. Direct measurements on e.g. traffic and usage vary in the points of measurement, which range from mobile terminal devices and the network air interface to other network interfaces and external servers. Questionnaires and surveys are applicable on both consumer and business users, for both evaluations of usage patterns and the collection of demographic background factors. Expert interviews, and discussion panels or seminar groups with various specialists can provide insights, inside information and estimates on the market situation. New information can also be derived by analyzing existing market data such as billing and retail information, and company annual reports.

Data collection conditions can also be used to differentiate mobile Internet research. Data collection in "field conditions", such as from live networks and from samples of real users provides data on actual usage. Data collection in "laboratory conditions" includes the use of test networks or an easily accessible but poorly representative user population. In addition, data can also be collected from case studies ranging in scope from a single server or a company to a whole nation or a geographical market.

The **focus of mobile Internet research** varies from economic or techno-economic to purely technological. Studies with a techno-economic focus entail aspects such as the enabling characteristics for mobile Internet (e.g. mobile terminal base), consumers' propensity to use mobile data services, and the effects of pricing on usage. Business models and the adoption of certain services, for instance, might also be studied. More technologically-oriented mobile Internet studies concentrate on topics such as the performance of networking protocols and quality of service in the mobile environment, Mobile IP, and radio interface –related issues, and might also be used in the testing and development of new systems.

Finally, mobile Internet research is conducted by many **research organizations**. These include a range of research institutes, both commercial and academic, as well as industry actors, such as equipment manufacturers, mobile operators, consulting firms and investment banks. In addition, telecommunications regulators and independent researchers also conduct their own research.

Out of all mobile Internet –related research those featuring similar research methods to this research are presented. Moreover, previous research with differing methods but similar objectives to this research is also treated.

2.4.1 Previous Operator Reporting System –Based Measurements

Operator reporting based measurements are, as the name implies, conducted by mobile operator's reporting function weekly, monthly, or yearly depending on the issue at hand, or at a separate request. As the information is gathered straight from the operator's subscriber management systems or charging data records (see chapter 2.2), they might not even be considered as measurements by the operator. Thus, a lot of the information published in mobile operators' annual report, press releases, and also in secondary studies based on any operator-originated information is actually likely to be based on reporting systems. As relevant previous studies using specifically reporting systems as data collection method are hard to identify, they will be treated along with other relevant previous research (see chapter 2.4.4).

2.4.2 Previous Packet Data Traffic Trace Collection

GPRS network operators have probably conducted detailed measurements on their networks using some form of traffic trace collection or more sophisticated commercial traffic analyzer equipment. While an operator conducting this kind of measurements has very few incentives to publish any of the findings, some previous research based on GPRS traffic trace collection and subsequent analysis has been published on few occasions. However, details of the measured traffic profiles have usually not been published due to their sensitive nature.

Several short and long term GPRS network traffic trace measurements were conducted during 2002 and 2003 in three commercial European Vodafone GPRS networks. Kalden (2004) used nine of these measurements, which lasted from 50 to 900 hours. The objectives for these measurements were to study the application and session usage in GPRS (also Kalden et al. 2003), to derive models for application flows based on UDP and TCP packet flows (also Kalden & Ekström 2004), to understand the user's mobility and its correlation to the application usage (also Kalden & Sanders 2004), and to assess the self-similarity property of GPRS traffic (also Kalden & Ibrahim 2004). Furthermore, Varga et al. (2004) used two Vodafone measurements to extend the understanding of WAP traffic patterns. All these measurements were conducted using the same setup with two measurement points enabling the linking of IP traffic to subscriber information. IP traffic measurements were conducted at the GPRS network Gi interface, whereas GSN event measurements took place in the SGSN. The open source tcpdump tool was used for IP packet trace collection, whereas the SGSN measurements were done with a proprietary tool. Several proprietary, commercial and custom developed software tools were used for the post-processing and analysis of the traffic trace data.

Kilpi (2002, 2003a, 2003b) studied GPRS/GSM sessions by measuring GPRS traffic on Finnish mobile operator Radiolinja's (currently Elisa) network in May 2002, about half a year after the operator had launched its commercial GPRS service. Over 2800 tcpdump traces were collected from a monitoring interface of a firewall router during 11 days. Each trace contained a fixed number of packets and the start time of each tcpdump run was also recorded. In addition, Kilpi and Lassila (2005) studied the variability of round trip times of TCP flows on a 30-hour traffic trace from one of GGSN node of Elisa's Finnish GPRS/UMTS network measured in November 2004. The measurement data showed that the effect of the portion of subscriber population having flat rate packet data tariffs from 18:00 till 06:00 on evening and night-time traffic volume was significant.

Benko et al. (2004) conducted passive traffic monitoring in seven European and Asian countries to study the end-to-end performance of TCP connections. Monitoring was made at the core network Gi interface with the measurement duration varying between a few days to four weeks.

Ricciato et al. (2005) collected packet traces at the UMTS core network of Mobilcom Austria to study bottlenecks in the network. The one-month measurements were conducted using the Gn interface as the point of measurement. Vacirca et al. (2006) passively monitored all core network interfaces (Gi, Gn, Gb, Iu) of the Mobilkom Austria GSM/UMTS network to analyze the frequency of spurious retransmission timeouts and the timeout probability in different situations.

Ivanovich et al. (2004) measured GPRS traffic at Telstra's Australian network to build detailed traffic profiles and study existing and new traffic models. Measurement traces were captured in four one-week periods between November 2002 and May 2003 at the Gb interface using a "Tektronix K1205 Protocol Analyzer". Post-processing of data was conducted using locally written scripts and standard Windows and Linux tools.

Dahmouni et al. (2005) performed measurements on the live operational Orange France GPRS network to evaluate the impact of data traffic composition on GPRS network performance. GPRS data traces were collected on two different Gb interfaces during a ten day period in July 2004 (about 5 hours per day between 11 a.m. and 4 p.m.). The collected traces contained TCP/IP and UDP/IP header information and associated timestamps.

2.4.3 Previous Handset Usage Measurements

A continuous panel study is a series of measurements on the same sample of test units over an extended period of time. Panel research is used extensively both in Europe and the U.S. to study consumer purchase patterns (Aaker et al. 2004). Handset usage measurements conform to the above definition of panel studies as a large sample of willing participants must be collected for the continuous handset usage monitoring panel. However, the implementation of these measurements is quite different.

Handset usage monitoring, or other direct measurements of handsets usage are not very common due to the relative novelty of smartphones that make the development of software enabling such measurements possible. Nevertheless, at least one 3rd party measurement software platform called ContextPhone has been developed on the Symbian S60 platform at the University of Helsinki (Raento et al. 2005). The platform was used in Eagle's research (2005) on social networks and human behavior. However, due to the proprietary nature of parts of the S60 platform the data acquired with this platform is rather general regarding handset usage factors.

Verkasalo (2005) treats extensively the statistical methods used when analyzing data collected with a handset usage research platform developed by Nokia (see chapter 3.3.2), with the main data set encompassing a group of 300 over 18-year-old Nokia 6600 users with postpaid GPRS subscriptions using mobile content services in the U.K. and Germany. Several projects have been implemented in laboratory environment related to the early development of the research platform. Large-scale field studies have been implemented since 2004. Currently, further studies using the platform are taking place on several countries in Europe, Asia, and Americas. Each new study has featured an updated version of the research platform with some added functionality. Nevertheless, comparisons to previous studies are possible concerning the older functions of the software. No studies involving multiple mobile operators have been implemented before this study. (Verkasalo 2005) Heimo Consulting, a Finnish consultative research company and a former subcontractor in the development of the monitoring software conducted a mobile handset usage study between March and December in 2005. The study concentrated on 30 to 44 year-old active users of both mobile handsets and Internet, with a sample size of around 50 consumers in Helsinki metropolitan area. An earlier version of the monitoring software was preinstalled to Nokia 6600 handsets, which were given to the participants for the duration of the study. Another monitoring application was also installed on the participants' PCs to study the usage of fixed Internet. Among the studied issues were usage of SMSs, MMSs, and email; usage of service numbers; usage of phone applications, and usage of mobile and fixed Internet. (Heimo 2005)

2.4.4 Other Mobile Internet Usage Research

Much of the published research on mobile Internet usage is originated from commercial research institutes, consulting firms, and investment banks (e.g. Forrester, Nomura, Ovum, Gartner, Deloitte). The results are usually presented in commercial research reports and are typically based on surveys and expert interviews. Moreover, these reports do not often provide a careful analysis on the significance of potential sources of errors. Nevertheless, some relevant previous research using different data collection methods but similar research objectives to this research is presented next.

The Finnish Association of Electronic Wholesalers (Finnish: *Elektroniikan Tukkukauppiaat r.y., ETK*) publishes monthly statistics on the sales volumes and average prices of mobile phones and other consumer appliances in Finland. Kotek is a collaboration forum of Finnish consumer electronics and appliance industry actors, such as ETK, Kodintekniikkaliitto r.y., and the major retailer chains. Kotek publishes similar statistics on a monthly basis, including sales volume and growth figures, and average prices for all mobile phones and smartphones in particular, gathered together with ETK and the globally acting market research company GfK’s Retail & Technology division. However, statistics on the actual mobile terminal base in use or more detailed information on the shares of certain advanced handset features are not published.

The Finnish Ministry of Transport and Communications publishes figures on mobile phone penetration in Finland. A report (Snellman 2005) prepared by eBird Scandinavia for the ministry gives some further insights on the Finnish installed base of mobile phones. Based on publicly available statistics and data from interviews of Finnish mobile operators, service providers and other industry experts, estimations are given for the percentage shares of phones with color screens, camera phones, email capable phones, smartphones, and 3G subscriptions for the end of 2004. It was stated that the share of data services, while in rapid growth (26%), was less than 10% of the value of the whole mobile service market in 2004. The value of data services was dominated (95%) by packet-switched services while the share (and volume) of circuit-switched services is decreasing rapidly.

Ketamo et al (2005) studied the mobile communications patterns of Finnish youth with a web survey of over 2400 (mostly) teenage respondents. The study revealed differences in mobile operator and handset manufacturer market shares, most popular handset models, change rates of handsets and subscriptions, and voice call and SMS usage in various teenage age categories.

A report prepared by Forrester Research Inc. (Van Veen 2005) gives some percentage information on the 3G penetration in different European countries based on interviews of 25 European mobile operators. These figures show Finland lagging behind its European counterparts. Similar, though less detailed figures on the penetration of certain mobile terminal features in Finland compared to average European level are provided by Finnish operators in their annual reports (e.g. Elisa 2006).

Minges (2005) recognizes the difficulties of doing international comparisons on mobile Internet usage and proposes several standard indicators to mobile operators, telecommunications policy makers and government statistical offices that can assist inter-country comparability and enhance understanding of trends. Indicators in usage, users/subscribers, infrastructure, revenues and pricing –categories are used by Minges to classify seven developed economies on the Asia-Pacific region with publicly available information.

3 Measuring Mobile Internet Usage

Three fundamentally different methods for measuring Internet usage by mobile subscribers are presented in this chapter. These methods are: data collection with operator reporting systems, packet data traffic measurements in mobile networks, and handset – based usage measurements. A comparison of the presented methods is also provided.

3.1 Data Collection with Operator Reporting Systems

Mobile operators' charging-oriented billing systems provide an excellent source for general level service usage data. Two separate types of reporting systems, namely CDR databases and billing systems can be used for collecting data on the entire subscriber base.

The **CDR databases / data warehouses** give access to vast amounts of data, as the CDRs provide diverse information on the chargeable events generated by the user. In principle, postpaid and prepaid subscribers can be distinguished by the different IMSI / SIM card ranges used, traffic to different packet networks counted by the network element (GGSN) in question, and the subscriber's mobile device by the IMEI / TAC code of the used terminal. Thus, packet data traffic volumes of different service operators' postpaid and prepaid subscribers' to (e.g.) Internet during a certain period of time could be divided by terminal models simply by using CDR data. However, one major limitation of CDR-based data collection is that the network basically has no knowledge on subscription types, or tariff schemes for that matter. Thus, measuring specifically (e.g.) business subscribers' traffic is not possible unless their traffic is directed to a specific network element (e.g. GGSN of a corporate APN).

Mobile operator **billing systems** provide the actual contact channel to end users. The data obtainable via these systems is usually limited to the postpaid subscribers of the service operator in question, as prepaid subscribers are not actually billed and might thus be located in a different information system. Subscription types can be further divided to consumer and business subscriptions. Consumer subscriptions are registered to an identity number or social security number, whereas business subscriptions are identified by a business ID (Finnish: *Y-tunnus*). However, both identity numbers and business IDs only refer to the bill payer of the subscription, not to the actual end user. Billing systems also provide further information on the subscription, such as the tariff schemes used for different services (voice calls, SMSs, packet data transfer…), and possibly some additional background information on consumer subscribers such as gender and age. However, legislation might limit the usage of such data. Billing systems provide access to a variety of data items related to service usage, including usage volumes and time-of-usage information in addition to the amount of money billed. Billing system also provides information on roaming activity albeit with somewhat less accuracy. General statistics for different subscriber types can be generated for most of the services using the billing system. Using billing systems as a source for data collection has some additional limitations. For instance, a more detailed investigation of Internet-bound packet data traffic is not realizable, as only the aggregate amount of chargeable packet data traffic volume is usually available. Moreover, the availability of data might also be limited to the moment when the subscriber is actually billed for the usage, not to the moment of usage. Finally, as billing systems are not specified by (e.g.) 3GPP standards, different systems provide very differing functionalities making the comparison of data from different systems challenging.

Considering the pros and cons of the above methods, **combining CDR and subscriber information** appears as an attractive option. As chargeable usage information is collected in standard CDR format separately for each subscriber, this does not seem to be an impossible task. If such functionality is not already implemented in the reporting systems, one could easily envision merging the two resulting data sets "manually" using some key identifying the subscriber at both systems (e.g. the IMSI code). Furthermore, such merging should be easy to automate. In practice, however, the possibilities of accurately linking CDR data to billing data depend on the integration of the used information systems and the resources available for this type of efforts. The ever-present cost-benefit balancing may well prove combining the data too laborious, even though it might be technically possible.

3.2 Packet Data Traffic Measurements in Mobile Networks

Standard Internet traffic measurement methods and tools for data collection, processing, and analysis can be utilized outside the PLMN, after the GGSN has detunneled the GTP tunnels to normal IP traffic. More thorough discussion on the matter can be found in e.g. Peuhkuri (2003).

3.2.1 Types of Network Traffic Measurement

IP network traffic measurements have been thoroughly studied at various levels. These measurements are performed for a variety of **measurement objectives**, including traffic and network characterization, network monitoring and network control. The objectives dictate much of how the measurement should actually be implemented. Measurements can also be classified according to two factors: location and granularity.

Traffic measurements can take place in many locations, ranging from backbone network, and access networks to the edge of core network and specific servers. The selection of **measurement location** is made according to the research objectives as the selected location must provide representative data. Both full monitoring (census) and sampling – based measurements can be implemented. Moreover, the measurement can take place in a single location or it may take place in several locations. A network consists of network nodes, each of which having one or more network interfaces connected via links to interfaces in other nodes. Some simple monitoring, i.e. passive and local measuring is possible at each interface, but this does not produce any overall view on network status or end-to-end performance. Node-to-node measurements are typically used in performance monitoring, using either real application traffic or test traffic. Certain measurements in a telephone network are carried out on the call level (CDRs, as seen in the chapter above). In IP traffic the equivalent is a flow-based measurement. Packets are grouped into flows by their source and destination IP addresses, protocol, and source and destination port numbers. For each flow a set of details is collected. There are a huge amount of flows in the Internet as e.g. a retrieval of one web page may result in several TCP flows. Thus, flow measurements produce a lot of data.

Measurement granularity is also an issue dictated by measurement objectives. Network metrics can be classified into at least four categories: utilization metrics, performance metrics, availability metrics, and stability metrics. The utilization metrics include among others packet and byte counts, protocol, and application distribution. Considering the time dimension, traffic measurements can be either continuous or sample-based. The time scale of the measurements varies depending on measurement objectives. Network planning requires measurement periods in the order of months; capacity management reacts on the time scale of hours or days, whereas real-time network control can be manual or automatic and works in minutes or shorter time scales.

The measurements can be divided into two categories: active measurements and passive measurements. **Active measurements** involve sending either real application data or test data to the network, and measuring the response and/or some other factors at the sending end or at both ends. An active measurement can cause disturbance to the network by introducing excess load, and affect the performance of the network. **Passive measurements** do not add traffic to the network, except for the transfer of results, which can also be done offline. As passive measurements do not interfere with the normal operation of the network, they are preferable in many business critical networks. Privacy issues present another problem with passive measurements if the measurements are done on an operational network. Some measurement efforts may also combine passive and active measurements. When measurement data is collected by the network elements themselves some care should be taken considering the load caused to the operational network elements as no performance degradation should be allowed. Thus, the use of external measuring equipment is often recommendable.

3.2.2 Passive Packet Data Traffic Trace Collection and Processing

The network hardware technology in use dictates how the data capture or **packet trace collection** can be done in practice. Regarding software tools, the most popular packet trace collector is tcpdump, which uses a highly portable library libpcap to capture packets. The program runs under the command line and is available for multiple operating systems, typically for different UNIX variants. While other tools do exist, tcpdump is the de facto standard in research. Trace files written by tcpdump, “pcap”, are a common format to exchange packet traces. (Peuhkuri 2003) Ethereal is another fairly popular protocol analyzer software. Ethereal provides similar functionality, but unlike tcpdump, also includes a graphical user interface front-end. In addition, there is also more sophisticated commercial traffic analyzer software capable of e.g. real-time analysis of high bandwidth traffic. However, these solutions are out of the scope of this study.

After the measurement, the measurement data is exchanged either online or offline. Before an analysis of the data is possible some **processing** is most likely needed. The amount of data traffic can be quite large and the finite memories available set some limitations on the granularity and time scale of the measurements. When planning the measurements one should know what data is truly necessary and plan the measurements accordingly. Furthermore, there is quite a lot of redundancy in the data that can be compressed away either in implementation or post-processing. Compression also keeps the amount of data manageable. Due to confidentiality requirements set on network operators, it might be necessary to remove sensitive information such as sender IP addresses from the data. This can be achieved by replacing the addresses with contrived information (with e.g. tcpdpriv), which makes the mapping of the sender to data impossible while still maintaining other useful information. However, when dynamic IP address allocation is used by the operator the sanitation of sender IP addresses might not be necessary, as users are not identifiable based only on IP addresses anyway.

3.2.3 Analysis of Packet Traffic Traces

Various types of information could be extracted from packet traces, even if it only includes basic TCP/IP header data. Naturally, basic traffic quantities such as the number of transferred bytes, flows and packets are derivable from header data. Data on used application protocols, and visited web sites can also be extracted in a fairly straightforward manner, at some level of accuracy. A more advanced analysis also enables an identification of the operating system that generated the traffic.

Every TCP and UDP protocol header features the source and destination port numbers **identifying the application protocol** used. Currently IANA lists several thousand assigned TCP and UDP port numbers for different application protocols. Thus, TCP or UDP port number seems to be an ideal choice for application identification. There are, however, several problems with port-based application identification. Some applications (e.g. RTP) do not use fixed port numbers at either end but select both ports dynamically. There are also common protocols that are used in non-standard ports. For instance, with URLs some schemes allow specifying other port numbers with host names instead of the default 80 of HTTP. Moreover, an application may use non-standard ports to masquerade as another application in order to e.g. pass simple firewalls that filter by port numbers or gain higher priority when this is done based on port numbers. Applications can also have different roles. SSH supports terminal connections, file transfer and arbitrary TCP tunneling, and uses the same port for all these. Actually there might be multiple different connections in one TCP connection. Similar problems exist with other tunneling protocols. Thus, foolproof identification of applications using only stateless information, i.e. port numbers is difficult. The accuracy of application identification could be enhanced by observing how different applications communicate. Applications can be differentiated by comparing packet sizes, packet inter-arrival time distributions and flow lengths. (Peuhkuri 2003) Another method to identify application protocols is to use transport protocol payloads, i.e. by capturing the first few bytes of TCP or UDP payload a fingerprint for the application protocol can be identified. Such a method would solve some of the problems of port number –based identification, whereas the use of encryption would make it less accurate.

Internet Protocol datagram headers contain the source and destination IP addresses of the packet. Typically, domain name system (DNS) is used to determine what IP address is associated with a given hostname. A reverse DNS lookup can be used to find out what the associated hostname for a known IP address is. Thus, the IP address can be used to identify the organization hosting a particular service (application protocol). By inspecting the server IP addresses and associated domain names for web protocol traffic (e.g. HTTP, HTTPS), the **most visited web sites** can be identified at some level of accuracy. While traffic volume can be derived with this method fairly accurately, the number of times each web site (or domain) was visited is not so clearly defined. The number of traffic flows to the domain can be used as an indicative measure for site visits. Generally, the number of files downloaded from the site is (equal or) larger than the number of flows from the site, which in turn is (equal or) larger than the number of actual site visits. One should also notice the possible effect of caching on this. Again, TCP payload containing the web protocol traffic itself provides even more accurate data on the visited web sites.

Operating system (OS) **fingerprinting** is based on analyzing the information sent by a remote host while performing usual communication tasks. The underlying principle is that every operating system's TCP/IP stack has its own idiosyncrasies. There are subtle differences between TCP/IP stacks, and sometimes certain implementation flaws that, although harmless, make those systems quite unique. OS fingerprinting is mainly used on TCP traffic, as UDP implementations are relatively simple and do not differ substantially in different operating systems. In order to be able to identify OSs, a database with identified OS fingerprints has to be built. Then traces of the packets sent by the remote system are captured and compared with those stored in the database. OS fingerprinting can be done both actively (e.g. queso and nmap) and passively (e.g. p0f). The passive technique is based on the collection and analysis of packet traces, whereas active tools rely on sending a variety of malformed packets to the identified host and capturing and analyzing the responses. OS fingerprinting is not 100% certain, as the identifying accuracy depends on the scope of the fingerprint database. Moreover, it should be noted that it is possible to change TCP/IP stack settings to either avoid identification, or appear as some other system. While OS fingerprinting can be used to identify e.g. remote proxy firewalls as they rebuild TCP connections for the clients, this same property prevents the identification of the real OS of the user. (Honeynet 2002)

3.3 Usage Measurements on Individual Handsets

Measuring handset usage directly from the handsets has not been possible until recently due to the hardware-based limitations, as handsets have lacked adequate capability of independently registering usage events. On the other hand, while PDAs might have had the necessary computing capabilities for some time, telephony is a rather new addition to the range of PDA functions. Nowadays smartphones also include sophisticated operating systems and software platforms that enable the development of more advanced application software. Nokia has developed a market research platform, which enables data collection directly from Symbian S60 handsets. Direct usage measurements are possible with specialized monitoring software that runs as a background process in the smartphone. As the development of this kind of software necessitates knowledge on the protected lower level software interfaces, no private party can implement it independently. While the Internet features a variety of different malware used among other purposes to collect information on usage behavior, this software is not malicious as it is purely used for the anonymous monitoring of users who have given their consent for the monitoring. Nevertheless, strictly technological understanding does not suffice for implementing measurements successfully. As the "big brother" effect inherent to the monitoring of individual users might results in bad publicity, requirements are set on both legal and marketing capabilities as well.

Handset usage studies using the research platform are essentially panel studies, where a set of smartphone users is recruited for monitoring. This set, the *panel*, is then monitored continuously for a period time, usually several months. The monitoring software on each participating *panelist*'s handset logs usage events and sends the logs regularly over the network to a server, treating the panelists anonymously (see chapter 3.3.2). A supplementary survey is made by the panelists at the beginning of the study, and optionally at the end of the study. The common phases of panel studies include: panelist recruitment, data collection, and processing and analysis.

3.3.1 Panel Recruitment for Handset Usage Research

No established methodology exists for launching panel studies due to their relative rarity and different areas of application. Nevertheless, four general stages can be identified in all panel research: sample design; sample selection; sample contact, classification and task description; and sample recruitment. (Blyth 1991)

Sample design

The **recruitment rate**, the share of the contact sample actually joining the panel is a key issue in sample design. Recruitment rates to panel studies are not as high as response rates to random sample survey studies. Recruitment rates depend on the perceived scale of the task in terms of time; the complexity of the task; the perceived personal and social relevance or value of the task; and the value of the incentive offered. Moreover, recruitment rates also vary by demographic sub-group, depending on the subject and the data collection methodology. All these issues affect the possibility for making generalizations on panel data. (ibid)

People usually need some kind of **incentive** for participating in any research. In the case of intense monitoring of individual user actions and decline of privacy, the use of sufficient compensation can increase the recruitment rate considerably. When considering handset monitoring studies, both the user's expenses and personal effort must be compensated. This means both compensating the user's increased data transmission costs for the uploading of usage logs from handsets, and providing addition compensation as an incentive for actual participation. The available options include e.g. providing cash compensations or free talk time for the participant, and organizing a marketing lottery with some attractive goods as rewards. However, it should be noted that national legislation might limit the organization of lotteries for marketing purposes. Some more radical forms of incentive can also be envisioned. In some previous studies, free handsets were given to the panelists for the duration of the study. While using this method provides an easy way to recruit an adequate number of participants, it might have an effect on usage behavior and is not a feasible method economic-wise. When targeting some specific group of people, higher participation might be obtained by appealing to people's duty as members of that community or organization. It is also recognized that some forms of compensation might appeal to certain user segments more than to others, resulting in disproportional representation in user segments. To limit the biasing effect of incentives on behavior, the compensation should be handed over as a reward after the data collection has ended, only if the user has presented sufficient participation effort.

Sample selection

The sampling process (e.g. Aaker et al. 2004) starts with the **identification of the target population**. Typical target populations are people living on a certain geographic market (e.g. Finland) or the members of a large organization. Similarly, customer segments (e.g. consumer/business subscribers, age) can also form the basis for the population. Age can also limit the population, as recruiting minors to consumer panels might include legislative issues. Furthermore, the target population of handset usage studies of this type is limited to smartphone users, as direct handset monitoring is possible only on smartphones. Smartphone users can be considered as an advanced “early adopter” user segment currently not representing the whole handset user population.

Next a **sampling frame**, a list or other record of the chosen population used to obtain a sample, must be defined. When the study is conducted in cooperation with a mobile operator a natural sampling frame is the operator’s subscriber base whereas members or the buyers of new handsets by bundling a recruitment message with the packages of new handsets might be used in other situations. Sometimes the personnel of an organization can also form a suitable sampling frame. In addition, as the recruitment of panelists can be outsourced to 3rd party agencies, some 3rd party customer databases might also form the sampling frame. Furthermore, with indirect contacting the users of some service, such as a web portal or a physical sales outlet, define the sampling frame used. Sometimes the decision on the contacting channel affects the sampling frame, as the contacting might be limited to those customers who have given their express consent for receiving certain kinds of direct marketing communication (e.g. email, SMS) or to those who have not explicitly forbidden it. Possible population and sampling frame differences should also be dealt with.

Next a **sampling method** or procedure is selected. Major probability sampling methods used include simple random sampling in which each population member has equal probability of being selected, and stratified random sampling which ensures adequate representation of certain predefined groups in the sample. Given the low recruitment rates and their differentials regarding demographics and other key variables, some form of quota sampling should be preferred in panel studies.

When attempting to generate a representative sample, determining the relevant **sample size** is a key consideration. While statistical theory provides some insight to solving an adequate size, the costs involved in executing the research might also set upper limits to the sample size. The high costs of sending thousands of letters or setting up physical locations for participant recruitment might be avoided by using electronic contacting channels. However, the number of people requiring compensation for participation sets another cost-related upper limit to the sample size. As seen above, recruitment rates are also relevant to sample size. While a highly targeted study might have high recruitment rates, the rates will be significantly lower in studies using of random sampling based contacting on non-interactive electronic channels and requiring plenty of effort from the respondents, despite the use of tempting incentives.

Sample contact, classification and task description

The next stage of panel launch includes contacting the selected people, classifying whether or not they are suitable for the panel, and describing the tasks required from them. All this can be achieved in one interactive encounter, or in separate phases using different communication channels. Piloting different contacting methods is advisable as they might also affect various demographic groups differently, which enables differential issuing of contacts respective to recruitment rates. Each approach features its own pros and cons making it ultimately a question of balancing inputs and outputs. Considering the alternatives for realizing a handset monitoring panel, several options exist.

Some form of **direct personal contacting** of the selected people is probably the most efficient contacting method in terms of recruitment rate, as a high level of interaction enables the recruiter to persuade the recruited to join the study and answer most of his questions and doubts during the encounter. Making voice calls is probably the most valid alternative, though selecting valid customers is very important to avoid unnecessary calls and negative feedback. However, this method is likely not used as a primary contacting method due to the higher costs involved. **Impersonal direct contacting** refers to sending impersonal invitations to the selected people and then passively waiting for the motivated to look for more information. Mass mailing with electronic channels, such as SMS or email, is a low-cost alternative. However, the use of email is somewhat questionable as the amount of unwanted spam email is ever increasing. Moreover, user email addresses are not necessarily collected to customer registers and while the phone numbers for SMS sending might be readily available, the actual sending of SMSs has legal limitations (see chapter 3.4.3). Traditional mail is a valid option as advertising through the post has no major legal restrictions. Mobile operators possess customer postal address information as billing letters and customer magazines are sent regularly. However, as the billing process is a fairly automated process and customer magazines are sent rather seldom, the sending of a separate letter might end up being the only suitable option. Sending thousands of letters through the post incurs relatively high costs. All direct marketing efforts result in some amount of **negative feedback**, or backlash from the targeted people. The people receiving the recruitment invitation might feel disturbed by it, find some aspects of the study illegal, or just require additional information before joining. Thus, all parties named in the recruiting message and e.g. at a related online site should give the customer service function instructions on how to deal with questions and feedback related to the study. Providing contact information on the recruitment message or at the registration site should also channel the feedback to one party responsible for these enquiries.

Indirect contacting methods essentially rely on the same principle as impersonal direct contacting described above, i.e. having some form of advertisement inviting people to find out more about the study. These methods include: advertising banners on company web sites, web/wap portals and related Internet discussion forums; separate flyers in handset packages; and stands or posters in retail outlets and trade fairs. Recruitment rates for these types of campaigns are obviously low and might also provide disproportional representation to some user categories. Nevertheless, indirect methods might be the only ones available and are usually low-cost, especially in the case of electronic methods. Indirect contacting can also be used as a complement to direct contacting.

Other approaches include using 3rd party agencies that might have proprietary access to some customer bases and have a variety of direct and indirect contacting methods in use. However, the costs of using 3rd party agencies for the recruiting users are high, and do not necessarily result in a truly representative group of panelists. One should also consider the effect of word-of-mouth, as the people receiving information on an interesting (or irritating) study share their experiences with coworkers or on some other forum.

As joining the study should be made as easy and simple as possible for the contacted people, the **change of communication channel** should be minimized. As the web-based registration system used with the handset research platform (see chapter 3.3.2) necessitates the use of both a PC and the handset, the use of email and SMS, and web/wap–based indirect methods should be preferred from this point of view. Similarly, doing the whole recruiting and joining process during a single person-to-person encounter at e.g. a retail outlet would also be preferable. However, it is recognized that cost-benefit considerations ultimately drive the channel decision.

Sample recruitment

Sample recruitment refers to the actual agreement to join the panel. When the contacted people agree to join the panel, they must register. In handset monitoring studies this can be done with a PC at the online web recruitment system, or at a physical location. During the registration process, the user completes the necessary **registration forms**, which include both accepting a contract with the party actually conducting the study and filling a beginning questionnaire with some background information (e.g. demographics) on himself, if such information is not available by other means. The **installation of the monitoring software** to the handset is done during the registration either remotely over-the-air (OTA) or at the physical location. The user must accept the installation of the software with his handset. Pre-installing the monitoring software to new handsets is also an option.

In typical panel studies this stage would also involve **panelist training**, i.e. what the researcher wants people to do. However, handset monitoring studies differ in this respect, as data collection is automated and instructions are needed only to install the monitoring software to the handset, and eventually to uninstall it. Thus, most of the training can be handled with online instructions and FAQ pages, and only some form of user support is needed for error situations.

Maintaining and ending the panel study

Panel research typically necessitates high care in panel maintenance in order to minimize the panelist mortality or **dropout rate**. There are numerous possible reasons for dropout, such as reporting fatigue or a sense of insignificance of the study, which can be countered with (progressive) incentives and adequate communication during the panel. In long-running panel studies a small panelist turnover is normal and the dropouts should be replaced by new panelists. In previous handset monitoring studies the dropout rate has been fairly low, most likely due to the transparency of the monitoring software and the automation of usage data collection (see chapter 3.3.2). Moreover, if the length of the study is only few months, there might not be enough time to recruit new panelists and collect enough usage data from them. Despite the potentially low dropout rate, some **user support** for the panelists should be provided by the study organizers.

The **study ends** after a predefined duration, which is usually several months. The panelists can also leave the study any time simply by stopping the use of the monitored handset or by uninstalling the monitoring software. Thus, some limits on the sufficient level activity must be defined. Changing the SIM card of the handset does not affect panel participation, although the use of a different mobile network (and SIM) is observable in the usage data. A **final questionnaire** containing questions on e.g. usage experiences might be presented to the panelists at the end of the study. Obviously there is less motivation for the panelists to fill this second questionnaire unless it is somehow tied to receiving the compensation.

3.3.2 Research Platform for Usage Data Collection

The typical **data collection** methods in panel research vary from diaries, interviews, smart card or point of sale data capture, and electronic data capture. (Blyth 1991) In handset monitoring studies, the actual collection of usage data, which commonly comprises major efforts in panel research, is completely automated with the handset usage research platform.

While being a key component in handset usage research, the monitoring software is only a part of the whole **research platform**. The acquisition and retrieval of usage data from handsets is implemented with an information system connected to Internet. First the user registers for the study while in interaction with the information system with a PC over the Internet. The information system then verifies that the software can be uploaded to the handset in question. The user must confirm the upload process with his handset after which the monitoring software upload commences. During the registration each user is provided with a unique identification number. In all subsequent phases only the user identification number is used for the mapping of users and data thus preventing the identification of the true personalities behind the individual data cases. Panel management and the surveys (beginning and final questionnaire) are also handled by the information system, which features some overall reporting functionality too. However, a more comprehensive separate analysis is performed on the data after the study. At the lowest level of the information system architecture handsets send the data on usage factors through GPRS link either on a daily or weekly basis, depending on the factors in question. The monitoring software uploads around 50 kB of usage reports per average user. Verkasalo (2005) provides a comprehensive outline of the information system and the process of data acquisition and retrieval. (Verkasalo 2005)

The information system used in data acquisition archives usage data from the functional point of view, resulting in dozens of raw data tables. The **raw usage data** includes all software platform level data from the smartphone, such as application launches, data transfer session, voice calls, setup actions etc. with detailed records on each of the different function used, including attributes such as timing and duration, related network signal coverage and the identification number –based anonymous identification of the people the panelist has communicated with. The information system enables the extraction of a complete usage data dump for analysis any time during the panel.

The collected **raw usage data is processed** in several separate phases from locating outliers and handling missing data to the aggregation of variables and the creation of relational databases. More applied statistical methods can then be applied to derive descriptive reports and graphs, to construct data structures or to perform specific analysis from the data. Further discussion on the usable statistical methods can be found in Verkasalo (2005).

3.4 Comparison of Network and Handset-Based Measurements

The three different methods for measuring mobile Internet usage presented in the previous chapters are now compared. First, the nature of measurable data in each measurement is presented. Second, the resources required in measurement implementation are evaluated. Third, the legal constraints present in all measurements are treated. Fourth, the major benefits and drawbacks of the measurements are compared.

3.4.1 Nature of Measurable Data

The strength of data collected using operator **reporting systems** is in its large scope, as the operator's entire subscriber base and their terminals can be included in the measurement. Data obtained with these means is available up to a high granularity level, as data items can be divided by subscription types (postpaid/prepaid, consumer/business) and terminal devices (terminal models, terminal radio capabilities). Data items attainable with billing systems include the numbers of subscribers, chargeable packet data traffic volumes, and general activity measures on packet data usage and roaming activity, among other topics. CDR-based data enables the collection of numbers of terminals, general packet data traffic volumes, and gives insights on the distribution of packet data traffic between different APNs. Moreover, additional analysis on CDR-based data gives access to e.g. the distribution of mobile terminal base by terminal manufacturer, specific terminal feature, and terminal age.

Similarly to the reporting system –based measurement, the asset of **packet data traffic measurement** is in its scope. Capturing traffic at an access point effectively captures all network users' traffic going via the access point in question. If the measured access point is selected well, this can include the vast majority of the operator's subscribers and their terminals. A separate analysis enables dividing packet data traffic by operating system. The measurable data items include traffic volumes (bytes, flows, packets) by transport protocol and traffic direction, server TCP and UDP port numbers, and visited web server IP addresses. Further analyses can refine these data items to application protocols and web site domain names.

Usage measurements on individual handsets are somewhat limited in scope, as the panels typically consist of a few hundred Symbian S60 smartphone using participants. While this might be a statistically adequate number of people to represent the smartphone using subscribers, the data is nevertheless sample based. Moreover, observations made on data on smartphone using panelists generally apply only to other smartphone users, unlike the data of the network level measurements. On the other hand, the possibility to collect extensive background information on the panelists enables the linking of gender, age, type of subscription, or any other collected background variable to handset usage data.

Moreover, some other usage characteristics, such as handset radio capability and time of usage are collected automatically by the research platform. Collected usage data can be presented in the form of descriptive statistics, such as tables and charts, but also provides the possibility of performing more advanced statistical analyses on handset usage.

The observations on the nature of data items measurable with the three measurement methods are presented in the table below.

Table 1 Comparison of measurements: Measurable data items

	Reporting system – Based Data Collection	**Packet Data Traffic Measurement**	**Usage Measurements on Individual Handsets**
Scope of data (subscribers)	All MSO subscribers, though mainly postpaid	All MNO network users with traffic via the point of measurement	Few hundred Symbian S60 smartphone using participants
Scope of data (terminal types)	All subscriber terminals	All subscriber terminals with traffic via the point of measurement	Symbian S60 smartphones
Data granularity	Subscription types (postpaid/prepaid, consumer/business), subscriber tariff types Terminal devices, terminal radio technologies	Terminal operating system	Gender, age, type of subscription, any other collected background variable Handset capability, time of usage
Measurable / collectable data items	Number of subscribers, chargeable packet data traffic volumes, general usage & roaming activity Number of terminal models, packet data traffic volumes, share of packet data traffic / APN	Traffic volumes (bytes, flows, packets) by transport protocol and traffic direction, server TCP/UDP port numbers, visited web server IP addresses	Any individual usage events, e.g. calling, messaging, application usage, browsing (domain), packet data traffic
Derivable data items	Distributions on terminal manufacturers, terminal features, and terminal age	Traffic distribution by application protocols, visited web site domain names	Descriptive statistics, more advanced statistical analyses on handset usage

3.4.2 Resource Requirements of Measurement Processes

As the **reporting systems** used by mobile operators for subscriber management are not standardized, the possibilities and limitations of the specific system(s) to be used must be known before detailed measurement specifications can be formulated, i.e. before a compromise can be made between the kind of data wanted and the kind of data possible and feasible to produce. However, as all the raw data is collected for charging purposes anyway, no separate setup is required, although it is recommended to carefully communicate each requested item of the specifications to the people actually implementing the measurements to avoid misunderstandings. The resources needed for the actual implementation of the measurements depend on whether already existing reports or separately run custom reports are used. Setting up custom report runs entails some additional work, and the data warehouse runs themselves on the operator's entire subscriber base over a long time period are rather resource intensive, despite the fact that they produce more accurate data specifically suited for the purposes of a certain study. Some manual work might also be needed in data aggregation, whether using existing or custom reports. While the duration of data warehouse runs producing custom reports is in the order of hours, the actual data collection cannot be completed until the charging data is collected, i.e. after the studied time period ranging from one week to one month has passed. Further waiting of about one month might be necessitated by delays in roaming data delivery from foreign operators, as well as a possible 1-2 month billing delay, depending on the used billing system's functionality and level of integration to other information systems. Considering the analysis of raw data, some phases might include lots of manual work. The collection of comprehensive lists mapping TAC codes to specific terminal model, and detailing the supported features of specific terminal models is very laborious, provided such lists are not already available. Furthermore, combining operator-specific data in multi-operator measurements might necessitate the use of much calculation and supporting assumptions to ensure commensurability of data sets.

As the methods used in the **packet data traffic measurement** are fairly well-established, both the preparation of measurement specifications and planning of the measurement are fairly straightforward. The actual measurement setup is also technically simple, although some separate hardware and software are need for it. As the measurements are conducted on a live operational network, considerable caution is required. The actual implementation of these days-to-weeks lasting measurements is not too laborious either, even starting and stopping of measuring can be automated with a timer beforehand. Some effort is required to transfer the high volume of captured measurement data to the site of analysis in a secure way. Processing and analyzing the raw measurement data can be mostly done with the help of several well documented open source applications, and some perl scripts separately prepared for the task.

The resource needs of the **handset–based usage measurements** are emphasized at the early stages of the process, as the final panel profile deciding panel representativeness and the background variables available in various analyses are both defined by the decisions made in the planning stage. Furthermore, as handset-based measurements feature more legal concerns than the network-based measurements, the input of lawyers and other legal specialists is required. The actual panelist recruitment process is rather laborious and time-consuming, which naturally depends on the recruitment methods used. Moreover, localization of the research platform to a new country necessitates some first-round setup, such as versioning of the monitoring software and the online registration site to conform to local language(s) and legislation. Once the panelist recruitment and registration has started successfully, usage data is collected pretty much automatically from panelists' handsets by the research platform for a duration of some weeks to several months. As there is an abundance of available handset usage data, the main concern in analysis is to decide which phenomena are actually worth studying. Presenting descriptive statistics by several background variables is relatively simple, provided that suitable software tools are available. Constructing more advanced statistical models on usage is naturally more laborious.

The resource requirements of the three measurements are summarized in the table below.

Table 2 Comparison of measurements: Process resource requirements

	Reporting system – Based Data Collection	**Packet Data Traffic Measurement**	**Usage Measurements on Individual Handsets**
Measurement planning	Laborious preparation of detailed measurement specifications by iterating between what is wanted and what is possible/feasible with the used reporting systems	Preparing measurement specification simple as measurement method is fairly well-established Measurement planning technically straightforward with detailed specifications	Central as panel profile decides panel's final representativeness Planning of initial questionnaire central Multiple legal issues require lawyers' input
Measurement Setup	No setup needed, charging data is collected anyway Communication of specifications to operator people implementing data collection required	Some resources needed for setting up the required hardware and software Caution required, as measurements conducted on an operational network	Panelist recruitment process laborious First-round research platform setup laborious (software language, registration site…)
Measurement implementation	No implementation needed if existing reports used Some work in custom data warehouse run setup Data warehouse runs on entire subscriber base are resource intensive Manual work needed in data aggregation with existing / custom reports	Starting and stopping measuring can be automated with a timer Moving high volumes of measurement data to site of analysis necessitates separate arrangements	Automatic data collection from recruited panelists' handsets using existing research platform
Measurement duration	Data collection possible only after study period 1 month delay of foreign roaming data delivery 1-2 month billing system dependent delay	Measurement duration from days to weeks	Measurement duration from weeks to months
Processing and analysis of measurement data	Comprehensive mappings of TAC code – terminal model – terminal feature are laborious to collect Combining operator-specific data sets might necessitate lots of work	Open source, well-documented tools usable for analysis Preparation of separate perl scripts for specific analyses necessitates some work	Abundance of data → most work in deciding what is worth studying Deriving descriptive statistics simple Advanced statistical models more laborious

3.4.3 Legal Constraints Involved

The **reporting systems –based data collection** has some serious legal restrictions, as collecting the kind of information described in chapter 3.1 relies heavily on the use of identification information (Finnish: *tunnistamistieto* or *teletunnistetieto*). Usage of identification information is heavily regulated, and Finnish legislation (Sähköisen viestinnän tietosuojalaki 516/2004) stipulates by whom, how, and for what purposes it can be used. While identification information itself is sensitive, it can be utilized for billing purposes and while settling subscriber misuse cases. Moreover, identification information can also be used for marketing purposes, if the subscriber has given his/her consent for the type of usage, and for technical development of products or services. Regarding this measurement, the usage of identification information is acceptable as its usage can be interpreted being related to service development. However, the raw measurement data containing identification information cannot leave the operator's premises. While analyzing the collected data, a very high level of granularity concerning rare subscription and contract types or terminal models should be avoided, as individual subscribers and terminal users can potentially be identified in cases of small sample sizes.

Regarding the **packet data traffic measurement**, the sensitive nature of some of the captured traffic items should be considered. In IP headers, the sender's and recipient's IP addresses identify communication parties to host level, meaning in most cases *personal data* (Personal Data Act 523/1999). Regarding TCP and UDP headers, protocol port numbers can be considered sensitive as they can be used to identify the used application and thus possibly also the person using them when the number of users in the network is small. Moreover, the payload carried by TCP and UDP may contain sensitive information, such as private email. In large-scale packet data measurements where no transport protocol payload is captured, sensitivity issues are limited to sanitizing subscriber terminal IP addresses in the case where static IP addressing is used. More discussion on sensitivity of IP traffic can be found in Peuhkuri (2003).

There are several legislative issues concerning direct **measurements on individual handsets**. First of all, the very realization of this kind of individual-level measurements might be restricted in some countries. In Finland there are no major obstacles when proper precautions are taken. These include (but are not limited to) voluntary participation to the study, adequately notifying the participants on the type of information collected, and measures to ensure the confidentiality of personal data (e.g. names, phone numbers) and communication. (Tietosuojavaltuutettu 2005) Participation (and subsequent study phases) might also be limited by participant age, as adulthood is required in many situations. Different phases of the recruitment of target users also encompass certain legal restrictions. Sending electronic mass scale messages is restricted in some cases, depending on whether it involves direct marketing in a client-principal relationship or other type of communication to customers. (ibid) The sampling of customer base is also restricted to some extent, depending on the methods used. For instance, the handset model a customer is using can be used in sampling if the information is readily available in customer registers. If the information must be acquired from the communication network, the focus moves from the use of customer registers to the use of identification information. Again, identification information can be used when the reason of use is technical product or service development. The Finnish legislation concerning both electronic communication and personal data is currently (1/2006) not quite up-to-date. Responsibility in controlling the compliance is shared by the Finnish Communications Regulatory Authority (Finnish: *Viestintävirasto*) and the Office of Data Protection Ombudsman (Finnish: *Tietosuojavaltuutettu*), depending on whether the matter concerns identification information (the former) or personal information (the latter).

The legal issues constraining the realization of the three measurements are presented in the table below.

Table 3 Comparison of measurements: Legal constraints

	Reporting system – Based Data Collection	**Packet Data Traffic Measurement**	**Usage Measurements on Individual Handsets**
Measurement implementation	Usage of identification information for data collection allowed for technical service development	TCP and UDP payloads are sensitive, and should not be captured IP addresses should be sanitized if static IP addressing for subscriber terminals is used	Study realizable when: voluntary participation, adequate notifying on collected information, confidentiality of personal data and communication, no minor participants SMS-based recruitment allowed if not as direct marketing Subscriber sampling using identification information allowed for technical service development
Measured / collected raw data	Individuals potentially identified if high level granularity used with rare subscription types and terminal models	None as dynamic IP addressing is used None as only packet headers are captured	Panelists identified using impersonal ID codes, specific application usage might reveal panelist identity
Processing and analysis of measured data	Raw data containing identification information cannot leave operator premises	–	Privacy issues with data items (applications, web sites) used by a small number of panelists

It is recognized that there are differences in international legislation concerning these issues. However, further treatment of issues concerning legislation is clearly out of the scope of this study. Furthermore, it should be noted that the collected data might have sensitivity issues business-wise. Especially the market level data collected in the reporting system –based part is of strategic nature to the contributing operators. Thus, some operator-specific information must also be sanitized from the results.

3.4.4 Summary of the Advantages and Disadvantages of Measurement Methods

The three methods provide **measurement data** of different nature, each with lots of potential and many concerns. The clear advantage of both reporting system and packet data traffic measurement is in their broad scope, as the data is collected at a nationally representative level in both cases. Reporting system –based data can also be considered very reliable, as subscriber billing and charging is based on the same source data. However, in multi-operator measurements, the validity of the data can be questioned as different operators use somewhat differing reporting systems making comparing of operator-specific data challenging. On the other hand, the validity concerns could be largely fixed by making customized reporting runs instead of using existing reports. Nevertheless, the collected data enables many additional analyses, depending on the data item in question, and its validity. In addition to the large scope, another asset of packet data traffic measurements is in the good comparability in multi-operator measurements, as data is measured in standard format. Comparability naturally requires the use of comparable points of measurement. A downside of the measurement method is in its inability to distinguish even the most general subscription types (e.g. business and consumer) from the traffic. On the other hand, similarly measured standard format data also enables the use of more advanced traffic analysis methods giving more accurate results, at the expense of increased complexity. Measuring usage at the handset level presents several advantages to other measurement methods. While measurements conducted at the network level on various network nodes give information on traffic flows and usage activity, it is extremely laborious to relate data flows and timing accurately to individual users, not to mention the legal issues potentially involved. Individual usage patterns are best captured at the individual level, at the handsets. The abundance of individual-level usage data enables a multitude of different advanced analyses not possible with other measurement methods. Moreover, the recruitment of the measured population individually also permits the collection of user background data, and thus enables mapping user demographics and other background variables such as subscription types on usage data. There are no comparability issues in multi-operator measurements, provided that the same research platform is used for all usage data collection. The only major limitation of handset-based measurement is in its scope, as usage data is obtained from a relative small sample of smartphone users. While a panel of few hundred smartphone users might be adequate to represent the current smartphone user population, all the segments of the whole population are not necessarily represented equally. While larger sample sizes are possible at the expense of higher costs related to running the panel, the research results are still limited to smartphone users only, not to the entire mobile user population.

The three **measurement process**es also entail some differing constraints. Considering who can actually implement such measurements, i.e. who has access to relevant source data is the primary question. Both network level measurements are available only to mobile operators. While service operators (or MVNOs) might have outsourced the management of some reporting systems, the “ownership” of the subscribers and related usage data still resides at the operator’s. Similarly, packet data traffic measurement requires access to the core network, thus limiting the possible measuring parties to network operators (or MVNOs). Handset-based measurement can be implemented by practically any party having capable research software at their disposal. However, the recruitment of hundreds of mobile subscribers is much facilitated by the participation of mobile operators in the recruitment effort. The resources required in actual data collection in reporting system –based measurements depends on the amount of custom reporting runs conducted, although some operator resources are needed in any case to collect the data. The implementation of packet data traffic measurements is technically rather straightforward. Moreover, it does not produce significant costs as basic hardware and open source software can be used for measuring. However, measuring traffic at live operational networks does necessitate extra caution. Regarding handset-based measurements, recruiting hundreds of voluntary participants requires a major effort from study organizers, especially when such measurements have not been conducted before in the market in question. As the most of the phases of the measurement process involve legal and/or sensitivity issues, much caution is required throughout the process. Moreover, the infamous “big brother” effect referring to the pervasive surveillance and the reduction of personal privacy might result in setbacks to participating companies’ reputation and brand name.

The advantages and disadvantages of the different measurements are presented in the table below.

Table 4 Comparison of measurements: Summary

	Reporting system – Based Data Collection	**Packet Data Traffic Measurement**	**Usage Measurements on Individual Handsets**
Measurement data • Scope • Comparability • Granularity • Subsequent analyses • Uniqueness • Reliability and validity	Large scope provides representative data at the national level Collected data is very reliable, as charging is also based on same source data. Validity of multi-operator data not quite same due to differing reporting systems, comparability concerns could be largely fixed with custom reports Collected data enables many additional analyses	Large scope provides representative data at the national level Good comparability as measured data is in standard format, provided that uniform points of measurement used No solid connection to subscription types possible Similarly measured data enables more advanced traffic analysis as well, at the expense of complexity	Usage data at individual level not attainable through other means Collected data enables a multitude of different advanced analyses and possibility to link usage data to background variables (subscription type, demographics…) All usage data comparable while same research platform used Scope limited to a small sample of smartphone users, although larger samples are possible
Measurement process • Resource requirements • Legal issues • Other issues	Mobile service operators (or MVNOs) able to measure Data collection is resource intensive for operators, depending on the usage of custom/existing reports	Mobile network operators (or MVNOs) able to measure Setup technically simple, some hardware and open source software needed Measurements on live operational networks require caution	Parties with capable research software able to measure, mobile operator participation facilitates recruitment Panelist recruiting complicated, especially in 1st time measurements Caution required due to sensitive nature of almost all measurement phases, potential bad publicity due to “big brother” -effect

All in all, each of the measurement method seems to have its own pros and cons. The methods provide partly overlapping data, as some data items are provided by two (or three) of the measurements. These cases could be considered as measures of reliability and validity. In cases where no such overlapping occurs, the scope of the measurements is in turn extended. The three measurement processes also burden the measuring parties in different ways. Thus, the resources needed for measurement implementation should be balanced with the value of the measurement results to each participant.

4 Measurement Setup

Three fundamentally different types of mobile Internet usage measurement were conducted in fall 2005. Two of the measurements were realized at the mobile networks, whereas the third measured usage directly from smartphone handsets. The network level measurements were based on two different methods, data collection using mobile operators' charging-oriented reporting systems and a separate measurement of mobile network packet data traffic. These measurements were conducted by all participating mobile operators according to a harmonized specification. The handset –based measurement used a separate research platform, essentially consisting of a client application or monitoring software which was installed to a group of handsets.

This chapter describes what kind of data was sought in each of the measurements, how the measurements were actually conducted, and what kind of processing and analysis tasks were necessary to obtain the final results. In addition, all the identified potential sources of error related to both measurement and analysis phases are also evaluated in this chapter. Actual measurement results are presented in chapter 5.

4.1 Operator Reporting System –Based Measurement

Realizing the operator reporting system –based measurements consisted of several phases. The process started with the preparation of a version of the measurement specifications indicating the kind of information that was sought. As measurements were conducted using mobile operators' existing information systems each of them having its own possibilities and limitations, the final detailed form of the specifications had to be iterated together with mobile operator personnel actually carrying out the measurements. The differences between operator-specific data had to be carefully described and documented. Finally, measured operator-specific data was combined and the resulting market-wide data was analyzed.

4.1.1 Measurement Specifications

The purpose of reporting system –based measurements was to collect general level data on mobile Internet usage at the national level. All data collection focused on **subscribers of participating MSOs** (Sonera, Elisa, DNA Finland), even though some data might have been available at the network more broadly. The primary scope of these measurements was on packet data usage of postpaid consumer subscribers, though in some cases data on business and/or prepaid subscribers was also required.

Data describing **mobile terminal installed base** was sought in the form of operator-specific complete or top 100 -lists of most popular terminal models with the percentage shares of total of each model, or corresponding data per terminal TAC code. Additionally, the total number of terminals was also required to enable combination of operator-specific terminal bases.

General level data on Finnish **mobile subscribers** was also to be measured. Among the studied data items were: number of subscribers by different types of subscription (postpaid/prepaid, consumer/business, packet data tariff plans), packet data traffic volumes by subscription type, packet data usage by mobile terminal radio technology, and general packet data usage and roaming activity. More specifically, the specifications included several simple questions requesting some relative or absolute figure on mobile subscribers or their packet data usage.

More detailed measurement specifications are presented in Appendix A. As the sought information relied on two fundamentally different data items, i.e. CDRs generated at the network and subscriber information located at the billing system, the questions in the specifications were divided accordingly to different subchapters.

4.1.2 Measurement Description

The information requested in the measurement specifications was acquired from both **existing and custom reports** produced by the information systems of the participating mobile operators. Existing reports were used whenever the data was already available in them, whereas separate custom reports were run in other cases. As the running systems of the operators were used, no separate setup of measurement equipment was necessary. However, lots of attention was required to assure sufficient level of comparability of data obtained using the three differing systems of three operators.

Data related to terminal models and terminal radio technologies was acquired from the **CDR data warehouses** (or databases), as a CDR including the IMEI code of the originating terminal is generated at the network after any chargeable transaction. An evaluation of the significance of Internet traffic was also possible at this level by comparing the packet data traffic going through the Internet APN to the packet data traffic of all APNs combined. Separation of postpaid and prepaid subscribers is possible at the network level, as a different IMSI range is allocated for prepaid subscribers by the operator. However, no information on the subscription type (e.g. consumer/business) of postpaid subscribers is available at the network level.

Information related to subscription types was obtained from operators' **billing systems**, and was thus limited to postpaid subscribers and volumes of chargeable packet data traffic. As registration of prepaid subscribers is not compulsory in Finland, more detailed data on them was not available from all the operators. However, as prepaid subscribers represent only a minority of all mobile subscribers and arguably have differing usage patterns than postpaid subscribers, scoping them out of most of the analysis was acceptable. Moreover, a more specific focus on Internet-bound traffic was not possible, as only the aggregated packet data traffic volumes are available at the billing system.

Mobile terminal base

Functionally, the **basis for data collection** was similar for each operator. Each terminal creating a certain type of CDR during the measurement period was taken into account once. Different terminals were identified by their IMEI codes, and were summed by the associated TAC codes giving the number of different terminals observed per TAC code. The TAC codes were then mapped to corresponding terminal models, and summed to model level as some terminal models correspond to several TAC codes, depending on policies of terminal manufacturers.

However, the **data sets** had some differences, as operators' existing reports on the mobile terminal base were used. These differences resulted from the types of CDRs included, and differing measurement periods. The first data set included all MSO's subscribers' terminals with made voice calls or sent SMSs during September 2005. The second data set included all MSO's subscribers' terminals with at least one transaction (phone call, SMS, other transaction) during September 2005. The third data set included all terminals observed at the network on week 34 (end of August) with no particular requirements on transactions or subscriptions as merely turning on the phone creates a CDR.

Differences in measurement period resulted in some error, as there was both churn and purchases of new handsets during the month-long measurements, and between the last week of August and end of September. According to Numpac (2006), some 115 000 mobile numbers were ported during September 2005 in Finland. This corresponds to less than 3% of the total of measured mobile terminal base, which might result in a small overlap between data sets if subscribers have moved from one measured operator to another while using the same handset. On the other hand, Numpac's churn numbers also include the 10-20% of mobile subscribers completely outside the measurements, which decreases the potential overlap. Moreover, one can assume that the terminal profile of number porting subscribers does not differ from general terminal profile in any significant way. Thus, churn should not bias the results much. On the other hand, churn also includes the cases where the operator is changed without porting the phone number, i.e. when a subscription is closed and another is opened. The significance of this type of churn is not known, but it supposedly mainly concerns prepaid and business subscribers and should be much less than the amount of ported numbers. Considering new handset purchases, 161 200 new handsets were delivered to retail in Finland during September 2005 according to the Association of Electronics Wholesalers (Digitoday 2005b). This might lead to a 4% overlap in the mobile terminal base data. However, these deliveries of new handsets should not have as large effect on the terminal base, as it is estimated that this number actually includes large proportions of transit traffic to Russia (e.g. ITviikko 2005). On the other hand, the problem of potential overlap in the terminal base caused by this can be overcome by simply considering that both the old and the new handsets that were used during the one-month measurement period truly belong to the Finnish mobile terminal installed base.

Differences in operator-specific data sets were another source of error. There is only a small difference between the first two data sets mainly resulting from the decision whether to consider uplink traffic CDRs only or to also include downlink traffic. However, the third data set differs from the others, as it potentially includes not only foreign roamers but also other operators' subscribers' terminals due to the requirement on emergency call readiness while not covered by the home network. Based on the subscriber data collected during the measurement, there is a maximum of 2,5% excess in total measured terminal base due to this. As the tendency to use other operators' networks for emergency call readiness is not related to specific handset models but to network coverage, this should not have a major effect on the distribution of different mobile terminal models. Foreign roamers, on the other hand, do cause some bias, as their terminals indeed do not belong to Finnish mobile terminal base.

Another issue worth noting is the fact that a mobile user might have **multiple mobile terminals and subscriptions** in use. For instance, one could have separate handsets for business usage and private usage, and yet another terminal (e.g. data card) specifically for data usage with subscriptions and/or SIM cards for each terminal. Nevertheless, this does not affect the results in any way, as it is reasonable to say that as long as these terminals and subscriptions are actively used, they indeed belong to the Finnish mobile terminal base and mobile subscriber population.

Mobile subscriber population

At a general level the measurement specifications detailed rather simple questions, for which operator-specific answers should mostly exists already. However, the basis for calculating even the most common key figures describing the mobile subscriber population differ, as was seen above with the differing data sets for mobile terminal base. Moreover, operators' use of non-uniform information systems and organizational differences make providing commensurable data very challenging. Thus, the obtained subscriber data was more or less heterogeneous, and lots of harmonization was required in the analysis phase.

Source data received from mobile operators consisted of relative and absolute numbers on numbers of subscribers and volumes of packet data traffic, accompanied by more or less detailed explanations on how the reports were generated, i.e. which parameters and conditions were used in the data warehouse runs. Data was collected from weeks 34, 38 and September 2005, with some variation depending on the operator and the data item in question.

There was a **billing delay** in at least some of the measurements as data on usage was available at the billing system only when the subscriber was billed for it. Subscribers are normally billed monthly if a euro threshold in bill size is reached, or later when the bill size is adequate. Some operators estimated that almost 90% of voice calls are billed after 2 months and about 95% after 3 months of the call. Assuming that packet data users also tend to be among the more active callers, their usage should be well captured in the measurements, as billing data was collected over 2 months after the latest studied period. Numbers of users and usage volumes on certain studied weeks can be extracted from the data once it is available in the billing system.

As with mobile terminal base related data, **differences in measurement period** resulted in some potential error. Obtained data was from the specified time periods in the majority of cases, with only a few exceptions having data on monthly instead of weekly level. Similarly to measurement data on mobile terminal base, churn during measurements might affect the data especially during longer study periods. While the terminal profile of people actively changing mobile operator is not known, it is reasonable to assume these people to be price sensitive. As prepaid phone numbers cannot be ported, high churn rates are probably largely caused by postpaid subscribers. Moreover, one could also assume churn to be somewhat more consumer-driven, as consumers are likely to react more to marketing campaigns. While this could lead to small overweighting of postpaid consumer subscribers in the data sets, the effect of this price sensitivity on distribution of subscribers by packet data tariff alternatives and packet data usage itself is unclear.

Moreover, as most of the data was from one week, the possible effect of churn is significantly lower than in the case of mobile terminal base.

The **passivity of subscribers** was also an issue in the measurements as data from some operators concerning subscriber numbers included active subscribers only, i.e. those with certain transactions during the measurement period, whereas other data contained all subscriptions regardless of their activity. These figures do not give quite the same results, as a subscriber might have been passive, and thus out of measured data during the study period. Naturally, passivity depends on the length of the study period itself and the type of usage studied. Voice calling and SMS are more extensively used than e.g. MMS and packet data, which is in fact the center of attention in research items related to packet data usage activity. However, there might be some subscribers who truly have remained passive for the entire study period biasing some of the more general level analyses concerning the entire subscriber population. While this “true” passivity concerns all subscribers, prepaid subscribers are subject to yet another type of passivity. As prepaid subscriptions are valid 12-14 months after the last credit recharge, a high proportion of them might in reality be abandoned by users, while still being considered as subscribers by operators. Thus, the number of prepaid subscriptions is presumably too high.

There were some **representativeness issues**, as few of the requested data items related to mobile subscriber population and packet data traffic volumes were not delivered by all operators due to high operational costs involved in acquiring them. In these rare cases further analysis was based on the data sets of two of the operators. While the smaller sample in these cases might have some effect on representativeness, no further analysis on this was conducted due to sensitivity reasons.

Another type of **imprecision regarding data on prepaid subscribers** was also present. While data on prepaid subscribers mostly covered the true number of subscribers, some data actually included the number of different terminals (identified by IMEI code) used by people with prepaid subscriptions. As people might have several terminals and subscriptions in use simultaneously, these two figures are not quite the same thing, as described before. Moreover, the issue of passive users presented above was also present in this case. While these issues made combining the data sets rather challenging, it was still possible by making several simplifying assumptions in the analysis phase.

There were some **differences in mobile operators’ packet data tariff plans**, which affected both data collection and combination of operator-specific data sets. Mobile operators offer various different packet data plans for consumer subscribers, which are not all entirely comparable with each other. Moreover, many subscribers might be billed according to an old tariff plan no longer publicly offered. This multitude of alternatives did not always conform to the general categories listed in measurement specifications, and data was thus obtained in varying level of accuracy. In some cases data was available for each and every tariff alternative whereas in others alternatives were grouped into a few aggregated categories. However, the potential bias caused by these differences was avoided by further aggregation of tariff alternative categories in the analysis phase.

There were **multiple reliability issues with roaming data**. While regarding general roaming activity, a part of the data sets included voice call roaming only, whereas other parts included SMS roaming as well. While previous research suggests that SMS roaming is used more actively than voice call roaming by Finns (Taloustutkimus 2005), partial measurement data showed no major differences between the two. Another somewhat similar difference between the data sets was whether only subscriber originated voice calls (and SMSs) were included or were both originated and received calls (and SMSs) considered. A broader issue concerning packet data roaming as well was again related to billing. In some data sets actual roaming during study period was considered, whereas in others roaming billed during study period (or actually a month after the study period) was taken into account. Billed roaming clearly isn't the same as actual roaming, but it gives a usable figure for weekly roaming assuming average roaming levels remain constant during non-peak seasons. Another issue concerning the representativeness of roaming data is related to the delays in the delivery of roaming data from foreign operators. This delay depends on several factors (e.g. home operator, foreign operator, type of roaming usage), and might have lead to the exclusion of some roaming data from the measurements. While this should not influence packet data roaming figures as long as home GGSN roaming is used, it does affect voice call roaming. Finally, the representativeness of the roaming figures in general is questioned by the possible variation in roaming activity between random weeks, as only one non-summer-vacation-biased week of roaming (week 38) was studied. In conclusion, the above differences between source data and the general representativeness issues question the overall accuracy of the roaming-related data.

4.1.3 Processing and Analysis of Measurement Data

Operator-specific measurement data concerning both the mobile terminal installed base and mobile subscriber population necessitated some further processing before a unified data set forming a basis for the subsequent analysis phase could be obtained and the analysis performed.

Mobile terminal base

The obtained three **data sets** had some subtle differences necessitating some further processing before uniting them was possible. The sole differentiating factor between the first two data sets was in their scope, as one data set gave a list of top 100 terminal models and the other a complete list of all terminal models. The third data set also gave a complete list of all observed terminals, but depicted terminal TAC codes instead of terminal model names for all cases, although terminal model names were also given for the majority of TAC codes. Most of TAC codes for which terminal model was not already available were mapped to appropriate model names using either listings published by BABT (British Approvals Board of Telecommunications) or data given by terminal manufacturers on their web sites. All lists included information on the total number of terminals, and the respective share of each terminal model (or TAC code) of the total. The three data sets were then combined by model information in order to have a file describing the Finnish mobile terminal installed base.

The mobile terminal base was then ranked by the share of total of each terminal model, and selected **terminal features** for all models representing cumulatively over 99% of terminal base total were sought. Web sites of terminal manufacturers were the primary source for the features of specific terminals. In addition, web pages of all major terminal manufacturers were examined in order to identify additional WCDMA capable terminals outside the 99% threshold, and a similar analysis was made concerning major smartphone operating systems (e.g. Symbian, Microsoft, Palm). These searches were complemented with several random web searches to ensure that the grand majority of all WCDMA and smartphone terminals outside the most common 99% were properly identified.

There were some 4,4% of **unidentified terminals** in the data sets. A part of terminal TAC codes could not be mapped to terminal models either by operators or during the analysis phase. Apparently all terminal manufacturers do not deliver the TAC code – terminal model mappings to TAC allocating organizations in real time, the results of which is that model information for most recent terminals is not up-to-date. Another part of unidentified terminals was formed by those terminal models outside the top 100 listing in one data set. Based on other data sets, the most recent terminal features are slightly less represented among the top 100 terminals than outside it. Thus, the unidentified terminals should be somewhat more advanced than identified terminals in general.

Another source of inaccuracy resulted from **lacking terminal feature data**. Some specific feature information was not collected or found for about 1 – 1,5% of all identified terminals. This mostly results from the selected 99% threshold, and only has a marginal effect on the distributions of various features. It is also worth noting that the process of manually searching specific feature information for over 100 different models is prone to subjective error by the researcher.

While the obtained terminal model level data already included manufacturer information, the **data on terminal manufacturer** given by TAC code -based mapping to terminal models was somewhat more ambiguous. As terminals of major manufacturers are made in multiple locations, TAC code ranges for a large terminal manufacturer also seem to be allocated for different factories and manufacturing facilities separately. Even when these manufacturing locations were aggregated together, over 100 different terminal manufacturers were still identified. Several contract manufacturers were identified among these manufacturers. These cases were accounted for the manufacturers whose brand is used during sales by using the data on model name and contract manufacturer to find the actual brand manufacturer. Possible error in this phase should only have a marginal effect on the results.

The data on **terminal model's year of introduction** was not completely reliable. Data for this analysis was mainly collected from the web site mobile.softpedia.com, which provides a year of introduction entry for the majority of common mobile handset models. Whenever necessary, this data was supplemented by data from terminal manufacturer press releases. The source of Softpedia's data is not indicated on the web site, although its data on other terminal features conforms to manufacturer-originated data. More generally speaking, the "year of introduction" is not very well defined. Whether this refers strictly to official introduction dates of specific terminal models or also to introductions of more "accidental" nature, is unclear. Moreover, the delay from terminal model introduction to start of sales has increased during recent years, and depends on manufacturer and the market, among others things. Thus, one should definitely question the reliability of this data while planning any further analyses on e.g. average terminal holding times.

Nevertheless, while the accuracy of this data can be questioned on multiple levels, using it can still give some useful indicative results.

Mobile subscriber population

Combined data set with average figures describing the entire Finnish mobile subscriber population was obtained by combining the operator-specific data sets. This was accomplished by counting together data sets' absolute numbers of subscribers and volumes of packet data traffic on all measurement periods for which data was available.

In cases where data items from different operators were not completely commensurable, **simplifying assumptions** were made to enable the calculation of average figures for the entire Finnish mobile subscriber population. Most critical commensurability issues resulted from differing measurement results related to the passivity of subscribers, as some measurement data included all subscribers and some active subscribers only. This could be partly managed by assuming that the share of passive postpaid subscribers during the study week is the same for all operators, regardless of whether actual subscribers or the number of terminals is considered. Moreover, as there might be numerous passive prepaid subscribers, their share was assumed to be in the range between 0 to 50% of all prepaid subscribers when calculating the results, again regardless of whether actual subscribers or their terminals were considered,

As a variety of alternative packet data tariff plans are offered to mobile subscribers, consumer subscribers were grouped into three **aggregated packet data tariff categories** to enable comparability between operator-specific data. Aggregation to a small number of groups was due to the constraint of source data mentioned earlier as well as sensitivity reasons, i.e. to prevent the operators from counting detailed figures on each others' subscriber base from more accurate level data. The category "No fixed fee" included subscribers having operators' default packet data tariffs, basically meaning subscribers with purely usage-based tariffs or those who hadn't activated packet data capability at the time of the study (currently packet data capability is enabled by default in all subscriptions of each of the operators). The "Small fixed fee" –category encompassed subscribers with a usage-based tariff which is decreased by a small monthly fixed fee (no longer offered), subscribers with a chargeable packet data activation during the study period, and the 2 to 50 MB (2, 10, 20 or 50 MB) block-based plans meaning plans where the tariff includes some monthly fixed fee for traffic volume up to the block size and an additional usage-based fee for the traffic exceeding the block size. The "Large fixed fee" –category included block-based plans from 100 to 500 MB (100, 150, 200, 300 or 500 MB) per month, and either partly (fixed fee for all usage during evening and night time, usage based fee during office hours) or fully flat-fee plans. Computational euros per megabyte packet data traffic tariffs for each category were also calculated with the actual tariffs used during the study period (Ficora 2005). Average tariffs were calculated by weighting each operator-specific tariff alternative within a tariff group by its actual number of subscriber. For block-based alternatives, per megabyte values were counted by dividing the monthly fixed fee by the average size of the block similarly to Ficora's computations. Similarly, the size of flat-rate plans was approximated to be 500 MB, i.e. the largest available block-based alternative. Due to the inaccuracy of the calculations, the results obtained with these calculations should be considered indicative at best.

Differences in the length of the measurements made combining the figures challenging. While one can assume that average weekly packet data volumes per some subscriber subcategory are multipliable to monthly level (and *vice versa*) in a straightforward manner, the same does not apply to the number of subscribers. The number of subscribers using a service at least once a month is more than the number using it at least once a week, but clearly less than four times the weekly number as part of the weekly users tend to use the service repeatedly. In theses cases data was transformed from monthly to weekly level by using the minimum (1/4 of monthly users) and maximum (all of monthly users) values for the possible weekly number of subscribers, and then giving the results with some level imprecision.

4.2 TCP/IP Header Collection –Based Measurements

The setup and implementation of traffic trace collection measurements essentially involved three stages. First, the measurement specifications detailing the kind of information sought were prepared. Then, the actual measurement organization was planned and the measurements were carried out. Finally, raw measurement data was processed and analyzed to obtain meaningful results.

4.2.1 Measurement Specifications

The aim of the measurement was to study consumer-originated mobile network packet data traffic. In addition to **general traffic characteristics** such as traffic direction (uplink/downlink) by transport protocol (TCP/UDP), **application protocol distribution** and **most popular web sites** were also studied. Furthermore, the intention was to identify the underlying subscriber **terminal device operating system**, and to associate the data on application protocols and web sites to this information. This identification would also enable an evaluation of the share of laptop/PC traffic from purely handset-originated traffic.

More detailed measurement specifications with both a description of the desired data and practical instructions for conducting the measurements are presented in Appendix A.

4.2.2 Measurement Organization

Traffic trace measurements were conducted at the nation-wide GSM/UMTS networks of two (out of three) Finnish mobile network operators. More specifically, two major Finnish MSOs (Sonera and DNA Finland) and about **50-60% of all Finnish mobile subscribers** were in the scope of the measurements. No data on other operators (Elisa, Saunalahti, TeleFinland, others) was obtained.

In order to capture as much of consumer generated Internet-bound packet data traffic as possible, the Internet APN was selected as the **point of measurement**. Technically said, the point of measurement was the core network Gi interface. While the points of measurements used in the two measurements were not identical, they are comparable as the traffic quantities (bytes, flows) of the first measurement were later multiplied by the actual number of GGSNs used. Thus, proper weights for the operators' traffic were obtained. Traffic going via Internet APN includes traffic from postpaid and prepaid subscribers, as well as business and consumer subscribers. In fact, other measurements conducted during the study reveal that about 90% of all packet data traffic in the mobile network (traffic going via all APNs) goes via Internet APN. Thus, the measurements included about 50% of all Finnish mobile network packet data traffic during the measurement period. The used measurement setup is described in more detail in the figure below.

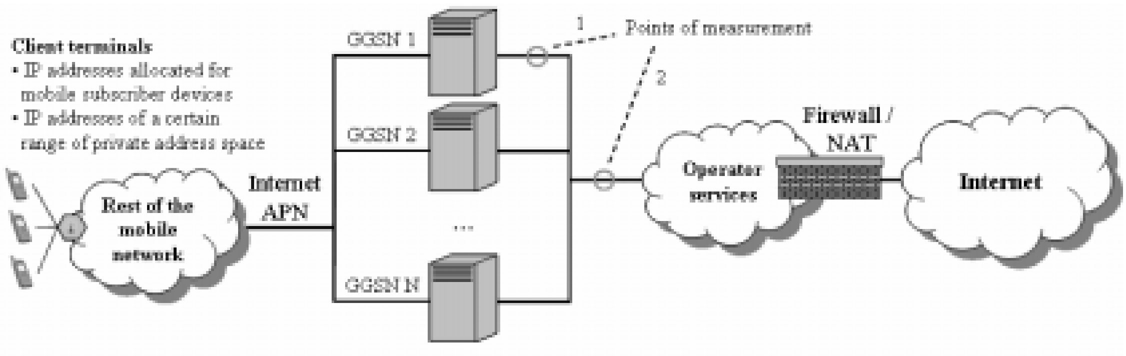


Figure 8 Measurement setup of packet data traffic trace collection

The *de facto* standard, open-source **software tool** tcpdump was run to capture TCP, UDP and IP packet headers during the measurement period. More specifically, all TCP packet headers of packets establishing and closing TCP connections (SYN, FIN, RST) with corresponding IP headers, and all UDP packet headers with corresponding IP headers, with all header fields of TCP, UDP and IP were captured. While HTTP-level header data would have given more accurate information on web traffic patterns, the more imprecise method relying on server IP addresses was chosen instead, as organizing the measurements this way was less complex, involved significantly less confidentiality issues. Moreover, IP address level data was actually deemed to be more stable in time than data on detailed web site addresses.

The measured data is free of the possibly biasing **influence of roaming**, as both operators used home GGSN roaming at the time of measurements, i.e. all packet data roaming traffic generated by an operator's subscribers is routed via home network GGSN. Thus, all packet data roaming traffic by the operators' own subscribers is included in the measured data, whereas no foreign roamers' packet data traffic is included.

The **measurement period** was set to one week and the measurements were conducted non-synchronically between weeks 38 and 40 (September 2005), which can be considered representative as major seasonal distortions to usage mainly result from summer holidays and the Christmas period. Moreover, the 1,5 week time difference between the measurements should not affect the comparability of the two data sets. Furthermore, no out of ordinary marketing campaigns, which could considerably affect mobile packet data usage, were observed during the study. Thus, the measurement period should represent an average week on fall 2005 rather well.

4.2.3 Processing and Analysis of Raw Traffic Data

The measurements resulted in about 20 Gb of raw data in pcap format. Some **pre-processing of raw data** was first conducted. The raw data originally in 6-7 daily dump files per measurement was merged into two larger week-long dump files to facilitate further analysis. These weekly dumps were trimmed down to include exactly seven consecutive 24-hour days of traffic data. Operator-specific measurement data was treated separately from each other until the last phases of analysis.

Raw data was then analyzed to determine the **client-server relationships**, i.e. which IP addresses belonged to mobile subscriber terminals (clients), and which were used by their destination servers. The subscriber terminal IP address range of one operator was publicly known, as the operator used public IP addresses. The other operator used private addresses (NAT), in which case the private address range used by mobile subscriber terminals was determined by direct observation of raw data. A later verification confirmed that all traffic flows in both data sets were either originated from or destined to these IP address ranges. All other IP addresses were considered servers. The information on client-server roles was later used when identifying application protocols and web servers. As both operators used dynamic IP addresses, no sanitation of IP addresses to protect subscriber specific usage information was necessary.

Subscriber terminal operating systems (OS) were identified using passive **TCP operating system fingerprinting**. Operator-specific weekly dump files were inputted to the open source TCP OS fingerprinting tool p0f (2005) with options to consider only flows originating from the subscriber terminal IP addresses. In other words, the aim was to identify only the subscriber terminal OSs. p0f determines the underlying OS for each TCP flow by comparing the traffic trace to its database of OS TCP fingerprints. Only the most reliable identification mode (the SYN mode) of p0f was used, meaning that only packets with the TCP SYN flag on were used in OS identification. This analysis resulted in a text file that contained a table giving the source operating system (if identified) of each subscriber terminal originated TCP flow, current terminal IP address, and the timestamp for the SYN packet (start of the TCP flow).

Open source CoralReef utility's crl_flow (or crl_traffic2) tool was used on the weekly dump files to **aggregate packet level data to flow level**. This analysis resulted in a text file containing a table of all TCP and UDP flows each described by a complete 5-tuple (transport protocol, source and destination port numbers, source and destination IP addresses), byte/flow/packet counts for the flow, and the associated timestamps giving the starting and ending time of the flow.

The **combining of operating system and flow level information** was done using a perl script separately prepared for the task. Essentially the script created a table from a p0f result file indicating the time frames each OS had resided in each specific client IP address. More specifically, a client IP address was accounted for a certain OS from the first subscriber terminal originated TCP flow identifying the OS until the next flow giving contradictory information. The table was then used to allocate the traffic flows in CoralReef output files to different OSs based on flow timestamps, and source or destination IP addresses. This last part of the script had three variants for three types of analysis. The first variant studied **general traffic patterns**, and counted the byte, flow, and packet sums of both uplink and downlink traffic per transport protocol (TCP/UDP) for all identified and unidentified subscriber terminal operating systems. The second variant was aimed at studying **application protocols**, and counted the byte, flow, and packet sums per transport protocol and server side transport protocol port number for each operating system. The third variant concentrated on **web traffic**, and included only flows with server side transport protocol ports 80, 8080, 8000, 8888, and 443 while counting byte, flow, and packet sums per transport protocol and server IP address for each operating system. UDP traffic was later removed to include TCP-based web traffic only. The three script variants were run on the p0f and CoralReef result files of both operators, each run resulting in a large text file with tables containing the variables described above.

The text files resulting from different script runs were imported into the SPSS software (Statistical Package for Social Sciences), where some **further processing** was conducted. The byte, flow, and packet counts of the result files of one operator were multiplied by the actual number of GGSNs used at the Internet APN by the operator to take into account the differences in measurement setup. Then, the operator-specific dump files were combined. The resulting data files were processed and aggregated in several ways to discover usage patterns in data. All graphs describing the measurement results were also drawn using SPSS.

Identification of application protocols was done using transport protocol server port numbers. This task proved to be ambiguous, as one of the operators used public IP addresses for subscriber terminals, which could sometimes result in reversed client-server roles. Indeed, some 64 000 UDP and 55 000 TCP server ports were observed, proving the presence of client ports in the range of IP addresses considered as servers. Nevertheless, those port numbers that could reliably be identified were mapped onto common application protocols. While it is known that (e.g.) some malware and P2P applications also use ports reserved for other applications, a reasonable assumption would be that the application protocols were mostly identified correctly. It is recognized that more accurate methods exist for identifying application protocols (e.g. traffic characterization). However, these methods were out of the scope of this study. Identified application protocols were further divided into 8 general aggregate categories, such as VPN, web, and email. Some 31% of traffic volume (bytes) and 18% of traffic flows were left uncategorized, mostly reflecting unidentified application protocols.

Analysis of web traffic patterns necessitated some additional processing. After removing all UDP traffic, different server IP addresses were extracted from the combined web traffic results files in SPSS. Another script was then prepared to resolve the domain names for these 75 000 different IP addresses. In cases where an IP address was hosting multiple domain names, the first PTR record was taken into account. This run resulted in a file mapping the IP addresses to some 73 000 different sub domain names. These results were once again imported into SPSS, where sub domain names were further aggregated into 19 000 different domain names. Information on domain names and domain categories was then merged with original web traffic data. Finally, most significant domain names (up to 90% of bytes and flows) for all major OSs were categorized into 14 categories based on researcher's subjective evaluation of the domain name and/or the contents of the first page seen in browser.

The **operating system identification possibly includes some bias**. The TCP fingerprinting method itself should be adequately reliable, as the fingerprint database should have entries for most of the common PC and smartphone operating systems.

While TCP fingerprinting identifies some 80% of traffic from fixed Internet traffic traces, the accuracy is probably somewhat lower in the mobile network due to the presence of a variety of devices with proprietary OSs. However, the process used in OS identification probably induces more bias. As only subscriber terminal OSs are identified, both uplink and downlink traffic are accounted for the terminal OS in question. Thus, the OS identification is based on uplink TCP traffic only, which corresponds to only about 37% of all flows and 5% of all bytes in traffic. In other words, downlink TCP flows and all UDP flows are accounted for different OSs based on uplink TCP flows. While the effect of the 63% of non-identified flows on OS identification accuracy is not known, one can assume a terminal's OS to be identified correctly following the first uplink TCP flow as long as the subscriber has the same IP address in use. Moreover, VPN traffic should be mostly allocated on the correct OS if VPNs are not on as a default, i.e. if there is at least one uplink TCP flow before VPN usage is started. The OS identification related to TCP based application protocols (e.g. web, email) should be more reliable.

There is some **potential error in the web traffic analysis**. First of all, some non-web traffic could be included in this data due to the inaccuracy of identifying web traffic by server port numbers. Moreover, a distorting effect results from the domain name resolving method, which essentially gives the domain names of hosting service providers instead of the actually visited hosted services. Finally, the categorization based on the domain name instead of more accurate sub domain name is not only very subjective, but ambiguous and prone to human errors as well, and results in somewhat overlapping categories. All these reasons lead to a rather high uncertainty, which is illustrated by the 40% share of unknown and uncategorized domains.

4.3 Handset Usage Measurement Setup

Measuring handset usage consisted of three separate phases. The pre-measurement planning and recruiting phase is central to making successful research, similar to all panel studies. Actually collecting usage data directly from handsets is challenging, even though it can be automated to a large extent. Finally, with usage data of sufficient accuracy on key variables, analysis on a variety of subjects is possible. All the above phases include some potential sources of error, the effect of which on the obtained results must be evaluated separately.

4.3.1 Panelist Recruitment Process

A **common approach** to panelist recruitment was attempted by the three participating mobile operators. However, due to some differing interpretations of the current legislation, alternative subscriber sampling and contacting methods were used.

The objective set for the recruitment was to acquire 100 to 200 **adult smartphone-using consumer subscribers** per participating operator to join the panel. As the recruitment rate was estimated to be rather low, each operator selected 10 000 subscribers from their subscriber bases to be contacted. For eligibility, the subscriber had to be an over 18 years old consumer subscriber, a user of one of the eight suitable Nokia S60 handsets, and either to have given his/her permission for SMS contacting or not explicitly prohibited it (depending on the operator). Apart from the criteria described above, the selection was random.

The **incentive** for joining the study had two elements. A 20 € voucher as a compensation for the increased data transfer costs incurred by the sending of the usage log files was sent to all participants showing reasonable participation effort. In addition, a lottery of 10 Nokia N70 smartphones was organized among the participants.

Two **contacting methods** were used. A common SMS message was sent to over 20 000 of the selected subscribers. The SMS message contained an invitation to join the study organized by TKK, a teaser on the possibility of winning a new smartphone, and a web address of the online registration site. More information on the study was provided at the online site. In addition, numerous operator customer service voice calls were made to the sampled subscribers to complement the recruitment. The called people were also instructed to visit the site for more detailed information. No classification of the contacted people was made, and simply the first 200 people per operator were able join. Based on the difference in time between the two contacting campaigns in September – October 2005, about 80% of the panelist were recruited by the SMS messages and resulting word-of-mouth, whereas the remaining 20% were due to the voice calls. However, the SMS recruitment rate could also have been somewhat higher as there was some excess of willing participants who did not fit into the operator-specific quota of 200 panelists.

The actual **sample recruitment** was completely handled by the online registration system. Registration was open to public, meaning that all users arriving on the site could join the study without any invitation, provided that they had a suitable handset in use. Completing the online registration process was somewhat laborious, taking about 10-15 minutes and required the combined use of a PC and the handset. Registration as a subscriber of an operator was closed after the operator’s quota of 200 was full.

Ending the panel was managed by sending a letter containing the 20 € voucher and a recommendation to fill the closing questionnaire at the study’s online site. The letter was sent to a panelist after he/she had been in the panel for two months. After completing the closing questionnaire, the panelist was provided with online instructions on how to remove the monitoring software from the handset.

There was probably some **bias due to the recruiting method** used regarding how well the panel represented actual Finnish smartphone user population. First of all, different types of incentives appeal to people in a different manner. All sorts of competitions, including the kind of marketing lottery used in this study, allegedly appeal more to younger people. On the other hand, the somewhat oversized 20 € compensation probably affected some people even more. The contacting channels also appeal differently to different people. While people react to a short SMS invitation in a different manner, this is further affected by the attitude towards mobile marketing in general, i.e. whether the person had given permission for SMS-based marketing efforts or had not separately denied it. Moreover, the voice calls made to recruit some people to the panel brought in panelists for slightly different reasons. How all this is related to panelist background demographics is yet another question. In previous studies it was found that the targeted users’ willingness to participate depended more on the used contacting channel than the size of the compensation. The above representativeness issues were recognized already while planning the recruitment, however, taking these issues into account in planning would have complicated the recruiting process significantly. Moreover, these imperfections can be considered negligible compared to the effect of the unknown amount of panelists who originally learned about the study by word of mouth or by other means. Finally, the used online registration method also acted as a screen in the selection of participants. As the would-be panelist had to install the monitoring on the handset him/herself albeit with the help of online instructions, the capability and willingness to install applications on the handset was a kind of prerequisite for participation.

4.3.2 Collection of Usage and Background Data

The monitoring software collects **usage information** on a variety of handset functions. In this study, however, mainly packet data and browsing related data was used. Moreover, some data on more general application and communication service related usage was also used to position the significance of data usage in the general handset usage context.

In addition, some **background information** on the panelists was collected in the beginning and closing questionnaires presented to the panelists. The variables collected in the beginning questionnaire included data on panelists' gender and age, and handset bill payer, whereas the main variable obtained from the closing questionnaire was the packet data tariff plan panelists had during the panel. Many other questions were also posed to the panelists in both of the questionnaires, but were not used in subsequent analysis. Completing the beginning questionnaire was a part of the registration process required from all panelists. While actual answering to all questions was not enforced, the background variables from the beginning questionnaire were obtained from over 90% of the panelists. As completing the closing questionnaire was only recommended, the answering rate dropped to about 62%.

While it is reasonable to assume that all possible erroneous entries in the usage data were successfully removed during the processing phase, there could be some **bias in questionnaire data** for several reasons. False data on background variables could result from panelists' mistakes and downright lying, or from the fact that the panelist did not know the right answer to a question. The latter especially applied to the question on panelists' packet data tariff plans, as not all consumers and even fewer business subscribers actually remember or know this.

4.3.3 Processing and Analysis of Usage Data

The collection of handset usage data resulted in about 200 MB of **raw data** in about 30 text files. All this data was inputted into SPSS where further processing was performed. The usage data was collected from over 500 Symbian S60 handsets that were monitored between September and December 2005. As a two-month panel participation target was embedded in the recruiting process, the first 8 panel participation weeks (panel weeks), or 56 days to be more exact, were taken into account for each panelist to enable equal weighting of individual usage.

Panelist activity was an issue during the analysis, as previous research (Verkasalo 2005) has shown that some panelists might be using the monitored handset only as a secondary handset. Panelist activity was determined by counting each panelist's number of active days, meaning days during which the panelist had had some calling or messaging activity, data usage or application launches with the monitored handset. Several panelists with passive non-usage periods ranging from 1 to 20 days during their panel participation were observed. Some panelists also had passive periods of over 5 days immediately after joining the panel after which more stable daily usage continued until the end of the panel. In these cases the panelist's first panel day was moved to the first active day after the passive period. Two separate measures for activity also used by Verkasalo were then utilized to determine the sufficiently active panelists. Firstly, a simple *active day threshold* was set at 28 days, meaning that each panelist had to have 28 active days during their first 56 panel days to be considered as active panelists. Secondly, an *activity ratio* was counted for each panelist by dividing the number of active days by the number of days in the panelist's active weeks, i.e. 7 times the number of panel weeks during which the panelist had active days. A threshold of 50% was set to this ratio. The two activity measures gave very similar results. While the second measure would comfortably include those panelists with some usage days during each panel week, the strictness of the first measure effectively made the second measure useless. The utility of the second measure is in panels with longer durations and less defined joining and quitting times for panelists, when less strict activity thresholds must be used. Eventually, 482 sufficiently active panelists were included in the final panel data. However, despite the use of the activity measures some passive periods were still observed in the data of active panelists. One can only assume this to be normal handset usage behavior, meaning that for some reason or another, some people indeed do not use their handsets in each consecutive day.

Before starting the actual analysis, **excess data was excluded** from each data table. First of all, data with faulty date information was excluded. In other words, cases with a timestamp outside the known panel duration or with no timestamp at all, were removed. Then, data on the passive panelists, and data on the active panelists outside their panelist-specific 56 day panel periods were excluded. Finally, some erroneous entries such as duplicate cases and outliers identified and defined by earlier studies were removed from the data. The entries related to the monitoring software itself were also removed. Further filtering of specific applications, for instance, was also performed in several analyses. Operator-specific usage data was not treated at any phase of the analysis.

5 Results

Results of the operator reporting system, TCP/IP header collection, and handset –based usage measurements are presented separately from each other in this chapter.

5.1 Results of Operator Reporting System –Based Measurement

Results of the data collection effort from three Finnish mobile operators' charging-oriented reporting systems are presented in this chapter. Data on some 80-90% of both Finnish mobile terminals and subscribers is presented.

5.1.1 Mobile Terminal Installed Base

Finnish mobile terminal installed base is described by distributions of terminals by individual model, terminal feature, terminal manufacturer, and terminal year of introduction.

Terminal distribution by model

The observed mobile terminal installed base was first studied by individual terminal model. The distribution of 50 most common terminal models with corresponding shares of total terminal base are presented in the figure below.

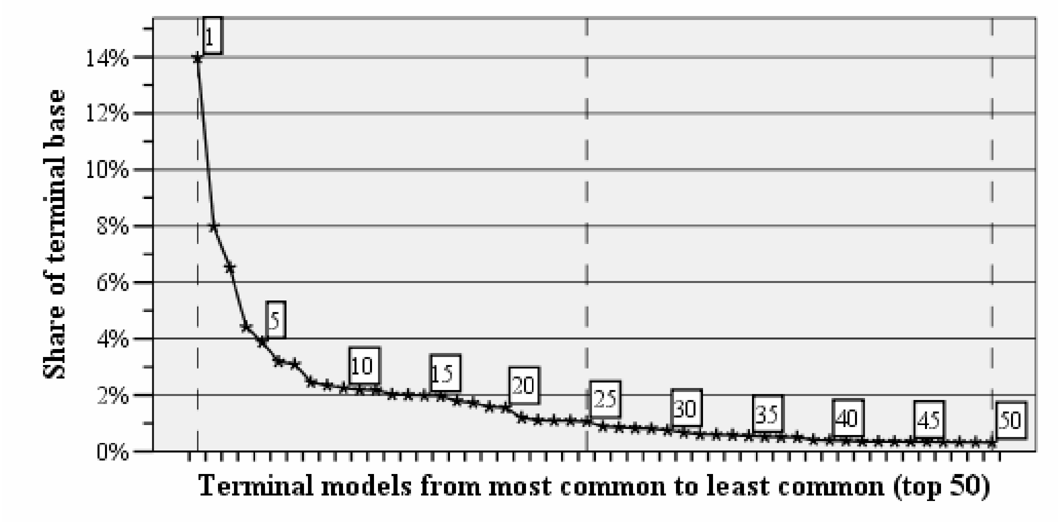


Figure 9 Mobile terminal distribution by individual model

Similarly to the above figure, the distribution of 150 most common terminal models with their cumulative shares of total terminal base are presented in the figure below.

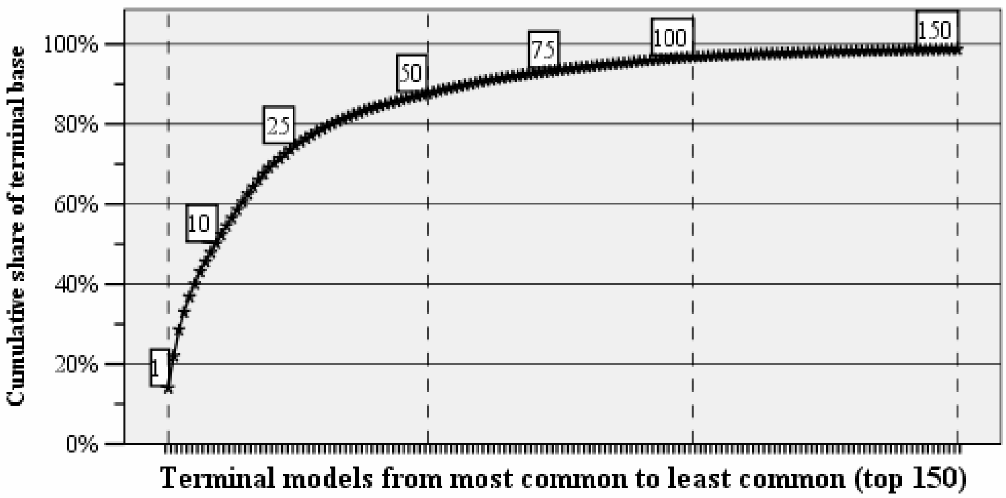


Figure 10 Mobile terminal distribution by individual model, cumulative

Judging by the above two figures, the terminal base seems to be fairly concentrated, as the top 50 terminal models make up 88% of all terminals and only 3% of all terminals are outside the top 100 models.

In total, over 1000 different terminal models were identified from the source data. The single most popular mobile terminal model was Nokia 3310 with a 14% share of all terminals. The first camera phone was the 11th, first smartphone the 21st, first WLAN capable terminal the 37th, and first WCDMA capable terminal the 54th most popular model.

Terminal distribution by feature

Mobile terminal installed base is also characterized by the penetration shares of certain key features. The share of all terminals supporting selected mobile terminal features relevant to packet data usage are presented in the figure below.

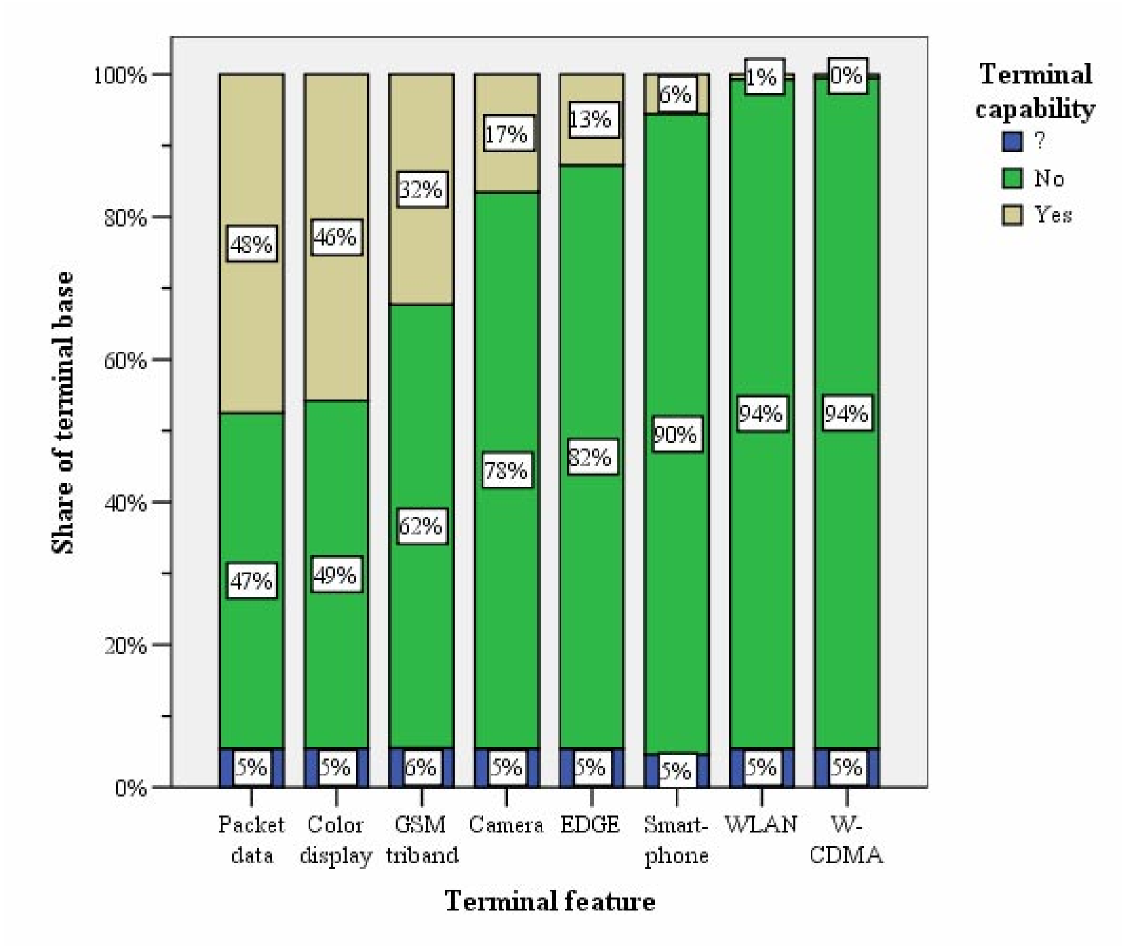


Figure 11 Mobile terminal distribution by selected features

A straightforward conclusion from the figure is that the key features for packet data usage are not widely spread. Packet data capability itself is in only 48% of terminals, whereas radio technologies supporting higher data transmission speeds have low penetration figures with the share of EDGE capable terminals being at 13%, WLAN capable terminals at 0,7%, and WCDMA capable terminals at 0,5%. The high share of WLAN compared to WCDMA is explained by the unusually large presence of Nokia communicators (see next sub-chapter). Moreover, about 5,5% of all terminals are smartphones, while there are color displays in 46% and cameras in 17% of terminals. Considering the GSM capability of terminals, the share of older single band terminals at 4% and dual band capable terminals at 58% seem rather reasonable. However, the share of GSM triband terminals is surprisingly high at 32% considering the 1900 MHz bands are only used in Americas. While some people specifically acquire triband capable handsets with traveling in mind, a more viable explanation for the high share could be the inclusion of GSM triband capability into most mainstream handsets as a generic feature. Finally, some 0,6% of the observed mobile terminals are not handsets or mobile phones, but some other mobile network using devices such as data cards and GSM/GPRS modems/modules. These terminals are included in all feature-specific considerations.

Considering the blue area representing the share of unidentified terminals or terminal features at the bottom of each bar, the share of all terminals supporting certain features is actually 0-6% higher than indicated by the brown area. More detailed analysis on the observed terminal base suggests the unknown to probably have similar or somewhat more advanced profile than identified terminal base.

Another remark is that all features will probably never be present in all terminals due to the growing variety of hardware and software –based features and the (e.g.) size and battery-life limitations of the terminals. Thus, each feature will probably approach some upper limit as its penetration grows. This limit depends on the segmentation strategies of manufacturers and the market in question, among other factors. One should also remember that the observed terminals also include a multitude of other interesting and potentially relevant hardware and software –based features not considered in this study. These features/capabilities include (but are not limited to) Bluetooth, video recording, HTML-browsing, Java, built-in hard drives, double screens, etc. One could also envision making further suggestions on these penetration shares based on the figure above, as the presence of some features is a prerequisite for more advanced features (e.g. camera / color screen, Java / GPRS and color screen, etc.).

Terminal distribution by manufacturer and smartphone type

The market shares of terminal manufacturers also give a worthy description of the mobile terminal base. Such a distribution is presented in the figure below.

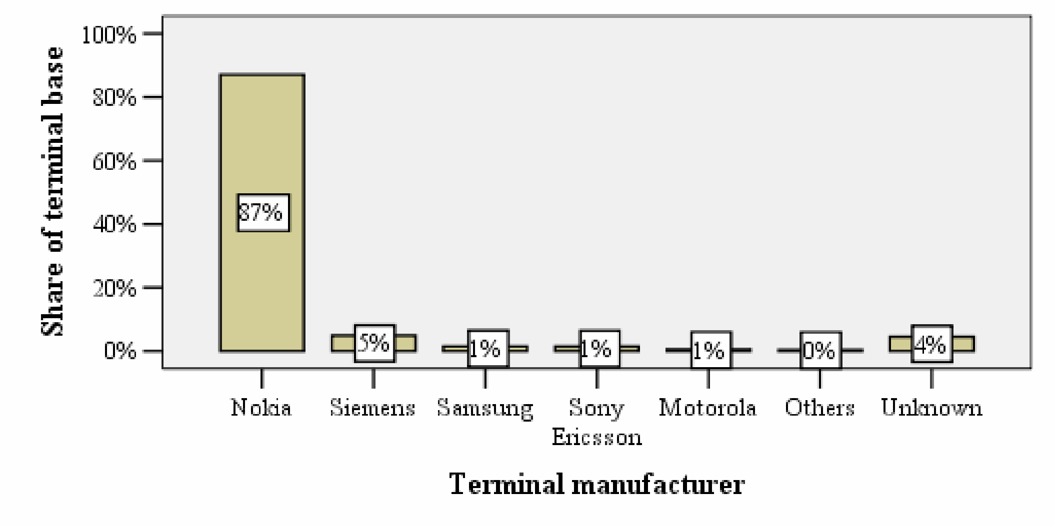


Figure 12 Mobile terminal distribution by terminal manufacturer

The first observation from the figure is Nokia's remarkable 87% market share in Finland. The first non-Nokia terminal model is actually only the 30th most common model. This is obviously explained by Nokia's Finnish origins. Siemens possesses the clear 2nd place, partly due to the presence of Siemens' GSM/GPRS modules, which are also considered as a part of the mobile terminal base. The shares of globally significant Motorola, Samsung, and SonyEricsson seem to be surprisingly low at about 1% each. The 4% share of unidentified terminals should raise these shares somewhat, although these terminals might as well include terminals manufactured by Nokia. Moreover, it is useful to remember that the collected data refers to the mobile terminal installed base, i.e. the terminals currently in active use, not the current market share of sales.

The distribution of advanced smartphone handsets to different smartphone types is an interesting addition describing the high-end part of the terminal base. The distribution of smartphone terminals by smartphone type is presented in the figure below.

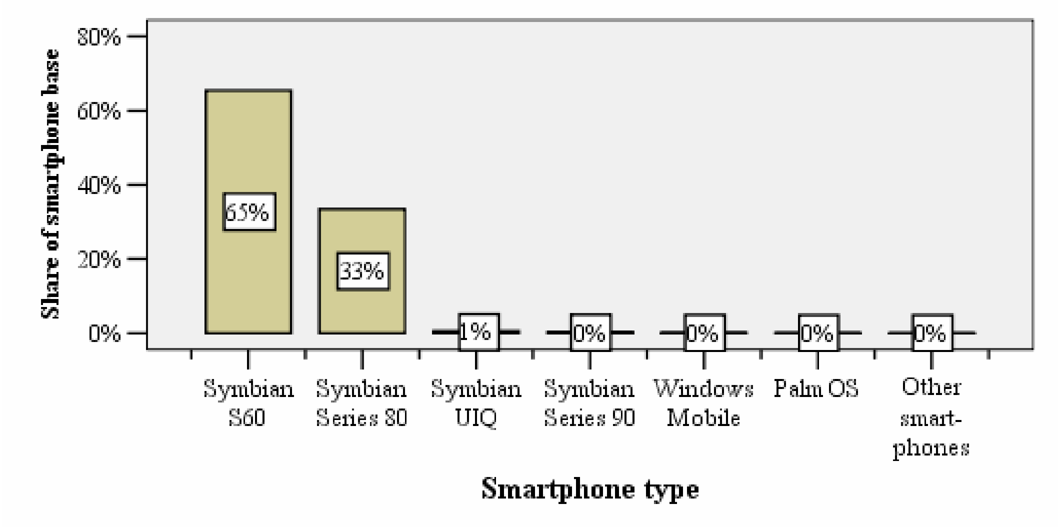


Figure 13 Mobile terminal distribution by smartphone type

Nokia's large market share in Finland is further illustrated by the smartphone distribution, as over 99% of all smartphones are Nokia-manufactured Symbian-based handsets. Another remark underlying the particularity of the Finnish scene is the notable 33% share of Nokia communicators (Symbian Series 80) of all smartphones. All other smartphone types, including Symbian UIQ and Series 90 handsets, Windows Mobile and Palm OS using handsets seem to be marginal.

However, the distribution of smartphone handsets is not entirely reliable, as even a small presence of smartphone among the unknown terminals (4% of all terminals) could significantly change the balance between smartphone types. Despite this, Nokia Symbian terminals will nevertheless represent the clear majority of all smartphones. Moreover, the shares in the above figure describe those mobile devices using the mobile network with a mobile operator's subscription and SIM card. Thus, all (e.g.) Palm OS using devices without mobile network access are not included in this data.

Terminal distribution by year of introduction

Finally, mobile terminal base is also characterized by the age of the terminals. A distribution of mobile terminals by their year of introduction type is presented in the figure below.

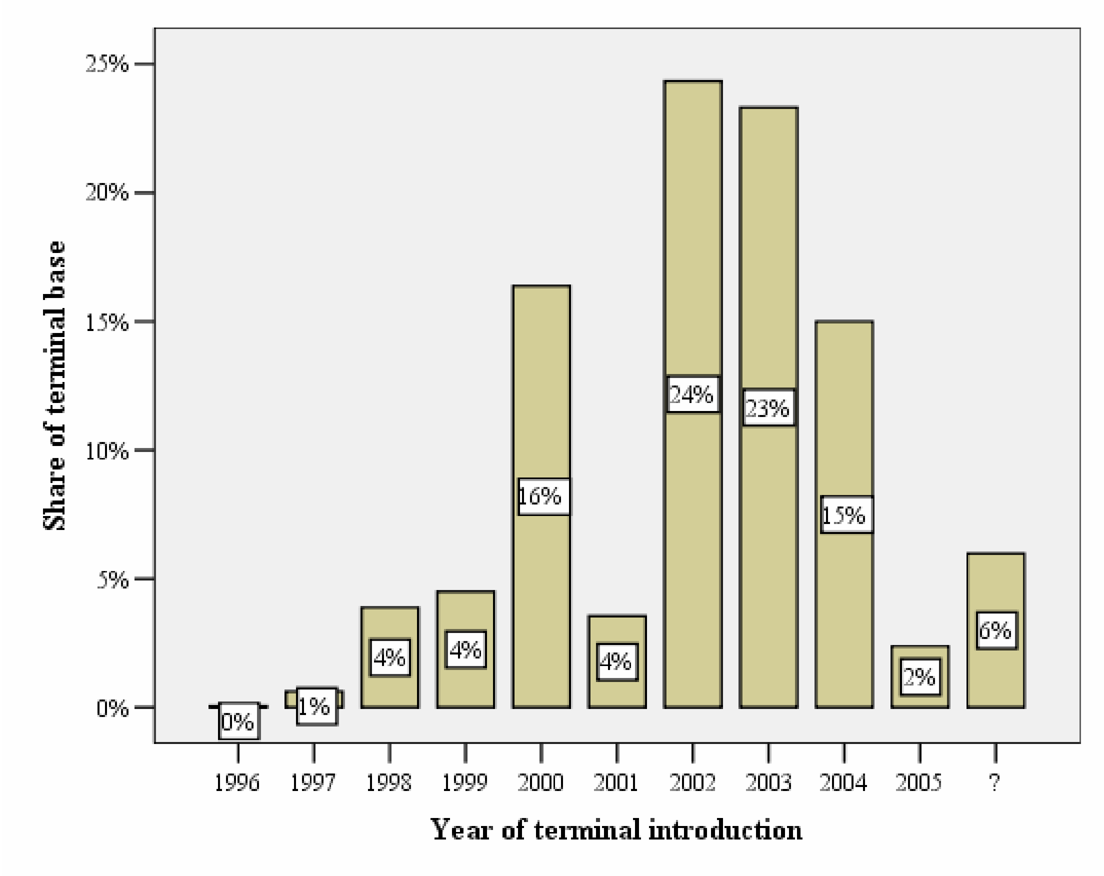


Figure 14 Mobile terminal distribution by year of introduction

A general remark is that Finnish mobile terminals are old, as the average (mean and median) year of introduction of all terminals is 2002. The share of most recent terminals, i.e. those introduced in 2005 seems rather small. However, this can be explained by the difference between date of introduction and the start of sales, as all handsets introduced in 2005 are currently (3/2006) still not sold, not to mention fall 2005 when the data was collected. At the other end of the figure, there are also some terminals introduced in 1996, which are still in use.

Regarding the general shape in the figure, the distribution seems to form a rather nice bell curve conforming to theory (product life cycle), apart from year 2001. What happened in year 2001? Could this result from burst of economic bubble at the turn of the millennium, or from terminal manufacturer actions, as introductions of the first WAP and GPRS capable terminals were done close to 2001? Or could this be explained by the huge popularity of some models introduced in 2000 or 2002 compared to models introduced in 2001? While further speculation is possible, no definite answer for the question can currently be provided.

5.1.2 Mobile Subscriber Population

Finnish mobile subscriber population is described by distributions of mobile subscribers by type of subscription; mobile subscriber packet data usage by terminal radio technology, consumer subscribers by packet data tariff alternative, and consumer subscriber packet data usage and roaming activity.

Mobile subscribers by type of subscription

The breakdown of all mobile subscribers to postpaid and prepaid subscribers, and postpaid subscribers to consumer and business subscribers with the associated shares of postpaid subscribers' packet data traffic are presented in the table below.

Table 5 Mobile subscribers by type of subscription

Type of subscription	Share of all subscriptions	Type of subscription	Share of postpaid subscriptions	Share of postpaid subscribers' packet data traffic
Postpaid	92 – 94%	Consumer	75%	38%
		Business	25%	62%
Prepaid	6 – 8%	N/A	–	–

The share of postpaid subscribers is very high compared to prepaid subscribers in Finland, as was expected. Moreover, consumer subscribers represent the majority of postpaid subscribers with a 75% share. Despite this, business subscribers generate the majority of packet data traffic, which means that average packet data usage volumes are significantly higher among business subscribers.

Mobile subscriber packet data usage by terminal radio technology

A distribution of mobile subscriber terminals by terminal radio capabilities is presented in the table below. The table contains the share of terminals with different radio capabilities (GSM/GPRS, EDGE, WCDMA) out of all packet data capable terminals (2^{nd} column), out of all postpaid subscriber terminals actually used for packet data (3^{rd} column), and the share of postpaid subscriber packet data traffic by radio capability (4^{th} column).

Table 6 Mobile subscriber packet data usage by terminal radio technology

Terminal radio technology	Share of packet data capable terminals	Share of postpaid subscriber terminals actually used for GPRS	Share of postpaid subscriber packet data traffic*
GSM/GPRS capable terminals	100%	100%	100%
EDGE capable terminals	27%	>27%	35%
WCDMA capable terminals	2%	>2%	16%

* Distribution of traffic by terminals of different radio capabilities, not the bearers actually used

The table shows that users of more capable terminals seem to use packet data more actively than those using less capable terminals. Terminals with higher radio capability represent a relatively higher share of terminals actually used for packet data than implied by their share of terminal installed base (exact figures for column 3 are not shown due to sensitivity reasons). Similar trend is visible in the last column, as more capable terminals generate relatively higher volumes of packet data traffic. This is especially clear for

WCDMA capable terminals. An easy explanation for this is that WCDMA capability is currently acquired specifically for packet data usage (e.g. 3G data cards), whereas EDGE capable terminals are possibly (also) bought to be used as simple phones.

Consumer subscribers by packet data tariffs

As explained in chapter 4.1.3, consumer subscribers were divided into three aggregated groups based on packet data tariff alternatives. In sum, “No fixed fee” basically included subscribers with purely usage-based tariffs, “Small fixed fee” encompassed subscribers with a usage-based tariff plus a small monthly fixed fee and the 2 to 50 MB block-based plans, while “Large fixed fee” included block-based plans from 100 to 500 MB and either partly or fully flat-fee plans. The distribution of consumer subscribers and their share of packet data traffic by alternative packet data tariff groups is presented in the table below. The table also presents values for average per subscriber chargeable packet data traffic volume for each tariff group.

Table 7 Consumer subscribers by packet data tariff groups

Packet data tariff group	Share of subscribers	Share of packet data traffic	Average chargeable packet data traffic volume per subscriber
No fixed fee	99,0%	18%	0,01 MB / week
Small fixed fee	0,6%	44%	3,93 MB / week
Large fixed fee	0,4%	38%	7,33 MB / week

The major remark from the table is that 99% of consumer subscribers have the operators’ default usage-based GPRS tariff alternative, while the remaining 1% of consumer subscribers in the fixed fee categories creates 82% of all consumer packet data traffic. In addition, judging by the last column of the table it seems that the volume of packet data usage increases as it gets relatively cheaper.

Consumer subscriber packet data usage and roaming activity

General consumer subscriber packet data (GPRS) usage activity and consumer subscriber packet data roaming activity are presented in the tables below.

Table 8 Consumer subscriber packet data usage activity

Packet data usage by consumer subscribers during study week	
Share of consumer subscribers using GPRS	8 – 9%
Average packet data volume per weekly GPRS using consumer subscriber	0,8 MB / week

The table shows that almost 10% of consumer subscribers used packet data during the study week. Considering the results of the previous sub-chapter, this seems somewhat higher than expected as only 1% of consumer subscribers had some fixed fee tariff for packet data transmission. On the other hand, this only means that people do use packet data also with relatively more expensive usage-based tariff plans.

Table 9 Consumer subscriber packet data roaming activity

Roaming by consumer subscribers during study week	
Share of consumer subscribers using voice (and SMS) roaming	3 – 4%
Share of consumer subscribers using GPRS roaming	0,1%
Ratio of GPRS roamers to voice roamers	2 – 6%
Average packet data volume per GPRS roaming using consumer subscriber	0,4 – 0,8 MB / week

Some 3-4% of consumer subscribers traveled abroad and used roaming during the study week. While roaming, GPRS seemed to be less used as only 2 – 6% of roaming subscribers used packet data. Moreover, the volume of packet data usage per roaming subscriber was also somewhat smaller than the usage volume at home. These smaller figures might result from the relatively higher prices of GPRS roaming compared to both voice (and SMS) roaming, and GPRS usage in the home network. Finally, the results related to roaming activity should be considered indicative only, as there were multiple reliability issues regarding the collection of this data.

5.2 Results of TCP/IP Header Collection Measurement

Results of the packet data traffic measurement on two Finnish mobile operators' core networks are presented in this chapter. First, some general packet data traffic patterns are presented. Then, the focus moves to the application protocol content of the traffic. Finally, the observed web traffic is studied.

5.2.1 General Packet Data Traffic Patterns

General packet data traffic patterns including the traffic distribution by operating system, transport protocol and traffic direction are now introduced.

Traffic distribution by operating system

The distribution of packet data traffic by operating system (lower graph) and the distribution of smartphone traffic by smartphone operating system (upper graph) by bytes and flows are presented in the figure below.

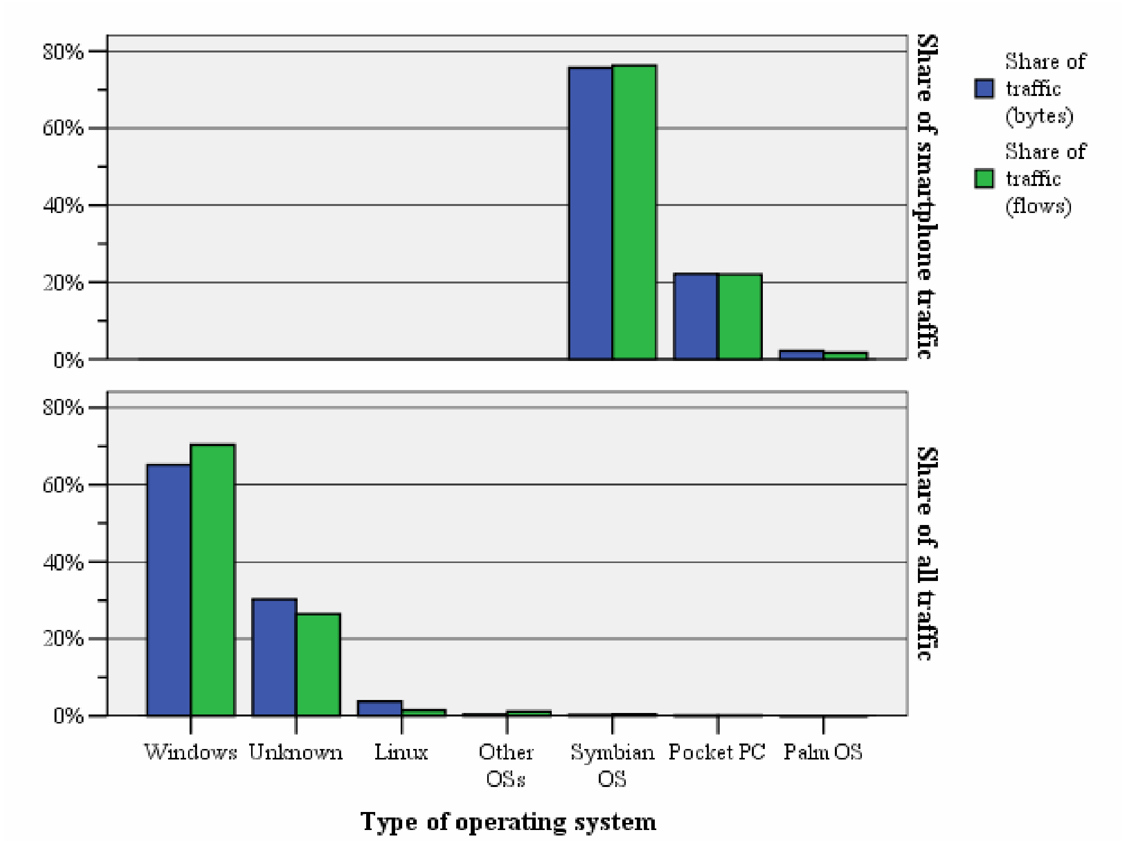


Figure 15 Packet data traffic distribution by operating system

The most striking result of the figure is that Windows originates (and receives) 65% of all packet data traffic in the mobile network. The presence of Windows can be explained by the usage of handsets as a modem via Bluetooth/cable, data cards, and GPRS modems, which are often used with laptop computers. The high share of Windows, in turn, is explained by the fact that a few PCs create more traffic than many mobiles, i.e. PCs tend to create lots of traffic, whereas handsets use simplified and packed mobile content specifically tailored for their limited capacity. Nevertheless, this underlines the importance of operating system identification, if truly mobile usage profile is to be uncovered. Another notable category is the Unknown operating system with a 30% share of all traffic, which aggregates over 500 unidentified OS TCP fingerprints. This category includes traffic of all non-smartphone handset, but possibly also some additional Windows and smartphone traffic as the OS identification process is not 100% accurate. Moreover, an unknown amount of telematics, machine-to-machine communications, alarm terminal, remote camera etc. type of traffic is also included in the Unknown category. Another issue is the potential altering effect of an intelligent modem / GPRS module, VPN usage or firewall on the TCP fingerprint used to identify the operating system responsible for the traffic. These issues make the high share of Unknown traffic somewhat troubling. The category Other OSs includes all other identified operating systems, such as Mac OS, but no proprietary mobile operating systems.

Regarding the smartphone traffic, the over 20% share of Pocket PC seems surprising considering its less than 1% share of smartphone terminals observed in the network (see chapter 5.1.1). However, there are supposedly many Pocket PC devices without GSM capability that use other devices for network access, which could lead to their larger share of traffic. One could also speculate that Pocket PCs tend to generate more traffic than Symbian terminals. Despite these arguments, one should be somewhat skeptical towards the results regarding the distribution of smartphone traffic, as even 1% of Symbian traffic in the Unknown category would alter the smartphone traffic distribution considerably.

Traffic distribution by transport protocol and traffic direction

The distribution of packet data traffic by transport protocol (TCP and UDP) and traffic direction (uplink and downlink) for the major identified operating systems is presented in the figure below.

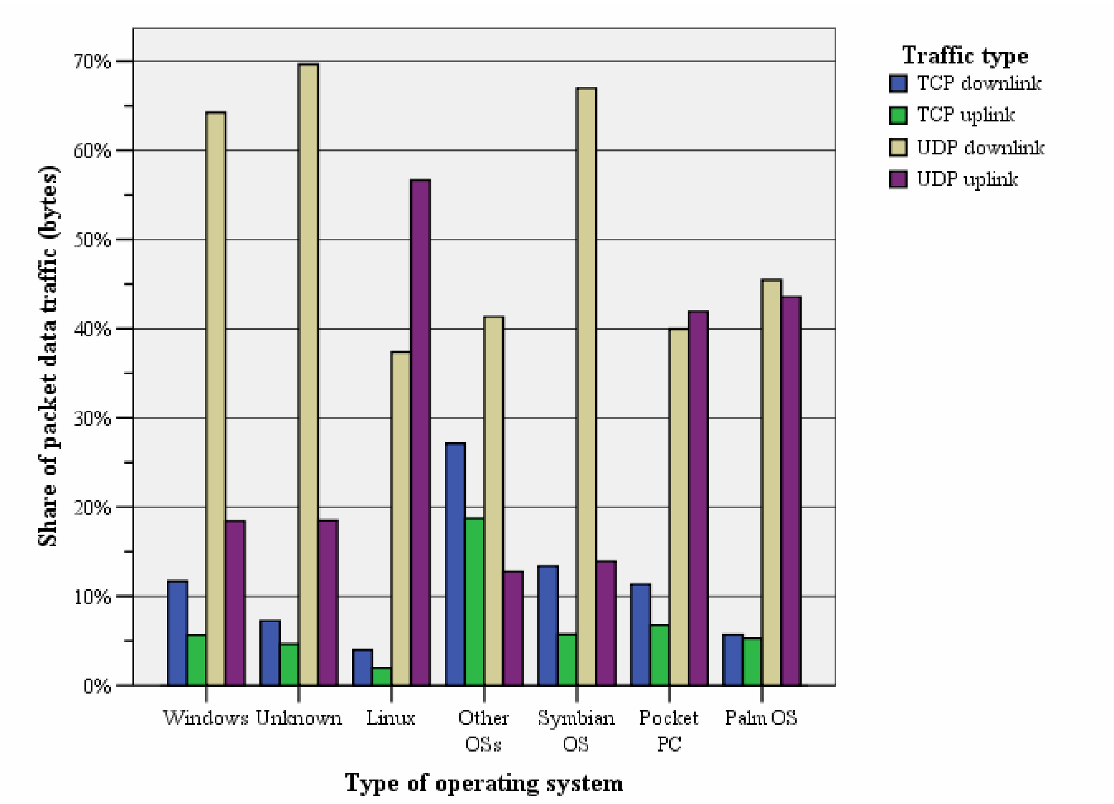


Figure 16 Packet data traffic distribution by transport protocol and traffic direction

The remarkably high 85% share of UDP traffic, illustrated by the two rightmost bars of each operating system, is the most surprising result. This share is pretty much the opposite in favor of TCP in the case of typical fixed network packet data traffic. In general, UDP is used by VPN protocols (for NAT traversal), DNS, WAP, MMS, and streaming, among other application areas. In this case, the high share is explained by VPN and DNS, as described in chapter 5.2.2. Downlink traffic is more important than uplink traffic in most cases, as could be expected.

Comparing the traffic profiles of different operating systems, Windows and Unknown seem to be fairly similar at this level of analysis. Could the Unknown traffic actually be dominantly Windows traffic as well? Moreover, the Symbian profile seems rather similar to Windows profile and different from the profiles of other smartphones, which in turn are very similar with each other. The profile of Symbian is partly explained by the presence of VPN traffic similarly to Windows, which is absent from Pocket PC and Palm OS traffic, as is seen later in chapter 5.2.2. Finally, the profile of Linux (4% of all traffic) is the most extreme with over 90% share of UDP traffic dominantly in uplink direction. No reasonable explanation for this can be provided, apart from the possible presence of Linux server(s) in mobile subscriber terminal IP address space.

5.2.2 Packet Data Traffic by Application Protocol

Regarding application usage, packet data traffic patterns are described by application protocols and a categorization of these protocols. A combination of application protocol categories and operating systems is also introduced.

Traffic distribution by application protocol

Application protocols were identified by TCP and UDP server port numbers, as described in chapter 4.2.3. The share of bytes and flows to and from the most significant TCP and UDP server ports are presented in the table below.

Table 10 Packet data traffic distribution by application protocol

Rank*	TCP port	Share of TCP traffic		UDP port	Share of UDP traffic	
		Bytes	Flows		Bytes	Flows
1.	80	80%	74%	53	7%	54%
2.	443	7%	8%	2746	17%	0%
3.	135	1%	3%	4500	15%	0%
4.	110	1%	2%	10000	13%	0%
5.	143	1%	1%	370	12%	0%
6.	445	1%	1%	500	8%	0%
7.	8080	1%	1%	4672	0%	7%
8.	1863	0%	1%	0	3%	0%
9.	25	0%	1%	5003	2%	0%
10.	7171	0%	0%	32555	2%	0%
11.	6346	0%	0%	6346	0%	2%
12.	4662	0%	0%	39273	1%	0%
13.	4283	0%	0%	9183	0%	1%
14.	28467	0%	0%	45991	1%	0%
15.	139	0%	0%	123	0%	1%
Others	Others	6%	7%	Others	16%	35%

* Ranked by TCP/UDP port's combined share of bytes and flows

Assuming the observed most common server ports actually correspond to the mainstream application protocols, most of the traffic contents can be explained. TCP traffic consists dominantly of web traffic (HTTP: 80 and 8080, HTTPS: 443), although email (POP3: 110, IMAP: 143, SMTP: 25) also has a small share of traffic. In addition, some Windows “self-initiated” traffic (135, 445) is also present. UDP traffic includes mainly DNS (53) and VPN traffic (CheckPoint UDP Encapsulation: 2746, IPsec / NAT-Traversal: 4500, Network Data Management Protocol / Cisco IPsec VPN: 10000, ISAKMP / IKE: 500). Moreover, some anti-virus traffic was observed (F-Secure updates / BackWeb: 370). UDP traffic also seems to be less concentrated to few application protocols, as its share of traffic to/from other server ports beyond the top 15 is much higher than for TCP.

Traffic distribution by application protocol category

The application protocols identified by transport protocol port numbers were then grouped into 8 aggregate categories to facilitate comparisons. A description of the basis of categorization is presented in appendix B.1. The aggregated categories with the major included protocol ports and their combined share of traffic (bytes and flows) are presented in the table below.

Table 11 Packet data traffic distribution by application protocol category

Rank*	Application protocol category	Major protocol ports included**	Share of traffic	
			Bytes	Flows
1.	Web	TCP: 80, 443, 8080	13,6%	64,0%
2.	VPN	TCP: 10000 UDP: 500, 2746, 4500, 10000, 1194	45,5%	0,1%
3.	DNS	UDP: 53	6,2%	12,0%
4.	Multimedia / IM	TCP: 1863, 5001, 6667, 554 UDP: 5003, 5001, 5000	2,7%	0,8%
5.	Email	TCP: 110, 143, 25, 993, 995	0,4%	2,7%
6.	P2P / file transfer	TCP: 6346, 4662, 21, 20, 1214 UDP: 4672, 6681	0,5%	2,4%
7.	WAP	TCP: 9200, 2949, 2805, 2923, 4035, 2948, 4036, 9202, 9201 UDP: 9201, 9203, 9202, 9204, 2805, 9200, 2923, 4036, 4035	0,1%	0,0%
8.	SSH / telnet	TCP: 22, 23	0,0%	0,0%
	Other	TCP: 135, 445, 7171, 4283, 139, 28467, 19977, 8081, 50123, 1435 UDP: 370, 0, 32555, 39273, 45991, 6346, 48000, 49000, 10001, 9183, 9181, 8889, 123, 434, 137, 9872, 12345	31,1%	17,8%

* Ranked by the category’s combined share of bytes and flows
** TCP and UDP server ports with at least 0,5% of the total bytes or flows in the category

VPN traffic seems to be in a central role with its 46% share of traffic volume. Its proper identification is supported by the very small number of flows, as VPN connections are typically contained in a single traffic flow. Apart from VPN, web and DNS are also major applications with 25% and 11% shares of non-VPN traffic, respectively. The over 30% share of uncategorized protocols illustrates the ambiguity related to the identification method used. The uncategorized protocols include self-initiated Windows traffic, traffic to/from client ports, malware, P2P, etc, and probably also some “Multimedia / IM” protocol traffic (audio/video conferencing protocols, streaming, IM, IRC). The observed share of email is fairly low, but both VPN and web (webmail) are likely to include lots of email traffic. Otherwise, the share of P2P is much smaller than in typical fixed network measurements, although such share does depend a lot on the point of measurement used in the fixed network measurements. Finally, the share of WAP is very small, as WAP APN traffic was not measured in the study.

Regarding entire application protocol profile of the observed traffic, it is advisable to remember that there is another and possibly different protocol profile inside the VPN. However, it is reasonable to assume the VPN usage to be almost entirely generated by business subscribers. Thus, the observed protocol profile can be considered representing “consumer natured” usage, i.e. either usage of consumer subscribers or usage of business subscribers for non-occupational purposes.

Traffic distribution by application protocol category and operating system

Combining the application protocol categories and operating system identification presented in previous chapters enables an analysis of traffic profile specifically for Windows and Symbian terminals. These operating systems were chosen, as Windows is by far the most significant observed OS, and Symbian is the largest identified handset OS. The traffic profile of the Unknown operating system is not further studied due to its heterogeneity, although further analysis might reveal something on the terminals contained in it. Moreover, Windows and Symbian are supposedly identified correctly and have enough traffic volume to feasible analyses, which is not the case with the other identified smartphone operating systems. Despite this, the data contains lots of uncertainty as the error present both in application protocol categorization and OS identification is combined in this analysis. A figure presenting the distribution of packet data traffic to the application protocol categories introduced above for both Windows and Symbian is included in appendix C.1.

This figure underlines the fact that Windows imposes itself on the general traffic profile. Another remark is that VPN seems to be used also on Symbian devices. Indeed, some of the nearly 100 000 Nokia communicators in the Finnish mobile terminal installed base do have VPN clients (Nokia Mobile VPN Client, related to Check Point technology).

5.2.3 Web Protocol Traffic Patterns

Similarly to the previous chapter, web protocol traffic patterns are studied by simple ranking of most popular domains, a categorization of the domains, and a combination of the categorization with operating system identification.

Most popular web domains

The most popular web domain names were identified from the captured traffic, including the server IP addresses of all TCP flows with server ports 80, 8080, 8000, 8888, and 443. The most popular web sites (excluding operator domains) ranked by the share of web traffic volume (bytes) and web site visits (approximated by number of flows) are presented in the table below.

Table 12 Most popular web domains

Rank*	Domain name of site	Share of web traffic volume	Share of web site visits**
1.	mtv3.fi	4,3%	3,9%
2.	doubleclick.net	4,9%	2,4%
3.	basefarm.net	2,5%	3,4%
4.	irc-galleria.net	3,1%	2,6%
5.	luukku.com	1,6%	1,5%
6.	google.com	0,9%	1,5%
7.	hotmail.com	1,4%	0,9%
8.	nebula.fi	1,0%	1,1%
9.	yahoo.com	1,2%	0,9%
10.	sihteeriopisto.net	1,1%	1,0%
11.	adtech.de	1,0%	1,0%
12.	sampo.fi	0,9%	0,8%
13.	htv.fi	0,8%	0,8%
14.	yle.fi	0,9%	0,8%
15.	akamaitechnologies.com	0,7%	0,8%
	Other identified domains	45,4%	47,7%
	Unknown addresses	22,1%	24,3%
	Private addresses	6,2%	4,9%

* Ranked by the domain's combined share of bytes and flows
** Share of TCP flows: # of web site visits <= # of flows <= files downloaded from site

The table suggests that mobile network web traffic is not very concentrated to few domains, as only about 48% of all traffic volume (including operator domains) goes to/from the top 10 domains and 77% to/from top 100 domains. Otherwise, the table mostly includes the domains of well known Finnish and international companies. However, the relatively high share of unknown (22%) and private (6%) addresses increases the uncertainty of these figures. The unknown addresses represents (web/other) server IP addresses for which no reverse DNS entry was available, probably also including some non-web traffic, e.g. P2P traffic using web ports to traverse firewalls.

Most popular web domains by category

The web traffic domain names identified above were then grouped into 10 aggregate categories to facilitate comparisons. A description of the basis of categorization is presented in appendix 0. The aggregated categories with the major included domain names and their combined share of traffic (bytes and flows) are presented in the table below.

Table 13 Most popular web domains by category

Rank*	Site category	Major sites included**	Share of web traffic volume	Share of web site visits
1.	Information	mtv3.fi, yle.fi, sanomawsoy.fi, almamedia.fi, helsinginsanomat.fi	12,9%	12,8%
2.	Entertainment	irc-galleria.net, veikkaus.fi, sm-liiga.fi, telkku.com	6,0%	5,9%
3.	Operator site	–	5,6%	6,3%
4.	Advertising	doubleclick.net, adtech.de, tradedoubler.com	7,4%	4,5%
5.	Messaging	luukku.com, hotmail.com, gmail.afraid.org, msn.com, passport.com, passport.net	5,3%	4,4%
6.	Adult content	sihteeriopisto.fi, seksitreffit.fi	3,6%	2,7%
7.	Web search	google.com, yahoo.com	2,1%	2,5%
8.	Banking	sampo.fi, eQonline.fi, op.fi, nordea.fi	2,1%	2,4%
9.	E-commerce	huuto.net, mobile.de, infosto.fi, thomann.fi, ebay.com, verkkokauppa.com	1,0%	1,1%
10.	Mobile content	jippii.net, jamster.com, mobilenator.com, buumi.net	0,6%	0,8%
	Hosting / corporate site	basefarm.net, nebula.fi, akamaitechnologies.com	14,2%	15,3%
	Other	–	10,8%	11,8%
	Unknown	–	22,4%	24,5%
	Private	–	6,2%	4,9%

* Ranked by the domain's combined share of bytes and flows
** Sites with at least 5% of the total bytes or flows of the category

The combined share of information and entertainment, infotainment, is about 20% of all web traffic. The share of advertising (e.g. pop-up windows) is also a notable 7%, while the share of sites controlled by Finnish operators (mobile 75%, fixed 25%) is about 6%. The high share of hosting is explained by the domain name resolving method used. There is, however, some overlapping in the categories (e.g. information and entertainment, messaging and web search, mobile content and operator sites…), as the categorization was based on the domain name instead of more accurate sub domain name. One should also notice the presence of multiple sources of error regarding these results that arise from the application protocol identification and web domain categorization. Thus, these results should be considered as indicative only. Finally, it is worth remembering that the web traffic potentially included in the VPN traffic might have an entirely different web traffic profile.

Most popular web sites by category and operating system

Combining the web domain categories and operating system identification again enables an analysis of traffic profile specifically for Windows and Symbian terminals. Similarly to the case of combining application protocol data with operating system data, the potential sources of error multiply in the combination process. A figure presenting the distribution of web traffic to the web domain categories introduced above for both Windows and Symbian is included in appendix C.2.

This figure again shows how Windows imposes itself on the general web profile. The Symbian device web traffic profile in turn could be considered to indicate the web-based services actually usable on a small handset display. Such an assumption would lead to the conclusion that web messaging, banking, and e-commerce are not really usable with the current Symbian devices.

5.3 Results of Handset Usage Measurement

Results of the handset usage measurement on some 500 Finnish Symbian S60 smartphone handsets are presented in this chapter. First, a description of the monitored panel of users is provided. Second, some general handset usage patterns are presented. Third, general packet data usage patterns are studied. Fourth, the packet data generating application usage is treated. Fifth, some smartphone browsing patterns are introduced.

5.3.1 Description of the Panel

The panel started on 7.9.2005 and was closed on 7.12.2005. In total 482 sufficiently active panelists were included. The three participating mobile operators were roughly equally represented in the data and the geographic distribution of panelists was not known. More detailed distributions of panelists by gender, age, handset capability, packet data tariff, and bill payer are presented next.

Panel by gender and age

The distribution of panelists by gender and age is presented in the figure below.

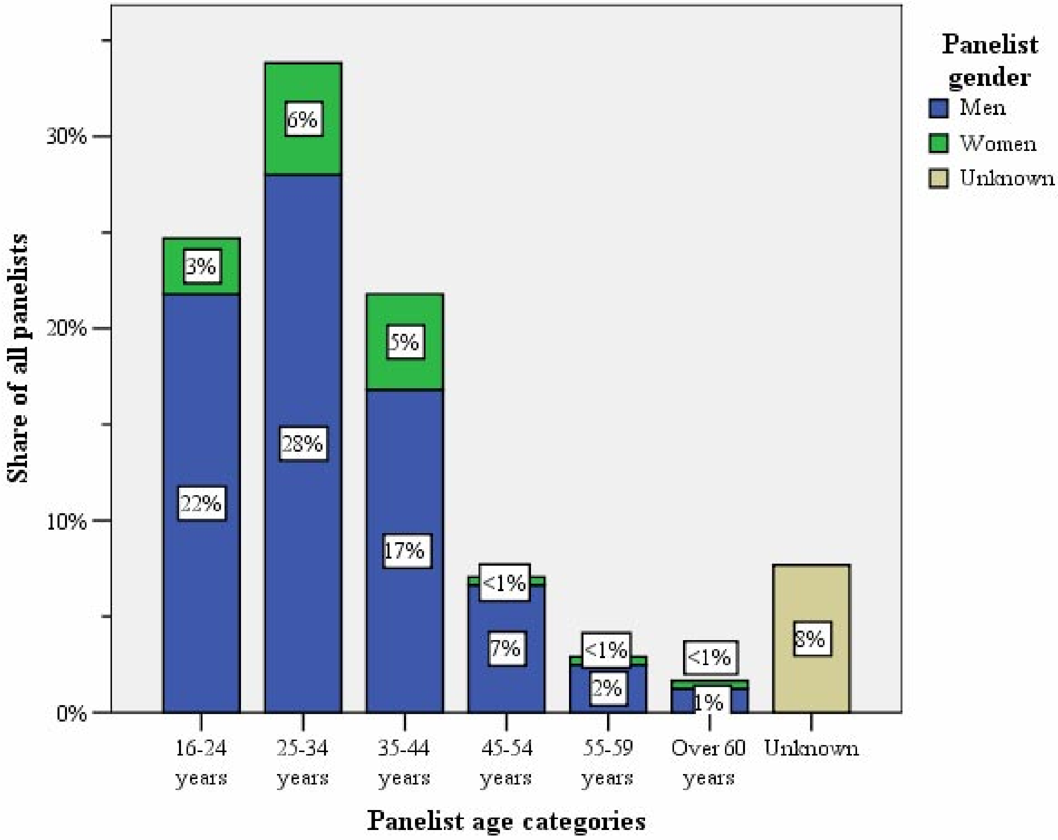


Figure 17 Panel by gender and age

The panelists were dominantly young to middle-aged men, as the share of men was 77% and the share of under 34-years-olds about 59%. The share of women in the more mature age categories was particularly low. For 8% of the panelists no demographic background variables were available. These cases were excluded from the data when analyzing the effects of demographics to usage.

The overall distribution in the above figure surely does not correspond to the distribution of general mobile subscriber population, as almost all Finns have mobile subscriptions nowadays. However, a more relevant question is whether or not this corresponds to the demographic distribution of smartphone users or e.g. users of advanced data services. In other words, does this result reflect the profile of the current Finnish smartphone user population, or is it partly a result of the recruiting method used?

Panel by handset type

The panelists used 8 different Nokia Symbian S60 handset models to participate in the study. At least one third of the handsets used by the panelists were introduced almost three years ago, which implies that the panel did not consist only of the most recent high-end models. The handsets were classified by the handset's highest radio capability (henceforth: handset capability). Thus, WCDMA refers to handsets with WCDMA, EDGE and GSM/GPRS capability, whereas EDGE to handsets with EDGE and GSM/GPRS capability. Moreover, in all subsequent analyses WCDMA, EDGE and GSM/GPRS refer to the capability of the handset, not to the bearer actually used. The distribution of panelists by handset capability is presented in the figure below.

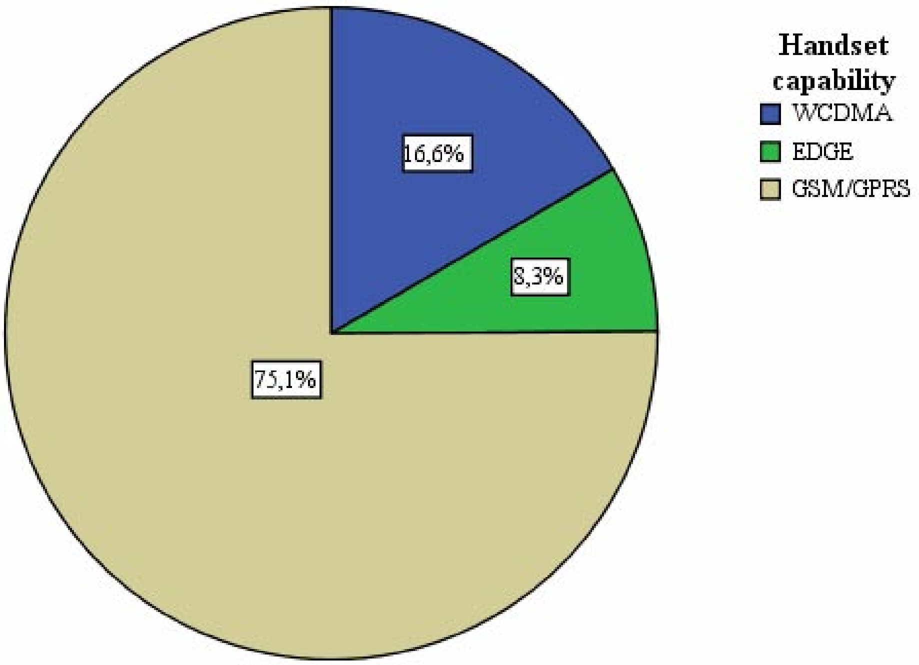


Figure 18 Panel by handset capability

The panelists were dominantly users of GSM/GPRS capable handsets. In addition, the share of EDGE capable handsets was relatively small compared to other handsets, which made some analyses on the usage of EDGE capable handsets somewhat less reliable due to the smaller sample size.

Panel by packet data tariff category and bill payer

Panelists were divided into three aggregated categories based on their alternative packet data tariff plans, similarly to the categorization used in reporting system –based measurements. “No fixed fee” included purely usage-based tariff plans, “Small fixed fee” included block-based plans with monthly block sizes from 2 to 50 MB, and “Large fixed fee” included blocks from 100 to 500 MB and flat-rate plans. Moreover, the panelists for whom background information on packet data tariffs was not available formed a fourth category. The distribution of panelists by packet data tariff category and handset capability is presented in the figure below.

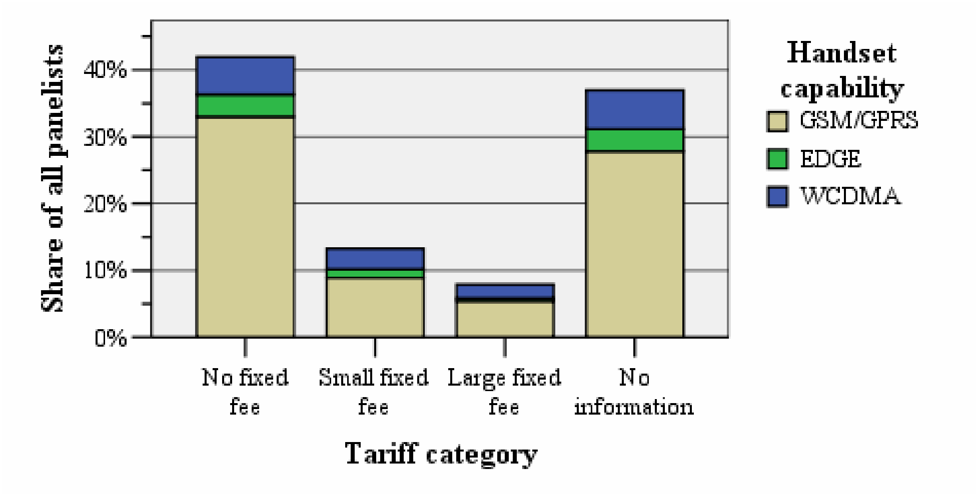


Figure 19 Panel by packet data tariff category

Similarly, the panelists were divided into three categories based on handset usage bill payer. “Pays self” includes panelists who pay the handset bills themselves. “Pays partly self” includes cases where the employer or someone other pays a part of the bill and cases where the bill payer is someone in the family (e.g. spouse, parents). “Employer pays fully” includes cases where the employer pays the entire bill. Again, a fourth category was formed from panelists for whom this information was not available. The distribution of panelists by bill payer and handset capability is presented in the figure below.

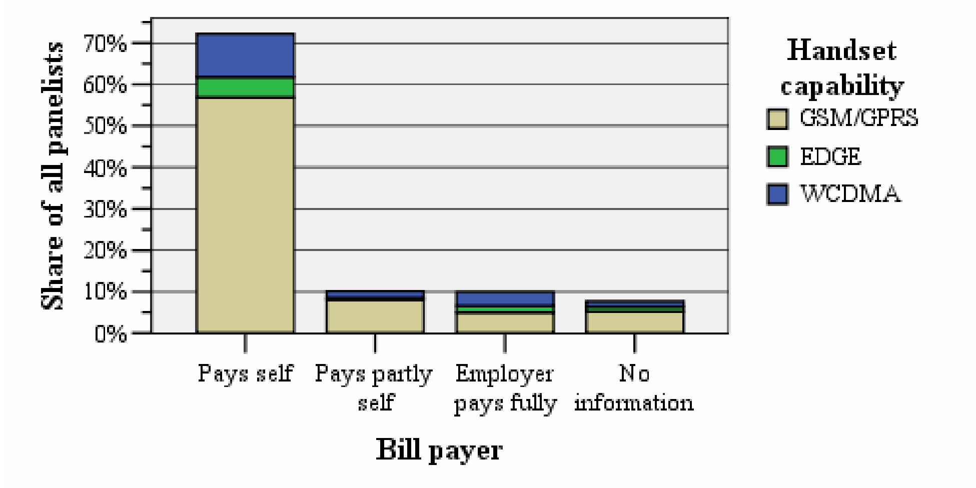


Figure 20 Panel by bill payer

The “No fixed fee” -category accounts for 42% of all panelists, whereas the fixed fee categories together represent about 20% of the panelists. Although 37% of panelists are without background information on tariff category, the share of fixed fee categories is clearly higher in this sample of smartphone users than in the general mobile subscriber population (see chapter 5.1.2). At least 72% of panelists pay handset bills themselves and for a further 10% the bill payer might be somebody else in the family. This corresponds rather well to the share of consumer subscribers in the entire mobile subscriber population (again, see chapter 5.1.2), although consumer subscribers were specifically targeted in panel recruiting. Handset capabilities seem to be distributed to alternative packet data tariff and bill payer categories fairly equally, as illustrated by the colors on the two figures. This decreases the potential effect of tariff category and bill payer on handset capability specific analyses. The share of “No information”-category is significantly higher for data concerning packet data tariff categories, as this information was collected in the voluntary closing questionnaire instead of the compulsory beginning questionnaire. While combining the two background variables, the “Large fixed fee” tariff category was proportionally highest in “Employer pays fully” bill payer category.

5.3.2 General Handset Usage Patterns

Some general handset usage patterns were studied to position or evaluate the significance of packet data usage among other type of usage by describing daily handset usage patterns, general communication service usage, and handset application usage activity.

Daily handset usage patterns

The distribution of daily usage activity of general application usage, outgoing voice calls and messaging (SMS and MMS), and packet data usage is presented in the figure below.

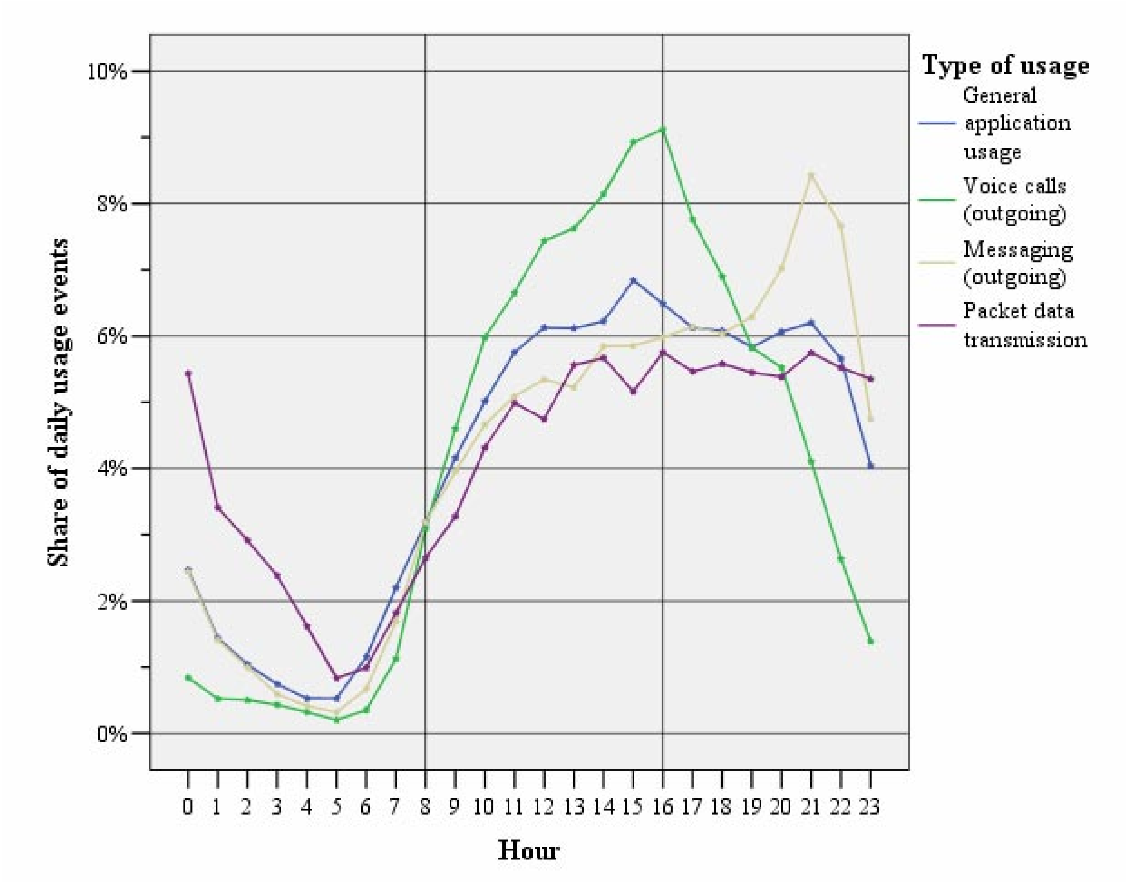


Figure 21 Daily handset usage patterns

Voice calling seems to be concentrated mostly on office hours with a peak at 4 PM, whereas messaging activity increases in the evening with the most active period being around 9PM. Data transmission activity is also present during night-time. General application usage activity forms a sort of an average for the different types of usage, with the influence of calling and messaging apparent on the curve at peak times. In addition, personal information management (PIM) applications form an important part of general application usage.

Communication service usage

The usage frequencies of several communication services were studied. For voice calling, SMS, and MMS only outbound calls/messages were included as they represent usage activity originated by the handset user himself. Bluetooth messaging includes Bluetooth usage such as business cards, calendar updates, image sending, but not e.g. using handset as a modem or headset / hands free usage. For email, both inbound and outbound mails were considered, however, data was available only on platform email application usage and not on 3rd party email applications and webmail. Instant messaging (e.g. Agile Messenger) and push-to-talk usage was derived from packet data traffic generated with identified IM and PoC applications. The usage frequencies of communication services at monthly (28 days), weekly, and daily level are presented in the figure below.

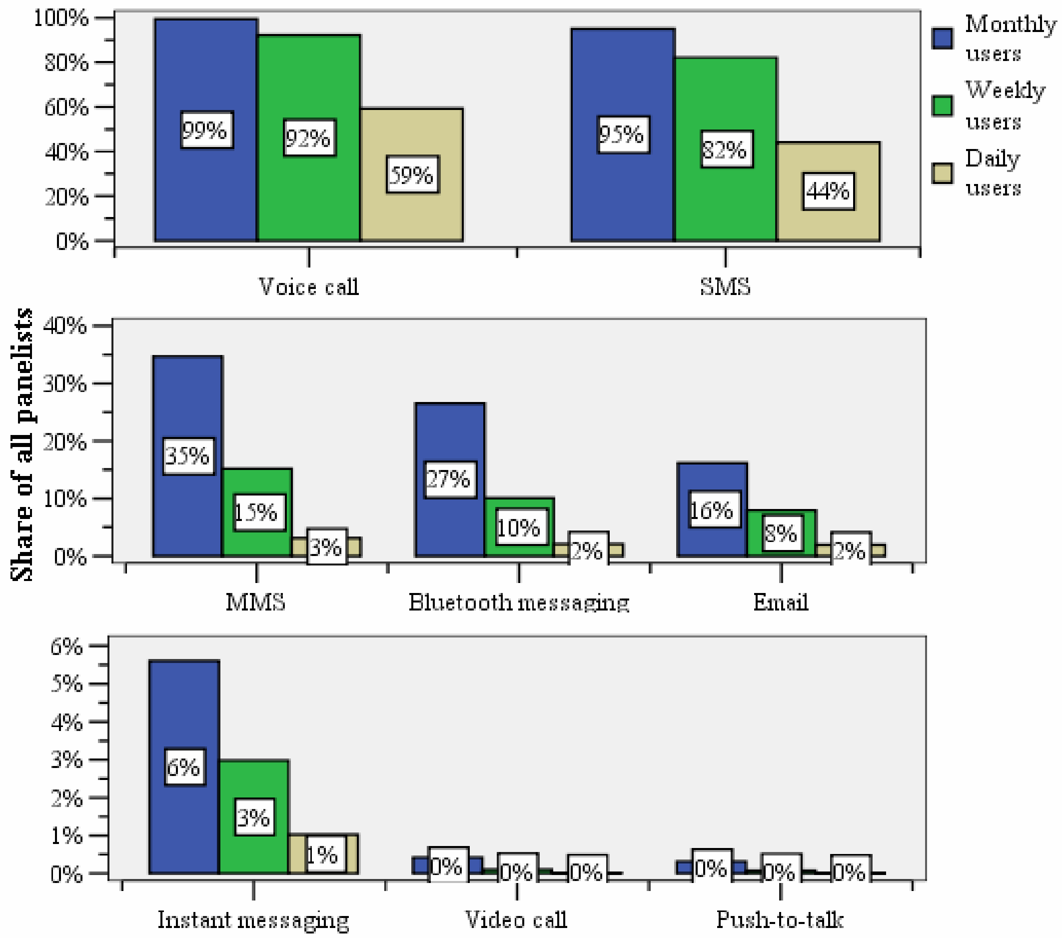


Figure 22 Communication service usage frequencies

Considering the three graphs, it seems that traditional voice calling and SMS messaging still dominate other communication services. Voice calls seem to be used slightly more often than SMSs, whereas other communication services are used seldom. Out of other means of communications, MMS is used the most, although its usage numbers are still meager. The third graph also shows that instant messaging is used monthly by 6% of the panelists. This figure is remarkably high, considering the practically non-existent marketing compared to that made by the operators for e.g. MMS. On the other hand, one should remember that this only means that about 30 particularly active people happening to participate in the panel were using instant messaging during the panel. Video call and push-to-talk usage was marginal, as their availability was limited during panel. More detailed analysis on usage data reveals some additional aspects on communication service usage. Emails are received on handsets generally more often than sent from them, implying that handset usability regarding writing emails is still not sufficient. Another aspect out of the scope of this study would be to evaluate the temporary effect of marketing such as MMS campaigns on communication service usage.

Handset application usage activity

Handset application usage activity was analyzed to study the significance of packet data generating application usage compared to other applications. Some “always on” applications (e.g. Phone, Menu) were removed from the data, as were too short (<3 seconds) and too long (>3 hours) usage instances. A figure presenting the distribution of handset application activations and applications’ active time for three application categories (Browsing, Messaging, Other applications) is included in appendix D.1.

The figure shows that applications able to packet data generation account for almost a half of all handset application activity, as they make up 37% of all application activations and 53% of applications’ combined active time. These applications are mostly browsing and messaging related applications. One should notice, however, that all activity on packet data generation capable applications is not necessarily related to packet data traffic. Rest of usage is mainly due to personal information management (e.g. phonebook, calendar, clock…), multimedia (e.g. camera application, media players…), and utility (e.g. GPS applications, file browsers) applications.

5.3.3 General Packet Data Usage Patterns

General packet data usage patterns were studied by describing the accumulation of packet data traffic, general packet data traffic volumes, significance of modem usage, usage of bearers on packet data traffic, and packet data traffic by tariff category and bill payer.

Accumulation of data traffic by panelist

The accumulation of packet data traffic by panelist activity is presented in the figure below. The upper graph includes all packet data traffic, whereas modem traffic has been excluded from the lower graph. Colors indicate panelists using handsets of different capability.

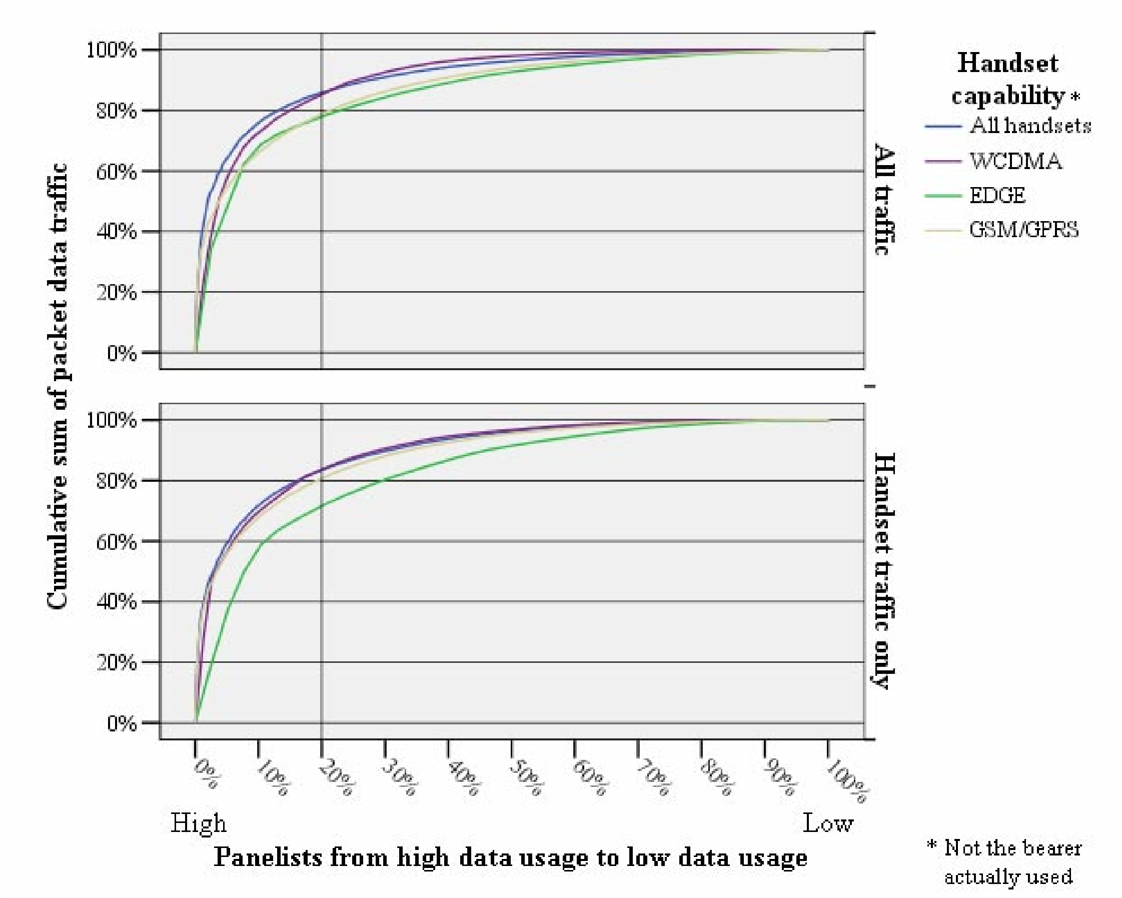


Figure 23 Accumulation of packet data traffic by panelist

The figure shows that 20% of the most active panelists create 80% of packet data traffic. These active 20% most likely include panelists with WCDMA capable handsets, those using the monitored handset as a modem, those with flat or large fixed fee packet data tariff plans, those whose employer pays the handset bills, and probably also other types of heavy using panelists. However, the 20/80-rule seems to apply also when modem usage is excluded from the data, as is shown by the lower graph. The relation is, though, somewhat weaker for those using EDGE capable handsets, which might partly result from the smaller sample of EDGE users.

This result illustrates the challenges encountered when counting average usage figures for the panelists. Using simple averages (arithmetic mean) gives too much weight on the heavy using 20% of panelists while the median describing "average panelist" instead of "average traffic" would give more truthful results in some cases. Despite this, the average (mean) is used in the analyses, unless otherwise stated.

Packet data traffic volumes

The average monthly packet data traffic volume per panelist is presented in the figure below. The upper graph illustrates "average traffic", i.e. mean of monthly traffic volume per panelist, whereas the lower graph gives similar data for "average panelist" counted using median instead of mean.

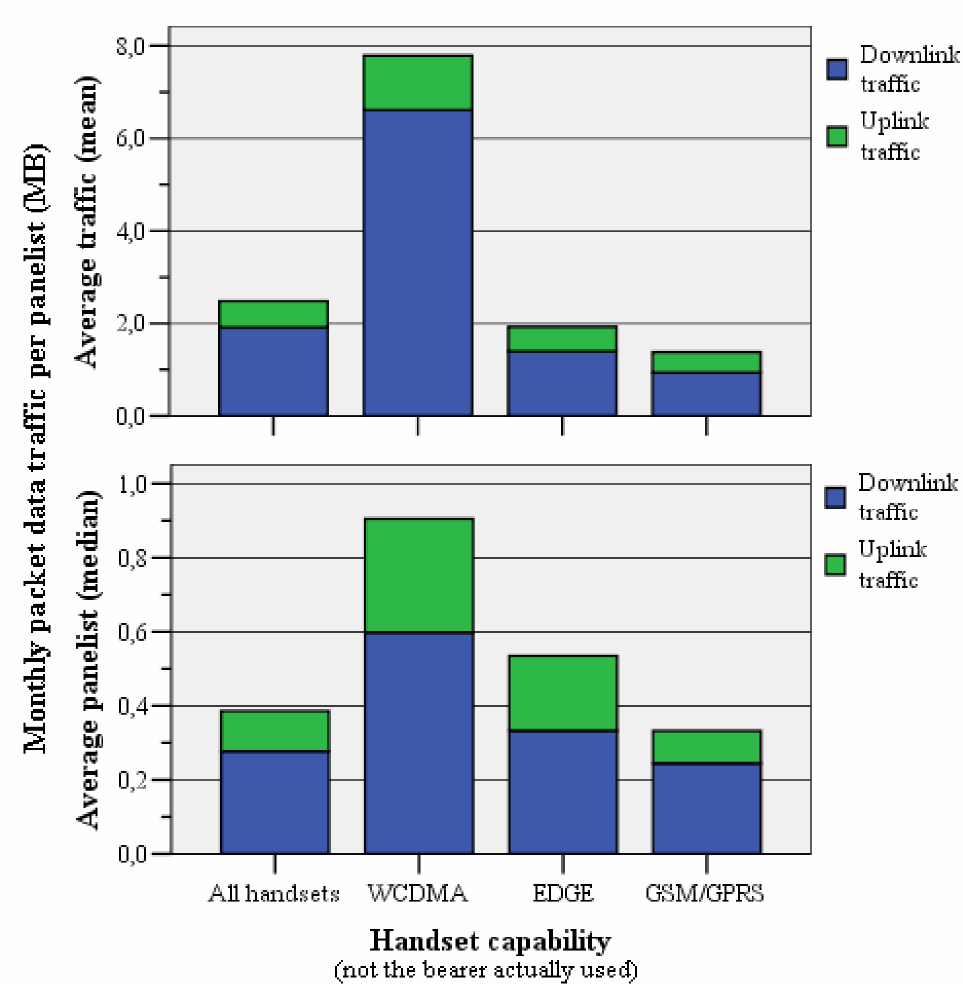


Figure 24 Average monthly packet data traffic volumes

The figure shows clearly that regardless of which average figure is used, average traffic volumes are higher among users of more capable handsets, especially with WCDMA. Does high usage lead people to acquire capable handsets, or does using a capable handset increase data usage volumes? Determining the causality in this situation is hard. Another remark when comparing the two graphs is that average traffic volumes per panelist (mean) are much larger than traffic volumes of "average panelist" (median). This is a straight result of overweighting of heavy users mentioned above.

Significance of modem usage

Identifying the packet data traffic generated using the handset as a modem is ambiguous with the available data. Thus, three traffic categories were defined: "Handset originated traffic" includes traffic positively generated with the handset's Symbian applications, "Non-handset originated traffic" includes traffic positively generated with applications external to the handset, and "Handset or modem" traffic includes the ambiguous cases where a distinction between handset and modem usage could not be made. The shares of the three traffic categories for an average panelist, again counted using both means (upper graph) and medians (lower graph), are presented in the figure below.

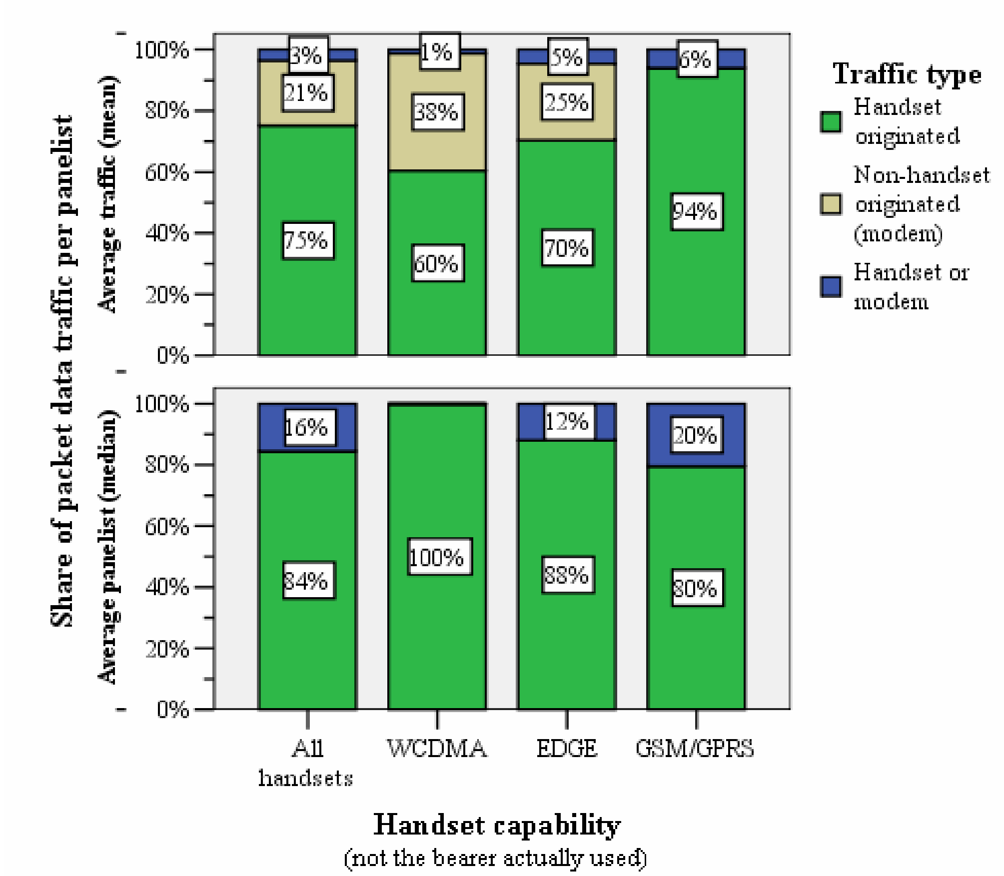


Figure 25 Significance of modem usage

As illustrated by the figure, about 21 to 25% of a panelist's packet data traffic is modem traffic on average (mean). However, an average panelist (median) has less or no modem traffic at all. Thus, it seems that a few modem users create a large share of all traffic, especially while using WCDMA capable handsets. While the share of modem traffic is quite high, it is significantly smaller than the share of Windows traffic observed in the associated network level packet data traffic measurement (see chapter 5.2.1). This is mainly explained by heavy data card usage in mobile networks, which is completely outside direct handset measurements. Moreover, another explanation is the relatively high share of consumers in the handset panel, as modem usage can be considered as a dominantly business driven usage type.

Due to the ambiguity in identifying of modem traffic, the number of modem users could not be accurately determined from the usage data. Based on the above categorization, about 3 – 92% of all panelists used their handset as a modem at least once during the panel. Similar figures are 16 – 69% for WCDMA capable handset users, 5 – 87% for EDGE capable handset users, and 0 – 100% for GSM/GPRS capable handset users.

Usage of bearers on packet data traffic by handset capability

The usage of different bearers on packet data traffic by panelist handset capability is presented in the table below.

Table 14 Usage of bearers on packet data traffic by handset capability

Bearer used	WCDMA capable handsets	EDGE capable handsets	GSM/GPRS capable handsets
WCDMA	45%	–	–
EDGE	10%	30%	–
GSM/GPRS	45%	70%	100%

The figures on the table show that the most capable bearers are not fully used by the handsets for packet data transmission. The primary reason for this is probably in network coverage, as WCDMA (and EDGE) networks cover neither the entire geographic area nor all mobile subscribers in Finland. Moreover, WCDMA network access was not fully provided by one of the participating operators during the panel. Some lesser reasons for the low usage of capable bearers can also be envisioned. The subscription might affect this as it is possible that older SIM cards do not support WCDMA usage. Moreover, the user can also have some effect on this, as the preferably used network can be selected from the handset by the user. Operators might also have influence over this as they might prioritize some subscription types regarding 2G/3G network usage, or resort to load balancing or prioritizing of certain services (e.g. voice calls).

Furthermore, the usage share of EDGE bearer on EDGE capable handsets is surprisingly small. As EDGE network coverage should generally be better than WCDMA coverage, this share could be expected to be at least 55% equaling the combined shares of WCDMA and EDGE bearers on WCDMA capable handsets. Could the built EDGE capacity be so small that there is not enough room for all EDGE users under EDGE coverage, or could there be handset type specific problems in EDGE usage? In any case, this can also be explained simply by the small sample size of EDGE capable handsets, which was actually even smaller in this specific analysis.

Packet data traffic by tariff category and bill payer

Average (mean) monthly packet data traffic volume per panelist for different packet data tariff (upper graph) and bill payer categories (lower graph) are presented in the figure below. Handset capability is indicated by the colors.

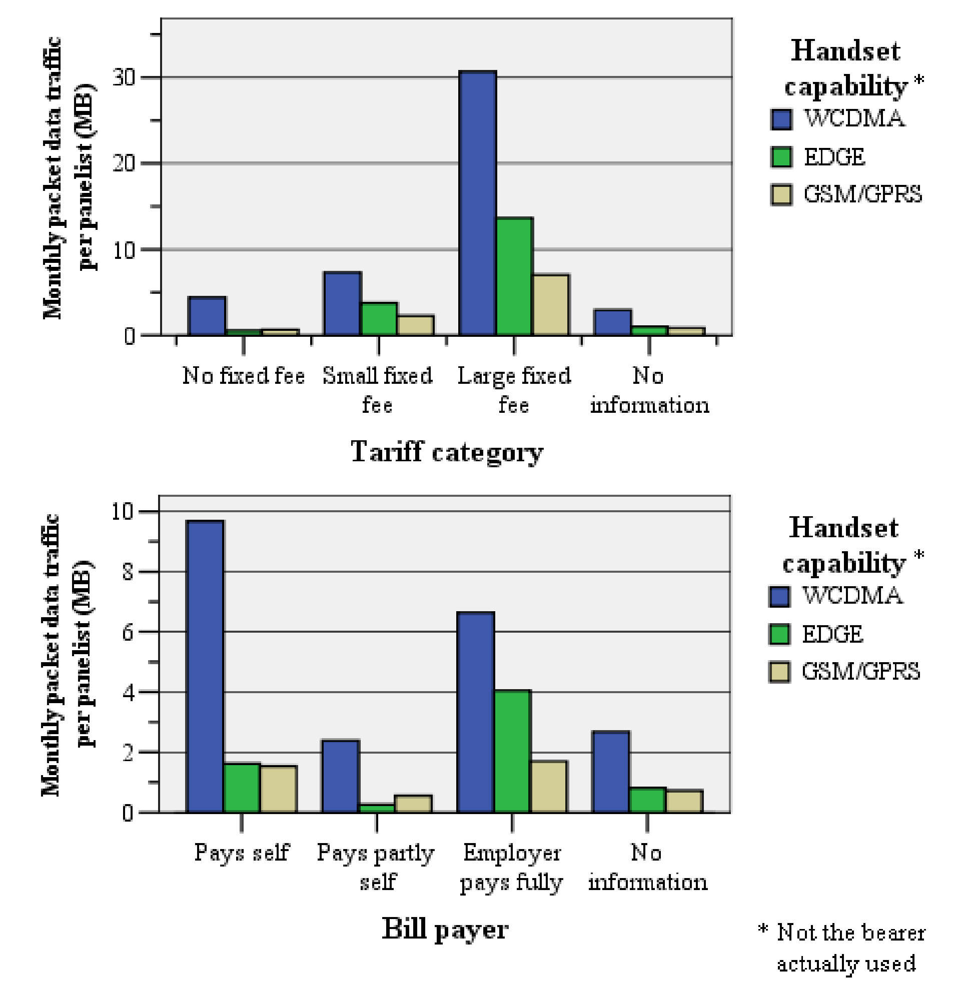


Figure 26 Packet data traffic by tariff category and bill payer

Handset capability seems to monotonically drive packet data usage in all but one of the categories. This is especially clear for WCDMA capable handset users. Moreover, similar results were obtained when using medians instead of means for counting the averages. Regarding the upper graph, panelists with higher fixed fees have more data usage than those with lower or no fixed fee. Does high volume data usage lead people to choose relatively cheaper larger fixed fees, or does the effectively flat fee pricing on packet data increases usage volume further? This is again a situation where the causality cannot be determined at this level of analysis. The lower graph, in turn, does not give a clear indication on the effect of bill payer on packet data usage volume. Packet data usage volume might increase when the employer pays the bills, but what is the case when the bill is a part of the salary? Another speculation could be that heavy users use packet data in high volumes regardless of who pays the bill.

5.3.4 Packet Data Generating Application Usage

The observed packet data generating applications were divided into functional categories based on application name, or its absence. The category *Browsing* mainly included Nokia platform browser, Opera, and NetFront. *Messaging* –category included platform messaging applications, in addition to all identified instant messaging (e.g. Agile Messenger, WirelessIRC, IM+…) and email (Profimail…) applications. *Other applications* included all other applications with an application name not already in Browsing or Messaging (e.g. Anti-virus, PuTTY, Symella, Nowire Teletext…). Finally, *External applications* included all cases where an application name was not registered. This category is also referred to as “modem traffic”, as it combines two traffic categories seen before when treating modem traffic (“Non-handset originated traffic” and “Handset or modem traffic”).

Packet data application usage frequencies and volumes were analyzed based on these categories, after which packet data application usage by panelist activity and panelist demographics was studied.

Packet data application usage frequencies

Packet data application usage frequencies were first studied. A figure presenting the share of monthly, weekly, and daily users for different application categories and handset capability groups is included in appendix D.2.

The figure shows that in general, messaging applications are used most frequently. Moreover, browsing application usage frequency seems to increase with handset capability, whereas messaging frequency is not as clearly driven by handset capability. No major differences between handset capability groups are visible on other applications’ usage, as the category itself is rather heterogeneous. However, other applications are in general used less often than messaging and browsing applications. As some messaging and browsing –related application are always installed in new handsets, a simple assumption could be that usage of the handset browser is a sort of a key to other usage, i.e. other new applications are found via browsing.

Packet data application usage volumes

Packet data application usage volumes were also calculated. A figure presenting the distribution of packet data traffic between application categories and handset capability groups is included in appendix D.3. The figure includes graphs for all packet data traffic (upper graph) and handset-originated traffic only (lower graph).

When modem usage is excluded, the share of browsing usage is highest for WCDMA capable handset users. For users of EDGE and GSM/GPRS capable handsets, the distribution of traffic is fairly similar. In total, browsing generates about 70% of all packet data traffic. This share is relative high, as the share of browsing of packet data generating applications’ active time was only 13%. This implies that not only are browsing sessions active and intensive, but they also entail volume-intensive non-text content.

Packet data application traffic by panelist activity

A distribution of packet data traffic between application categories by panelist activity is presented in the figure below. Again, the figure includes graphs for all packet data traffic (upper graph) and handset-originated traffic only (lower graph).

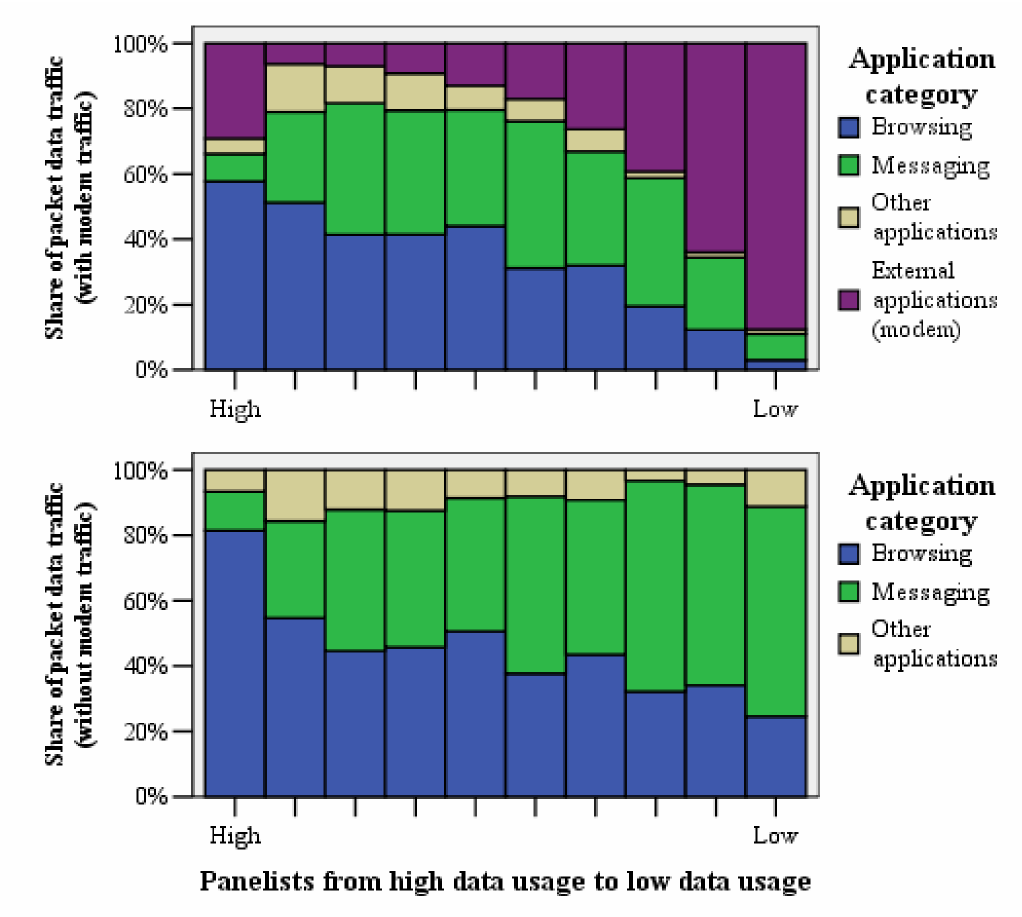


Figure 27 Packet data application traffic by panelist activity

Two major remarks can be made from the figure. First, the share of browsing increases as data usage volume increases. Second, the share of messaging increases as data usage volume decreases. Otherwise the share of the heterogeneous other applications –category remains at a fairly constant level among all user groups. Considering the external applications category present in the upper graph, the high usage share among the high volume users is indeed likely to be modem usage. However, the very high share of modem traffic for low volume users probably results from erroneous usage event logs instead of true modem usage.

Packet data application traffic by gender and age

The packet data application traffic by panelist gender and age was the subject of another analysis. A figure presenting the distribution of packet data traffic between application categories by panelist gender (upper graph) and age (lower graph) is included in appendix D.4.

The figures imply that men are slightly more active data users than women. Regarding browsing women seem to be more active users than men. Furthermore, there seems to be no significant differences in messaging, while other data applications and external applications (modem) are used more by men. Young people are more active packet data users than old people, especially in browsing. The age group of 25-44 year-olds is clearly most active group of data users. Age does not seem to have a major impact on messaging activity. The age distribution illustrates also well the fact that whenever modem usage is present, it forms a major part of all usage volume. As modem usage mainly consists of a small number of sessions by a small number of people, the results cannot be generalized.

However, the problems associated with the relatively small sample size are increased in this type of analysis, where the sample is divided into a number of small categories. Moreover, there were significantly less panelists in the higher age categories, as was seen in the panel description. Thus, a few heavy users can distort the results remarkably. Indeed, one woman created over 70% of women's browsing traffic, which led to women's high browsing averages seen above. In another case one 55-59 year-old created over 85% of all 55-59 year-olds' modem traffic in just one session, which in turn led to high modem usage averages for that age category. Unfortunately, using median figures in this analysis did not remedy problems caused by small sample size and outliers in data.

5.3.5 Smartphone Browsing Patterns

Smartphone browsing patterns were studied by observing the number of web/wap site accesses made by the monitored handsets' Nokia platform browser and Opera. Over 99% of all browsing traffic was made with these browsers.

Most popular browsing destinations by domain

Most popular browsing destinations were first studied. All operator-specific sites were removed from the analysis due to sensitivity reasons. Similarly, all sites visited by less than 10 panelists were removed due to privacy reasons. The most popular non-operator browsing destinations visited by at least 10 panelists are presented in the table below.

Table 15 Most popular smartphone browsing destinations by domain

Rank	Domain name of site	Share of web site visits*
1.	mtv3.fi	9,5%
2.	suomi24.fi	3,6%
3.	iltasanomat.fi	2,2%
4.	google.fi	1,8%
5.	google.com	1,6%
6.	weatherproof.fi	1,5%
7.	nokia.com	1,3%
8.	yle.fi	1,2%
9.	doubleclick.net	1,0%
10.	maf.fi	0,9%
	Others	75,4%

* Web site accesses by handset default browser or Opera, not transferred data volumes

The top 10 list mostly includes some well known service providers. Nevertheless, browsing traffic is not very concentrated to few domains, as the top 10 destinations (including all sites) make up about 44% of all site visits and the top 50 destination about 72% of all site visits.

Most popular browsing destinations by domain category

The browsing destinations were then grouped into 10 aggregated site categories to facilitate comparisons, similarly to the method used in the network level measurements. A description of the basis of categorization is again presented in appendix 0. As the subjective domain name based categorization of web/wap sites had several potential sources of error, such as overlapping categories (information and entertainment; web search and messaging; operator site, mobile content, and entertainment), these results should be considered as indicative only. The aggregated categories with the major included domain names, their combined share of site visits and the share of panelists having visited sites of the category during the panel are presented in the table below.

Table 16 Most popular smartphone browsing destinations by domain category

Rank	Site category	Major sites included*	Share of web site visits	Share of panelists**
1.	Operator site	–	31,8%	68,9%
2.	Information	mtv3.fi, iltasanomat.fi, weatherproof.fi, yle.fi	21,5%	58,3%
3.	Entertainment	suomi24.fi, wamli.net, mbnet.fi, subtv.fi, veikkaus.fi	11,4%	36,5%
4.	Adult content	–	8,7%	12,3%
5.	Web search / portal	google.fi, google.com, motionbridge.com	3,7%	35,4%
6.	Mobile content	maf.fi, buumi.net, inpoc.com, zed.fi, funman.fi	3,2%	28,1%
7.	Advertising	doubleclick.net	1,6%	6,0%
8..	E-commerce	huuto.net	0,8%	6,0%
9.	Banking	nordea.fi, sampo.fi, op.fi	0,8%	11,2%
10.	Messaging	–	0,5%	6,5%
	Other	–	9,6%	53,7%
	Hosting / corporate site	nokia.com, opera.com	6,4%	41,4%

* Sites visited by at least 10 panelists, and with at least 5% of all site visits of the category
** Share of browsing panelists having visited sites of the category during the panel

Operator sites seem to be actively accessed, as they constitute over 30% of all site visits. Moreover, almost 70% of all panelists visited sites of this category during the panel. In a way, this represents the share of operator “portal traffic”, as site visits of this category were over 95% constituted of mobile operator –specific sites. This high share probably results from handset embedded bookmarks and includes lots of mobile content as well. Otherwise, the share of infotainment (information and entertainment) is significant, as is the share of adult content. At the bottom of the table, the browser-based usage of banking, e-commerce and messaging seems marginal. One could interpret this as an indication that using these services with the handset browser is still too awkward.

Individual level browsing patterns

A figure presenting the distribution of an individual’s web traffic to his/her top browsing destinations (i.e. primary, secondary…) with the share of each destination (lower graph) and cumulative sum of the shares (upper graph) is included in appendix D.5.

The upper graph shows that at the individual level, browsing is concentrated into a few domains as the top 5 destinations make up 69% of all site visits. Moreover, when studying the panelists individually (without any averaging), browsing is even more concentrated. On average, a panelist made 41 visits to 9 different domains during the panel.

6 Discussion

The measurement results are discussed in this chapter by comparing the results with each other, and comparing them with previous knowledge on research subjects. Some implications of the results are also discussed.

6.1 Measurement Results Compared with Each Other

The results of the three measurements were introduced separately from each other. This chapter combines the results wherever possible. First, Symbian operating system packet data traffic and handset (smartphone) usage patterns are compared. Second, volumes of packet data traffic measured with different methods are considered. Third, some insights into per subscriber traffic volumes and the resulting revenues are given.

6.1.1 Symbian Traffic and Usage Patterns

Symbian usage was measured in header collection –based and handset –based measurements. Both application and web browsing –related usage patterns provided by the two measurements are compared.

Application protocol traffic versus smartphone application usage

The application protocol profile of Symbian operating system –originated traffic observed in header collection –based measurements (appendix C.1) can be compared to the application usage profile observed in handset –based measurements (appendix D.3).

The first clear observation is the over 30% presence of VPN traffic in the header-based data, which is completely absent in the handset usage data. In fact, no VPN applications were observed in the handset measurements. An evident explanation for this difference lies in the scope of the measurements, as the network level header collection included all identified Symbian S60 and Series 80 terminals and the handset measurements were limited to S60 handsets only. Indeed, VPN software is often used on Symbian Series 80 –based Nokia communicators, which are widely used in Finland.

Distributions of non-VPN application protocol traffic and handset packet data application traffic by application category is presented in the table below.

Table 17 Non-VPN Symbian application traffic

Application category	Symbian operating system traffic (Symbian S60 / Series 80)	Handset packet data application traffic (Symbian S60)
Browsing	36%	70%
Messaging	6%	20%
Other applications	58%	10%

While considering the non-VPN traffic in the header collection measurements, about 36% of traffic volume could be accounted for browsing –related usage (web, wap, and DNS). The share of messaging is roughly 6% (email and multimedia / IM combined), whereas the share of other application protocols is about 58%. These figures differ drastically from handset –based measurement data, where the share of browsing is over 70% of total traffic volume, the share of messaging about 20%, and the remaining share of other applications less than 10%. Some possible explanations for this difference in application usage can be provided. First, as WAP and MMS APNs were not in the scope of header collection measurements, the combined share of browsing and messaging, respectively, should supposedly be about 10% higher. Second, the application usage in S60 and Series 80 –based handsets might indeed differ significantly, and contribute to the observed difference in the profiles. Third, the application (or application name) is not the same as application protocol, as some applications can encompass traffic of multiple protocols. Browser is the first candidate for being such multi-purpose application. Fourth, the method used for application protocol identification and categorization in the header collection measurements was not entirely accurate, which might in turn partly influence the observed difference.

Web protocol traffic versus smartphone browsing

The same domain name –based categorization was used on both header collection –based web protocol traffic data and handset –based browsing data. Thus, a straightforward comparison of the distributions of Symbian operating system –originated site visits (approximated by number of TCP flows) and handset browsing destinations (domain or site accesses) could be made. These distributions are presented in the figure below.

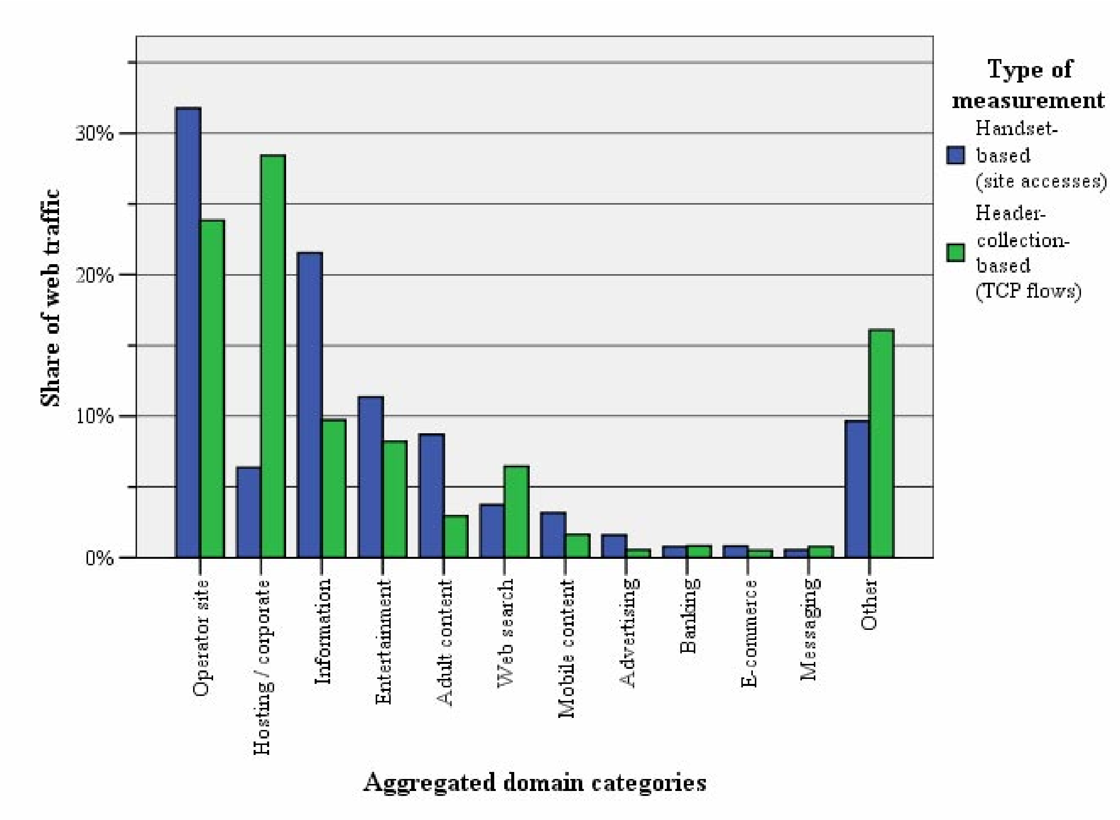


Figure 28 Comparison of Symbian handset browsing patterns

The share of operator sites is the most significant category in both measurements. The figure for handset-based data is almost 10% higher, even though the 16% of private addresses in header data were counted as operator sites. Moreover, the operator sites in header data also included some broadband operator sites (10% for Symbian traffic), while handset –based operator site category mainly contained mobile operator sites (95%). Nevertheless, the difference in these figures seems rather reasonable considering again the fact that the WAP APN with 0-10% of total traffic volume and dominantly consisting of WAP traffic to operator sites was out of the header measurements. The absence of WAP traffic from header data also explains its lower share of mobile content.

The category “Hosting / corporate” also entails a huge difference between the measurements. This is, however, easily explained, as the domain name resolving method used in header measurements resulted in a large amount of hosting provider sites at the expense of the sites of true service providers, as already described in chapter 4.2.3. This could also explain the higher shares of information and entertainment in handset data, i.e. many infotainment sites are likely included in the hosting category in the header data.

Other –category is much higher for header data, as it has a 10% share of completely unknown addresses. This could also partly explain the lower share of adult content given by the header data, i.e. many dubious adult content sites are not so easily identified by reverse DNS resolving. No solid explanations for the higher web search share of header data and higher advertising share of handset data can currently be provided. In addition to the above categories, the shares of browser-based banking, e-commerce, and messaging were equally low in both of the measurements.

In sum, it seems that while the used measures (number of flows vs. browser accesses) were not entirely comparable, they do provide rather similar results. No major conflict between the results was observed. Regarding browsing studies in general, handset-based data is clearly more accurate, but the broader scope of header data evens up the balance.

6.1.2 Measured Packet Data Traffic Volumes

The measurement methods and obtained results can be evaluated by comparing the measured total packet data traffic and Symbian-based smartphone traffic volumes.

Total packet data traffic volume

The operator reporting system –based measurement and the TCP/IP header collection –based measurement both provide means to evaluate the total volume of packet data traffic in mobile operator networks.

Considering the TCP/IP header collection –based traffic measurements, an estimate for the total packet data traffic volume of all operators can be calculated by weighting the traffic volume measured from two operator’s networks by the number of subscribers of all operators. Moreover, the fact that only the Internet APN covering about 90% of all packet data traffic volume was measured must be taken into account. Furthermore, the differing lengths of the measurements (week vs. month) must be balanced in calculations. The total traffic volume from reporting system measurements entails more obscurity, as data volumes were partly acquired from the CDRs and partly from the billing systems, thus partly including a subscriber's all packet data traffic and partly his/her chargeable packet data traffic only. Nevertheless, it was possible to calculate an estimate for the total volume of packet data traffic generated by the postpaid subscribers of all three operators.

Comparing the two figures for total traffic volume, a somewhat troubling six fold difference in favor of the reporting system –based figure was observed. While the primary source of reporting system data was the billing system, it might be that the total volumes in these figures also include the packet data traffic due to separately billed services or content, such as MMSs. While this kind of content was not measured in the header collection (WAP and MMS APNs were not measured), it should have already been taken into account in the calculation. Thus, no explanation for the difference in the measured volumes is provided by this. The measurement periods in the two measurements were somewhat different, but it seems unlikely that the variation in traffic volume between random weeks in September could be so large. Moreover, as the prepaid subscribers were in the scope of the header collection measurement, but outside the volumes from reporting system measurements, the difference between the two measurements should actually be even a bit larger. The only logical explanation remaining is that some traffic was left outside the measurement, for some reason or another. This might have resulted from flawed measurement organization not capturing all the targeted traffic, although this explanation seems unlikely as similar differences were found in the data of each of the operators. As only TCP and UDP traffic was captured, some IPSec VPN traffic might have been left outside the measurement. While the share of VPN traffic was found to be significant, an addition of one order of magnitude in VPN traffic volume seems unlikely. Thus, this difference in measured volumes remains largely unexplained.

Smartphone packet data traffic volume

The total volume of Symbian-based traffic can also be estimated by combining the results of the measurements.

According to the results on mobile terminal installed base, 5,5% of all Finnish mobile terminals are Symbian –based smartphones. The average (mean) monthly packet data traffic volume for the monitored panelists was about 2,5 MB, about 21-25% of which being modem traffic. On the other hand, an average panelist (median) used about 0,4 MB per month with a 0-16% share of modem traffic. These volumes can again be compared to the traffic volumes of the header collection measurement, but this time to the share of identified Symbian traffic.

In this case, the difference in magnitude is a staggering 480 and 90 in favor of the handset-based measurements, when mean and median figures on average usage were used respectively. However, this time some more feasible explanations can be provided. First, the traffic volumes given in handset usage –measurements include all traffic to and from the handset, including possible overhead. While all monitoring software –related traffic was excluded from the handset usage data, at least some signaling traffic is no doubt still included. Second, the handset usage measurement only included Symbian S60 handsets, whereas one third of Finnish Symbian smartphones are actually Symbian Series 80 –based Nokia communicators, as observed with results concerning mobile terminal installed base. It could well be that S60 handset users generate a lot more traffic than Series 80 users. Third, the share of Symbian might also be played down in the header collection measurements, if all versions of Symbian were for one reason or another not identified in the operating system identification. Fourth, the type of people participating in this type of studies, responding to the kind of recruiting methods used, and able to independently (albeit with instructions) install 3rd party software on the handsets are arguably early adopter users, i.e. even more advanced data service users than other smartphone users. Thus, their usage volumes are probably also higher than the volumes of average smartphone users. Fifth, the share of Symbian packet data traffic going via Internet APN might be smaller than the share in general for all traffic. While Internet APN traffic in general is dominated by modem and VPN traffic, the advanced multimedia features in most Symbian handsets might result in higher MMS usage. By using some rough percentual estimates on the effect of these issues, a difference of similar magnitude as found in the previous chapter on total packet data traffic volumes remains.

6.1.3 Per Subscriber Data Traffic Volumes and Revenues

Subscriber-level data traffic volumes and the resulting operator revenues can be evaluated by comparing reporting system and handset-based measurement results, and combining these results with the pricing in effect during the panel.

Consumer subscriber data traffic volumes

Comparing traffic volumes by packet data tariff category, it appears that apart from the "No fixed fee" –category, the aggregate level values given by reporting system –based data are 2 to 4 times larger than those given by handset-based measurements (when possible modem traffic is not excluded from handset-based data). This is troubling, as the analysis above suggests that the total volumes of packet data traffic measured on handsets should be significantly higher than those given by the reporting systems. Moreover, as the figures for handset-based measurements are for all panelists, i.e. also the 10-28% of business subscribers in the panel, these traffic volumes should be even higher. An explanation for this is that the true data card / modem users outside the handset –based measurements have a high tendency to acquire larger fixed fee or flat fee packet data tariff plans, and also use these plans extensively. Indeed, total packet data traffic in mobile networks seemed to be dominated (65%) by Windows originated-traffic.

Revenues per consumer subscriber packet data categories

Some average values for consumer subscriber packet data traffic tariffs can be calculated for the three packet data tariff categories. Computational values for average block sizes, and average fixed and variable tariffs in each category are presented in the table below.

Table 18 Computational values for packet data tariff categories

Packet data tariff group	Average block size	Average fixed tariff	Average variable tariff
No fixed fee	0 MB / month	0,0 € / month	4,9 € / MB
Small fixed fee	10 MB / month	6,0 € / month	1,7 € / MB
Large fixed fee	179 MB / month	16,6 € / month	2,0 € / MB

A rapid calculation combining this data with the data from table 7 of chapter 5.1.2 gives insights into the packet data revenues the operators gain from subscribers with different tariff plans. An average subscriber in “No fixed fee” plan creates traffic worth 0,2 € per month. Subscribers with “Small fixed fee” alternative generate traffic over their average fixed fee block size and thus spend monthly about 18€ on packet data traffic. Subscribers with a “Large fixed fee” in general do not go over their block sizes and escape with just the fixed 17€ monthly tariff. Weighting these figures with the number of subscribers tells that no fixed fee subscriber bring 54%, small fixed fee 30% and large fixed fee 16% of the operators’ overall packet data revenues. However, average per subscriber revenues from small fixed fee subscribers seem particularly high, and might partly result from larger variation among the alternative tariff plans within the category. Thus, these results should only be considered indicative due to the inaccuracies present in analysis.

Revenues from smartphone usage

Average values for smartphone users’ volume of communication service usage can also be calculated. Using simple averages (arithmetic means) over the entire panel, each panelist originated about 188 voice call minutes monthly, sent 61 SMSs and 4 MMSs monthly, in addition to the packet data traffic usage. Average revenue stream of a single subscriber can be calculated by multiplying these usage volumes with exemplary average prices in fall 2005, i.e. 0,069 € / call minute, 0,069 € / SMS, 0,38 € / MMS. Thus, an average subscriber used voice calls, SMSs, and MMSs for about 19 € each month.

Considering the packet data traffic volumes of smartphone users (see chapter 5.3.3), it seems that on average smartphone users with fixed fee tariffs do not use packet data up to the maximum block size, whereas those without a fixed fee plan create roughly 1,1 MB of traffic monthly. Using the tariffs indicated in the above table, the packet data traffic revenues from smartphone users are 5,4 € for no fixed fee users, and 6,0 € and 16,6 € for users fixed fee plans.

Comparing the packet data revenues figures to the monthly communication service revenues leads to rather promising conclusions, at least from the point of view of operators. Smartphone users’ data usage does indeed bring additional revenues comparable to e.g. SMS usage, at least with current pricing. Moreover, potential increase in usage of separately-charged mobile content could further increase mobile data revenues, even if the prices of traditional communication services keep on falling.

6.2 Measurement Results Compared to Previous Knowledge

The results of the three measurements are compared to previous knowledge in this chapter, whenever such data on specific research items was available. First, the characteristics Finnish mobile terminal base and subscriber population are compared to other data describing the Finnish market, and data on international level. Second, differences in traffic patterns in mobile and fixed networks are evaluated. Third, the observed smartphone usage patterns are compared to previous research results on smartphone usage.

6.2.1 Mobile Terminals and Subscribers

Some mainly survey-based information on both Finnish and European mobile terminal bases and on the Finnish mobile subscriber population was available for comparisons.

Mobile terminals in Finland and abroad

The report prepared by eBird Scandinavia (Snellman 2005) gives some figures for the penetration of certain mobile terminal features. According to the study, the share of phones with color display was 37%, camera phones 12%, smartphones 4%, and the share of “3G subscriptions” less than 1% at end of 2004. These figures are rather well in line with the measurement results, which gave the percentage shares of 46%, 17%, 6%, and less than 1% respectively in fall 2005. Elisa (2006) presents some figures for mobile terminal feature penetration in Finland. Accordingly, the share of packet data capability is 52% and 3G 0,3% of all mobile terminals. These figures are very close to the obtained measurement results, which should be no surprise as Elisa’s data was one of the used data sources in this study.

The results on mobile terminal installed base included information on manufacturer market shares in the Finnish mobile terminal base. According to manufacturer representatives’ estimates at the end of 2004 (Digitoday 2005a), the share of Siemens was about 10% and SonyEricsson about 6-7%, whereas Samsung targeted a 20% share by 2006 and was allegedly close to the target during some months already at that time. These estimates differ drastically from the actually observed terminal base, which indicates the combined share of all non-Nokia manufacturers to be at maximum about 13%. While estimating the difference between the two sources, one should remember that Digitoday’s figures were given by manufacturers’ representatives, who are not the most objective source of information imaginable. Moreover, they represented the market shares on sales of handsets and not on the shares of actually used mobile terminals (including some 0,6% of non-handset terminals). Ketamo et al. (2005) have also studied the shares of manufacturers on the handsets actually in use, focusing especially on teenagers. Their results indicate Nokia’s share to be between 79% and 92%, depending on the age category, and also reveal the second place of Siemens among the smaller manufacturers. Theses figures are already much closer to results obtained in this study.

Ketamo et al. (2005) also presented data on most popular handset models, with the most popular models Nokia 3510 and Nokia 3310 having a combined share of over 30% among the youngest people and a gradually decreasing share of total in more mature age categories. The same models are also at the top of the measured mobile terminal base. Handset age is also briefly treated by Ketamo et al., as their study discusses the recycling of handsets in the family. Essentially, it is suggested that a handset’s user changes every 1-2 years, but the handset itself remains in use for a much longer period of time. This would implicate average holding times of over 3 years for handsets, which is consistent with measured results.

The penetration of certain mobile terminal features on the western European level is evaluated in a report prepared by Forrester Research Inc. (Van Veen 2005). The report estimates that 3G phones represent 1% of total terminal base in Finland, and about 6% on average on 17 western European countries. Moreover, the share of GPRS capable phones is estimated to be about 67% of all terminals in these countries at the end of 2005. According to Elisa's estimate (Elisa 2006), the share of GPRS capable phones in "Europe" is 72%. Both of these estimates, while somewhat vague and differing from each other, suggest that the Finnish terminal base is not at the average western European level concerning the penetration of advanced terminal features.

All in all, a significant growth in the usage of new services necessitates a relatively high penetration for the supporting terminal features. In the case of SMS in Finland, the required penetration was 15% to 30% before the network externalities began to have an effect. However, the required penetration depends on the service in question. While person-to-person communication services require that the service is supported by the terminals on both sides of communication, most of the mobile content services can be used as long as the user's handset supports that type of usage. (Snellman 2005) Thus, high volume usage of advanced data services at the individual level is not dependent on the general penetration of certain terminal features, even though penetration is definitely a factor at the aggregate level.

Mobile subscribers in Finland

Previous knowledge on mobile subscribers similar to the items studied in this research is mainly related to subscription types and service usage frequencies.

The high share of postpaid subscriber in Finland verified by this research is a rather commonly known fact (e.g. Snellman 2005). However, regarding the distribution of postpaid subscriber to consumer and business subscribers no previous knowledge was found. Having a handset provisioned by the employer is not quite the same thing as having a business subscription, but might give some insight into the share of business subscriptions. According to a survey research (Otantatutkimus 2005) made by Otantatutkimus Oy for Ficora in November – December 2005, about 20% of people have employer-provisioned handsets, although half of those people also own a private handset.

The same study also provides data on data service usage. Accordingly, 91% of the studied people use SMS, 23% MMS, and 14% Internet with their handsets. These figures are rather consistent with the measurement results. Obtained communication service usage figures of smartphone users for SMS and MMS were 95% and 35%, respectively (see chapter 5.3.2). The slightly higher figures are explained by the "early adopter" bias of smartphone users, especially regarding MMS usage. The reporting system –based figures indicated that about 8-9% of consumer subscribers used packet data during the study week (see chapter 5.1.2), which means that the overall share of users is somewhat higher. This is also close to the figure given by the survey research, although packet data usage is not quite the same as using Internet with the handset.

6.2.2 Traffic Patterns in Mobile and Fixed Networks

Similarities and differences in mobile and fixed network traffic patterns are evaluated by comparing the measured application protocol traffic and browsing patterns to previous research results.

Application protocol traffic in mobile and fixed networks

The application protocol profile in fixed networks depends a lot on the point of measurement, i.e. whether the traffic is dominantly created by consumer or business users. In general, it can be said that business traffic has a large share of VPN traffic, whereas the share of P2P and/or web traffic should be large in the case of consumer traffic. Core network measurements (e.g. Viipuri 2004) are yet another case, where the traffic of different types of users is supplemented by background traffic independent of user actions.

In the header collection –based measurement made in mobile networks, the measured traffic had a high share of VPN, which also explained the unusual domination of UDP over TCP traffic. Apart from VPN, web and DNS were also major applications. The over 30% share of “other” application protocols, meaning all unidentified TCP and UDP protocol ports, however, makes accurate comparisons challenging. The share of observed P2P traffic is particularly low, although the unidentified traffic probably includes P2P traffic as well. As there are not many mobile P2P software (e.g. Symella) and the storage capacity limitations in current mobile devices imply that there is not yet much to share, P2P usage necessitates the usage of a PC. Moreover, the mobile subscriber most likely needs a fixed fee plan for packet data for high volume file sharing to be reasonable. Such plans were used by some subscribers during the measurements. Thus, the real share of P2P traffic in the mobile network might actually be higher than observed.

Mobile and fixed browsing patterns

A comparison of mobile and fixed network browsing could reveal some additional mobile-specific usage patterns. Considering the measurement results, the observed Windows traffic patterns should be quite similar with its fixed network counterparts, whereas the Symbian profiles compared in chapter 6.1.1 reflecting truly mobile usage could indeed have some differences to fixed network or PC browsing.

TNS Gallup provides weekly data on the most visited Finnish web sites as a part of its TNS Metrix study. Data is collected directly from end-users using browser cookies. The sites of service providers having a contract with TNS and allowing the data to be published on TNS web site are included in data. The most popular Finnish PC browsing destinations are presented in the table below.

Table 19 Most popular PC browsing destinations

Rank	Service provider	Share of sessions*
1.	MTV3	15%
2.	MSN.fi	12%
3.	IRC-Galleria front page	12%
4.	Suomi24.fi	9%
5.	Ilta-Sanomat	7%
6.	YLE	6%
7.	Iltalehti	6%
8.	Telkku.com	4%
9.	Helsingin Sanomat	4%
10.	Huuto.net	2%
11.	MBNet	2%
12.	Subtv.fi	2%
13.	Nettiauto.com	2%
14.	Eniro.fi	2%
15.	Kauppalehti Online	1%
	Others	14%

* Share of all browsing sessions to sites of the reported 97 service providers

Comparing the results on mobile browsing to the above TNS data, some similarities and differences can be observed. However, the percentage shares of most visited sites are not comparable, as all web sites (by far) are not included in TNS data.

Considering the domain categories used to aggregate measurement results, the share of infotainment category sites is much larger with PCs. This results at least partly from the measurement method that focuses on established service providers only. Moreover, these results (on week 38) do not entail many operator portals and hosting service providers or foreign domains due to the measurement method, nor advertisement and adult content domains due to the lacking incentives of providers of such services to participate in the study. Thus, a more fruitful approach to comparisons would be to use the domain name level data.

A comparison to mobile network PC browsing traffic can be done using the reported general header collection –based data, which essentially corresponds to Windows-originated traffic in mobile networks. Some of the most important sites, MTV3 and IRC-Galleria are present in both the measured and TNS's results, as are the sites of Helsingin Sanomat, Telkku.com, and Huuto.net. Most of the other sites in the above PC browsing top 15 are also at the top of mobile network PC browsing, when considering similar infotainment oriented Finnish service providers only. The most significant difference between the results is the absence of MSN.fi among the most visited sites. One is tempted to explain this with the business-oriented nature of PC usage on mobile networks, but the presence of teenage-focused IRC-Galleria vitiates this argument.

A comparison of fixed PC browsing to smartphone browsing reveals somewhat surprisingly a better match between the browsing patterns. MTV3, Suomi24, Iltasanomat, and YLE are major sites on both lists, and MBNet, SubTV, and Huuto.net also rank high in both results. However, an explanation for this is straightforward. The major Finnish service providers primarily in the scope of TNS's fixed browsing measurement provide a variety services, including e.g. information and content specifically for mobile users. These services are previously known to smartphone users and are often promoted by the operators. Thus, smartphone users use the same services on PCs and mobiles. One major difference between the results is the absence of MAF.fi from PC browsing results. This is not surprise, as MAF focuses specifically on providing mobile content.

6.2.3 Smartphone Usage Patterns

Smartphone usage patterns are evaluated by comparing the handset-based measurement results to data from previous research.

Handset usage in previous studies

The same research platform has been used before to measure handset usage on Symbian S60 handsets (see Verkasalo 2005). The main differences between the studies are in panel size, measurement period (year, length), measured markets, and the version of the used measurement software, as the large Finnish panel had a more constrained 3-month duration than the 8-month lasting somewhat smaller panels in the U.K. and Germany that also used an earlier version of the research platform. Despite these differences, similar results were obtained regarding multiple research items.

The distributions of panelists on several background variables were surprisingly similar, as both panels consisted dominantly of young men and those who pay handset bills themselves. This cannot be explained by only the used recruiting method, as they were similar only regarding the required online registration. A more solid explanation is in the demographic distribution of smartphone users, and the propensity of certain types of people to participate in this type of research in general.

Regarding usage of advanced services, such as application and packet data usage, men and young people proved to be more active users in both researches. Communication – related services were also used quite similarly, as voice calls were mostly placed during day time and messaging activity increased in the evening according to both results.

Smartphone browsing patterns

A research by Kamvar and Baluja (New Scientist 2006) on Google web searches provides some insight into the importance of adult content in the mobile context. By analyzing one million searches made using Google's mobile search software, it was found that adult material constituted 20% of searches on handsets ("cell phones") and only 5% on PDAs, whereas the share of adult content on fixed PCs is around 8,5%. It was argued that this results from the fact that people regard their handsets as personal devices and may feel more comfortable searching adult material.

Somewhat similar results were obtained in the handset-based measurements, where the share of adult content, while being only 9% of all site visits, was 2-3 times higher than in the mobile network Windows-based traffic.

6.3 Implications of Measurement Results

The obtained measurement results imply that Finnish mobile data usage is currently business driven. Business subscribers create five times as much packet data traffic than consumer subscribers on average. Moreover, business-oriented usage, such as modem and VPN traffic, represents high shares of total packet data traffic volume. Regarding the mobile terminals, one third of Finnish smartphone handsets are Nokia communicators specifically intended for business users. On the other hand, consumer mobile data usage is not very intense. Mobile data usage is currently not widely supported by the mobile terminals, and consumer subscribers do not seem to be too keen to pay additional fees on data usage either.

The share of mobile Internet traffic appears to be rather significant according to the measurements. The share of Internet APN traffic is about 90% of all mobile network packet data traffic and the share of browsing to sites not controlled by mobile operators is about 70% of all smartphone web accesses. Packet data usage seems to be very concentrated at the moment, as 20% of the monitored smartphone users and a mere 1% of all consumer mobile subscribers create over 80% of panelists' and consumer subscribers' total packet data traffic, respectively. This could imply that mobile Internet is currently used primarily by technologically apt early adopter users who prefer using free Internet over operator-provided services, and also know how to do it. Thus, the share of traffic to operator portals could actually increase in the future, as the masses of consumers preferring easy access to selected operator-controlled services start using mobile data.

Using terminals with higher data transmission capability seems to imply higher packet data usage volumes, especially in the case of 3G terminals. Similarly, having relatively cheaper partly of fully fixed fee packet data plans implies higher data usage. However, it is not evident to prove the causality between active packet data usage, and high capacity terminal usage or flat-rate packet data tariffs (hypotheses 1 and 2), i.e. to prove that data usage is really driven by either of the two factors. The simple statistics provided in this study merely show that data usage volumes are higher for such subscribers. The handset bundling allowing subsidies on 3G handsets starting in April 2006 is likely to have a renewing effect on the Finnish mobile terminal base. According to the above, advanced handsets enabling new data intensive services and fast data transmission should also lead to increased data usage activity by consumer subscribers, which should be observed on consumers' data tariff plans as well.

Browsing is a central application area in mobile Internet, as it is in fixed Internet. The handset browser is probably not only used for searching information or entertainment content, but it also functions as the primary tool for finding and downloading other 3rd party applications. Thus, the ability to browse the mobile Internet is a prerequisite for other types of usage, apart from using the preinstalled platform messaging applications.

7 Conclusions

Conclusions are presented in this chapter. A summary of the major findings made in this research are presented, after which the reliability and validity of the research results are discussed. Finally, some tangible recommendations to industry actors are provided before giving suggestion for further research.

7.1 Summary of Findings

The operator reporting system –based measurements included about 80-90% of all Finnish mobile subscribers and their terminals in fall 2005. The observed Finnish mobile terminal installed base is relatively old compared to European reference countries. Key terminal features for packet data usage are not widely spread, as packet data capability itself is only in 48% of all terminals. The share of EDGE capable terminals is 13%, and WCDMA capable terminals less than 1%. Nokia's market share of mobile terminal installed base is a remarkable 87%. About 6% of all terminals are smartphones, over 99% of which are Nokia manufactured Symbian –based handsets. Moreover, about one third of all smartphones are actually Nokia communicators. At the level of individual models, the terminal base is fairly concentrated, as the 50 most common models make up 88% of all terminals. Regarding the mobile subscriber population, 92-94% of all subscribers are postpaid subscribers. While consumers represent about 75% of postpaid subscribers, business subscribers create 62% of packet data traffic volume. About 99% of consumer subscribers have the operators' default usage-based packet data tariff plan. The remaining 1% with some kind of fixed fee plan creates 82% of all consumer subscriber packet data traffic.

The TCP/IP header collection –based measurements captured about 50% of all Finnish mobile network packet data traffic during one week in fall 2005. The most striking finding made in these measurements was that Windows operating system seems to originate about 65% of all packet data traffic in mobile networks due to the usage of data cards and GPRS modems. This also means that the truly mobile usage profile is effectively hidden by the Windows traffic. It was also discovered that VPN usage creates 46% of packet data traffic, which leads to the very high 85% share of UDP traffic compared to the situation in fixed networks. According to operators' estimates, about 90% of all packet data traffic, meaning the packet data traffic of all APNs, goes via the Internet APN. Moreover, web is also a major application with 14% of data traffic volume, and 25% share of non-VPN data traffic volume.

In the handset-based measurements some 500 Finnish Symbian S60 handsets were monitored for a period of 3 months in fall 2005. The measured usage data showed that people with higher radio capability (GSM/GPRS → EDGE → WCDMA) handsets used packet data more frequently and in higher volumes. The effect of pricing was also observed, as data usage volumes were higher for users with relatively cheaper larger fixed fee packet data plans. Regarding browsing, operator sites and infotainment dominated web/wap site visits with 32% and 33% shares of all visits, respectively. Individual level browsing was concentrated into just a few sites. Using handset as a modem forms a significant 21-25% part of all smartphone-based packet data traffic volume. Furthermore, it was observed that the most active 20% of packet data users create 80% of traffic, which applies even when modem traffic is excluded from the data. Considering application usage, browsing was the most important data application area with a 72% share of non-modem traffic. The relative share of browsing increased with usage volume.

Measuring mobile Internet usage with three fundamentally different methods proved worthy the effort, as the differently measured data items complemented each other in a satisfactory manner. No seriously contradictory results were observed, apart from the differences in measured traffic volumes which necessitate additional analysis in the future. The obtained results were also rather consistent with previous knowledge, although the mainly survey-based previous research provides somewhat less accurate data and is also much more restricted in scope, i.e. many of the measured data items cannot be studied using survey-based methods.

Considering the above measurement results, it seems that Finnish mobile data usage is currently business driven. Mobile data usage is not widely supported by Finnish mobile terminals at the moment and consumer subscribers do not seem to be too keen to pay extra on data usage. Mobile Internet in the form of off portal traffic appears to be rather important and the share of operator-controlled is less than expected. However, this might change as the masses of consumer subscribers start using mobile data actively. Browsing is a central application area also in mobile Internet. The usage of 3G terminals and effectively flat-rate packet data tariffs seems to increase packet data usage considerably, although the causality between these two factors and active packet data usage is hard to prove.

All in all, the research questions with the associated research sub-questions were answered by reaching the related research objectives. Moreover, the supporting research hypotheses were all confirmed.

7.2 Reliability and Validity

The reliability and validity of research methods and results are discussed in this chapter.

Reliability

There were some potential sources of error in all measurements affecting the reliability of the measured data. First, several research phases might have included random error. The operator reporting system –based data collection necessitated the setup of data warehouse runs and subsequent aggregation of the collected data by the operators' personnel. In the handset –based measurements, the participants filled an online questionnaire detailing their background demographics during which some false data might have been given either intentionally or accidentally. Indeed, some false background data on panelist's operator variable was observed. Finally, the analysis phase in all three measurements included many potential places for error from the researcher. Second, some systematic error might also have been present throughout the measurements. In operator reporting system –based measurements the operators might have used somewhat differing definitions for what constitutes the requested research item, e.g. how the share of prepaid subscribers is defined. The measurement setup might also entail systematic error if the measurements were implemented differently by the operators, or for instance in the case of bad or double entries made by the handset monitoring software. Such error was also present in various analyses regarding the header collection measurements. The methods used for subscriber terminal, operating system, application protocol, and web traffic identification, and the categorization of application protocols and web domains probably all included a degree of systematic error. As some measurements included more identified error sources than others, certain results can be stated to be more reliable than others. More specifically, those research items with error in measured data or some early analysis phase include the error in all subsequent analyses. Thus, the results derived using data from multiple analyses, such as operating system identification and application protocol categorization in the header collection measurements are clearly less reliable.

Much of the concerns on measurement reliability were eased with the use of comprehensive descriptions on measurement implementations and evaluations of the potential sources of error throughout these descriptions. All in all, the operator reporting system –based measurements were clearly the least transparent part of the measurements, and thus are the place where measurement reliability could most likely be questioned.

Validity (internal)

This research provided primarily observation-based descriptive results without too many notions on causal relationships. Thus, the question of internal validity is not very relevant. Causality was somewhat of an issue when considering the effect of terminal device's higher data transmission capability and subscriber's fixed fee packet data tariffs on packet data usage volumes. As already stated, causality might in fact be opposite, i.e. the tendency for high volume usage might lead people to acquire more capable terminals and relatively cheaper fixed fee plans. As the conclusions based on measurement results were made rather carefully and were founded only on the most reliable data items, the validity of the conclusions need not be questioned.

Representativeness and generalizations (external validity)

Considering the representativeness of the measurement samples, some remarks are in order. The operator reporting system –based measurement included about 80-90% of all Finnish mobile subscribers and terminals during one week/month free of major seasonal variations. This is largely sufficient for any generalizations on the Finnish market, even though the somewhat differentially positioned mobile operators TeleFinland and Saunalahti were not included in measurement data. The header collection measurement encompassed 50-60% of all Finnish mobile network packet data traffic to the Internet during the measurement week. As circuit-switched mobile data usage is becoming marginal in Finland, this should also be a sufficient sample size to describe Finnish mobile data usage, although including the third operator and performing the measurement in two one-week periods would remove the remaining concerns on representativeness. The handset-based measurement differed from the two network level measurements in the sense that it was the most dependent on a sample of measured subjects instead of measuring practically all subjects. Nevertheless, the final sample size of 482 panelists should be enough to represent the some 180 000 Finnish Symbian S60 users, providing that panelist demographics correspond adequately to the actual demographic distribution of smartphone users. While the contacting method and used incentives might have included some bias in the sample, the most biasing aspect of recruitment was no doubt the requirement to be capable of installing new applications on the handset that was in practice set on the participants. This led to an even more grave bias in panelist population towards so-called early adopter users, which was verified in the very high average 3rd party application installation frequency among the panelists, even though over 30% of panelists' handsets were not high end smartphone models. Thus, while the handset –based data provides quite representative results on Finnish smartphone usage, direct generalizations are somewhat problematic. However, results concerning traditional communication service usage can be considered more representative than those on application or data usage.

In sum, the network level measurement results can be comfortably generalized to represent the situation at the entire Finnish market. The handset measurement –based results represent the Finnish Symbian S60 user population rather well, although some representativeness issues prevent making direct conclusions based on the handset data. No further generalizations based on the measurement results on mobile subscriber behavior and usage patterns can be made at the international level, although in this regard the handset –based results on smartphone usage might actually be most easily generalized.

7.3 Recommendations

The data collected in the three measurements and the associated analyses could be very valuable to various mobile industry actors. Thus, some concrete recommendations can be provided.

For **mobile operators**, a clear recommendation would be to include some of the header collection –based research items on off portal mobile Internet traffic to their regular management reporting. This would entail information on the application protocol and web traffic profiles corresponding to the usage of mobile Internet services, and potentially also encompass the identification of different subscriber terminal operating systems. Sufficiently accurate subscriber terminal identification, based on either TCP fingerprinting or terminal TAC codes, could also be used in price differentiation when separating modem traffic from truly mobile terminal –originated usage. The handset –based data could be utilized by mobile operators in a number of ways including mobile subscriber segmentation and development of subscription types. The data could also be used for targeted marketing campaigns and differentially priced service offerings to subscriber segments, as the usage of specific services and applications with related usage times can be associated with the subscriber type.

Network manufacturers could be interested in some of the abovementioned issues concerning network level traffic profiles, whereas both **network manufacturers and operators** could use the handset-based data for analyzing network traffic loads.

Handset manufacturers and handset operating system / software platform developers might be interested in the information on various usability aspects present in the handset usage data. Such data could be used e.g. in S60 platform and product user experience planning and design. Handset –based data also enables assessing the potential drivers and bottlenecks related to new technologies and services. These studies could also be sold to or traded with mobile operators to facilitate the sales of handsets and network operation services.

The **regulator** might also find some of the collected data useful. Potential instances for this include both the drafting of new laws and the monitoring of compliance to these laws. In fact, the regulator is known to commission 3rd parties to conduct studies providing similar type of data.

3rd party software companies could use the handset-based data in smartphone application development. Moreover, the analyses made in conjunction with the header collection –based measurements could be productized by a 3rd party, and sold to operators as a tool or managed service.

7.4 Suggestions for Further Research

Considering the measurement methods used in this study, some general suggestions can be given. First and foremost, similar measurements should be repeated regularly to obtain time series data on the development of mobile data usage in Finland. Considering the measurements themselves, the used sample size of three major operators is largely sufficient, unless other mobile Internet –related network technologies are to be included. Adding new measured items (e.g. communication service usage volumes, mobile content usage) can also be envisioned. Some improvement of measurement and analysis phases is required to decrease the amount of uncertainty included in the results. More detailed measurement-specific suggestions are presented below.

Further operator reporting system –based measurements

In future data collection efforts, some enhancements on the used measurement setup could be envisioned. Future measurements could involve the counting of not just the number of different terminal models (by their TAC codes) observed in the network during a certain time period, but counting the volumes of generated packet data traffic volumes per terminal model as well. This would provide another way of evaluating the share of modem traffic out of total traffic volume, although the usage of handsets as modems would not be included in these figures. Moreover, the share of smartphone data usage volumes compared to less capable handsets could also be studied with this type of data collection. Furthermore, this measurement could also be extended to consider Internet APN specific data usage only. Considering the analysis phase, data on some additional features, such as Bluetooth and Java, could be included in feature penetration analyses. Collecting more accurate data on the ages of specific terminal models is also advisable.

Finally, this type of data lays ground for many further research topics. As an example, the effect of the upcoming (April 2006) handset bundling in Finland on mobile terminal installed base and mobile packet data usage could be evaluated by comparing the terminal base and usage volumes before and after the bundling has started.

Further TCP/IP header collection–based measurements

TCP/IP header collection –based measurements could be developed further by improving the measurements and/or the associated analyses.

Some of the inaccuracy involved in identifying application protocols could be cleared using traffic characterization. As each application (protocol) has its own logic, it will result in a certain type of network traffic when considering the traffic beyond the level of individual packets. (Peuhkuri 2003) Whether or not these characteristics could be used to increase the accuracy in identifying applications should be studied.

Considering browsing studies, the used domain name categorization should be refined with more standard categories and a more comprehensive domain name – domain category mapping. More accurate results could also be obtained with the same setup by including HTTP headers (or parts of them) in header collection. While such approach gives more accurate results, it would also complicate the analysis considerably, not to mention the legal aspects involved as e.g. password and phone numbers might be included in plain text in the HTTP traffic. Thus, HTTP header measurements will probably be deemed unfeasible also in the future.

The terminal operating system identification also included a substantial amount of ambiguity due to unidentified operating systems. While some of this will probably be corrected “naturally” as the fingerprint database of the used identification software (p0f) gradually grows, some new terminal operating systems will probably emerge as well. One suggestion to cope with this would be to make sure all major Symbian version releases are identified correctly by p0f. Moreover, some common Nokia proprietary operating systems could be added to the fingerprint base “manually” with a separate arrangement, as the share of e.g. Nokia Series 40 handsets is currently significant in Finland.

Previous research (Kilpi et al. 2005) has separated GSM/GRPS traffic from UMTS traffic from similar traffic traces using differences in TCP flow round trip times (RTT). The applicability of such separation methods in future research should be evaluated.

Kalden (2004) has studied user mobility and (among other issues) its effect on application usage with a measurement setup including a separate SGSN-based measurement in addition to a similar GGSN measurement used in this study. The applicability and potential value of such measurements regarding usage measurements should be evaluated.

Further handset–based usage measurements

Several handset usage measurements using the same research platform used in this study have been planned to the near future. While planning these panels, some issues should be considered. Quota-based recruitment should be considered to avoid representativeness concerns, although background data on the demographics of general smartphone user population is needed as well. As the mandatory beginning questionnaire is the primary method to collect background data on the panelists, its proper formulation is of central importance. A standard set of background questions should be developed to ensure sufficient comparability between different panels. The final questionnaire could be better used to make an additional, supporting survey study on the panelists. Usage of advanced statistical analysis methods beyond descriptive statistics could be envisioned not only on data-specific usage but on all handset functions, although mobile data usage is clearly one of the central areas of interest in handset usage. On the other hand, focused studies with more homogeneous panel participation (only one handset model, focused demographics, members of target organization…) could also lead to interesting discoveries of more specific nature. The adaptation and usage of some new pilot services could also be monitored using the same monitoring technique.

The handset-based data also provides means to study several specific research questions. Handset user mobility and the context of usage of different mobile services is one such topic, i.e. does the usage depend on user location (home, work), type of usage (free time, business), or time (weekday, time of day). Doing comprehensive international comparisons using data collected in separate panels on different markets is another tempting topic. Further research on both of the abovementioned topics has been planned.

Verkasalo (2005) also presented some additional research topics concerning the adoption of new handset features and agglomeration of usage actions. New handset features usually become available to users when they change the handset. It seems that handset usage activity develops in time, meaning there is a short-term testing period during which users try out the new handset features and ultimately select the ones they continue using. After this testing period usage volumes stabilize to a somewhat lower level and applications are used less diversely. This topic could be studied further. Another suggestion is that smartphone usage actions are agglomerated together into usage session, i.e. when a user starts using the handset he will perform several actions and use several applications successively during the same usage session. A more detailed study could reveal if particular usage events trigger certain kind of usage, e.g. if receiving an email on the mobile triggers other mobile Internet usage.

All in all, the overall potential of the research platform increases in the future, as smartphones represent an increasingly large share of the total mobile terminal installed base.

References

The references are divided into four parts: Bibliography, Online Resources, Standards and Specifications, and Software tools.

Bibliography

Aaker David A., Kumar V., Day George S. (2004). *Marketing Research*, 8th ed., John Wiley & Sons Inc.

Benko P., Malicsko G., Veres A. (2004). A large-scale, passive analysis of end-to-end tcp performance over gprs, in *Proceedings of IEEE Infocom 2004*, Hongkong, China.

Blyth W. G. (1991). Panels and Diaries, in Birn R., Hague P., Vangelder P. (Eds.), *A Handbook of Market Research Techniques*, Kogan Page, London, United Kingdom.

Comer D. (2000). *Internetworking With TCP/IP – Vol. I : Principles, Protocols, and Architecture*, 4th ed., Prentice-Hall Inc., New Jersey, US.

Dahmouni H., Rossé D., Morin B., Vaton S. (2005). Impact of Data Traffic Composition on GPRS Performance, in *Proceedings of 19th International Teletraffic Congress – ICT19*, Beijing, China.

Eagle N. (2005). Machine Perception and Learning of Complex Social Systems. Doctoral dissertation, Program in Media Arts and Sciences, Massachusetts Institute of Technology, USA.

Funk J. (2001). *The Mobile Internet: How Japan Dialed Up and The West Disconnected*, ISI Publications Ltd., Kent, United Kingdom.

Ivanovich M., Li J., Neame T., Yin J., Fitzpatrick P. (2004). GPRS Data Traffic Modelling, in *proceedings of the 16th ITC specialist seminar*, Antwerp, Belgium.

Kalden R. (2004). Mobile internet traffic measurement and modeling based on data from commercial GPRS networks. Doctoral dissertation, University of Twente, The Netherlands.

Kalden R., Ekstrom H. (2004). Searching for mobile mice and elephants in GPRS networks. *Mobile Computing and Communications Review*, Vol. 8, No. 4 (October), pp. 37-46.

Kalden R., Ibrahim S. (2004). Searching for self-similarity in GPRS in *Proceedings of the Passive and Active Network Measurement: 5th International Workshop – PAM 2004*. Pp. 83–92.

Kalden R., Varga T., Wouters B., Sanders B. (2003). Wireless service usage and traffic characteristics in GPRS networks, in *Proceedings of the 18th International Teletraffic Congress – ITC18*, Vol. 2, pp. 981-990.

Kalden R., Sanders B. (2004). Cell reselection inter-arrival time investigation for GPRS, in *proceedings of the WiOpt04 Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Cambridge, United Kingdom.

Ketamo H., Multisilta J., Haaparanta H., Suominen M. and Vasama J. (2005). Matkapuhelin vai laajakaista: Nuoret ja viestintä 2005, Research Report 3, Tampere University of Technology. Pori, Finland.

Kilpi J. (2002). A Portrait of a GPRS/GSM Session. Technical Report 279TD(02)040, COST-279, 2002.

Kilpi J. (2003a). Distributional Properties of GPRS/GSM Session Volumes and Durations. Technical Report 279TD(03)023, COST-279, 2003.

Kilpi J. (2003b). A Portrait of a GPRS/GSM Session, in *proceedings of the 18th International Teletraffic Congress – ITC18*, Berlin, Germany.

Kilpi J., Lassila P. (2005). Micro- and macroscopic analysis of RTT variability in GPRS and UMTS networks, in *Proceedings of Networking 2006*, Coimbra, Portugal.

Korhonen J. (2001). *Introduction to 3G Mobile Communications*, Artech House Inc., Norwood, MA, USA.

Minges M. (2004). Superstars of Mobile Internet. Telecommunications Management Group Inc., Arlington, VA, USA.

Minges M. (2005). Is the Internet Mobile? Measurements from the Asia-Pacific region, *Telecommunications Policy*. Vol. 29, Issues 2-3 (March-April 2005), pp. 113-125.

Raento, M., Oulasvirta, A., Petit, R., Toivonen, H. (2005). ContextPhone – A Prototyping Platform for Context-Aware Mobile Applications. *IEEE Pervasive Computing*, Vol. 4, No. 2 (April-June 2005), pp. 51-59.

Ricciato F., Vacirca F., Karner M. (2005). Bottleneck Detection In UMTS Via TCP Passive Monitoring: A Real Case, in *Proceedings of ACM CoNEXT 05*, Toulouse, France.

Salokangas T. (2002). Teknillinen korkeakoulu opiskeluympäristönä sukupuolten tasa-arvon näkökulmasta. Master’s thesis, Faculty of Social Sciences, University of Helsinki, Helsinki, Finland.

Snellman K. (2005). Mobiilipalvelumarkkinat Suomessa 2004 (Mobile services market in Finland 2004), *Publications of the Ministry of Transport and Communications 34/2005*, Edita, Helsinki, Finland.

Vacirca F., Ziegler T., Hasenleithner E. (2006). An Algorithm to detect TCP Spurious Timeouts and its Application to Operational UMTS/GPRS Networks, *Computer Networks Journal*, Vol. 11, 2005.

Van Veen N. (2005). European Mobile Forecast: 2005 To 2010. Forrester Research Incorporated.

Varga T., Haverkamp B., Sanders B. (2004). Analysis and modeling of WAP traffic in GPRS networks, in *proceedings of the 16th ITC specialist seminar*, Antwerp, Belgium.

Verkasalo H. (2005). Handset-Based Monitoring of Mobile Customer Behavior. Master's thesis, Networking Laboratory, Helsinki University of Technology, Espoo, Finland.

Viipuri T. (2004). Traffic Analysis and Modeling of IP Core Networks. Master's thesis, Networking Laboratory, Helsinki University of Technology, Espoo, Finland.

Online Resources

BABT (2005). BABT – Certification [online]. British Approvals Board of Telecommunications (BABT), updated June 2005 [accessed July 2005]. Available in www-format: <URL: http://www.babt.com/gsm-imei-number-allocation.asp>

Digitoday (2005a). Samsung rynnistää myös Suomessa [online]. SanomaWSOY Plc., updated January 7 2005 [accessed March 10 2006]. Available in www-format: <URL: http://www.digitoday.fi/showPage.php?page_id=12&news_id=39272>

Digitoday (2005b). Kännykkämyynnin vuosiennuste nousi 1,8 miljoonaan puhelimeen [online]. SanomaWSOY Plc., updated October 14 2005 [accessed March 10 2006]. Available in www-format: <URL: http://www.digitoday.fi/showPage.php?page_id=12&news_id=49130>

Elisa (2006). Elisa – Tilinpäätös 2005 [online]. Elisa Plc. [accessed March 15, 2006]. Available in www-format: <URL: http://www.elisa.fi/ir/docimages/attachment/sijoittajat/Q4%202005%20Suomi.pdf>

Ficora (2005). Matkaviestinnän datapalveluiden hintavertailu [online]. Ficora, December 2005 [accessed February 15, 2006]. Available in www-format: <URL: http://www.ficora.fi/suomi/document/mobiilidatapalvelut_12_05_JULK.pdf>

FNC (1995). Definition of "Internet" [online]. US Federal Networking Council, updated 30/10/1995 [accessed June 20, 2005]. Available in www-format: <URL: http://www.nitrd.gov/fnc/Internet_res.html>

GSMA (2005). GPRS Class Types [online]. GSM Association, updated 2005 [accessed September 29, 2005]. Available in www-format: <URL: http://www.gsmworld.com/technology/gprs/class.shtml>

GSMA (2005a). GSM World – Using GSM – Tapping the Potential of Roaming [online]. GSM Association, updated 2005 [accessed August 25, 2005]. Available in www-format: <URL: http://www.gsmworld.com/using/billing/potential.shtml>

GSMA (2005b). GPRS Roaming Guidelines [online]. GSM Association, updated 2005 [accessed August 25, 2005]. Available in www-format: <URL: http://www.gsmworld.com/documents/ireg/ir33.pdf>

Heimo (2005). Heimo Tribe [online]. Heimo Consulting, Helsinki, updated January 2005 [accessed October 5, 2005]. Available in www-format: <URL: http://www.heimotribe.com>

Honeynet (2002). Passive Fingerprinting [online]. Honeynet Project, updated March 4 2002 [accessed August 2005]. Available in www-format: <URL: http://project.honeynet.org/papers/finger/>

ITviikko (2005). Matkapuhelinten markkinatiedot paranevat [online]. SanomaWSOY Plc., updated January 7 2005 [accessed March 10 2006]. Available in www-format: <URL: http://www.itviikko.fi/uutiset/uutisalue.asp?alue=ratkaisut&UutisID=65377>

LEAD (2006). LEAD – Optimal Rules for a Leading Mobile Data Market [online]. Helsinki University of Technology, Espoo, updated February 17 2005 [accessed March 22, 2006]. Available in www-format: <URL: http://www.netlab.hut.fi/tutkimus/lead>

Microsoft (2005a). Mobile Solutions, Mobile Applications, and Handheld Devices from Microsoft Windows Mobile [online]. Microsoft Corporation, updated 2005 [accessed July 25, 2005]. Available in www-format: <URL: http://www.microsoft.com/windowsmobile/>

Microsoft (2005b). Windows CE Home Page [online]. Microsoft Corporation, updated 2005 [accessed July 25, 2005]. Available in www-format: <URL: http://msdn.microsoft.com/embedded/windowsce/default.aspx>

mTLD (2006). dotmobi – mobile internet switched on [online]. mTLD Mobile Top Level Domain Ltd., updated 2006 [accessed April 4, 2006]. Available in www-format: <URL: http://pc.mtld.mobi/>

New Scientist (2006). Cellphones search more sex than ever [online]. Reed Business Information Ltd., updated February 25 2006 [accessed March 10 2006]. Available in www-format: <URL: http://www.newscientist.com/channel/sex/mg18925406.200>

Nokia (2005). Nokia – Platform Architectures [online]. Nokia Corporation, updated 2005 [accessed September 29, 2005]. Available in www-format: <URL: http://www.nokia.com/nokia/0,,62611,00.html>

Numpac (2006). Matkapuhelinnumeroiden siirtomäärät vuosina 2003, 2004 ja 2005 [online]. Numpac, January 2006 [accessed February 15, 2006]. Available in www-format: <URL: http://www.numpac.fi/tiedostot/Siiirtotilastot_020106.ppt>

Otantatutkimus (2005). Telepalvelututkimus 2005 [online]. Otantatutkimus Oy [accessed March 2006]. Available in www-format: <URL: http://www.mintc.fi/oliver/upl103-Telepalvelututkimus%202005.pdf>

Taloustutkimus (2005). Matkapuhelimen käyttö ulkomailla, kevät 2005, Taloustutkimus Oy [accessed March 2006]. Available in www-format: <URL: http://www.ficora.fi/suomi/document/roaming2005.pdf>

Tietosuojavaltuutettu (2005). Tietosuojavaltuutetun toimisto – Data protection in Finland [online]. Office of the Data Protection Ombudsman [accessed July 15, 2005]. Available in www-format: <URL: http://www.tietosuoja.fi/>

Standards and Specifications

3GPP TS 23.002. Third Generation Partnership Project, Technical Specification 23.002. Network architecture, V6.8.0 (2005-06).

3GPP TS 23.003. Third Generation Partnership Project, Technical Specification 23.003. Numbering, addressing and identification, V6.7.1 (2005-07).

3GPP TS 23.101. Third Generation Partnership Project, Technical Specification 23.101. General Universal Mobile Telecommunications System (UMTS) architecture, V6.0.0 (2004-12).

3GPP TS 29.060. Third Generation Partnership Project, Technical Specification 29.060. General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface, V6.9.0 (2005-06).

3GPP TS 32.101. Third Generation Partnership Project, Technical Specification 32.101. Telecommunication management; Principles and high level requirements, V6.1.0 (2004-12).

3GPP TS 32.240. Third Generation Partnership Project, Technical Specification 32.101. Telecommunication management; Charging management; Charging architecture and principles, V6.0.0 (2004-09).

Shirley R. (2000). RFC2828 –Internet Security Glossary, The Internet Society.

WAP Forum (2001a). WAP Architecture, WAP-210-WAPArch-20010712, Version 12-July-2001.

WAP Forum (2001b). Wireless Profiled TCP, WAP-225-TCP-20010331-a, Version 31-March.2001.

WAP Forum (2001c). Wireless Profiled HTTP, WAP-229-HTTP-20010329-a, Version 29-March.2001.

Software tools

TCPDUMP network traffic capturing and analysis tool, available online: <URL: http://www.tcpdump.org>

CoralReef network traffic collecting and analysis software suite, available online: <URL: http://www.caida.org/tools/measurement/coralreef/>

p0f passive operating system fingerprinting tool, available online: <URL: http://lcamtuf.coredump.cx/p0f.shtml>

Appendix A Network-Based Measurement of Mobile Internet Traffic

This document specifies the techno-economic traffic analysis to be performed in the LEAD project by TKK and the partner operators TeliaSonera, Elisa, and DNA. Operators use their reporting processes and collect the traffic data to generate the operator-specific reports as defined in this document. TKK will combine the operator-specific data and publish averaged reports to characterize the status of the Finnish market. The operator-specific data is considered confidential according to the existing LEAD contract and will be handled confidentially by the relevant TKK persons N/N. Additional non-disclosure agreements will be signed by the TKK persons as necessary.

Time Period

Traffic data will be collected during two full-week periods to secure at least one successful analysis. The selected measurement periods are:

- Week 34, Aug 22-28, 2005
- Week 38, Sep 19-25, 2005

A full week (from Monday 00:01 AM to next Monday 00:01 AM) is needed to absorb the weekend vs. weekday traffic variation. The time between the periods should be enough to analyze the results of the first period and reiterate the reporting requirements for the second period. It is recognized that some unpredictable marketing campaigns of operators may distort the results. Furthermore, deviation of traffic volume on measurement periods from weekly averages will be evaluated separately.

The final results are planned to be ready by 12/2005 when the LEAD project finishes.

Types of Measurement

Two types of measurement are likely to be needed:

- A measurement based on regular traffic reporting
- A measurement based on additional TCP/IP header collection

The operator's regular charging-oriented traffic reporting can be used to retrieve most of the needed data. However, a detailed analysis of Internet traffic profiles (top-10 Internet protocols, top-10 Internet sites, per operating system type) requires additional data collection and fingerprinting of TCP/IP packet headers.

Types of Questions

The operator-specific reporting questions will be based mainly on percentage information. However, to secure correct averaging some basic operator-specific information on volumes is needed during the measurement periods.

A.1 Regular Traffic Reporting –Based Measurements

The functioning operator-specific reporting systems must be understood sufficiently to ensure comparability of the reports. This includes evaluating how accurately the regular reporting can single out targeted subscriptions/terminals/traffic and how representative sample is used for the reports. Separate questions will be posed for this.

Types of Subscriptions

Subscriptions will be studied to understand the status of active packet data subscribers. All subscriber questions refer to the MNO's service and brand operators' (called "internal MSO" for clarity) postpaid subscribers. While the focus of the traffic analysis is in private consumer subscribers, some information on business subscribers is also needed.

GPRS tariff alternatives are classified as follows: usage-based (euro/MB, with and without monthly fee), block-based (2, 10, 20, 100, 200, or 500 MB, and other possible alternatives), flat-rate (fixed monthly rate), and roaming (all tariff classes).

Out of all (postpaid internal MSO) subscriptions:

What is the total number of MSO subscriptions? (for averaging only)
- *What is the percentage of consumer subscriptions?*
- *What is the percentage of business subscriptions?*

Out of all (postpaid internal MSO) consumer subscriptions:

What is the percentage of consumer subscriptions by GPRS tariff alternative?
- *GPRS not activated?*
- *Usage-based (without monthly fee)?*
- *Block-based (10 MB)?*
- *Block-based (100 MB)?*
- *... other possible block sizes*
- *Flat-rate?*
- *Roaming (all tariff classes)?*

What is the percentage of consumer subscribers using GSM voice roaming?
What is the percentage of consumer subscribers using GPRS roaming?

An *active packet data subscriber* is defined as a (postpaid) subscriber having generated any packet data traffic during the measured one-week period. While the measurement period is not sufficient to capture the number of all active packet data subscribers, it gives information on average weekly usage.

What is the percentage of active consumer packet data subscribers?

Out of all (postpaid internal MSO) subscriber packet traffic:

What is the total packet traffic volume (KB) generated by all subscribers?
- *What is the percentage of traffic generated by consumer subscribers?*
- *What is the percentage of traffic generated by business subscribers?*

Out of all (postpaid internal MSO) consumer subscriber packet traffic:

What is the percentage of packet traffic volume per GPRS tariff alternative?
- *GPRS not activated?*
- *Usage-based (without monthly fee)?*
- *Block-based (10 MB)?*

Block-based (100 MB)?
... other possible block sizes
Flat-rate?
Roaming (all tariff classes)?

Types of Terminal

The current installed base of all mobile terminals (with and without packet data capabilities) will be studied to test the hypothesis that the Finnish terminal base is lagging behind competition.

Information from the following questions will be used to derive the Finnish top–50 list of mobile terminals, the percentage shares of terminal vendors, terminal operating systems, Nokia Series 60 –based terminals, packet data capable terminals, and the shares of different radio technologies (GSM / EDGE / WCDMA, and WLAN) for the whole terminal base. Further analysis for certain terminal features is also possible based on the collected information.

Out of all online mobile terminals:

What is the total number of mobile terminals? (for averaging only)
What is the top-100 list (number of terminals) of mobile terminals, including:
Terminal model or TAC-code?
Percentage of total for each model?
What is the percentage share of the rest of the terminal base?

Ideally, the top-100 list above should include the terminals that have been online anytime during the measurement week, including the postpaid and prepaid subscribers of all MSOs in the MNO network and excluding foreign roamers. However, it is recognized that the information might not be available at this extent from all operators. Thus, information on only postpaid subscribers or “internal MSO” subscribers (see “Types of Subscriptions” above) is also sufficient. Similarly, a report considered representative by the operator is satisfactory, even if it is based on a period shorter or longer than a week. These departures from the ideal situation should be pointed out with the results.

Types of Traffic

Packet traffic will be studied to understand the usage profile of active packet data users. The focus of the measurements in this section is postpaid internal MSO subscribers using the MNO network. If possible, all external MSO subscribers and foreign roamers using the MNO network should be excluded from the results.

In questions referring the cellular radio technologies (GSM/GPRS, EDGE, and WCDMA) of mobile terminals, multiradio terminals are classified by the fastest radio.

Out of all (internal MSO) network users:

What is the total number of mobile network users? (for averaging only)
What is the percentage of postpaid internal MSO subscribers?
What is the percentage of prepaid internal MSO subscribers?

As prepaid subscribers are a minority user category with presumably different usage patterns from postpaid subscribers, they should be separated in the measurements. Thus, all subsequent questions refer solely to postpaid subscribers of the internal MSO using the MNO network.

Out of all postpaid (internal MSO) network users:

An *active packet data user* is defined as a mobile network user having generated any packet data traffic during the measured one-week period. While the measurement period is not sufficient to capture the number of all active packet data users, it gives information on average weekly usage.

What is the percentage of active postpaid packet data users?
What is the percentage of active postpaid packet data users using ...
... GSM/GPRS capable terminals?
... EDGE capable terminals?
... WCDMA capable terminals?

Out of postpaid (internal MSO) network users' packet traffic:

What is the total volume (KB) of the packet traffic (for averaging only)?
What is the percentage of packet traffic volume generated by...
... GSM/GPRS capable terminals?
... EDGE capable terminals?
... WCDMA capable terminals?

"Internet traffic" in this context refers to consumer-generated GPRS packet data traffic going to independent 3rd party sites not controlled by the MNO or MSOs. It is recognized that such traffic might not be completely separated from other packet traffic for accurate measurements. While this "Internet traffic" might be somewhat heterogeneous (e.g. include both consumer and business subscriber traffic), the following question should be answered using the most suitable point of measurement (e.g. Internet APN).

What is the percentage of Internet traffic?

How well this result represents consumer Internet traffic will be evaluated separately.

A.2 TCP/IP Header Collection –Based Measurement

This additional measurement is needed for data that cannot be tracked via regular reporting mechanisms. The point of measurement should be chosen so that the traffic is consumer-only, Internet-only, and fingerprint-enabled (no scrambling of headers other than sender IP address). Furthermore, it should capture a representative portion of mobile Internet traffic created by private consumer subscribers. The representativeness of captured traffic will be evaluated separately. The collected TCP/IP headers (without payload) should be analyzed off-line.

Measurement Targets and Methods of Analysis

Information retrieved from the following operator-specific questions will be used to derive the top-10 of application protocols, the top-10 of visited Internet sites, and the corresponding top-10 lists for all operating systems, aggregated for the traffic captured at all participating operators' sites.

Out of all MNO (consumer) Internet traffic:

What is the total volume (KB) of captured Internet traffic (uplink and downlink)?
What is the total number of sites visited (each visit counts as one)?
What is the top-20 of application protocols?
  Protocol, percentage of total traffic?
What is the top-50 of Internet sites?
  IP address, percentage of all site visits?

The top-lists can be extracted from the collected header data using a separate software/script prepared for the task.

What is the percentage of traffic per operating system version?
  All operating system versions

The `p0f` passive OS fingerprinting tool (http://lcamtuf.coredump.cx/p0f.shtml) can extract this from the collected header data.

What is the percentage of laptop traffic?
What is the top-20 of application protocols per operating system version?
  All operating system versions: protocol, percentage of total traffic?
What is the top-50 of Internet sites per operating system version?
  All operating system versions: IP address, percentage of all site visits?

These questions can be answered using a combination of the two analysis methods mentioned above (separate software/script, and `p0f`).

Header Data Collection in Practice

In order to be able to perform the analysis described above, the following TCP/IP header information should be collected:

- All TCP packet headers of packets establishing and closing TCP connections (SYN, FIN, RST) and corresponding IP headers, with all header fields of both TCP and IP included.
- All UDP packet headers and corresponding IP headers, with all header fields of both UDP and IP included.

The `pcap` file format (used by e.g. `tcpdump` and `ettercap`) should be used in the collection of header data as it enables the use of several existing tools in subsequent analysis. Sensitive information such as sender IP addresses can be removed from the headers (stored in `pcap` format) by replacing them with anonymous addresses using the `tcpdpriv` tool (http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html).

The methods used in the actual implementation of the packet header data collection depend on the network equipment available at each operator's site.

Appendix B Categorization of application protocols and web domains

B.1 Categorization of Application Protocols

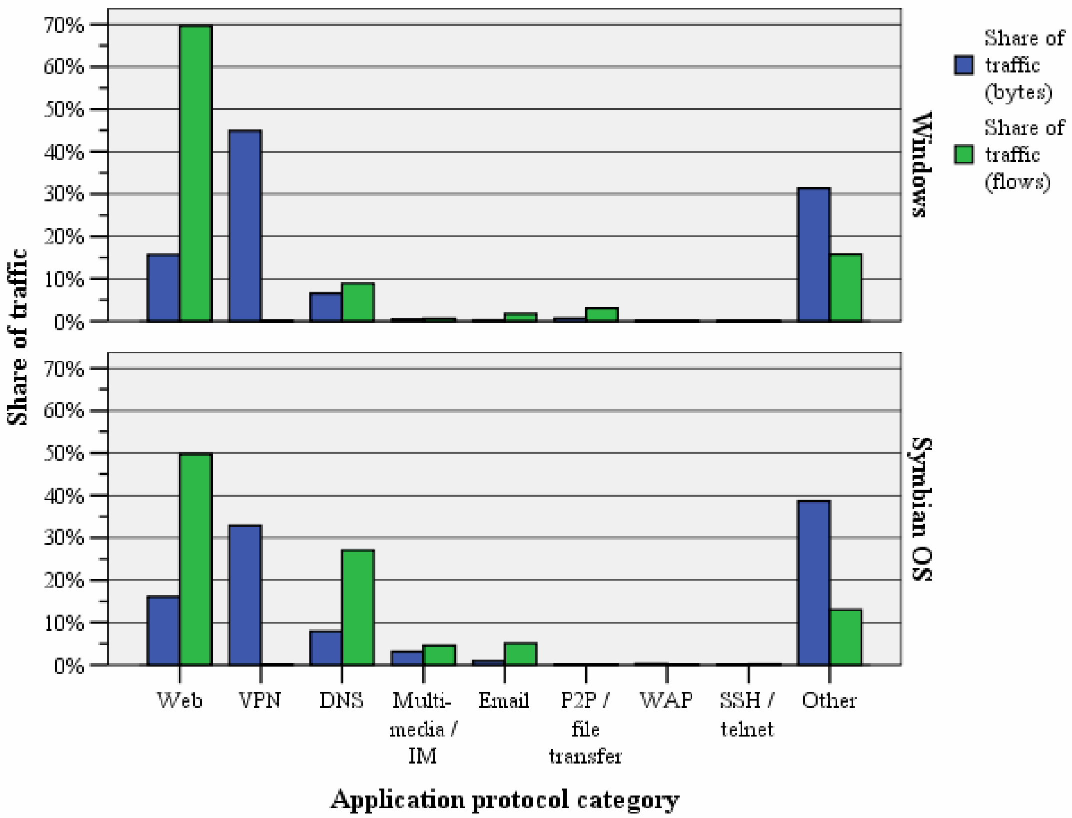
Site category	Type of application protocol included	Exemplary TCP and UDP ports included
Web	HTTP, HTTPS, HTTP Alternate	TCP: 80, 443, 8080
VPN	ISAKMP, cpudpencap, IPSec NAT-T, NDMP, Open VPN	TCP: 10000 UDP: 500, 2746, 4500, 10000, 1194
DNS	DNS	UDP: 53
Multimedia / IM	Audio / video conferencing protocols, streaming, IM, IRC	TCP: 1863, 5001, 6667, 554 UDP: 5003, 5001, 5000
Email	SMTP, IMAP, POP3	TCP: 110, 143, 25, 993, 995
P2P / file transfer	FTP, several potential P2P protocols	TCP: 6346, 4662, 21, 20, 1214 UDP: 4672, 6681
WAP	All registered WAP protocols	TCP: 9200, 2949, 2805, 2923 UDP: 9201, 9203, 9202, 9204
SSH / telnet	SSH and telnet	TCP: 22, 23
Other	All other protocols and unidentified protocols	–

B.2 Categorization of Web Domains

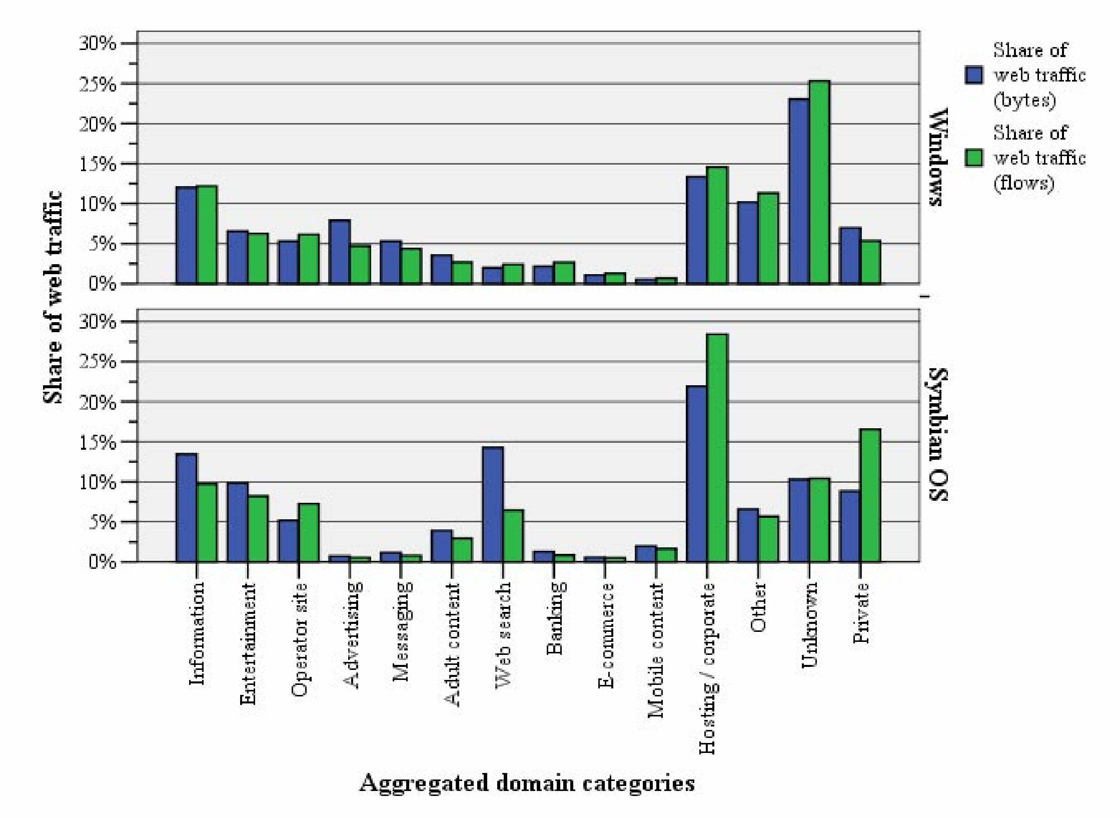
Site category	Type of domain included	Exemplary domain names included
Adult content	Sites with adult content, sex related dating/chat services	sihteeriopisto.fi, seksitreffit.fi
Advertising	Pop-up sites	doubleclick.net, adtech.de, tradedoubler.com
Banking	Banking, insurance, and real estate related sites	sampo.fi, eQonline.fi, op.fi, nordea.fi
E-commerce	E-commerce, auctions, online rental services, and other similar sites	huuto.net, mobile.de, infosto.fi, thomann.fi, ebay.com, verkkokauppa.com
Entertainment	Music, movies, TV, sports, games, gambling, other hobby related	irc-galleria.net, veikkaus.fi, sm-liiga.fi, telkku.com
Hosting / corporate site	Web hosting providers, mobile email/calendar hosting, load balancing servers Commercial sites not included in other categories	basefarm.net, nebula.fi, akamaitechnologies.com
Information	Newspapers, public sector sites, weather, timetables	mtv3.fi, yle.fi, sanomawsoy.fi, almamedia.fi, helsinginsanomat.fi
Messaging	Email and IM related sites	luukku.com, hotmail.com, gmail.afraid.org, msn.com, passport.com, passport.net
Mobile content	Sites providing ring tones, mobile games, logos…	jippii.net, jamster.com, mobilenator.com, buumi.net
Operator site	Finnish mobile/fixed operators	–
Other	Domains not fitting to any other category, and uncategorized domains	–
Private	Private addresses	–
Unknown	Unidentified domains	–
Web search	Web search sites / portals	google.com, yahoo.com

Appendix C TCP/IP Header Collection –Based Results

C.1 Distribution of Application Protocol Traffic by Operating System

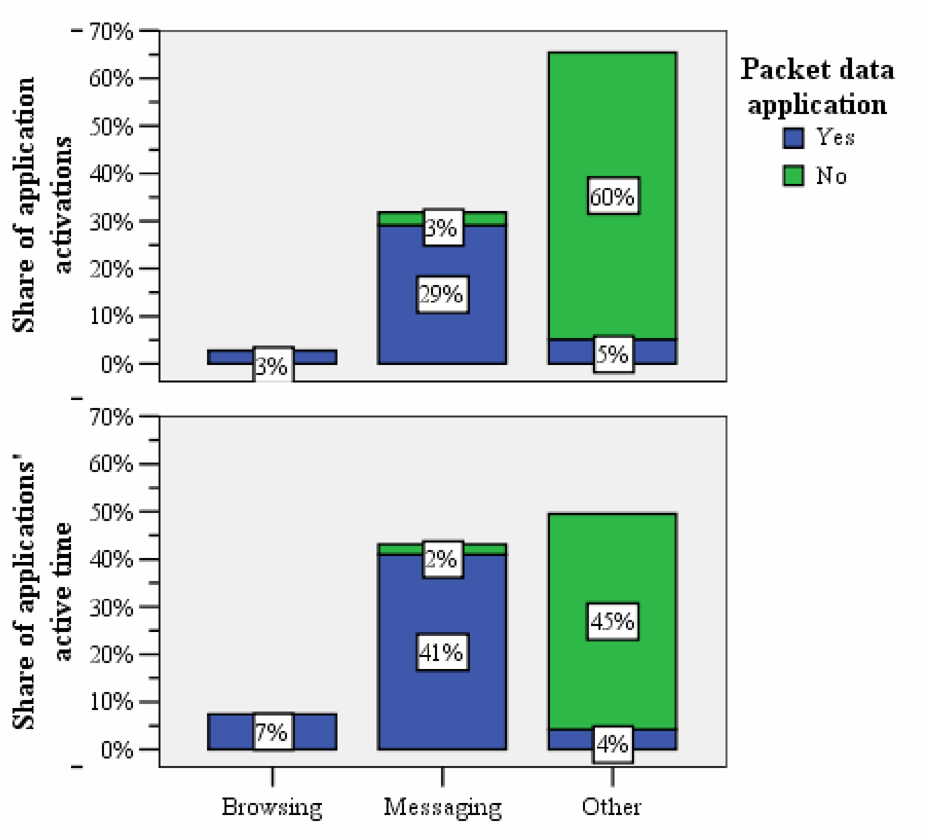


C.2 Distribution of Web Protocol Traffic by Operating System

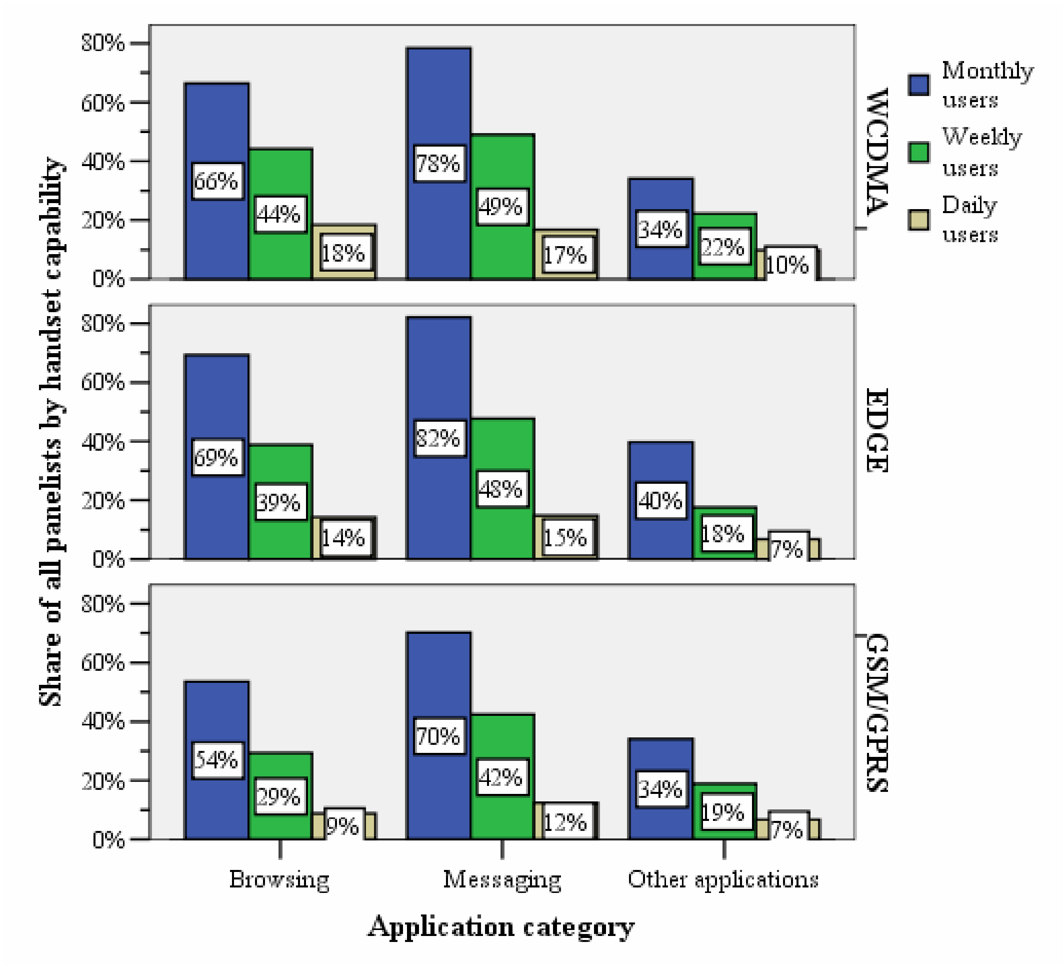


Appendix D Handset Usage Measurement Results

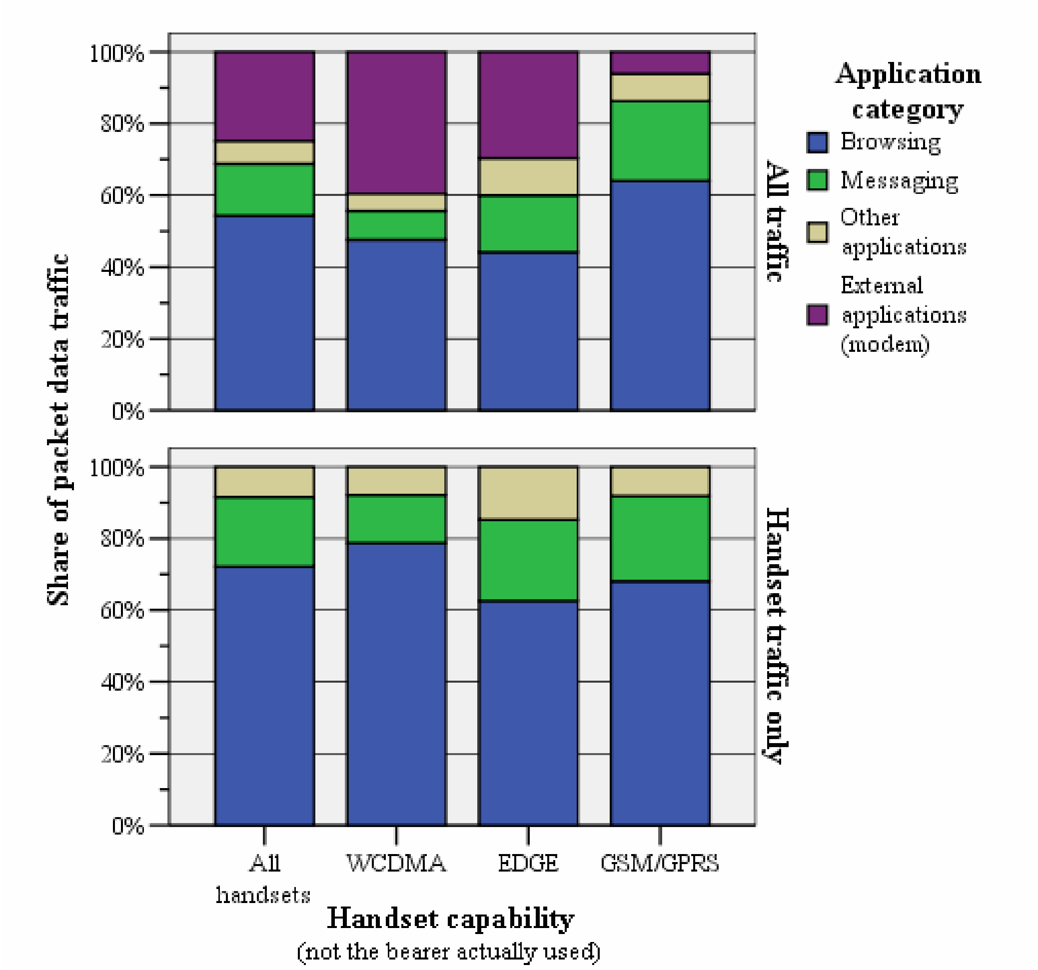
D.1 Handset Application Usage Activity



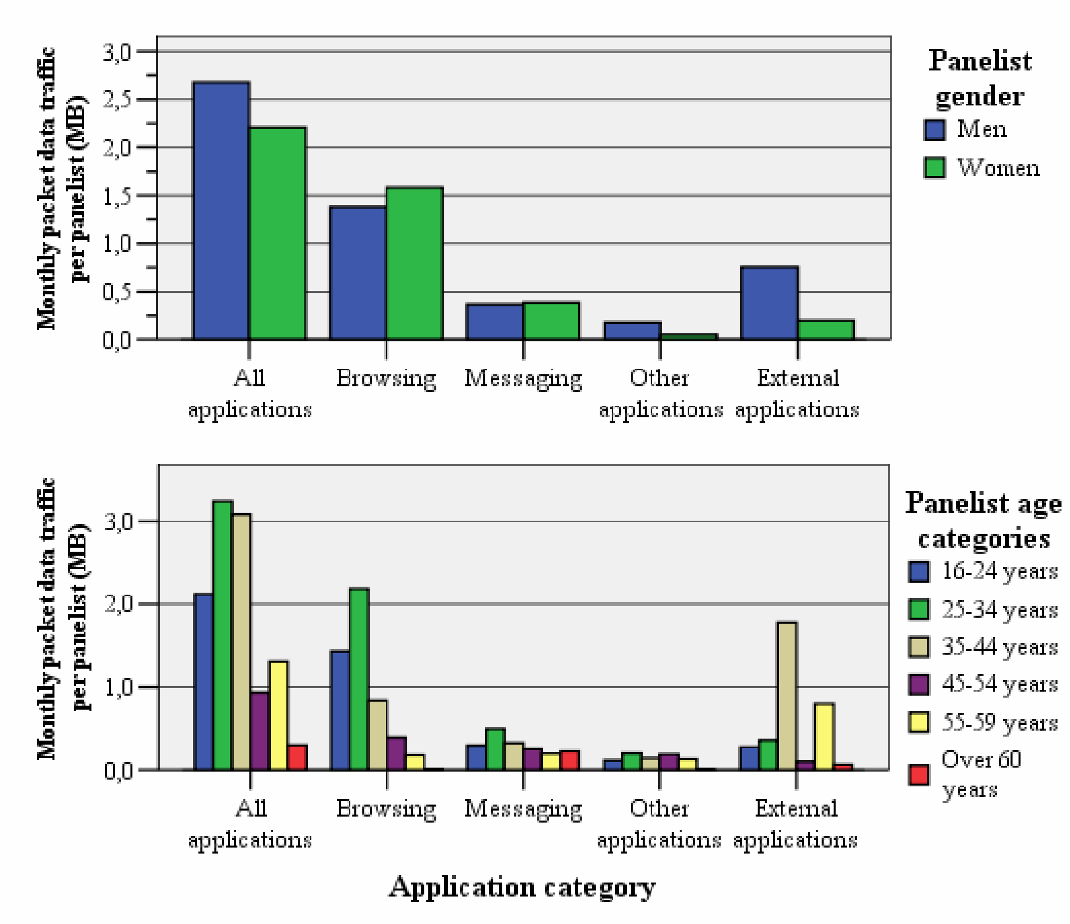
D.2 Packet Data Application Usage Frequencies



D.3 Packet Data Application Usage Volumes



D.4 Packet Data Application Usage by Gender and Age



D.5 Individual Level Browsing Patterns