HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Communications and Networking

Antti Riikonen

MOBILE INTERNET USAGE - NETWORK TRAFFIC MEASUREMENTS

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

Espoo, September 1st, 2009

Supervisor:	Heikki Hämmäinen Professor, Networking Business
Instructor:	Antero Kivi M.Sc. (Tech)

HELSINKI UNIVERSITY OF TECHNOLOGY

Abstract of the Master's Thesis

Author: Antti Riikonen

Title: Mobile Internet Usage - Network Traffic Measurements

Date: 1.9.2009 **Number of pages:** 9+92

Department: Department of Communications and Networking

Professorship: S-38 Networking Technology

Supervisor: Prof. Heikki Hämmäinen

Instructor: Antero Kivi M.Sc. (Tech)

Fundamental transformations are taking place in the telecommunication domain as the Internet and mobile industries are converging. Mobile phones are developing into multimedia computers and laptops are getting smaller with cellular connectivity, increasing the amount of mobile Internet capable devices. Furthermore, as mobile broadband prices have decreased and offered bandwidths increased, also the usage of mobile Internet has been increasing rapidly during the past couple of years. In search for new revenue sources, various industry stakeholders are interested in measurements that can help understanding the mobile Internet usage patterns.

This thesis focuses on mobile network traffic measurements and studies their applicability for providing market understanding for the different stakeholders. First, measurements from operational Finnish mobile networks are analyzed to provide factual statistics on the usage patterns of the Finnish market. Second, the properties of the existing measurement organization are analyzed, possible measurement design and development areas are classified, and recommendations are provided for further development of the measurements.

The factual statistics showed that most of the Finnish mobile Internet usage traffic volume is generated by computers, whereas the share of mobile handset generated traffic is less than one percent. Symbian operating system dominates the web oriented mobile handset usage. Traditional Finnish media houses, social media sites, and Nokia are among the most popular content providers for web usage. In addition, also other web traffic classes than web browsing, such as email and synchronization, were observed to be used by mobile handsets.

As different measurement points in a mobile network provide different data granularity, the choices related to the measurement have to be made according to the objectives of the measurement. If advanced analysis is needed, the measurements are recommended to be conducted at a point in the mobile network where user identification is possible, whereas total traffic level patterns from IP traffic are adequate for general market description. From a mobile operator viewpoint, automated and continuous data collection and analysis could enable utilization of the results in multiple corporate functions. In general, the possibilities of traffic measurements are vast. On the other hand, they may require a lot of resources to succeed in their full potential. Nevertheless, mobile network traffic measurements can provide intelligence and support for operators in their decision making and business development.

Keywords: Mobile Internet, Traffic Measurement, Mobile Service Usage

TEKNILLINEN KORKEAKOULU

Diplomityön tiivistelmä

Tekijä: Antti Riikonen

Työn nimi: Mobiilin Internetin käyttö - verkon liikennemittaukset

Päivämäärä: 1.9.2009 **Sivumäärä:** 9+92

Laitos: Tietoliikenne- ja tietoverkkotekniikan laitos

Professuuri: S-38 Tietoverkkotekniikka

Työn valvoja: Prof. Heikki Hämmäinen

Työn ohjaaja: DI Antero Kivi

Perustavanlaatuisia muutoksia on tapahtumassa tietoliikennetoimialalla kun Internet ja mobiili konvergoituvat. Matkapuhelimet ovat kehittymässä multimediatietokoneiksi ja kannettavat tietokoneet muuttuvat pienemmiksi ja sisältävät kasvavissa määrin liitettävyyden matkapuhelinverkkoon. Kun samaan aikaan mobiililaajakaistojen hinnat ovat laskeneet ja tarjotut kaistannopeudet kasvaneet, mobiilin Internetin käyttö on lisääntynyt nopeasti viimeisten parin vuoden aikana. Uusia tulonlähteitä etsiessä teollisuuden eri sidosryhmät ovat kiinnostuneita mittauksista jotka voivat auttaa ymmärtämään mobiilin Internetin käyttöä.

Tämä diplomityö keskittyy mobiiliverkon liikennemittauksiin ja niiden soveltuvuuden tutkimiseen markkinatiedon tuottamisessa eri sidosryhmille. Työssä analysoidaan suomalaisissa mobiiliverkoissa tehtyjä liikennemittauksia ja tuotetaan statistiikkaa mobiilin Internetin käytöstä. Lisäksi nykyisen mittausjärjestelyn ominaisuuksia analysoidaan, mahdolliset mittausten kehittämisalueet luokitellaan ja suosituksia esitetään mittausten kehittämiseksi.

Statistiikat Suomen mobiilii-Internetin käytöstä osoittivat että tietokoneet generoivat suurimman osan Suomen liikennevolyymistä, kun taas matkapuhelinten tuottaman liikenteen osuus on alle prosentti. Symbian-käyttöjärjestelmä dominoi matkapuhelinten käyttöä joka on web-orientoitunutta. Muitakin web-liikenteen luokkia kuin web-selailua, kuten email- ja muuta synkronointiliikennettä, huomattiin käytettävän matkapuhelimilla. Perinteiset kotimaiset mediatalot, sosiaalisen median sivustot ja Nokia ovat suosituimpien web-sisällöntuottajien joukossa.

Koska eri mittauspisteet mobiiliverkossa tuottavat eri tarkkuustason tietoa, täytyy mittausprosessiin liittyvät valinnat tehdä mittauksen tavoitteiden mukaan. Jos kehittynyt analyysi on tarpeen, mittausten olisi suositeltavaa suoritettavan pisteessä jossa käyttäjien tunnistaminen on mahdollista, kun taas pelkästä IP-liikenteestä saatavat tulokset ovat riittäviä yleisempään markkinan kuvaamiseen. Mobiilioperaattorin kannalta yhtäjaksoiset ja automatisoidut mittaukset mahdollistaisivat tulosten hyväksikäyttämisen useissa eri yrityksen toiminnoissa. Yleisesti liikennemittausten mahdollisuudet ovat laajat, mutta toisaalta täyden hyödyn saavuttamiseksi resurssivaateet saattavat olla suuria. Liikennemittaukset voivat kuitenkin tuottaa tietoa ja tukea operaattoreita päätöksenteossa ja liiketoiminnan kehittämisessä.

Avainsanat: Mobiili Internet, Verkkoliikenteen mittaukset, Mobiilipalveluiden käyttö

Preface

This Master's Thesis has been written as a partial fulfillment for the Master of Science degree in Helsinki University of Technology. The work has been conducted as a deliverable for the MoMI project in the Department of Communications and Networking. 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, I would like to thank Professor Heikki Hämmäinen for the opportunity to work in the team and to write the thesis under his guidance. I am grateful to Antero Kivi for his extensive assistance and support throughout the whole research process.

I would also like to express my gratitude to Markus Peuhkuri for the assistance and advice during the measurements. In addition, I would like to thank the other co-workers in the Networking Business -team for the comments and discussion regarding my work. I especially wish to thank Juuso Karikoski for reviewing the Thesis and Timo Smura for reviewing the project report from the measurements.

I would like to thank my family, friends, and fellow students for the support during my studies. To my dearest, Emppu, thank you for the strength and wisdom you gave me during the course of my work. Finally, I want to thank my parents, Aila and Raimo, for all the support they have provided me.

Espoo, September 1st, 2009

Antti Riikonen

Table of Contents

Table of figures

List of tables

Acronyms and terms

2G	2nd Generation
3G	3rd Generation
3GPP	3rd Generation Partnership Project
API	Application Programming Interface
APN	Access Point Name
ARPU	Average Revenue per User
AuC	Authentication Center
BG	Border Gateway
BLINC	Blind Classification
BSS	Base Station Subsystem
COIN	Dynamics of Competition and Innovation in the Converging Internet and Mobile Network
CS	Circuit Switched
DAG	Data Acquisition and Generation
DNS	Domain Name System
DPI	Deep Packet Inspection
Econ@Tel	Telecommunications Economics COST Network, COST605
EDGE	Enhanced Data rates for Global Evolution
ESP	Encapsulated Security Payload
FTP	File Transfer Protocol
Gb	Gb interface is a GPRS interface between a BSS and a SGSN
GERAN	GSM EDGE Radio Access Networks
GGSN	Gateway GPRS Support Node
Gi	Gi Interface is a GPRS interface which is located between the GGSN and the external PDN (Public Data Network)
GMM/SM	GPRS Mobility Management and Session Management
Gn	Gn interface is a GPRS interface which is located between the GSNs

Gp	Gp interface is a GPRS interface between GSNs within different PLMN
GPRS	General Packet Radio Service
GPS	Global Positioning System
Gs	Gs interface is an interface between a SGSN and a MSC
GSM	Global System for Mobile communications
GSN	GPRS Support node
GTP	GPRS Tunneling Protocol
GTP-C	GTP control traffic (see GTP)
GTP-U	GTP user traffic (see GTP)
GTVS	Ground Truth Verification System
HLR	Home Location Register
HSDPA	High-Speed Downlink Packet Access
HSPA	High speed packet data access
HSUPA	High-Speed Uplink Packet Access
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IMAP	Internet Message Access Protocol
IMEI	International Mobile station Equipment Identity
IMEISV	International Mobile station Equipment Identity and Software Version number
IMS	IP Multimedia Subsystem
IMSI	International Mobile Subscriber Identification
IP	Internet Protocol
IuPS	IuPS interface is a UMTS interface which links the RNC with a SGSN
LEAD	Optimal Rules for a Leading Mobile Data Market
LTE	Long Term Evolution
ME	Mobile Equipment
MIME	Multipurpose Internet Mail Extensions
MINTC	Ministry of Transportation and Communications Finland

MNO	Mobile Network Operator
MoMI	Modeling of Mobile Internet Usage and Business
MSC	Mobile Switching Center
MSS	Maximum Segment Size
MVNO	Mobile Virtual Network Operator
NDA	Non-Disclosure Agreement
NOP	No Operation
NSAPI	Network Layer Service Access Point Identifier
NTP	Network Time Protocol
OS	Operating System
P2P	Peer-to-Peer
PC	Personal Computer
PCAP	Packet Capture
PDA	Personal Digital Assistant
PDN	Packet Data Network
PDP	Packet Data Protocol
PLMN	Public Land Mobile Network
POP3	Post Office Protocol version 3
PS	Packet Switched
P-TMSI	Packet-Temporary Mobile Subscriber Identity
RA	Routing Area
RADIUS	Remote Authentication Dial In User Service
RAT	Radio Access Technology
RNC	Radio Network Controller
RTSP	Real Time Streaming Protocol
RTT	Round Trip Time
SackOK	Selective Acknowledgment Permitted
SAE	Service Architecture Evolution
SGSN	Serving GPRS Support Node

SMTP	Simple Mail Transfer Protocol
SNDCP	SubNetwork Dependent Convergence Protocol
SP	Service Provider
SPSS	Statistical Package for the Social Sciences
TCP	Transmission Control Protocol
TEID	Tunnel Endpoint Identifier
TEKES	Finnish Funding Agency for Technology and Innovation
TKK	Helsinki University of Technology
TTL	Time To Live
UDP	User Datagram Protocol
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USB	Universal Serial Bus
USIM	User Subscriber Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
WAP	Wireless Application Protocol
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
VOIP	Voice over IP
VPN	Virtual Private Network

1 Introduction

1.1 Background

Fundamental transformations are taking place in the telecommunication domain as the Internet and mobile industries are converging. Increasing competition is forcing traditional players to rethink their strategies as new entrants are emerging and there are new products and services increasingly offered. On one hand, mobile phones are developing into multimedia computers including advanced features, functionality, and connectivity. On the other hand, computers are getting smaller and more portable, which is another reason why the distinction between different mobile devices is blurring. In addition, as mobile access technologies are developing, mobile internet access is both, substituting and complementing the traditional fixed Internet connectivity and thus increasing traffic volumes in mobile networks.

The increase in the usage of mobile data services provides companies an opportunity to look for new revenue sources and improve profitability. Industry players, including mobile operators, device manufacturers, application or content producers, as well as consulting companies, academia, and regulators, are all keen on understanding the mobile Internet usage patterns and market characteristics.

There are different ways to measure mobile Internet usage and different stakeholders have access to different data, as mobile Internet usage is dispersed between devices and networks (Smura et al., 2009). Mobile core network provides one centralized measurement point to provide objective information about the mobile Internet usage patterns. However, network traffic measurements are an emerging method and previously they have been conducted mainly for technical purposes. Such large scale measurements have been rarely used to uncover mobile service usage level information, or even basic mobile traffic characteristics (Dahmouni, 2005).

The focus of this thesis is to study network traffic measurements and their applicability to complement the traditional market research approaches. This thesis analyzes usage data that has been gathered from Finnish mobile operators in the MoMI[1] project and its predecessors[2] at Helsinki University of Technology (TKK) since 2005. In addition to providing empirical data on the Finnish mobile Internet usage characteristics, this thesis gives suggestions on how the existing measurements should be developed to provide enhanced understanding of the market. Furthermore, recommendations are given for mobile operators on how to implement network traffic measurements to support business intelligence and decision making.

1.2 Research questions & objectives

Although the usage of mobile Internet has been growing all over the world, little empirical data about the usage is available. Nevertheless, industry players are interested on such data due to the high value of the mobile Internet domain and the transformation that is forcing the players to adapt and look for new revenue sources. As mobile network traffic measurements have traditionally been conducted to research

[1] This thesis has been conducted in the MoMI project and Econ@Tel context. MoMI (Modeling of Mobile Internet Usage and Business) is a national research project (2008-2009) funded by the Finnish Funding Agency for Technology and Innovation (TEKES), Nokia, Elisa, DNA Finland, AinaCom, Yle, Accenture, MINTC, Sanoma, and TKK Comnet. Econ@Tel (Telecommunications Economics COST Network, COST605) is a European research network consisting of research units in 20 countries.

[2] Predecessors of the MoMI project were LEAD (Optimal Rules for a Leading Mobile Data Market) and COIN (Dynamics of Competition and Innovation in the Converging Internet and Mobile Networks) projects between 2004-2005 and 2006-2007, respectively.

technical issues, few generally known methods or tools are available to analyze network from a service usage and user behavior perspective. As the MoMI project has been conducting such measurements annually and the possibilities of the method have been found promising, further development alternatives needed to be classified and concrete recommendations provided.

The main research question of this thesis is:

Q_1: How should mobile network traffic measurement and analysis be designed and developed

a) to provide understanding on mobile Internet usage patterns and user behavior

b) to support a mobile operator's strategic decision making and business planning

The following research objectives were set in order to answer the research questions and to provide concrete objectives for the thesis:

O_1: Provide factual statistics of the mobile Internet usage in Finland

O_2: Classify the design and development areas of network traffic measurements

O_3: Provide recommendations for the development of IP traffic measurements conducted at mobile operator core network

a) From the MoMI project viewpoint
b) From a mobile operator viewpoint

1.3 Scope of the research

The term "mobile Internet" can be defined in multiple ways as mobile Internet usage has been dispersed between different devices and access network technologies. Smura et al. (2009) have classified mobile devices and networks for analyzing mobile service usage. Their categorization of mobile devices is based on the physical size of the device, the capability to make 2G/3G ($2^{nd}/3^{rd}$ Generation) voice calls, and the operating system (OS) of the device. According to these criteria, there are five types of mobile devices, namely mobile phones, smartphones and PDAs (Personal Digital Assistant), ultra-mobile PCs (Personal Computer), laptops and tablet PCs, and other devices such as iPods and digital cameras. Furthermore, Smura et al. (2009) divide wireless and mobile networks based on their geographical coverage. In addition to cellular based nationwide mobile networks, such as GSM/UMTS (Global System for Mobile communications/Universal Mobile Telecommunications System), also wireless access networks, such as WiMAX (Worldwide Interoperability for Microwave Access) and WLAN (Wireless Local Area Network), can be used to access the Internet.

In this thesis network access technologies are limited to mobile, i.e., cellular network technologies, as the centralized measurement point has been chosen from the 2G/3G core network. Thus, connectivity to the Internet via other wireless access networks, such as WLAN, is out of the scope of this thesis. However, in practice the same measurement setup used in this thesis could also be utilized with other wireless access networks. Mobile devices are not limited by any means, as long as they are mobile network capable devices, or using mobile network capable data terminals in connecting to the Internet. Thus, all devices from basic mobile phones to computers are in the scope of the thesis.

As Finnish mobile network operators were measured in this thesis, the scope is restricted to the Finnish mobile market and Finnish mobile subscribers. However, no geographical limitations exist, as also the roaming data of the measured operators' home subscribers is transferred via the home core network and the measurement point.

1.4 Research methods

First, a *literature review* is carried out to give an overview on the underlying technologies. In addition, previous research in the areas operating mobile network traffic measurements and IP traffic measurement analysis is reviewed.

Second, an *analysis of operating mobile network IP traffic measurements* is conducted to provide descriptive statistics on the mobile Internet usage in Finland***.*** The analysis includes multiple methods and tools, varying from statistical tools, such as SPSS (Statistical Package for the Social Sciences) and Excel, to proprietary Perl scripts.

Last, the literature review and experiences from the measurements are utilized in a *qualitative case analysis*, classifying the development and design areas for the measurements, as well as providing concrete recommendations. Two viewpoints are taken, one concentrating on the measurements conducted in the MoMI project and the other considering a mobile operator point of view.

1.5 Structure of the thesis

This thesis consists of seven chapters. The structure is presented in Figure 1.

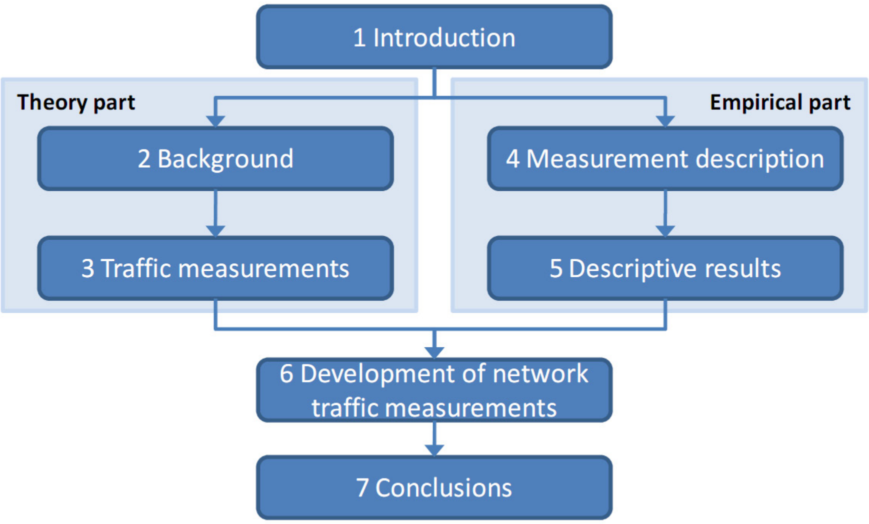

Figure 1 Structure of the thesis

Chapter 2 provides background information about mobile network architecture and the basics of data transmission in mobile networks and the Internet. It describes the mobile network entities, data transmission procedures, and the TCP/IP (Transmission Control Protocol/Internet Protocol) suite.

Chapter 3 provides information about IP traffic measurements in general, as well as in the context of mobile domain. A holistic comparison of different measurement analysis methods is also presented.

Chapter 4 includes the measurement organization and analysis description from fall 2008 and spring 2009, respectively. The chapter discusses the objectives, measurement setup, trace data, analysis methods, and the measurement and analysis processes.

Chapter 5 provides descriptive statistics about mobile internet usage in Finland. The results include information about general mobile internet traffic patterns, as well as more detailed application and mobile web browsing usage patterns.

Chapter 6 classifies the main development and design areas of IP traffic measurements used in operating mobile networks. In addition, concrete recommendations about future development are provided.

Chapter 7 provides a summary of the main findings. In addition, the results are discussed and further research suggestions are presented.

2 Background

IP based traffic has dominated the fixed Internet for basically its whole lifetime and it has also been adapted to the mobile Internet. On the other hand, mobile Internet access requires enhanced features from the access network. Thus, data transmission in mobile networks requires also proprietary protocols used with Internet traffic, and measurement traffic at different points provide also different trace data.

First, this chapter gives overview on IP traffic characteristics. Second, mobile network architecture is described from a mobile Internet point of view.

2.1 IP traffic

The objective of IP is to enable an interconnection between heterogeneous networks and their interoperation (Comer, 2000). IP is based on a layered model where multiple protocols handle different level of connectivity between two hosts.

2.1.1 Protocol layers

TCP/IP suite describes the different layers of Internet traffic. The layers from bottom-up are the physical layer, data-link layer, network layer, transport layer, and application layer, each including their own protocols that are used for the interconnectivity (Forouzan, 2002). The protocols in different layers encapsulate and decapsulate the traffic when it is being transmitted from a sender to a receiver. Encapsulation means that at the sender side a lower level protocol takes a higher level protocol packet and without changing it adds a standardized header of its own to the beginning. In decapsulation at the receiver end the headers at each layer are removed and the original application layer packet stays untouched. The TCP/IP suite and exemplary protocols are presented in Figure 2. The figure also demonstrates the encapsulation and decapsulation process.

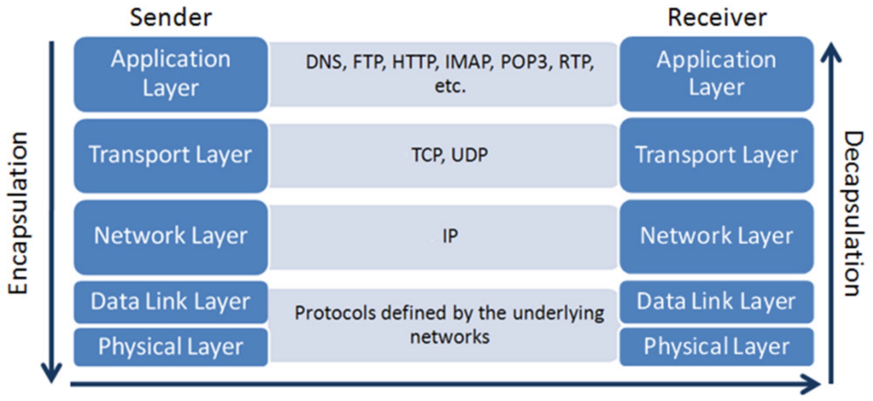

Figure 2 Internet model and encapsulation of traffic (adapted from Forouzan et al., 2002)

On the data-link and physical layers there are no specific protocols defined. These layers support the proprietary protocols of the underlying networks and take care of the link- and physical level connectivity.

The IP on the network layer handles the host-to-host traffic by mainly offering the addressing for it. It uses IP-addresses, which are fixed length of 32 bits, and transmits the data unreliably over one or several networks in IP-datagrams. Unreliability means that the protocol does not include error checking or tracking of the datagrams, for instance. As the maximum packet sizes in different network links may vary, the IP-datagram may have to be split into smaller packets, which is why IP protocol also offers fragmentation and reassembly (Postel, 1981a).

The transport layer on top of IP layer handles the delivery of messages from an application to another, whereas IP protocol handles the delivery of messages from a host to another. There are two main protocols on the transport layer, the User

Datagram Protocol (UDP) and the Transmission Control Protocol (TCP), from which the latter is more used in the Internet.

UDP is a connectionless protocol, which does not include any reliability, flow control, or error recovery, meaning that segments are sent to the network without any guarantees. UDP is a straightforward and light protocol, which is why its header is also small, carrying only the obligatory information, as a lot of responsibility is given to the applications. UDP and its header format are described in the RFC 768 (Postel, 1980).

TCP is a connection-oriented and reliable protocol. It identifies every connection uniquely, retransmits packets if needed, and has a flow control method. TCP header includes source and destination ports, as well as a checksum field like UDP does. In addition, there are other fields, such as sequence and acknowledgment numbers and different flags, which take care of the reliability and connectivity of the protocol. TCP protocol header description is provided in the RFC 793 (Postel, 1981b).

The highest layer of TCP/IP suite, called the application layer, is the most visible part for the end user. Its function is to enable users to access the network by providing user interfaces, support for services, and file access, for instance (Forouzan et al., 2002). The applications then access services across the Internet by invoking the transport layer protocol, which then handles the delivery from that on as described earlier. In comparison to the network layer with one main protocol (IP) and the transport layer with two main protocols (TCP and UDP), the application layer includes a variety of different protocols serving a variety of applications, which is why Internet traffic, in general, is rather heterogeneous.

Examples of widely used protocols on the application layer are the Hypertext Transfer Protocol (HTTP) used by browsers and web servers, and the Domain Name System (DNS) for mapping IP addresses and readable domain names. In addition, application layer protocols include mail protocols, e.g. SMTP (Simple Mail Transfer Protocol), POP3 (Post Office Protocol version 3), and IMAP (Internet Message Access Protocol), the File Transfer Protocol (FTP), and protocols involved with media streaming, for instance, RTSP (Real Time Streaming Protocol).

2.1.2 Structural levels of traffic

In addition to the different layers of traffic protocols, there are three principal structural levels of traffic in the Internet (Crovella & Krishnamurthy, 2006). From lowest to highest they are the packet level, train or flow level, and session level, as presented in Figure 3.

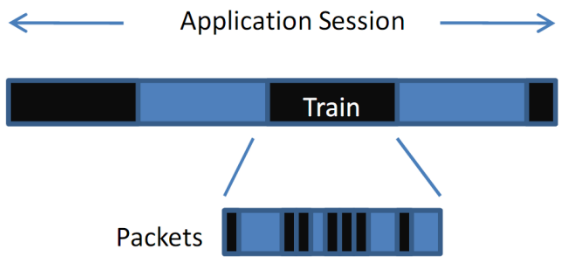

Figure 3 Levels of structure in Internet traffic (adapted from Crovella & Krishnamurthy, 2006)

The lowest level includes packets that are either transferred in the network or not. The second level includes trains of packets, often called as flows, as rather rarely only one packet is sent between a source and destination. A flow or a packet train can, for example, represent a download of one file or a web page. At its simplest a packet train or a flow is a burst of packets from a certain source arriving to a certain destination. If a timeout between two packets exceeds a certain interval, the packets belong to a different packet train or flow (Jain & Routhier, 1986). One of the most popular ways of defining a flow is a five-tuple, which includes source and destination IP addresses, source and destination port numbers, and the protocol number (Ilvesmäki et al., 1998).

Thus all the packets having the same characteristics of the five-tuple and arriving inside a certain interval belong into the same flow. One popular time-out in studies recently has been 60 seconds (see, e.g., Claffy, 1995; Zhang, 2002), but also longer or shorter timeouts ranging from a few seconds to half an hour have been used. In addition to the unidirectional flow defined above, sometimes a more appropriate way with client-server traffic is to use bidirectional flow, in which traffic to both ways between the two hosts is included into a single flow.

The highest structural level of IP traffic is the session level (Crovella & Krishnamurthy, 2006), which describes a single execution of an application by including a set of flows. Often traffic in the Internet has a clear beginning and an end that correspond with the usage time that a human is spending with the application. However, the increased P2P (peer-to-peer) traffic, for instance, has changed this behavior during the last years, as many P2P applications are used also while the end user himself or herself is inactive. In these cases the application session times last longer and do not represent a user session anymore.

2.1.3 Web traffic

Web content is one of the major traffic generators in the Internet, including mobile Internet. Web traffic can include variety of content, such as web pages, pictures, video, web applications (e.g., java), and so on. HTTP traffic is based on a client-server relationship where a client, usually a web browser, initiates (HTTP) requests that the web server responds to, offering different resources for the client. These web resources can be static files, such as HTML (Hypertext Markup Language) pages, or dynamic content resources. Dynamic content can be, for instance, a live web camera gateway, an online web store, or an internet search engine. Due to the variety of possible data transferred over HTTP the type of the content is tagged in a data format label called

MIME (Multipurpose Internet Mail Extensions), originally used with electronic mail systems, hence the name. (Gourley & Totty, 2002)

The client can make a request to a certain web resource by using URIs (uniform resource identifier) which point out the exact domain name (server), and the location of the resource in the server. The most common URI is a Uniform Resource Locator (URL), consisting of three parts, the scheme (e.g., http://), the server internet address (e.g., www.example.fi), and the resource location (e.g., /pictures/example.jpg). This exemplary URL for a picture resource would thus be http://www.example.fi/pics/pic.jpg. A more detailed URI description can be seen in Berners-Lee et al. (1998).

Nowadays it is common that requesting a single web page involves multiple transactions with multiple servers. An exemplary HTTP request of a web page can include only the frame and text of the web page from the requested server and the client may have to fetch pictures included in the page from content distribution network (CDN) servers, and advertisement banners from advertisement servers. Furthermore, these servers can be chosen dynamically through load balancing or DNS lookups. (Crovella & Krishnamurthy, 2006) For example, Smith et al. (2001) studied web traffic and noticed that 35% of web pages included resources from multiple servers. In addition, they noticed that 70% of consecutive top level page references go to a different server than the previous reference, due to load balancing.

Redirection is nowadays used in HTTP with device identification adapting content according to the device. Every HTTP response includes a status code which can also be a redirect code for the client to look for the resource from a given alternative address. For instance, mobile handsets can be identified from their user agent field in the HTTP request and redirected to a mobile optimized web page with the equivalent content. An example would be a mobile handset trying to get a web page from http://www.example.fi and being redirected to a mobile optimized site http://m.example.fi. In addition, sometimes redirecting can be conducted on other levels also, such as DNS, TCP, or even lower levels (Crovella & Krishnamurthy, 2006). This makes web traffic complex of nature, as requesting one resource or a web page can include data from multiple servers and via multiple redirections. One feature of the newer HTTP version 1.1, called pipelining, enables more simple traffic on a transportation level. It allows one TCP connection to be used for multiple HTTP requests and responses, thus reducing the amount of TCP connections needed in HTTP transactions.

2.2 Mobile network

A mobile network (PLMN, Public Land Mobile Network) offers telecommunication services to the public and consists of multiple logical and physical parts and interfaces. There are two main logical domains in a mobile network, a core network and an access network. (3GPP TS 23.002) The core network consists of a packet switched domain, a circuit switched domain, and an IP Multimedia Subsystem (IMS) concentrating on multimedia specific IP services, such as video, audio, and chat. However, as this chapter describes the basics of mobile networks, IMS is not discussed further. The rest of the network entities are presented in Figure 4.

2.2.1 Mobile device & subscriber

A subscriber has a physical (mobile) device with variety of functionalities to take a radio connection to the access network. The device can be a mobile handset or a laptop, for instance. In UMTS terms the device is called a User Equipment (UE), which consists of a Mobile Equipment (ME) and a User Subscriber Identity Module (USIM). The mobile equipment, which can be further divided into subcategories, includes applications and handles the radio transmission, for instance. (3GPP TS 23.002, 3GPP TS 23.101).

The USIM, often referred in spoken language as a SIM card, is a possibly removable smart card that can be used in different kind of equipments. The SIM card allows the subscriber to be securely identified independent of the equipment, by providing data and procedures for it. In more detail, a number called an International Mobile Subscriber Identification (IMSI) number is used in the subscriber identification procedure. For subscriber identity confidentiality purposes also temporary subscriber identification methods are used in the mobile network, such as P-TMSI (Packet-Temporary Mobile Subscriber Identity) in the packet switched domain. In mobile networks also the user equipment can be uniquely identified from the International Mobile station Equipment Identity (IMEI) or the International Mobile station Equipment Identity and Software Version number (IMEISV), identifying the manufacturer, the model, and uniquely the device. More detailed description of the subscriber and device identifiers in mobile networks is provided in 3GPP (3rd Generation Partnership Project) specifications (see 3GPP TS 23.003)

2.2.2 Access network

Via the access network a mobile device may take connection to both, circuit switched and packet switched core networks one at a time or both simultaneously. There are two possible access networks used widely in operational mobile networks at the moment in Western Europe, a GSM EDGE (Enhanced Data rates for Global Evolution) Radio Access Networks (GERAN) and an UMTS Terrestrial Radio Access Network (UTRAN), both of which can be deployed in parallel in practice. In addition, High speed packet data access (HSPA) can be implemented into UMTS, including enhanced protocols to provide faster data transfer rates.

2.2.3 Core network

The core network consists of two main domains, packet switched (PS) and circuit switched (CS). They share some entities handling general information such as mobility management and subscriber information, as well as have certain domain specific entities. There are three main interfaces in the packet switched core network; one to access network, one to other PLMN's, and one to external packet data (or other) networks. External packet data networks are usually either private corporate networks or public networks, i.e., the Internet. The logical architecture of GPRS (General Packet Radio Service)/UMTS networks is presented in Figure 4, including the main interfaces and network entities.

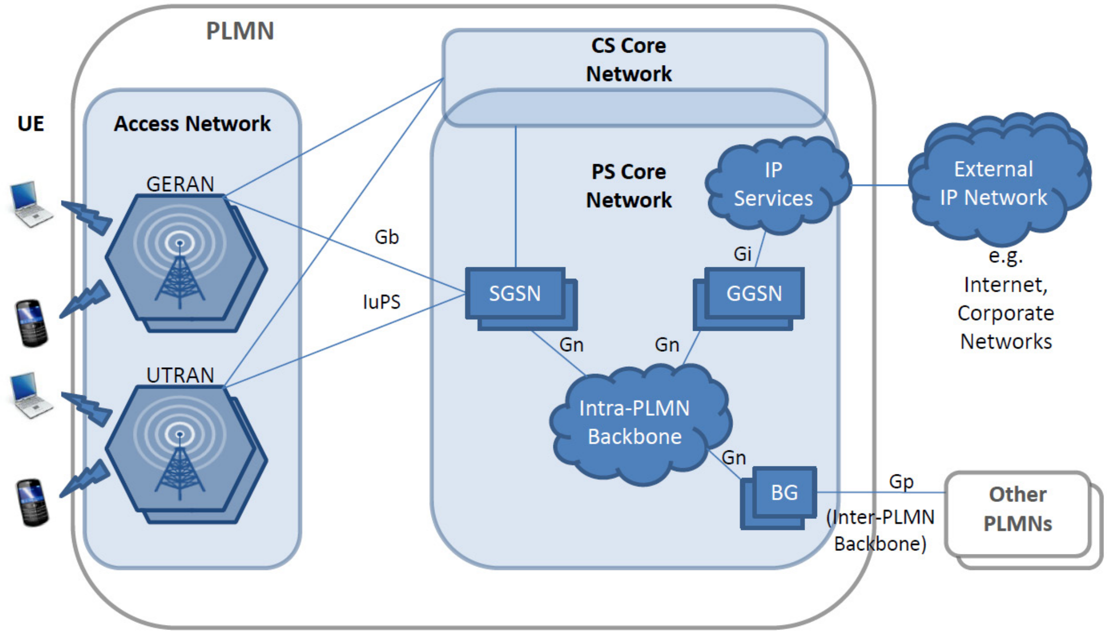

Figure 4 GPRS/UMTS architecture with the main interfaces (adapted from Kivi, 2006)

The packet switched core network consists mainly of GPRS Support nodes (GSN), including Serving GPRS Support Nodes (SGSN) and Gateway GPRS Support Nodes

(GGSN), which handle the packet transmissions between the main interfaces. The GSNs are interconnected via a GPRS backbone network usually based on IP.

As a central part of the packet switched core network, SGSN has multiple tasks. It takes care of routing traffic, session and location management, as well as includes authentication and charging functions, for instance. It is connected to the access network via *Gb* (GERAN) or *IuPS* (UTRAN) interfaces, forwarding traffic from a mobile device to the right GGSN (*Gn* interface) through the core network backbone (3GPP TS 23.060). Thus, information that SGSN needs to store include both, subscriber data, including subscription information such as IMSI and PDP (Packet Data Protocol) addresses (usually IP), as well as location information, for example, routing area information and addresses of GGSNs with active PDP contexts. (3GPP TS 23.002)

GGSN's main function is to provide interworking with external PS networks. Also GGSN stores both, subscriber and location data, including IMSI, possible PDP addresses, and the SGSN addresses where UEs are registered. GGSN and SGSN functions can be in separate physical entities, or it is also possible to integrate them into a single physical entity. (3GPP TS 23.060) In reality there is often a many-to-many relationship between SGSNs and GGSNs, meaning that a SGSN may provide service to many GGSNs and one GGSN can be in contact with many SGSN's to route traffic to all the geographically dispersed mobile stations. (3GPP TS 29.060)

In addition, there is a third component called a border gateway, which is an entity connecting PLMNs in the GPRS/UMTS architecture. In practice, border gateways (BG) are SGSNs or GGSN, and the *Gp* interface includes the same functionalities as the *Gn*, except that some security needed for inter-PLMN transmission is added (3GPP TS 23.060).

2.2.4 Data transmission in mobile networks

Data transmission via a mobile network to outside IP networks, such as the Internet, is more complex than in fixed Internet access networks where, for instance, no session or mobility management functions exist. Even though mobile devices can nowadays use the TCP/IP protocol suite, the process of transmitting the data to external networks requires also protocols specified by 3GPP. The basics of data transmission in mobile networks are described next.

First, before data transmission can take place a mobile device has to make it available by doing a GPRS attach to a SGSN, which checks that the user, and possibly the device, is authorized and assigns a P-TMSI for the user. After this, the mobile device applies for a PDP address, which is allocated by a GGSN chosen by the SGSN. After the creation of the PDP context, the mobile subscriber is visible to external networks via the PDP address. (Bettstetter, 1999) It is possible for a user to have multiple simultaneous PDP contexts active, meaning that every application can have its own PDP context attached to a different GGSN. This is why there are two kinds of PDP contexts possible; primary and secondary. The primary PDP context has a unique IP address and other context information, whereas secondary PDP contexts adopt the IP address and context information from the primary PDP context. Simplified GPRS attach and PDP context activation procedures are presented in Figure 5.

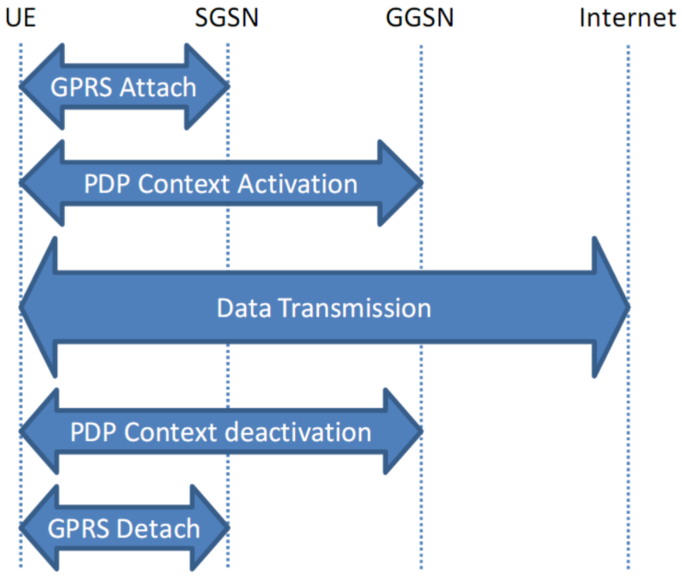

Figure 5 Simplified GPRS attach & PDP context creation procedures (adapted from Svoboda, 2008)

Inside the GPRS/UMTS backbone, between the GSNs (*Gn* interface), traffic is tunneled with a GPRS Tunneling Protocol (GTP). In practice it means that the traffic is encapsulated with a GTP header between SGSNs and GGSNs. Thus the backbone network is transparent to the user, as the tunnel is decapsulated in the GGSN after which the traffic continues as plain IP traffic. GTP is transmitted on top of UDP and is used with both, control (GTP-C) and user traffic (GTP-U). Control plane provides functionality to create, delete, and modify GTP tunnels, whereas user plane only transfers user data and takes care of routing it. (Bannister et al., 2004)

From a subscriber viewpoint, an Access Point Name (APN) is used in a PDP context as a reference to a GGSN, meaning that a user does not have to care about which SGSN and GGSN to use. The APN selection is handled by DNS functionality in the core network, which returns the IP address of a correct GGSN for the SGSN, which handles a certain geographical area. The APN is in a readable format including two parts, an APN Network Identifier and an APN Operator Identifier. They identify the external network that the GGSN is connected to, and in which PLMN GPRS backbone the GGSN is located in, respectively. From these the latter is optional. (3GPP TS 23.003) Thus, often Internet APN is in readable form, such as “internet” or “internet.operator”, providing a simple access point name for the subscriber to be used. Also other access points, such as WAP (Wireless Application Protocol), act in a similar manner and only the access point name differs for the end user in connecting to operator on-portal services.

3 Internet traffic measurements

Internet traffic measurements are conducted for a variety of reasons and objectives. They are often performed to characterize, to monitor, or to control the network. For a network operator, monitoring and controlling the network are essential parts of its operation and maintenance. On the other hand, longer time-scale traffic characterization is often related to network research activities. (Peuhkuri, 2003) These characterizations can bring both, technical and commercial information on how to run the network and what kind of traffic is transferred in the network. User behavior and usage characteristics interest many entities, including academic, regulative, as well as corporate entities, but these issues have not been studied academically as much as the more technical aspects.

First in this chapter, different measurement methods are classified. Second, different user, device, and application identification approaches are discussed. Third, traffic measurement research in the mobile context is reviewed and measurement points in mobile networks are discussed. In addition, extractable information from each measurement point is described.

3.1 Measurement types

The measurement type or setup is mainly affected by the measurement objectives. As traffic measurements can be conducted in multiple parts of a network, the measured trace data has to be representative for the phenomenon that is studied. Thus, the choices of the measurement location and granularity of the data are important.

Location describes the part of the network and the traffic that is measured. In general, measurements can be conducted at different nodes or links in the network and in single or multiple locations. On the other hand, granularity of the data, i.e., the level of detail, defines the metrics that can be used. (Peuhkuri, 2003) If Internet usage characteristics are measured, utilization metrics, such as byte count and application distribution, can be used. However, as traffic volume as such, for example, is not always the best metric to describe usage, new metrics may have to be created to match the objectives better. Time scale of a measurement means both, the scale of the measurement period, as well as the continuity of it. Some measurements are sample based by nature, whereas sometimes sampling is conducted to save resources, for instance. However, Smith et al. (2001) argue that web traffic measurements, for example, have to be at least hours long to capture fully the long tail of distribution.

The measurement hardware needs to be physically attached to the point of measurement. Viipuri (2004) describes two common ways to connect measurement hardware to the network. The first method is network tapping, which includes a passive method of splitting either an optical or a copper link. Another method is port mirroring, in which one or more ports in a switch are mirrored to another port connected to the measuring device. This method requires certain functionality from the switch, which is the method’s weakness, whereas network tapping is applicable to almost all kinds of networks.

3.1.1 Active & passive measurements

Another measurement categorization can be done between active and passive measurements. In active measurements there is always some excess traffic sent to the network and analyzing the response. Examples of metrics analyzed in active measurements are latency, jitter, and packet loss.

However, as unobtrusiveness is often one objective of measurements, active measurements would bias the results. Another way is to conduct passive measurements, in which no additional traffic is sent to the network. Passive methods are usually preferred and active or hybrid solutions are only used when passive measurements are not applicable. An example is a case where not enough traffic streams are present (Viipuri, 2004)

3.1.2 Trace data & post-processing

The measurement analysis can be done both, online or offline (Williamson, 2001), affecting whether the result of the measurement is trace data for further analysis, or a direct report of the real time analysis. As the latter requires high processing power, the traffic is often recorded on-location and the analysis is made afterwards. From the measured packets different information can be recorded. In general, the measurement can record protocol headers, protocol payload, certain fields in the headers or payload, or simply all the data.

Packet level data can be provided by, for instance, network and transport level header measurements. One reason for recording only headers is the huge amounts of data traffic in high speed network links. However, from high bandwidth links, even the headers produce high amounts of trace data. Thus, one way of post-processing packet level traces is to aggregate them to a flow level to compress the amount of data. There are multiple free tools to do the aggregation. For instance, a tool called CoralReef[3] takes *de facto* pcap-format (Packet Capture) packet trace data as input and outputs five-tuple flows with a wanted timeout and an interval period.

[3] CoralReef: http://www.caida.org/tools/measurement/coralreef/ [Accessed 29 May 2009]

If application level traces are needed, also the transport protocol payload, in other words the application level data, is measured. As application level protocol headers may be of highly varying lengths, either only certain amount of the application level data is recorded from the beginning, i.e. the headers, or all data is saved. If real time processing is conducted, only the significant fields can be saved for later use, decreasing the amount of trace data.

3.2 Identification of users & devices

There are only few ways to identify unique users and devices directly from IP traffic. User identification can be done directly from IP addresses if static IP address pooling is used. However, as dynamic IP address allocation is popular in mobile networks nowadays, this method cannot be always relied on. In mobile networks, where user authentication is conducted in the core network, the measurement point can be selected from a part of the network where user identification information is available. On the other hand, these measurements are often no longer plain IP traffic measurements as there are mobile network specific protocols in use, meaning that proprietary tools have to be developed for the purpose. Nevertheless, there are solutions to identify the operating system (OS) of a mobile device from plain IP network traffic, which can give indicative information about the device. The method is rather simple, yet some trade-offs have to be made as single user level identification accuracy is not achieved.

3.2.1 Operating system identification

If unique user identification or device identification is not possible, an alternative is to identify the OS by TCP fingerprinting, which can be conducted directly from TCP and IP headers. Another possibility would be to use HTTP protocol user agent field that provides information on the used browser, operating system, and possibly even device. However, this method is bound to HTTP traffic only, creating issues with other protocols.

The principle of OS fingerprinting is that every operating system has its own idiosyncrasies in the TCP/IP stack that enable identifying them from each other. Thus the traces obtained by measurements can be analyzed by comparing certain IP and TCP headers to previously known signatures of different operating systems. Smith & Grundl (2002) have listed four and Peikari & Chuvakin (2004) eight TCP areas that can be inspected in TCP fingerprinting for operating system determination. The common fields in both these lists are the initial TTL (Time to Live), the Window Size, and the "Don't fragment" flag. In addition, Peikari & Chuvakin (2004) mention the Maximum Segment Size (MSS), the Initial Packet size, as well as the No Operation (NOP), Selective Acknowledgment Permitted (SackOK), and Window Scaling options. Operating systems tend to use default values for these fields and the combination of all these fields tend to be unique with each operating system.

There exist several active and passive tools to do OS fingerprinting, e.g. Ettercap (passive), p0f (passive), and Nmap (active). According to Zalewski (p0f, 2006), p0f was the first application to start using a SYN[4]-based passive fingerprinting technique and provides rather accurate results.

However, there are some limitations with TCP fingerprinting. The manipulation of the different fields is rather easy and can be done manually by the end-user, by an application, or by a network node such as a gateway or a firewall. However, manual field

[4] SYN is a control flag in a TCP header which is used in establishing TCP connections (Postel, 1981b)

manipulation is not by any means presumable by an average end user, as it requires detailed knowledge on how to do it. In addition, there has to be a database of known fingerprints to compare the network traffic with. The process of updating the fingerprints requires time and resources, as well as an access to the latest devices and operating systems.

3.3 Application identification

There are different ways of doing application identification from traffic measurements. This chapter introduces three different methods that have been studied and used in past literature.

3.3.1 Port based method

A simple method in application identification is to use well known protocol port numbers. An advantage of identifying applications by port numbers is that the process is straightforward and easy to implement. Many well known applications use registered ports, for example, HTTP (port 80) and POP3 (port 110), which are publicly listed by the Internet Assigned Numbers Authority (IANA)[5]. The port numbers are divided into three ranges. The well known ports include the range 0-1023, the registered ports the range 1024-49151, and the rest of the ports (49152-65535) are private ports not assigned to any specific applications. Port based application identification can be done by proprietary scripts, as well as with commercial or freeware tools, for instance, CoralReef. The CoralReef API (Application Programming Interface) consists of methods to manage application rules, which are given as input, and to search them for matching applications to the protocol and source and destination ports.

[5] IANA Port Numbers. http://www.iana.org/assignments/port-numbers [Accessed 27 February 2009]

Recently the use of the port based method has been decreasing. One reason for this are the uncertainties that arise when trying to get application level information from the transport level traces (Smith et al., 2001). Many emerging applications use dynamic ports and some applications may try to masquerade as web traffic, because firewalls sometimes limit other application protocols except HTTP (port 80). In addition, network address translators and gateways can hide the original port information. Thus, the method is at its best when identifying conventional applications. Some authors accuse the port based application identification method to be too inaccurate with the emerging applications and its use alone in application identification Internet traffic has been questioned (see, e.g., Moore & Papagiannaki, 2005; Kim et al., 2005). To increase the accuracy of the port based method additional port information can be gathered. However, this includes often a lot of manual work as reliable references are hard to find.

3.3.2 Payload inspection

One alternative for port based identification is to inspect the TCP/UDP payload, i.e., application level data. These methods are also called Deep Packet Inspection (DPI). They include finding signatures from the application level and matching them to signatures of previously identified applications. In practice, they can be hybrid solutions using also transport level information. For instance, Moore & Papagiannaki (2005) were able to increase the accuracy of application identification with the payload method in comparison to a plain port based method. In addition, Sen et al. (2004) created a DPI classification approach based on individual packets and noticed significant improvement in P2P connection identification.

Payload inspection introduces also multiple issues. Moore & Papagiannaki (2005) argue that often almost all the payload has to be studied to be able to find the correct application. Payload inspection is resource-intensive, meaning it scales poorly to increasing bandwidths, and does not work on encrypted traffic. Furthermore, problems on legal and privacy side arise as the application level data may include sensitive information. In addition, the classifier needs to know the payload syntax of all traffic it identifies. As new applications emerge all the time the need of manual work for the syntaxes is increasing (Nguyen & Armitage, 2008). For the problem of increased manual work, some automatic application signature solutions have been developed (Haffner et al., 2005). However, many of the fully automated methods have been able to identify only conventional applications (Kim et al., 2008). Thus a trade-off between the accuracy and available resources has to be made.

3.3.3 Advanced statistical and machine learning methods

Lately also other advanced ways of identifying applications, relying neither on plain port numbers nor application payload, have been developed. These methods often use statistical patterns of easily observable attributes from the transport or network levels, such as packet lengths, inter-packet arrival times, etc. (Nguyen & Armitage, 2008). The amount of recent research on these areas is increasing and some authors have compared and categorized alternative approaches.

Kim et al. (2008) make a distinction between two general traffic classification approaches that use different features in the classification, namely host-behavior based and flow features based methods. Host-behavior based methods use the “social interaction” information of hosts to define the applications, including also the ports. One application of this method is BLINC (Blind Classification). BLINC inspects multiple flows at a time during a certain interval and studies the host to host behavior during that interval (Karagiannis et al., 2005). Thus the measurement point for this method is important, as it needs all traffic between two hosts. BLINC includes a substantial downside of misclassifying rarely transferred flow types in the network as unknown.

This error needs to be corrected by, for instance, using some flow-level machine learning method (flow features based method) for the unknown part of the traffic. Thus, in practice two methods have to be implemented when using BLINC. In addition, Kim et al. (2007) have accused the parameter tuning in BLINC to be too cumbersome.

Flow-features based methods use information such as duration, number and size of packets per flow, as well as packet inter-arrival times utilizing machine learning. The amount of features that are used in flow-based classification can vary from a few (Crotti et al., 2007) to hundreds (248 features identified in Moore & Zuev, 2005). The simplest methods can be basically enhanced port based identification with additional rules, whereas the complex ones require high processing and memory capabilities. There are some software suites available that have been used with machine learning methods, such as WEKA[6].

Nguyen & Armitage (2008) divide machine learning methods into two general groups; unsupervised (also called clustering) and supervised learning. The supervised machine learning methods are more capable of classifying earlier identified applications, whereas unsupervised can also be used in finding new applications.

The unsupervised approaches cluster flows with similar characteristics together. Thus no training for the classifier is needed. An advantage of unsupervised methods is the automatic discovery of classes through the recognition of clusters in the data. However, these clusters have to be somehow manually labeled afterwards. In addition, the clusters may not map one to one with the applications classes wanted. Thus the amount of clusters can be larger than application classes, for instance. On the other hand,

[6] Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 27 May 2009]

unsupervised methods can identify also emerging applications with beforehand unknown traffic characteristics.

In supervised learning methods a pre-labeled training data exists, and a model is produced to fit the training data. Kim et al. (2007) have classified supervised learning methods into five approaches based on the used model. Examples are Bayesian (e.g., Bayesian Networks) and function (e.g. Support Vector machine) based methods. In supervised learning the labeling of the training data is classified manually in advance, which requires time and resources. On the other hand, these supervised algorithms can then be tuned to detect also subtle differences between flows that match the applications or application groups, which is uncontrollable with unsupervised methods. In addition, supervised methods can label a flow right after its termination, which is not possible with the clustering methods that need the whole dataset (Kim et al., 2008). According to Nguyen & Armitage (2008), using Bayesian trained neural networks in supervised learning an accuracy of up to 99% is possible.

Machine learning techniques with IP network traffic application identification have some general open questions. It is not known how well they perform with behavior such as packet loss, fragmentation, and jitter. In addition, they are not an absolute answer with newer applications like P2P and video streaming that are harder to identify. Kim et al. (2008) have summarized general challenges of machine learning methods into three. First, it is hard to find the key features that capture the fundamental characteristics of different applications. Second challenge is the trade-off between accuracy of the method and the computational cost, as the methods tend to get complex. Third, finding datasets with representative data to provide the fundamental characteristics of applications is difficult.

3.3.4 Other advanced methods

In addition to the approaches presented above, also other advanced approaches exist, often combining different methods. One recent combinatorial approach presented by Canini et al. (2009) is called GTVS (Ground Truth Verfication System) and has multiple iteration rounds with different methods. The identification is made on application level information, flow statistics, host-level connection statistics, host name information, and transport-layer behavior level, thus including basically all the aforementioned methods.

Canini et al. (2009) have used their application identification method GTSV also for HTTP traffic classification, including 14 categories that are presented in Table 1. Motivation for this is the expanded role of HTTP protocol in other than plain web browsing use. The advantage of this method is the added information about web traffic, which often is interpreted as web browsing as analysis methods do not have the functionality to do more accurate identification. Different classes are identified with combinations of different HTTP header fields, such as user agent, host name, and content type. An interesting detail about the research by Canini et al. (2009) is that they have planned to publish the tool and the rule sets to be used by other researchers in the future.

Table 1 Web traffic classification categories (adapted from Canini et al., 2009)

Class	Activities
Web Browsing	visiting web pages using a web browser
Web App	applications via web interface: e.g. Java applets, web gadgets
Crawler	bots crawling web pages
File Download	file downloading over HTTP
Webmail	web based e-mail services
Advertising	advertisement on a web page or embedded in software
Multimedia	streaming media or viewing media files on web pages
SW update	software update over HTTP
News Feeds	RSS feeds
Link Validator	automated link validators
Calendar	calendar application based on web, e.g., ical, gcal
Attack	malicious traffic over HTTP
IM	Windows Live Messenger
Monitoring	network monitoring

In addition to general application identification, also more specific research on certain application categories has been conducted, concentrating rather heavily on P2P applications. For instance, Svoboda et al. (2009) have used a cross layer method in VoIP (Voice over IP) classification, Karagiannis et al. (2004) used transport level information and the first 16 bytes from the payload to identify P2P applications, and Sen et al. (2004) used an application level P2P classifier to identify five common P2P applications.

3.3.5 Comparison of the application identification methods

As the research especially on flow-features based methods is rather active at the moment and lots of publications are available, comparative studies have also been conducted. For instance, Kim et al. (2008) and Nguyen et al. (2008) have listed and compared different flow features based methods. In addition, Kim (2007) has done accuracy comparison. The four application identification methods compared in this chapter can be divided into two generic groups, header measurement based and payload based. The latter group includes payload inspection and the former port based, host behavior based and flow-features based methods. Table 2 lists important properties of each method and is based on comparison by Nguyen et al. (2008), which was conducted between the port based, host-behavior based, and flow-features based methods.

Table 2 Comparison of application identification methods (adapted from Nguyen et al., 2008)

	Network and Transport Layer			Application Layer
	Port Based Method	Host-behavior Based Methods (BLINC)*	Flow-Features Based Methods	Payload Inspection
Key Features	Protocol & protocol port	Behavior patterns of hosts & ports	E.g. ports, protocols, TCP flags, packet size	E.g., HTTP header fields
Input Data	Packet header or flow data	Flow data & payload bytes per flow	Training data with the key features	Packet headers & payload
Rules Buildup	Manual	Manual + parameter tuning	Automatic	Manual
Configuration complexity	Low	High	Case dependant	Medium/High
Advantages	Easy to implement & fast Good with conventional protocols	Detects new & encrypted data Good with P2P flows	Highly accurate Automatic learning	Application level header inspection possible Web traffic subclassification possible
Challenges / Limitations	Problems with dynamic port use Problems with traffic masquerading	High computational requirements Manual parameter tuning	High computational requirements Finding right features Obtaining representative training data	High computational requirements Encrypted traffic
Legal Issues	Low	Low	Low	High

* BLINC classifier introduced in Karagiannis et al. (2005)

Port based method is the simplest method to implement and is still an alternative, especially with conventional applications. The most work is needed for the rule buildup, which has to be done manually due to lack of trustworthy references. On the other hand, manual rule buildup is also needed with the other methods, except the flow-feature based methods using machine learning. Nevertheless, automatic rule buildup in machine learning requires training data and possibly multiple training rounds. Payload inspection suffers from the headers of different sizes and the multitude of protocols on the application layer, if the signature matching is not conducted in real time. In addition, one of the biggest challenges relating the approach is the sensitivity issue when operational networks are measured. Nevertheless, there are cases when no other methods are applicable than payload inspection, such as when sub-classification of web traffic is needed.

None of the methods can be said to be better than the others in all aspects. Machine learning methods have been argued to reach an accuracy of about 99% (Ngyen & Armitage, 2008) and also full payload inspection enables very accurate results (Moore & Papagiannaki, 2005) but requires extensive manual work. Thus, the choice of a method is always a trade-off between accuracy, simplicity, computational power, technical knowledge, and resources available. In addition, the measurement setup and the trace data may limit or enable the methods that can be used.

3.4 Traffic measurements in mobile networks

There has been a variety of different operating mobile network traffic measurements that have been published. In addition, mobile operators conduct their private measurements, but no information of these is usually publicly available. However, as most of the existing literature has been focusing on the technical properties of mobile networks, the results and analysis methods are rarely comparable to our usage perspective. Thus, this chapter compares the different measurement setups in the literature and the descriptive statistics are not described in detail. In addition, the extractable information from different measurement points is discussed.

3.4.1 *Measurement points and tools*

There are different measurement point alternatives in mobile networks all providing different trace data. There are three main measurable interfaces in the mobile core network where user traffic is transmitted, in addition to the control traffic links and nodes. The interfaces are presented in Figure 6.

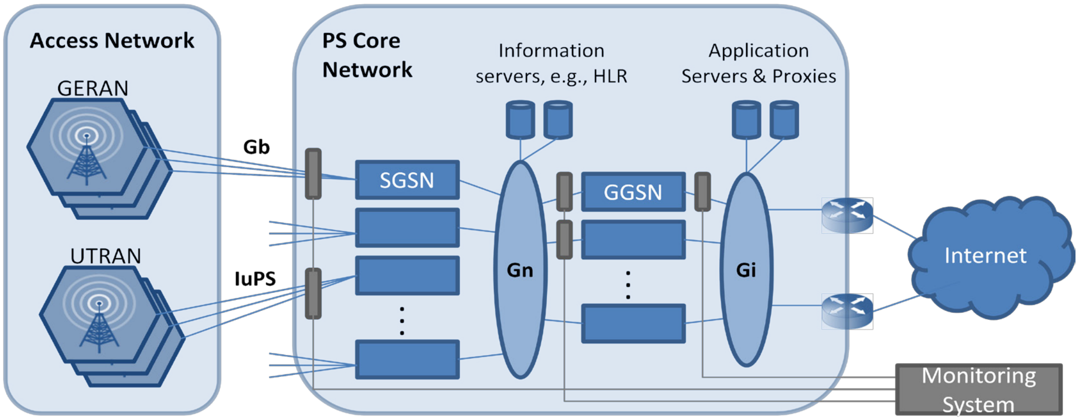

Figure 6 Main measurement interfaces in a mobile network (adapted from Ricciato et al., 2006)

One difference between the interfaces is that in the *Gi* interface traffic has been decapsulated to plain IP traffic and thus the measurements can be conducted by using *de facto* measurement tools developed for Internet traffic measurements. The other interfaces, *Gn* and *Gb/IuPS,* have protocol stacks specified by 3GPP.

Different research projects have developed their proprietary measurement and analysis tools for mobile network traffic measurements. For instance, Darwin and Metawin – projects in Austria have created a Linux based tool that can measure all the main

GPRS/UMTS mobile core network interfaces[7], including *Gn*, *Gi*, *IuPS/Gb*, and *Gs.* The system is able to store full traces with also the payload. Measurements have been conducted in the 3G network of Mobilkom Austria. The measurement system has been lately commercialized and is provided by Kapsch Carrier named as DataXTender (Svoboda, 2008). Many of the publications have been concentrating on traffic traces from the *Gn* interface, studying, for instance, traffic composition (Svoboda et al., 2006; Svoboda, 2008), capacity bottleneck detection (Ricciato et al., 2005; Svoboda et al., 2007a), email traffic modeling (Svoboda et al., 2007b), analyzing TCP spurious timeout detection (Vacirca et al., 2006), and to do TCP RTT (Round Trip Time) analysis (Romirer-Maierhof et al., 2009). Traffic from both, *Gn* and *Gi* interfaces, has been lately utilized by Svoboda (2008), which studied user related composition of GPRS/UMTS traffic, application flow patterns, and different application level traffic models. Romirer-Maierhof et al. (2008) utilized measurements from *Gn*, *Gi*, and also the *Gb* interface when studying delays of GPRS/UMTS networks.

In a joint research project (MeaDow) by Ericsson and Vodafone measurements on two interfaces, *Gi* and *Gb,* were conducted in three European Vodafone networks (see, e.g., Kalden, 2004). In the *Gb* measurements a proprietary tool called *Moniq*[8] was used. The measurement was actually conducted at the SGSN node itself, not on the link traffic. The reporting tool was integrated into the operating system of the SGSN and the measurement setup enabled the capturing of also SGSN internal information on, for instance, PDP contexts. The same tool was also used later in the anonymization and analysis of the traces. The *de facto* TCPDUMP tool was used for measuring the *Gi*

[7] Metawin monitoring system: http://userver.ftw.at/~ricciato/darwin/#metawin [Accessed 25 May 2009]

[8] Developed by Ericsson Hungary and Ericsson global services business unit

interface. The interface was measured for two traffic types, the basic user IP traffic and the control traffic between GGSN and a RADIUS[9] server. Traffic traces from the *Gi* interface by this monitoring system has been used for GPRS traffic composition (Kalden & Ekström, 2004), to investigate self-similarity properties (Kalden & Ibrahim, 2004). The traces from both, *Gi* interface and the SGSN node (*Gb*) measurements have been utilized and mapped to reveal application usage by session duration, packet size distribution, data volumes, user mobility, and application usage frequency (Kalden et al., 2003), and modeling application flow lengths (Kalden & Haverkort, 2006). These areas have been discussed also in Kalden (2004).

In addition to the aforementioned projects, single point *Gi* measurements have been conducted by several authors in several other projects. For instance, Kilpi (2003) studied GPRS sessions, Kilpi & Lassila (2005) RTT variability, and the MoMI project has been studying mobile Internet usage properties (see, e.g., Kivi, 2006; , Kivi, 2007a; Kivi 2008; Riikonen & Kivi, 2009), all of these in Finland. Furthermore, Benko (2004) has measured passively several Asian and European countries to study TCP end-to-end connections. Madsen (2004) analyzed *Gi* traces for flow and application composition and to investigate user sessions directly from IP traffic.

3.4.2 Extractable information from different interfaces

The nodes between the main mobile network interfaces have different functions. Different protocols are used in the interfaces and thus different information is extractable from different measurement points. As mobile networks usually have a tree-

[9] RADIUS (Remote Authentication Dial In User Service) can be used to provide centralized IP address allocation for the GGSNs, for instance, in addition to other authentication, authorization, and accounting services. RADIUS traffic is sent on top of UDP.

form structure[10], the representativeness of a single measurement point increases when going to an interface closer to the last node between the core network and the Internet. On the other hand, interfaces closer to the access network include detailed control plane traffic, and possibly enable more advanced analyses.

The *Gi* interface between GGSN and the Internet includes user plane IP traffic. However, if the IP address allocation for mobile devices is conducted separately from GGSN, it is possible to record this signaling traffic from the *Gi* interface, providing session and some level of authentication information. In general, from the user traffic in the *Gi* interface the authentication information is limited to IP address and operating system identification (by TCP fingerprinting) of the mobile device. Thus, the main metric to be used is traffic volume. By using a timeout, traffic could also be divided into sessions and analysis conducted on a single session level. However, if dynamic IP address allocation is used in the network, no interrelation between two consecutive sessions of a single user can be analyzed.

Traffic in the *Gn* interface is GTP traffic, including both control plane and encapsulated user plane IP traffic. As PDP context creation is conducted in this interface between SGSN and GGSN, some session level information is accessible. In addition, users are authenticated in GTP traffic uniquely. However, the identification in the link traffic is conducted via TEID (Tunnel Endpoint Identifier), which has a one-on-one relationship with IMSI and NSAPI (Network Layer Service Access Point Identifier), but has only local meaning and the computation is implementation dependent. Thus, extracting IMSI information out of the TEID from each GTP packet could require complex algorithms or

[10] The amount of network entities usually decreases from the access network entities to SGSNs, and from SGSNs to GGSNs

measuring the node directly. In newer UMTS release specifications RAT (Radio Access Technology) type flag has been introduced for GTP, meaning that also the access network type may be identifiable from GTP traffic (Svoboda, 2008). The location information of the subscriber is restricted to Routing Area (RA) level meaning that location identification is rather limited. In addition, both, RAT and RA fields are only conditional, meaning they are partly implementation specific.

The interfaces between the access network and SGSN, *IuPS* interface in UMTS and *Gb* in GPRS, differ from some parts. On the control plane both, GPRS and UMTS use GMM/SM (GPRS Mobility Management and Session Management), but the underlying protocols differ. On the user plane UMTS uses GTP-U also in the IuPS interface, whereas GPRS relies on SNDCP (SubNetwork Dependent Convergence Protocol). By measuring the *Gb/IuPS* interfaces location information can be extracted either on cell (GPRS) or routing area (UMTS) level. On the other hand, unique user identification based on IMSI and P-TMSI is possible from the link traffic from the interfaces. However, as temporary identifiers are used, stateful tracking of the identifiers is needed. On the other hand, the *Gb* interface may provide more accurate subscriber location and session information than the other interfaces with user plane traffic. It is also optional in *Gb* and *IuPS* interfaces to do IMEI check, which could reveal unique device identification. (3GPP TS 23.060)

There are also other possible interfaces to measure from the mobile network not presented in Figure 6. However, these interfaces, such as the *Gf* interface between SGSN and HLR/AuC (Home Location Register/Authentication Center; entities of the GSM architecture), include only control plane traffic. These could be measured mainly for subscriber related data to add value for the measurements by providing additional information, or by providing unique user or device identification.

3.5 Issues with traffic measurements

Multiple challenges arise in network traffic measurements. The same issues are valid also with mobile network measurements, in addition to some mobile domain specific problems. The challenges include more than just technical aspects, such as legal and socio-political issues involving multiple stakeholders. These stakeholders are the (mobile) network operators, the infrastructure manufacturers, end-users, and naturally researchers, among others (Clegg et al., 2008).

3.5.1 Operational issues

There are several operational issues involved in network traffic measurements. First of all, cooperation is needed between different parties including communication and interactions thus causing possible delays. As the measurements are conducted in operational networks, the availability of the service for the end users is always the first priority. The service has to be guaranteed for the subscribers, allowing no breaks due to measurements. In addition, mobile networks include proprietary protocols of the infrastructure manufacturer which may require operator or measurement specific preparation, which are not needed with standardized solutions. As the traffic in the Internet increases all the time also the data storage requirements rise at the same time. In addition, increasing link speeds require increased processing power. (Clegg et al., 2008)

The issue of clock synchronization arises only with multiple simultaneous measurements in different points of the network. The accuracy of real-time clocks in computers is not good, possibly including an error of several seconds per day (Peuhkuri, 2003). In addition, Network Time Protocol (NTP) is not suitable in short term accuracy and due to the possibility of delay variations, as the traffic using NTP is transferred in an IP network

(Paxson, 1997). Peuhkuri (2003) suggests GPS (Global Positioning System) to be used as an out-of-band timing solution for accurate clock synchronization, as even low cost applications can reach an accuracy of ±1 µs. In addition, GPS has been utilized with mobile network traffic measurements and multiple simultaneous measurement points (see, e.g., Svoboda, 2008).

3.5.1 Privacy

Privacy and anonymity are relevant issues that need to be well considered when handling data from operational networks, as the traffic can include sensitive information about the end user, such as passwords and other identification or personal information. This issue is especially relevant in mobile networks where subscribers are uniquely authenticated. In addition, although the user data would not be recorded, also the plain communication related data can provide sensitive information about the end user, for instance, identification information. There are two main legislative issues involved with traffic measurements, concerning interception of communications and data protection.

Clegg et al. (2008) lists the different aspects that affect the legal positioning of a traffic measurement. These include the purpose of the measurement, ownership of the data and its storage location, possible anonymization, the source and nature of the data that is captured, and the form that it is stored in.

The Finnish legislation, in short, says that if a message is not intended for public, it is private (Peuhkuri, 2003). However, it is not always unambiguously defined which laws are applied and the interpretation is case-specific. There are two laws closely related to traffic measurements in Finland. Act on the protection of Privacy in Electronic Communications (516/2004) and the Personal Data Act (523/1999) are involved with identification information and personal data, respectively. In short, the former specifies that identification information can be used for the implementation, use, and technical development of products and services, as well as billing purposes, marketing, and statistical analysis with the user’s consent. However, for example, from statistical research no unique persons can be identified. On the other hand, Personal Data Act, which had the latest amendments in 2009, specifies that personal data can be used for special purposes without the need to ask for acceptance, such as, historical and scientific research or statistical purposes, direct marketing or similar addressed shipments, as well as market research. However, user has the right to unambiguously prohibit the usage of his or her personal data.

In TCP/IP suite protocols most header fields are insensitive. From IP, the addresses of the source and the destination are sensitive due to possible user identification. As many mobile networks use dynamic IP address allocation nowadays, also the sensitivity issues regarding it are rather low. On the other hand, transport level protocols TCP and UDP include port number information, which in very small networks can be sensitive as they reveal information about the application. (Peuhkuri, 2003) However, operational mobile networks are of a size that does not have this problem. Thus, the most sensitivity in IP protocols relate to the application level protocols.

4 Measurement description

This chapter introduces the measurement organization and specifies the analysis methods that were used in the measurements.

4.1 Measurement organization & specification

The measurement setup in 2008 was adopted from the previous research made at TKK. In the MoMi project and its predecessors, packet data traffic measurements have been conducted during four consecutive years, 2005-2008. The earlier measurement specifications and results have been documented by Kivi (2006; 2007; 2008). The data from previous years was also used in this thesis to enable longitudinal results.

The main objective of the measurement was to provide objective statistics on the Finnish mobile Internet usage. In more detail, tasks included general traffic characterization by traffic direction and transport protocol, application protocol distribution, and the most popular web sites visited by mobile handsets. Terminal device operating systems were also identified, and this information was used in the other objectives to provide comparison between handheld and larger devices, such as laptops. In addition, trace data timestamps were utilized in providing distributions of general and application usage according to the time of a day and the weekday.

The measurements were conducted at the national GSM/UMTS networks of two Finnish Mobile Network Operators (MNO) Elisa and DNA Finland. There are three MNOs in Finland, which means that the scope of the measurements included roughly 40-50% of the Finnish mobile subscribers. The point of measurement was adapted from previous years (see, e.g., Kivi, 2006) to be at the Internet APN, i.e., at the *Gi* interface (Figure 7).

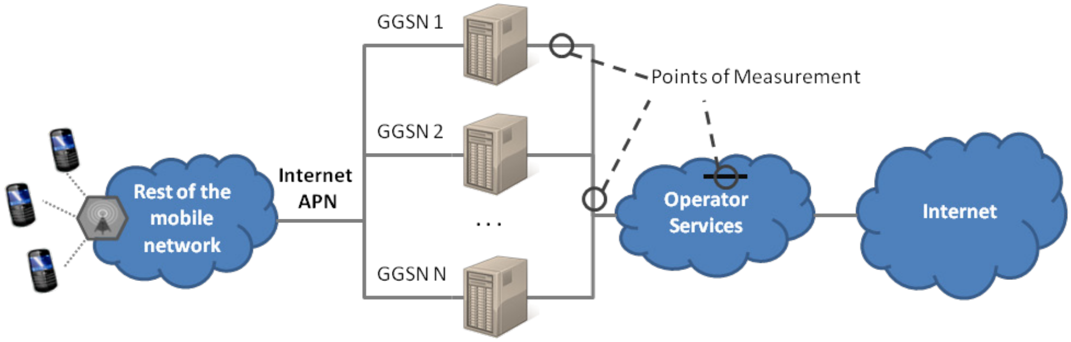

Figure 7 Measurement point (adapted from Riikonen & Kivi, 2009)

All traffic going via the measurement point was captured and no sampling was used. However, the measurement point does not include all the traffic going through Internet APN as the traffic load is divided between multiple GGSNs. However, if an assumption is made that the GGSN's are equally distributed geographically and traffic load-wise, the captured data is a representative subset of the total traffic in the network (Kalden, 2004). As over 90% of all data traffic (via all APNs) goes via Internet APN (Kivi, 2008), and if the Finnish subscriber generated traffic is assumed to be equally distributed between the operators, the measurements at the two MNO's represent the Finnish mobile Internet usage.

As both of the measured operators used home GGSN roaming during the measurements, the home subscribers' roaming data was included in the captured traffic. Likewise, foreign roamers' traffic generated in Finland was not included in the measurements.

The measurement hardware included Endace DAG (Data Acquisition and Generation) cards, which were connected to the network by an unobtrusive method. In practice, the measurement hardware was sent to the operators' premises, where the physical attachment was done by the operators' staff. The measurement software consisted of proprietary measurement software developed at the Department of Communications and Networking at TKK. The proprietary capturing tool utilizes the UNIX pcap library libpcap. The same library is used by the tool TCPDUMP that has been used in previous years.

From every packet going through the point of measurement the network and transport level headers were captured. Exceptions were DNS request and responses, from which also application level information, such as requested domain name and returned IP addresses were captured. In the case of DNS requests and their responses on-the-fly processing was done, in addition to the basic recording of the requests. The domain names and the corresponding IP addresses were retained in memory and every eight hours the domain names that had had over 40 requests were saved to the hard disk. The reason for this are privacy issues as by leaving the domain names with less than 40 requests out, individual users cannot be identified from the web browsing information.

The packet trace data was provided in the pcap format. In addition, DNS requests and responses were provided in a format including domain names, timestamps, and the matching IP addresses.

4.2 Measurement process

The implementation of technical research methods to provide market and usage related understanding has been one purpose of the techno-economical perspective of the MoMI project. On a holistic level, the network traffic measurements conducted in the MoMI project follow traditional market research phases rather well. Thus, an indicative framework from market research literature has been adapted in describing the measurement process in this thesis. The same framework is used also later in classifying and analyzing the possible development areas of mobile network traffic measurements.

Aaker et al. (2007) have divided a market research process into seven generic phases:

1) *Agree on the research process,*
2) *Research objectives,*
3) *Estimate the value of information,*
4) *Research design,*
5) *Data collection,*
6) *Data analysi*s, and
7) *Reporting.*

On a more general level, the first three phases are called as the preliminary planning in this thesis, research design as the planning, and the last three phases as the implementation. The first phase, *Agree on the research process*, is not discussed in more detail in the process description, as it has been part of a bigger entity, the agreement of the MoMI project. Neither is the third phase, *Estimate the value of the information*, included in the process description, as the measurements have been conducted annually and the value of the research has been noticed to overcome the costs of it.

Figure 8 represents a simplified Gantt-chart of the MoMI measurement process with an indicative time span. Some additional tasks have been added to the process description in the figure as specific for the measurement process used in the MoMI project. The data collection phase is divided into two separate data collection tasks and the supporting subtasks, relating to server administration and physical transferring of the measurement system to operator premises and back after the measurements (Transf. HW).

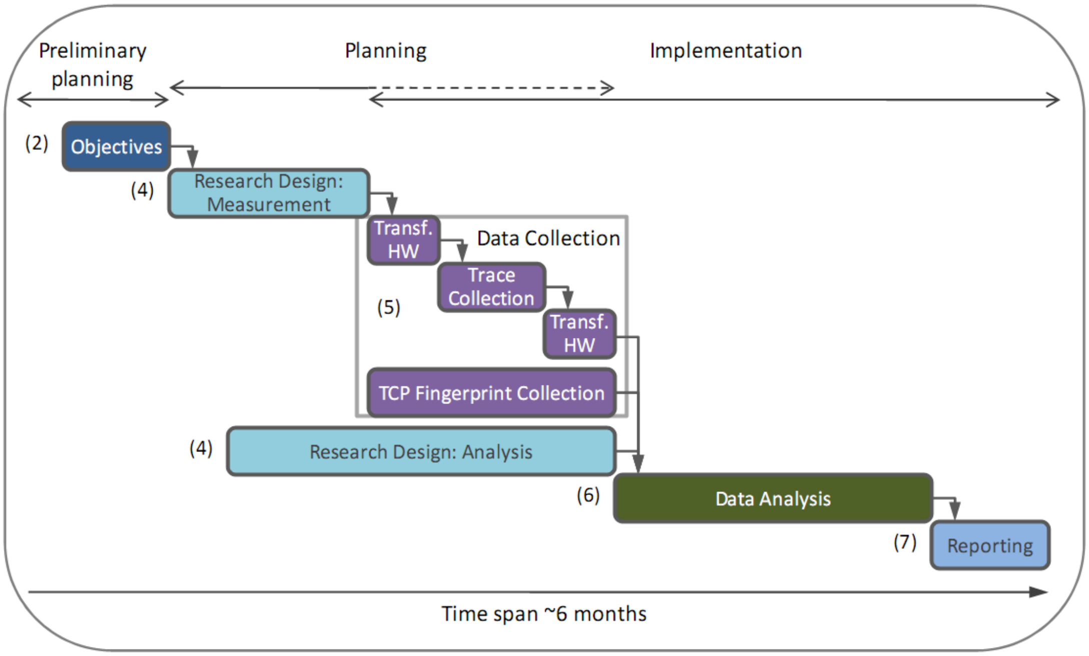

Figure 8 Measurement process

There are multiple stakeholders involved in the measurement process, mainly from either TKK or mobile operators. In addition, some external consulting, such as legal advising, or agreements with other project affiliates is necessary in the first phases of the process. The main TKK stakeholders are the researchers and a technical measurement administrator (later called the measurement admin) who is a the MoMI specific technical specialist involved in all measurement related tasks at TKK. From the operator side, the affiliates are divided into two, business affiliates (project coordination) and network administration affiliates, who are usually both involved in this sort of a market research process.

4.2.1 Phases of the measurement process

The first phase in Figure 8 is the (Research) *Objectives* phase. The phase includes mainly the formation of the research questions and boundaries of the study (Aaker et al., 2007). Involvement is needed mostly from the non-technical affiliates, including researchers, operator business side, and other project affiliates in agreeing the research questions. In addition, legal counseling may be needed if objectives are to change the measurement setup. The phase is rather short, but as communication between multiple parties is involved, delays are possible.

The first task of the *Research design* phase (Research Design: Measurement) is related to the measurement organization and specification planning. It demands attention from the more technically oriented affiliates, including the measurement admin., operator network admin., and the researchers. Some legal counseling may be needed also in this phase as, for instance, the trace data format is agreed. As the measurements in the MoMI project have been conducted annually and only minor changes have been made to the measurement setup, also this phase has been rather short, mainly consisting of updating the measurement hardware and software. At the same time the measurement planning and the data collection are conducted, researchers can begin to design the data analysis phase (Research Design: Analysis) and choose or create the tools that are used. The time span of this task depends on the amount of script and tool development that needs to be conducted. If existing tools can be used, the phase can be rather short. In the MoMI project the amount of trace data has been large, which has required TKK developed scripts to be used for the analysis and annual code optimization.

There have been two different tasks in the *Data collection* phase; the mobile network trace data collection and the TCP fingerprint collection. In addition, as the physical trace data collection location is at operator premises, time has to be reserved for the physical transferring of the monitoring system (Transf. HW). The trace data collection is rather straightforward, including mainly the physical tapping of the monitoring software to the network by the operator network staff. During the measurement period few resources are needed, as even the start and stop of the measurement can be automated with a timer (Kivi, 2006).

The second task in *Data collection* phase, the TCP Fingerprint collection, includes setting up an HTTP server (described in chapter 4.2.2) where the operating system fingerprints are collected to. In addition to the participating end-users providing the fingerprints, involvement is needed from the researchers, to manage the collection, and the measurement admin, to manage the server.

The *Data analysis* phase (described in chapter 4.3) cannot begin before the data collection and the analysis planning have been conducted, making it the most critical transformation in the process. If any of the aforementioned phases are late the entire process is delayed. This phase is mostly conducted individually by the TKK researchers and includes running the scripts and tools in the creation of the results and their evaluation.

Reporting in the MoMI project measurements has been conducted via technical reports and presentations at TKK and affiliate premises. The phase requires mainly researcher resources as reporting of the results has been conducted manually.

4.2.2 TCP fingerprint collection

The software used for the TCP fingerprinting (p0f) has not been updated officially since 2006. For this reason the fingerprint database has been updated annually in the MoMI project to include the latest fingerprints. The TCP fingerprint measurement organization includes a server that records participants’ HTTP user agents and the needed TCP/IP protocol header fields for the TCP fingerprint. In addition, the participants have also been asked to fill a few text fields to provide information about the device and the operating system for confirmation.

The participants for the process were contacted and requested directly by email and through social networks. As Kivi (2009a) has identified, almost 90% of mobile handsets in the Finnish mobile market in 2008 were manufactured by Nokia[11]. On an operating system level, almost 70% of mobile handsets were either Symbian or Nokia S40 based devices. For these reasons the fingerprint collection focused on Symbian S60 (only 1% of the devices have other Symbian based OSs: S80 or UIQ) and S40 devices. Fingerprints for all the newer Symbian S60 editions were recorded. In addition, some other major mobile operating systems were measured. However, for an unknown reason all the Nokia Series 40 fingerprints ended up as invalid. Thus, devices with Nokia Series 40 OS could not be identified in the measurement analysis.

4.3 Data analysis

This chapter describes the *Data analysis* phase of the measurements in more detail. Post-processing and analysis include multiple phases, as there are different forms of trace data used in the analysis, the packet level trace data and the TCP fingerprint data. The data analysis process of 2008 is presented in Figure 9.

[11] Nokia: http://www.nokia.com/ [Accessed 28 July 2009]

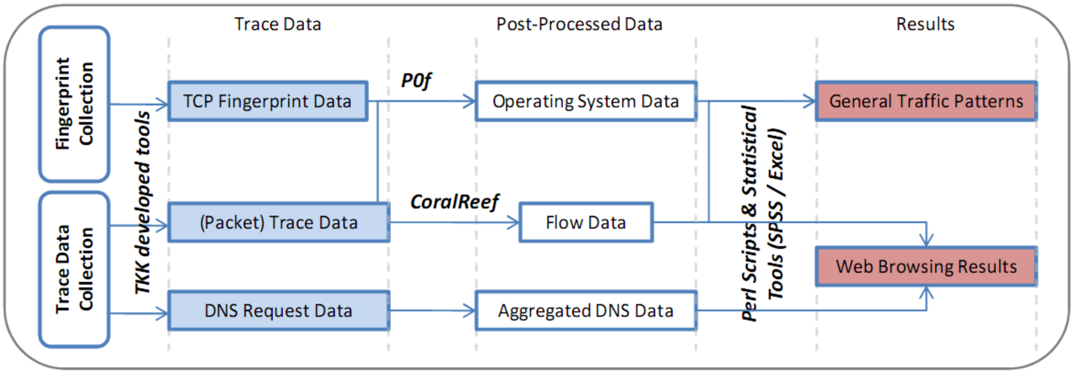

Figure 9 Data analysis phases and tools

Measurement software tools providing the trace data are developed at TKK. In the analysis, the p0f tool uses the fingerprint database and the packet level trace data in the identification of operating systems for each TCP SYN packet. In addition, packet level trace data is aggregated to flow level by using CoralReef and the DNS request data is aggregated from the request level to a domain name level by proprietary tools.

After the post-processing of the data, a combination of TKK developed Perl scripts and statistical tools, such as SPSS and Microsoft Excel, are used in generating the end results. These include general traffic patterns, such as operating system and application distributions, and the most popular web browsing sites. The generation of web browsing results also utilizes operating system and application identification.

4.3.1 Preservation of data

As measurements in operational networks may include sensitive data, the trace data, post-processed data and some the results have to be protected accordingly. In the MoMI project the trace data is preserved at TKK servers that are in a distinct internal network. The network is accessible only from a computer that has restricted access only to the involved researchers. At the point in the analysis when statistical tools are used, the post-processed data is transferred to the hard disks of the researchers' computers. All the data and results are stored on encrypted partitions. In addition, researchers have signed non-disclosure agreements (NDA) with the project affiliates. Thus, all the data is well protected and not accessible by others than the researchers involved in the analysis.

4.3.2 Classification of operating systems

The traffic was divided into different categories based on the device type provided by the TCP fingerprinting. The reason for making a distinction between operating systems was to see if there are differences in mobile Internet usage based on the device type. The chosen categories were Computer, Handset, and Others/Unidentified. The category Computer included operating systems Windows, Mac OS, and all identified BSD (Berkeley Software Distribution) variants. The Handset category included Symbian OS based traffic, as other identified mobile handset operating systems generated insignificant levels of traffic. Nevertheless, as the exclusive distributor of iPhone, Sonera, was not measured in 2008, no data about iPhone usage was available.

4.3.3 Application identification

Application identification was conducted with a port based identification method based on TKK developed Perl scripts. More precisely, applications were identified from server side TCP or UDP ports on a flow level. Applications were classified into five categories, namely Web, Email, P2P, Other, and Unidentified. The category Web includes all traffic on HTTP port 80 and HTTPS (Hypertext Transfer Protocol Secure) port 443. Email includes several ports used by protocols SMTP, POP3, and IMAP. Also the category P2P includes multiple ports. Examples are Bittorrent ports 6881-6889 and dc++ port 412. In addition, the Other category includes all other identified traffic, such as VPN (Virtual

Private Network), file transfer, control, and news traffic. More exemplary ports are provided in Table 3.

Table 3 Exemplary Ports Used in Application Identification

Application category	Major transport protocol ports included	
Web	TCP	80 (HTTP), 443 (HTTPS)
Email	TCP	e.g., 25 (SMTP), 110 (POP3), 143 (IMAP), 993 (IMAP/SSL), 995 (POP/SSL)
P2P	TCP	e.g., 411-412, 1214, 1412, 4661-4662, 6881-6889, 7777
Other *(e.g. VPN, WAP, file transfer)*	TCP	e.g., 20 (FTP), 119 (NNTP), 9200-9207 (WAP)
	UDP	e.g., 53 (DNS), 4500 (VPN), 1812 (RADIUS)

4.3.4 Mobile handset web traffic identification

First, mobile handset generated web traffic was identified by using the application identification method and operating system identification. Web domain identification was conducted on a server, i.e., IP address level. Web traffic was mapped to domain names by combining the server IP addresses to the domain names provided by the aggregated DNS data. If a certain IP address could be mapped with multiple domain names, manual inspection was conducted.

Web traffic analysis was divided into three; HTTP traffic web browsing analysis, HTTPS "secure web" analysis, and mobile optimized web browsing analysis. The HTTP analysis provided the mobile handset web browsing top list of the most popular web sites. Manually identified "non-browsing" sites were filtered from the list, as this analysis method would otherwise include also traffic to "web infrastructure", as discussed in Kivi (2007d). In the HTTPS analysis focus was to identify the most popular mobile Internet web sites or services using encrypted traffic. Last, only web domains beginning with "*.m*" or "*wap.*", or ending with "*.mobi*" were included into the mobile optimized web browsing top list.

5 Descriptive statistics on mobile Internet usage

This chapter presents the statistics and findings from the analysis of the measurements. A summary of the statistics has also been presented in Riikonen & Kivi (2009). First, a general overview of the Finnish mobile market is given, providing statistics from also external sources. Second, general network traffic patterns are described. Then the application and operating system distributions and diurnal usage patterns are presented. Last, the most popular web browsing sites are listed.

5.1 Overview on the Finnish mobile market

The Finnish mobile market has been known as progressive and especially the traditional voice services have been highly adopted in Finland. There were over 6.8 million mobile subscriptions in Finland in July-December 2008 (Ficora, 2009), corresponding to a subscriptions per capita ratio of 130%.

Finland is a rather exceptional country in the sense that the mobile handset market is dominated by almost one player only. According to Kivi (2008), about 90% of mobile handsets used in Finland are manufactured by Nokia. The share of Nokia devices from the whole Finnish handset population has been decreasing, though with a marginal pace. From the whole mobile handset population, one fourth has 3G connectivity, and almost 80% has packet data connectivity of some kind, meaning that mobile Internet is becoming more and more available to mobile subscribers. One explanation for the increasing popularity of 3G capable handsets could be the bundling of mobile handsets and mobile broadband subscriptions, which was permitted for 3G handsets in Finland in 2005 (Tallberg et al., 2007).

The market shares of the three Finnish MNOs were 38% (TeliaSonera)[12], 37% (Elisa), and 23% (DNA Finland) in the end of 2008 (Ficora, 2009). The remaining 2% market share is divided between a few mobile virtual network operators (MVNO) and Service Providers (SP). In addition, there is a fourth regional MNO, Åland Mobiltelefon Ab, which owns GSM- and UMTS-licenses to operate networks in Åland, an autonomous island group belonging to Finland.

There are three main pricing schemes for mobile Internet in Finland, usage based pricing, block based pricing, and flat rate pricing. In addition, one operator (Sonera) offers a hybrid pricing instead of usage based pricing. It is an hourly flat rate pricing, which includes unlimited data use for an hour from the first packet transferred. Furthermore, Sonera offers a daily flat rate in a similar manner. The mobile broadband subscriptions in Finland include unlimited traffic, but the data transfer rate can be limited. Ficora (2008) has been studying European telecommunication pricing annually and in comparison of mobile broadband subscriptions Finland was ranked as the cheapest in 2008[13].

5.1.1 Total traffic volume in the Finnish mobile networks

Mobile Internet usage in Finland has been increasing in a slow pace until the most recent years. This view is supported by the network traffic statistics by our measurements; the increase in total data traffic volume in the Finnish mobile networks was close to linear before the second half of 2007. After this the growth has been

[12] TeliaSonera operates with the brand name Sonera in Finland

[13] Ficora's definition of a mobile broadband included 1) nominal data transfer rate of minimum 384 kbps, 2) flat rate pricing, and 3) unlimited traffic. Thus only 13 countries were included in the study, as many European operators have certain traffic limitations in their mobile broadband subscriptions

exponential, approximated as close to 900% between falls 2007 and 2008 (Kivi, 2009c). These statistics can be seen in Figure 10 and a more detailed discussion is provided in Kivi (2009c).

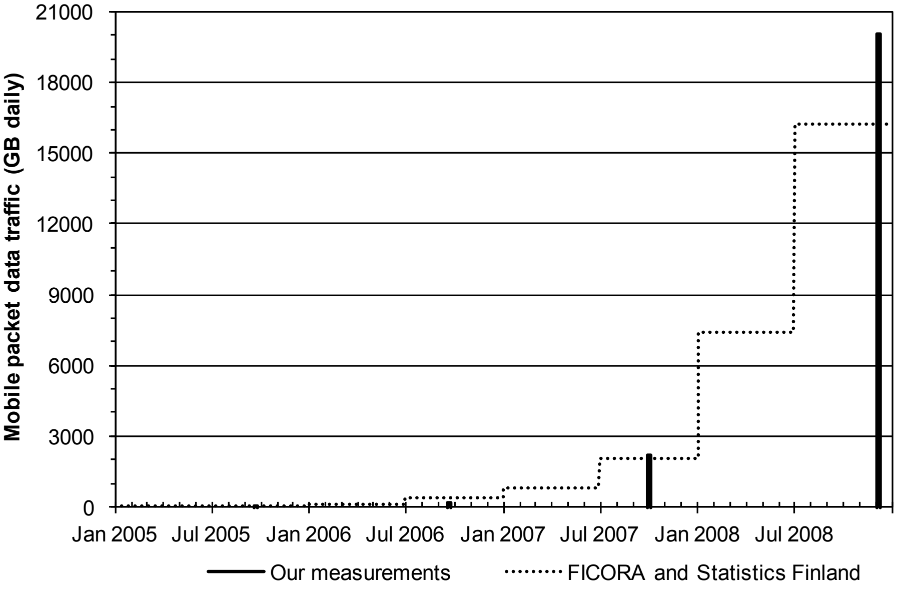

Figure 10 Total traffic volumes in mobile networks 2005-2008 (Kivi, 2009c)

The fast exponential increase in the network traffic can be explained with a few market changes between 2006 and 2008. The bundling of mobile subscriptions and 3G handsets was allowed in Finland from April 2006 onwards. At the same time operators conducted HSDPA (High-Speed Downlink Packet Access) upgrades to their networks and as of that they have been increasing the coverage of their networks. In addition to offering better transmission speeds for mobile handset users, aggressive marketing of bundled 3G/HSDPA-capable data cards and flat rate data subscriptions for laptops were introduced in 2007. First, the bundle was offered with USB (Universal Serial Bus) modems targeted for laptop use and mini-laptops with embedded 3G/HSDPA modules were added to the offered portfolio in 2008. The amount of mobile broadband subscriptions has also been growing fast since the second half of 2007. (Kivi, 2009) The number of them multiplied in 2008; from less than 150 000 to almost half a million. This resulted in mobile broadband subscriptions having a share of about 23% of all broadband connections in Finland in the end of 2008 (Ficora, 2009). One explanation for this is also the pricing; cheapest mobile broadband subscriptions have also been the cheapest of all broadband connectivity available (Kivi, 2009).

The rest of the descriptive results are based on relative byte shares. Due to sensitivity reasons, no statistics are provided about the absolute traffic volumes.

5.2 General traffic patterns

Traffic in the Finnish mobile networks is mainly towards mobile devices, i.e., downlink traffic. Roughly 75% of all traffic volume was downlink traffic in our measurement in the end of 2008. The corresponding numbers from previous years have been 84% (2005), 73% (2006), and 63% (2007). In mobile networks there is more capacity provided towards downlink, so the traffic is more downlink oriented by default. In addition, the adoption of HSDPA to the UMTS has happened before the adoption of HSUPA (High-Speed Uplink Packet Access) in Finland, increasing the gap between downlink and uplink capacities[14].

On a protocol level, mobile network traffic is dominated by TCP with a share of almost 95% of all traffic volume. The rest of the traffic is mainly UDP (< 5%), and other protocols, such as ESP (Encapsulated Security Payload) protocol, which generate less

[14] HSDPA can provide a capacity of up to 14 Mbps and HSUPA 5.8 Mbps in the uplink. http://www.3gpp.org/HSPA [Accessed 18 July 2009]

than 0.5% of the traffic volume. Other protocols than TCP and UDP were excluded from the rest of the analysis phases.

5.2.1 Traffic by operating system

The distribution of the Finnish mobile Internet traffic between different operating systems is provided in Figure 11. The most striking result is that computers generate over 98% of the total traffic in mobile networks. Handsets generate only 0.6% of the traffic and the rest of the operating systems and unidentified traffic generate 0.9%. Windows alone generates about 93% of the total traffic volume. On the other hand, Symbian based operating systems (S60, S80, UIQ) are the dominant handsets of Finnish mobile Internet usage generating majority of the Handset traffic.

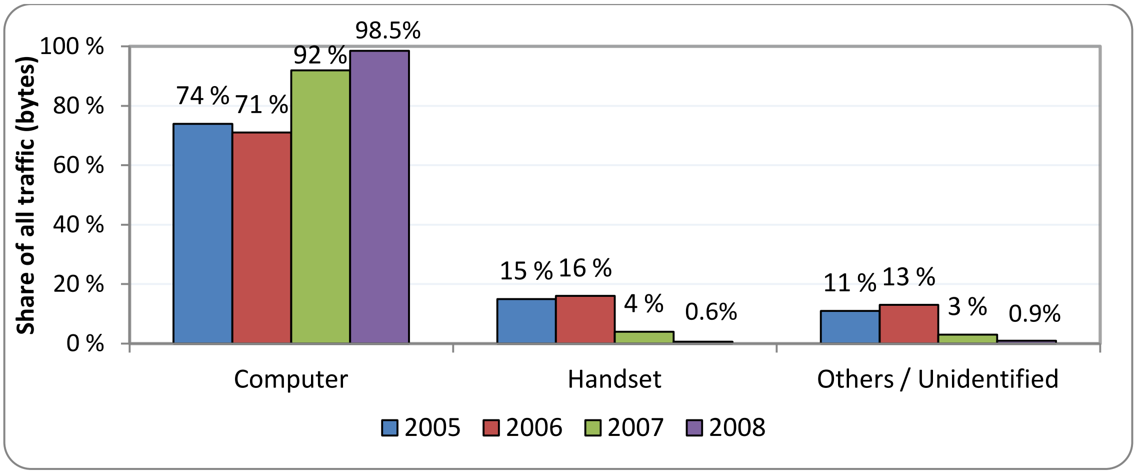

Figure 11 Traffic in mobile networks by operating system

As computer generated traffic increased from 92% to over 98%, a large share of the increase in total traffic volume was also computer generated in 2008. Absolute mobile handset traffic volume approximately doubled from 2007, even though the relative share of handset generated traffic decreased. These results confirm the trend that has been visible during the previous years.

According to Kivi (2009c) the share of data terminals is roughly 5% of all mobile devices (mobile handsets: 95%). Thus computers generate hundreds of times the traffic than mobile handsets. The increase in computer generated absolute traffic volume can be explained by the market changes described earlier. The amount of data terminals for laptops and 3G/HSDPA capable mini-laptops in the operator portfolio has increased and they have been marketed aggressively and for a competitive price. On the other hand, the increase in mobile handset based traffic can also partly be explained by the pricing. In addition, Kivi (2007b) in his research found that a large renewal of the Finnish mobile handset base happened during 2005 and 2006 and the amount of Symbian OS based devices doubled. Since then operator marketing has focused on advanced models and the development in handsets and their features has provided, for instance, improved usability and larger screens improving also the user experience. In addition, the amount of mobile optimized and improved content has been increasing.

5.2.2 Traffic by day and hour

Next, Computer and Handset categories were compared by providing diurnal distributions of the traffic they generate. The distributions are presented in Figure 12.

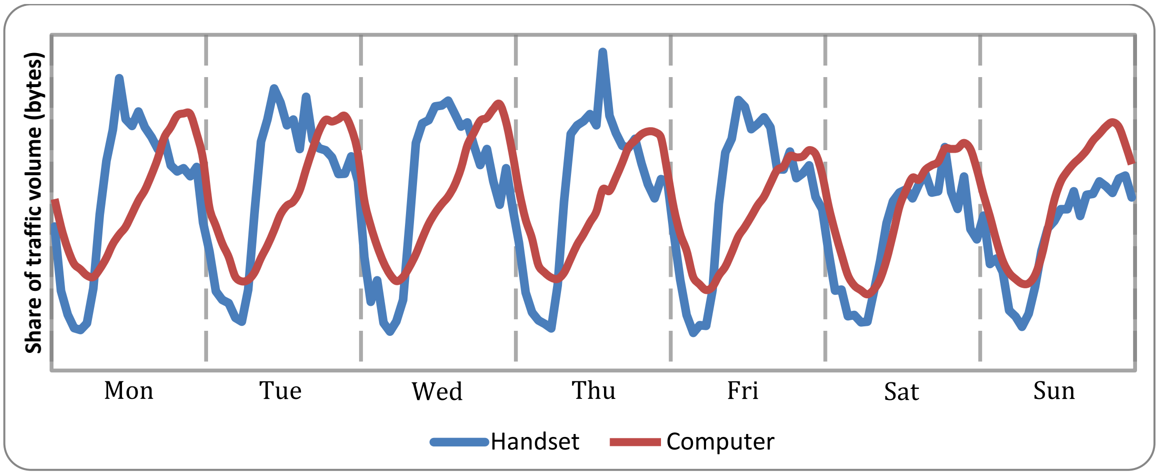

Figure 12 Traffic by day and hour

Handset traffic has high variations during the day. This could be explained by it reflecting human activity more than computer traffic. On the other hand, as the total volume of handset generated traffic was lower than computer traffic, single users have larger weight on the results. However, it is clearly visible that handset traffic tends to peak earlier than computer traffic, usually in the morning. For example, during the measured Monday, Tuesday, and Friday the busiest hour was from 9 am to 10 am. On the other hand, handset traffic decreases during the weekend, which refers to business oriented use, to the use of substitutive devices during weekends such as the home computer, or to mobile device usage being concentrated to “on the move” use.

Computer traffic is more continuous than handset traffic and peaks in the evening between 6pm and 10pm. The pattern follows the results from the previous measurements (Kivi, 2008). In the mobile domain this could be explained by computers with mobile access actually used mostly at home, as part of the mobile broadband subscriptions are substituting fixed internet broadband subscriptions. On the other hand, there may be increased capacity in the network during the evening when less calls are made. In addition, computer usage does not decrease during weekends as handset traffic does.

5.2.3 Computer usage of applications

Computer application usage results include comparison of total traffic volumes and diurnal distributions between different applications.

Computer traffic has two main categories with a cumulative share of about 96%, namely Web and Unidentified (Figure 13). The amount of web traffic has increased from year 2007, but as web traffic can include variety of traffic types, no obvious explanation can be provided on this granularity level. However, Internet video traffic has been forecasted to constitute the majority of all IP traffic in the future (Cisco, 2009), which could explain partly the observed increase.

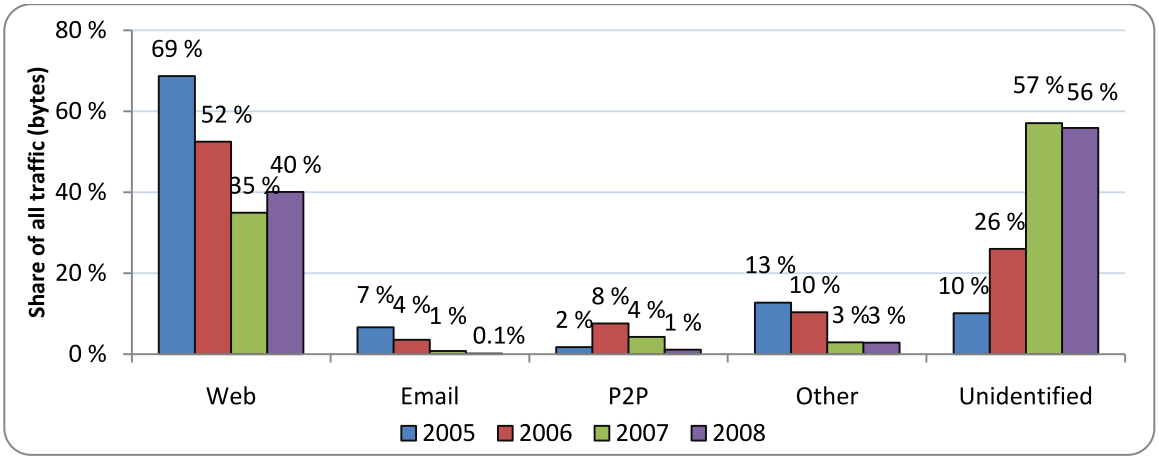

Figure 13 Computer traffic by application

On the other hand, an interesting remark is the amount of unidentified traffic and the low share of P2P traffic. The reason may be the difficulties of P2P traffic identification, as same sort of results were achieved also last year (see Kivi, 2008), as P2P traffic applications tend to use many features not suitable for the port based identification, resulting to some unidentified P2P traffic. This assumption is supported by the fact that the P2P ratio of downlink traffic per all traffic seems to match well with Unidentified (close to 50%), whereas other categories are more downlink-oriented. In addition, when the diurnal distribution of P2P and Unidentified traffic are compared, they seem to match rather well, as can be seen in Figure 14. P2P and Unidentified traffic are also more evenly distributed than Email and Web categories.

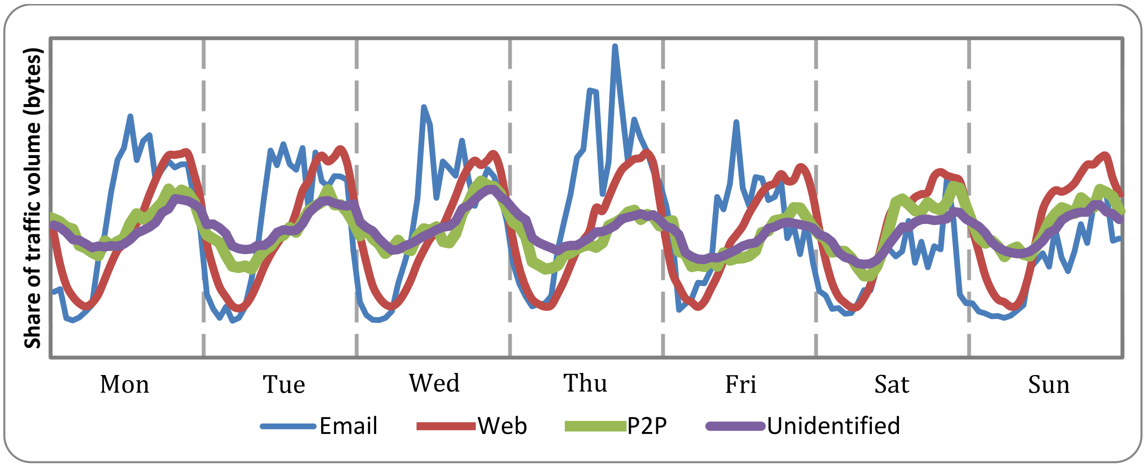

Figure 14 Computer applications by day and hour

Computer Web traffic seems to match the distribution of Computer traffic in general, peaking in the evening, whereas Email traffic concentrates more on the working hours and decreases during the weekend. It is possible that Email is more business oriented than the other categories, which would explain the behavior partly.

5.2.4 Handset usage of applications

The total traffic and diurnal application distributions for mobile handsets are presented in Figure 15 and Figure 16, respectively.

Mobile handset traffic is Web oriented with a share of 69% of all handset traffic. This trend has also been seen during the previous years in our measurements and in 2007 the traffic was actually even more web oriented. In addition, the share of Email traffic has been decreasing annually by multiple percents. One explanation for the decrease in web traffic volume is the fact that Email traffic does not require as much capacity as the web traffic. On the other hand, Email traffic is also transferred on top of HTTP, which means that part of the Web traffic can actually be Email. Unidentified traffic was the second largest category of mobile handset usage for the first time in 2008 measurements. No correlation with any of the other application categories was noticed and thus little can be said about the traffic of this category.

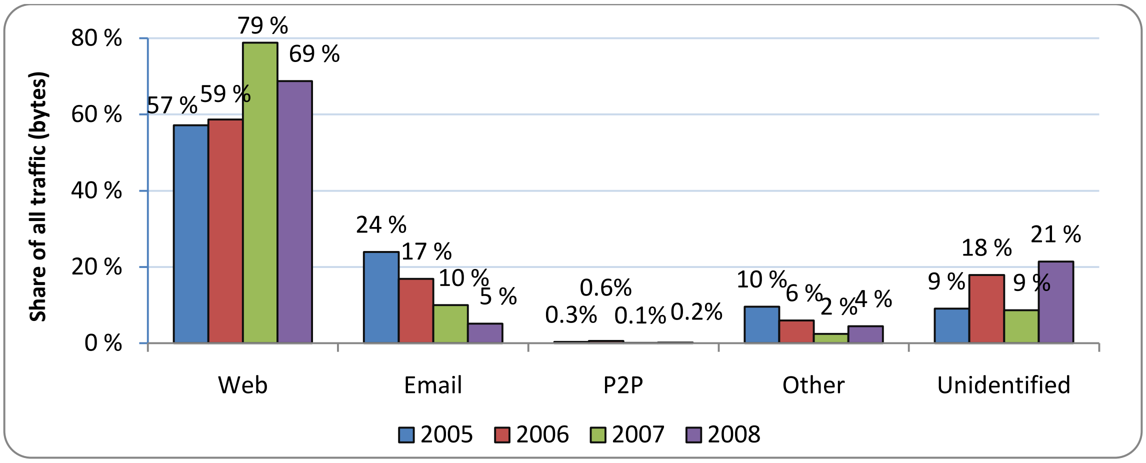

Figure 15 Handset traffic by application

For closer comparison, the two largest identified categories were chosen; Web and Email. Figure 16 shows the diurnal distribution of these categories during a time period of one week. An interesting observation is that web traffic is rather evenly distributed during daytime on workdays, as there are no clear peak hours. There also seems to be noticeable web usage during weekends. On the other hand, Email traffic has a clear weekday and working hour orientation and Saturday has the lowest usage.

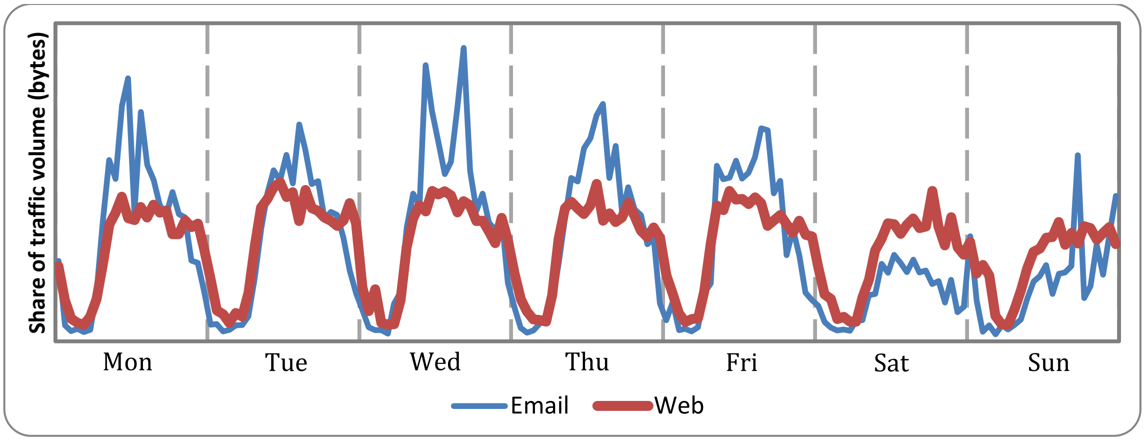

Figure 16 Handset applications by day and hour

5.3 Mobile handset web traffic results

As the mobile handset traffic has been observed to be dominated by web traffic, this chapter discusses this traffic category in more detail. Three viewpoints are taken, mobile handset web browsing (HTTP traffic), mobile handset "secure web" browsing (HTTPS traffic), and "mobile optimized" browsing, i.e., browsing to mobile handset optimized web sites. The ranking in all lists provided in this chapter are also based on traffic volumes. Mobile operator sites were excluded from the web browsing results due to sensitivity reasons.

5.3.1 Mobile handset web browsing

The objective of the ranking list, including HTTP traffic (85% of all web traffic), was to provide the top web browsing sites with most traffic by mobile handsets. The ranking list presented in Table 4 shows that traditional media and familiar websites from the fixed Internet are also used with mobile handsets, including content from traditional

Finnish media houses. In general, nine of the top fifteen web domains were Finnish sites, indicating high interest of Finnish content midst the Finnish mobile subscribers. Other observed categories were, for instance, social media and adult content. However, mobile handset web browsing seemed not to be concentrated, as the rank number one (iltalehti.fi) was noticed to have a share of only 4% of all HTTP traffic.

Table 4 Ranking list of top handset traffic HTTP domain names

Rank	Domain name*	% of HTTP traffic	Information
1	iltalehti.fi	4%	Traditional media
2	kauppalehti.fi	2%	Traditional media
3	mtv3.fi	2%	Traditional media
4	suomi24.fi	1%	Social media
5	opera-mini.net	1%	Opera Mini browsing
6	tube8.com	< 1%	Adult content
7	irc-galleria.net	< 1%	Social media
8	facebook.com	< 1%	Social media
9	bigbrother.fi	< 1%	Traditional media
10	sihteeriopisto.net	< 1%	Adult content
11	hs.fi (includes oikotie.fi)	< 1%	Traditional media
12	flickr.com	< 1%	Social media
13	ilmatieteenlaitos.fi	< 1%	Information (weather)
14	wikimedia.org	< 1%	Social media (mostly uplink)
15	blogger.com	< 1%	Social media
* Operator sites not included			

5.3.2 Mobile handset "secure web" traffic

Encrypted web traffic, constituting about 15% of the total web traffic, was more concentrated than the HTTP traffic. About one fourth of the encrypted web traffic during the measurement was generated to or from a nokia.com domain name, which was manually inspected and is presumably an email server. In total six of the top ten domains are email or synchronization related servers, making email the biggest category in secure web traffic. Other categories that were observed from the data are media management, gambling, and m-banking, which use encrypted traffic to provide privacy for the end users. The whole ranking list can be seen in Table 5. Even though encrypted web traffic was only about 15% of all web traffic, the top five domain names would place themselves in the top 15 ranking list of HTTP domain names also, indicating strong presence of Email traffic in the mobile Internet usage of the Finnish mobile subscriber population.

Table 5 Ranking list of top handset HTTPS domain names

Rank	Domain name*	% of traffic	Information
1	nokia.com	25%	Email
2	sok.fi	7%	Email
3	fmdm.net	3%	Media management
4	logica.com	3%	Email
5	veikkaus.fi	3%	Gambling
6	op.fi	2%	Nokia Intellisync
7	turku.fi	1%	Email
8	hus.fi	1%	M-banking
9	eqonline.fi	1%	F-Secure Mobile Service
10	f-secure.com	1%	Email
* Operator sites not included			

5.3.3 Mobile handset optimized web traffic

Last, the ranking list for mobile handset optimized web sites was provided, as there is an increasing amount of mobile optimized web sites being introduced at the moment, but little information about their usage is available. The top 20 ranking list can be seen in Table 6. In general terms, the amount of mobile optimized sites found from the DNS data was still rather low in 2008, as in total 110 different mobile optimized domain names were found[15].

Table 6 Ranking list of "mobile web" sites

Rank	Domain name*	% of mobile web traffic
1	nokia.mobi	50%
2	m.facebook.com	19%
3	m.hs.fi	8%
4	lataa.yle.mobi	5%
5	m.youtube.com	2%
6	www.foreca.mobi	1%
7	wap.jamba.fi	1%
8	m.volvooceanrace.org	1%
9	wap.sp.fi	< 1%
10	www.ovi.mobi	< 1%
11	wap.aftonbladet.se	< 1%
12	m.espn.go.com	< 1%
13	m.ebay.com	< 1%
14	m.note.nokia.com	< 1%
15	wap.eniro.fi	< 1%
* No operator sites included		

Mobile optimized browsing was observed to be highly concentrated, as half of the traffic volume is either to or from a nokia.mobi domain name. However, this could be explained by the resource intensive application downloads, which the nokia.mobi server, generating most of the traffic, presumably hosts. In addition, m.facebook.com was placed second and m.youtube.com fifth, indicating visible usage of social media sites. In addition to Nokia and social media, traditional Finnish media houses are in the

[15] With more than 40 DNS requests made during the measurement period

top places. YLE and Sanoma, which ranked well in general web browsing ranking list as well, were the most popular media houses.

5.4 Summary of the descriptive statistics

Traffic in the Finnish mobile networks was multiplied in 2008. However, most of the total traffic in mobile networks was generated by computers (98.5%) and mobile handsets originated only 0.6% of the total network traffic volume. This is explained by the increased amount of data terminals used by laptops, combined with the fact that one computer generates easily hundreds of times the traffic that a mobile handset generates. Nevertheless, also the absolute amount of mobile handset generated traffic is increasing and it approximately doubled during 2008. Symbian OS devices were observed to be the dominant mobile handset devices using mobile Internet. Computer generated traffic consisted mainly of web traffic (40%) and unidentified traffic (56%), from which possibly a large share is P2P traffic. On the other hand, mobile handset traffic is dominated by web, consisting of almost 70% of the total handset traffic volume. Handsets were noticed to be more morning and working day oriented in comparison to computer traffic that peaks in the evening.

Familiar web sites from the fixed Internet are popular also in mobile browsing, such as Finnish media houses and social media. However, mobile web browsing is not very concentrated. In addition, email and synchronization were noticed to be popular in mobile Internet use. The browsing to mobile optimized sites is more concentrated than with web browsing in general, Nokia as the most popular mobile optimized web content provider.

6 Development of mobile network traffic measurements

This chapter presents the classification of the development areas for mobile operator network traffic measurements. First, limitations of the measurement process are identified. Second, based on the limitations and the literature review the possible development areas are classified. Last, possible development paths and concrete recommendations are given from two different viewpoints; MoMI project and mobile operator.

The market research process phases by Aaker et al. (2007), were divided in this thesis into three more general areas; preliminary planning (*Agreeing on research* & *Research objectives,* and *Estimate the value of the information*), planning (*Research design*), and implementation (*Data collection, Data analysis,* and *Reporting*). The analysis in this chapter is conducted by utilizing the market research framework.

6.1 Limitations

This chapter identifies the limitations that have been identified during the traffic measurement process in the MoMI project. The knowledge from this chapter is later used in the analysis of development areas and recommendations.

6.1.1 Preliminary planning

Good preliminary planning and defining the objectives and research problems are requirements for a good implementation of network traffic measurements. In the phases of preliminary planning basically all of the major stakeholders are involved. In the MoMI project the amount of industry affiliates has been large, making the communication between different parties rather complex and time demanding, which partly limits the amount of preliminary planning that can be conducted.

6.1.2 Research design

There were two tasks in this phase, research design of the measurement and research design of the data analysis. The former, measurement planning, requires close collaboration between the measurement administration on TKK side and the operators' network administration affiliates, as mobile operators may have multiple operator and infrastructure vendor specific solutions in the network. The amount of this communication and knowledge exchange has been to some extent limited in the MoMI project and the operator specific measurement implementations have not been fully transparent.

6.1.3 Data collection

Data collection in the MoMI project has been conducted for a period of roughly one week and the results have been generalized to describe each year's usage. The measured trace data from the *Gi* interface has been limited to the user plane traffic and to the network and transport layer protocol headers, i.e., IP and TCP or UDP headers. The only exception has been DNS request and response traffic, from which application layer information has been recorded. TCP/IP headers do not provide any control plane information, such as mobility and session management, or identification of the subscriber, which limits the analysis to a total traffic volume level.

From a technical measurement viewpoint, the possible limitations of a measurement concern mainly processing and memory requirements. No evidence of congested buffers was visible in the trace data chosen for the analysis. However, there was one technical issue in the 2008 measurements. Four hours (Monday 7am-10am) from the trace data from one operator was corrupted for an unknown reason. These hours were substituted in the analysis with concurrent hours from another weekday with similar traffic patterns, adding possible bias in the results. Nevertheless, previous measurements (Kivi, 2006; Kivi, 2007a, Kivi, 2008) have shown that Monday has a fairly similar traffic profile in comparison to other weekdays, and the level of analysis in 2008 was total traffic level. Thus the bias from the four hours was assumed to be insignificant. Another technical issue in data collection is the delay in the radio interface and the backbone (Kilpi, 2003). According to Romirer et al. (2008), user mobility and flow control in a SGSN cause larger delays than the queue related delay in GGSN. However, all these delays are small in comparison to the level that the analysis is conducted when studying general usage patterns.

6.1.4 Data analysis

One limitation of the data analysis is the amount of time that processing large trace data files requires. Processing of the data can take days, which clearly limits the amount of different analyses that can be conducted. However, as the measurements have been conducted annually, this has not been a significant issue.

The format of the trace data, including only user plane data and protocol headers, limits the analysis methods and metrics that can be used. No unique user, device, location, or session identification, can be conducted. However, device identification by operating system identification is rather accurate. Certain level of session analysis could also be conducted from plain user data, but including some uncertainties.

Application identification does include uncertainties leading to a certain amount of unidentified traffic. One limitation of the method is also the inability to identify subcategories, such as different web traffic types (e.g., web browsing, application download, instant messaging).

The largest uncertainty in the data analysis comes from the web domain identification and web traffic analysis. The mapping of web servers and domains was conducted on server level, resulting in web traffic that could not be identified to a single web domain. This is due to the reason that hosting of multiple web sites from a single server is rather popular nowadays (Edelman, 2003). As the chosen metric in the analysis is traffic volume, some bias is possible also because of this. Volume based ranking of web pages stress sites that are byte-wise large, such as sites including multimedia.

6.1.5 Reporting

Due to the frequency of the measurements, also the reporting has been conducted annually by publications and presentations, roughly between a few months and half a year after the data collection. Thus, the large time span between data collection and reporting limit the topicality of the results. In addition, the frequency of the reporting (and data collection) being a year, many changes in user patterns between the measurements are possible.

6.2 Design & development areas

Both, measurement process phase-specific development areas, as well as some more general areas affecting multiple phases were identified. The design and development areas of mobile network traffic measurement are presented in Figure 17. On the left side of the figure the more general choices are presented. The analysis from the mobile operator viewpoint concentrates on these general design and development areas. The MoMI project viewpoint in the analysis concentrates on the right side of the figure, which includes the measurement process phases and some major exemplary variables or choices relating to each phase. The arrows in the figure represent the causal relationships between the different phases, as well as between the phases and the more general choices on the left. On a high level, the general choices affect the more specific measurement phases and their implementation. On the other hand, the decision relating the early measurement process phases affect the following phases and either limit or enable the choices that can be made in them.

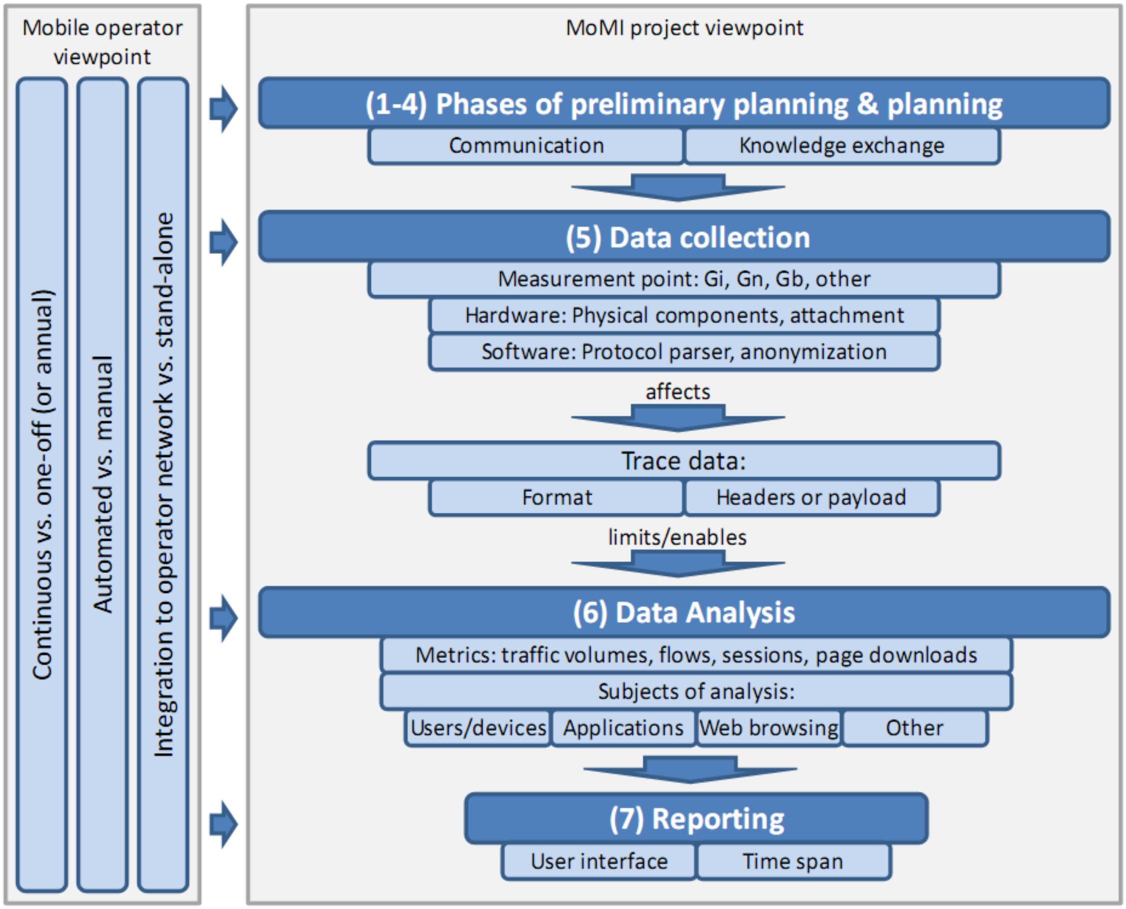
Mobile operator viewpoint
Continuous vs. one-off (or annual)
Automated vs. manual
Integration to operator network vs. stand-alone
MoMI project viewpoint
(1-4) Phases of preliminary planning & planning
Communication
Knowledge exchange
(5) Data collection
Measurement point: Gi, Gn, Gb, other
Hardware: Physical components, attachment
Software: Protocol parser, anonymization
affects
Trace data:
Format
Headers or payload
limits/enables
(6) Data Analysis
Metrics: traffic volumes, flows, sessions, page downloads
Subjects of analysis:
Users/devices
Applications
Web browsing
Other
(7) Reporting
User interface
Time span

Figure 17 Design & development areas of traffic measurements

6.2.1 General areas

Figure 17 includes three general aspects of mobile network traffic measurement that affect all the measurement phases. These interrelated areas are the level of integration to operator network, level of automation in the measurement, and the continuity of the measurement.

First of all, traffic measurements can be conducted continuously, or as one-off measurements as in the MoMI project. The advantage of continuous measurements is that there is more freedom in choosing the time period or periods that are analyzed. On the other hand, continuous measurements require either a lot of real-time processing or a lot of disk space where to save the trace data. By introducing automation to the measurement phases the amount of recurring manual work decreases, but the development of the needed automation requires more resources in the beginning. The third area is related to the amount of integration between the traffic measurement and the existing reporting and business intelligence systems of the mobile operator. On one hand, integration can provide advanced analysis, but on the other hand, it may require cooperation with the network infrastructure manufacturer to be able to be implemented. In addition, highly integrated measurement setup is dependent on the properties of the infrastructure, and thus may need development if the infrastructure is updated.

6.2.2 Measurement phase specific areas

During the first four phases of the measurement process (preliminary planning and planning), it is important to enable communication and knowledge exchange between the affiliates. This way the research questions, resource usage, as well as the outcome of the *Research design* can be optimized.

The different measurement point alternatives of *Data collection* phase were presented in Chapter 3.4, all providing different levels of information to be used in the data analysis. The measurement point affects also the hardware and software choices in the measurement, by setting demands to the bandwidth that needs to be processed by the monitoring system. The question of what to measure affects the trace data and its format, including whether user or control data is measured. On protocol layers, the common alternatives are either to measure network and transport level protocol headers or also the payload, i.e., the application level data. The time scale of the data collection is mainly affected by the objectives of the measurement.

Choices in the *Data analysis* are mostly related to the metrics, analysis methods, and the subjects of analysis. Development in this phase is highly influenced by the trace data format and thus the data collection methods. The development of the analysis methods at the lowest levels is code optimization and on higher levels algorithm or tool selection. On a higher level, the analysis can either be made on general total traffic level, or with enhanced trace data on a single user basis.

The development of the *Reporting* phase is interrelated to the earlier phases. The time span required for the reporting can be shortened by shortening the earlier phases. Automation of the reporting is another factor that can enhance the usability of the measurement system, as well as speed up the reporting phase.

6.3 Recommendations

This chapter utilizes the information provided in the previous chapters and provides concrete recommendations. First, the case MoMI project is discussed. Second, a mobile operator viewpoint is taken.

6.3.1 Case MoMI project

There is a need for increased knowledge exchange between different affiliates mainly in the *Research design* phase so the measurement implementation would be more transparent and the planning of the data analysis could begin earlier. In addition, parallel planning of data collection and data analysis (tasks of the *Research design*) are needed to enable shortened time span for the entire measurement process. Thus, reporting could be conducted closer to the *Data collection* providing more topical results.

There are two high level choices for developing the MoMI project traffic *Data collection* phase based on the measurement point. The alternatives include measurement points at the *Gn* (between SGSN and GGSN) or the *Gi* (between GGSN and the Internet) interface. However, the most recommendable solution to develop the measurements would be to measure all the main interfaces, including *Gb/IuPS, Gn*, and *Gi*. The measurement from *Gb/IuPS* could provide stateful user and possibly device identification information that can be mapped to the decapsulated IP traffic from the *Gi* interface and session information from the *Gn* interface. In addition, certain level of location information could be extractable from the *IuPS* interface. However, this would require multiple protocol parsers, clock synchronization, and a lot of resources for measuring multiple links, which is why it may not be feasible. Thus, also less resource intensive more applicable choices were identified for short-term development.

All in all, six development paths were identified (Figure 18). The development paths have been prioritized based on the granularity of trace data (user plane vs. user and control planes / headers vs. headers and payload) the link can provide.

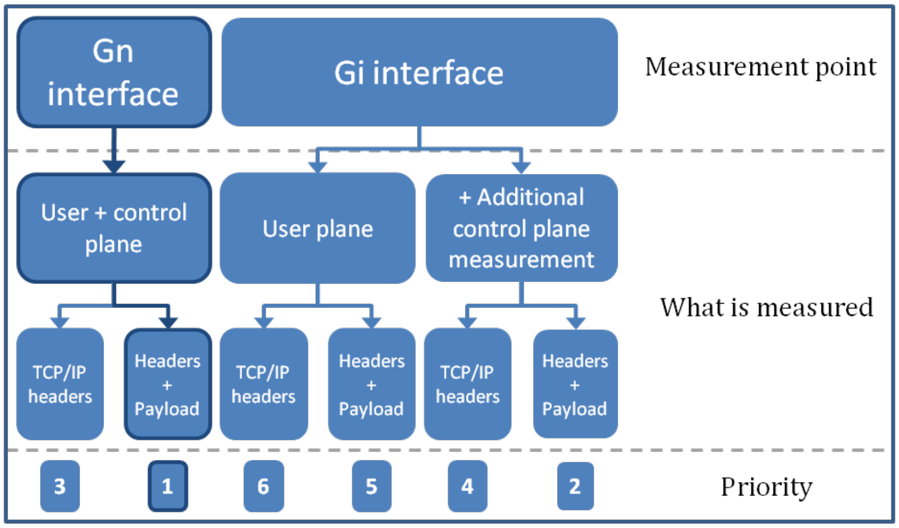

Figure 18 Development paths for the MoMI project traffic measurement

The prioritized recommendation for the *Data collection* is to measure the *Gn* interface for both, protocol headers and payload (path 1). According to Svoboda et al. (2007b), the *Gn* interface is a good trade-off between the available control plane information and the representativeness of the data. An alternative for the *Gn* measurement is a *Gi* measurement with an additional control traffic measurement (see, e.g., Kalden, 2004) from a link or a node, providing session and identification information. This path (2) would provide almost the same granularity of data as the first recommendation, but the two simultaneous measurements may need more resources and clock synchronization. The same analogy can be found between the next development paths (3 and 4); The *Gn* interface measurement is preferred, as it should provide the same (or better) data granularity from a single measurement. Path 5 would include measuring the *Gi* interface as in this thesis, but including also payload inspection. This could enable more advanced application identification and web traffic analysis, for example. The last development path (6) would include development only in the *Data analysis* phase, as the measurement organization would remain unchanged.

Considering the development paths from the *Data analysis* viewpoint, there are differences on the available metrics and methods based on the development paths. The development paths with user and control plane data, as well as payload (path 1), would enable the largest variety of analysis methods. Measuring control traffic enables more detailed session level analysis (and possibly unique user and device identification), whereas payload data can enable more accurate classification and identification of applications and content. About the application identification methods, it seems that the most accurate ones are advanced statistical, e.g., machine learning, or combinatorial methods including also certain amount of payload inspection. For instance, the combinatorial method, GTVS by Canini et al. (2009), is to some extent publicly available to be used by other researchers. However, the port based identification is still applicable with traditional applications and, for instance, by combining payload inspection to port identification or simple statistical analysis of packet or flow features, rather accurate results could be acquired. From the web traffic analysis, independent of the chosen development path, one development recommendation is to develop the domain identification to work on a DNS request level. This could remove some of the uncertainties involved in the web browsing results.

As many of the development areas require extensive technical knowledge, collaboration with instances, for example, inside TKK could be considered to ease the resource requirements. An alternative would be to prioritize the recommendation areas of the chosen path and do them iteratively during a longer time span.

6.3.2 Case mobile operator

The first recommendation for all operators is to start measuring network traffic to support their other business intelligence. All operators should utilize traffic measurements at least on a descriptive level, thus providing valuable information to complement their business intelligence. Value added service operators should seek for advanced per user based analysis, including subscriber identification and some level of integration with the existing reporting systems, whereas a bit pipe operator could manage with a total traffic volume level understanding. Even though high levels of integration to the existing network infrastructure could provide detailed information, a stand-alone measurement system is recommendable, as then it is more decoupled from possible hardware and software changes of the network infrastructure. However, in the analysis and reporting phases some integration or mapping of results to existing reporting systems and further subscriber information would enable better utilization of the measurement system in strategic decision making and business development. In addition, the analysis and reporting should be developed as highly automated to ease the use and utilization of the system and to work as on-demand basis without large one-off investments and operating costs. When the measurement can be utilized in many areas and by multiple departments, such as marketing, CRM (Customer Relationship Management), and corporate performance management, the cost of the measurement is low in comparison to the perceived value.

On a more general level, a mobile operator needs to decide whether the measurements and the development of the monitoring system are conducted as internal, in cooperation with academia as in the MoMI project, or outsourced to a company offering commercial solutions for traffic monitoring. The choice between these alternatives affects the implementation from some parts, but the decision should be based on an analysis of the value versus costs ratio of each alternative. If the time span available for the development of the measurement system is short, outsourcing or a direct purchase of a commercial solution should be considered.

7 Conclusions

This chapter summarizes the main findings of the thesis. In addition, recommendations for further research are provided.

7.1 Summary of findings

First, this thesis provided descriptive statistics on the mobile Internet usage in Finland. They showed that traffic in Finnish mobile networks multiplied in 2008 and most of the total traffic was generated by computers, whereas mobile handsets originated less than 1% of total traffic volume. From both of these categories, dominant operating systems were found, Windows (Computers) and Symbian (Handsets), generating the majority of the traffic of the category. Computer and mobile handset generated traffic were observed to differ by their application and diurnal distributions. Mobile handset traffic is largely web traffic, generating almost 70% of all traffic volume and peaking during mornings and working days, whereas computer traffic peaks in the evening and the share of web traffic is smaller. Furthermore, mobile handset web traffic was observed to consist of different traffic classes. For example, email and other synchronization and application downloading were observed.

Second, based on the literature review and experiences of the measurement, a classification of possible measurement development areas was conducted and concrete development recommendations were provided. The best solution would be to measure all the main interfaces, which could enable advanced user, as well as some level of device and location identification. However, as this alternative would require a lot of resources, six more feasible high level development paths were identified for short-scale development in the MoMI project and were prioritized based on the granularity of the information they could provide. The main differences of the paths are the interface that is measured, the nature of the data that is measured (user data or user and control data), and the parts of the protocols that are measured (network and transport level protocol headers or headers and payload). The MoMI project should consider measuring the *Gn* interface which includes also control traffic, or complement the existing *Gi* interface measurement with additional control plane data collection. This could enable advanced analysis and a wider variety of problems to be researched, such as ones involved with detailed session level analysis and possibly user identification. In addition, application, content, and web domain identification could be developed by measuring application level data and enhancing the analysis methods.

From an operator viewpoint, network traffic measurements could bring added value to their decision making and business development. The full advantage of a traffic measurement system could be utilized by conducting the measurements as highly automated and continuous. If the data analysis was also integrated to the existing reporting systems, at least to some extent, advanced analysis would be possible. For instance, mapping the results to further subscriber data could provide utilization of the results in multiple corporate functions, from optimizing marketing campaigns to corporate performance management.

On a general level, the most important aspect in designing network traffic measurements is to fit the measurement to match the objectives, as different granularity levels are provided from different interfaces. Careful consideration of the value versus cost ratio of different development paths is suggested.

7.2 Discussion

This thesis had three objectives, providing factual statistics on the Finnish mobile Internet usage, classifying the development and design areas of the network traffic measurements conducted to provide the statistics, and providing recommendations for the future development of such measurements.

The factual statistics and the analysis of the measurement organization provided information about the limitations of the existing measurement. Even though multiple limitations were identified, it should be remembered that network traffic measurements always include certain amount of uncertainty and no analysis methods are available providing 100% accuracy. Thus, even the existing measurement provides highly representative and reliable data on a total usage level. However, by developing some of the identified areas certain uncertainties could be decreased, new metrics utilized, and new research areas enabled.

The main recommendations that were provided have been, to some extent, conducted in previous research. Svoboda (2008) has been able to identify unique users, and Kalden (2004) has executed additional *Gi* interface measurements for control plane traffic to complement IP traffic trace data. However, as only temporary identifiers in the *Gn* interface are mandatory (in PDP context activation and deactivation procedures) and there can be implementation dependent differences in the computation of the identifiers, user identification could be complex or even unfeasible without measuring also the *Gb/IuPS* interfaces. If a *Gi* interface measurement was complemented with an additional control plane measurement, further knowledge exchange between TKK and the operators would be needed in deciding the most optimal measurement link or node.

For mobile operators the recommendations were provided on a higher level, concentrating on the issues that could enable the operator to utilize the measurements as a support for decision making and business planning. Nevertheless, the recommendations for the MoMI project are also valid in the mobile operator viewpoint, and the development paths are adaptable for mobile operators.

7.3 Further research

There are multiple areas where further research could be conducted. First of all, the measurements should be continued also in future to provide longitudinal data on the development of the Finnish mobile market. The results have created value by providing factual statistics available nowhere else.

One future research track could concentrate on the technical aspects of the measurements and their development. As described in the recommendations, some enhancements on the measurement or analysis level should be conducted to enable better understanding of the mobile Internet usage patterns. These research tasks could include, for example, code optimization and development of more accurate algorithms. In addition, it would be important to find and exclude unwanted traffic from the network, such as SYN flooding, as it may bias results of a user behavior level analysis.

If future measurements enable individual user and device identification, some new research areas could be enabled. Instead of general network level usage patterns, user behavior aspects could be studied. Thus, user populations of different devices and services could be studied. User level or single service user population based analyses could enable for example, demand estimation. Users could also be segmented based on the actual usage patterns instead of demographics. The usage based segments could be linked with additional subscriber and billing information, which could enable pricing optimization and profitability related studies. In addition, for a mobile operator usage based segmentation could provide enhanced understanding, for example, to the CRM or sales in targeting products to the most applicable segments. These all could help mobile operators in looking for new revenue sources and optimizing costs due to falling ARPUs (Average Revenue per User).

Accurate device identification could enable interesting studies on mobile device convergence area. As the differences between mobile Internet capable devices are blurring, it would be interesting to study if there are any differences in the usage of converging devices and to identify the factors or features that affect the usage patterns. In addition, device identification could enable mapping of usage data with, for example, retail sales data to provide pricing information.

Studying mobile web usage was a large part of this thesis and the domination of web in handset usage seems to continue in the future. Thus, one research area could be concentrated on studying the possibilities of mobile network traffic for mobile web analytics. For example, automated web content identification methods could be developed or mobile advertising studied.

This thesis concentrated on 2G and 3G networks, leaving LTE (Long Term Evolution) and SAE (Service Architecture Evolution) out of the scope. One research topic could be to study the next generation network architectures and their possible effects on the network traffic measurements.

References

Laws, standards & specifications

3GPP TS 23.002. Third Generation Partnership Project, Technical Specification 23.002. Network architecture, V8.4.0 (2008-12).

3GPP TS 23.060. Third Generation Partnership Project, Technical Specification 23.101. General Packet Radio Service (GPRS) Service description, V8.3.0 (2008-12).

3GPP TS 23.101. Third Generation Partnership Project, Technical Specification 23.101. General Universal Mobile Telecommunications System (UMTS) architecture, V8.0.0 (2008-12).

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, V8.6.0 (2008-12).

Act on the protection of Privacy in Electronic Communications (516/2004). [Online] Finlex, 17 July 2009. Available at: http://www.finlex.fi/fi/laki/ajantasa/2004/20040516 [Accessed 26 July 2009].

Personal Data Act (523/1999). [Online] Finlex, 1 September 2008. Available at: http://www.finlex.fi/fi/laki/ajantasa/1999/19990523 [Accessed 28 July 2009]

Bibliography

Aaker, D., Kumar, V. & Day, G., 2007. *Marketing Research*. 9th ed. John Wiley & Sons, New Jersey, US.

Bannister, J., Mather, P. & Coope, S., 2004. *Convergence Technologies for 3G Networks*. John Wiley & Sons, New Jersey, US.

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.

Berners-Lee, T., Fielding, R., Irvine, U. C. & Masinter, L., 1998. *RFC 2396 - Uniform Resource Identifiers (URI): Generic Syntax*. [Online] Internet Society, August, 1998. Available at: http://www.ietf.org/rfc/rfc2396.txt [Accessed 13 July 2009].

Bettstetter, C., Vögel, H.-J. & Eberspächer, J. , 1999. GSM Phase 2+ General Packet Radio Service GPRS: Architecture, Protocols, and Air Interface. *IEEE Communications Surveys and Tutorials*, 2(3), pp. 2-14, 1999.

Canini, M., Li, W., Moore, A. W. & Bolla, R., 2009. *GTVS: boosting the collection of application traffic ground truth*. Technical Note. University of Cambridge.

Cisco, 2009. *Cisco Visual Networking Index: Forecast and Methodology, 2008-2013*. [Online] Cisco Systems, Inc., 9 June 2009. Available at: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf [Accessed 13 July 2009].

Claffy, K., Braun, H-W. & Polyzos, G., 1995. A Parameterizable methodology for Internet traffic flow profiling. *IEEE Journal on Selected Areas in Communications (JSAC)*, 13(8), pp. 1481-1494, October 1995.

Clegg, R., et al., 2008. Challenges in the capture and dissemination of measurements from high-speed networks. To appear in: *ET Communications, Special Issue on Simulation, Analysis and Measurement of Broadband Network Traffic*.

Comer, D., 2000. *Internetworking With TCP/IP – Vol. 1: Principles, Protocols, and Architecture*. 4th ed. Prentice-Hall, New Jersey, US.

Crotti, M., Dusi, M., Gringoli, F. & Salgarelli, L., 2007. Traffic Classification Through Simple Statistical Fingerprinting. *Computer Communications Review*, 37 (1), pp. 7–16, January 2007.

Crovella, M. & Krishnamurthy, B., 2006. *Internet Measurement: Insfrastructure, Traffic & Applications*. John Wiley & Sons, New Jersey, US.

Dahmouni, H., Rossé, D., Morin, B. & Vaton, S., 2005. Impact of Data Traffic Composition on GPRS Performance. In: *Proceedings of the 19th International Teletraffic Congress (ICT19)*, 29 August - 2 September 2005, Beijing, China.

Edelman, B., 2003. *Web Sites Sharing IP Addresses: Prevalence and Significance*. Berkman Center for Internet & Society. [Online] Harvard Law School, 12

September 2003. Available at: http://cyber.law.harvard.edu/people/edelman/ip-sharing/ [Accessed 28 July 2009].

Ficora, 2008. Matkapuheluhinnat 2008 - Kansainvälinen vertailu. [Online] Ficora, 30 May 2008. Available at: http://www.ficora.fi/attachments/suomial/5h3ijZXav/Matkapuheluhinnat_2008_-_KV-vertailu.pdf [Accessed 1 February 2009].

Ficora, 2009. *Market Review 2008: Communications Markets in Finland*. [Online] Ficora, 31 March 2009. Available at: http://www.ficora.fi/attachments/englantiav/5gl4IASfV/vivi_mk_08_ENG_lowres.pdf [Accessed 7 July 2009].

Forouzan, B. & Chung, F., 2002. *TCP/IP Protocol Suite*. 2nd ed. McGraw-Hill.

Gourley, D. & Totty, B., 2002. *HTTP: The Definitive Guide*. 1st ed. [Online] O'Reilly Media, Inc., 27 September 2002. Available at: http://proquest.safaribooksonline.com/1565925092 [Accessed 23 July 2009].

Haffner, P., Sen, S., Spatscheck, O., & Wang, D., 2005. ACAS: Automated Construction of Application Signatures. In: *Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*, pp. 197 – 202, August 22-26, 2005, Philadelphia, Pennsylvania, US.

Ilvesmäki, M., Luoma, M. & Kantola, R., 1998. Flow Classification Schemes in Traffic-based Multilayer IP Switching - Comparison Between Conventional and Neural Approach. Computer Communications, 21, pp. 1184-1194, May 1998.

Jain, R & Routhier, Shawn A., 1986. Packet Trains – Measurements and a New Model for Computer Network Traffic. *IEEE Journal on Selected Areas in Communications (JSAC)*, 4(6), pp. 986-995, September 1986.

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, 31 August – 5 September 2003, Berlin,Germany.

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. & Ekström, H., 2004. Searching for mobile mice and elephants in GPRS networks. Mobile Computing and Communications Review, 8(4), pp. 37-46, October 2004.

Kalden, R. & Ibrahim, S., 2004. Searching for self-similarity in GPRS. In: *Proceedings of the Passive and Active Network Measurement: 5th International Workshop (PAM)*, pp. 83–92, 19-20 April 2004, Antibes Juan-les-Pins, France.

Kalden, R. & Haverkort, B., 2006. Measuring and modelling of application flow length in commercial GPRS networks. In: *The 3rd International Conference on the Quantitative Evaluation of Systems*, pp. 83-92, 11-14 September 2006, Riverside, CA, USA.

Karagiannis, T., Broido, A., Faloutsos, M., & Claffy, K. C., 2004. Transport Layer Identification of P2P traffic. In: *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pp. 121-134, 25-27 October 2004,Taormina, Sicily, Italy.

Karagiannis, T., Papagiannaki, K. & Faloutsos, M., 2005. BLINC: Multilevel Traffic Classification in the Dark. *ACM SIGCOMM Computer Communication Review*, 35(4), pp. 229-240, October 2005.

Kilpi, J., 2003. A Portrait of a GPRS/GSM Session. In: *In: Proceedings of the 18th International Teletraffic Congress – ITC18, Vol. 2, pp. 981-990, 31 August – 5 September 2003, 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.

Kim, M.-S., Won, Y., & Hong, J., 2005. Application-Level Traffic Monitoring and Analysis on IP Networks. *Electronics and Telecommunications Research Institute (ETRI) Journal*, 27(1), pp.22-42, February 2005.

Kim, H., et al., 2007. Comparison of Internet Traffic Classification Tools. In: *ANF Workshop*, 2 November 2007.

Kim, H., et al., 2008. Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices. In: *Proceedings of the ACM CoNEXT Conference*, December 2008, Madrid, Spain.

Kivi, A., 2006. *Mobile Internet Usage Measurements - Case Finland*. Master's Thesis. Networking Laboratory, Helsinki University of Technology.

Kivi, A., 2007a. *Mobile Data Service Usage Measurements - Results 2005-2006*. COIN project report.

Kivi, A., 2007b. Mobile Data Adoption in Finland 2005-2006. In: *Proceedings of the 6th Conference on Telecommunication Techno-Economics*, 14-15 June, 2007, Helsinki, Finland.

Kivi, A., 2007c. Measuring Mobile User Behavior and Service Usage: Methods, Measurements Points, and Future Outlook. In: *Proceedings of the 6th Global Mobility Roundtable*, 1-2 June 2007, Los Angeles, California, US.

Kivi, A., 2007d. Diffusion and Usage of Mobile Browsing in Finland 2005-2006. In: *Proceedings of the 4th CICT Conference*, 29-30 November 2007, Copenhagen, Denmark.

Kivi, A., 2008. *Mobile Data Service Usage Measurements - Results 2005-2007*. COIN project report.

Kivi, A., 2009a. *Mobile Handset Population in Finland 2005-2008*. MoMI project report.

Kivi, A., 2009b. Measuring mobile service usage: methods and measurements points. *International Journal of Mobile Communications*, 7(4), pp. 415-435.

Kivi, A., 2009c. Diffusion of Mobile Data in Finland. Accepted for publication at *NETNOMICS: Economic Research and Electronic Networking*, 2009.

Lambert, M., 1995. RFC 1857 - A Model for Common Operational Statistics. [Online] Internet Engineering Task Force, October 1995. (Informational). Available at: http://tools.ietf.org/html/rfc1857 [Accessed 13 July 2009].

Madsen, T., Schwefel, P., Hansen, M., Bogh, J. & Prasad, R., 2005. *On Traffic Modelling in GPRS Networks*, 2005, Denmark.

Moore, A. W. & Papagiannaki, D., 2005. Toward the Accurate Identification of Network Applications. In: *Proceedings of the Sixth Passive and Active Measurement Workshop (PAM)*, pp. 41-54, March/April 2005, Boston, MA, US.

Moore, A. W. & Zuev, D. 2005. *Discriminators for use in flow-based classification*. Technical report, Intel Research, Cambridge.

Nguyen, T. & Armitage, G., 2008. A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys and Tutorials, 10(4), 2008.

Paxson, V., 1997. Measurements and Analysis of End-to-End Internet Dynamics. Technical report, University of California, Berkeley, April 1997.

Peikari, C. & Chuvakin, A., 2004. *Security Warrior*. O'Reilly Media, Sebostopol, CA, US.

Peuhkuri, M., 2003. *Internet Traffic Measurements – Aims, Methodology and Discoveries*. Licentiate Thesis. Networking laboratory, Helsinki University of Techynology.

p0f, 2006. the new p0f: 2.0.8. [Online] Zalewski, M., 6 September 2006. Available at: http://lcamtuf.coredump.cx/p0f.shtml [Accessed 14 July 2009].

Postel, J., 1980. *RFC 768 – User Datagram Protocol*. [Online] USC/Information Sciences Institute, August 1980. Available at: http://www.ietf.org/rfc/rfc768.txt [Accessed 13 July 2009].

Postel, J., 1981a. *RFC 791 - Internet Protocol.* [Online] DARPA Internet Program Protocol Specification. USC/Information Sciences Institute, September 1981. Available at: http://www.ietf.org/rfc/rfc791.txt [Accessed 10 July 2009].

Postel, J., 1981b. *RFC 793 - Transmission Control Protocol.* [Online] DARPA Internet Program Protocol Specification. USC/Information Sciences Institute, September 1981. Available at: http://www.ietf.org/rfc/rfc793.txt [Accessed 10 July 2009].

Ricciato, F. Vacirca, F. & Karner, M., 2005. Bottleneck Detection in UMTS via TCP Passive Monitoring : A Real Case. In: *Proceedings of 1st International Conference on Wireless Internet*, Budapest, July 2005, Hungary.

Riikonen, A. & Kivi, A., 2009. *IP Traffic Measurements 2008*. MoMI project report.

Romirer-Maierhofer, P., Ricciato, F. & Coluccia, A., 2008. Explorative Analysis of Oneway Delays in a Mobile 3G Network. In: *The 16th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN)*, 3-6 September 2008, Cluj-Napoca, Romania.

Romirer-Maierhof, P., Ricciato, F., D'Alconzo, A., Franzan, R., & Karner, W., 2009. Network-Wide Measurements of TCP RTT in 3G. In: *Proceedings of the 1st International Workshop on Traffic Monitoring and Analysis (TMA)*, 11 May 2009, Aachen, Germany.

Sen, S., Spatscheck, O. & Wang, D., 2004. Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures. In: *Proceedings of the 13th international conference on World Wide Web (WWW)*, pp. 512-521, May 2004, New York, NY, US.

Smith, F. D., Campos, F. H., Jeffay, K. & Ott, D., 2001. What TCP/IP Protocol Header Can Tell Us About the Web. *ACM SIGMETRICS Performance Evaluation Review*, 29(1), June 2001.

Smith, C. & Grundl, P., 2002. *Know Your Enemy: Passive Fingerprinting*. [Online] Honeynet project, 4 March 2002. Available at: http://old.honeynet.org/papers/finger/ [Accessed 14 july 2009].

Smura, T., Kivi, A. & Töyli, J., 2009. A Framework for Analysing the Usage of Mobile Services. Accepted for publication at *INFO - The journal of policy, regulation and strategy for telecommunications, information and media*, 2009.

Svoboda, P., Ricciato, F., Hasenleithner, E. & Pilz, R., 2006. Composition of GPRS/UMTS traffic: snapshots from a live network. In: *4th Internationall Workshop on Internet Performance, Simulation , Monitoring and Measurement (IPS-MOME),* 27-28 February 2006, Salzburg, Austria.

Svoboda, P., Ricciato, F. & Rupp, M., 2007a. Bottleneck Footprints in TCP over Mobile Internet Accesses. *Communication Letters*, 11(11), pp. 839–841, 2007.

Svoboda, P., Karner, W. & Rupp, M., 2007b. Modeling E-Mail Traffic for 3G Mobile Networks. In: *The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 3-7 September 2007, Athens, Greece.

Svoboda, 2008. *Measurement and Modelling of Internet Traffic over 2.5 and 3G Cellular Core Networks*. Doctoral Dissertation. Faculty of Electrical Engineering and IT, Vienna Univeristy of Technology.

Svoboda, et al., 2009. Detection and Tracking of Skype by Exploiting Cross Layer Information in a Live 3G Network. In: *Proceedings of the Traffic Monitoring and Analysis Workshop*, March 2009, Aachen, Germany.

Tallberg, M., Hämmäinen, H., Töyli, J., Kamppari, S. & Kivi, A., 2007. Impacts of handset bundling on mobile data usage: The case of Finland. *Telecommunications Policy*, 31(10-11), pp. 648-659, November-December 2007, Elsevier.

Vacirca, F., Ricciato, F. & Pilz, R., 2005. Large-Scale RTTMeasurements from an Operational UMTS/GPRS Network. In: 1st *International Conference on Wireless Internet (WICON'05)*, July 2005, Budapest, Hungary.

Vacirca, F., Ziegler, T. & Hasenleithner, E., 2006. An algorithm to detect TCP spurious timeouts and its application to operational UMTS/GPRS networks. *Elsevier Computer Networks*, 50(16), pp. 2981–3001, November 2006.

Viipuri, T., 2004. *Traffic Analysis and Modeling of IP Core Networks*. Master's Thesis. Department of Electrical and Communications Engineering, Helsinki University of Technology.

Williamson, C., 2001. Internet traffic measurement. *IEEE Internet Computing*, 5(6), pp. 70–74, November 2001.

Zhang, Y., Breslau, L., Paxson, V. & Shenker, S., 2002. On the Characteristics and Origins of Internet Fow Rates. In: *Proceedings of ACM SIGCOMM*, August 2002, Pittsburgh, US.