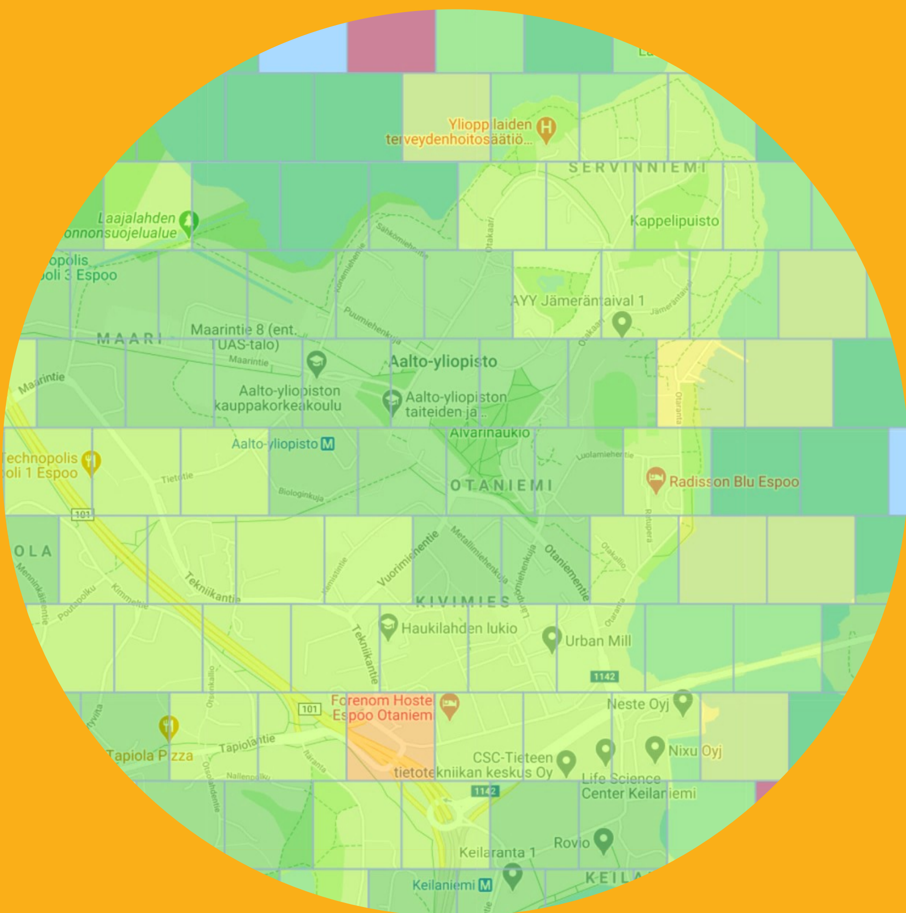
Department of Communications and Networking

A hybrid approach to quality measurements in mobile networks

Eren Boz



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 147/2020

A hybrid approach to quality measurements in mobile networks

Eren Boz

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall AS1 of the school on 28 October 2020 at 12.

Aalto University
School of Electrical Engineering
Department of Communications and Networking

Supervising professor
Professor Jukka Manner, Aalto University, Finland

Thesis advisor
Professor Jukka Manner, Aalto University, Finland

Preliminary examiners
Professor Peter Reichl, University of Vienna, Austria
Assistant Professor Aaron Ding, Delft University of Technology, Netherlands

Opponent
Professor Jari Porras, LUT University, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 147/2020

ISBN 978-952-64-0058-7 (printed)
ISBN 978-952-64-0059-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0059-4

Unigrafia Oy
Helsinki 2020

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author
Eren Boz

Name of the doctoral dissertation
A hybrid approach to quality measurements in mobile networks

Publisher School of Electrical Engineering

Unit Department of Communications and Networking

Series Aalto University publication series DOCTORAL DISSERTATIONS 147/2020

Field of research Networking technology

Manuscript submitted 25 March 2020 **Date of the defence** 28 October 2020

Permission for public defence granted (date) 18 August 2020 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

The mobile-first era is here. Due to constantly increasing demand for mobile data, cellular network infrastructures running on limited radio spectrum are struggling to keep up. Constant monitoring and measurements are necessary to ensure service availability and quality as saturated networks are not able to deliver consistent experience. Such measurements are especially crucial for mission-critical use cases such as public safety as they increasingly rely on commercial networks. However, measuring mobile networks at scale bears some fundamental problems. Although there are significant improvements in capabilities of mobile networks (e.g. bit rate, latency), measuring them is still rather complicated task compared to fixed networks given that in mobile networks, performance is a result of complex interaction between momentary cell load, adjacent cell interference, shadowing, fading, mobility and user device capabilities. The adoption of commercial 5G networks is expected to increase the variability even more as it depends on smaller cells. Active measurements that inject large amounts of traffic into the network for the sole purpose of measuring are costly in terms of both bandwidth and energy. Passive mechanisms are lightweight but miss the information of why a certain bit rate is received or sent by the end device. They can not tell whether the performance bottleneck is in the network or in the service itself.

On the other hand, modern mobile platforms provide a great diversity of networked applications that have varying network resource demands. It is often hard to estimate which aspect of mobile network performance is relevant to the quality that users experience. For example, just providing an excess of downlink bandwidth at the expense of packet loss or latency might not cut it for highly interactive real-time applications.

By combining active and passive measurements in a novel way, this work focuses on a hybrid measurement approach to quality measurements in mobile networks. First, an efficient and scalable hybrid methodology is proposed and evaluated for mobile network Quality-of-Service measurements. Then building on top of it, mobile Quality-of-Experience and its predictability in the field via smartphones is empirically investigated by carrying out extensive field studies. The thesis concludes by evaluating the findings to establish future work necessary to achieve pervasive mobile measurements that are capable of both reflecting user-perceived quality and predicting mobile performance as an enabler of network adaptive applications.

Keywords mobile networks, performance measurement, quality of service, quality of experience

ISBN (printed) 978-952-64-0058-7 **ISBN (pdf)** 978-952-64-0059-4

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2020

Pages 145 **urn** http://urn.fi/URN:ISBN:978-952-64-0059-4

Preface

Firstly, I'd like to thank my supervisor Professor Jukka Manner for providing me a perfectly balanced environment in which I got to do both research and the innovation on an interesting set of problems that are presented in this work. Without such an environment and his continuous support, I would potentially never undertake doctoral studies and as a result, I'd miss on deeply invaluable skills and experience I gained during the process.

I also would like to thank Professor Antti Oulasvirta for providing me additional funding (via *EMERGENT* project) and extending my horizon in understanding the human aspect of quality that is Quality-of-Experience.

Then, I have to express my deepest gratitude towards my co-authors Benjamin Finley and Lennart Schulte for their collaboration and persistence in pulling the research and papers forward, especially during the times I was extra busy with implementations and maintenance of our data pipelines. It was an honor and absolute pleasure to work with such hardworking and competent researchers.

Most naturally, the next big thanks goes to Netradar/Nettitutka team: Arttu Tervo, Jukka-Pekka Virtanen, Sebastian Sonntag, Antti Jaakkola, and Tommi Tuura. Together, we got to build a lot of interesting features onto the platform and enjoy it at the same time. I learned a lot from them and I will cherish our time together in the project forever.

During my tenure in the department, I have met many people, had many interesting collaborations and conversations. I particularly would like to thank Heikki Hämmäinen, Kalevi Kilkki, Timo Kiravuo, Le Wang, Nuutti Varis, and Gautam Moktan for humoring my various curiosities.

Also, I especially need to thank Suphi Özgür Cankurt, Hüseyin Yiğitler, and Müge Tetik for being the Finnish branch of my support system. And I have to acknowledge Colm Mc Caffrey for much needed refreshing adventures we had together in Japan during our research exchange. It was the most peculiar coincidence that we went to Tokyo on separate exchange programs, right after being teammates in Bit Bang course.

As I moved onto the industry before finalizing this work, I also would like to acknowledge and thank my manager Eliina Vornanen for accommodating various leaves that helped me finish.

Lastly, I am endlessly grateful to my mother for putting up with a lousy son who rarely visits. None of this would be even remotely possible if it wasn't for your tireless efforts and patience.

Espoo, September 20, 2020,

Eren Boz

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I L. Schulte, E. Boz, N. Varis, and J. Manner. On detecting TCP path saturation in LTE networks. *Wiley International Journal of Communication Systems (IJCS)*, 30(16):e3334, April 2017.

II E. Boz and J. Manner. A hybrid approach to QoS measurements in cellular networks. *Elsevier Computer Networks*, 172:107158, May 2020.

III B. Finley, E. Boz, K. Kilkki, J. Manner, A. Oulasvirta, and H. Hämmäinen. Does Network Quality Matter? A Field Study of Mobile User Satisfaction. *Elsevier Pervasive and Mobile Computing*, 39:80-99, August 2017.

IV E. Boz, B. Finley, A. Oulasvirta, K. Kilkki, and J. Manner. Mobile QoE prediction in the field. *Elsevier Pervasive and Mobile Computing*, 59:101039, October 2019.

Author's Contribution

Publication I: "On detecting TCP path saturation in LTE networks"

This article studies the efficiency of TCP in mobile networks. The author's contribution was covering delay characteristics in LTE radio access networks which included literature review, detailed report and tabulation of delay specifications as the basis of the methodology. Additionally, the author carried out extensive analyses on parameter search for the resultant algorithm. The scientific writing, editing, and styling were mainly done by Schulte and reviewed by the author.

Publication II: "A hybrid approach to QoS measurements in cellular networks"

This paper presents modeling, implementation, and empirical evaluation of a novel probe-based measurement methodology for mobile networks using smartphones. The author did all the design, modeling, software implementations, and analyses. The paper was written by the author and reviewed by Manner.

Publication III: "Does Network Quality Matter? A Field Study of Mobile User Satisfaction"

This paper investigates mobile user satisfaction via an active network measurement tool and survey questionnaire. The survey and study were designed jointly. The author implemented the measurement tool, survey, and data collection set-up; and participated in data analysis. The author focused on writing the sections describing network measurement methodologies and features as well as QoS-QoE mapping aspects. The writing, editing, and styling of the paper were led by Finley and reviewed by the author.

Publication IV: "Mobile QoE prediction in the field"

This article studies the predictability of QoE in the field via hybrid network measurements. The author designed and implemented a panel application based on novel measurement methodology. The experience sampling was designed jointly and validated by the author. The data analyses were jointly done by the author and Finley. The author focused on writing the sections describing methodology, algorithms, and features as well as their interactions. The writing, editing, and styling of the paper were led by Finley and reviewed by the author.

List of Abbreviations

2G	2nd Generation
3G	3rd Generation
4G	4th Generation
5G	5th Generation
API	Application Programming Interface
AQM	Active Queue Management
BO	Buffer Occupancy
CDN	Content Distribution Networks
DNS	Domain Name System
EDGE	Enhanced Data GSM Environment
FTP	File Transfer Protocol
GPRS	General Packet Radio Service
HSDPA	High Speed Downlink Packet Access
HSPA+	Evolved High Speed Downlink Packet Access
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
IBA	Index Balanced Accuracy

IETF	Internet Engineering Task Force
IP	Internet Protocol
IPv4	Internet Protocol Version 4
LTE	Long Term Evolution
MTU	Maximum Transmission Unit
NIC	Network Interface Card
OS	Operating System
OWD	One Way Delay
PaaS	Plaftorm as a Service
PGM	Probe Gap Model
PRM	Probe Rate Model
QoE	Quality of Experience
QoS	Quality of Service
RF	Random Forest
RLC-AM	Radio Link Control - Acknowledged Mode
RRC	Radio Resource Control
RTT	Round-Trip Time
SlopS	Self-loading Periodic Streams
SNMP	Simple Network Management Protocol
SVM	Support Vector Machine
TCP	Transport Control Protocol
TOPP	Trains of Packet Pairs
UDP	User Datagram Protocol
VoIP	Voice over IP
VPN	Virtual Private Network

WCDMA	Wideband Code Division Multiple Access

1. Introduction

Nowadays, it is quite uncommon to come across a person who does not own a connected device. Personal computers, tablets, smartphones are everywhere. Even watches are going smart along with other types of wearables. Every single one of them is connected, one way or the other, changing how we live our daily lives, solve basic problems and communicate. According to Ericsson, as of 2018, the number of smartphone mobile subscriptions has reached 5.1 billion. Projections show that this number will grow to 7.2 billion as of 2024. Not only the number of users increase but also demand per user is also growing considerably. The mobile data traffic has grown by 82% in 2018. Mobile data traffic is expected to grow 3.5x between 2018 and 2021 [25].

Increased mobile access to the internet, allowed us to free ourselves from offices, buildings, even crowded cities altogether. It even enabled a considerable number of workforce to function fully remotely as telecommuters [74]. Telecommuting is only a business-critical example case where a certain level of service is required. As the variety and the appeal of the content on the internet increases, they replace old fashioned habits such as watching TV or listening to the radio, because these modern services have the edge of personalization and on-demand nature.

We are living in a mobile-first world now both in computing and telecommunications. People want to be able to do anything and keep up with everything on the run. Such transformation is stressing the current commercial mobile networks to their limits and hiking the expectations from next-generation networks. Mobile networks, by nature, are heterogeneous and highly dynamic. They exhibit complex behaviors depending on network load, mobility, radio environment, and traffic patterns required by a variety of applications. Consequently, in practice, mobile internet performance is quite unpredictable due to a variety of constantly changing factors. Mobile internet measurements help us to characterize these variations and provide useful insights on the predictability of their performance.

On the other hand, network performance is only one pillar in the overall mobile experience user gets because mobile platforms feature a diversity of applications which might demand different levels of network resources at various phases of the user session. For instance, such demands might range from strictly low latency, no packet loss for a Voice over IP (VoIP) call to high bandwidth with moderate latency for over the top video streaming. Yet we might not be able to anticipate the amount of degradation in mobile experience if the network fails to deliver it at a random time. One practical example would be the initial buffering time for video applications. If the network particularly fails to provide a reasonable amount of bandwidth at the beginning of a video viewing session, the user perceives significantly longer delays as opposed to a bandwidth reduction later. Consequently, at the end of the day, it is equally important to understand the demand characteristics of particular applications to be able to approximate user experienced quality. For the purpose, practical models that investigate the relationship between mobile network performance, applications and user expectations are essential.

1.1 Motivation

To study the performance of mobile networks, there exists a wide range of commercial tools, crowdsourcing based services, and mobile apps [27]. The common aspect of these solutions is that they all provide some sort of on-demand active measurements with more or less fixed methodologies. Prescribed fixed methodologies tend to generate the same pattern of traffic and measure only a certain subset of network behavior. Let us imagine a smartphone traveling at a speed of 60km/h inside a bus, halfway done downloading a file, it is quite a dare to say it would download the other half of the file in a similar amount of time. It is even hard to say if the file download will ever be completed at some point in time. Simply, the bus might drive the phone away from the coverage.

The point is that in a highly variable and complex environment, a few numbers of measurement data points will not tell us much about the network. To achieve a statistical significance upon the measurements with respect to free variables affecting the resultant network performance, a great number of measurements should be drawn from varying scenarios and locations along the route of that particular bus. The question is whether these tools can collect that much measurement data to explain the expected behavior of a commercial cellular network in a crowded city. Let alone the number but, it is unknown if typical situations where measurements are done by using an active tool can be used to predict the actual usage scenarios at large. Consequently, it is difficult to claim that the measurements derived from these active network measurement tools depict the whole picture. On the other hand, the traffic injected into the network for bandwidth measurements introduces a considerably large amount of overheads both in terms of bandwidth and energy.

Passive methods are more lightweight on the network and they can provide information about actually experienced network performance. However, in practice, they are harsh on computational resources as they require extensive packet processing to obtain deeper information about underlying applications that originates the traffic. Also often such level of detailed information is not available in mobile platforms, practically rendering it unusable for crowdsourcing purposes. Finally, without application-level protocol information, passive methods often fail to detect the particular bottleneck in the overall performance. They can not pinpoint whether it is a slow application server or poor network connectivity that causes degradation.

1.2 Research questions and methodology

This work essentially divides the research problem into two distinct parts. The first part is about measuring the in-context mobile network quality out in the field in terms of objective Quality of Service (QoS) metrics. The second part is focusing in particular how these metrics relate to user-perceived quality, namely Quality of Experience (QoE).

In the first part, we review existing mobile network measurement methodologies commonly used in crowdsourced measurement platforms to identify their shortcomings. Then we investigate probe-based measurements and their applicability to mobile networks to lay the foundation of a better mobile measurement solution. We propose, implement and evaluate a hybrid measurement methodology for smartphones that solves the addressed shortcomings.

In the second part, we carry out extensive user studies (in Finland) to investigate factors influencing mobile user experience via custom smartphone applications. First, we carry out a survey study analyzing traditional active network measurements to get an initial understanding. Then we design a more detailed panel study by implementing proposed hybrid methodology as well as collecting detailed metrics available in Android smartphone platform. We also employ an exit survey to collect detailed demographic information from participants.

In summary, the following questions are asked in this work:

- How quality of mobile networks can be measured accurately with as little resource usage (e.g. time, bandwidth, energy) as possible?
- What are the measurable factors influencing mobile smartphone user experience in the field and how predictable is application QoE based on such factors?

And then the findings of the first question are used to help answer the second question. Hence we present a holistic quality measurement concept.

1.3 Contribution

The overall motivation behind this work is about ensuring excellence in mobile experience. Finland is proudly known to provide great mobile access for majority of its citizens. The author was part of the team behind Nettitutka [54], a Finnish smartphone-based crowdsourced mobile measurement platform utilizing traditional methodologies to map mobile network quality all over Finland. The author has noticed that the rate of incoming measurements was not significant enough to provide reliable information per location in modern mobile networks that are highly dynamic and heterogeneous. Consequently, the author took up the challenge to come up with a novel solution that can generate a significant amount of quality measurements that would help the mission of excellence.

Our contribution is split into 4 publications that focus on different parts of the problem. In Publication I, in-depth analysis of packet delay, buffering and Transmission Control Protocol (TCP) performance characteristics in mobile networks have been investigated in light of Nettitutka measurements. Such analysis allowed us to gather great insight on how to utilize probe-based approaches on modern mobile networks which led us to work on hybrid measurements. In Publication II, specifics of probe-based approach are laid out to come up with a novel methodology that's cost-efficient, scalable and accurate. This approach was implemented and empirically evaluated in the field to provide proof of concept. A hybrid approach is described and shown in detail that it is superior in measuring experienced network quality by the user, unlike traditional active measurement approaches.

Secondly, Publication III revealed how user-triggered measurements relate to self-reported experiences of Finnish mobile users. The results have shown that a certain subset of networking characteristics is more dominant than others while some non-network factors have even more impact on user satisfaction. This study also has shown that people tend to interpret measurement results they see and there is varying measuring behavior among users.

Finally, in Publication IV, utilizing the outcome of previous research, the author implemented a user panel to investigate the predictability of per application perceived quality. This paper presents the findings and challenges of QoE predictions in the field using previously proposed hybrid measurements and smartphone platform provided data.

1.4 Thesis structure

From this point on the thesis structured as follows. Each of the chapters includes material and essential findings from one or more of the publications.

Chapter 2 starts by discussing mobile network evolution. Then, relevant aspects of internet measurement methodologies are briefly covered while an entire section is dedicated to present TCP performance issues in mobile networks and our findings. This chapter provides the conceptual background of the problems that we work within the thesis. Related work is also covered in this chapter. Chapter 3 describes the modeling, implementation and empirical evaluation of a novel hybrid QoS measurement methodology for mobile networks using smartphones. Chapter 4 presents QoE aspect of this work in which predictability of mobile experience is investigated in depth. First, we longitudinally study the concept of user satisfaction in mobile networks by looking at the quality factors affecting satisfaction in a longer time window via active measurements. In the second study, learning from the first study is utilized for designing a shorter time window application-level QoE study based on our novel hybrid approach.

Finally, Chapter 5 starts by summarizing the findings concerning research questions, then moves on to practical implications of the contribution in this work. In the end, future work is discussed to cover open questions and research vision.

2. Mobile internet measurements

This chapter starts by discussing the mobile network evolution. Then, relevant aspects of internet measurement methodologies are briefly covered while an entire section is dedicated to present TCP performance issues in mobile networks and our findings. In the remainder of the chapter, background topics and issues ranging from crowdsourcing, probe-based QoS measurements to QoE are discussed. Lastly, related work is reviewed and examined. This chapter is mostly based on Publication I, Publication II and Publication IV.

2.1 Mobile network evolution

In 2nd generation (2G) mobile networks, data transfer capabilities were very rudimentary as they were optimized for circuit-switched services such as voice calls. Only with the deployment of General Packet Radio Service (GPRS), commercial networks started to support packet-switched internet access with very modest user data rates of 20 Kbps [51, 42]. Within the same generation, Enhanced Data GSM Environment (EDGE) improved the rates to the order of several 100 Kbps. Yet these early technologies were also featuring significantly high radio access delays of 400-700 milliseconds [8]. Only after the adoption of advanced 3G technologies (also dubbed as "3.5G") such as High Speed Packet Access (HSDPA) user data rates and latencies reached plausible levels of 10+ Mbps and 100- milliseconds respectively.

Even though the mobile data rates were closing up to typical levels that commercial fixed internet services offered, mobile networks still had peculiarities compared to fixed networks. In order to overcome challenges that arise from mobility and radio communications, mobile network technologies utilize complex methods for optimizing both system capacity and user experience together. They strive to conserve scarce radio and spectrum resources by using methods such as multi-radio access technologies, scalable communication channels, scheduling, RRC State machines, discontinuous reception, and transmission, etc. As a result of the interplay of such methods along with mobility, mobile data services tend to create a rather different experience compared to fixed networks. Often mobile users experience additional delays for establishing a connection, slower and highly variable data rates. However, such experiential difference has started to slowly vanish with the wide adoption of 4G technologies as in the case of LTE [78]. Increased reliability in the radio access layer brought latencies down to the level of 10 milliseconds and user experiences up and closer to fixed networks.

Regardless of the convergence in capabilities of fixed and mobile networks, measuring mobile networks is still rather complicated given the nature of wireless communications. The resultant performance of a connection in mobile networks is a complex interaction between momentary cell load, adjacent cell interference, shadowing, fading, mobility and user device capabilities. Along with the upcoming adoption of 5G networks, heterogeneity and variability in radio access are expected to even grow further since improvements heavily depend on elements such as smaller cells, fragile millimeter-wave bands, etc. Consequently, complete quality mapping requires frequent measurements with both time and space diversity, which can only be managed by end-user device based measurements schemes.

2.2 Measurement methodologies

Measurement methodologies can be classified into two basic categories as active measurements and passive measurements [59, 17]. The measurement methods which actively inject data to the network it is measuring are classified as active measurements. Depending on the metric which is being measured, active measurements usually require co-operation from both ends, in end-to-end measurement scenarios. Metrics such as latency, jitter, throughput and packet loss can be measured easily using active measurement methods. The availability of a wide variety of tools, protocols, and accuracy is the biggest advantage of active measurement methods. Unfortunately, it comes at the cost of network capacity usage. Especially in case of throughput measurements, the network capacity and service is blocked by the tool itself, rendering user unable to communicate for the duration of the measurement. Also considering that bandwidth costs money in certain situations, sometimes it is more beneficial not to measure, e.g. mobile data plans with high peak data rates and low data quotas. Another downside is that if the tool is not aware of the cross-traffic flows, it might impair the accuracy of the measurement results based on the additional network usage during the measurements.

In contrast to active measurements, passive measurements do not introduce data traffic to the network it is measuring. It is non-intrusive by nature. The methods of passive measurements involve capturing traffic generated by other users or applications. The process of capturing does not necessarily include all the bits that go through the point of measurement, but can be and should be limited to only relevant information such as packet traces including protocol headers that is of interest. The amount of information captured simply can be as little as the sampling of counters of incoming and outgoing bytes. One of the most used passive measurements is probably the collection of traffic statistics per network interface in network devices, as it is simplified by Simple Network Management Protocol (SNMP) [15] for accounting and managerial purposes. Another interesting example is measuring latency using only passive measurement by having synchronized captures from various points in the network to approximate the delay experienced by individual packets, in a certain path.

Passive measurements might require knowledge of individual protocols depending on the measured metric. For instance, a latency measurement can be approximated knowing that a three-way handshake of a TCP connection effectively involves two round-trip delays (of two distinct origins) between the particular server and the client. Similarly for the throughput, if a TCP flow is known to transfer a rather large bulk of data, e.g. a file download using File Transfer Protocol (FTP) [7] or Hypertext Transfer Protocol (HTTP) [29]; the stable speed for the connection can be effectively treated as TCP throughput from the server to client, given that bottleneck is the network between, not the computation required for file serving in the server system. The disadvantage comes from the fact that the measurer does not have any control over the passively observed traffic, so it is impossible to measure a certain subset of parameters if no suitable traffic happens. Consequently, the timing of the measurement can not be controlled.

A measurement method does not have to be purely active or passive. Sometimes, it is more efficient to carry out measurements comprising both active and passive elements to approximate a metric. Crovella and Krishnamurty call it fused measurements [17]. RFC 7799 by IETF names it as hybrid measurements [53] where a hybrid measurement is defined as the combination of metrics derived from passive and active measurements to produce a measurement result. The example for a hybrid measurement could be measuring one-way delay to a set of endpoints from passive measurements in which there is a frequent traffic exchange and it can be complemented with active measurements from endpoints with less frequent traffic to generate a delay map. That is to say, active measurements can be used to remedy the disadvantages of passive measurements. Furthermore, hybrid methods are classified into two categories based on how many distinct streams of interest are involved. Methods that utilize a single stream of interest are dubbed as Hybrid Type I. However if it involves multiple distinct streams of information, it falls into category of Hybrid Type II. In this work, we motivate and design Hybrid Type II methods which involve many different streams of information.

2.3 Transmission Control Protocol in mobile networks

TCP is the standard protocol of reliable data transport on the internet and it carries the majority of internet traffic in mobile data as well. Naturally,

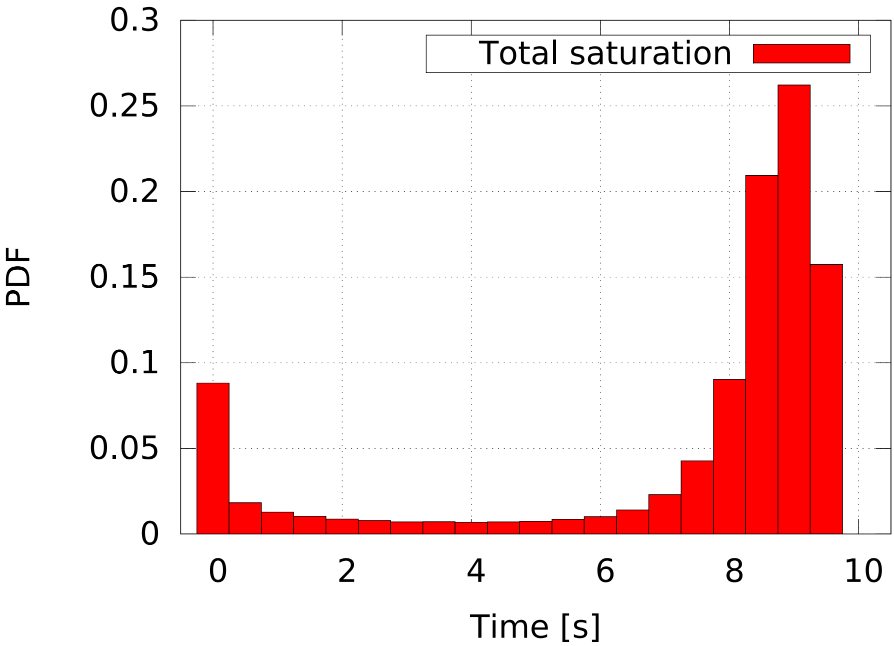


Figure 2.1. PDF for time spent saturated (of max. 10 seconds measurement). From Pub. I.

it is an indispensable part of the internet and quality measurements as well. However, it is widely known that there are performance issues with TCP in mobile networks, and many papers state possible reasons, including variable bottleneck bandwidth, excessive buffering, high bandwidth-delay product, high network delay and packet loss unrelated to congestion [84, 88, 39]. But end-to-end measurement studies, especially in deployed mobile networks, are mostly missing. So the question is: are TCP algorithms responsible for the poor performance in mobile networks or do other reasons prevail, such as the mobile device, network issues, or just network capacity? We carry out our measurement study on Finnish commercial Long Term Evolution (LTE) networks and analyze suboptimal performance by separating TCP performance issues from other factors.

For gathering the data we use the crowdsourced measurement platform called Nettitutka (formerly Netradar) [54, 76, 75]. The app measures against the Nettitutka servers on the internet. The main data that we are using in this study is the 10-second full bulk TCP download of random data. On the measurement servers, we record packet traces, from which we extract round-trip time (RTT) and throughput samples, as well as TCP related events.

We also find that even in modern LTE networks, TCP performs poorly on various occasions and is unable to saturate network bandwidth. Figure 2.1 shows the degree of saturation among the connections in our dataset. We see that a huge part achieves a saturation of over 8 seconds. But it is also visible that there are about 20% of connections that have an even lower saturation,

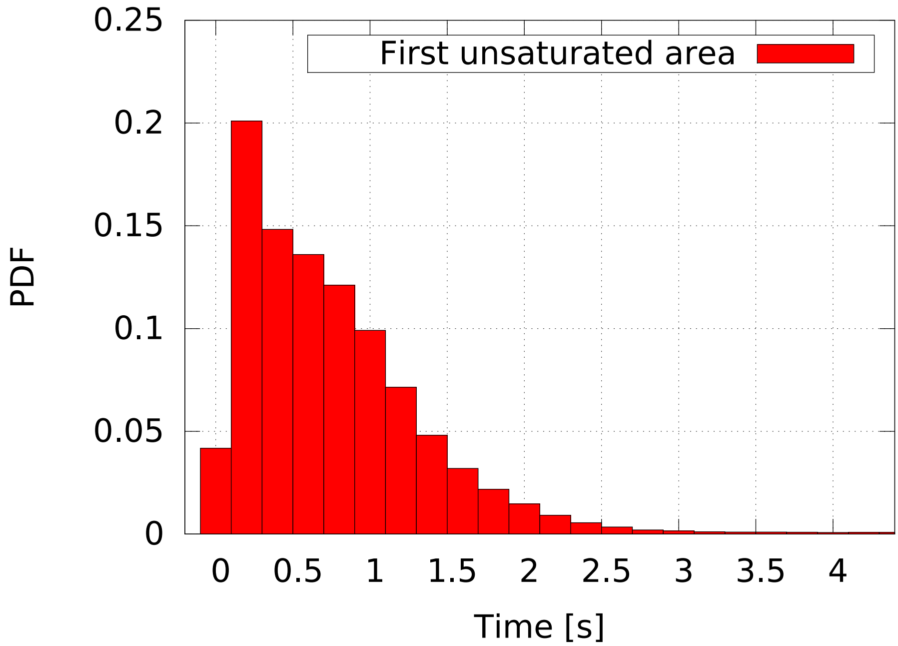


Figure 2.2. Time until first saturated. From Pub. I.

and that the pinnacle is at 9 seconds which means that most connections leave 10% of the connection time with unused network resources. Surprising is the 8% of connections which have a saturation time of 0 seconds, i.e. they are never saturated. Analyzing these connections shows that the culprit is a high number of recovery due to lost packets while the RTT is not visibly increased. While early iterations of mobile networks were affected by data loss on the wireless link, the reliability of LTE networks nowadays is almost as good as in wired networks, as "Radio Link Control - Acknowledged Mode (RLC-AM) serves the important purpose of ensuring (almost) error-free data delivery to TCP" [19]. Although the losses might still be due to an unreliable wireless link, another possibility is a misconfigured queueing system. Furthermore, there are system-level dynamics such as horizontal and vertical handovers that can cause losses as well.

At the beginning of each connection, TCP performs slow start to probe the network. Hence, each connection starts unsaturated. In theory, even with high network data rates, this should not take more than a few hundred milliseconds. Figure 2.2 shows the time until the first saturation, which is what the slow start algorithm tries to achieve. We see that many connections are spending significantly more time than is to be assumed.

Overall, we have found several situations where TCP behaved suboptimally: ramping up the sending rate during slow start, problems with deployed client-side optimizations for bufferbloat mitigation, and changes in sending rate so that TCP either reacts too slow or is forced into a recovery. Additionally, on occasion, we observed TCP was not able to track the sudden increases in network bandwidth.

Well-known bufferbloat problem in mobile networks particularly creates a handicap for the case of utilizing a single TCP connection as an indicator of network quality, since it's shown that competing network flows affect TCP performance characteristics severely [43, 3]. Such findings suggest that major crowdsourced measurement services (e.g. Speedtest.net [56], Rootmetrics [66]) that utilize TCP as its underlying transport can be error-prone and biased in networks that have issues discussed above. Consequently, having measurement methodologies directly measuring bandwidth of the mobile link is particularly useful to mitigate the situation.

2.4 Probe based measurements

Available bandwidth and capacity estimation on a path have beneficial uses such as selecting better routes on an overlay network, QoS verification, and traffic engineering. In literature [60], various probe-based methodologies are proposed to measure capacity or available bandwidth in different fixed network settings such as single-hop or multi-hop. Essentially these methods can be classified into two main groups as probe rate model and probe gap model [77], based on the approach of measurement.

2.4.1 Probe gap model

The probe gap model (PGM) makes use of the time difference between the arrival of two successive probes at the receiver. The model denotes the time difference of consecutive probes sent in the sender side as Δ_{in} and the time difference of reception of those at the receiver side as Δ_{out}. Along with the assumption of single bottleneck (tight link) along the path of measurement, the relation of Δ_{in} and Δ_{out} constitutes a measure of congestion in the tight link as depicted in Figure 2.3, $\Delta_{out} - \Delta_{in}$ is the time it takes to transmit the cross traffic at the tight link.

Furthermore, if the capacity information of the tight link is available, one can estimate the available bandwidth A using Equation 2.1.

$$A = C \times (1 - \frac{\Delta_{out} - \Delta_{in}}{\Delta_{in}}) \tag{2.1}$$

In a situation with a constant bit rate cross traffic, a small number of probe pairs would yield a plausible estimate by averaging samples calculated by the equation. However, cross-traffic rarely occurs at a constant bit rate. It is also important to note that consecutive probe packets might arrive back to back, even out of order that might generate erroneous and negative samples at times. Naturally, those cases might need to be handled separately to improve estimates. Moreover, most of the cross-traffic on the internet is bursty, so the real challenge for the PGM is to adapt the sending process of probe packets.

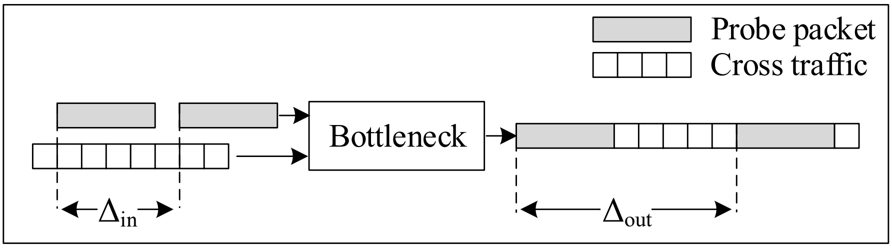


Figure 2.3. Probe Gap Model (PGM) illustration. From Pub. II.

Hence, PGM tools adaptively send several packet pair trains, which has many probes spaced variably to generate a rather unbiased estimation. Spruce [77], Initial-Gap-Increasing [36] and Delphi [65] are example tools for estimating available bandwidth, which tries to overcome these challenges.

PGM tools typically have the advantage of low bandwidth usage to produce a measurement, but Lao et al. [46] shows that PGM can underestimate the available bandwidth of multi-hop paths even if there is a single tight link. Additionally, even though a single tight link assumption is plausible for most of the internet paths, the requirement for the knowledge of the capacity of the tight link is impractical. In conclusion, PGM tools fail to provide the desired reliability and practicality for most applications.

2.4.2 Probe rate model

Unlike the probe gap model, probe rate model (PRM) relies on the self-induced congestion on the network path to measure available bandwidth. Congestion on the tight link is going to cause the following packets on the path to experience additional queuing delays. Hence, injecting more traffic at a higher rate than available bandwidth as depicted by Figure 2.4, results in an increasing trend in one way delay (OWD). PRM tools use this basic premise to measure available bandwidth.

Self-loading periodic streams (SlopS) methodology is a typical instance of the probe rate model [41]. In SlopS methodology, the source sends a fixed number of equal-size packets to the destination at a certain rate R. The receiver side monitors and reports the OWD trend back to the sender, so that the sender can statistically evaluate the trend and adjust the sending rate R accordingly to estimate available bandwidth A. The adjustment is similar to a binary search if OWD is increasing, hence $R > A$, the rate R is decreased for the next iteration, if OWD is non-increasing in the case of $R < A$, R is increased. Once two successive sending rates are close to each other to a degree of ϵ, the algorithm estimates $A = R$. This type of approach requires a considerable number of iterations, especially in the case of highly varying available bandwidth.

Melander et al. [48] proposed a slightly different methodology than self-loading periodic streams. The proposed methodology is built around the idea of trains of packet pairs (TOPP). In TOPP, the sender sends many packet pairs at gradually

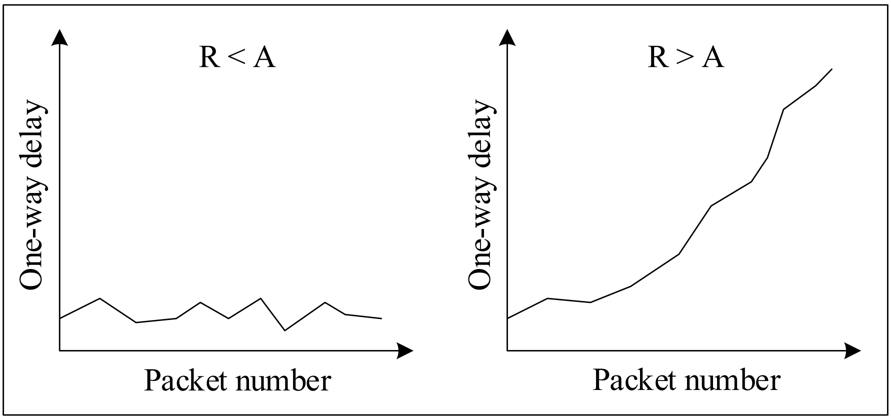


Figure 2.4. Probe Rate Model (PRM): the relation between one way delays and available bandwidth. From Pub. II.

increasing rates, in other words gradually decreasing inter-packet interval Δ_S. The offered packet rate at a packet pair, R_o is equal to L/Δ_S where L is the packet size. If $R_o > A$, the receive rate R_m is expected to be equal to R_o. Consequently, TOPP estimates available bandwidth as the maximum R_o where $R_o \approx R_m$.

Additionally, TOPP is able to estimate capacity of the narrow link in the path using Equation 2.2 where R_c is the rate of cross traffic. The capacity C can be estimated from the slope of R_o/R_m vs R_o.

$$R_m = \frac{R_o}{R_o + R_c} \times C \tag{2.2}$$

Even though TOPP has the benefit of multi-rate probing, it loses the temporal behavior of queuing caused by bursty traffic due to rather large spacing between the probe pairs.

2.4.3 Challenges in mobile networks

For practical purposes, it is beneficial to concentrate and summarize the reasons why probe-based measurements are challenging in mobile network environment given the discussions up until this point. The common feature of all probe-based measurements is that they use the delays experienced by the probe to estimate the available bandwidth in the communication path. The assumption for this is that the change in the delay is only caused by the change in the available bandwidth. This assumption is broken in mobile networks because the delay experienced by probe packets can be chiefly affected by interruptions caused by all sorts of handovers be it vertical or horizontal as well as radio resource control (RRC) state changes. Additionally, varying radio conditions and signal quality can change packet loss characteristics of the channel and finally probe packets might be retransmitted which in return adds a multiple of retransmission base delay. Finally, in a loaded network along with the varying radio conditions, scheduling at the base station can immensely affect the delays experienced by each packet.

Secondarily, traffic patterns might affect how mobile networks behave. Moreover, the behavior dynamically might change in different operator's network, which depends on deployment and systems governing the radio resources. A tangible example is that RRC state transitions are typically triggered by what is called buffer occupancy (BO). BO, the increase in the buffered traffic at the base station, is an indicator of more traffic coming than the current service rate, hence it justifies the switch to a better RRC state and higher capacity channel. BO is also a configurable RRC parameter which is usually a value close to 1 KB [58].

Lastly, even if the cell has no load and the entire cell capacity is free for the user. It is very typical that, within the same RRC State, all radio resources for the maximum capacity is not allocated right away. Instead, parallel to BO idea, it might be gradually increased. The reason is that there might be a cost associated with deallocation and reallocation of the resources to possible incoming users. Spreading codes in WCDMA is an example of such a case.

In conclusion, for all these reasons, the delay and the capacity of the channel to a particular user systematically vary. To handle these variations, the probe methodology should be aware of the possible state of the network and ensure that samples derived from measurements belong to what is intended to be measured.

2.5 Quality of experience

In the era where mobile services were limited to a fewer number of applications such as voice calls and messaging, the notion of quality in mobile networks was largely captured by the network-centric concept of QoS. However, the usage of personal mobile devices and the diversity of mobile data-based applications in everyday life has increased dramatically over the past decade. For example, quantitatively the average US adult now spends over 1.5 hours daily on their smartphone [81]. Consequently, QoS centric network management has started to fall short especially in areas like customer experience management. Given the ubiquity, the quality of mobile experiences has become an important factor for both users (that want to maximize their own utility) and mobile network operators (that want to optimize their service offerings to increase customer satisfaction and prevent churn) [70]. Research into the quality of mobile experiences falls into the research domain of mobile QoE. A domain with roots in the areas of mobile networks.

Specifically, the mobile QoE domain encapsulates the theory, methods, and metrics for evaluating mobile services in a way that aligns with how end-users actually experience those services. For example, the use of non-linear scales, such as mean opinion score, and contextual information in QoE aligns with the non-linearity and contextual dependence of human perception systems. QoE contrasts with the quality of service domain which, nowadays, encapsulates technical and often more easily accessible network measures such as throughput, delay, and loss.

Traditionally, telephone networks were planned by utilizing purpose-built QoE models such as ITU-T Rec. G.107 [6, 52]. It is also known that, in general, QoS metrics have a logarithmic relationship to QoE due to psychophysics of human perception [63, 28]. However, for multi-modal applications such as web browsing, establishing such a relationship based on user-perceived delays or impairments is not straightforward [24]. Modern smartphone applications utilize even more diverse methods that decouple momentary network QoS from the user experience. For example, many social media applications prefetch media content on anticipation before the user is ready to consume it. Consequently, building practical QoE models as a function of network quality metrics is particularly challenging for such adaptive applications.

Additionally, in cases of longer usage sessions, the psychological temporal factors tend to creep into the dynamics of QoE. Retrospective retrieval of experience is subject to well-known primacy, recency, and peak-end temporal psychological effects [83, 32, 35, 34]. Consequently, when dealing with longer time-window measurements and perceived quality it's important to allow QoE measurements and models to incorporate such temporal factors in [72].

QoE is a common term in mobile research yet the definition of QoE has been difficult to pin down as many formulations have appeared over the years [49]. Though recent work, such as the Qualinet project, has made progress in creating an intuitive yet relatively complete definition [11]. In our studies, we take a relatively broad view of QoE that accommodates experiencing over both shorter and longer time frames to have a more complete understanding of it in various settings.

2.6 Related work

In this section, first, we go through related work on in-context mobile network measurement methodologies and then we visit relevant studies on QoE aspects and compare them to our work. In 2006, Diaz et al. [23] visions and proposes the architecture of a crowdsourced measurement system that is made very possible today by the fact that most of the end-user devices include a GPS chip. The architecture is simply composed of soft-agents installed to user terminals that run measurements at appropriate times. Measurement results along with the location, time and service type information are fed back to a server where the information is processed. Through analysis of the centralized information, the system would be able to provide real-time network diagnosis, service performance degradation monitoring, and service availability monitoring. Essentially, this type of architecture is used by many measurement services as it is the most logical and convenient one for the purpose.

Yao et al. [87] conducted an eight-month-long study to inspect bandwidth predictability in a 3G cellular network. Experimenters drove through the same 26 km route many times while doing measurements at different points of time, to account for a consistent realistic scenario of mobility. Bandwidth measurements are carried out using a fast converging probe-based measurement to achieve cost efficiency and granularity. They have found out that using past measurement data along with the location information is a good predictor for the bandwidth.

Yet on another measurement study called Mobile Internet Services Test, Wittie et al. [85] approached the issue from the perspective of mobile application developers, on the fact that underlying network characteristics directly impact the user experience of an application. Considering the fact, that the user experience and usability are the most critical aspect of a mobile application, it is important to understand the issues regarding the interaction of network and software. To characterize cellular data network performance as it is experienced by individual mobile devices, they developed a mobile application to measure interactions between network technology and a variety of mobile devices and platforms. Gember et al. [31] conducted a study to obtain in-context measurements on the premises that to quantify user experience, the relevant measurements need to be carried out during the relevant contexts. They found out that recent usage of the network due to user activity, macro-environment (e.g. indoors and stationary), micro-environment (e.g. in the user's hand) matters as the context of measurements and changes the results.

Mobiperf is also another related measurement study by Huang et al. [37]. One of the significant contributions by Mobiperf is that they have introduced an RRC State Machine inference methodology in their applications to discover parameters of network state changes. Using the parameters they were able to assess performance and energy consumption characteristics of 3G and 4G networks [61, 38]. Similarly, Garcia et al. [30] acknowledges the complexity of mobile data contexts and studies delay characteristics of commercial mobile networks on a large scale by investigating delay characteristics from network layer to TCP handshakes and flows. Mobilyzer [55] is yet another proposed mobile network measurement platform by Nikravesh et al. designed to carry out controlled and isolated in-field network measurement to reveal large scale performance of mobile networks. Related studies mentioned so far, try to address the stated problems to some extent, except the cost-efficient bandwidth measurements in mobile networks.

In literature, probe-based methodologies are proposed to achieve fast and efficient bandwidth measurements, although it is heavily studied on fixed networks, their behavior in mobile networks is rather unexplored [17, p. 130]. Moreover, several studies and surveys already make the case and portray the need for adopting hybrid measurements [50, 45, 5]. In this work, we specifically focus on modeling and building a hybrid probe-based crowdsourced network quality measurement solution.

On the other hand, there is also a set of passive measurement studies. The MopEye [86], Haystack [62], and AntMonitor [73] approaches use the device-based Android virtual private network (VPN) APIs to inspect per-app sent and received traffic and thus measure network quality. Their collected network quality metrics are similar to active metrics including delay and throughput. Unlike others, these approaches can associate network measurements to specific app traffic, though this comes at the expense of significant computational and energy overheads (due to the need to maintain an additional IP network stack and crunch a lot of data). Additionally, these approaches occupy the device's VPN slot, thus preventing users from using an actual (remote) VPN for privacy. All the studies also include characterization of network quality for a group of test or crowdsourced users. However, these characterizations do not include QoE as their focus is on presenting the approaches.

Unsurprisingly, large volumes of research exploring the relationship between QoE and mobile network performance has been published [49]. Most of these studies have tended to focus on specific use cases such as QoE of mobile video [4] or specific mobile applications [16, 69, 14, 13, 2]. Relatively few studies have examined the related area of user satisfaction with mobile network concepts such as speed or availability [57]. Furthermore, studies that have examined this area have been customer satisfaction studies based only on user questionnaires without network data [57, 44, 79]. Additionally, related studies have also typically considered only a small set of network performance features such as download goodput, jitter, and latency and have considered non-network features (such as device type or device quality) out of scope (sometimes because non-network features were unavailable) [16, 69, 14, 13]. Certain non-network features are important to include because they can affect the perception of network quality since, for example, users might incorrectly attribute device-related application performance issues to the network.

Significant research has focused on specifically modeling QoE via network QoS and/or contextual data [1, 33, 14, 12]. Furthermore, such QoE modeling research has been profitably performing field-based experiments with device-based methodologies (data is collected directly from a mobile device through an app). Similarly to prior work in this area, we also examine the relationship between QoS measurements and QoE evaluations through modeling. However, existing device-based QoE modeling research has limitations regarding the level of realism, sample size, and network quality and therefore may overestimate the practical performance of such models. Specifically, prior studies have instructed users to perform predefined application (app) actions, such as scrolling the Facebook news feed; thus the situations are only semi-realistic. Additionally, sample sizes in terms of users and sessions have typically been small (less than 50 users and 1000 sessions). Finally, several of the studies did not include sufficient sessions over LTE or Wifi connections, thus providing only a limited view given the widespread deployment of LTE and Wifi in most developed countries.

2.7 Summary

This chapter started by describing the mobile network and data capabilities evolution. Existing traditional measurement methodologies for mobile networks, that are mostly active measurements, are impractical. The convergence in quality characteristics of modern mobile and fixed networks enable us to design better methodologies for crowdsourcing mobile network measurements.

Probe-based measurement methodologies, that has been traditionally utilized in fixed networks, can be also reused in modern mobile networks. However, mobile networks still introduce additional challenges due to their complex nature and such efforts require caution in tailoring measurement methodologies.

Along with the evolution of mobile data capabilities, the use of mobile networks has diversified greatly as well. During 2G era, mobile devices were chiefly used for voice calls and messaging. Nowadays, we are living in a mobile-first world where the majority of personal computing happens in smartphones. Consequently, the nature and needs of mobile measurements have evolved beyond the classic notion of QoS, which can not capture the essence of quality. QoE has been put forth to amend the shortcomings and capture the relationship between the technical capabilities of networks along with varying in-context needs of users.

3. Hybrid approach for QoS measurements

This chapter describes the modeling, implementation and empirical evaluation of a novel probe-based QoS measurement methodology for mobile networks using smartphones. This chapter is based on Publication II.

3.1 Model

In order to devise a cost-efficient, reliable and accurate method which is capable of measuring user mobile link capacity in mobile networks with smartphones, we model a new approach to capture the best of both worlds, i.e. to be as cheap as PGM but as accurate as PRM at the same time.

The main requirement for the model is that the only single tight link is the user's access link which is the wireless link from base station to user device for cellular network case. Thus, the model measures the capacity of the tight link. The packet delay in the path from probe sender to the tight link is assumed to be near-constant and does not vary within a measurement. Also, the tight link is assumed to behave like a first-in-first-out queue per user traffic. The model requires the ability to measure the amount of cross-traffic between any two moments. All sorts of information related to cross-traffic such as packet count and packet sizes are useful, but not necessarily required.

Figure 3.1 depicts the send buffer of tight link where a measurement can be done upon reception of the whole content by the user device. $d_{rcv}[n]$ denotes

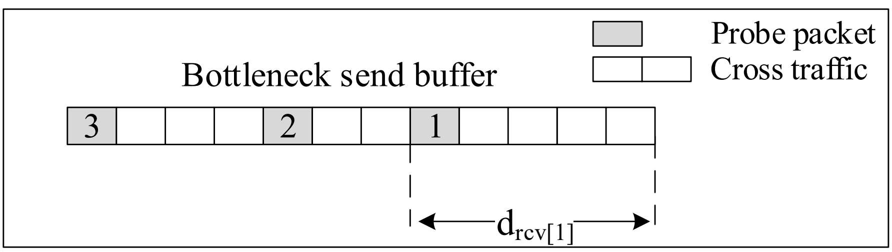


Figure 3.1. Base measurable model. From Pub. II.

queuing delay experienced by the probe number n. Let $r[n]$ be the total number of bits received by the user at $t[n]$ which is the time of reception of the probe number n. Momentary link capacity can be simply estimated by Equation 3.1 where $d_{rcv}[n] > d_{rcv}[n-1]$, $(r[n]-r[n-1]) > 0$ and $(t[n]-t[n-1]) < d_{rcv}[n]$. That is the condition for probe number n and $n-1$ to be in the same queue and had some amount of cross traffic between them. Final estimation can be calculated by merging all viable samples within a meaningful time frame. Within this model individual samples are not expected to prove accurate mainly due to packetization effects of the network and it's important to note that individual samples derived with Equation 3.1 become highly skewed for probes that arrive almost back to back causing the denominator to become infinitesimally small. However, this does not necessarily create a problem when final estimation is calculated via a weighted average of samples where the weight is the denominator (duration of the sample) in the equation. Consequently, those samples will have an insignificant effect on the final estimation provided that there are significant amount of adjacent samples.

$$C(t[n]) = \frac{r[n]-r[n-1]}{d_{rcv}[n]-d_{rcv}[n-1]} \tag{3.1}$$

However, in practice, the exact queuing delays are not known to receiver side. $d_{rcv}[n]$ can be approximated by $d[n]-min(d)$ where $d[n]$ is one way delay of probe number n and $min(d)$ is the minimum OWD which happens at the point where the packet sees the queue empty. Yet again, measurement of OWDs requires clock synchronization between the sender and receiver which is impractical. Fortunately, absolute OWDs are not necessary, relative delays can be used to approximate queuing delays of probe packets, simply because $d_{rel}[n] = d[n] + clock_{offset}$. Naturally, this is the case only if the clock skew is negligibly small.

The base measurable state only describes the minimal conditions to obtain a measurement sample. The validity of the samples is greatly impacted by the implementation environment which includes all network aspects, user device hardware/software, probe sender/receiver implementation, etc. So in short, it is essential to filter measurement samples at appropriate situations to improve the reliability of measurement results.

A simple framework for filtering measurement samples can be based on conditions enforced on measured values per sample. For example, the samples where there is no certain number of cross-traffic packets between consecutive probes can be filtered out, because it might simply be an anomaly caused by radio retransmissions. Similarly, if the probe packets received too closely in time relative to their separation at the probe sender, this situation can also be an anomaly caused by another element in the implementation stack. However, as we are aware that in packet networks the traffic is highly non-fluid. Hence, excessive filtering would result in inaccurate measurements.

One essential filter that could be utilized, is to establish a delay margin for minimal relative delays in the measurement sample. This delay margin can account for the delay variability for the reasons other than queuing in the tight link. At the same time, the margin should be small enough so that it is possible to capture measurement samples at all, especially in cases where the buffer allocated to the user at the base station is small.

The fundamental problem with passive measurements is that the measurement tool must be alert and running to observe the measurable state. To do so, the tool must continuously track the cross-traffic on a level that can detect the measurable state. Continuous tracking has an associated processing cost. Furthermore, on portable devices like smartphones, battery is a way too valuable resource to waste.

On the other hand, the measurable state, in this case, is the state where the send buffer is grown enough that the communication channel is sending the traffic at its momentary achievable rates i.e. capacity to the user. To confirm such a state, the tool must be able to decide when to initiate a probe train and then inspect the probe delays to check if the measurable state as in Figure 3.1 is available. As it is well-known that most of the internet traffic is carried via TCP connections and most of these connections are short-lived; anticipating those short moments and appropriately injecting a probe pattern to achieve measurable state is an optimization problem of its own.

3.1.1 Inherent selection bias

Given that, as part of TCP's congestion control approach, the majority of internet traffic rates exhibit a gradually increasing pattern, especially shorter data transfers have a higher probability of observing lower speeds. Apart from the traffic pattern, our methodology can not measure what is not demanded. So the distribution of the demand for speed, i.e. bulk data transfer will directly influence the measurements. Under normal circumstances, such influence is also likely to favor lower speeds due to bottlenecks in other points e.g. slow server. Additionally, burstiness of traffic is yet another point that possibly can shape the selection bias. Relatively short bursts combined with small buffers can cause probes to miss measurable states altogether. Yet again, in case of highly varying bandwidth where TCP fails to fill the pipe, queuing growths are more associated with decreasing speeds than increasing. Consequently, such factors result in a strong selection bias for our methodology.

3.1.2 Dummy-way vs Smart-way

Apart from the inherent selection bias, the logic of opportunity detection itself has the potential to introduce another kind of bias, because probes will be injected only when conditions dictated by the logic are met. In most dummy scenario, for saving on probe costs, static traffic rate thresholds could be utilized to decide when to inject probe packets. However, if more parameters are available to sample such as packet size and counts, the logic can be improved by enforcing observed average packet sizes to be close to maximum transmission unit (MTU). Thus, cross-traffic is more likely to be a bulk transfer that has more potential to saturate the link. Yet another improvement would be to maintain dynamic thresholds calculated from previous speed observations, given that link capacity has known statistical properties such as range of variation. Consequently, the design of opportunity detection is a trade-off point between cost and accuracy of the measurements.

3.2 Implementation

In this section, the prototype implementations of the hybrid model for realizing the goals of this study are described in sufficient detail.

3.2.1 Probe server

Essentially, the measurement can be achieved using a simple ping-pong server including send and receive timestamps in the packets. However, such an approach is challenging in our mission to achieve precision in probe intervals, mainly due to asymmetry in downlink and uplink. The additional variation in delays due to uplink decreases the timeliness of downlink probes. Furthermore, probe requests per each probe are a waste of resources, especially in downlink measurements. Hence, the purpose of the probe server is to serve variable probe trains upon request to achieve measurable states at the user end. To ensure accurate queuing delay estimations, a set of geographically distributed servers should be utilized as shown in Figure 3.2. User prototype application sends its probing request to the closest server by resolving geographical DNS enabled domain name. Closer probe servers are desired in order to meet the near-constant delay requirement as well as to be able to quickly serve probe trains upon request.

The probe server has two different modes of functioning. For uplink measurements, user application sends a UDP request packet per probe packet in various intervals where server replies with a probe packet which includes a timestamp of the server's monotonic clock. Differences in the timestamps are used to calculate the queuing delays experienced by the requests in the uplink direction to ultimately collect measurable samples. On downlink measurements, the aforementioned optimized approach is followed because downlink traffic usages are much bigger and more frequent than uplink ones. Consequently, downlink measurable states occur much more often. For downlink probing, a subscriber model is followed. User application request a probing stream with a certain interval i, and the server sends a probe pair every i ms. User application needs to send keepalives to keep probes coming otherwise the subscription times out. Once user application decides the probe stream is no longer necessary it

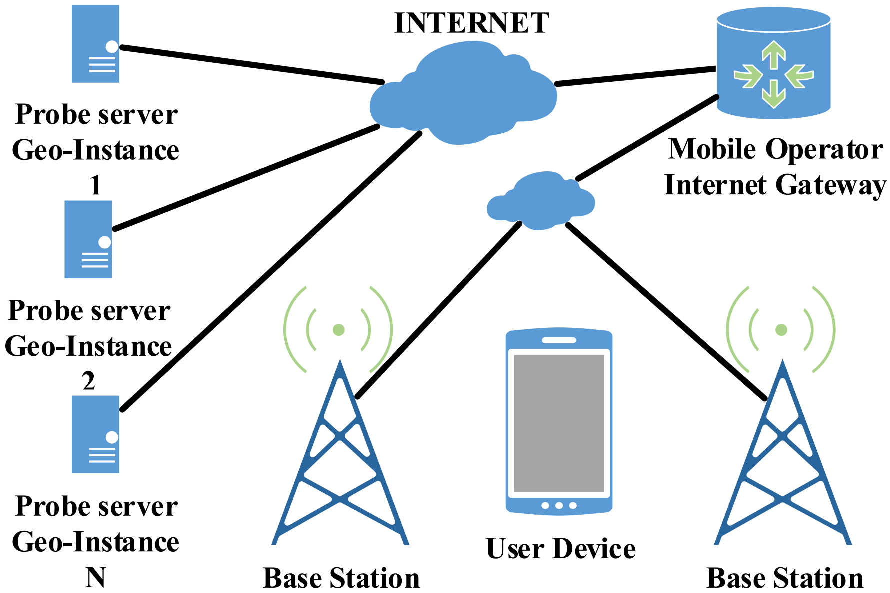


Figure 3.2. Probe servers network diagram. From Pub. II.

can request it to be ceased.

A probe pair is two probe packets sent 20 ms apart. The value 20 ms is arbitrarily chosen by experimenting on Finnish commercial networks providing 3G and 4G connectivity, but each parameter in the implementation could be optimized per radio access technology to make the most out of it. The significance of probe pair separation is that the queuing delays need to grow above this value so that there is a chance of observing a measurable state. Assuming that there is enough cross-traffic to utilize the full capacity of the measured user, the chance of observing the transient measurable state is directly proportional to how often the probe pairs are sent. So this is a trade-off point between probe bandwidth overhead and the chance of obtaining measurement samples.

3.2.2 Measurement client

Android OS is an inherently eligible mobile platform for the implementation of our prototype measurement tool. It provides all the requirements demanded by our measurement model already on its developer framework. Moreover, it is the most popular platform which powers a great variety of devices.

The prototype is developed as an Android library which can be easily plugged into any other android application. The measurement service is implemented using Android background service [21] capability which enables application code to run in the background silently without intruding user activity. The background service is configured to start automatically on device boot up and run indefinitely.

As a matter of fact, the application does not continuously run the code, but

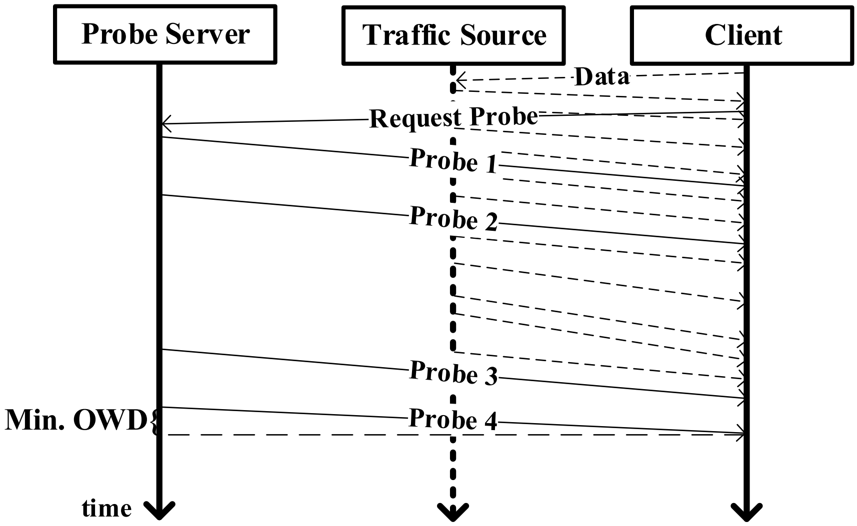


Figure 3.3. Sequence diagram for a short downlink measurement. From Pub. II.

sleeps in the background and listens for some trigger to wake up and start processing. As the code detects data activity, it starts sampling the rate of incoming and outgoing traffic using traffic stats API [22]. However, in the uplink, the cross-traffic can not be sampled correctly as the traffic stats are derived from the amount of traffic passed the NIC-kernel boundary. Hence, the cross-traffic data is approximate for the uplink case.

The sampling is the phase where the application needs to decide if a probe from the server should be requested or not. We have implemented a dummy approach here by checking if downlink or uplink traffic is above some threshold. These thresholds are 100 Kbps and 200 Kbps respectively. Once the threshold for the downlink rate is reached, the application requests a probe stream with a 100 ms interval and periodically sends keepalives until the traffic rate goes below half of the threshold for 2 seconds. For the uplink, as covered in probe server discussion, the application sends a probe echo request from the server every 60 ms. Yet again similarly if uplink traffic rate goes below half of the original threshold for 2 seconds the echo requests are stopped. These stop points are called idle phase. We define a ***session*** to be started by the sampling phase and ended by an idle phase where a form of probe conversation occurs. For the time between the application records all the traffic rates as well as probe packets information which includes sequence numbers, timestamps from both server and client application itself. Additionally, signal strengths, network technology, cell ID and location information that corresponds to the session are collected and bundled into a data structure. A list of these data structures piled until they reach a certain size, then the whole data is serialized and sent to the research server along with some detailed device information.

3.2.3 Bandwidth estimations

The queuing delay of packets is estimated by calculating the relative delay of each probe packet. In the context of a session, the relative delay is calculated by using the probe packet with minimum offset as the difference between server send timestamp and application receive timestamp. This indicates that the particular probe packet was the fastest in that session. Typically, it is one of the last packets in the tail of the session where cross-traffic is almost non-existent as the session will become idle soon as shown in Figure 3.3. One interesting fact about the fastest probe packet is that it might not be the absolute fastest which is physically possible. This effect only changes the number of observed measurable states because the minimum offset is slightly bigger than the real value, so the calculated relative delays are slightly lower.

The samples which satisfy the base measurement state are determined by using their relative delays, arrival times. The total number of bytes received and the duration between measurement samples are calculated. These two components are essential results of a measurement sample which at least a bandwidth estimation can be produced by. The reduction of the measurement samples into speed dimension is not desired because the duration of samples is crucial information regarding the calculation of a proper estimate based on weights by duration. The samples are known to carry the effects of packetization both in radio and IP layer as discussed.

However, these samples are not used right away but filtered by a set of conditions to eliminate radio access and implementation-specific effects. There was no delay margin filter imposed for this study, but to meet the model a filter is used to check the constraint of having at least a certain amount of cross-traffic between the arrival of probe pairs. This is especially desired in cases where there are few measurement samples so that the effects of interruptions in wireless communications(e.g. handovers, RRC state changes, etc.) are removed. The parameters for this filter is arbitrarily chosen as 3 packets and 1000 bytes. Additionally, in downlink, probe pairs received within 1 ms were filtered. These essential filters were applied to all samples and constitute a ***base***.

Lastly, to explore the effects of a simple filter based on probe pair spacings, we separately define Equation 3.2 as means of further eliminating potential "noise" caused by radio hiccups and user-space processing. The packets can be buffered in the kernel before delivery to the application, so fluctuations in the buffer result in different additional delays experienced by probe packets. These variations ultimately affect the calculated speeds. Such a filter restricts the measurement samples to be smoother where their spacing at the sender and receiver is similar. Hence, it aims to eliminate any samples affected by distortions in buffering or processing. The base measurements which are further filtered by this are referred as ***filtered*** in our analyses.

$$|\frac{(t_{rcv}[n] - t_{rcv}[n-1])}{t_{snd}[n] - t_{snd}[n-1]} - 1| < 0.5 \tag{3.2}$$

3.2.4 Example measurements

A hybrid measurement is comprised of many data points to be able to make accurate estimations. Therefore, it is helpful to first visualize them. A plot schema as in Figures 3.4 - 3.6 with two Y-axes was designed to describe prototype measurements. Stars represent estimation samples where red and purple stars belong to samples filtered out by Equation 3.2 for downlink and uplink probes respectively. Vertical lines show the average of corresponding samples so that we can see if there is any substantial difference between filtered samples and regular samples. Additionally, queuing delay experienced by probes over time can be seen by the black dots. However, note that RTT measurements, which are denoted by black triangles, are only available where we perform probing for uplink measurements due to downlink overhead optimizations in our measurement methodology. Relative delays (compared to minimum OWD in downlink) and RTT in the plots help us to follow the growth of queuing delays at particular instances.

We hand-picked some arbitrary familiar usage scenarios to showcase the functionality of the measurement tool and methodology. The figures below show results of those example measurements detected by the prototype application from an LG Nexus 4 device with a 4 Mbps limited HSPA+ connectivity. The network is known to be capable of achieving much higher speeds at the location of these measurements. Hence, 4 Mbps is the expected correct value of the achievable download speeds. There wasn't an artificial limit enforced for the upload speeds. In Figure 3.4, it is observable that the tool is capable of detecting achievable rates even with lower average speeds shown by green download speed indicator. This is a result of burstiness in the traffic and a probe pair coinciding with this burst wave in the user send buffer of the base station. As discussed in the challenges of cellular networks, it also shows a clear example of how a radio interruption such as handover or RRC state change can affect measurement as indicated by probe delay peak of 150 milliseconds at the early phase of the session.

The proposed simple algorithm seems to catch and filter out a seemingly unreliable sample around the interruption, however, it also eliminated many more fine measurement samples. So essentially filtering is a trade-off and optimization problem in itself to achieve more reliable estimates for short-lived measurement opportunities where we have very few samples. Those few initial samples by nature of bursty traffic process and mobile data context are more prone to deviation. For example, in the beginning of Figure 3.6 there is a download sample around 25 Mbps filtered. More often these odd samples are caused by some buffering process within user device operating system, while our measurement samples are collected on application level. If the measurement session did not generate any more viable sample after that point and we had not filtered that particular erroneous sample, we would end up with highly inaccurate estimation. Figure 3.5 shows the measurement of a YouTube video

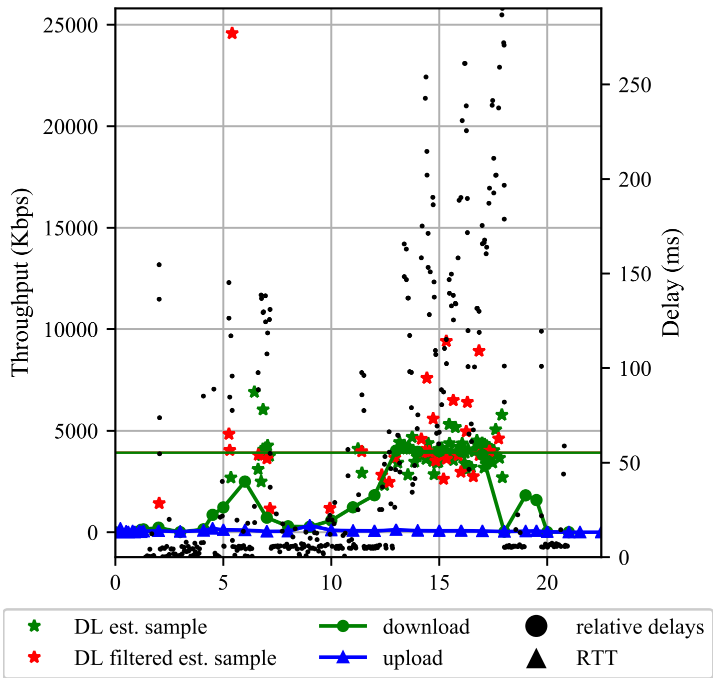


Figure 3.4. A typical web browsing session. From Pub. II.

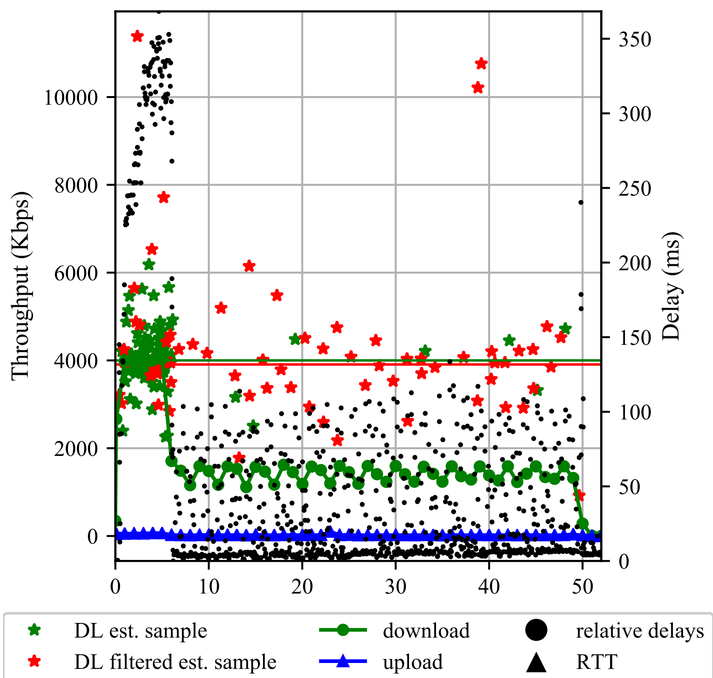


Figure 3.5. A youtube video viewing session. From Pub. II.

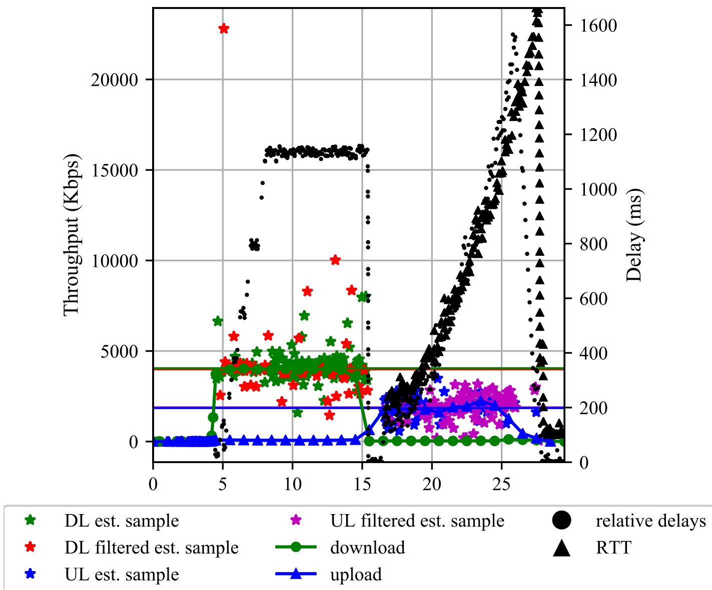


Figure 3.6. A hybrid measurement session corresponding to an active measurement as cross traffic source. From Pub. II.

streaming session which has its particular bursty traffic pattern. Initially, the client buffers a certain amount of data to ensure the smooth playout of the media. Subsequently, we observe periodical bursts and an example of coincidental interaction of probes and bursty traffic patterns.

We observe that our algorithm managed to carry out highly accurate estimations in these controlled example measurements by tackling varying traffic patterns and cellular network specific issues. However, we also observe that having a low link capacity allowed us to get an ample amount of samples in these example measurements to make accurate predictions.

3.3 Results

In this section, we evaluate our methodology via the prototype and present our findings. The prototype was evaluated with an Android application which also featured active goodput measurements. We have collected data for roughly 2 years which resulted in about 9 million LTE and HSPA+ sessions each, from 3796 users, 260 operators[1] and 522 distinct device models.

3.3.1 Accuracy

In order to investigate the accuracy of the hybrid measurements, first, we have analyzed how active goodput measurements compare when they act as a cross-

[1]Majority of the data originated from Finnish operators.

traffic to generate hybrid measurement samples as in Figure 3.6. Then we visit the differences in statistical properties of hybrid measurements as opposed to active measurements.

Firstly, we would like to acknowledge that the most granular analysis of accuracy would be to compare base station data to momentary estimations. However we can not practically carry out such experiments on a large scale with multiple operators, hence our experiment is set up to explore it via a proxy. Hybrid measurements as implemented based on NIC traffic counters measure IP throughput, however, the experimental setup does not have such active measurement but expectedly closest proxy that we can utilize is active goodput measurements. Such comparison is certainly weak in the aspect that TCP can not guarantee full link utilization especially in the cases of unstable connectivity and poor buffer policy leading to starvation. However, goodput values provide a definite lower bound for both TCP and IP throughput. Hence the comparison allows us to roughly explore the typical ceiling for the accuracy figures in different scenarios. Also, goodput comparability is of great interest to understand how our methodology relates to typical measurement tools out there.

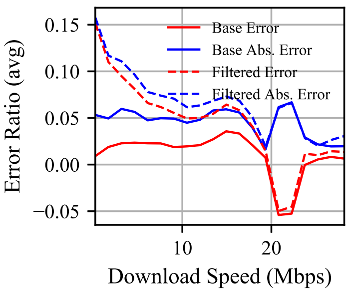
$$ErrorRatio = \frac{0.97 \times Estimation_{t>5s}}{Goodput_{t>5s}} - 1 \qquad (3.3)$$

We use a factor of 0.97 to account for the approximate overhead of IPv4 and TCP to make the comparison more sensible. To avoid the slow-start portion of the active measurements, we only compare the second half of the 10-second measurements. Equation 3.3 defines an arbitrary signed error metric to be able to assess the occurrences of overestimation and underestimation. Such a definition also helps us to make sense of the direction of average errors where underestimations and overestimations might cancel each other out.

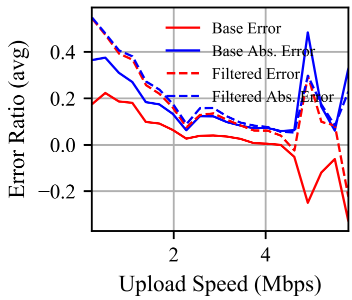
At this point, we have two separate datasets to work with. ***Dataset A*** comprised of the hybrid measurements in which we inspected its characteristics in the wild. ***Dataset B*** is composed of active goodput measurements and associated hybrid measurement sessions.

To further refine our comparative analysis we limit the Dataset B to only include measurements where goodput has more than 4 seconds non-zero samples and estimation has more than 0.2 seconds. Such elimination is useful to limit our comparison to relatively stable connectivity which includes a considerable period of saturation. However, the difference in the sample sizes would still lead to inflated error ratios where the throughput variation is high. Additionally, we evaluate the simple filtering condition as defined in Equation 3.2 to explore the effects of filtering based on probe arrivals.

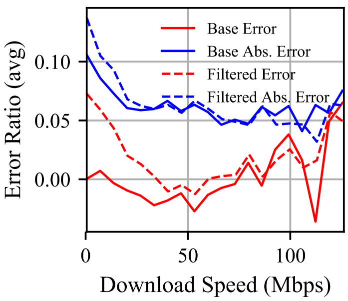
Figure 3.7 shows typical accuracy numbers changing with the speed. We observe that filtered measurements usually have higher error rates especially at lower speeds where they tend to overestimate. Not only they increase error rates but they also decrease the number of measurement samples which adds to



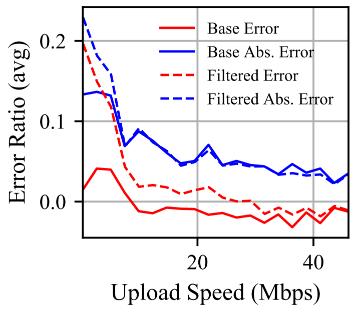
(a) DL estimations (HSPA+)



(b) UL estimations (HSPA+)



(c) DL estimations (LTE)

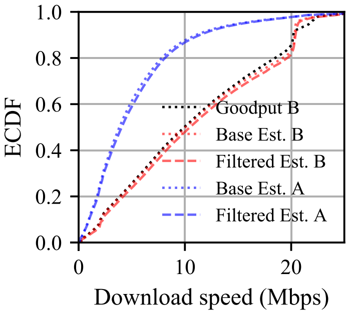


(d) UL estimations (LTE)

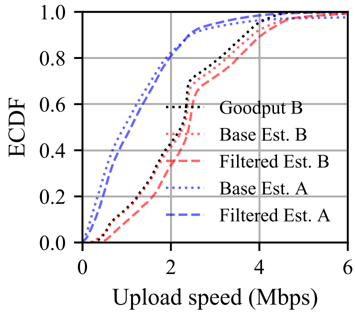
Figure 3.7. Accuracy of estimations based on active measurement comparisons. From Pub. II.

sampling error in general. The uplink measurements experience higher error ratios as an expected result of NIC stat reading issues described in section 3.2.2. Another interesting observation is the spikes in underestimation on physical channel limits such as 21.1Mbps for HSPA+. This can be attributed to sampling error gaining a direction in such cases as samples can not exceed the maximum transmission speed significantly.

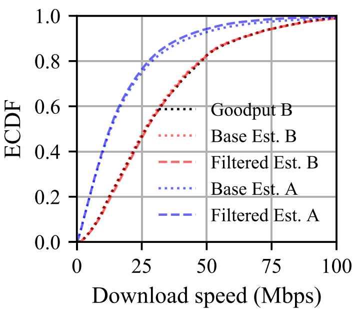
Finally, we take a look at the visible effects of selection bias in our hybrid measurements due to the opportunistic approach. To be able to inspect and compare the results, we produced empirical distribution plots of our datasets in various cases. It is crucial to note that although datasets are derived from the same user base, they certainly do not belong to the same population in statistical sense as we have not done any stratification or matching in any direction to make them more comparable. Dataset B, where active throughput measurement acts as cross-traffic, is characterized by a comparatively very small number of measurements featuring its own biases due to user measuring behavior. On the other hand, Dataset A is dominated by the selection bias of the opportunistic approach where we expect lower speeds to be caught more. Figure 3.8 confirms such expectation on a very high-level comparison. In spite of estimations in Dataset B are accurate on predicting goodput, we observe that in Dataset A measurements resulted in much lower numbers for the reasons discussed in section 3.1.1.



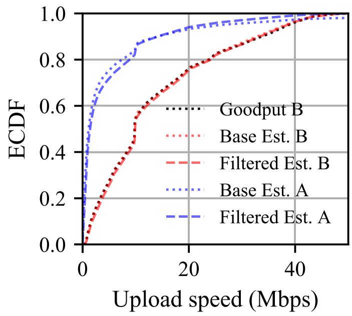
(a) DL estimations (HSPA+)



(b) UL estimations (HSPA+)



(c) DL estimations (LTE)



(d) UL estimations (LTE)

Figure 3.8. Cumulative distribution comparisons of datasets and estimations. From Pub. II.

3.3.2 Efficiency

The resultant measurement service is very cost-efficient, probe session incurs less than 10 Kbps overhead and less than 1% battery usage. The total bandwidth and battery usage by the service are naturally proportional to mobile data usage itself. So static probe patterns as we have implemented in our research prototype typically result in a few megabytes measurement overhead per day.

However, the overhead can be hugely improved by smarter deployments of the methodology where probing is seldom performed during network usage. We have briefly experimented using a dynamic exponential backoff method for the probing threshold based on maximum achieved speeds in previous measurements. There was about 30 fold improvement in overhead. However, such fundamental changes would certainly result in different selection biases. Consequently, comparative studies of the measurements must be based on careful statistical modeling of the data. Furthermore, given that our methodology bears the potential to gather a large amount of data, a machine learning approach can be taken to explore the predictability of the bandwidth using additional data points such as location, device, operator, signal levels, cell ID, time of the day, etc.

3.3.3 Limitations

Accuracy figures and selection biases have the potential to significantly deviate among different cellular network deployments. This is majorly due to different mechanisms such as active queue management (AQM) or scheduling disciplines that might alter the traffic arrival characteristics and buffering process. As a result, in such cases, sanity checks for those issues should be carried out to make sure that they don't systematically skew the measurement results. Additionally, in this study majority of the measurements are characterized by a short distance to probe server which results in relatively accurate queuing delay estimations throughout the dataset. Consequently, it is naturally expected for measurement results to become less reliable when the assumptions of the model are not properly met.

Although, in our evaluation study we mainly focused on hybrid bandwidth estimations, probing data can also be used to estimate QoS metrics such as RTT, loss, jitter with little effort. As opposed to commonplace tools, those metrics can be explored for different traffic patterns. For example, jitter originating from the network can be analyzed for cases of constant or variable bit rate real-time communications such as VoIP. Furthermore, loss spikes combined with cross-traffic observations can be used to infer user experienced connectivity losses to approximate availability.

3.4 Summary

This chapter described a novel probe-based measurement approach for mobile networks. By utilizing the capabilities of smartphones, specifically the ability to sample incoming and outgoing traffic rates at the mobile network interface, it is possible to have a capacity measurement model having the best properties of both probe gap and rate models. On top of that, thanks to smartphone platforms, additional contextual information such as location, device, operator, signal levels, cell ID, time of the day, etc. enable fairly advanced analytics as well. Furthermore, having always-on passive measurements provide us usage characteristics and the actual quality experienced by the user.

The measurement prototype is empirically tested and shown to be cost-efficient and scalable. Although the capacity measurements are comparably accurate when directly tested against active goodput measurements, the passive nature of sampling always introduces complexity and bias. Consequently, carrying out sparse and optimized measurements using this approach requires careful statistical modeling to achieve better results.

4. Predicting mobile experience in the field

This chapter presents QoE aspect of this work in which predictability of mobile experience is investigated in depth. First, we study the concept of user satisfaction in mobile networks by looking at the quality factors affecting satisfaction in a longer time window via active measurements. In the second study, learning from the first study is utilized for designing a shorter time window application-level QoE study based on our hybrid measurements. This chapter is based on Publication III and Publication IV.

4.1 Mobile user satisfaction

We study the network and non-network predictors of user satisfaction with network speed and availability in the context of user device based network measurements. Specifically, we combine user questionnaire responses and those questionnaire respondents' empirical mobile network measurements from the device based network measurement platform Nettitutka [75]. This combination allows us to determine the significant predictors (aka features) of network satisfaction for those end users that themselves actively measure the network and observe the reported results.

4.1.1 Methodology

In order to study the aforementioned user satisfaction, we implemented a pop-up questionnaire in the Android, iOS, and Windows Phone versions of the Nettitutka application. Importantly the questionnaire first prompts the user whether they want to participate and thus the questionnaire is completely optional and not required to utilize the application. The participation prompt always appears after the first user-initiated cellular network measurement, and if the user dismisses the prompt then the prompt reappears with a 15% probability after subsequent user-initiated cellular measurements. If the user completes the questionnaire then the prompt does not reappear for three months. We released the updated Nettitutka applications that include the questionnaire

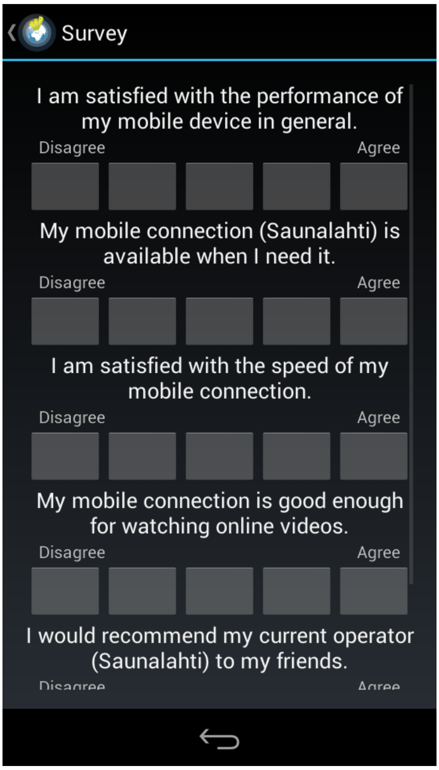


Figure 4.1. Screenshot of the Questionnaire from the Android Nettitutka App. From Pub. III.

during the course of summer 2015.

The questionnaire includes five statements with each statement having a five-point Likert scale that gauges whether the user agrees or disagrees with the statement. The questions are enumerated below and a screenshot of the questionnaire within the Nettitutka app is shown in Figure 4.1.

Q1. I am satisfied with the performance of my mobile device in general.

Q2. My mobile connection (current operator) is available when I need it.

Q3. I am satisfied with the speed of my mobile connection.

Q4. My mobile connection is good enough for watching online videos.

Q5. I would recommend my current operator (current operator) to my friends.

We construct a series of ordinal logistic regression models to determine which features are statistically significant with respect to predicting responses for each questionnaire question. We first briefly explain our procedures for multicollinearity testing, missing data imputation, and model building. Then we present the resulting models and describe the results.

For convenience hereafter, we refer to the Likert scale values for the questions by numerical shorthand values of 1 to 5 with 1 representing disagree and 5 representing agree. Also, we denote aggregations of scale values by parentheses; in other words, (1,2) vs (3,4,5) indicates a simple dichotimization of the responses.

Overall as of December 2015, we find that 5,550 users completed the questionnaire. In terms of response rate, the per user response rate is approximately 15.23%. In other words, 15.23% of users that saw the questionnaire prompt at least once eventually completed the questionnaire. The respondents are then filtered to create a more homogeneous subset of respondents for analysis. First, We filter respondents to keep only those based in Finland. We considered a respondent to be Finland-based if the respondent's operator is one of the three major Finnish operators[1] (Elisa, Sonera, or DNA). Over 95% of all respondents as of December 2015 were Finland-based respondents giving a total of 5,274 respondents. Then in order to have an adequate number of measurements per user from which to derive features we further filter the dataset by keeping only respondents with at least five measurements and with at least one of those measurements being valid (i.e. successfully measured all network metrics). We find that 42.17% of respondents in the dataset meet this criteria thus giving a total of 2,224 respondents.

4.1.2 Results

In terms of overall model results, Table 4.1 details the significance levels of each feature for each model. Overall, we find that the minimum, median, and last TCP down features are significant across many models. In fact, the high significance ($p \leq 0.1\%$) of last TCP down in so many models especially highlights the importance of temporal effects in such satisfaction evaluations.

Interestingly, we find that very few of the TCP upload features are significant in any of the models. This result can be understood in context that for a typical user the download share of their mobile traffic is much larger than the upload share. For example, Sandvine [68] reports that for a median European user the download-upload mobile traffic split is 88%-12%.

Furthermore, relatively few of the latency based features are highly significant. We postulate that many users might not be aware of the nature of latency and high latency might not directly affect their opinions. In addition, latency below a certain threshold is valuable in only a relatively small subset of applications, thus in our case, the effect of latency might not be perceivable to many users. In a specific example, Wac et al. [82] describe a variety of latency requirements for different mobile applications from literature with the lowest requirement level being 100 ms for first person shooter video games. About 90% of our respondents already have median RTT below 100 ms.

In terms of non-network features, the number of user-initiated invalid measurements is only weakly significant in a few models. The weak significance of invalid measurement counts might be partly due to the relative infrequency

[1]The three major operators have over 99% combined market share of mobile subscriptions in Finland.

Table 4.1. Feature Significance Levels[c] for Logit Models. From Pub. III.

Feature	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**
Median TCP Down		*	**	**	
St. Dev. TCP Down	*				*
Last TCP Down	***a		***	***a	***a
Min TCP Down		**a	***a	*	
Median TCP Up	**			*	
Last RTT	*	*		*	
Min RTT					*a
Max RTT		*		*a	*
Number Frequently Measured Locations		**	*	**	**
Number User-Initiated Measurements	*				
Number Invalid User-Initiated Measurements			*	*	*
Device Platform	*b	*	*	*b	***b
Mobile Network Operator		**		*	***b
Device Quality	***				
Device Type (Tablet or Smartphone)		**			

[a] Feature did not satisfy parallel lines assumption, reported significance level is for baseline of (1,2) vs (3,4,5).
[b] Feature did not satisfy parallel lines assumption and feature is categorical, reported significance level is for baseline of (1,2) vs (3,4,5) for most significant comparison between any two categories.
[c] $*$: 5%, $**$: 1%, $***$: 0.1%

of invalid measurements overall. Specifically, 46% of users do not have any invalid measurements at all and for the median user with 13 user-initiated measurements, only 1 of these are invalid.

Unsurprisingly, the custom device quality feature is highly significant in the Q1 model as would be expected. However, interestingly the feature is not significant in any of the other models, thus suggesting that device quality is, in and of itself, not predictive of network availability or speed satisfaction.

Since adjusted probabilities in ordinal cases are only intuitive for binary outcomes we again collapse the ordinal dependent variable to a binary dependent variable by considering only the (1,2) vs (3,4,5) comparison. We term the (3,4,5) state as satisfaction. We then re-estimate the models as simple binary logistic regression models. Furthermore, since all models are simple binary logistic regression models we no longer need to differentiate between normal and generalized models. In addition due to technical limitations we do not utilize multiple imputations but instead utilize complete case analysis. This change though does not substantially alter the outcomes.

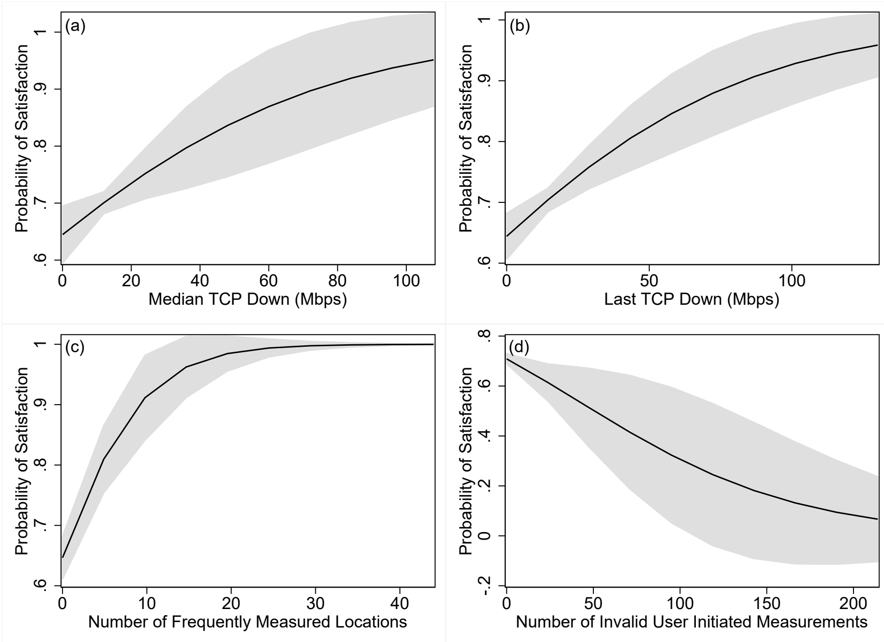


Figure 4.2. Predicted Probability Functions for Several Significant Features of Q3: (a) Median TCP Down, (b) Last TCP Down, (c) Number of Frequently Measured Locations, (d) Number of Invalid User Initiated Measurements. The Stata command `mcp` was utilized to compute the probability functions [67]. From Pub. III.

We focus on network speed satisfaction (Q3), Figure 4.2 illustrates the predicted probabilities of satisfaction for Q3 over four features: median TCP down, last TCP down, number of frequently measured locations, and number of invalid user-initiated measurements. As expected in both the median and last TCP down cases we find that the probability of satisfaction increases rapidly as the goodput increases but shows constant diminishing returns. In terms of locations, we find that the probability of satisfaction increases strongly with the number of frequently measured locations. Finally, the number of user-initiated invalid measurements suggests a strong effect but as previously discussed relatively few users have significant numbers of invalid measurements and thus no strong conclusions can be drawn.

4.2 Hybrid measurement based mobile QoE

We model and analyze mobile (including LTE, Wifi, and HSPA+) app-level QoE data collected from a large (292 users with 64,036 experience ratings), a diverse group of participants during Summer 2017.

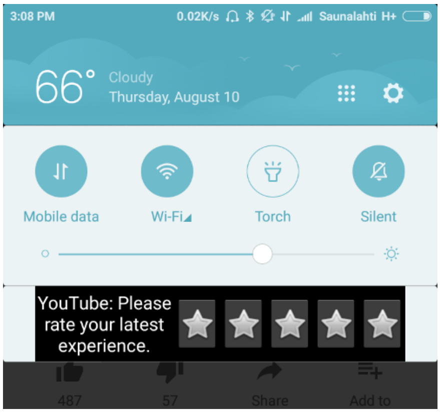


Figure 4.3. Screenshot of an experience rating (for YouTube) from the QoE-Client. From Pub. IV.

4.2.1 Methodology

The participants installed a custom measurement app that both monitors network quality and prompts participants to rate their experiences with other apps during or after usage of those other apps. Therefore the method is a version of the experience sampling method [47, 18]. Additionally, participants were instructed to use their devices normally (without any pre-defined instructions or actions), thus facilitating realistic interactions.

We perform both a descriptive statistical analysis and machine learning, specifically random forest (RF) and support vector machine (SVM), based modeling with a focus on imbalanced dataset learning[2]. Compared to prior work [12], this more greatly focus on imbalanced learning including using imbalanced learning specific metrics (rather than less suited general metrics) is theoretically and practically advantageous. For example, general accuracy as a metric can greatly reduce the impact of the minority class (which is of significant interest) [10].

The dataset was collected from a group of Finnish Android smartphone users that installed a custom measurement app known as the Aalto QoE app and hereafter the QoE-Client. The QoE-Client measures network quality during usage of other mobile apps and occasionally shows a notification prompting the user to rate their experience with the specific app (denoted as an experience rating). In other words, the QoE-Client is a network monitoring client that also employs experience sampling. Experience sampling is a widely used social science method frequently applied in mobile contexts [80]. Additionally, we use an exit survey to collect demographic information of the participants.

The participants were recruited through several different media channels including email lists, word-of-mouth, online ads, and newspaper ads to acquire a diverse group. Some ads were specifically targeted towards media channels of rural areas to attempt to recruit such users under the assumption that their

[2]The imbalance ratio between our normal and bad QoE modeling classes is about 26.

network quality might be lower than urban users.In terms of representativeness, Table 4.2 compares the demographics of the participants (collected through the exit survey) to Finnish smartphone users in general. Despite attempts to acquire a diverse participant group, the participant demographics are still skewed towards younger users in the Uusimaa region[3]. Therefore, this discrepancy should be considered when making generalizations.

The QoE-Client network measuring method is based on a hybrid methodology described in Chapter 3. Specifically, the QoE-Client continuously performs passive network measurements whenever the display is on. Such passive measurements are mainly comprised of network interface level traffic samples and contextual information such as the foreground app, location, network type, and signal strength. Additionally, the QoE-Client employs active latency measurements during specific interesting moments based on a set of heuristics.

In terms of the actual experience rating interface, Figure 4.3 illustrates an example experience rating. The use of a five-point star scale is based primarily on both the prevalence of the scale in everyday use [71] and display size considerations. Additionally, given the prevalence of the scale, users do not need to be trained to use the scale in contrast to the mean opinion score scale.

Additionally, we crawl the Google Play app store (hereafter Play store) for data related to the specific app evaluated in each experience rating. Specifically, we collect Play store category, Play store mean review score (scale of [1,5]), number of Play store reviews, and number of Play store downloads (minimum of the download range).

Table 4.2. Comparison of demographics between participants and Finnish smartphone users. From Pub. IV.

Demographic	Participants	Finnish Smartphone Users[a]
Young Participants (% 19-32 years old)	67.80	25.56 (1.90)
Gender (% male)	47.44	50.26 (2.00)
Home region (% Uusimaa)	67.12	32.64 (1.92)
Smartphone usage[b,c] (years)	6.88 (3.07)	-
Sub limited[d] (% yes)	44.03	-
N	292	4,237,310

[a] Finnish smartphone user demographics are from the Finnish sub-population that reported owning a smartphone (n=713) from Eurobarometer 87.1 survey of March 2017 [26]. Proportions also include linearized standard errors for population estimates.
[b] Number of years the participant has used a smartphone
[c] Means also include standard deviations
[d] Whether the participant's subscription is limited to a maximum DL throughput. In Finland, many mobile subscriptions are differentiated by artificial maximum throughput caps rather than data caps. Note that these throughput caps are not the same as the simple technological limits of certain network technologies (for example HSPA, LTE, etc.).

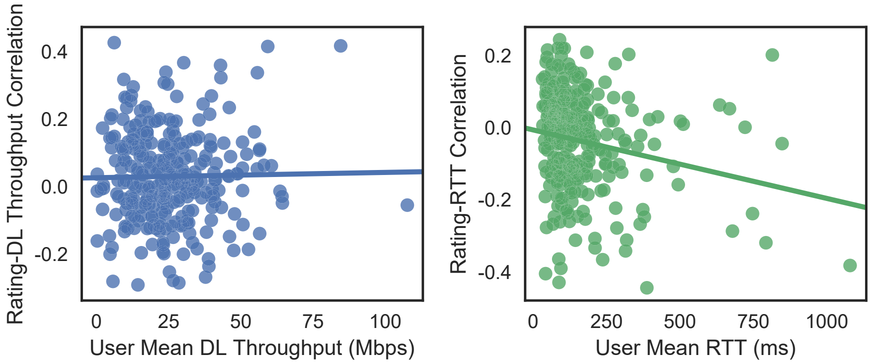


Figure 4.4. DL throughput and delay Spearman correlations versus normalized rating at the user level. From Pub. IV.

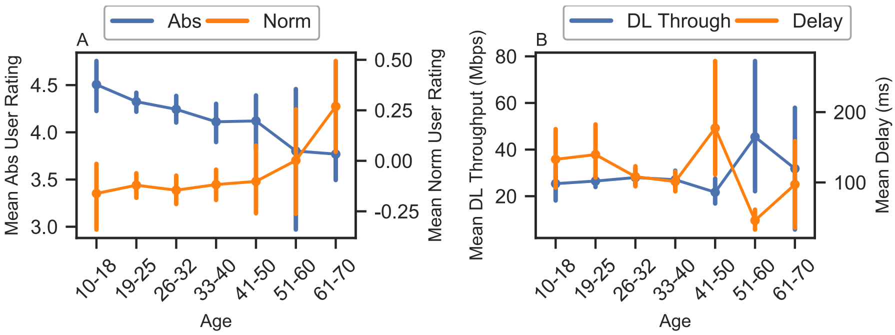


Figure 4.5. A) Mean absolute and normalized user rating for different age ranges, and B) Mean DL throughput and delay for different age ranges. From Pub. IV.

4.2.2 Results

Figure 4.4A details the user-level correlations between DL throughput and normalized rating versus the user-level mean DL throughput. Interestingly, we find no clear significant relationship ($\rho = -0.04$, $p = 0.53$), even for users with low mean DL throughputs of less than 8 Mbps ($\rho = 0.06$, $p = 0.79$). One possible explanation is that these low DL throughput users adapt their usage (and expectations) to the low throughput environment (either manually or through apps that automatically adapt).

In comparison, Figure 4.4B shows the user-level correlations between delay and normalized rating versus the user-level mean delay. The relationship is clearly stronger and more significant ($\rho = -0.22$, $p = 0.0002$). This significant relationship is potentially because adapting delay-sensitive apps to higher delays is generally more difficult than adapting throughput-sensitive apps to lower throughputs.

Additionally, at the user level, we can analyze the relationship between these

[3] The Uusimaa region includes Helsinki and surrounding municipalities including Espoo which contains Aalto University.

Table 4.3. Mean test results for general models (real and baselines). From Pub. IV.

	RF[a]	GBF[b]	SVM[c]
G_{mean}	**0.598 [0.534, 0.662]**	0.597 [0.526, 0.668]	0.586 [0.485, 0.687]
$IBA_{G_{mean}}$	**0.355 [0.274, 0.435]**	0.354 [0.264, 0.444]	0.339 [0.216, 0.463]
F_1	0.106 [0.064, 0.148]	**0.107 [0.061, 0.154]**	0.105 [0.06, 0.151]
	Baseline-UC[d]	**Baseline-SC[e]**	**Baseline-IF[f]**
G_{mean}	0.497 [0.463, 0.530]	0.182 [0.120, 0.244]	0.338 [0.276, 0.400]
$IBA_{G_{mean}}$	0.248 [0.213, 0.282]	0.031 [0.010, 0.052]	0.122 [0.081, 0.164]
F_1	0.070 [0.044, 0.097]	0.035 [0.012, 0.057]	0.063 [0.044, 0.082]

[a] Random forest model
[b] Gradient boosted forest model
[c] C-Support vector Machine model
[d] Baseline (dummy) model with uniform class prediction
[e] Baseline (dummy) model with stratified class prediction
[f] Baseline model with isolation forest anomaly detection

aggregate features and the user demographics collected from the participant exit survey. Figure 4.5 illustrates the mean absolute and normalized user rating for different age groups. We find that absolute rating tends to decrease with age, but normalized rating stays roughly the same. This suggests that older users might be using the rating scale differently (potentially using the center three-star rating as a baseline rather than the five-star rating). In further support, as Figure 4.5 also illustrates, we do not find clear patterns for network quality differences between age groups. We did not find significant differences in terms of ratings or features for different genders, smartphone usage years, or home region.

We perform supervised learning for predicting the normalized QoE classes. Specifically, we perform two concrete modeling tasks. First, we train and evaluate a general model that includes all app meta-categories and predicts either the Normal-QoE or Bad-QoE class (binary classification). Second, we train and evaluate a separate model (also Normal-QoE or Bad-QoE class prediction) for each specific meta-category. These meta-category specific models help quantify the potential performance on these diverse meta-categories.

For additional context, we first describe the metrics for evaluation of the models, the evaluated model types (e.g. RF, SVM), the training procedure, and the baseline models for comparison. Finally, we detail the modeling task results in Table 4.3.

For evaluation, we use several common imbalanced learning metrics including G-Mean, index balanced accuracy (IBA) of G-Mean, and F_1 [10]. G-Mean is the geometric mean of sensitivity (true positive rate) and specificity (true negative rate). Therefore, G-Mean, and essentially all binary imbalanced learning metrics, measure the balance (or trade-off) between positive and negative accuracy. We

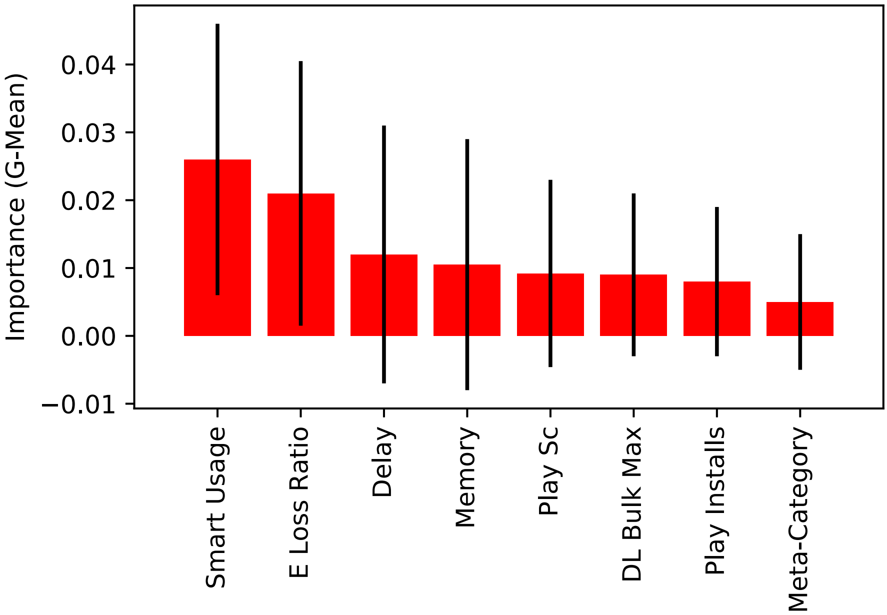


Figure 4.6. Feature importances (out-of-bag permutation importance of the G-Mean) of the RF model (including standard deviations of importances). From Pub. IV.

find that the models provide a significant but small advantage over the baselines with only about a 20% higher G-Mean.

In terms of individual features, the backward feature elimination approach leaves eight features in the final RF model including device memory available, DL bulk max (an estimate of application throughput regardless of bottleneck being server or network), delay, echo loss ratio, Play store mean review score, Play store number of downloads, smartphone usage years, and meta-category. For illustration, Figure 4.6 depicts the feature importance of each model feature (as quantified by the out-of-bag permutation importance of the G-Mean).

The most important feature is smartphone usage years which potentially acts as a proxy for technological sophistication through the proclivity for early adoption. Such technological sophistication likely affects the user's expectations of quality, an important factor influencing mobile QoE [64].

We also find that the echo loss ratio is one of the most important features. This finding is intuitive given that a significant connection loss is the most egregious network problem and might affect the QoE of all connection-reliant apps. Contrastingly, given the generally high DL throughputs and low delays, even moderate changes in these features might not affect a user's QoE. Though, we note that DL bulk max and delay still have modest importance in the model.

Interestingly, we find that both the Play store number of downloads and the Play store mean review score are important features. We hypothesize that these app store features could act as proxies for general app quality. For example, these features could capture issues with app user interfaces such as glitches and poor design that are not network related. Similarly, as hypothesized earlier, available device memory is also important.

However, we note that the resultant features and their spread of importance are far too wide to allow any generic conclusions for arbitrary app QoE within our study. Our data was highly imbalanced and insufficient for deeper focused analyses on subcategories. This indicates that such high-level data with no actual contextual features from application level is potentially not sufficient to develop strong QoE models. This is somewhat expected as we have previously discussed many smartphone apps feature multi-modal use cases from browsing, streaming to calls. Some are latency-sensitive, some sessions make use of prefetched cached content that decouples experience from network metrics, etc. At the end of the day, these applications are designed for pleasant user experience for varying conditions and deteriorations in underlying systems and devices. Consequently, this study is a first to explore and paves the way to uncharted territories between large-scale QoS measurements and tightly controlled QoE experiments on lab conditions. Our results indicate that future studies should focus on inference of modality in the applications themselves and enrich the app-specific contextual data to allow stronger and practical QoE models that can be used in smartphone based in-field measurements.

4.3 Summary

In this chapter, we investigated factors affecting mobile user experience. Our findings in the user-initiated active measurement study have shown the latest download throughput is most deterministic in user-reported satisfaction whereas no such significant relation was there in hybrid measurement (results invisible to user). This suggests that user-initiated measurements bear some peculiarities and are detached from actual usage experience dynamics in terms of mobile network quality. This can be partially attributed to user measuring behavior and users giving ratings based on their perception of the visible results.

In hybrid measurements, latency and network availability turned out to be more deterministic in app QoE along with other non-network factors (e.g. device, application quality). On the other hand, we found that the profile of the user influences overall self-reported rating characteristics. For instance, significant differences in reported ratings were found between different age groups.

Although both of the studies were fundamentally challenged by extremely good Finnish mobile networks, hybrid measurements, and machine learning approach was able to yield significant gains in predictive power. Yet it might prove even more useful in poorer network conditions. However, our findings suggest that the predictability of mobile user experience remains a hard problem without complete end-to-end visibility due to the fact that mobile device ecosystem and usage are extremely diverse.

5. Discussion

In previous chapters, our research questions, motivations as well as associated studies were presented. This chapter discusses our overall conclusions. First, the findings are covered for each research question that has been posed. Then, the practical applications of this work are discussed. Finally, open questions and related future work are presented.

5.1 Research implications

The mobile-first era is here and demand for mobile data is growing stronger than ever. Regardless of the improvements in the capabilities of mobile networks, measuring them is still rather complicated task given the nature of mobile communications. The resultant performance of a connection in mobile networks is a complex interaction between momentary cell load, adjacent cell interference, shadowing, fading, mobility and user device capabilities. Along with upcoming commercial adoption of 5G networks, heterogeneity and variability in radio access are expected to grow even further due to the fact that improvements heavily depend on elements such as smaller cells, fragile millimeter-wave bands, etc. Consequently, accurate quality mapping requires frequent measurements with both time and space diversity that's representative of how the networks are actually used by end-users.

At the beginning of the work, we divided our research questions into two distinct groups with respect to our overall problem. Firstly, we focused on network QoS measurements and methodologies in mobile networks where we started by investigating shortcomings of existing ways of measuring, especially for the purpose of crowdsourcing network quality. In Chapter 2, we have discussed and shown that existing measurement methodologies are either too costly and non-scalable (e.g. TCP throughput tests, probe rate model) or too inaccurate (e.g. probe gap model, passive measurements) for many purposes. Additionally, we have shown that even TCP tests often fail to reflect the true capacity of mobile link due to fundamental assumptions in TCP congestion control algorithms. As motivated by all these problems, in Chapter 3, we have proposed a hybrid

QoS measurement model and methodology that has the potential to meet such challenges. Furthermore, we prototyped a proof-of-concept Android application. Finally, we carried out a large-scale evaluation study that has shown how not only continuous QoS measurements that provides cost-efficient network capacity measurements but also mobile data usage characteristics can be gathered at scale with such little overheads.

In the second group of research questions, we focused on user aspect of the network quality in mobile smartphone usage. In chapter 2, we have briefly discussed the paradigm shift from network-centric QoS based quality management to user-centric QoE focused quality measurements as well as the fundamental problems related to such transition. For example, definition and scope of QoE, measurement metrics and methodologies for QoE, and QoS to QoE modeling are some of those problems. Although generically QoE is known to be logarithmically related to QoS for real-time applications such as voice/video calls and IPTV, for complex applications such as web-browsing the relationship is found to be not so straightforward. Modern smartphone applications bring in even more variety on how network resources are utilized throughout the usage session. However, when we have investigated the relevant literature as covered in related works, we have found out that similar studies were overly simplistic, their results were too good to be true and not rigorous enough to convince us in terms of generalizable predictability of in-field QoE by just applying machine learning along with few features. Consequently, we decided to focus on this research gap to explore the predictability of QoE out in the wilderness in order to identify the challenges and gauge the potential of a hybrid measurement approach that's harnessing all the capabilities smartphone platforms have to offer.

In Chapter 4, we first started by investigating the generic factors (both network and non-network) that affect in-field mobile user satisfaction in a longer time window via a user study based on user-initiated active measurements. In the study, users were able to carry out measurements on demand and see the results before they were prompted to answer various survey questions. Unsurprisingly, most recent download throughput was the most deterministic factor for network satisfaction. However, in the second study where we have focused on short time window per-app QoE via hybrid measurements, we could not find such a clear relationship between user experienced download speeds and QoE ratings. Given that Finnish networks are mostly performant enough for many usage scenarios, such finding is completely plausible as well. So it can be concluded that user-initiated measurements are detached from actual usage experience dynamics in terms of mobile network quality. Such finding suggests that active measurements along with self-reported ratings where users interpret measurement results might not reflect the actual impact of network quality on user experience where it matters. In hybrid measurements, latency and network availability are shown to have more impact in app QoE along with other non-network factors (e.g. device, application quality). At the same time, our findings suggest that profile of the user influences overall self-reported rating characteristics. For example, we have found significant differences in reported ratings between different age groups and similarly, there might be varying experiential dynamics among different groups as well.

Our empirical efforts for modeling QoE based on network QoS were fundamentally limited by good quality Finnish mobile networks. Consequently, the measurements we could gather were highly imbalanced. However, such limitation also enabled us to get a glimpse of in-field predictability of QoE degradation on rare cases of network issues. Our findings suggest that it is possible to obtain significant gains in predicting QoE degradations even in good quality networks. However, diverse non-network factors are often much more dominating in such cases, especially, considering the variety in tasks and usage scenarios of different types of applications. For example, a typical social media application might feature text messaging or varying media content delivery in realtime or non-realtime. Without in-app state and context, it would be a useless effort trying to come up with an accurate QoE model capable of reflecting the momentary experience. In this respect, our opportunistic generic hybrid approach to quality measurements was not strong enough to lend itself to apply advanced techniques such as pattern recognition to act as a proxy to in-app contexts and modality (e.g. detecting video streaming session in a well-known social media app as opposed to live video call). It's clear that measurements need to be able to adapt to detected "quality" context to collect a complete set of relevant network quality metrics and avoid information scarcity as we have faced often in our studies.

5.2 Practical impact and implications

The research problems we have worked on in this dissertation were originally motivated by the search for practical ways to map mobile network quality at scale. At each step, we have carefully evaluated the practical applicability of various approaches and built complete implementations to test them. Consequently, our work can be readily and easily integrated into mobile platforms as a service and can be exposed to application developers to improve the networking aspects of their apps. Android network capabilities API that provides coarse bandwidth information is a strong example of such a use case [20].

Moreover, today many networked applications are backed by the "Cloud" that comes in many forms. Content Distribution Networks (CDN), Infrastructure as a Service (IaaS), and Plaftorm as a Service (PaaS) are some of the examples of cloud ecosystems that modern applications are built on. The reliability and quality of such services also have a direct impact on the overall user experience. For example, degraded performance of a networked application can be caused by slow servers, congested networks or device issues. Consequently, constant monitoring of relevant aspects is essential in building great experiences. Given that app developers are capable of tracking in-app state and greater contextual details, our measurement framework would enable constructing built-in

QoE models that might be used for user experience monitoring. Similarly, our work can be utilized by mobile operators to ensure service quality and manage customer experience better at terminal devices as opposed to the core of their network.

5.3 Future work

Fundamentally, our work so far has been about exploring the feasibility and potential of a hybrid measurement approach on user devices for the purpose of evaluating experienced quality at scale. Due to resource restrictions on mobile devices such as time, energy and bandwidth; measurement solutions always need to care about the value of the measurement itself compared to the costs. As we have explored the relationship between network measurements and the user-perceived quality in detail, it is evident that in order to achieve most value, future measurement efforts that target to estimate QoE via QoS has to be optimized end-to-end from sampling of the QoS measurements to QoE model. Following such an approach enables us to only measure the elements that provide information gain over what we already know and not waste resources on low-value measurements. In this respect, a centralized system can be established for the purpose of crowdsensing rather than crowdsourcing. Such a system can manage the parameters for network measurements with respect to contextual variables (e.g. location, device, signal levels, user-profile, current application, task) informed by the user device and measure only what's required to improve QoE approximation of the particular user. For instance, network measurements could be focusing on latency, jitter and packet loss for a video call given that bandwidth has been good enough historically for that particular location and radio technology.

Eventually, along with such improvements in scalability, information can be utilized to develop adaptive methods for improving user experience on the application level by projecting quality changes over time [40]. As a practical example, a study done by Borkowski et al. [9] shows that prefetching media content based on network quality prediction improves user-perceived delays up to 95%.

References

[1] V. Aggarwal, E. Halepovic, J. Pang, S. Venkataraman, and H. Yan. Prometheus: Toward quality-of-experience estimation for mobile apps from passive network measurements. In *Proceedings of the 15th ACM Workshop on Mobile Computing Systems and Applications*, HotMobile '14, pages 18:1–18:6, New York, NY, USA, 2014.

[2] A. Ahmad, L. Atzori, and M. G. Martini. Qualia: A multilayer solution for QoE passive monitoring at the user terminal. In *Proceedings of the 2017 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2017.

[3] S. Alfredsson, G. Del Giudice, J. Garcia, A. Brunstrom, L. De Cicco, and S. Mascolo. Impact of TCP congestion control on bufferbloat in cellular networks. In *Proceedings of the 14th IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 1–7, 2013.

[4] M. Alreshoodi and J. Woods. Survey on QoE\QoS correlation models for multimedia services. *AIRCC International Journal of Distributed and Parallel Systems (IJDPS)*, 4(3):53–72, 2013.

[5] E. Atxutegi, J. O. Fajardo, E. Ibarrola, and F. Liberal. Assessing Internet performance over mobile networks: from theory to practice. In *2016 ITU Kaleidoscope: ICTs for a Sustainable World (ITU WT)*, pages 1–8, 2016.

[6] J. Bergstra and C. Middelburg. The e-model, a computational model for use in transmission planning. *ITU-t recommendation G*, 107, 2006.

[7] A. Bhushan. *File Transfer Protocol (FTP) status and further comments*. IETF, RFC 414, Dec. 1972.

[8] T. Blajić, D. Nogulić, and M. Družijanić. Latency improvements in 3G long term evolution. In *Proceedings of the 2006 MIPRO International Convention on Information and Communication Technology, Electronics and Microelectronics*, pages 1–6, 2006.

[9] M. Borkowski, O. Skarlat, S. Schulte, and S. Dustdar. Prediction-based prefetch scheduling in mobile service applications. In *Proceedings of the 2016 IEEE International Conference on Mobile Services (MS)*, pages 41–48, 2016.

[10] P. Branco, L. Torgo, and R. P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2):31:1–31:50, 2016.

[11] P. L. Callet, S. Móller, and A. Perkis. Qualinet white paper on definitions of quality of experience. Technical report, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, 2013.

[12] P. Casas, A. D'Alconzo, F. Wamser, M. Seufert, B. Gardlo, A. Schwind, P. Tran-Gia, and R. Schatz. Predicting QoE in cellular networks using machine learning and in-smartphone measurements. In *Proceedings of the 9th IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, May 2017.

[13] P. Casas, B. Gardlo, M. Seufert, F. Wamser, and R. Schatz. Taming QoE in cellular networks: from subjective lab studies to measurements in the field. In *Proceedings of the 11th IEEE International Conference on Network and Service Management (CNSM)*, pages 237–245, 2015.

[14] P. Casas, R. Schatz, F. Wamser, M. Seufert, and R. Irmer. Exploring QoE in cellular networks: How much bandwidth do you need for popular smartphone apps? In *Proceedings of the 5th ACM SIGCOMM Workshop on All Things Cellular: Operations, Applications and Challenges*, AllThingsCellular '15, pages 13–18, 2015.

[15] J. Case, M. Fedor, M. Schoffstall, and J. Davin. *Simple Network Management Protocol (SNMP)*. IETF, RFC 1157 (Historic), May 1990.

[16] K.-T. Chen, C.-Y. Huang, P. Huang, and C.-L. Lei. Quantifying skype user satisfaction. In *Proceedings of the 2006 ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '06, pages 399–410, 2006.

[17] M. Crovella and B. Krishnamurthy. *Internet Measurement: Infrastructure, Traffic and Applications*. John Wiley & Sons, Inc., New York, NY, USA, 2006.

[18] M. Csikszentmihalyi and R. Larson. Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology*, pages 35–54. Springer, 2014.

[19] E. Dahlman, S. Parkvall, J. Skold, and P. Beming. *3G Evolution; HSPA and LTE for Mobile Broadband*. Academic Press, 2007.

[20] A. Developers. Android network capabilities. `https://developer.android.com/reference/android/net/NetworkCapabilities`. [Referenced: 2020-01-18].

[21] A. Developers. Android services. `https://developer.android.com/guide/components/services`. [Referenced: 2019-09-11].

[22] A. Developers. Android traffic stats. `https://developer.android.com/reference/android/net/TrafficStats`. [Referenced: 2019-09-11].

[23] A. Diaz, P. Merino, A. Gil, and J. Munoz. x-AppMonitor μAgent: a tool for QoS measurements in cellular networks. In *Proceedings of the 3rd IEEE International Symposium on Wireless Communication Systems*, pages 343–347, 2006.

[24] S. Egger, P. Reichl, T. Hoßfeld, and R. Schatz. "Time is bandwidth"? Narrowing the gap between subjective time perception and quality of experience. In *Proceedings of the 2012 IEEE International Conference on Communications (ICC)*, pages 1325–1330, 2012.

[25] Ericsson. Ericsson Mobility Report 2019. `https://www.ericsson.com/49d1d9/assets/local/mobility-report/documents/2019/ericsson-mobility-report-june-2019.pdf`, June 2019. [Referenced: 2019-09-11].

[26] European Commission and European Parliament, Brussels. Eurobarometer 87.1 (2017), TNS opinion, Brussels [producer], 2017.

[27] A. Faggiani, E. Gregori, L. Lenzini, V. Luconi, and A. Vecchio. Smartphone-based crowdsourcing for network monitoring: opportunities, challenges, and a case study. *IEEE Communications Magazine*, 52(1):106–113, 2014.

[28] M. Fiedler, T. Hossfeld, and P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. *IEEE Network*, 24(2):36–41, 2010.

[29] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. *Hypertext Transfer Protocol – HTTP/1.1*. IETF, RFC 2068 (Proposed Standard), Jan. 1997. Obsoleted by RFC 2616.

[30] J. Garcia, S. Alfredsson, and A. Brunstrom. Delay metrics and delay characteristics: A study of four Swedish HSDPA+ and LTE networks. In *Proceedings of the 2015 IEEE European Conference on Networks and Communications (EuCNC)*, pages 234–238, 2015.

[31] A. Gember, A. Akella, J. Pang, A. Varshavsky, and R. Caceres. Obtaining in-context measurements of cellular network performance. In *Proceedings of the 2012 ACM Conference on Internet Measurement (IMC)*, IMC '12, pages 287–300, New York, NY, USA, 2012.

[32] X. Geng, Z. Chen, W. Lam, and Q. Zheng. Hedonic Evaluation over Short and Long Retention Intervals: The Mechanism of the Peak-End Rule. *Wiley Journal of Behavioral Decision Making*, 26(3):225–236, 2013.

[33] J. Hosek, P. Vajsar, L. Nagy, M. Ries, O. Galinina, S. Andreev, Y. Koucheryavy, Z. Sulc, P. Hais, and R. Penizek. Predicting user QoE satisfaction in current mobile networks. In *Proceedings of the 2014 IEEE International Conference on Communications (ICC)*, pages 1088–1093, June 2014.

[34] T. Hoßfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler. The memory effect and its implications on Web QoE modeling. In *Proceedings of the 23rd IEEE International Teletraffic Congress (ITC)*, pages 103–110, 2011.

[35] T. Hoßfeld, M. Seufert, C. Sieber, and T. Zinner. Assessing effect sizes of influence factors towards a QoE model for HTTP adaptive streaming. In *Proceedings of the 6th IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 111–116, 2014.

[36] N. Hu and P. Steenkiste. Evaluation and characterization of available bandwidth probing techniques. *IEEE Journal on Selected Areas in Communications*, 21(6):879–894, 2003.

[37] J. Huang, C. Chen, Y. Pei, Z. Wang, Z. Qian, F. Qian, B. Tiwana, Q. Xu, Z. Mao, M. Zhang, et al. Mobiperf: Mobile network measurement system. *Technical Report. University of Michigan and Microsoft Research*, 2011.

[38] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. A close examination of performance and power characteristics of 4G LTE networks. In *Proceedings of the 10th ACM International Conference on Mobile systems, applications, and services*, pages 225–238, 2012.

[39] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck. An in-depth study of LTE: effect of network protocol and application behavior on performance. *ACM SIGCOMM Computer Communication Review*, 43(4):363–374, 2013.

[40] W. Hummer, S. Schulte, P. Hoenisch, and S. Dustdar. Context-aware data prefetching in mobile service environments. In *Proceedings of the 4th IEEE International Conference on Big Data and Cloud Computing*, pages 214–221, 2014.

[41] M. Jain and C. Dovrolis. End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput. In *ACM SIGCOMM Computer Communication Review*, volume 32, pages 295–308, 2002.

[42] M. Jaloun and Z. Guennoun. Wireless mobile evolution to 4G network. *SciRP Wireless Sensor Network*, 2(04):309, 2010.

[43] I. Järvinen, B. Chemmagate, A. Y. Ding, L. Daniel, M. Isomäki, J. Korhonen, and M. Kojo. Effect of competing TCP traffic on interactive real-time communication. In *Proceedings of the 2013 Springer International Conference on Passive and Active Network Measurement*, pages 94–103, 2013.

[44] Y.-F. Kuo, C.-M. Wu, and W.-J. Deng. The relationships among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. *Elsevier Computers in Human Behavior*, 25(4):887 – 896, 2009.

[45] B. Landfeldt, P. Sookavatana, and A. Seneviratne. The case for a hybrid passive/active network monitoring scheme in the wireless Internet. In *Proceedings of the 2000 IEEE International Conference on Networks (ICON 2000). Networking Trends and Challenges in the New Millennium*, pages 139–143, 2000.

[46] L. Lao, C. Dovrolis, and M. Sanadidi. The probe gap model can underestimate the available bandwidth of multihop paths. *ACM SIGCOMM Computer Communication Review*, 36(5):29–34, 2006.

[47] R. Larson and M. Csikszentmihalyi. The experience sampling method. In *Flow and the foundations of positive psychology*, pages 21–34. Springer, 2014.

[48] B. Melander, M. Bjorkman, and P. Gunningberg. A new end-to-end probing and analysis method for estimating bandwidth bottlenecks. In *Proceedings of the 2000 IEEE Global Telecommunications Conference*, volume 1, pages 415–420, 2000.

[49] K. Mitra, A. Zaslavsky, and C. Åhlund. QoE Modelling, Measurement and Prediction: A Review. *ArXiv e-prints*, 2014.

[50] V. Mohan, Y. J. Reddy, and K. Kalpana. Active and passive network measurements: a survey. *AIRCC International Journal of Computer Science and Information Technologies*, 2(4):1372–1385, 2011.

[51] W. Mohr and W. Konhauser. Access network evolution beyond third generation mobile communications. *IEEE Communications Magazine*, 38(12):122–133, 2000.

[52] S. Möller. *Assessment and prediction of speech quality in telecommunications*. Springer Science & Business Media, 2012.

[53] A. Morton. *Active and Passive Metrics and Methods (with Hybrid Types In-Between)*. IETF, RFC 7799, May 2016.

[54] Nettitutka. `http://www.nettitutka.fi/`, 2019. [Referenced: 2019-09-11].

[55] A. Nikravesh, H. Yao, S. Xu, D. Choffnes, and Z. M. Mao. Mobilyzer: An Open Platform for Controllable Mobile Network Measurements. In *Proceedings of the 13th ACM Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '15, page 389–404, New York, NY, USA, 2015.

[56] Ookla. Speedtest.net – the global broadband speed test. `http://www.speedtest.net/`, 2019. [Referenced: 2019-09-11].

[57] A. Özer, M. T. Argan, and M. Argan. The effect of mobile service quality dimensions on customer satisfaction. *Elsevier Procedia-Social and Behavioral Sciences*, 99:428–438, 2013.

[58] P. H. Perala, A. Barbuzzi, G. Boggia, and K. Pentikousis. Theory and practice of RRC state transitions in UMTS networks. In *Proceedings of the 2009 IEEE GLOBECOM Workshops*, pages 1–6, 2009.

[59] M. Peuhkuri. *Internet traffic measurements – aims, methodology, and discoveries*. Licentiate thesis, Helsinki University of Technology, Finland, May 2002.

[60] R. Prasad, C. Dovrolis, M. Murray, and K. Claffy. Bandwidth estimation: metrics, measurement techniques, and tools. *IEEE Network*, 17(6):27–35, 2003.

[61] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. Characterizing radio resource allocation for 3G networks. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet measurement*, pages 137–150, 2010.

[62] A. Razaghpanah, N. Vallina-Rodriguez, S. Sundaresan, C. Kreibich, P. Gill, M. Allman, and V. Paxson. Haystack: In situ mobile traffic analysis in user space. *CoRR*, abs/1510.01419, 2015.

[63] P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo. The logarithmic nature of QoE and the role of the weber-fechner law in QoE assessment. In *Proceedings of the 2010 IEEE International Conference on Communications (ICC)*, pages 1–5, 2010.

[64] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank. Factors influencing quality of experience. In *Quality of Experience: Advanced Concepts, Applications and Methods*, pages 55–72. Springer, 2014.

[65] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, B. Hendricks, and R. Baraniuk. Multifractal cross-traffic estimation. In *Proceedings of 2000 ITC Specialist Seminar on IP Traffic Measurement*, 2000.

[66] RootMetrics. Rootmetrics | measuring mobile performance for you, and with you. `http://www.rootmetrics.com/`, 2019. [Referenced: 2019-09-11].

[67] P. Royston. marginscontplot: Plotting the marginal effects of continuous predictors. *Stata Journal*, 13(3):510–527(18), 2013.

[68] Sandvine. The global internet phenomena report: 2h 2014. Technical report, Sandvine, 2014.

[69] R. Schatz and S. Egger. Vienna surfing: Assessing mobile broadband quality in the field. In *Proceedings of the 1st ACM SIGCOMM Workshop on Measurements Up the Stack*, W-MUST '11, pages 19–24, 2011.

[70] R. Schatz, M. Fiedler, and L. Skorin-Kapov. Qoe-based network and application management. In S. Möller and A. Raake, editors, *Quality of Experience: Advanced Concepts, Applications and Methods*, T-Labs Series in Telecommunication Services, pages 411–426. Springer International Publishing, 2014.

[71] M. Schöffler. *Overall Listening Experience - a new Approach to Subjective Evaluation of Audio*. PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2017.

[72] Y. Shen, Y. Liu, Q. Liu, and D. Yang. A method of QoE evaluation for adaptive streaming based on bitrate distribution. In *Proceedings of the 2014 IEEE International Conference on Communications Workshops (ICC)*, pages 551–556, 2014.

[73] A. Shuba, A. Le, E. Alimpertis, M. Gjoka, and A. Markopoulou. Antmonitor: System and applications. *CoRR*, abs/1611.04268, 2016.

[74] N. Sikes, K. Mason, and S. VonLehmden. Telecommuting advantages and challenges for it management and staff. In *Proceedings of the 49th ACM SIGMIS Annual Conference on Computer personnel research*, pages 20–25, 2011.

[75] S. Sonntag, J. Manner, and L. Schulte. Netradar – Measuring the wireless world. In *Proceedings of the 11th IEEE International Symposium on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt)*, pages 29–34, 2013.

[76] S. Sonntag, L. Schulte, and J. Manner. Mobile Network Measurements – It's not all about signal strength. In *Proceedings of the 2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 4624–4629, 2013.

[77] J. Strauss, D. Katabi, and F. Kaashoek. A measurement study of available bandwidth estimation tools. In *Proceedings of the 3rd ACM SIGCOMM Conference on Internet measurement*, pages 39–44, 2003.

[78] A. Toskala and H. Holma. *WCDMA for UMTS – HSPA Evolution and LTE*. John Wiley & Sons, fifth edition, 2010.

[79] O. Turel and A. Serenko. Satisfaction with mobile services in canada: An empirical investigation. *Elsevier Telecommunications Policy*, 30(5-6):314 – 331, 2006.

[80] N. van Berkel, D. Ferreira, and V. Kostakos. The experience sampling method on mobile devices. *ACM Computing Surveys*, 50(6):93:1–93:40, 12 2017.

[81] H. Verkasalo. Metrics that matter: Two new approaches from huawei connect europe. `http://www.vertoanalytics.com/metrics-that-matter-two-new-approaches-from-huawei-connect-europe/`. [Referenced: 2020-03-25].

[82] K. Wac, G. Pinar, M. Gustarini, and J. Marchanoff. Smartphone users mobile networks quality provision and volte intend: Six-months field study. In *Proceedings of the 16th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–9, June 2015.

[83] B. Weiss, D. Guse, S. Möller, A. Raake, A. Borowiak, and U. Reiter. Temporal Development of Quality of Experience. In S. Möller and A. Raake, editors, *Quality of Experience: Advanced Concepts, Applications and Methods*, T-Labs Series in Telecommunication Services, pages 133–147. Springer International Publishing, 2014.

[84] K. Winstein and H. Balakrishnan. TCP Ex Machina: Computer-generated Congestion Control. *ACM SIGCOMM Computer Communication Review*, 43(4), 2013.

[85] M. P. Wittie, B. Stone-Gross, K. C. Almeroth, and E. M. Belding. Mist: Cellular data network measurement for mobile applications. In *Proceedings of the 4th IEEE International Conference on Broadband Communications, Networks and Systems (BROADNETS'07)*, pages 743–751, 2007.

[86] D. Wu, R. K. C. Chang, W. Li, E. K. T. Cheng, and D. Gao. Mopeye: Opportunistic monitoring of per-app mobile network performance. In *Proceedings of the 2017 USENIX Annual Technical Conference (USENIX ATC 17)*, pages 445–457, Santa Clara, CA, 2017.

[87] J. Yao, S. S. Kanhere, and M. Hassan. An empirical study of bandwidth predictability in mobile computing. In *Proceedings of the 3rd ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization*, WiNTECH '08, pages 11–18, New York, NY, USA, 2008.

[88] Y. Zaki, T. Pötsch, J. Chen, L. Subramanian, and C. Görg. Adaptive congestion control for unpredictable cellular networks. *ACM SIGCOMM Computer Communication Review*, 45(4), 2015.

Aalto-DD 147/2020

ISBN 978-952-64-0058-7 (printed)
ISBN 978-952-64-0059-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Communications and Networking
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS