Czech Technical University in Prague

Faculty of Electrical Engineering
Department of Telecommunication Engineering

TRACKING USERS IN MOBILE NETWORKS: DATA ACQUISITION METHODS AND THEIR LIMITS

Doctoral Thesis

Michal Ficek

Prague, June 2013

Ph.D. Programme: Electrical Engineering and Information Technology
Branch of study: Telecommunication Engineering

Supervisor: Dr. Lukáš Kencl

Abstract

In this thesis we investigate tracking methods in mobile networks and their principal limits.

We study various methods of tracking (i.e., periodical positioning) of a number of mobile network users. We consider two basic options of tracking methods in mobile networks — terminal-based and network-based. Terminal-based techniques require user's cooperation and special hardware or software on the side of the localized mobile terminal. Network-based tracking, generally reaching all subscribers, is implemented in the network in either active or passive manner. Active tracking is based on queries of the network about the tracked device, whilst passive tracking uses operating data, which are generated and stored automatically by the network for all users.

Our original contribution to the area of *network-based active tracking* is a detailed study of a particular method of active tracking, the SMS-based one, using our proof-of-concept tracking platform connected to a live mobile network. Based on a large-scale measurement, we build a model of tracking process, and simulate how many users a tracking platform is able to track simultaneously without overloading the mobile network. Finding out that the principal limitations lie in the radio access network, we express the scalability of the method at various network-infrastructure levels and point out some pitfalls of active tracking, such as user mobility.

To the area of *network-based passive tracking* we significantly contribute by deriving a novel, probabilistic Inter-Call Mobility model, which overcomes the main limitation of passive tracking data — the poor temporal granularity of Call Data Records (CDRs). Our Inter-Call Mobility model spatio-temporally fits the aggregated mobility behavior of a large user-pool and significantly improves the CDR-based deduction of user's presence at some place in time: from the timestamped cell coordinates of mobile phone communication records, to a probabilistic distribution of user's position in between consecutive communication records, over time. On the example of user-proximity probability we demonstrate large disproportions in expected user's position in space and time among different mobility models, concluding that the Inter-Call Mobility model outperforms existing modeling techniques.

Finally, we investigate the limits of cooperative *terminal-based tracking* (crowdsensing) in discovering the mapping of mobile-network cell identifiers to geographic locations. Based on a real-world trace, we propose a novel data-driven mobility model to express the number of unique mobile-network cells a user is capable of visiting during one day. The model describes users' daily patterns, captures the fine-grained temporal characteristics of human movement during a day, and quantifies daily user-cell associations. Synthetic traces of user mobility from the model serve as an input for a large-scale simulation in an approximation of a mobile network. We show how crowdsensing may serve as a fight-back solution against a particular mobile-network-topology obfuscation method.

These three topics, studied in this thesis, illustrate the extent to which a particular technology or tracking method is applicable. Individually, they present relevant information about exploring and modeling human mobility based on mobile-network data, and the proposed mobility models can be used in future research.

Abstrakt

V této dizertační práci zkoumáme metody sledování uživatelů v mobilních sítích a limity těchto metod.

Pojmem „sledování“ rozumíme opakované zjišťování polohy účastníka mobilní sítě. Podle místa, kde probíhá zjištění a vyhodnocení pozice uživatele, rozlišujeme dvě základní metody sledování: terminálové a síťové. Terminálové metody využívají mobilního telefonu (terminálu) uživatele, k čemuž je obvykle zapotřebí spolupráce uživatele, nebo speciální technické či programové vybavení sledovaného zařízení. Síťové metody jsou realizovány uvnitř mobilní sítě, využívají jejich specifických možností a umožňují zjištění polohy obecně libovolného uživatele sítě. Způsob sledování pomocí síťových metod je dvojí: aktivní a pasivní. Během aktivního sledování se síť aktivně dotazuje na polohu účastníka, resp. jeho terminálu, pasivní sledování využívá provozních dat, která jsou v mobilní síti automaticky vytvářena a ukládána.

Náš původní přínos do oblasti aktivního síťového sledování spočívá v detailním zmapování konkrétní metody založené na posílání textových zpráv (SMS), kterou zkoumáme na existující implementaci připojené do mobilní sítě. Průběh zjištění polohy účastníka modelujeme pomocí diskrétní simulace s parametry získanými na základě rozsáhlého měření. Simulace slouží ke zjištění kritického množství uživatelů, které by bylo možné sledovat v síti najednou, bez negativních důsledků pro mobilní síť. Podrobně se věnujeme stěžejním limitům aktivního sledování, které pramení z omezené kapacity rádiového rozhraní přístupové sítě, a popisujeme, do jaké míry je nasazení aktivního sledování použitelné v aplikacích náročných na velký počet uživatelů.

K výzkumu v oblasti pasivního síťového sledování významně přispíváme vytvořením pravděpodobnostního Inter-Call Mobility (ICM) modelu, jenž popisuje pohyb uživatele mobilní sítě mezi místy, kde komunikoval. Tento model překonává základní nedostatek provozních dat mobilních sítí — záznamů volání (Call Data Records) — jejichž velké časové rozestupy znemožňují dostatečně přesné určení polohy uživatele kdykoli během dne. ICM model je založen na agregovaných datech pohybu mnoha uživatelů; významně rozšiřuje možnosti použití záznamů volání pro zjištění polohy uživatele mobilní sítě: od zaznamenaných poloh uživatele ve chvíli komunikace až po pravděpodobnostní rozložení předpokládané polohy mezi dvěma po sobě následujícími záznamy. Použití a výhody modelu ilustrujeme na příkladu odhadu pravděpodobnosti setkání uživatelů a ukazujeme, že ICM model překonává současné způsoby modelování pohybu uživatelů na základě záznamů volání.

V poslední části dizertační práce zkoumáme limity sledování účastníků za použití terminálových metod, konkrétně použití kooperativního sběru dat z mobilních telefonů, tzv. crowdsensing. Crowdsensing je často používán ke zmapování topologie mobilní sítě, tj. přiřazení geografických souřadnic jednotlivým buňkám sítě, a tím pádem k vytvoření databáze buněk pro lokalizační služby. Na základě rozsáhlého záznamu pohybu uživatelů v mobilní síti odvozujeme model, který umožňuje odhadnout počet buněk, které uživatel navštíví během jednoho dne. Tento model popisuje vzorce chování uživatele z pohledu jeho přesunů mezi buňkami mobilní sítě, zachycuje tendenci k přesunům v průběhu dne, a vyčísluje očekávaný počet buněk, ke kterým je terminál uživatele během dne připojen. Pomocí modelu generujeme charakteristické chování uživatele během dne a tento výstup simulujeme v uměle vytvořené topologii mobilní sítě. Výsledkem je odhad za jak dlouho může určitý počet uživatelů navštívit předem stanovené množství buněk v síti. Závěrem ukazujeme úspěšnost kooperativního sběru dat v případě, že je v mobilní síti použita konkrétní metoda zabraňující mapování poloh buněk.

Výše uvedená tři témata, kterými se zabýváme v této dizertační práci, ukazují, pro které aplikace a do jaké míry je konkrétní technologie sledování uživatele v mobilní síti vhodná. Jednotlivě prezentují řešené problémy významný přínos do oblasti zkoumání a modelování lidské mobility na základě dat z mobilní sítě. Představené modely mobility mohou být použity v další výzkumné práci.

Acknowledgments

First of all, I would like to express my gratitude to my advisor, Dr. Lukáš Kencl, for his support and research supervision during my studies. Without his persuasive arguments and his confidence in me I would never have chosen a research career. From him I have learned the art of writing a good research paper and presenting my work at the top conferences. His advice was always apt and valuable.

I thank Microsoft Research Cambridge for supporting me through its PhD Scholarship Programme. It meant great motivation and commitment for me, and enabled me to focus solely on research, the way I enjoy so much. I am extremely thankful to my Microsoft Research supervisors, Milan Vojnovic and Thomas Karagiannis, who have always found some time to discuss my results and future steps anytime I came to visit them in Cambridge.

I would like to thank my colleagues and friends at the Czech Technical University in Prague, namely, but in no particular order: Petr Vláčil, Jan Staněk, Jan Oppolzer and Petra Marešová for being friendly and funny officemates; Pavel Mach, Zdeněk Bečvář and Michal Vondra for being always there, just across the hall; Robert Bešťák, whose scientific problems were always a welcomed distraction from my work; and František Křížovský, for interesting and illuminating discussions about queuing theory. I also thank Boris Šimák and the staff of the Department of Telecommunication Engineering, who have been there to help me all along the graduate school.

I would never have managed to conduct the work on active tracking without a unique signaling platform, the SS7Box, built by many gifted colleagues of mine. I sincerely thank Kateřina Dufková, Jan Kouba, Petr Vláčil and Martin Tomek, who significantly contributed to the SS7Tracker application. Last but not least, I thank Vodafone Czech Republic a.s. for their support through the work of Michal Frýda, Martin Tomek, Tomáš Svoboda, Ladislav Soldán, Vladislav Veselý and Pavel Dušek.

I thank to all co-authors of our papers — they know it was not always easy to work with me, but they endured and we could celebrate every paper accepted and published. Gladly, I thank to all the anonymous reviewers of the papers I have co-authored over the years — their constructive comments and suggestions have helped to make this thesis stronger. I especially value and thank Srinivasan Keshav for his thorough shepherding of our MobiSys paper. Big thanks belong to Richard Fuller for he has always been helpful in discussing any idea in the field of location-based services with me.

My hearty thanks belong to my dear friend Kateřina Dufková, whose first-class output in every activity has always been motivating for me. I also thank Tomáš Pop for being good friend, dutiful co-author of our papers, and someone who has always been open to grumble with me at the sad fact that Ph.D. studies are not all roses.

Very special acknowledgment goes to my family, and especially my parents. It was truly relieving to know that they have always been there to help me by all means in case I would have experienced any trouble or hardship.

Finally, I deeply thank Pavla Kosinova and Hana Kosinova for their unconditional support and the faith they have kept in me. And to Hanka, I owe a peculiar thanks — for the never-ending stream of witty and artful text messages, she has been sending to me for years, and for her loving mockery of me.

Contents

List of Figures

List of Tables

List of Acronyms

AECID	Adaptive Enhanced Cell-ID
A-GPS	Assisted Global Positioning System
AOL	Age Of Location
API	Application Programming Interface
ATI	Any Time Interrogation
BSC	Base Station Controller
BSS	Base Station System
BTS	Base Transceiver Station
CDR	Call Data Record
Cell-ID	Cell Identifier
CGI	Cell Global Identity
CN	Core Network
CTU	Czech Technical University in Prague
E-OTD	Enhanced Observed Time of Difference
GPS	Global Positioning System
GSM	Global System for Mobile communication
GoS	Grade of Service
HLR	Home Location Register
ICM	Inter-Call Mobility
ICMP	Internet Control Message Protocol
IMSI	International Mobile Subscriber Identity
LAC	Location Area Code
LBS	Location Based Service
LBSN	Location-Based online Social Network

LMU	Location Measurement Unit
LTE	Long Term Evolution
MAP	Mobile Application Part
MSC	Mobile Switching Center
MSISDN	Mobile Subscriber ISDN Number
NSS	Network Switching Subsystem
OTDoA	Observed Time Difference of Arrival
PSI	Provide Subscriber Info
RDC	Research and Development Centre
RMD	Reality Mining Dataset
RMSE	root-mean-square error
RNC	Radio Network Controller
SDCCH	Standalone Dedicated Control Channel
SMS	Short Message Service
SMSC	SMS-Center
SRI	Send Routing Info
SS7	Signaling System Number 7
TA	Timing Advance
TCH	Traffic Channel
TDMA	Time Division Multiple Access
TRX	Transmitter/Receiver
UE	User Equipment
USSD	UnStructured Service Data
U-TDOA	Uplink Time Difference of Arrival
UMTS	Universal Mobile Telecommunications Service
VLR	Visitors Location Register

Chapter 1

Introduction

According to the International Telecommunication Union there were 6 billion mobile-phone subscriptions worldwide by the end of 2011, and mobile phone penetration topped 100% of population in many countries [111]. This huge worldwide mobile-phone pervasiveness is increasingly turning the mobile network into a gigantic ubiquitous sensing platform, enabling large-scale analysis and applications exploiting data acquired from the network. In recent years, mobile data-based research has reached important conclusions about various aspects of human characteristics, such as human calling patterns [99, 202], virus spreading [197, 102], social networks [40, 192, 213, 68], human daily activity patterns [164], urban and transport planning [194, 196], network design [211] and others. Particularly, information about movement of network users is of utter interest to many researchers [88, 212, 108, 95] because mobility description and prediction may have a profound effect on various fields of science, including telecommunications [64], human and time geography [21], urban studies [20] and energy-efficient networks [65]. Examples of practical applications of user-tracking using mobile phones are roaming optimization [66], tracking criminals and suspects [178], traffic-monitoring [41] and targeted advertising [136].

For the purpose of acquiring data about movement of users, we consider the task of simultaneously tracking a high number of mobile network users or, more specifically, of their mobile terminals. By *tracking* we mean collecting continuous information on the user's geographical position by means of various *positioning* techniques. The outcome of this process is a timestamped history of users' positions in the network and their geographical coordinates. The main problem with tracking in a mobile network is that suitability of existing methods for large-scale tracking, network-wide application, and their technological limits are often not discussed or remain unknown.

There are two principal technology options for tracking users in mobile networks:

- *Network-based methods* rely on the mobile-network infrastructure, which performs measurements and calculates the position of a subscriber;
- *Terminal-based methods* refer to the terminal's activity in carrying out the measurements and position calculation.

Network-based tracking methods hold the advantage that they are generally applicable to all network users. Two main approaches to network-based tracking can be used:

- *Active tracking* is based on queries of the network about the tracked device, so the network *actively* gathers information about users' terminal physical coordinates.
- *Passive tracking* uses operating data, such as Call Data Records (CDRs) or network logs, which are generated and stored automatically by the network for *all users* for billing and network troubleshooting reasons, without causing additional traffic in the network.

Terminal-based solutions are suitable for smartphones and evolving mobile devices, but often require user's cooperation or software installation, which prevents universal coverage.

1.1 Motivating Problems

The problem we solve in this thesis is to find and describe particular limits of the three tracking methods above—network-based active tracking, network-based passive tracking and terminal-based tracking.

Network-based active tracking delivers the position of mobile-network users with unprecedented temporal granularity, which has proved useful for tracking of criminals and suspects [178], and studies about mobility patterns [20] and urban dynamics [155]. However, it is a complex process that involves many nodes in the mobile network and wastes resources at the air interface between cell towers and a mobile terminal. Although computing power of network nodes and bandwidth are not usually limiting, the air interface is still a valuable resource that is hard to scale. The main question we try to answer is "How many users can be tracked simultaneously in a mobile network, and how often?" This is becoming a serious issue when thousands of subscribers are to be tracked in order to deliver representative and statistically significant studies of human mobility and location-based services. Knowing such limits would help to adjust the number of users or their tracking interval to a level that would not hamper the usual network traffic.

Network-based passive tracking has become immensely popular as a unique source of large-scale data about individual mobility [88], calling patterns [99], patterns of tourists' spatial mobility [21] and urban analysis [173], mainly because of the lack of appropriate active-tracking solutions or the cost of active tracking. Posterior interpretation of large-scale CDR datasets is perhaps even more important for purposes such as urban and transport planning [194, 196], network design [211], opportunistic spectrum access [202] and user mobility [95]. Unfortunately, whereas spatial precision of such passive-tracking data is acceptable, the accuracy in temporal dimension is substantially low. The main reason is that the position of a user is recorded *only* at places where the user's communication events occur (text messages, calls, data sessions), thus it depends on the communication frequency of an individual. During time of no communication activity it is not clear *where* the user is geographically located. This represents a problem for applications or analyses assuming ubiquitous and continuous user-tracking capability, such as opportunistic data dissemination [97] or epidemiology [102].

Terminal-based tracking remains a viable alternative to network-based methods when active-tracking infrastructure and passive tracking data are not available. It is advantageous in delivering rich data from contemporary sensor-enabled smartphones, and therefore is often sought after in academic research [68], but also in environmental [67], infrastructure [103, 147, 144] and social [69, 174] applications, in which interest groups actively participate on data collection—a method called *crowdsensing*. Crowdsensing proved useful especially in mapping fixed structure of mobile network cells and wireless-network access points to reference databases [52, 90], which are subsequently used for geolocation of mobile terminals and various other location-based services. The quantitative limits of such approach, however, remain in question: "What is the required minimal size of a user group needed for obtaining a critical mass of knowledge about the mobile infrastructure? And, how much time is needed to do so?" The answers may help in deciding whether crowdsensing is a viable solution for mapping country-wide networks infrastructure, for example for building a new geo-location business such as [52], or a hint for mobile network providers, whether they should try to monetize their costly infrastructure and apply mechanisms [62, 201, 30] preventing unauthorized use of such geo-location in their networks.

1.2 Objectives of the Thesis

We aim to deliver convincing results about limits of mobility-related data-collection methods, which may help understand the extent to which a particular technology or tracking method is applicable. The work presented in this thesis concentrates on addressing the above-mentioned problems and related limits. The primary objectives of this thesis are:

- Limits of network-based active tracking—to show, on an existing tracking solution in a live mobile network, how many users can be tracked simultaneously and how often. To understand the scaling ability of the active-tracking method in country-wide applications.
- Passive tracking data utilization—to examine the accuracy of network-based passive-tracking data in spatial and temporal dimensions and to propose an extension of the data towards more accurate interpretation.
- Coverage capabilities of crowdsensing—to assess how effective may crowdsensing be in mapping a mobile network infrastructure, measured in the size of a user-pool and the time needed to map critical mass of knowledge about the mobile network.

1.3 Outline of the Thesis

In Chapter 2, we introduce some basic background knowledge about mobile networks, together with a detailed description of different tracking methods and of the existing work related to particular topics investigated in this thesis.

In Chapter 3, we study the limits of network-based active tracking from various points of view, using an existing active-tracking solution. Based on a measurement on a large-scale tracking, we describe performance statistics of the tracking platform and of the adjacent network nodes. We implement a faithful model of the tracking process by means of a queuing network and run a discrete-time simulation which shows that our tracking solution is capable of tracking thousands of users with a period of minutes. We analyze various limiting factors of active tracking, including basic constraints of the method, constraints of the location server and of the mobile network, finding out that the principal limitation of network-based active tracking is the radio access network. We calculate the limitation of the radio access network on different network-infrastructure levels, and show how mobility of tracked users, and thus their possible concentration at one place, may affect the performance of the radio access network. Finally, we show our contribution on roaming optimization in mobile networks—a practical use-case of active tracking.

In Chapter 4, we address the limits of passive-tracking data—their limited accuracy in temporal dimension—by building a probabilistic model of users' position in between communication events. Through the analysis of coarse-grained CDR-based trajectories and corresponding finer trajectories of user-cell associations we show that the nature of human mobility between communication events is in strong contrast with assumptions of existing modeling methods. We formulate a new probabilistic *Inter-Call Mobility* (ICM) model, spatio-temporally fitting the aggregated mobility behavior of users in a real-world trace. Finally, on the example of user proximity probability we demonstrate that the ICM model outperforms different existing mobility models.

In Chapter 5, we study the limits of terminal-based tracking from the point of view of its potential for mapping of mobile network cells to geographic locations. Based on a detailed dataset, we build a model which describes user-cell association in a mobile network over a day. Using the model we generate thousands of synthetic yet realistic traces of user movement applied by a large-scale simulation to an approximated mobile network topology. By this we determine the fraction of mobile-network cells visited by a fixed number of users over a time interval. We apply the results to a practical use-case in which crowdsensing serves as a fight-back method against a particular method of mobile-network topology obfuscation.

Finally, in Chapter 6 we conclude and we outline possible future objectives of research work in the field of mobile data acquisition and utilization.

Chapter 2

Preliminaries

This chapter presents the general background knowledge and the state-of-the-art on which our work is based on.

Section 2.1 briefly introduces the structure of mobile networks. In Sections 2.2 and 2.3 we summarize the main approaches to network-based and terminal-based tracking, respectively. In Section 2.4 we review previous research and related works in the areas relevant to this thesis.

2.1 Mobile Network Primer

A mobile network is a wireless network made up of a number of radio cells—the basic geographic units of cellular systems—each served by at least one fixed-location transceiver. In this section we show the basic structure of mobile networks, location management and positioning methods, on the example of a Global System for Mobile communication (GSM) network. Although the share of contemporary and evolving networks such as Universal Mobile Telecommunications Service (UMTS) and Long Term Evolution (LTE) is constantly increasing, GSM technology has by far the widest reach and covers more than 85 percent of the world's population today [74].

2.1.1 GSM Network Structure

The network structure is divided into the Network Switching Subsystem (NSS) (often called *core network*) and the Base Station System (BSS) [96], see Figure 2.1.

The Network Switching Subsystem is a wired backbone that enables basic functionality of the mobile network—communication between mobile terminals and with ter-

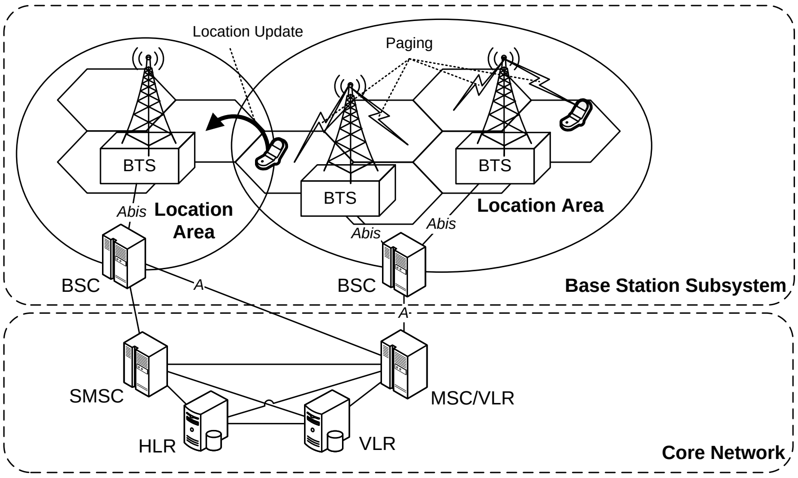


Figure 2.1: GSM network structure.

minals in other networks. Core network consists of the Mobile Switching Center (MSC), which are primary service delivery nodes responsible for handling voice calls and other services. A special type of MSC is a SMS-Center (SMSC), which supports sending and receiving text messages—Short Message Service (SMS). In addition, various databases are present in the core network: the Home Location Register (HLR) is a central database that manages information about users authorized to use the network; Visitors Location Register (VLR) are databases of temporary data about users currently present in a particular set of cells.

The Base Station System, which consists of the Base Transceiver Stations (BTSs, also known as cell towers or simply base stations) and Base Station Controllers (BSCs), operates the radio part of the network and handles traffic and signaling between mobile terminals and the core network. Each cell in the network is uniquely identified by a Cell Global Identity (CGI) number which consists of four numeric fields: Mobile Country Code, Mobile Network Code, Location Area Code and Cell Identifier (Cell-ID).

GSM networks are controlled by signaling protocols, carried out-of-band, in separate signaling links that use message switching [117]. Signaling protocols used in telecommunication networks worldwide are grouped in the Signaling System Number 7 (SS7) standard [109].

2.1.2 Location Management in Mobile Networks

At any time a mobile terminal is connected to a network over a single serving cell. However, for bandwidth-saving and overhead reasons, networks do not continuously maintain information about terminal association at the level of individual cells. More complex mechanisms are thus used to find a cell that the user is currently attached to.

Location management in cellular networks incorporates the *location update* and *paging* mechanisms (see Figure 2.1). *Location update* is a process of reporting the mobile terminal's serving cell to the network and storing its code in network registers. This is performed when a user crosses boundaries of the so called *location areas* (these are geographically large, consisting of tens to hundreds of cells) or after a significant time (on the order of hours). *Paging* is a procedure of actively searching for a terminal within the last known location area. It is performed every time the mobile-terminated communication (call, SMS, data) is being established. Therefore, when no communication is in progress, only the location area in which the mobile terminal idles is known. The accurate knowledge of a serving transceiver in the network is thus guaranteed to be up-to-date *only when paging is performed or the user is actively communicating.*

2.1.3 Positioning in Mobile Networks

Positioning, the act of obtaining an approximate geographical position of a mobile terminal and its user in the mobile network, can be achieved by several techniques. These can be classified as network-based or terminal-based, depending on the site that performs measurements and calculates the position. A hybrid approach, called terminal-assisted, is possible when the measurements are made by the terminal and the position is calculated by the network.

Terminal-based positioning methods require special hardware or software in the terminal and usually rely on external satellite infrastructure.

The best known representative is the *Global Positioning System* (GPS) [114], which provides high accuracy but poor availability (as observed in [126]), mainly because GPS does not work inside buildings.

Assisted Global Positioning System A-GPS [114], utilizes additional assistance information from the mobile network which leads to shorter time to fix the first position of the terminal. It provides better accuracy than GPS and increased receiver sensitivity.

Location pattern matching or *fingerprinting* [112, 47], uses received signal strength from mobile base stations or wireless access points nearby to match it with terminal's physical coordinates. Reference database of the beacons' patterns and corresponding positions allows for estimating the terminal position with high accuracy [126].

Less accurate, but always available when in mobile network signal coverage, is the *Cell-ID positioning* [124]. Every mobile terminal knows the Cell Identifier (Cell-ID) of the cell it is currently attached to, and thus no additional hardware is needed. Since the mobile terminal does not have the information about cell tower coordinates, it is provided by third parties' Cell-ID databases [52, 18].

Terminal-assisted positioning methods, such as *Enhanced Observed Time of Difference* (E-OTD) [14] in GSM networks or *Observed Time Difference of Arrival* (OTDoA) [12] in UMTS networks, work on a combination of circular or hyperbolical lateration with timing measurement – the terminal computes its position from signals emitted by a number of base stations. To achieve this, a Location Measurement Unit (LMU) has to be deployed at every (or every third, fifth [1]) base station in the network for measuring time offsets and achieving a posteriori synchronization between base stations.

Network-based positioning methods differ in the extent of network updating needed and face the trade-off between positioning accuracy and implementation costs.

Uplink Time Difference of Arrival (U-TDOA) [16, 13] is based on hyperbolic lateration similar to the terminal-assisted E-OTD method. But conversely to it, the time measurements are applied in the uplink, which means that the signal emitted from a terminal is observed by the network. Just as for the E-OTD method, the network has to be equipped with LMUs.

The *Cell-ID* method relies on the fact that the mobile terminal can be attached to only a single cell at a time. Knowing the Cell-ID, the actual physical coordinates of a mobile terminal within the cell can be approximated by the position of the corresponding cell tower, or estimated from the expected cell coverage. The accuracy of the reported position varies, depending on the cell size [189].

The *Timing Advance* (TA) [3, 15] improves accuracy of Cell-ID positioning by combining it with auxiliary measurements—TA is used in GSM/GPRS networks to compensate for the propagation delay as the signal travels between a mobile terminal and a serving base station, and thus roughly corresponds to the distance between them. The Cell-ID+TA enhancement is, however, limited by the fact that they are always calculated by the serving base station, so the distance information can not be obtained from three or more base stations to enable lateration.

The last option, the *Adaptive Enhanced Cell-ID* (AECID) [200] method, combines knowledge about Cell-IDs of the serving base station and of the neighboring cells, the measured TA and the quantized signal strength measurements to automatically build a radio fingerprint of the whole network. The AECID method yields better accuracy in comparison with the plain Cell-ID+TA method [181], but is more complicated and it utilizes a dedicated BSSAP-LE [17] protocol, which must be implemented in the network.

2.2 Network-based Tracking Methods

Two main approaches to network-based tracking can be recognized: *passive* and *active* [20]. *Active tracking* is based on network-based positioning queries of the network about the tracked device, so the network actively gathers information about users' terminal physical coordinates or its presence in end nodes of network segments, for example in cells. Conversely, *Passive tracking* uses operating data, such as billing records or network logs that are generated and stored automatically by the network for billing and network troubleshooting reasons.

2.2.1 Active Tracking

To obtain the user's actual position, the network must actively seek it, which brings additional traffic. The easiest way to approximately locate a user within a mobile network is to obtain the Cell-ID of a cell the user is currently attached to. To track a mobile terminal using Cell-ID positioning means that information about the current serving cell must be kept up-to-date to be obtained from the network registers at any time, which *is not* a default feature of the mobile network (as explained in Section 2.1.2). To achieve this, the network must be forced to page the terminal every time the position is requested, to propagate the Cell-ID of its current serving cell into the network registers. Such proactive stimulation of the mobile terminal is called *active tracking*.

There are several options how to force a network to perform terminal paging. The latest standards of contemporary mobile networks propose signaling primitives such as the Any Time Interrogation (ATI) message in GSM [10] or the paging request in UMTS or LTE networks [11], that are capable of triggering the terminal paging procedure. Nevertheless, these primitives are often enabled in the core network infrastructure only according to former standards. Their advanced capabilities are not necessary for providing the basic call and data services and are thus considered expendable.

A general method to force a network to page *any* terminal is to *invoke the mobile terminal's communication.* In [66] we proposed an *SMS-based* solution that delivers an "invisible" network-originated text message to the user's terminal.

The **SMS-based solution** exploits a special class of text messages, *SMS Class 0* (Message Waiting Indicator Group [6]), which is usually used to set the indication of voice mail, fax and e-mail active or inactive. The key advantage is that its delivery cannot be prevented or rejected by the user's mobile terminal. Moreover, SMS has been supported by *all* mobile networks and *all* mobile terminals since it became available in a very early version of GSM.

Apart from the SMS-based solution above, other methods of invoking mobile terminal communication activity could be considered:

Data oriented approach refers to the fact that one could try to send an ICMP echo request (known as a *ping*) to the user's IP address to make the mobile terminal communicate, which forces the network to page the terminal and leads to the current Cell-ID information propagation among core network registers. This method, however, is only applicable to data-active users.

USSD oriented approach uses UnStructured Service Data (USSD) that provides a two-way session-oriented exchange of textual data in mobile networks. USSD is a capability of all terminals [2]. An approach that exploits USSD messages for providing location information is described in patent [31]: an empty USSD message can be sent to a mobile terminal which results in a paging procedure initiated by the network.

CAMEL protocol [4] provides the Any Time Interrogation (ATI) and Provide Subscriber Info (PSI) signaling messages that would include the "current location" and "active location retrieval requested" parameters respectively to immediately invoke the paging procedure [154]. Nevertheless, in spite of other useful features, a full implementation of the CAMEL protocol is not commonly present in mobile networks.

Fake handover procedure can be established by a BSC by sending a handover command to a mobile terminal to simulate the necessity of handover [16]. The terminal tries to co perform handover procedure, however the BSC suppresses the handover acknowledgment—after a certain period the procedure is evaluated as failure but communication with the mobile phone has been established.

2.2.2 Passive Tracking

Passive mode of user tracking means that no additional data transition within the mobile network is needed in order to obtain user's position. This technique either relies on the network logs about users' communication events, or an installation of a passive probe on network interfaces (connection links between executive parts of the network) is required.

Call Data Records (CDRs) store an information about every mobile phone's call (received or made) or service used (SMS, MMS, data). Its fields contain the communication type (voice, SMS), timestamp, duration, calling and called party numbers, etc., but also the Cell Global Identity and thus also the approximate spatial information where the communication has been established and where it terminated. CDRs are stored in network registers and used for billing purposes and legal reasons. They are considered strictly private, however, they are often shared between mobile providers and researchers in an anonymized or aggregated form [20, 21, 88, 107, 34].

Interface Monitoring can provide useful information about users' location and communication activity within a mobile network. The *A* interface between the MSC and the BSC (see Figure 2.1) is capable of capturing location updates in the network, handovers between cells during calls, and the Cell-ID of the cell the user was attached to when a call (SMS) has been made (sent) or received. A probe on the *Abis* interface between the BTS and the BSC can provide the average signal quality, received signal strength and timing advance (TA). These are network parameters that can be used to calculate distance of the mobile phone from the corresponding BTS and thus can serve for better positioning estimation, as explained in Section 2.1.3.

2.3 Terminal-based Tracking Methods

Terminal-based tracking, a particular use of mobile phone sensing [127], is a process of continuous collection of geographical coordinates from the mobile terminal's (A)GPS, Cell-IDs of cells the terminal is attached to, or beacons from the surrounding wireless access points, out of the position of the mobile terminal is then calculated or estimated. To enable terminal-based data collection, numerous sensing applications and platforms are available to maintain data recording and its offload to a remote server to process. Simple mobile-network cell-switching can be recorded for example by CellTrack [81] or Netmonitor [161] applications. Complex sensing systems include for example Mobile Context Toolbox [128] for Symbian S60, and Funf [29] and MobiSens [204] for Android. Mobile sensing platforms such as PEIR [150] and Medusa [169] use mobile terminals to collect and upload data to server-side models that generate the web-based output for each participatory sensing participant and enable achievement of a collective task, respectively.

An obvious disadvantage of software add-ons is that they reduce the target user group of any data-collection campaign to cooperating users, only precluding universal population coverage. Even the quality of terminal-based data is often disputable: users tend to forget their mobile terminals at home, battery drain precludes data collection for hours till the next recharge, and even malicious users can do harm to data integrity.

Promising research activities span the area of crowdsensing and participatory sensing. *Crowdsensing* [85] refers to a process of collecting data from mobile phones that has become popular in the academic environment [29] and various interest groups [90]. As an addition to it, *participatory sensing* represents the vision of distributed data collection and analysis at personal, urban, and global scales, in which participants make key decisions about what, where, and when to sense [36].

In recent years, Location-Based online Social Networks (LBSNs) [215], such as Google Latitude [89] or Foursquare [83], have become very popular and even giving rise to a plethora of new research work [23, 198]. These applications drive the evolution of geo-location based services, but require software installation, enlisting in a particular LBSN, and rely on mobile terminals or even on direct user activity. The granularity of user-triggered tracking with LBSNs is much more limited than for example network-based periodic active tracking.

2.4 Related Work

In this section we provide the overview of the existing work related to particular topics investigated in this thesis.

2.4.1 Active Tracking

In Chapter 3 we demonstrate an implementation of SMS-based active tracking in mobile networks, a method that utilizes Cell-ID based positioning. We explore limits of the method and show its practical application.

Positioning of mobile entities in networks is a well-studied problem. Works [171] and [124] summarize the main approaches to positioning, including the Cell-ID technique. Alternative positioning techniques are discussed in [168, 126, 184, 210]. Comparison of network-based positioning techniques is available from sources [73, 91, 124, 162, 167, 190].

The limits of Cell-ID positioning in terms of positioning accuracy are discussed in [189, 209]. Authors of [20] provide a short study of Cell-ID+TA-based active-tracking accuracy as a trade-off between the tracking interval and tracking costs. The impact on power consumption of a tracked mobile terminal has been measured and discussed in [63]. In this thesis we focus on the properties of active tracking with Cell-ID positioning and its impact on the network in general.

Recent works describe different methods of mobile terminal activity excitation, a necessary prerequisite for active tracking. A similar approach to the SMS-based method has been used in [178] under the name "blind SMS". Apart from methods summarized in Section 2.2.1, a method of using signaling primitives in mobile network for mobile phone positioning was proposed in [55]. However, the authors propose a modification of the existing signaling primitives, which is almost impossible in a live mobile network. Compared to all options available, SMS-based active tracking we use in our work has important advantages over other network-based active tracking methods. Data oriented approach is more demanding because a complicated radio connection would have to be established, so the method would use radio resources much more extensively. USSD method [31] is generally faster than SMS, but an SMS-based solution offers the additional advantage of recognizing whether the mobile terminal is out of its home network and thus reduces interconnection costs and wasting of network resources. Other methods such as CAMEL-protocol approach [4] and fake handover procedure [16] require dedicated signaling protocols, which are not usually present by default in all networks. The SMS-based solution described in this thesis is advantageous in that it can be used in any mobile network technology such as GSM, UMTS and LTE, since the Short Message Service is supported across all types of mobile networks.

There have been significant standardization efforts [7, 157, 148] and corporate initiatives [35, 73] for Location Based Service (LBS). A prominent example of a LBS platform is the Ericsson Mobile Positioning System [72], which complies with the latest LBS standards. In comparison, the tracking solution presented in this thesis is simpler, but lightweight and deployable by adding only a single node into the network.

A proof-of-concept tracking solution can be built on various signaling platforms. These comprise complex, ready-to-deploy solutions [105], as well as developer-oriented solutions [100, 57]. Development on these platforms is either provided by the vendor or limited by the platform Application Programming Interface (API). Unlike these robust and business-ready platforms, we use off-the-shelf hardware equipped with basic signaling configuration—a solution which is intended to be lightweight and suitable for proof-of-concept applications in research and academic environment.

The use of active-tracking data in mobility and human activity research is still sporadic, mainly for the lack of active-tracking solutions and the costs. Researchers in Tallinn used active tracking to investigate daily-activities and mobility patterns of city inhabitants and commuters [20], but found the method costly. Similar research, based on data collected using the active-tracking platform we describe in this thesis, was focused on urban dynamics [155]. A study of mobility patterns was presented in [88]. Schmitz [178] used the SMS-based active tracking for tracking of criminals and suspects. Although the possibility of studying roamer retention in GSM network were discussed in [63, 66], in this thesis we propose the formalization of the tracking process and visualization of the data using a novel cell-weakness metric.

2.4.2 Passive Tracking and Available Datasets

In Chapter 4 we deal with extending the utility of passive tracking data by building a mobility model for a-posteriori analysis of Call Data Records.

Two principal sorts of passive tracking data exist—location updates and CDRs [199]. Both methods work with Cell-ID-precision in spatial dimension, but differ in the frequency in temporal dimension, depending on user's mobility and calling patterns.

Monitoring location updates proved to be helpful in transportation for automatically deriving origin–destination matrices in a studied region [39] and in various applications of intelligent transportation systems [120]. However, location-update data are strongly limited by user mobility since they are recorded only when a user crosses borders of location areas[1].

In recent years, mobile operators shared Call Data Records with researchers and academia either directly [88, 173, 21, 211, 214], or in data-mining contests [158, 34, 143].

[1] Periodic location updates apply, but the interval is in the range of tens of minutes to several hours, depending on network settings.

Nevertheless, although many dataset resources exist, for example in [121], a publicly available large-scale CDR dataset is still not present, except for few recordings of individual enthusiasts [33, 78]. For the reasons above in this thesis we use a substitution for CDRs derived from a real-world trace, the Reality Mining Dataset [68].

Our approach in extending the utility of passive tracking data by describing the user's movement between the places of communication is novel. There were several attempts in describing movement of an entity in space and time between two places in general, for example by means of linear weighted interpolation [92], space-time prisms [94] and probabilistic variants of space-time prisms [203]. However, the assumptions behind these models are contradictory to the nature of Call Data Records, as we demonstrate in Chapter 4. Gonzalez at al. [88] analyze human mobility patterns from CDRs and deliver a general point of view on the nature of human movement, whilst we focus on a finer level of detail. The idea behind our methodology is similar to the work of Pfoser and Jensen [163] [163]—they proved that increasing the sampling rate of GPS lowers the localization error. We adopt their conclusion as we compare coarse-grained call records with finer-grained ground-truth data from cell transitions to deliver more accurate movement description. The uncertainty in user's trajectories has been studied in [123, 122, 61], but with respect to classic time-geography, which hardly applies to CDRs. The description of spatio-temporal trajectories using a Gaussian Mixture Model, similar to the method we use in this thesis, was demonstrated in [42] on training humanoid robots.

Many authors focused on limitations of CDRs from different points of view. Ranjan et al. [172] discuss a potential source of bias in CDRs for human-mobility studies, authors of [88] observed a heavy-tailed distribution of time between user's communication events, Zang and Bolot [212] analyzed the privacy risks of sharing CDRs. In this thesis we focus on the principal limitation of Call Data Records, their poor temporal granularity, and address it with a probabilistic refinement in both temporal and spatial dimensions.

2.4.3 Crowdsensing and Modeling Human Mobility

In Chapter 5 we explore the limits of crowdsensing in the ability to map a mobile network topology. In other words, we aim at quantifying the sufficient amount of users to discover all cells in a mobile network.

Crowdsensing represents an important tool for collecting network associations and activity of mobile phones by volunteer individuals. An example of a publicly available dataset is the Reality Mining Dataset (RMD) [68], other datasets are bound with legal consent [143] or were kept unpublished in a raw form [156]. We use the NRC-Lausanne dataset, which has been made available for participants of the Mobile Data Challenge [143], to obtain information about mobility of network users.

There were numerous attempts to describe user's movement. Random models, such as Random Way Point, Random Walk [130], Random Direction and Truncated Levy Walk [175], use some probability distribution to determine next user's waypoint. Variants of random models, Markovian Way Point [104] and Gaussian-Markov [137] models, introduce Markovian probability description between waypoints. Freeway and Manhattan mobility models [28] and the Obstacle Model [113] incorporate geographic constraints to restrict the movements of users. Social models describe mobility of particular social groups, such as student at a university [119, 146], or user's attraction to particular landmarks or geographical nodes [139]. Except the mobility models mentioned so far, human mobility patterns are described in [88], or the models are extended to better capture human nature in visiting similar places in Self-similar Least Action Walk (SLAW) model [131].

In this thesis, we propose a trace-based mobility model for user-cell association that differs from the existing large body of mobility-modeling work in that it describes users' daily patterns, captures fine-grained temporal characteristics of human movement during a day and quantifies daily user-cell associations. According to a recent survey [115], there is no mobility model available that provides all the above-mentioned features concurrently. Existing models are used to predict future places of user's presence [64] or to recognize significant places in the mobility trace [125, 208], but these work usually on short-term outlook. On the contrary, the model presented in [182] captures user-place association and the strength of such ties, but it is designed for long-term mobility predictions on the order of months. Models focusing on *important places* reflect the fact that humans usually spend their time only at few locations during a day. However, spatial prediction often takes priority over temporal prediction [48]. The NextPlace model [176] claims to forecast users' next place as well as their arrival time and residence time. Similar results can be achieved by using a time-variant community model [101]. Closest to our approach in mobility modeling is the methodology recently proposed in [179], which examines human mobility on the basis of motifs from network theory [26]. Similarly to the WHERE [108] and Time-variant community (TVC) [101] models, we use a real trace and generate new, synthetic traces; but WHERE uses CDRs and focus only on metropolitan scales, and TVC demands a large number of parameters, which were not derivable from the dataset we used.

Limits of crowdsensing have been studied in [205] from the perspective of hardware heterogeneity, burden placed on users and network bandwidth demands. The problem of finding a large-enough user-pool for crowdsensing applications is addressed with various incentive schemes [135, 207]. Energy-efficiency in continuous sensing has been examined in [140, 142, 53] and addressed with various energy-efficient sensing solutions [133, 151, 118, 216]. The capability of a userpool in gathering data and other tasks has been addressed in [191] by developing statistical tools for reasoning about trade-offs between time and completeness. In this thesis we explore the limits of crowdsensing in its ability to map a mobile network topology, which may be interesting for various crowdsensing communities [90, 52, 18], mobile network providers [116] and hardware vendors [35].

The given problem, i.e., characterizing sufficient number of users to cover a given fraction of cells in a cellular network during a limited time, is related to some classical problems: the cover time, the coupon collector's, and the cardinality estimation problems.

The Cover Time Problem

The *cover time problem* represents coverage the capability of a random walk on a graph. Let us consider an undirected graph $G = (V, E), |V| = n, |E| = m$, and let the probability to go to any of neighbors of u be the same, and is $\frac{1}{deg(u)}$. Let τ_n denote cover time, the expected number of steps of a random walk to visit every vertex of the graph starting at an arbitrary node in a graph with n nodes. It holds that for any graph with n nodes $\tau_n \geq (1 + o(1))n \log n$ and $\tau_n \leq (1 + o(1))4/27n^3$ [75], and, for example, a complete graph has $\tau_n \approx n \log n$.

Crowdsensing in a mobile network can be modeled as cover time of u independent random walks, representing movement of u users, in a mobile network represented by an undirected graph—vertices denote cells and edges stand for possible handovers between cells. A direct application of cover time results to the crowdsensing capabilities is, however, limited by several reasons. First of all, it has been demonstrated that random walks hardly suffice to describe human's mobility [88, 131, 182]. Second, transition probabilities between cells are not time-homogeneous as they vary depending on the time of the day. Next, time in a random walk is described by the number of steps between nodes, whilst in real life, the number of cell transitions does not necessarily correspond to the time measured in hours and minutes. Finally, because more users are involved in crowdsensing, a study of independent random walks traversing the graph concurrently would be necessary. Although some work on the impact of multiple random walks on coverage time exist [54, 24], they are restricted to random walks on a graph.

The Coupon Collector's Problem

The *coupon collector's problem* is the following: Let us have a set of n different kinds of coupons. Assume a person wants to collect all the n different kinds, by randomly picking one coupon with probability $1/n$, until all different kinds have been selected at least once. What is the expected number of coupons that need to be picked to collect at least one coupon of each type? It can be shown, that the expected number of coupons drawn until at least one coupon of each type is present is $n \cdot H_n$, where $H_n = \sum_{i=1}^{n} 1/i$ denotes the n-th Harmonic number. Since it holds $\log n \leq H_n \leq \log n + 1$, we basically have to buy $n \log n$ lots. The general case of a nonuniform probability distribution are more complicated [177, 82].

Direct application of the coupon collector's problem to our problem related to crowdsensing could be formulated as follows: let us have a finite set of users and let each user discover exactly one cell out of n cells equally likely. How many users do we need to discover all cells in the network? This formulation, however, does not reflect reality because a user of a mobile network is typically connected to more than one cell during the day. So, let us assume that a user discovers exactly k cells during a day. This corresponds to an extension of the coupon collector's problem in which the coupons are drawn into random packs of k different coupons [76]. Then, the expected number of users to discover all cells in the network corresponds to the expected number of packs we have to draw to collect all n coupon kinds. Since it is not likely that each user would discover exactly the same number of cells k, we found the extended coupon collector's problem not applicable.

The Cardinality Estimation Problem

The *cardinality estimation problem* (also known as *distinct counting*) is a fundamental problem of estimating the number of distinct values in a set from a small sample. This is important for example in the database community to estimate the number of distinct values in a table [93], as well as in biology to estimate the number of species [45], or in crowd-sourced enumeration queries to reason about query completeness [191]. The problem is as follows: a sample is drawn at random from a population (e.g. the entire population of animals in an area) and based on the frequency of observed items (species), the number of unobserved items (number of missing species) is estimated. Several methods have been proposed in the last decades, including sampling-based approaches [93], and extrapolation techniques [51].

Let us assume we have a group of users who use their mobile phones for discovering cells in a mobile network. The cardinality estimation helps us to guess what is the total number of unique cells in the network, so when can the cell mapping be considered finished. This is, however, difficult due to non-uniformities in the arrival of the crowd-sensed data, i.e., users typically discover only a limited number of cells and they visit some of the cells more likely than others ones. The role of the different distributions of human's responses has been studied in [191]. The authors use the *Chao92* estimator [45], which uses sample coverage and coefficient of variance of the underlying distribution to to predict the total number of distinct elements, and make it more resilient against highly skewed distributions by eliminating significant outliers in the data. Cardinality estimation thus can be used to reason about the total number of cells in a network, but the number of users necessary to discover all the cells in the network would remain unknown.

Chapter 3

Active Tracking in Mobile Networks

Active tracking is a promising approach in tracking *any* mobile-network user with unprecedented temporal granularity. In this chapter we present the SMS-based active tracking—a particular type of network-based active tracking using Cell-ID positioning with SMS-based paging excitation. Cell-ID positioning represents the easiest way to approximately locate a user within a mobile network. It is available in all environments (indoor, rural and urban) when the terminal is on signal. Moreover, Cell-ID positioning and SMS-based paging excitation do require neither any dedicated functionality in the mobile terminal not any update to the mobile network, thus they work with any mobile terminal in any mobile network.

SMS-based active tracking enables numerous services, such as tracking suspects and offenders [178], mobility and human activity research [66, 20], mobile network user tracking for signal coverage diagnostics and roaming optimization [66, 63]. However, implementation and deployment of this method brings questions about its limits in terms of the number of users that can be tracked simultaneously and the impact on network performance when a large-scale tracking would run in the network. In this chapter we focus on addressing these concerns by carrying out an extensive study of the limits of SMS-based active tracking in mobile networks.

In the initial Section 3.1, we describe the principle of SMS-based active tracking. In Section 3.2 we demonstrate functionality of this method by describing our academic proof-of-concept implementation running in a live GSM network. Next, Section 3.3 shows the fundamental limits of the baseline implementation by simulating the tracking process and comparing the results with theoretical laws. In Section 3.4 we show the limiting factors of SMS-based active tracking from various points of view, including fundamental constraints of the method, the location server constraints and the mobile network constraints. Issues regarding radio access network capacity, scalability and user mobility are examined in detail. Finally, Section 3.5 shows a particular use-case of active tracking in a mobile network—improvement of the roaming traffic.

3.1 Principle of SMS-based Active Tracking

We briefly describe the SMS-based tracking process in a GSM network, according to works [63, 66, 80].

The Cell-ID of a cell currently serving a user can be obtained by sending three MAP[1] primitives in the network. The only input is the Mobile Subscriber ISDN Number (MSISDN) of the user whose Cell-ID is to be retrieved. Figure 3.1 shows the types of messages and their ordering, described in detail as follows:

First, the VLR currently maintaining the user record in the network needs to be found, which is done by the Send Routing Info (SRI) request from the Location Server to the user's HLR (message 1). If the user's mobile phone is off, the HLR responds with an error, otherwise the response message (message 2) contains the user's International Mobile Subscriber Identity (IMSI) and VLR number. The mobile network of the current user's residence (home network, abroad, rival operator) can be determined from the VLR number.

Second, if both the positioned user and the Location Server reside in the same network[2], the messageflow continues with a request to send an "invisible" *SMS Class 0* to the user by the SMS-Center, using the *Forward Short Message* (SMS) (messages 3–4) message. The process of SMS delivery (messages 5–18), performed by the SMSC in cooperation with HLR, VLR, MSC and the SMS recipient, is a standard GSM procedure. It involves *paging* the user's mobile terminal (messages 9–14) which results in updating the location information (CGI of the cell where the user is located) in the VLR (messages 12).

[1] Mobile Application Part (MAP) is the topmost part of the SS7 stack, it enables applications in the GSM core network

[2] It is possible to position the mobile terminal in an arbitrary network, however, because of interconnection costs between rival networks and restrictive agreements between rival operators, the messageflow usually stops at this point when the mobile terminal is located in the rival network.

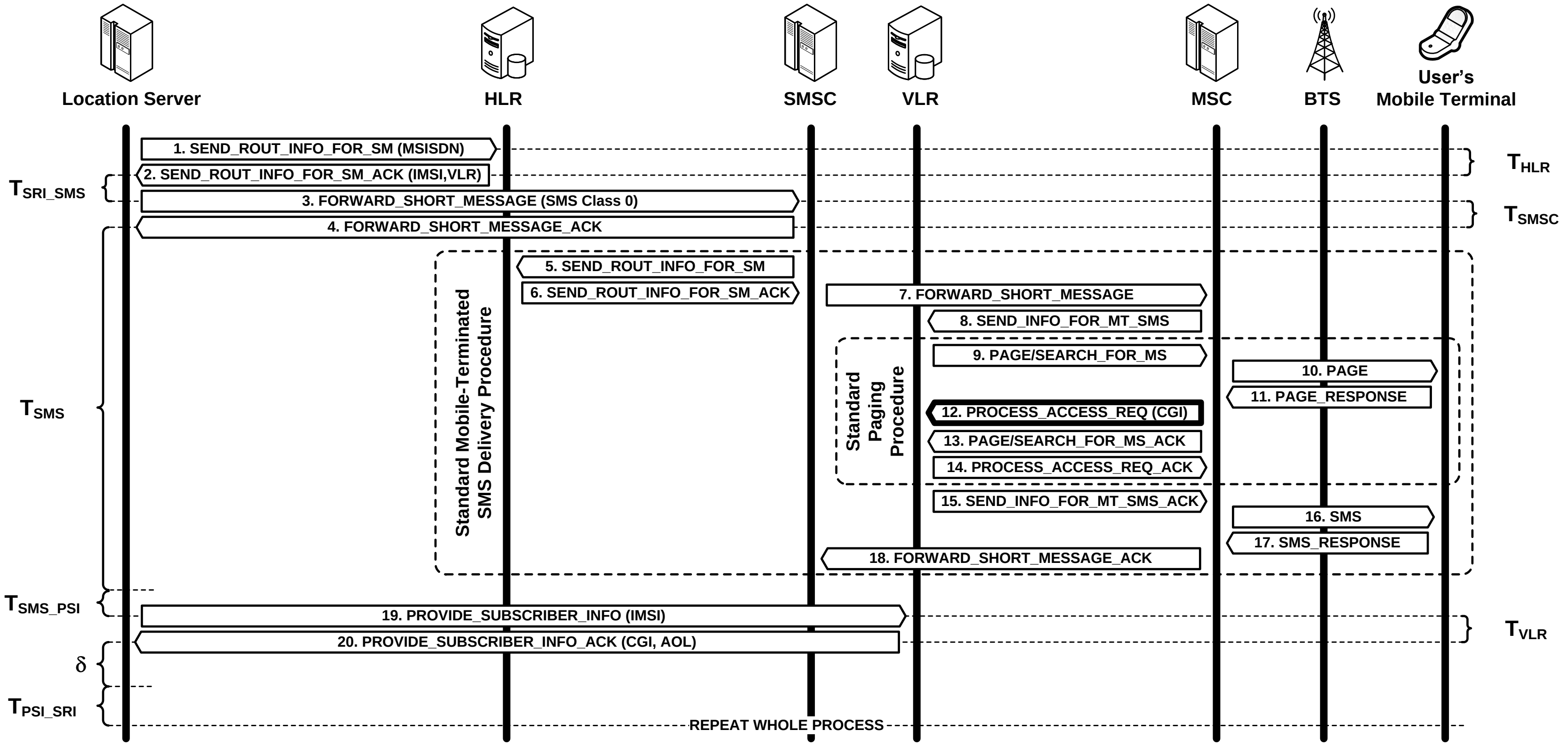

Location Server
HLR
SMSC
VLR
MSC
BTS
User's Mobile Terminal
1. SEND_ROUT_INFO_FOR_SM (MSISDN)
2. SEND_ROUT_INFO_FOR_SM_ACK (IMSI,VLR)
3. FORWARD_SHORT_MESSAGE (SMS Class 0)
4. FORWARD_SHORT_MESSAGE_ACK
5. SEND_ROUT_INFO_FOR_SM
6. SEND_ROUT_INFO_FOR_SM_ACK
7. FORWARD_SHORT_MESSAGE
8. SEND_INFO_FOR_MT_SMS
9. PAGE/SEARCH_FOR_MS
10. PAGE
11. PAGE_RESPONSE
12. PROCESS_ACCESS_REQ (CGI)
13. PAGE/SEARCH_FOR_MS_ACK
14. PROCESS_ACCESS_REQ_ACK
15. SEND_INFO_FOR_MT_SMS_ACK
16. SMS
17. SMS_RESPONSE
18. FORWARD_SHORT_MESSAGE_ACK
19. PROVIDE_SUBSCRIBER_INFO (IMSI)
20. PROVIDE_SUBSCRIBER_INFO_ACK (CGI, AOL)
REPEAT WHOLE PROCESS
Standard Mobile-Terminated SMS Delivery Procedure
Standard Paging Procedure
T_SRI_SMS
T_SMS
T_SMS_PSI
δ
T_PSI_SRI
T_HLR
T_SMSC
T_VLR

Figure 3.1: Message flow for obtaining the current Cell-ID of a user. The vertical lines represent time, note the definitions of time intervals $T_\star$ between the messages: $T_{\mathrm{SRI\_SMS}}$, $T_{\mathrm{SMS\_PSI}}$, and $T_{\mathrm{PSI\_SRI}}$ represent working times of the Location Server between different types of messages, T_{HLR}, T_{SMSC}, and T_{VLR}, denote service response times of adjacent network nodes, δ stands for the tracking interval.

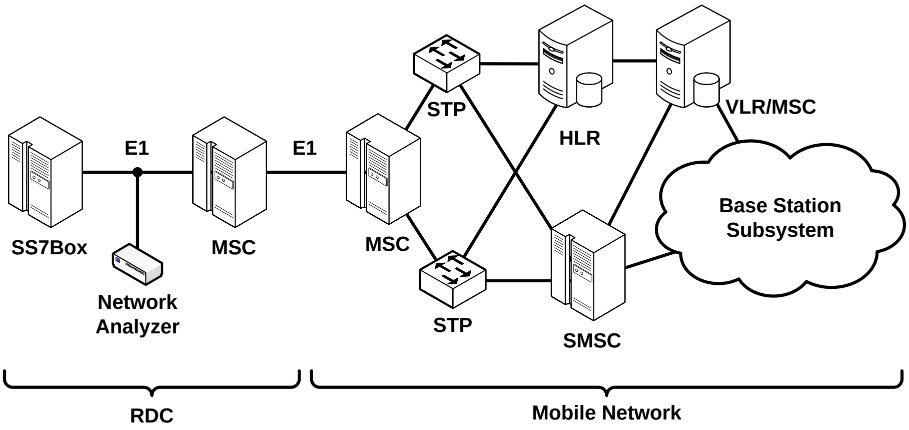


Figure 3.2: Interconnection of SS7Box into a mobile network.

Finally, an up-to-date Cell-ID and Age Of Location (AOL) is retrieved using a Provide Subscriber Info request (messages 19–20). The AOL value contains age of the location information in minutes; it is set to zero if the SMS was successfully delivered to the mobile terminal.

3.2 Implementation and Deployment

In this section we describe the implementation of an active tracking application called *SS7Tracker*, which we have built during our Master degree as a proof-of-concept of SMS-based active tracking [63, 66].

The SS7Tracker has been implemented on top of a modular signaling platform for fast telco application prototyping, called SS7Box, developed at Research and Development Centre (RDC) at the Czech Technical University in Prague [193]. The SS7Box is a universal signaling platform based on the SS7 protocol suite on top of which application modules realize business logic of telecommunication-oriented applications. A detailed description of the SS7Box is available in [80].

The SS7Box is interconnected with a live GSM network of Vodafone Czech Republic, and utilize no additional hardware or software elements in the core network. It resides in the RDC perimeter, together with its own MSC to which it is connected using one E1 line with one Pulse-Coded Modulation timeslot. The line is monitored by Tektronix K1205 network analyzer (see Figure 3.2).

Figure 3.3 shows the relation between the SS7Tracker modules, the services they implement are described as follows.

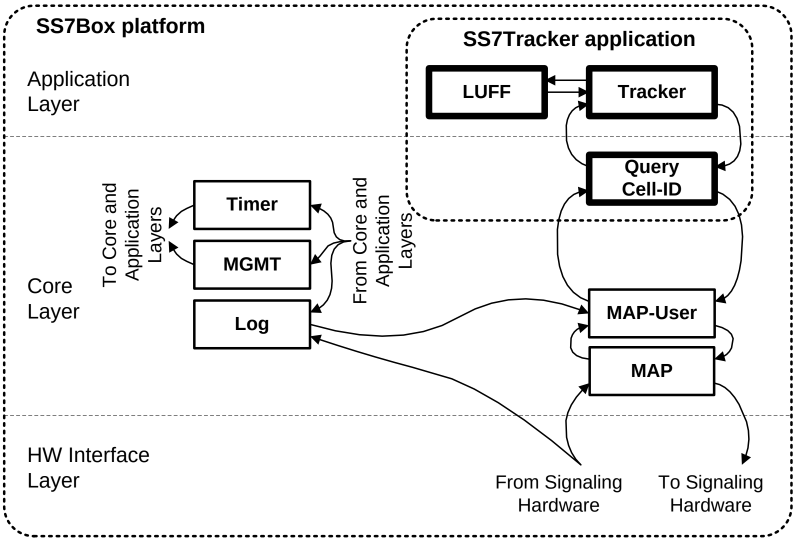


Figure 3.3: SS7Box architecture. Modules and important data flows of SS7Tracker.

The *Cell-ID retrieval service*, implemented in the *Query Cell-ID* module, deals with getting the location information from the network. It works according to the process described in Section 3.1.

The *Tracking service* periodically requests the Query Cell-ID module for the user's current cell identifier and returns a timestamped history of user-cell associations. The service is implemented in the *Tracker* module.

The *User selection service*, implemented as part of the *Location Update Feed Filter (LUFF)* module, provides the numbers (MSISDNs) of all users currently subscribed to the network. The list is based on Location Update events which are recorded on-line by the network operator during user movement between different location areas. The LUFF module filters the MSISDN feed according to the target group for tracking, and forwards it to the Tracker module.

Apart from those presented, the SS7Box platform runs other modules: *MAP* and *MAP-User* handle the SS7 and MAP protocol messages, and *Management*, *Timer* and *Log* are system modules maintaining operation of the platform.

3.3 SS7Tracker Performance Evaluation

In this section we examine the performance of the SS7Tracker in terms of the maximum number of users to track simultaneously.

The tracking process of a set of users is characterized by two main parameters. These are (1) the number of unique users tracked, denoted N, and (2) the tracking interval δ, i.e., the time between consecutive Cell-ID retrievals per single user. While it is usually desired to maximize the number of users served, the tracking interval typically depends on the tracking purpose and varies from tens of seconds to several minutes [66, 13].

We present the performance evaluation in the sense of a minimum deployment unit—the SS7Box's interconnection to the GSM network is realized by only a single E1 line with one timeslot (link) for signaling. This limits the data rate to 64 kbit/s and, moreover, the link utilization policies must be applied. According to ITU-T recommendation [110], each signaling link should provide extra capacity and thus its utilization ρ must not exceed a value of maximum utilization ρ_{max}, which usually lies between 0.2 and 0.4. Thus, only such a tracking process that utilizes the interconnection link to less than ρ_{max} is allowed.

In the rest of this section we provide estimation about how many users it is possible to track simultaneously, and how often, while utilizing the interconnection line at safe level. Since it may be harmful for the network to arbitrarily set the tracking parameters, measure the interconnection link utilization and adjust the parameters a posteriori, we propose a simulation model for system performance evaluation under different tracking parameters. First, we analyze network behavior, i.e. distribution of relevant service response times (Section 3.3.1). Afterward, simulation using measured distributions is performed to find values of tracking parameters N and δ that meet the interconnection link utilization limits (Section 3.3.2). Finally, the simulation is validated against link utilization measured by the network analyzer during tracking (Section 3.3.4).

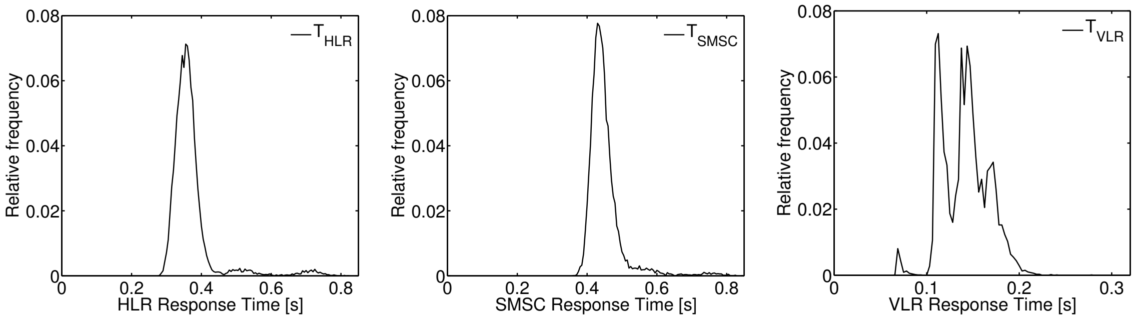
3.3.1 Tracking Measurement

To familiarize ourselves with time dimension values in the Cell-ID retrieval message flow (Figure 3.1), we have executed several measurements. Traces from the network analyzer monitoring the signaling link interconnecting SS7Box into operator mobile network have been analyzed in order to retrieve all time durations $T_\star$ from Figure 3.1 as well as lengths of the sent and received messages. We focused particularly on *service response time* distributions of adjacent network nodes (HLR, SMSC, VLR), SS7Box *working time* and *message lengths variance.* Except where explicitly specified, all data in this section come from a single measurement during 6 hours on a sample of 500 users with a tracking interval of 2 minutes. Thus, the total count of sent and received messages is about 700,000.

Table 3.1 summarizes network service response times of adjacent network nodes (HLR, SMSC, VLR), Figure 3.4 shows the corresponding distributions. The length of

	Network service response time [s]				
	min	**median**	**0.95-q**	**max**	**stdev**
T_{HLR}	0.2822	0.3567	0.4918	0.8212	0.0046
T_{SMSC}	0.3613	0.4397	0.5328	0.8434	0.0027
T_{VLR}	0.0667	0.1418	0.1824	0.2790	0.0006

Table 3.1: Measured network service response time



(a) *HLR* **(b)** *SMSC* **(c)** *VLR*

Figure 3.4: Measured distributions of network service response times. The distribution of T_{VLR} is multimodal, in contrast to the other distributions. Detailed analysis [80] of PSI messages revealed that the PSI response messages are sent by eight different Mobile Switching Centers, which are geographically distributed within the entire network, thus the response time depends on location of the currently tracked user in the network.

request messages sent from the SS7Box to the network is constant, the length of the response messages varies—see Figure 3.5. Interestingly, the network service response time for messages of various lengths differ more than the theoretical transmission time necessary to transmit a larger message. Figure 3.6 depicts and explains the difference in network service response time according to message length.

The working time of SS7Box modules, i.e., time between receiving message response from the network and sending a new request message, is summarized in Table 3.2. Different types of messages are processed by the SS7Box side with different distributions as depicted in Figure 3.7.

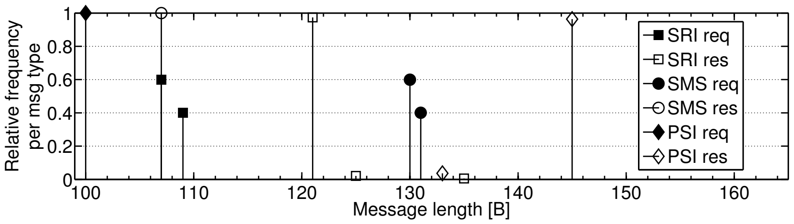
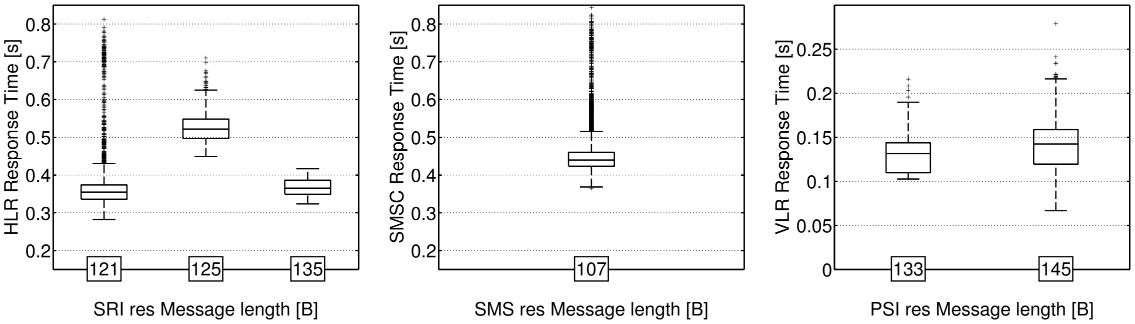
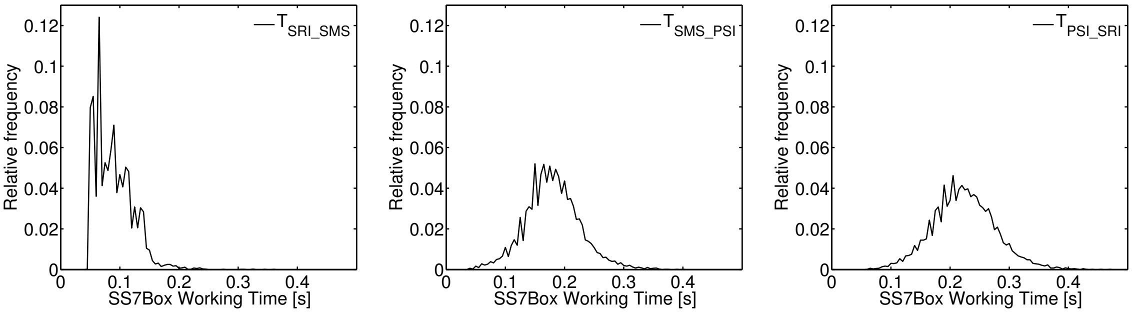


Figure 3.5: Message length distribution per message type. The lengths of request messages sent from SS7Box differ in one or two Bytes according to the MSISDN parameter length and message padding. The length of response messages differ significantly as a result of network nodes technological diversity and settings—some network nodes generate additional message fields.



(a) *HLR* **(b)** *SMSC* **(c)** *VLR*

Figure 3.6: Measured service response time as a function of response message length. The main difference for HLR response time results from different message processing on the network-node side, noticeably for messages of length 125 Bytes. These messages contain additional "protocol version info" field, which takes longer time to obtain from the core network.



(a) **(b)** **(c)**

Figure 3.7: Measured distributions of SS7Box working time. Time between receiving a response from the network and sending a new request. Distribution of $T_{\mathrm{SRI\_SMS}}$ differs from $T_{\mathrm{SMS\_PSI}}$ and $T_{\mathrm{PSI\_SRI}}$ because the SMS request is sent immediately after the SRI delivery whilst between other types of messages a system timer is called. As depicted in the messageflow in Figure 3.1, between SMS and PSI messages a fixed T_{SMS} delay is set to enable the SMS propagation in the network, and between PSI and SRI messages are divided by the tracking interval δ. The timer call results in context switching and re-scheduling of the module process in the operating system, thus a normal-like distribution can be expected.

	SS7Box working time [s]				
	min	**median**	**0.95-q**	**max**	**stdev**
$T_{\mathrm{SRI\_SMS}}$	0.0474	0.0859	0.1418	0.3583	0.0010
$T_{\mathrm{SMS\_PSI}}$	0.0385	0.1779	0.2585	0.4006	0.0021
$T_{\mathrm{PSI\_SRI}}$	0.0591	0.2220	0.3123	0.4696	0.0028

Table 3.2: Measured SS7Box working time

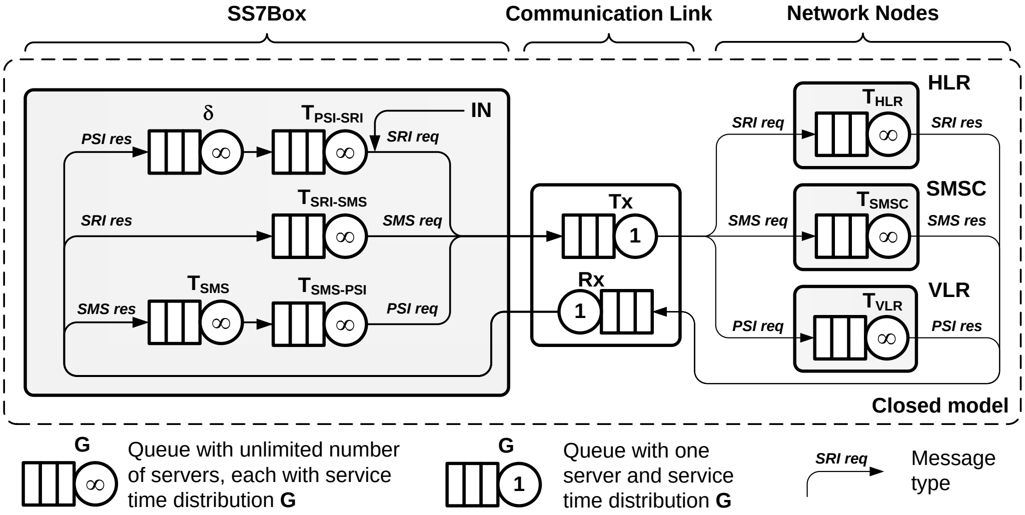


Figure 3.8: Simulation model. The communication link is modeled as a pair of queues for transmission and reception direction respectively, each with one server that processes messages with time of service equal to the ratio of message length and data rate. Network nodes are modeled as queues with unlimited number of servers thus no waiting time is applied. SS7Box working times, T_{SMS} time and tracking interval δ are modeled in the same way.

3.3.2 Simulation of the Tracking Process

We have modeled the communication process during active tracking in Matlab R2008b and SimEvents Library 2.3 using discrete-time queuing network simulation, schematically depicted in Figure 3.8. The model is *probabilistic* and *closed*, explained as follows.

The arrival process to each queue in the queuing network represents a particular message type in the active tracking message flow (recall Figure 3.1). Initial seed of processes corresponds to the tracking settings: the number of processes equals to the number of tracked users, all processes enter the model at the point denoted “IN”. Seeded arrival processes keep circulating in the model, entering the network-nodes queues and SS7Box queues according to the particular message type, and being delayed in the queues according to the particular service time distributions (Figures 3.4 and 3.7). The length of the message is selected from the corresponding distribution per message type (Figure 3.5). Signaling link utilization in one direction is interpreted as *server utilization* ρ , i.e., the proportion of the time the server is busy. According to [110], the direction (transmission or reception) with higher utilization stands for the overall signaling link utilization.

The reason for using a discrete-time simulation instead of application of queuing theory comes from the fact that common signaling link characteristics calculation based on M/G/1 models, recommended by [110], is not viable because of a violation of inter-arrival

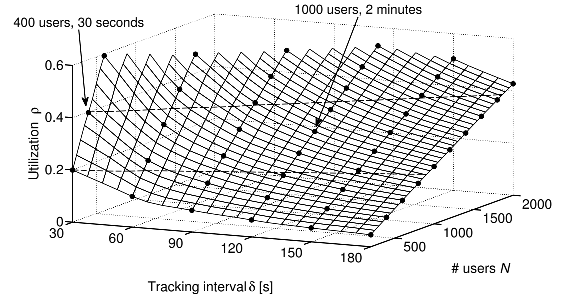


Figure 3.9: Signaling link utilization. Meshgrid interpolates simulation output values, represented with round points. Dashed lines mark limiting values of allowed utilization maximum ρ_{max}.

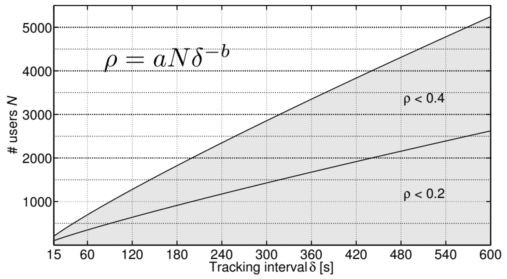


Figure 3.10: Area of safe operation. Each combination of number of users N and tracking interval δ falling into the area of safe operation (in gray) is feasible.

time distribution and the arrival process assumptions. The most general G/G/1 model is also inappropriate because the inter-arrival process distribution of Cell-ID retrieval messages is too complex and, besides, some of the service time distributions at network nodes and SS7Box working time are even multimodal (see Figures 3.4 and 3.7), making the summary statistics as the mean and standard deviation of service time deceptive.

3.3.3 Simulation Results

Figure 3.9 shows the dependency of signaling-link utilization ρ on the number of tracked users N and the tracking interval δ. As expected, a shorter tracking interval or an increasing number of tracked users cause higher utilization. Within the relevant intervals $\delta \in [30, 180]$ and $N \in [200, 2000]$, ρ can be closely approximated by a function that depends linearly on the number of users N and is inversely proportional to the tracking interval δ:

$$\rho = aN\delta^{-b}. \tag{3.1}$$

Using robust linear least-squares fitting, we found the values of coefficients to be $a = 0.0208 \pm 0.0003$ and $b = 0.8768 \pm 0.0032$ (95% confidence interval). The fit statistics $\mathrm{R}^2 = 0.9998$ indicate that the fit explains 99.98% of the variance, and the near-zero root mean squared error RMSE = 0.0020 implies that the fit is useful for prediction.

Figure 3.10 indicates which combinations of the number of users and tracking intervals are allowed with respect to ρ_{max}. We conclude that SS7Tracker implementation of SMS-based active tracking yields sufficient performance to track thousands of users with a period of minutes.

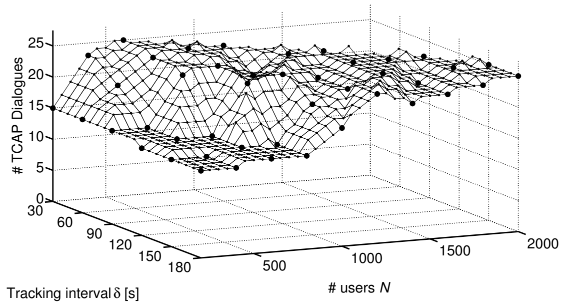


Figure 3.11: TCAP dialogs utilization. Meshgrid interpolates simulation output values, represented with round points. As number of dialogs is a discrete value, interpolated values are rounded up.

The simulation allows studying attributes not covered in standard queuing theory, such as the limits of signaling card in the SS7Box in terms of concurrently established outgoing dialogs. The hardware limit of the Dialogic SPCI4 card we use is 4,096 simultaneous outgoing MAP/TCAP dialogs [58]. This can be considered a significant limit, however, Figure 3.11 shows that the number of communication dialogs established at the same time is surprisingly not a concern at all.

3.3.4 Model Validation and Discussion

We validated the key simulation output, the signaling link utilization, against both live measurement and theoretical foundations.

We compared the results with measurements taken by the Tektronix K1205 analyzer Erlang application that computes signaling link utilization on-line, during live tracking. Utilization has been simulated with exactly the same input parameters as recorded on the analyzer during the tracking. The comparison between the simulated utilization and the measured utilization found the simulation result relative error to be 3.98%.

For the theoretical validation we used the Utilization Law by Buzen [38] that calculates the utilization U of an arbitrary device, *without* any assumption on inter-arrival time or service-time distributions, as:

$$U = \overbrace{\left(\frac{\text{Completions}}{\text{Total Time}}\right)}^{\text{Throughput}} \times \overbrace{\left(\frac{\text{Busy Time}}{\text{Completions}}\right)}^{\text{Mean Service Time}} = \frac{\text{Busy Time}}{\text{Total Time}} \tag{3.2}$$

The theoretical value of utilization for the same parameters as in the measured tracking, computed from this Utilization Law, is about 2.51% and 6.39% higher than the measured and the simulated utilization, respectively. This difference we attribute to real user behavior during the real tracking: their migration between local, rival and foreign mobile networks leads to an incomplete message flow and thus a lower number of sent messages (as described in Section 3.1). A lower number of messages naturally leads to lower utilization of the signaling link.

The operation region of the model, established by the measurement presented in Section 3.3.1, could be a source of inaccuracy between simulation and a real, live measurement on a bigger tracking extent (for example, 5,000 users). We claim that the simulation provides an *upper bound* of interconnection link utilization and that it will provide similarly accurate results even if the simulation ran with input parameters different from its operation region. Should we admit that one can hardly expect shortening of service and working times when tracking more users, the potential increase of the measured T_* values will definitely not affect negatively the key simulation performance characteristics: the signaling link utilization. This conclusion comes from the *closed* nature of the simulation model: every new location retrieval of a user can be provided if and only if the previous request for the same user was processed. Thus, the simulated utilization is the upper bound even if the T_* values rise. Longer service time decreases the total number of Cell-ID retrievals (”Completions” in Equation 3.2) during the observation time, but with no effect on utilization (see Equation 3.2). Moreover, fewer SMS per minute will be sent, which is favorable for the network. Although the number of TCAP dialogs, opened at the same time, will rise because of the longer message round-trip in the network, the total count of available dialogs may be considered acceptable.

3.4 Limiting Factors of SMS-based Active Tracking

In the previous section we have presented limits of a particular active-tracking solution. This section surveys the fundamental limits of SMS-based active tracking from the network perspective in terms of the minimal tracking interval size, interconnection to the network, constraints in the core and radio access networks, and scalability of the method.

3.4.1 SMS-based Active Tracking Constraints

The SMS-based active tracking principle is limited in terms of the achievable minimum time between two consecutive Cell-ID retrievals. The tracking history of a

	Time [s]				
	min	**median**	**0.95-q**	**max**	**stdev**
T_N	0.7595	0.9405	1.0714	1.7284	0.0109
T_{LS}	0.2167	0.4923	0.6422	0.8204	0.0077
T_{SMS}	3.6514	4.6888	5.8253	5.9649	0.2913

Table 3.3: Active tracking time characteristics

mobile terminal is a timestamped sequence of Cell-IDs, in which the time interval between consecutive timestamps t_i, t_{i+1} is not constant, i.e.,

$$|t_{i+1} - t_i| = \delta + T_N + T_{LS} + T_{\mathrm{SMS}}, \tag{3.3}$$

where δ denotes the fixed tracking interval, T_N denotes the variable network response time, T_{LS} denotes the variable Location Server working time, and T_{SMS} is the SMS delivery delay. While the tracking interval δ can be arbitrarily small, the network response time

$$T_N = T_{\mathrm{HLR}} + T_{\mathrm{SMSC}} + T_{\mathrm{VLR}} \tag{3.4}$$

is a sum of service response times of the adjacent network nodes and thus depends on the mobile network architecture and technology. Similarly, working time of the Location Server

$$T_{LS} = T_{\mathrm{SRI\_SMS}} + T_{\mathrm{SMS\_PSI}} + T_{\mathrm{PSI\_SRI}} \tag{3.5}$$

is dependent on implementation. Finally, the T_{SMS} delay is set to a fixed value on the order of seconds during which the SMS is most likely to be delivered.[3]

Based on the experimental tracking measurement from Section 3.3.1, Table 3.3 summarizes the total network response time and the SS7Box's working time in variables T_N and T_{LS} respectively. The delay corresponding to the SMS delivery process, T_{SMS}, has been measured during a shorter experimental measurement on 64 text messages with SMS delivery report enabled.

Considering the maximal values of the measured network response time ($T_N \approx$ 2 s), the SS7Box working time ($T_{LS} \approx 1$ s) and the SMS delivery time ($T_{\mathrm{SMS}} \approx 6$ s), we find that, for a zero-length tracking interval $\delta = 0$, the minimum time between two consecutive Cell-ID retrievals is limited to an approximate value of 9 seconds in our baseline

[3] Although an SMS-delivery report message may be used to inform the Location Server that the SMS has been delivered, and thus the location information is updated and can be requested, the fixed T_{SMS} saves valuable bandwidth of the signaling link.

implementation. Such a value is not limiting for the vast majority of tracking applications [66, 178, 5], which are satisfied with longer response times (2 minutes and more).

3.4.2 Location Server Constraints

System-wide limits of the SS7Box platform, on top of which runs the SS7Tracker active tracking application, are determined mainly by operational memory usage, CPU utilization and signaling hardware limits. We examined these areas in detail in study [80], concluding that demands of the SS7Tracker on RAM and CPU are negligible. Hardware limits of the Dialogic SPCI4 signaling card are not a concern either—the card limit in terms of the number of simultaneous active outgoing dialogs (up to 4,096 according to [59]) is far from being even approached for most of the reasonable combinations of tracking parameters, as shown in Section 3.3.3, Figure 3.11.

The principal limitation of the Location Server is its connection to the mobile network. We express this constraint in terms of the number of location retrievals that can be made through the minimal interconnection option, which is one timeslot with 64 kbit/s data rate. Assuming a zero-error condition on the link and no other communication proceeding concurrently on the signaling link during the tracking, the maximum number of location retrievals during a time period depends only on the signaling link speed. Let L_{tx} (L_{rx}) denote the sum of length of all messages transmitted (received) over the link during one single location retrieval. Then, the number of location retrievals over time period T and a signaling link with data rate S equals

$$\lfloor TS\rho_{max}/L \rfloor \,, \tag{3.6}$$

where ρ_{max} is the maximal allowed signaling link utilization and $L = \max(L_{tx}, L_{rx})$. According to the ITU-T [110], the direction (transmission or reception) with higher load is considered for calculation.

Let us consider L_{tx} to be the sum of the most frequent lengths of request messages in the outgoing direction, i.e.,

$$L_{tx} = \mathrm{SRI}_{\mathrm{req}} + \mathrm{SMS}_{\mathrm{req}} + \mathrm{PSI}_{\mathrm{req}} = 107 + 130 + 100 = 337 \text{ bytes}, \tag{3.7}$$

according to Figure 3.5. Similarly, let L_{rx} be the sum of the most frequent lengths of responses arriving from the network;

$$L_{rx} = \mathrm{SRI}_{\mathrm{res}} + \mathrm{SMS}_{\mathrm{res}} + \mathrm{PSI}_{\mathrm{res}} = 121 + 107 + 145 = 373 \text{ bytes.} \tag{3.8}$$

From Equation 3.6 it follows that the maximum hypothetical number of location retrievals through one 64 kbit/s link during one minute is about 514 (equation parameters $T = 60$ s, $S = 8{,}000$ B/s, $\rho_{max} = 0.4$, and $L = 373$ bytes). Tracking a higher number of users is possible only at the cost of lengthening the tracking interval δ, nevertheless, the number of position retrievals per time unit would remain the same.

The implementation and the signaling hardware are easily scalable. Contemporary high-throughput signaling hardware, for example Dialogic DSI SS7G32 [60], supports up to 192 links with 64 kbit/s data rate and could thus yield almost 10,000 location retrievals per minute. In addition, such hardware offloads signaling processing from application servers and thus saves their computing resources. Apart from the SS7 signaling we use, a similar performance could be achieved by enabling signaling over IP with the Stream Control Transmission Protocol (SCTP) by the SIGTRAN working group [50].

3.4.3 Network Constraints

The principal limits of SMS-based active tracking arise mainly from the constraints of the mobile network technology itself. In this section we examine the constraints of the different core network nodes and of the radio access part of the network.

Core Network

Active tracking based on sending SMS is a complex process that in GSM networks involves many core network nodes like HLR, VLR and MSC. Each of these nodes and their interconnections can be potential bottlenecks. However, nodes are usually designed for high performance and at least duplicated to guarantee availability in case of failure. The interconnection between nodes is capable of handling hundreds of millions of text messages and voice calls at peak times such as Christmas or New Year's Eve.

The SMS-Center may represent a bottleneck, but contemporary high-throughput solutions enable up to 25,000 SMS per second [106] which is easily sufficient for most tracking scenarios. The SMSC storage-buffer capacity is not a concern either, for the positioning procedure (described in Section 3.1) stops every time the mobile phone is off

Number of signaling	**Typical cell configuration**		
channels per cell	**1**	**4**	**12**
No. of SMS per minute	15	60	180

Table 3.4: Approximate number of deliverable text messages

or out of signal coverage (which precludes the SMS delivery) and thus the SMS is not stored in the SMSC for later delivery.

The tracking process definitely represents an overhead for the core network, but in comparison to average SMS/voice/data traffic, the impact is small. One location retrieval request amounts to one half of signaling messages needed for a mobile-to-mobile SMS [9], and to about three-fifths of the Mobile-Terminated Call signaling messages count [96].

Radio Access Network

The narrowest bottleneck of SMS-based active tracking in GSM networks is the *Air interface* between the Base Transceiver Station (BTS) and the mobile terminal [188]. SMS-based active tracking involves two dialogs transmitted over the Air interface at different scale: (1) *Paging*, transmitted by all BTSes in the location area where the tracked user resides, and (2) *SMS delivery*, performed at a particular BTS the user is attached to. These dialogs may result in potential congestion in a location area or in a cell. The maximum number of paging requests that can be served by a single BTS is dependent on the BTS configuration and ranges between 1,740 and 7,740 paging commands per minute [70]. Table 3.4 summarizes the approximate number of SMSes deliverable to one GSM cell per minute. This number depends on the cell configuration, i.e., the number of Standalone Dedicated Control Channels (SDCCHs) which carry SMS traffic and voice call establishment, and the fact that one SDCCH channel is typically occupied for 4–5 seconds during one SMS delivery [152]. Every SMS sent beyond these values would occupy signaling channels, thereby preventing voice traffic to or from the cell. We provide a detailed analysis in Section 3.4.4.

3.4.4 Scalability

The radio-access-network operation can easily be disrupted by high network-traffic load when tracking a high number of users residing in either the same cell or the same location area. However, mobile networks are planned and dimensioned to guarantee certain level of availability, the so-called Grade of Service (GoS), to all users. GoS, often called blocking rate or blocking probability, represents the maximum allowed ratio of requests blocked over a time period. In this section we show on an example of GSM network[4] how SMS-based active tracking changes the GoS with a growing number of tracked users, i.e., the scalability of the method.

Impact on Cell Capacity

First, we consider a particular GSM cell configuration[5], a cell with 2 Transmitters/Receivers (TRX). Since each TRX provide 8 TDMA carriers (timeslots), there are $2\times\ 8 = 16$ timeslots among which the traffic channels (TCHs) and control channels (such as SDCCH) are assigned according to the so called *SDCCH configuration*. We consider an SDCCH/8 configuration for a 2-TRX cell, which is composed of Broadcast and Common Control Channels in the first timeslot and 8 SDCCH sub-channels in the second timeslot, thus leaving 14 timeslots for TCHs.

Second, knowing the number of the signaling and traffic channels and the desired GoS[6] (0.5% for SDCCH), we use the Erlang B Table [160] to determine the maximum load capacity of SDCCHs and TCHs in the cell, which is $cap_{\mathrm{SDCCH}} = 2.73$ E, and $cap_{\mathrm{TCH}} = 8.20$ E, respectively. To estimate the load offered to SDCCHs and TCHs in a busy hour, we apply the BAS-1 Traffic Model which represents an average network according to Ericsson [71]. SDCCH resources are required by many events, such as Call Setup, SMS, Location Updates, Periodic Registration and IMSI Attach/Detach, whose load adds up. In total, the SDCCH load per user in an average cell (including 20% load margin for traffic peaks added) is estimated in the model to be $load_{\mathrm{SDCCH}} = 2.60$ mE; the TCH traffic per user is estimated to be $load_{\mathrm{TCH}} = 20$ mE.

Next, user capacity of SDCCH, i.e., the number of users that can be served by SDCCHs during busy hour, is calculated as a ratio of the channel capacity and the estimated load per subscriber: $usr_{\mathrm{SDCCH}} = \lfloor 2.73\ \mathrm{E}/0.0026\ \mathrm{E}\rfloor = 1050$ users. Similarly, user capacity of TCH equals $usr_{\mathrm{TCH}} = \lfloor 8.20\ \mathrm{E}/0.0200\ \mathrm{E}\rfloor = 410$ users. In order to perform a successful call setup, user capacity of SDCCH must be higher than the user capacity of TCHs:

$$usr_{\mathrm{SDCCH}} \geq usr_{\mathrm{TCH}}. \tag{3.9}$$

[4] Dimensioning for UMTS or LTE networks, which consider multi-class data traffic, is more complicated yet feasible [183].

[5] Similar methodology can be directly applied to any possible cell configuration [71].

[6] Conventionally used GoS values in GSM networks are 2% for Traffic Channels (TCHs) and 0.5% for SDCCH signaling control channels.

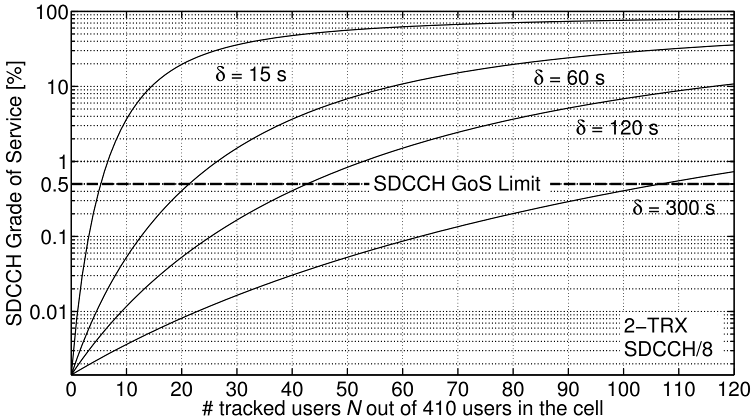


Figure 3.12: Impact of SMS-based active tracking on SDCCH GoS in a cell. Tracking interval is denoted δ.

Finally, the additional load caused by SMS-based active tracking, offered to SDCCHs over an hour, can be expressed as

$$N\bar{T}_{\text{SMS}}/\delta \tag{3.10}$$

where N denotes the number of tracked users in the cell, $\bar{T}_{\text{SMS}}$ denotes mean SMS delivery time (4.68 s, see Table 3.3) and δ is the tracking interval in seconds. A simple calculation shows that tracking *one* user with a 60 s tracking interval brings an additional load of 78 mE during busy hour, which corresponds to SDCCH load offered by 30 users during busy hour.

Figure 3.12 depicts GoS for SDCCHs as a function of the number of tracked users N and of the tracking interval δ. We assume there are 410 users in the cell, i.e., the maximum TCHs capacity at 2% GoS, and that exactly N of these users are tracked. The graph is calculated using the Erlang B formula for 8 SDCCHs and the offered traffic being a sum of the tracking load and the estimated load from all users in the cell. The impact of the increasing number of tracked users in the cell is significant: only 21 users, tracked every 60 seconds in the cell, suffices to exceed the desired GoS of SDCCHs. With 60 tracked users, the SDCCH blocking probability is above 10%, precluding every 10th user in the cell from being served on average. Although lengthening the tracking interval by every 60 s allows to track approximately 20 users more (at 0.5% GoS), the number of tracked users still remains only a fraction of all users in the cell.

Impact on Location Area Capacity

Positioning of an idle mobile terminal is always preceded by finding the cell the terminal is attached to. This is achieved by the paging procedure: a base station controller (BSC), serving a location area where a mobile terminal is registered, sends a Paging Command to *all* cells belonging to the location area and the mobile terminal responds from its actual serving cell. In the following paragraphs we show that periodic positioning increases paging load to such an extent that only a fraction of network users can be tracked with a considerably short tracking interval.

There are two principal types of components in the radio access network which can handle only limited paging load: base transceiver stations (BTSs) and base station controllers (BSCs). Paging capacity of a BTS ranges from 28 to 129 Paging Commands per second, depending on the cell configuration and paging strategy assumptions [70]. The BSC can be provided with paging capacity of about 8,500 Paging Commands per second [70]. Tracking is network-safe unless the number of Paging Commands per second remains below both the BTS and BSC maximum paging capacity.

Let us assume we have a BSC that serves a location area with 250 cells (BTS), each cell equipped with 2 TRX and configured as specified in the example in Section 3.4.4. Since one TRX can roughly carry 4.10 E of traffic during busy hour (Erlang B Table, 14 TCHs at 2% GoS), the total traffic capacity of the location area is 4.10 E/TRX × 250 cells × 2 TRX/cell = 2,050 E. Provided that, on average, one user offers load of 20 mE during busy hour [71], the location area can accommodate approximately 2,050/0.020 = 102,500 users, i.e., about 410 users at each cell on average.

According to the BAS 1 Traffic Model, paging load in the network may reach 0.0083 Paging Commands per second and Erlang traffic [70]. The paging load in the location area is then 2,050 E × 0.0083 Paging Command/(s×E) = 17.02 Paging Commands/s per BTS. Since there are 250 cells in the location area, a simple calculation shows that the BSC handles 17.02 × 250 = 4,255 Paging Commands/s.

Additional paging load, caused by periodic positioning of N users with tracking interval of δ seconds, can be expressed as

$$1.25 \cdot N/\delta \tag{3.11}$$

where N/δ represents the number of paging attempts per second and the multiplier 1.25

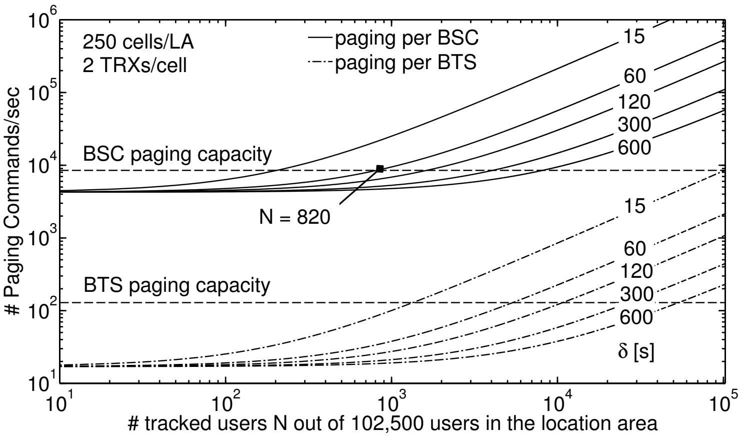


Figure 3.13: Impact of SMS-based active tracking on paging load. Tracking interval is denoted δ. BTS paging capacity is 129 Paging Commands per second, BSC paging capacity is 8,500 Paging Commands per second.

provides for the fact that on average 25% of paging attempts result in a second paging [70], thus the number of Paging Commands is higher.

Figure 3.13 shows the paging load during tracking as a function of the number of tracked users N and of the tracking interval δ. The number of Paging Commands rises with the number of tracked users, yet longer tracking interval results in slower growth. The graph provides a useful insight: since fewer tracked users suffice to exhaust the BSC capacity before the BTS capacity is exhausted, the bottleneck in the location area is the BSC. For example, tracking 820 users in the location area with 60 s tracking interval would disrupt the BSC paging functionality. Although 820 may seem a high number, it represent only 0.8% of all users in the location area. However, it could mean thousands or tens of thousands of users in the whole network, depending on the number of location areas in the network.

Interestingly, under the assumption of positioning an idle mobile terminal in the circuit-switched domain, similar results hold for *all* state-of-the-art network-based positioning methods (review in Section 2.1.3). Since every positioning method needs to establish a connection with the mobile terminal, the paging procedure is always necessary to locate mobile terminal's cell within a last known location area. We conclude that neither SMS-based active tracking nor any of the state-of-the-art network-based positioning methods can be used for large-scale tracking scenarios, such as tracking all users of a mobile network at the same time.

3.4.5 User Mobility

Mobility of tracked users who reside in the same geographical area, such as employees of one company or tourists, may represent a significant problem. As demonstrated in Section 3.4.4, a dense concentration of tracked users at one cell or at the same location area brings additional signaling load due to active tracking, which might render that particular network part inoperable. In this section we show that network congestion due to active tracking, caused by mobility of tracked users and their increasing concentration at a single cell, can happen on the order of minutes. In addition to that, we examine how to ease such situation by adopting a leaky-bucket traffic-shaping algorithm on the side of the Location Server.

We consider a cell with 2-TRX and the SDCCH/8 configuration, serving 300 users, in which the number or tracked users constantly increases over time as they arrive into the cell from the neighbor cells. Let λ denote the intensity of arrival of tracked users in the cell. Figure 3.14 shows, with solid lines, how GoS degrades over time when the tracked users keep concentrating in the cell. In consistence with results presented in Section 3.4.4, with $\lambda > 16$ tracked users arriving in the cell every minute, the desired SDCCH GoS can be exceeded in less than 2 minutes. Such intensity of arrivals can be observed for example before sport events, when tens of thousands of fans meet at a stadium within an hour or two.

To deal with the adverse impact of active tracking on signaling capacity in the cell, we suggest adopting a leaky-bucket traffic-shaping mechanism to limit the number of positioning requests. A leaky bucket [186] can be represented as a queue with the input flow of positioning requests. Arriving requests are enqueued, and then removed from the queue at a fixed rate r. Thus, the Location Server generates only r positioning requests per minute at a cost of lengthening the desired tracking interval δ. Figure 3.14 shows, with dashed lines, how traffic shaping with rate $r = 6$ positioning requests/min helps to keep GoS under the desired limit. However, because the arriving users bring additional signaling and traffic load, and not only the SDCCH load caused by active tracking, GoS degrades proportionally to the number of users in the cell nonetheless.

Figure 3.15 depicts the impact of the increasing concentration of tracked users at a cell on GoS of the traffic channels. Since active tracking brings additional load to SDCCHs only, the increase in TCH load and therefore the worse GoS is caused purely by new users in the cells. The TCHs load reaches 2% GoS when 410 users are present in the

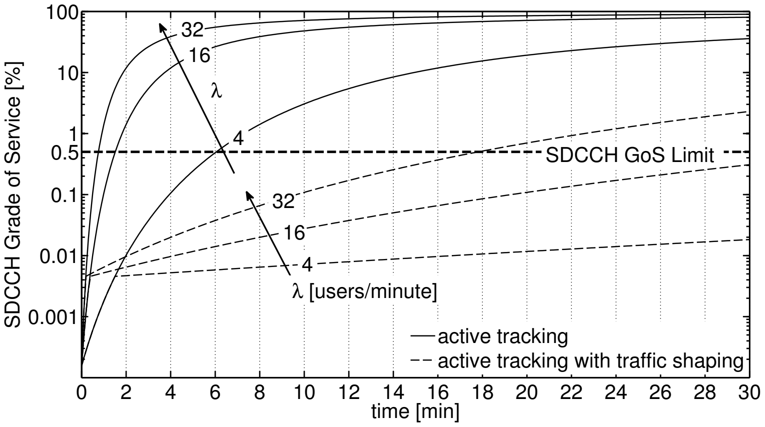


Figure 3.14: Impact of increasing number of tracked users on SDCCH GoS in a cell. An example for 300 users in the cell, $\delta = 60$ s tracking interval. The arrival rate λ of new users in the cell varies, traffic shaping is set to 6 positioning requests/min.

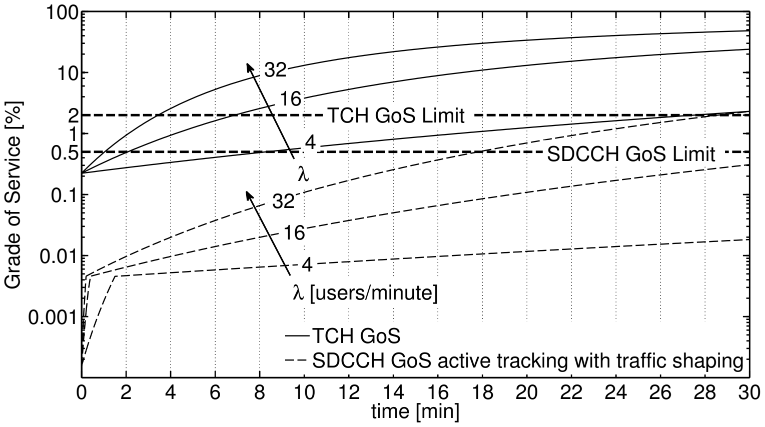


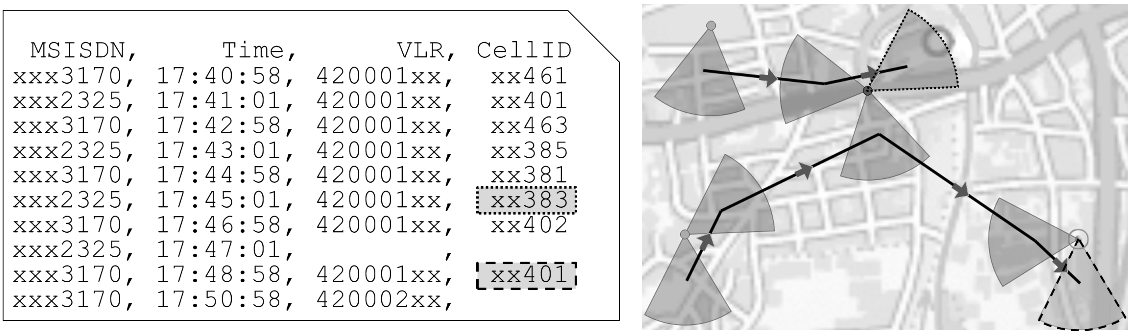
Figure 3.15: Impact of increasing number of tracked users on TCH GoS in a cell.

cell. The most important observation is that the expected SDCCH GoS for a particular λ hits the SDCCH GoS limit long after the TCH GoS limit is reached. As a result, active tracking with the leaky bucket traffic shaping mechanism can spare signaling capacity of the cell, but since arriving users would bring additional voice traffic load, the limiting factor becomes the capacity of the traffic channels nevertheless.

3.5 Case Study: Improving Roamer Retention

In this section we show how the network-based active tracking can serve as a tool for roaming optimization, and present our contribution to recognizing weak places in a mobile network in terms of roaming traffic.

Migration of mobile-network users between different countries and thus between different networks, simply called *roaming*, represents a significant revenue for mobile net-



```
 MSISDN,      Time,       VLR, CellID
xxx3170, 17:40:58, 420001xx,  xx461
xxx2325, 17:41:01, 420001xx,  xx401
xxx3170, 17:42:58, 420001xx,  xx463
xxx2325, 17:43:01, 420001xx,  xx385
xxx3170, 17:44:58, 420001xx,  xx381
xxx2325, 17:45:01, 420001xx,  xx383
xxx3170, 17:46:58, 420001xx,  xx402
xxx2325, 17:47:01,         ,
xxx3170, 17:48:58, 420001xx,  xx401
xxx3170, 17:50:58, 420002xx,
```

(a) *Records of user's active tracking.* **(b)** *Visualization on a map.*

Figure 3.16: User-Cell-ID association tracking output. Figure (a) shows a timestamped history of user-cell associations for two users. The log contains MSISDN, the mobile number of the user, Visitors Location Register number from which the provider's network is recognized, and Cell-ID of the cell the user was attached to. Mobile phone switched off is recognized by the missing VLR number, switch to a rival network is indicated by a different VLR number. Visualization on a map in Figure (b) shows that one user visited 3 cells and then switched the mobile phone off, while the other user visited 5 cells and then switched to a rival network; the *last cell before lost* is depicted with dotted and dashed lines, respectively.

work providers. New means of mobile network optimization, that would bring better roaming clients retention and thus a competitive advantage, are highly desired by network operators. Motivated by their needs in detecting places where *inroamers* (foreign roaming clients that subscribed to their network) leave to a rival network, we have used the SS7Tracker to track inroamers in a live GSM network in the Czech Republic and delivered a methodology to detect weak places in the network.

The goal of roaming optimization is to deliver recommendations for network planners *where* the weak places in their network are, in terms of possible insufficient signal coverage or unsuitable inter-cell handover scheme configuration. Network-based active tracking can deliver a list of cells, which has been visited by an inroamer, and a place where the user switched to a rival operator or lost the mobile signal (see Figure 3.16).

However, the last Cell-ID before the inroamer is lost to a rival network is not enough to draw conclusions about the cell's contribution to roaming losses. The reason is straightforward: because a user is tracked with period in the order of minutes, the geographical distance traveled by the user between the last known cell and the place of the actual loss to a rival network, recognized only later after the tracking interval, represent a significant measurement bias. Figure 3.17 illustrates that the last tracked cell before lost is not necessarily the cell that causes the inroamer's loss. Shortening the tracking interval is possible, but there is an obvious trade-off between length of the tracking interval and the number of tracked users, as presented in Section 3.3.3.

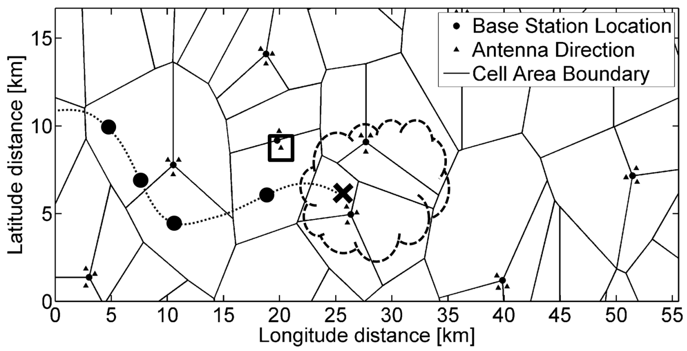


Figure 3.17: Misleading conclusion drawn from the "last cell before lost". The dashed line represents a trajectory of a tracked user, the circles mark the real user's position in a cell. The last tracked cell before lost is marked with a square. However, the real switch of user's mobile terminal to a rival network happened later at different place, denoted by the black cross. This loss may be caused by a weak signal from neighborhood cells, thus, the weak area spans more cells (indicated by the dashed cloud).

We address the above observation by formalizing the active tracking process, stating the problem of finding a weak place, and proposing a *cell-weakness metric* and an *appraisal function* that puts together weakness metrics for cells in the neighborhood to point out the geographical area of a weak place.

3.5.1 Active Tracking Formalization

Let $L \subset \mathbb{R}^2$ denote a set of *sites* (BTS locations) and let $C \subset \mathbb{N}$ be a set of all Cell-ID's in a studied network. For a given cell identifier $c \in C$ we denote a *cell* an area served by an antenna located at site $l_c \in L$. Let $U \subset \mathbb{N}$ denote a set of *users* (inroamers) subscribed to the studied network in the studied region. And let $S \subset \mathbb{N}$ denote a set of the following user's *states*: subscribed to the studied network (S_s) and subscribed to one of the rival networks in the studied region (S_r); i.e., $S = (S_s \cup S_r)$.

The output of an active tracking of a user u is a timestamped history of state-cell relations

$$T^u = \{(t_i, s_i, c_i)\}_{i=1}^{n} \tag{3.12}$$

described as follows: t_i denotes a timestamp (consecutive timestamps' difference correspond to the tracking interval); $s_i \in S$ denotes a state; and

$$c_i = \begin{cases} c \in C & \text{if } s_i = S_s, \\ \emptyset & \text{if } s_i \in S_r. \end{cases} \tag{3.13}$$

Accordingly, we denote $T^U = \bigcup_{u \in U} T^u$ as active tracking of a set of users U. Then (t_i^u, s_i^u, c_i^u) corresponds to i-th member of tracking T^u.

3.5.2 Problem Statement

The problem of revealing weak locations for roaming traffic is given as follows: Given a set L of site locations, a set C of cell identifiers, and an active tracking T^U of a set of users U, define an appraisal function $\mathbb{F}_M : (\mathbb{R}^2 \times \mathcal{P}(C)) \rightarrow \mathbb{R}$ incorporating a cell-weakness metrics $M : C \rightarrow \mathbb{R}$.

Weak locations will then be determined as a set of coordinates $W \subset \mathbb{R}^2$ satisfying $\forall x \in W : \mathbb{F}_M(x, A) > h$ for a given threshold value h and a set of cells $A \subseteq \mathcal{P}(C)$.

3.5.3 Cell-weakness Metric

We incorporate two basic facts in the cell-weakness metric:
(F1) users subscribed to a rival network will not generate any revenue for the provider of the studied network, and
(F2) places visited by a non-trivial number of users are supposed to achieve earlier return of resources invested in the network enhancement.
The cell-weakness metric M is then defined as follows:

$$M(c) = \left(\overbrace{\sum_{u \in U_c} \sum_{(i,j) \in I} |t_j^u - t_i^u|}^{\text{F1}} \right)^{\alpha} \cdot \overbrace{|U_c|}^{\text{F2}}, \tag{3.14}$$

where the set I contains indices of time intervals that a user u spent in rival networks after visiting the cell c until her return back to any cell in the studied network, i.e.,

$$\begin{aligned} I = \{(i,j) | i,j \in 1,2,...,|T^u|, i < j; \\ c_i^u = c \wedge \forall k = i+1, i+2, ..., j : s_k^u \in S_r \ \wedge s_{j+1}^u = S_s\}, \end{aligned} \tag{3.15}$$

a set U_c represents the users who visited the cell c,

$$U_c = \{u \in U | \exists i \in 1,2,...,|T^u| : c_i^u = c\}, \tag{3.16}$$

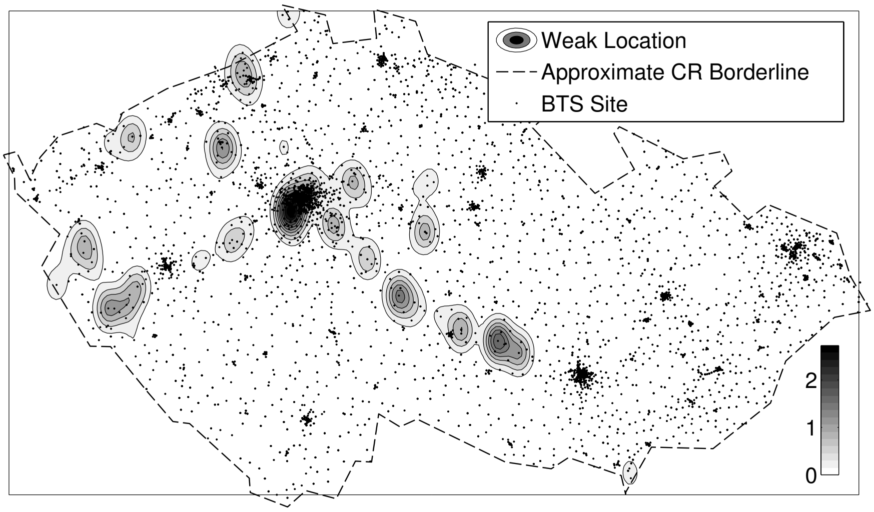


Figure 3.18: Visualization of suspicious weak locations. Suspicious weak locations in a live GSM network in the Czech Republic, according to the cell-weakness metric [79]. Color scale corresponds to different threshold values of the appraisal function that puts together metric values at each cell. The measurement has been conducted on 500 roaming users during 6 hours with 2 minutes tracking interval.

and $\alpha \in \mathbb{R}$ is a parameter which adjusts the weighted influence of the time factor[7] of the metrics.

3.5.4 Appraisal Function

We choose the appraisal function $\mathbb{F}_M(x, A)$ for cell-weakness metric values from all cells to be a weighted kernel density estimator [195] with a Gaussian kernel and the proposed "cell-weakness" metric M as a re-weighting function:

$$\mathbb{F}_M(x, A) = \sum_{c \in A} M(c) \frac{1}{|A|} K_h(x - l_c) \tag{3.17}$$

The kernel bandwidth parameter K_h controls the smoothness of the estimate. Using this type of function, an intuitive visualization of weak places in a mobile network can be prepared using a 2D-histogram, see Figure 3.18.

3.6 Conclusion

In this chapter, we have presented SMS-based active tracking in the mobile network, a network-based method for obtaining positioning data of users' terminals. We have

[7] Note that the time factor can be interpreted in the *roamer-day* units, explained as one day spent in a rival network by one roamer; similarly to an industrial unit of production, *man-day*.

demonstrated that a practical platform can be implemented and deployed in a live GSM network, using off-the-shelf computing equipment and common signaling hardware.

We have measured and described working and service response times of the tracking platform and adjacent mobile network nodes. These values were used in a discrete simulation of the tracking process, revealing the key performance characteristics of the tracking platform in terms of viable combinations of the number of tracked users and the tracking interval. The baseline implementation is capable of tracking thousands of users periodically on the scale of minutes.

Our study on SMS-based active tracking brings insight into the principal limitations of the method and how feasible it is to deploy it in large-scale tracking scenarios. On the basis of tracking measurement in a live network, we have estimated the minimal value of tracking interval and the maximal number of positioning retrievals achievable with the limited connection throughput. A detailed analysis of mobile network constrains revealed that the mobile-network radio access technology is the most limiting factor in active tracking because the number of SMSes and paging requests deliverable per transceiver of the radio access network cannot be increased beyond a relatively low technology-specific value. Nevertheless, considering the fact that there are thousands of transceivers in the mobile network, the total number of tracked users in the whole network could be on the order of tens of thousands.

The final part of the chapter presents a particular use-case of active tracking on roaming optimization in mobile networks. We have proposed a cell-weakness metric to express the losses in roaming clients from the cells, and a demonstrative visualization of weak places in the network.

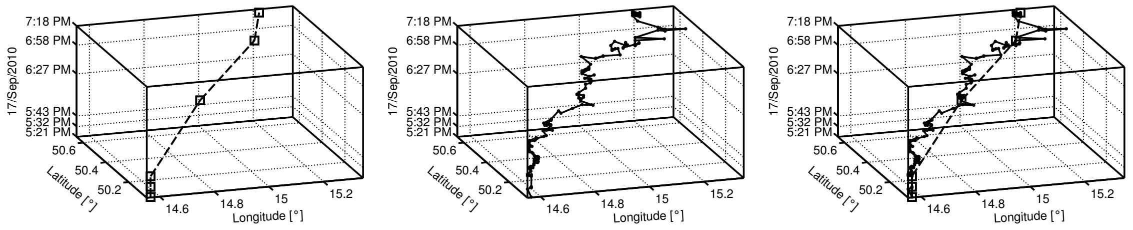
The SMS-based active tracking represents a solution for tracking a large number of mobile network users with unprecedented temporal granularity. From the research perspective, it can be used to build and verify accurate models for user mobility within mobile networks as well as among different geographic areas. Thanks to its fruitful features, mainly that it does not depend on user communication mode or terminal type, the active tracking may be adopted by a broad class of applications: network diagnostics, crime prevention, energy-consumption control, urban planning or sociological studies.

Chapter 4

Extending Utility of Passive Tracking Data

Passive tracking data, mainly the Call Data Records (CDRs), represent a vast amount of easy-to-collect data about *every* mobile network user, for they are automatically generated by telecommunication systems and archived for billing purposes and network troubleshooting. Whereas spatial precision of CDRs is determined the network-cell size, the accuracy in temporal dimension is substantially low. The reason is that the position of a user is recorded *only* at places where the user performs a communication event (SMS, call, data session). An average time between two communication events is 8.2 hours, as measured by [88] on a large-scale sample, which means that user's location is known on average only three-times during a day. During time with no communication activity it is not clear where the user is geographically located.

CDRs constitute an *event-based* motion description of a mobile user in space and time—we denote it a *call trajectory* and the communication event simply a *call.* It is a sequence of places, related to a single user, where a call (or text message) has been made (sent) or received, thus describing user's "hop" movements as he/she makes call after call. A natural refinement of the call trajectory is its corresponding ground-truth *movement trajectory*, i.e., a *continuous* trace with geographical coordinates of user's position (see Figure 4.1). GPS traces of users' movement provide great spatiotemporal accuracy but limited availability: one study found only 4.5% user-time coverage in tests with the device carried in users' pockets during a day [126]. A movement trajectory from active tracking of network users by the network provide offers identical spatial accuracy as CDRs but is



(a) *Call trajectory* **(b)** *Movement trajectory* **(c)** *Trajectories superimposition*

Figure 4.1: Call and movement trajectories explanation. (a) Six communication events (square marks) constitute a call trajectory in a space-time cube. The x- and y-axes represent space, the z-axis represents time. Line segments connect consecutive communication events—vertical line segments stand for stay at a certain location, sloped line segments indicate movement. **(b)** Movement trajectory of the same user, in this example made from cell transitions. Each dot represent geographical location and time of a cell the user was attached to. **(c)** Superimposition of trajectories clearly shows that the user does not follow a simple straight line between two consecutive communication events.

advantageous in higher granularity in the temporal dimension. However, its acquisition is bounded by network technological limits [80] and tracking costs [20]. The CDR-based call trajectory, despite it's temporal sparseness, is thus often the only source for numerous contemporary studies, such as on virus spreading [197], individual mobility and calling patterns [88, 99], urban and transport planning [194, 196], or network design [211].

Modeling techniques are used to describe the expected user position in between calls. *Linear interpolation*, a straight line between two points in space and time, is popular in movement objects databases [92]. However, this method is accurate only for sufficiently dense sampling of events, which calls definitely are not—for example, every other mobile user calls less than once per day [211]. Another model, the *space-time prism* [94], represents locations reachable by users, given only their origin and destination positions and maximum speed. Unfortunately, the maximum speed cannot be set for all users in general, which limits the model applicability.

In this chapter we present a probabilistic model that describes user's position between their consecutive communication events (call or SMS), the Inter-Call Mobility (ICM) model. This model enables sampling of user's geographical location at a particular time in between communication records and, vice versa, given geographical coordinates, probability of user's presence at a particular position over time can be derived. The model is based on a comparison between two representations of user's movement: the coarse-grained, CDR-based call trajectory, and the movement trajectory derived from ground-truth network-cell transitions. The rest of this chapter is organized a follows:

First, in Section 4.1 we shortly present the dataset we used. Second, we describe major observations of user inter-call mobility in Section 4.2. Next, in Section 4.3 we show how the model is build from the dataset and demonstrate how to use it with an arbitrary CDR dataset. In Section 4.4 we evaluate model's precision by comparing it with other state-of-the-art models. Finally, in Section 4.5 we show model's applicability on example of inferring proximity probability of two users.

4.1 Data Acquisition and Processing

To study users' inter-call movement, both call trajectory and its corresponding movement trajectory are needed. Mobile network providers, who may be willing to share CDR databases, do not record actual positions of their users for they lack the required trace-collecting technology and, in fact, do not need such fine-grained information. Publicly available datasets (e.g. in [121]) seldom contain both the call trajectory (CDRs with geographical coordinates of the network-cells) and a corresponding finer-grained movement trajectory.

At the time of writing this thesis, there was *only one publicly available dataset* providing data suitable for inter-call mobility analysis—the Reality Mining Dataset (RMD). The RMD is a real-world trace recorded at MIT that contains history of communication activity and network associations of mobile phones used by many volunteer individuals over several months [68]. This dataset contains call records, network-cell transitions and other records, but it originally lacks the geographical coordinates of user mobility—either GPS traces or positions of mobile network cells towers the users were attached to.

We have spatially extended the RMD by pairing cell identifiers in users' traces with their corresponding geographical coordinates (obtained from the Location API by Google [49]), we removed spatial outliers using a novel heuristic approach to agglomerative clustering, and provided a methodology to extract representative chunks of trajectories. This process, described in detail in Appendix A, delivers users' call trajectories and their corresponding ground-truth movement trajectories represented by the cell transitions (see an example in Figure 4.2).

In order to deliver a *general* description of user's position in between communication records, we describe user's mobility and detours relative to the inter-call distance. Therefore, in the rest of this chapter, we study only pairs of consecutive communication

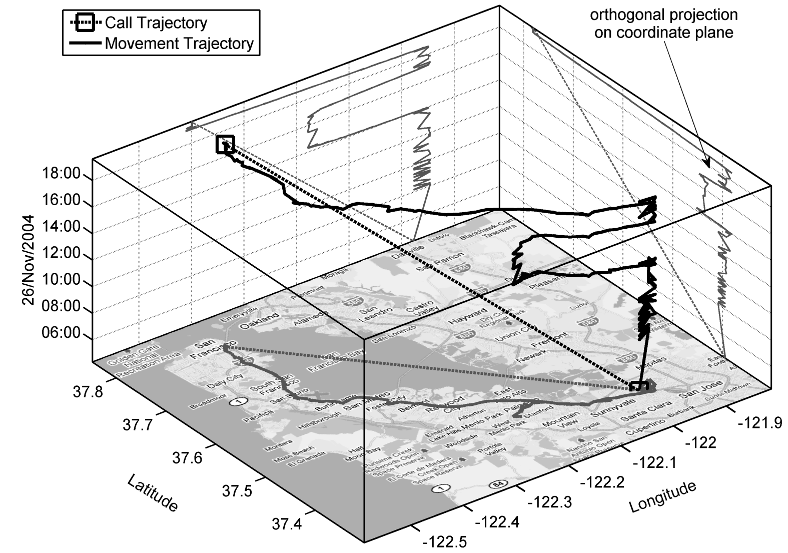


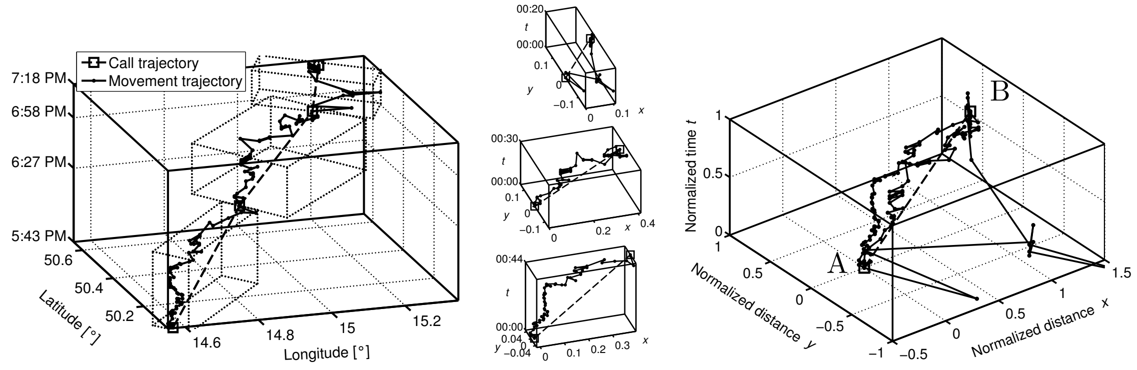
Figure 4.2: Difference between call and movement trajectories. The difference is demonstrated on a trip between San Jose and San Francisco downtown made by a user from the Reality Mining Dataset [68]. Call trajectory connects places of two consecutive communication activities: at 4 a.m. near San Jose, and around 7 p.m. in San Francisco downtown. Movement trajectory represents user's transitions between mobile network cells. It shows that around 2 p.m. the user moved to Stanford and went back to San Jose at about 5 p.m.

records *at distinct places.* Such selection is also based on our observation from the RMD that in the time interval between two consecutive communication events at the same place, a user does not, on average, depart from a call place further than 0.2 km (0.99-quantile $\approx$ 0.8 km). Therefore we can assume users to be static between communication events at same places. For our purposes, we consider distinct places to be places *more than three kilometers apart.* Conclusions drawn from lower inter-call distance can be affected by the accuracy of Cell-ID positioning method that delivers user's position only as an approximation of the geographical coordinates within a cell, while the exact position is not known[1]. We do not address movement between distinct call places less then 3 km apart for the limited Cell-ID-based positioning accuracy.

We work with an aggregated view of user movement between calls, derived as follows: first, we divide each call trajectory into pairs of *consecutive* events, which are geographical positions at the time a call (or text message) has been made (sent) or received[2] (see Figure 4.3a). The call trajectory reduces to a straight line between two points,

[1] Several studies, such as [189, 212], demonstrated the expected positioning error of Cell-ID-based positioning to be 400–500 meters in urban areas, 700 meters in suburban and 1 km in rural areas.

[2] For the lack of data-active users in the dataset, we have used only call and text message communication events. It is obvious that data sessions in Call Data Records could significantly improve the accuracy of user's position in time [172], for the data-active users generate more fine-grained footprint in network records in time. However, even for data connection that is always on, a mobile device with no data to send



(a) **(b)** **(c)**

Figure 4.3: Example of user's inter-call trajectories aggregation. **(a)** Four call places divide the call trajectory in three segments, the movement trajectory is divided by call segments in three inter-call trajectories. Each segment determines a separate reference frame (dashed boxes), the orientation of the coordinates is given by the direction of the segment. **(b)** Each inter-call trajectory is translated to coordinates origin. **(c)** Inter-call trajectories are scaled into a common reference frame, a normalized space-time cube.

but the corresponding *movement* trajectory represents a fine-grained *inter-call* trajectory. Further on, we transformed each inter-call trajectory to have a *common reference frame* (Figure 4.3b). Finally, we normalized the trajectories to have uniform origin A and destination B in space and time, $(x_A, y_A, t_A) = (0, 0, 0)$ and $(x_B, y_B, t_B) = (1, 0, 1)$, as indicated in Figure 4.3c.

In the Reality Mining Dataset, there are 901 inter-call trajectories, made by 56 users out of the 94 user sample Figure 4.4 shows these inter-call trajectories aggregated in the common reference frame. Such normalization brings an aggregated distance-time view of a position in between calls, but, at the same time, does not limit further inference from the data and allows for easy application to *arbitrary* CDRs, as demonstrated later in Section 4.3.4.

resides in a standby mode which implies that no information about its current cell is available within the current location area.

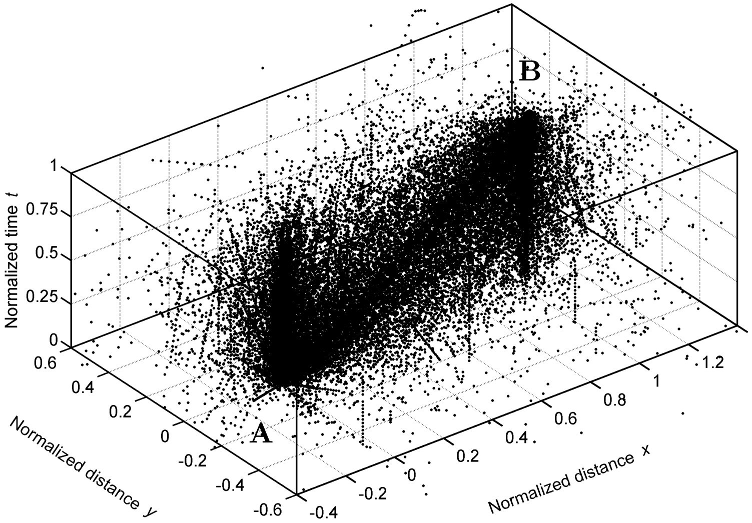


Figure 4.4: Aggregated inter-call trajectories. Each point represent user's position between calls (places A and B in normalized common reference frame) after aggregation of inter-call trajectories from the Reality Mining Dataset.

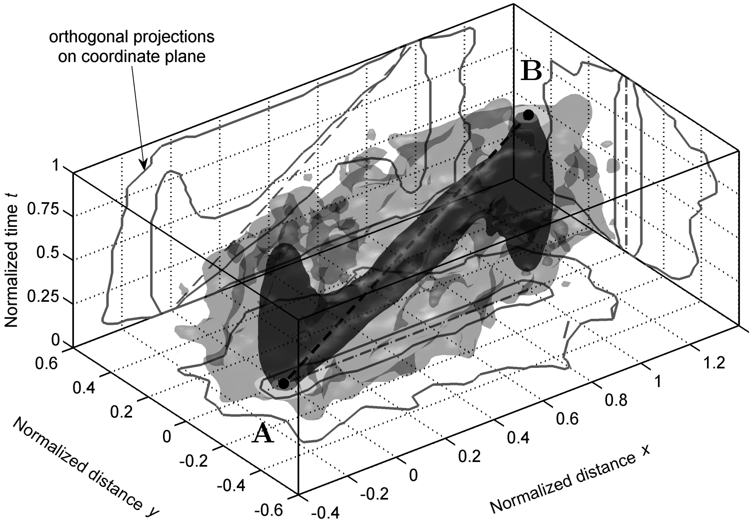


Figure 4.5: Spatio-temporal characteristics of inter-call movement. Kernel density estimation of the spatio-temporal probability distribution of users' position between calls (places A and B in normalized common reference frame). Isosurfaces enclose 0.5- and 0.9-quantiles (dark and light gray, respectively).

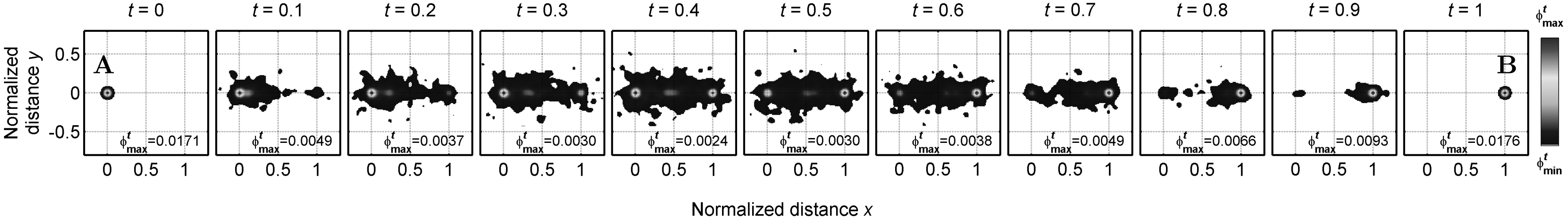


Figure 4.6: Inter-call movement in time-slices. Kernel density estimation of the PDF of finding a user at a position (x, y) at a time t between two consecutive communication records at distinct places A and B. Hot places represent higher concentration of users. The 2D-histogram bin size is 0.01×0.01 of normalized inter-call distance.

4.2 Inter-Call Mobility Observations

In this section we examine the spatio-temporal properties of user's movement between consecutive calls at distinct places, based on the aggregated inter-call trajectories from the Reality Mining Dataset, and we describe in detail our findings in the temporal and spatial dimensions separately.

4.2.1 Spatio-temporal Analysis

Figure 4.5 shows the kernel density estimation of the spatio-temporal probability distribution of user's inter-call movement: as time passes, users move from origin A towards destination B and take detours from the straight $A - B$ direction. The distribution shows three important aspects of inter-call mobility:

1. **Unskewed behavior in spatial dimension.** About half of users closely follow the direct $A - B$ linear interpolation line. This is observable in the symmetry of the 0.5-quantile projection on the xy- and yt-coordinate planes—it implies unskewed behavior with respect to detours from the straight $A - B$ course.

2. **Straight course between calls.** Some users take approximately the shortest path from the origin call-place A to the destination place B. This is indicated by symmetry in the spatial dimension (xy-coordinate plane) which does not change in time (yt-coordinate) and the fact that the 0.5-quantile projection on the xt plane encloses the direct $A - B$ interpolation line.

3. **Staying behavior at call places.** From the shape of the projection on the xt-coordinate plane at points $x_A = 0$ and $x_B = 1$ it follows that some users tend to stay at the origin call-place A before they move to destination B, or they leave the origin soon after the act of communication and stay in the vicinity of the destination place B. This may be more obvious from Figure 4.6 that depicts time-slices of the volumetric representation (in Figure 4.5) and thus represents the conditional Probability Density Function (PDF) of finding a user at a position (x, y) at time t. For example, at time $t = 0.5$ (half-way in between calls), about 12% of all users are still present at the origin call-place and 17% have already arrived at the destination call-place.

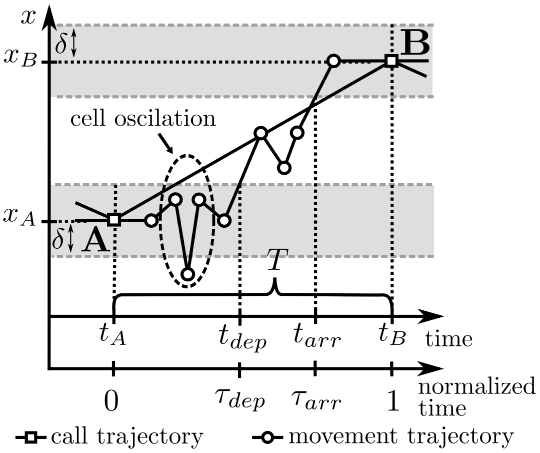


Figure 4.7: Schematic representation of inter-call trajectory. Trajectories between two call places A and B, for the sake of clarity with the spatial dimension limited to x coordinate. Time t_{dep} denotes departure after call, t_{arr} is time of arrival before call.

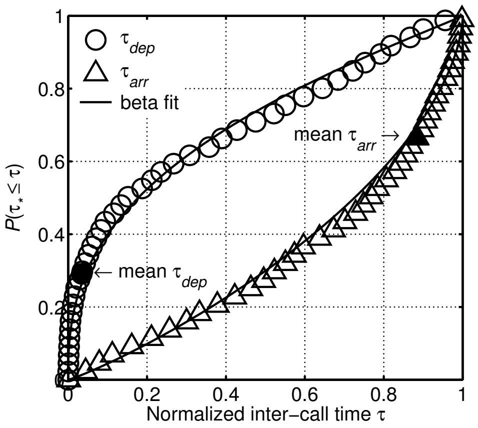


Figure 4.8: Empirical CDF of the *proportion* of inter-call time T: $\tau_{dep} = (t_{dep} - t_A)/T$, $\tau_{arr} = (t_B - t_{arr})/T$.

4.2.2 Temporal Dimension Analysis

In this section we describe in detail the staying behavior at call places, introduced in Section 4.2.1.

We consider a user to be leaving the origin call-place $A = (t_A, x_A)$ at time t_{dep} (see Figure 4.7) when he/she leaves the δ-neighborhood of x_A and *does not return back* into it before the consecutive act of communication at the destination call-place $B = (t_B, x_B)$. This "not returning back" request naturally excludes the so-called *cell oscillation*, a situation in which the mobile phone of a static user attaches itself to a different cell that can be even hundreds of meters apart. Similarly, the arrival time t_{arr} at the destination call-place $B = (t_B, x_B)$ is counted as the *first entering* of the δ-neighborhood of x_B after the preceding communication at the origin call-place A. We use $\delta = 0.5$ km as we consider it a reasonable value for the limited Cell-ID-based positioning accuracy, caused by variable cell size within urban and rural areas, albeit we understand that such threshold may be inconvenient for specific use-cases.

Figure 4.8 shows the CDF of τ_{dep} and τ_{arr}, the *proportions* of inter-call time T in which users leave the origin (after the call) for the last time and arrive at the destination for the first time (before making the call), respectively. With this simple classification, basic, but very useful observations about inter-call mobility can be made: on average, after 29% of inter-call time a user does not return to the original call place anymore, and he/she reaches the consecutive call place by 67% of inter-call time. More precisely, for another example, if a user communicates at work at 6 p.m. and the consecutive call record is at 8 p.m. from home, there is a 50% chance that he/she left work after 6:18 p.m.: $T = 120$ minutes, $P(\tau_{dep} > \tau \approx 0.15) = 0.5$ (see Figure 4.8), thus $\tau_{dep} \cdot T > \tau \cdot T = 0.15 \cdot 120 = 18$ minutes and so $t_{dep} >$ 6:18 p.m.. Similarly, with the same chance, he/she arrived home before 7:36 p.m.: $P(\tau_{arr} \leq \tau \approx 0.8) = 0.5$ thus $\tau_{arr} \cdot T \leq \tau \cdot T = 0.8 \cdot 120 = 96$ minutes and so $t_{arr} \leq$ 7:36 p.m..

The τ_{dep} and τ_{arr} can be approximated by beta distribution:

$$P(\tau_*) = \frac{\tau_*^{\alpha-1}(1-\tau_*)^{\beta-1}}{B(\alpha,\beta)} = \frac{\tau_*^{\alpha-1}(1-\tau_*)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du}, \tag{4.1}$$

with parameters α, β in Table 4.1. We have performed the Kolmogorov-Smirnov (KS) test and compared empirical data with the fitted distribution and 5,000 synthetic data sets generated from the best beta fit. At $\alpha = 0.05$ significance level the hypothesis that empirical and synthetic data are from the same distribution was rejected (p-value 0.0438 for τ_{dep} and p-value 0.0236 for τ_{arr}). Despite this fact, we consider this approximation close enough to rapidly improve time accuracy of CDR-based deduction of user's presence at call places over time: from a single timestamp of communication activity to an *analytical* expression of probability distributions of τ_{dep} (departure from call place) and τ_{arr} (arrival at consecutive call place).

4.2.3 Spatial Dimension Analysis

Because users do not move linearly over *time* between places of communication, the naïve approach of expressing user's *position* with the call trajectory and linear interpolation only suffers from a significant positioning error.

Figure 4.9 explains how much the "expected" position of a user, computed from the linear interpolation between A and B, and the actual position of a user, differ. It follows that about 20% of real user's positions are more than 10 km away from the corresponding

	τ_{dep}	τ_{arr}
α	0.39 ± 0.04	1.11 ± 0.10
β	0.96 ± 0.11	0.58 ± 0.06

Table 4.1: Beta distribution parameters with 95% confidence intervals.

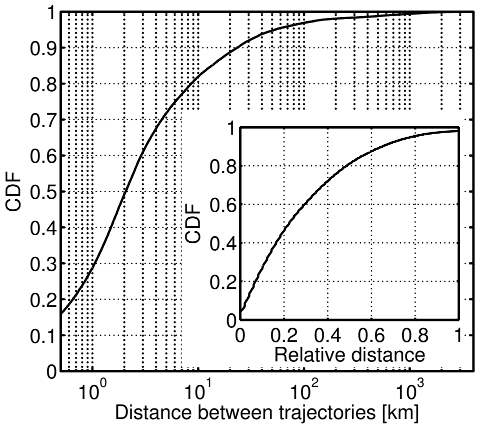


Figure 4.9: Positioning error. Geographical distance between the call and movement trajectories at identical moments during inter-call time. Distance less then 0.5 km is not considered for the limited spatial precision of the data. Inset shows proportion of call-movement trajectory distance to the total distance between calls.

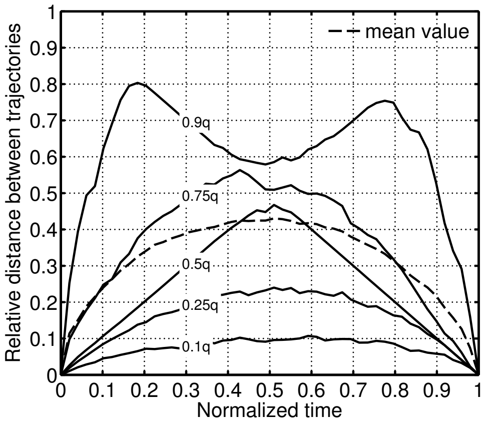


Figure 4.10: Positioning error over time. Quantiles of relative distance (proportion of call-movement trajectory distance to the total distance between calls) over the inter-call period.

point at the straight $A - B$ line. In other words, relative to the total distance between calls, 20% of real user's positions is *more than half the inter-call distance away* from the position "expected" by the linear interpolation (see inset in Figure 4.9).

Figure 4.10 shows the positioning error as a function of time. The relative distance between call and movement trajectory rises in the interval from 0 to 0.5 of inter-call time and culminates exactly in the middle of the inter-call period. However, the highest deviation of positioning error happens at relative time 0.2 and 0.8, for some users move soon after communication at the next-call place, and some stay at the origin place and move shortly before the next communication event, respectively. This observation supports our evidence about user's staying behavior at call places, described in Section 4.2.2.

4.3 Inter-Call Mobility Model

In this section we present the Inter-Call Mobility (ICM) model—a spatio-temporal probability distribution of user's position in space and time between two consecutive communication records at distinct places. The ICM model, simply an approximation of the aggregated inter-call trajectories of the RMD user-pool, is based on a finite Gaussian mixture model [145].

4.3.1 Gaussian Mixture Model Primer

A Gaussian mixture model (GMM) is a weighted sum of Gaussian PDFs, which are called *mixture components*. Let $\mathcal{Z}$ denote a set of d-dimensional random variables $\boldsymbol{z}$, i.e., $\mathcal{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N\}$ where $\boldsymbol{z}_i \in \mathbb{R}^d$. The PDF of Gaussian mixture model of K components is given by

$$p(\boldsymbol{z}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{4.2}$$

where π_k is a *mixing proportion* or *weight* of the kth component $\mathcal{N}(\boldsymbol{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, which is a Gaussian distribution defined by *mean* vector $\boldsymbol{\mu}_k$ and *covariance* matrix $\boldsymbol{\Sigma}_k$:

$$\mathcal{N}(\boldsymbol{z}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_k)}} e^{\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{z}_i - \boldsymbol{\mu}_k)}. \tag{4.3}$$

The mixing proportion π_k can be interpreted as *a priori* probability that a value of a random variable comes from the kth component, thus $0 \leq \pi_k \leq 1, (k = 1, \ldots, K)$, and $\sum_{k=1}^{K} \pi_k = 1$. A GMM of K components is completely defined by the vector $\boldsymbol{\theta}$ with all unknown parameters, represented as

$$\boldsymbol{\theta} = (\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \pi_2, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \ldots, \pi_K, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K). \tag{4.4}$$

The parameters $\boldsymbol{\theta}$ of the mixture model are usually estimated by the *Expectation-Maximization* (EM) algorithm [56]. For a given number of components K, it performs maximum-likelihood estimation of unknown parameters. The number of components K in a Gaussian mixture must be carefully determined. More components yield better fit, but more parameters are needed to fully define the mixture: the GMM of K components with non-restricted covariance matrices is given by $n = K(d + \frac{1}{2}d(d+1)) + (K-1)$ parameters, where d is the dimension of observation points. To select the "best" model from a set of candidate models with different number of components, various criteria are available (e.g. in [37]). The most commonly used are the *Akaike Information Criterion* (AIC) [22] and the *Bayesian Information Criterion* (BIC) [180].

4.3.2 Fitting GMM to the Reality Mining Dataset

We represent aggregated inter-call trajectories from the RMD as a dataset $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ of N spatio-temporal records $\boldsymbol{z_i} \in \mathbb{R}^3$, where each datapoint $\boldsymbol{z_i} = \{\boldsymbol{z}_{s,i}, z_{t,i}\}$ comprise spatial coordinates $\boldsymbol{z}_{s,i} \in \mathbb{R}^2$ and a temporal value $z_{t,i} \in \mathbb{R}$. The initial estimate

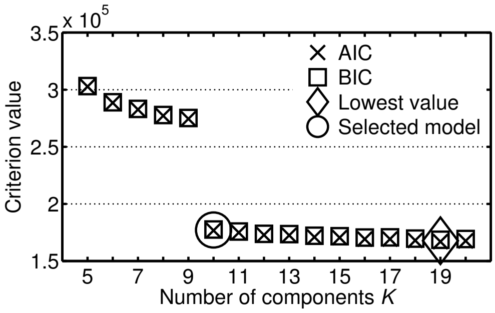


Figure 4.11: Gaussian mixture model fits.

of mixture parameters $\boldsymbol{\theta}$, needed by the EM algorithm, is given by *K-means* clustering. This technique divides the data in K partitions, according to the point-to-cluster-centroid distances, from which the initial estimation of π_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ are computed. Finally, we run the EM algorithm and retrieve the mixture parameter estimates and the values of model-fit criterions.

We fitted a set of GMMs with $K = 5, \ldots, 20$ components to the RMD inter-call trajectories and compared the resulting fit criteria AIC and BIC in Figure 4.11. Since a lower criterion value indicates a better fit, a model with the lowest criterion value is usually selected. In our case, it is the model with 19 components, which are fully described by $n_{K=19} = 189$ parameters. Nevertheless, we prefer the model with 10 components, for it needs significantly lower number of parameters ($n_{K=10} = 99$) at a cost of only slightly worse fit. The parameters are given in Table 4.2. Their standard error, estimated with parametric bootstrap ([145], p.68–70) with 1200 replications, is approximately three orders of magnitude lower than the parameter value.

4.3.3 Inter-Call Mobility Model Definition

We define the ICM model as follows. Given the origin position $(x_A, y_A, t_A) = (0, 0, 0)$ and the destination position at $(x_B, y_B, t_B) = (1, 0, 1)$, the probability $\Phi(x, y, t)$ of finding a user at coordinates $(x, y) \in \mathbb{R}^2$ at a time $t \in \mathbb{R}, 0 \leq t \leq 1$, is defined as a Gaussian mixture

$$p(\boldsymbol{z} = (x, y, t)^T; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{4.5}$$

of $K = 10$ components with parameters $\boldsymbol{\theta}$ in Table 4.2.

Figure 4.12 provides a visual representation of the ICM model. As the model approximates the inter-call trajectories of the RMD user-pool, its shape is similar to the estimation in Figure 4.5. Detailed model evaluation is presented in Section 4.4.

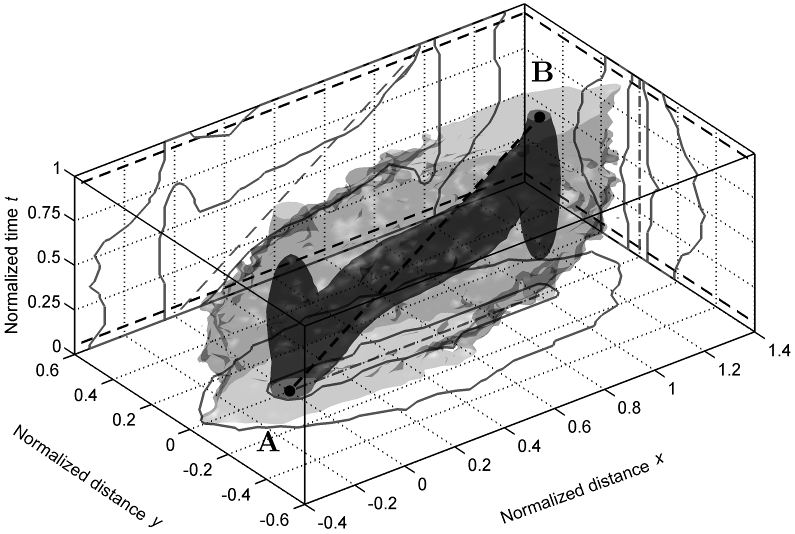


Figure 4.12: Inter-Call Mobility model. A spatio-temporal probability distribution of user's position between two communication records at distinct places A and B. Isosurfaces enclose 0.5- and 0.9-quantiles (dark and light gray, respectively). Dashed lines restrict areas of worse fit.

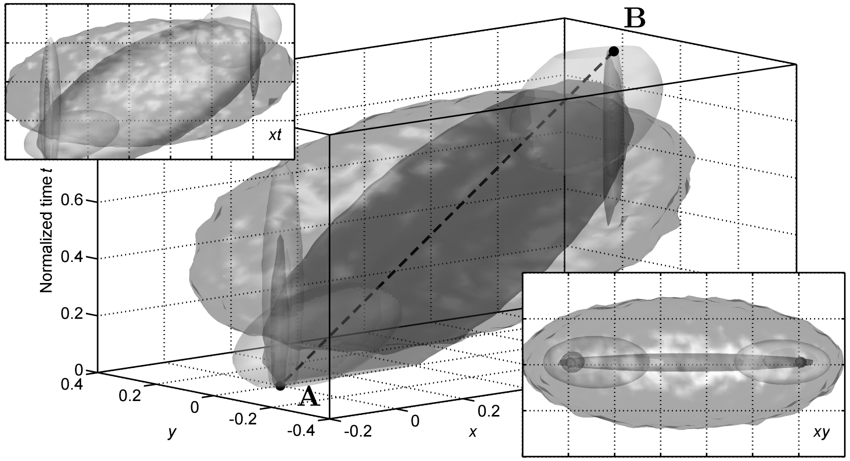


Figure 4.13: Components of the ICM model. There are 10 three-dimensional Gaussian components whose mixture constitutes the ICM model. The orthogonal projections on xt- and xy-coordinate planes are depicted in the upper left and lower right corner, respectively.

k	π_k	$\boldsymbol{\mu}_k \cdot 10^2$	$\boldsymbol{\Sigma}_k \cdot 10^3$
1	0.0418	$\begin{bmatrix} 0.00 \\ 0.00 \\ 21.64 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 43.11 \end{bmatrix}$
2	0.0314	$\begin{bmatrix} 0.01 \\ -0.03 \\ 22.75 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.01 & -0.01 \\ 0.00 & -0.01 & 46.12 \end{bmatrix}$
3	0.0844	$\begin{bmatrix} 0.70 \\ 0.10 \\ 34.47 \end{bmatrix}$	$\begin{bmatrix} 0.67 & 0.02 & 0.07 \\ 0.02 & 0.48 & 0.20 \\ 0.07 & 0.20 & 53.54 \end{bmatrix}$
4	0.0890	$\begin{bmatrix} 11.04 \\ 0.07 \\ 15.97 \end{bmatrix}$	$\begin{bmatrix} 13.61 & -0.08 & 3.23 \\ -0.08 & 3.09 & 0.11 \\ 3.23 & 0.11 & 6.61 \end{bmatrix}$
5	0.3590	$\begin{bmatrix} 47.84 \\ -0.25 \\ 48.95 \end{bmatrix}$	$\begin{bmatrix} 141.13 & -0.98 & 15.05 \\ -0.98 & 23.05 & 0.79 \\ 15.05 & 0.79 & 46.94 \end{bmatrix}$
6	0.0514	$\begin{bmatrix} 49.66 \\ 0.00 \\ 42.86 \end{bmatrix}$	$\begin{bmatrix} 96.17 & 0.00 & 52.10 \\ 0.00 & 0.00 & 0.00 \\ 52.10 & 0.00 & 71.09 \end{bmatrix}$
7	0.0419	$\begin{bmatrix} 58.97 \\ -7.55 \\ 48.67 \end{bmatrix}$	$\begin{bmatrix} 694.24 & 36.82 & 39.27 \\ 36.82 & 367.05 & -16.10 \\ 39.27 & -16.10 & 64.53 \end{bmatrix}$
8	0.0944	$\begin{bmatrix} 91.77 \\ 0.13 \\ 80.40 \end{bmatrix}$	$\begin{bmatrix} 11.40 & -0.23 & 2.61 \\ -0.23 & 2.21 & 0.24 \\ 2.61 & 0.24 & 12.59 \end{bmatrix}$
9	0.0532	$\begin{bmatrix} 99.83 \\ 0.10 \\ 68.52 \end{bmatrix}$	$\begin{bmatrix} 0.08 & -0.01 & -0.11 \\ -0.01 & 0.05 & -0.08 \\ -0.11 & -0.08 & 54.56 \end{bmatrix}$
10	0.1535	$\begin{bmatrix} 100.00 \\ 0.00 \\ 78.20 \end{bmatrix}$	$\begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 43.85 \end{bmatrix}$

Table 4.2: Inter-Call Mobility model parameters. An indication of position and shape of the components is depicted in Figure 4.13.

4.3.4 Using the Inter-Call Mobility Model with CDRs

The ICM model can be easily used with an arbitrary CDR dataset. Because the ICM model uses normalized coordinates, a transformation of the model's origin and destination places to the particular CDR-based call places is needed. Such transformation exists: the multivariate Gaussian distribution is invariant under *affine transformation with an invertible matrix* and thus the ICM model, as a mixture of multivariate Gaussian distributions, allows for this transformation as well. Affine transformation of Gaussian distribution works in the following way: if $Y = \boldsymbol{c} + \boldsymbol{DX}$ is an affine transformation of $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with translation vector $\boldsymbol{c} \in \mathbb{R}^{m\times 1}$ and transformation matrix $\boldsymbol{D} \in \mathbb{R}^{m\times n}$, then $Y \sim \mathcal{N}(\boldsymbol{c} + \boldsymbol{D\mu}, \boldsymbol{D\Sigma D^{-1}})$.

Let us now have the CDRs, i.e., a sequence of time and place of user's communication activity, and let us select any two consecutive records $A = (x_A, y_A, t_A)$ and $B = (x_B, y_B, t_B)$ at distinct places, i.e. $(x_A, y_A) \neq (x_B, y_B)$. Then the joint transformation for all Gaussian mixture components in the ICM model is given by the translation vector $\boldsymbol{c}$ and the transformation matrix $\boldsymbol{D} = \boldsymbol{SR}$ that combines scale ($\boldsymbol{S}$) and rotation ($\boldsymbol{R}$) matrices:

$$\boldsymbol{c} = \begin{bmatrix} -x_A \\ -y_A \\ -t_A \end{bmatrix} \qquad \boldsymbol{S} = \begin{bmatrix} d|_{xy}(A,B) & 0 & 0 \\ 0 & d|_{xy}(A,B) & 0 \\ 0 & 0 & t_B - t_A \end{bmatrix}$$

$$\boldsymbol{R} = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) & 0 \\ \sin(\varphi) & \cos(\varphi) & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{x_B - x_A}{d|_{xy}(A,B)} & -\frac{y_B - y_A}{d|_{xy}(A,B)} & 0 \\ \frac{y_B - y_A}{d|_{xy}(A,B)} & \frac{x_B - x_A}{d|_{xy}(A,B)} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.6}$$

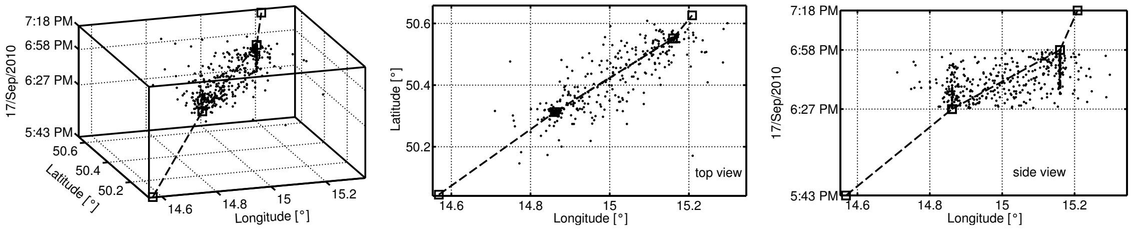
where $d|_{xy}(A, B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$ is Euclidean distance in spatial dimension and φ denotes the angle of rotation counter clockwise about the origin.

The probability $\Phi_{AB}(x, y, t)$ of finding a user at coordinates $(x, y) \in \mathbb{R}^2$ at a time $t \in \mathbb{R}, t_A \leq t \leq t_B$ is then a Gaussian mixture from Equation 4.5 with parameters

$$\begin{aligned} \boldsymbol{\theta}_{AB} = (&\pi_1, \boldsymbol{c} + \boldsymbol{D\mu}_1, \boldsymbol{D\Sigma}_1\boldsymbol{D^{-1}}, \pi_2, \boldsymbol{c} + \boldsymbol{D\mu}_2, \boldsymbol{D\Sigma}_2\boldsymbol{D^{-1}}, \\ &\ldots, \pi_K, \boldsymbol{c} + \boldsymbol{D\mu}_K, \boldsymbol{D\Sigma}_K\boldsymbol{D^{-1}}), \end{aligned} \tag{4.7}$$

where π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ come from the original ICM model parameter vector $\boldsymbol{\theta}$ (Table 4.2).

Two main applications of the ICM model—sampling users' positions between their calls and conditioning in spatial and temporal dimensions—are described in detail in the following paragraphs.



(a) *Sampling example* **(b)** *xy-projection* **(c)** *xt-projection*

Figure 4.14: Example of sampled user's position. Each black dot represents a probable position of a user between call places according to the ICM model.

Estimating user's position

Sampling from the $\Phi_{AB}(x, y, t)$ distribution delivers user's position between the communication events at places A and B. It can be achieved by common means of sampling from a mixture model in two steps: first, probabilities π_k represent the probability that a sampled point comes from the kth component; second, mean vector $\boldsymbol{c}+\boldsymbol{D}\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{D}\boldsymbol{\Sigma}_k\boldsymbol{D}^{-\mathbf{1}}$ define the kth component, a three-dimensional normal distribution from which sampling is usually done by applying the Cholesky decomposition.

Figure 4.14 shows an example of 500 user's positions sampled from the $\Phi_{AB}(x, y, t)$ distribution between two real-world call places $A = (14.86, 50.31,\ \text{6:27:41 PM})$ and $B = (15.16, 50.55,\ \text{6:58:08 PM})$.

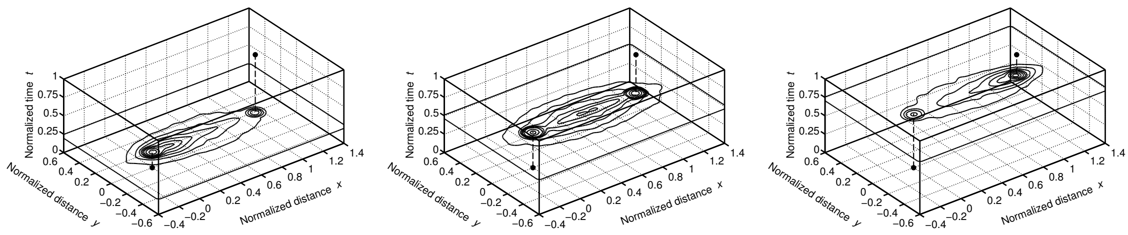
Conditioning in the Spatial Dimension

To estimate user's position at a particular time t, i.e., *where* the user is at a particular time in between consecutive calls, the ICM model allows for expressing conditional spatial distribution $p(\boldsymbol{z}_s|z_t = t)$. For each mixture component $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the corresponding spatial and temporal components can be expressed separately as

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{s,k} \\ \mu_{t,k} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_{ss,k} & \boldsymbol{\Sigma}_{st,k} \\ \boldsymbol{\Sigma}_{ts,k} & \Sigma_{tt,k} \end{bmatrix}. \tag{4.8}$$

Given a temporal value t, the expected spatial distribution $\Phi^t(x, y)$ is then

$$p(\boldsymbol{z}_s = (x, y)^T | z_t = t) = \sum_{k=1}^{K} \beta_k p_k(\boldsymbol{z}_s | z_t), \tag{4.9}$$



(a) $t = 0.25$ **(b)** $t = 0.50$ **(c)** $t = 0.75$

Figure 4.15: Example of conditioning in the temporal dimension. $\Phi^t(x, y)$ for different values of t.

where $\beta_k = p(k|z_t)$ is the probability that the kth component contains values for a particular temporal value z_t,

$$\beta_k = \frac{p(k)p(z_t|k)}{\sum_{i=1}^{K} p(i)p(z_t|i)} = \frac{\pi_k \mathcal{N}(z_t; \mu_{t,k}, \Sigma_{tt,k})}{\sum_{i=1}^{K} \pi_i \mathcal{N}(z_t; \mu_{t,i}, \Sigma_{tt,i})}, \tag{4.10}$$

and $p_k(\boldsymbol{z}_s|z_t)$ is the expected spatial distribution of the kth component given a temporal value z_t,

$$p_k(\boldsymbol{z}_s|z_t) = \mathcal{N}(\boldsymbol{z}_s; \hat{\boldsymbol{\mu}}_{s,k}, \hat{\boldsymbol{\Sigma}}_{ss,k}), \tag{4.11}$$

$$\hat{\boldsymbol{\mu}}_{s,k} = \boldsymbol{\mu}_{s,k} + \boldsymbol{\Sigma}_{st,k}(\Sigma_{tt,k})^{-1}(z_t - \mu_{t,k}), \tag{4.12}$$

$$\hat{\boldsymbol{\Sigma}}_{ss,k} = \boldsymbol{\Sigma}_{ss,k} - \boldsymbol{\Sigma}_{st,k}(\Sigma_{tt,k})^{-1}\boldsymbol{\Sigma}_{ts,k}. \tag{4.13}$$

Figure 4.15 shows and example of the expected spatial distribution $\Phi^t(x, y)$ for different values of t, derived as explained above.

Conditioning in the Temporal Dimension

For the case when the time estimation for a particular position is needed, i.e., one wants to know *when* a user will be at a particular position, we derived the expected temporal distribution $\Phi^{xy}(t)$ for spatial coordinates (x, y):

$$p(z_t|\boldsymbol{z}_s = (x, y)^T) = \sum_{k=1}^{K} \gamma_k p_k(z_t|\boldsymbol{z}_s), \tag{4.14}$$

where $\gamma_k = p(k|\boldsymbol{z}_s)$ is the probability that the kth component contains values for particular spatial coordinates $\boldsymbol{z}_s$,

$$\gamma_k = \frac{p(k)p(\boldsymbol{z}_s|k)}{\sum_{i=1}^{K} p(i)p(\boldsymbol{z}_s|i)} = \frac{\pi_k \mathcal{N}(\boldsymbol{z}_s; \boldsymbol{\mu}_{s,k}, \boldsymbol{\Sigma}_{ss,k})}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{z}_s; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{ss,i})}, \tag{4.15}$$

and $p_k(z_t|\boldsymbol{z}_s)$ is the expected temporal distribution of the kth component given spatial coordinates $\boldsymbol{z}_s$,

$$p_k(z_t|\boldsymbol{z}_s) = \mathcal{N}(z_t; \hat{\mu}_{t,k}, \hat{\Sigma}_{tt,k}), \tag{4.16}$$

$$\hat{\mu}_{t,k} = \mu_{t,k} + \boldsymbol{\Sigma}_{ts,k}(\boldsymbol{\Sigma}_{ss,k})^{-1}(\boldsymbol{z}_s - \boldsymbol{\mu}_{s,k}), \tag{4.17}$$

$$\hat{\Sigma}_{tt,k} = \Sigma_{tt,k} - \boldsymbol{\Sigma}_{ts,k}(\boldsymbol{\Sigma}_{ss,k})^{-1}\boldsymbol{\Sigma}_{st,k}. \tag{4.18}$$

4.4 Inter-Call Mobility Model Evaluation

In this section we discuss the ICM model's fit to the empirical RMD data, and compare the ICM model with an existing state-of-the-art model.

4.4.1 Goodness-of-fit Test

We performed a goodness-of-fig (GOF) test to examine the hypothesis that data sampled from the ICM model and the original RMD data share the same parent distribution, i.e., that the model provides true description of user's inter-call positions. We performed a multivariate distribution-free GOF test, based on the *mixed-sample method* [98].

The hypothesis that the RMD-empirical and the ICM-model distributions are identical was rejected at 95% confidence level, i.e., from a statistical point of view the two distributions differ. Such a result is not so surprising because even a visual comparison of Figures 4.5 and 4.12 shows that the model is an *approximation* of the empirical data. The model fits worse at the very beginning and end of the time budget (about the first and last 5% of total inter-call time), as it over-generalizes the position in the spatial dimension around the origin and destination places of the normalized inter-call trajectories.

More than 10 components (Gaussian mixtures) in the model would bring a better fit, but would add another 10 parameters *per component* and a risk of overfitting. We observed better fit when multivariate Student t-mixtures [145] were used instead of Gaussian ones in the ICM model. However, additional parameters are needed for the description of t-mixtures and working with a t-mixture model (sampling, conditioning) is more complicated. Therefore, we sacrifice the accuracy of the fit for the ICM model's fruitful features of sampling from Gaussian mixtures, conditioning in both spatial and temporal dimensions and simple application to any CDR traces.

We have used the Gaussian mixture model for it is a well-established method of statistical modeling, and it yields far the most accurate results among other approaches that we applied to approximate the aggregated inter-call trajectories of the RMD user-pool. A Hidden Markov Model [170] may be useful in better description of temporal variations of the data as it captures the transition probabilities between the Gaussian mixtures. However, such approach describes only a fixed number of temporal vales and is thus more suitable for learning and direct application, rather than building a general model. Inspired by Winter [203], who described his mobility model as a bivariate Gaussian distribution parameterized by a function of time, we tried to fit each time-slice of inter-call mobility distribution (Figure 4.6) separately as a two-dimensional Gaussian mixture. However, it led to an unacceptably wide range of standard errors in distribution parameters and thus we use the three-dimensional GMM.

4.4.2 Comparison with Other Models

To the best of our knowledge, there are no existing models describing inter-call mobility. Instead, only models that explain user's movement between two places in general are available, but they lack conditioning on communication activity at origin and destination places. Such models are as follows:

- *Linear weighted interpolation* [92], simply a straight line between two places, do not describe the real user's position precisely enough as demonstrated in Section 4.2.3.
- *Space-time prism* [94] represents positions reachable by a user in space and time as a volume in a space-time cube that a user is able to visit given his/her origin and destination positions, time budget, and maximum speed. Space-time prisms allow only for evaluation of binary statements: for example, the potential for encounter between two moving users exists if their corresponding space-time prisms intersect. Therefore, the space-time prisms are not comparable with the probabilistic ICM model.
- *Probabilistic extension of space-time prism* is the current state-of-the-art model by Winter [203]. It describes a non-uniform probability distribution within the space-

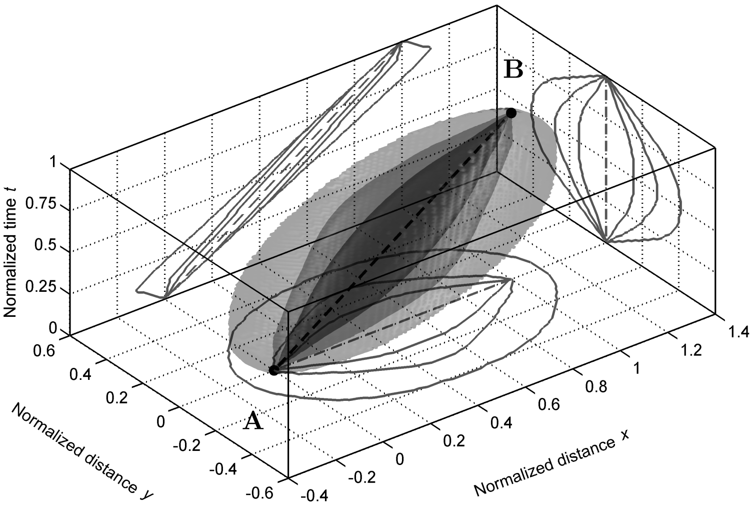


Figure 4.16: Probabilistic extension of space-time prism. Theoretical model of spatio-temporal probability distribution of user's position between two distinct places (A and B) by Winter [203]. Isosurfaces enclose 0.5-, 0.9- and 0.99-quantiles (dark to pale blue, respectively). Figure show distribution for user's maximum normalized speed $v_{\max} = 1.3$.

time prism (see Figure 4.16). The model is based on a biased random walk that creates a bivariate normal distribution of user's position at time t. Winter's model holds under the assumption that movement from origin to destination is linear over time. It implies that users are locally not more likely to be faster than slower (or vice versa) when compared to their average speed. Therefore, the observation of the "staying at call places" behavior of users within RMD (in Section 4.2.2) is in high contrast with the assumptions of the Winter's model, and is the reason why the shapes of volumetric representations in Figures 4.5 and 4.16 differ so strikingly.

Figure 4.17 shows the root-mean-square error (RMSE) measure of difference between the observed spatial probability distribution from the RMD, computed for 50 linearly-spaced time moments between calls, and the distributions given by different models. Based on this measure, the ICM model fits the empirical RMD data better than both the Linear interpolation approach and the state-of-the-art theoretical model by Winter.

4.5 Case Study: Inferring Proximity Probability

In this section we show how the Inter-Call Mobility model can be used for inferring proximity probability of mobile network users, based on their Call Data Records. We demonstrate, on the example of users' proximity probability in two hypothetical scenarios, that using the ICM model yields better performance than the general, theoretical model

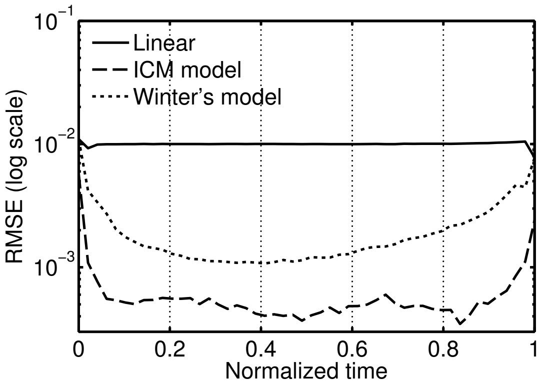


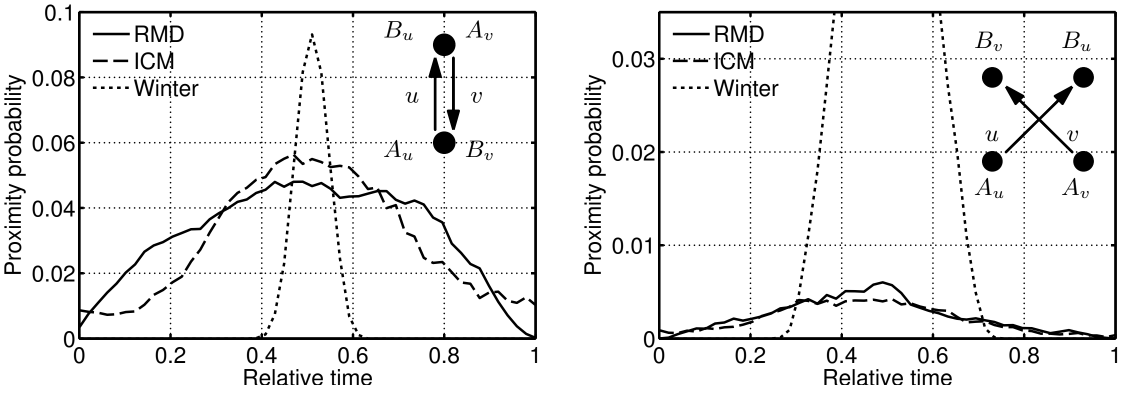
Figure 4.17: Comparison between models. The lower the RMSE value, the better accuracy of the model.

by Winter [203]. This is caused mainly by the fact that the ICM model captures the time variability in late departures and early arrivals, which is crucial for inferring user's position in between communication events.

Proximity probability is important for example in studies on Bluetooth viruses spreading [197] or opportunistic data dissemination [97]. Unlike the authors in [197], who computed the proximity probability from the expected number of users in a mobile network cell and the area covered by the cell, we use inter-call mobility characteristics[3]. Given CDRs, the ICM model can deliver probability of user's position between consecutive calls at particular time t, as we demonstrated in Section 4.3.4. By sampling a large number of positions between call places for two different users, and measuring the distance between sampled pairs of position, the proximity probability can be calculated as a fraction of tuples with distance shorter than a value given by a proximity definition.

By the term *proximity* we understand that two users are closer then x percent of the inter-call distance. Specifically, example in Figure 4.18 shows 5% proximity probability computed from the ICM model, Winter's model, and from the estimation of inter-call distribution from the RMD in Figure 4.5. The probability was computed using a Monte Carlo simulation with 10^7 sampled positions at 50 linearly-spaced time moments between calls. The overall proximity probability, computed as integration over the temporal domain, would result in only a single number, but we aimed to show *where*, over the inter-call time, the highest difference between mobility models and empirical data is. Figure 4.18 shows the *proximity probability* of two users between their calls in two *hypothetical* scenarios:

[3] Note that the distances considered in our proximity modeling are greater than those of Bluetooth proximity, but we are more concerned with general observations on modeling user proximity rather than this particular application.



(a) *Opposite travel direction*

(b) *Cross travel direction*

Figure 4.18: Proximity probability of two users. Users u and v travel between their calls at places A_u, B_u and A_v, B_v in two hypothetical scenarios (a) and (b). We assume that both users depart and arrive at the same time and travel the same distance.

Scenario (a), in which *users travel towards each other*, shows that the RMD data demonstrate several times higher proximity probability than the Winter's model, over majority of time. It is a consequence of user's staying behavior at communication places: both users can simply meet even at the very ends of the time budget, at the origin and destination places. Shapes of the ICM model and of the RMD data differ slightly, mainly because the Gaussian components fail in capturing rapid changes in space-time domain.

Scenario (b), with *non-identical user origin and destination places*, shows a significantly lower meeting probability for the ICM model and the RMD in comparison to Scenario (a). It follows from the fact that a meeting chance at origin or destination places, here impossible, contributes highly to the overall proximity probability. The assumption of linearity over time in Winter's model causes much higher meeting likelihood in the path-cross area—but in reality users tend to stay at their origin and destination places.

Note that the proximity probability in both scenarios is so low because we do not use any geographical background (roads, streets) to support the RMD data or the ICM model with - it is a general description of movement in space.

4.6 Conclusion

In this chapter, we have introduced the Inter-Call Mobility model, a spatio-temporal refinement of modeling user's movement between consecutive mobile communication records. It significantly improves the CDR-based deduction of user's presence at some place in time: from a single timestamped call coordinate to a probabilistic distribution of user's position between calls. Using a finite Gaussian mixture model, the ICM model approximates the ground-truth of users' inter-call movement, mined from the Reality Mining Dataset. It reflects three main findings in inter-call mobility: (1) unskewed behavior in spatial dimension, (2) straight course between calls (3) staying behavior at call places.

The ICM model outperforms two existing state-of-the-art models of movement in space and time, the linear weighted interpolation and the probabilistic extension of space-time prism, yet mainly because the assumptions of the models are incompatible with the characteristic inter-call behavior of the users.

Chapter 5

Exploring the Limits of Crowdsensing

In this chapter we investigate the limits of crowdsensing in discovering the mapping of mobile network Cell-IDs to geographic locations. More specifically, we focus on characterizing sufficient number of users to cover a given fraction of cells in a cellular network during a limited time.

Crowdsensing techniques are often utilized by communication market players such as Google and Apple to discover the structure of mobile networks. GPS-enabled phones of customers send their current GPS coordinates and Cell Identifiers (Cell-IDs) to a server that collects, clusters, fingerprints, and stores such data from all customers. Such a Cell-ID database can subsequently lead to the geolocation: given a Cell-ID, an approximated position inside the cell is returned. This enables services such as localization or friend proximity lookup, even for mobile phones without GPS receivers.

In this chapter we solve the problem of *discovering* the cells in a network, which means that only one GPS coordinate per Cell-ID is considered sufficient to represent the cell position. However, in a Cell-ID database, more GPS measurement could be associated with one Cell-ID and vice versa, even for one GPS coordinate there can be different Cell-IDs. A problem of identifying the Cell-ID for a cell, whose coverage is dominant over a given GPS coordinate with the largest probability, is not in the scope of this chapter.

We pose two key questions: What is the required minimal size of a user group needed for obtaining a critical mass of knowledge about the mobile infrastructure? And, how much time is needed to do so? We assume a probabilistic model of user mobility, infer parameters of the model using a real-world mobility trace, and simulate the model to estimate the fraction of cells visited by a fixed number of users over a time interval. This is vital to judge the ability of crowdsensing to build a new Cell-ID database, or rapidly update an inadequate, malfunctioning or obsolete Cell-ID database.

The rest of this chapter is organized as follows: First, in Section 5.1 we present describe our method, based on building and simulating a mobility model, and the available dataset. Second, in Section 5.2 we build a mobility model, based on the NRC-Lausanne dataset [129], that reflects both the temporal properties of human mobility patterns and the number of user-cell associations. Such model is necessary for sampling a high number of artificial, yet realistic user mobility patterns to serve as simulation input. Third, in Section 5.3 we perform a large-scale simulation of mobile user movements in an approximation of a mobile network, counting the number of distinct cells users visit over a period of time. Such simulation is vital because basic statistics on user movement (average number of unique cells visited during a day) cannot be used alone to form the mobility patterns, as different users are likely to visit the same cells during the day. Next, in Section 5.4 we summarize the simulation results. We discuss validity of the results with respect to the dataset and the simulation settings we used. Finally, in Section 5.5 we provide detailed analysis of simulation results to demonstrate the ability of crowdsensing to deal with obfuscation of the network topology by a particular mechanism, called Dynamic Cell-ID method.

5.1 Method Description

Any dataset that contains information about user-cell association over time gives away the coverage capabilities of the user pool for the time period in an area. However, the size of the userpool is critical, for tens of users cannot cover the whole mobile network (which typically consists of thousand of cells), so several-times more users' traces are necessary in a study of this kind. Therefore, we try to infer general coverage capabilities of the population. We build a model that captures general features of users' movement within the network to generate a high number of synthetic, yet realistic traces. Using simulation we apply the model-based traces of user movement to an approximation of a mobile network topology. By this we determine the fraction of mobile-network cells visited by a fixed number of users over a time interval.

5.1.1 Assumptions

We make three key assumptions regarding the act of crowdsensing in discovering the mapping of mobile network Cell-IDs to geographic locations.

A1) The crowdsensing campaign is short, at most one day long. This assumption precludes excessive use of users' mobile phone resources for a long time, thus preventing the battery drain. Moreover, for the motivating use-case of Cell-ID crowdsensing, described in Section 5.5, longer campaigns are useless.

A2) GPS receivers serve as the only tool for reporting geographical positions of the user. This approach represents the simplest terminal-based positioning solution. Although other positioning techniques or phone-sensed data may be available (such as WiFi beacons), we assume this wide-spread and accurate method.

A3) Mobile phones only report the Cell-ID of the currently attached cell. This assumption comes from the fact that the knowledge of neighboring Cell-IDs, although feasible to obtain, would not immediately bring any advantage and it needs more data to be transmitted between the mobile phone ad the Cell-ID database. Neighbor cells may be used in some future extensions to deliver the approximate position of a cell within the network.

5.1.2 Dataset

The data we use come from the NRC-Lausanne dataset, which consists of information about 168 users who participated in the Lausanne Data Collection Campaign [129]. This dataset was released for the purpose of the Nokia Mobile Data Challenge [143] and for specific reasons, it was divided in several parts. We were provided with a part which consist of information about 38 users, in the following text we will call it simply a *dataset*. It contains a timestamped sequence of Cell Global Identities (CGIs) per user with one record per every cell change during the campaign period (referred to as *cell trace*). Also, there is a timestamped GPS log for each user, yet it contains numerous gaps and does not cover the entire campaign period (we call it a *GPS trace*). The dataset covers a large part of Switzerland, including major cities and the countryside, as depicted in Figure 5.1. Table 5.1 provides basic summary statistics about the dataset.

Since there are only 38 users in the dataset, the concerns about validity of any statistics measured from such a small set of users could arise. Is the userpool with 38 users capable of providing relevant information about the network, i.e., could we estimate the

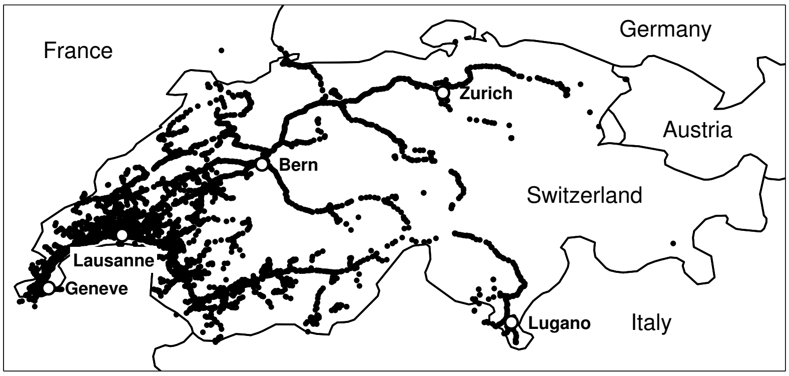


Figure 5.1: NRC-Lausanne dataset coverage. Each black dot represents one cell tower a user from the dataset was attached to. The map shows the coverage capability of 38 users during one year.

	min	**0.25-q**	**mean**	**0.75-q**	**max**	**stdev**
Days in survey per user	94	179	282	371	533	116.09
Cell transitions per user	2,540	7,781	16,899	23,212	38,187	9,314
Unique cells per user	202	684	1,206	1,758	2,723	693
Ratio of cells with GPS coords per user	15.67%	44.85%	68.12%	86.70%	97.73%	22.21%

Table 5.1: Summary statistics for the NRC Lausanne dataset. The total number of users is 38, period of collection from 30-Sep-2009 to 31-Mar-2011.

total number of cells in the network at least? This is related to the cardinality/species-richness estimation problem, presented in Section 2.4.3. Let us consider that each user represents an independently selected sampling unit that reports visited cells as incidence data, i.e., each visited cell in the network is simply noted as being present. Then, we can use such incidence data from multiple samples (users) in aggregate to estimate the total number of cells in the whole network. The estimator, conveniently named Chao2 [44], is defined for estimating species richness as:

$$S_{est} = S_{obs} + \left(\frac{t-1}{t}\right)\frac{q_1^2}{2q_2}, \tag{5.1}$$

where S_{obs} is the number of species observed in t samples, and q_1 and q_2 represent the number of species that occur in exactly one sample or in exactly two samples, respectively [51]. In our case and the dataset we use, $S_{obs} = 23{,}638$ is the number of unique cells observed by $t = 38$ users, and $q_1 = 17{,}141$ and $q_2 = 2{,}265$ are the numbers of cells observed by exactly one or by exactly two users, respectively. The estimated number of unique cells in the network is then $S_{est} = 86{,}791 \pm 44$ cells (with 95% confidence interval). Such a number seems adequate to the area of dataset cover: As reported in [129], all 168 participants of the Lausanne Data Collection Campaign observed in total 46,082 cells during the whole

			m_g			
t	S_{obs}	S_{est}	$g = 0.9$	$g = 0.95$	$g = 0.99$	$g = 1$
38 users	23,638 cells	86,791 cells	279 users	376 users	602 users	5,161 users

Table 5.2: Estimated sampling effort to discover all cells in the network. Abbreviations are: t, number of users (samples) in the NRC Lausanne dataset; S_{obs}, observed number of unique cells; S_est, estimated asymptotic number of unique cells; g, target fraction of S_{est}; m_g, the number of additional users needed to reach 90% ($g = 0.9$), 95% ($g = 0.95$), 99% ($g = 0.99$), and 100% ($g = 1$), respectively, of S_{est}.

period. So, how many users are required to reach the asymptotic number of unique cells estimated by Chao2 estimator? In [46], Chao derived that to reach a fraction g of S_{est} for sample-based data, the required number of additional samples m_g is

$$m_g \approx \frac{\log\left[1 - \frac{t}{t-1}\frac{2q_2}{q_1^2}(gS_{est} - S_{obs})\right]}{\log\left[1 - \frac{2q_2}{(t-1)q_1+2q_2}\right]}. \tag{5.2}$$

Table 5.2 illustrates the calculation of m_g, the number of users needed to achieve gS_{est} for the period of Lausanne Data Collection Campaign. As a result, we conclude that more users than 38 in the dataset we use would be more suitable for a mobility-related study, however, such a large dataset were not publicly available.

5.2 Data-driven Mobility Model

In this section we present our approach to building a data-driven mobility model.

5.2.1 Model Objectives

We aim to model user mobility patterns in terms of the number of unique cells visited *during one day*, starting from midnight. A common denominator of users' daily mobility is their presence at some *places*, such as home, work, the cinema, etc. for a substantial amount of time. Each of these places (which are not usually many in one day) is covered by a cell or a set of cells. A key observation here is that the *transitions* between places account for most of the total number of unique cells visited by a user during a day.

The objective of the mobility model is to capture

F1) the total number and ordering of places as they are visited by a user during a day,

F2) the start time of all user's transitions between places in the day, and

F3) the duration of transitions and their length, measured in the number of unique cells visited during the transition.

Such model would differ from the large body of similar work (see [115] for a survey and Section 2.4.3 for previously proposed models) in that it incorporates all of the features F1–F3 concurrently. As a result, it would describe the users' *daily patterns* (F1), capture *fine-grained* temporal characteristics of human movement *during one day* (F2), and quantify daily *user-cell associations* (F3). It is not an objective to capture any real-world counterparts of places, their semantic labels, geographical position or mutual distance. Instead, the metric of distance we use is the number of unique cells during transition, which we consider sufficient to simulate user-cell associations over a time period.

5.2.2 Model Description

We assume the following mobility model. We define a *place* as a set of neighboring cells (details are given later in Section 5.2.3) in which the user cumulatively spends a significant amount of time during a day. A *transition* between places is the act of leaving a place and visiting another, or the same place subsequently some time later. We enumerate places in each day-sequence according to the time of the first visit with numbers $L = \{1, 2, \dots\}$.

Users' daily patterns are represented by the number and ordering of different places visited by a user during a day. For example, a typical daily pattern for the majority of users would be 1—2—1, where 1 could represent "home", 2 could stand for "work" and dash (—) denotes transition between places. A user visits these places with different *probabilities* during the day. By $p_{i,j}^{\widetilde{t}}$ we denote the probability that a transition between places $i, j \in L$ *starts* during the time period $\widetilde{t}$, where $\widetilde{t} \in \{1, \dots, T\}$ represents one of T equidistant time slots in a quantized structure of time during the day. Note that this next-place selection process is not a time-variant Markov chain and that probabilities $p_{i,j}^{\widetilde{t}}$ for a given $\widetilde{t}$ do not constitute a transition probability matrix: Since it is possible to make a transition from a place i to the same place i, i.e., $p_{i,i}^{\widetilde{t}} \neq 0$, the probability that a user stays at a place i is then $1 - \sum_{j \in L} p_{i,j}^{\widetilde{t}}$. Moreover, the probability $p_{i,j}^{\widetilde{t}}$ is conditional on a set of places L' the user visited during time $t = 1, \dots, \widetilde{t} - 1$, and thus it does not have memoryless property.

The duration of transitions, measured as a fraction of a day, is given by distributions $f_{\text{new}}, f_{\text{same}}, f_{\text{old}}$, according to the transition classes. We distinguish three different transition classes, depending on the relationship between the origin O_i and destination D_i places during the day: A transition is classified as *new* if it ends at a new, not previously visited place, *same* if it starts and ends at the same place, and *old* if it is between places already visited. The length of transitions, expressed in the number of unique cells visited during the transition, is given by distributions by g_{new}, g_{same}, and g_{old}.

To describe user's movement during a day, we define a *transition sequence* $TS = \{(O_i, D_i, t_i, \delta_i, l_i)\}_{i=1}^{N}$, where N denotes the total number of transitions during one day, $O_i, D_i \in L$ are the origin and destination places of the i-th transition, $t_i, \delta_i \in [0, 1] \in \mathbb{R}$ represent the time (by a fraction of the day) of the transition start and its duration, and $l_i \in \mathbb{N}$ is the length of the transition expressed in the number of unique cells visited during the transition. The TS is empty for a user who spends the entire day at one place and makes no transitions. In consistency with the description above, $\delta_i \sim f_*$ and $l_i \sim g_*$, where $*$ denotes transition class according to O_j, D_j for $j = 1, \ldots, i$. Figure 5.2 shows an example of one day-sequence with three transitions T1–T3 of a user from the dataset. Its corresponding transition sequence[1] is as follows:

$$\begin{aligned} TS &= \{(O_i, D_i, t_i, \delta_i, l_i)\}_{i=1}^{3} \\ &= \{\mathrm{T1}, \mathrm{T2}, \mathrm{T3}\} \\ &= \{(1, 2, 11{:}40, 1{:}10, 43), \quad (2, 1, 16{:}10, 01{:}30, 64), \quad (1, 2, 21{:}20, 00{:}50, 46)\}. \end{aligned}$$

5.2.3 Data Processing and Recognizing Places

We process the data by dividing the cell trace of each user into day-long sequences, each starting at midnight. Days where the mobile phone was off were excluded. Because the daily routine of users and their mobility significantly differ between weekdays and weekends, these must be handled separately. Without loss of generality we restrict the dataset to weekdays only, however, weekends can be modeled in a similar manner. Finally, we consider all of the day-long weekday sequences to be independent, even when belonging to the same user. Handling the data in such way is viable because the mobility model is to describe *only* a period of one day and so weekday correlations need not be reflected. There is a total of 6,667 day-sequences in the dataset.

Important places for a user are recognized from mobile data by clustering methods [125, 208, 149]. Clustering is vital because a place is typically covered by overlapping

[1] Please note that time in the sequence is written in HH:MM format for better readability.

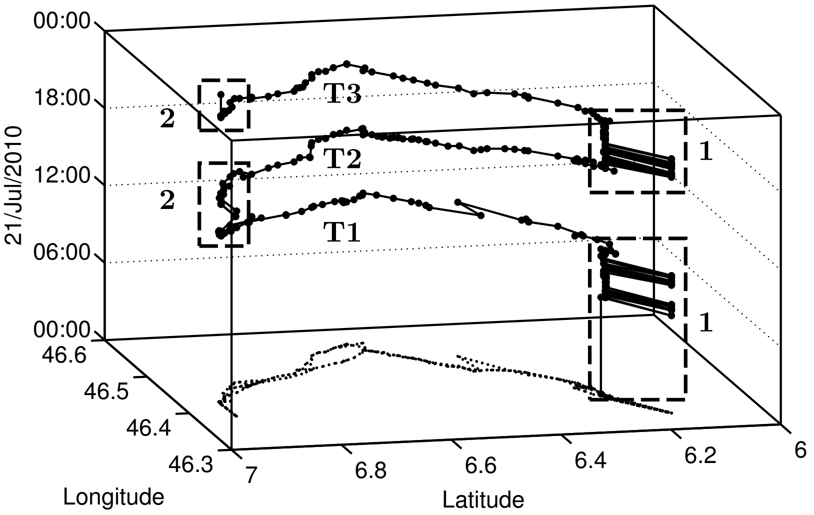


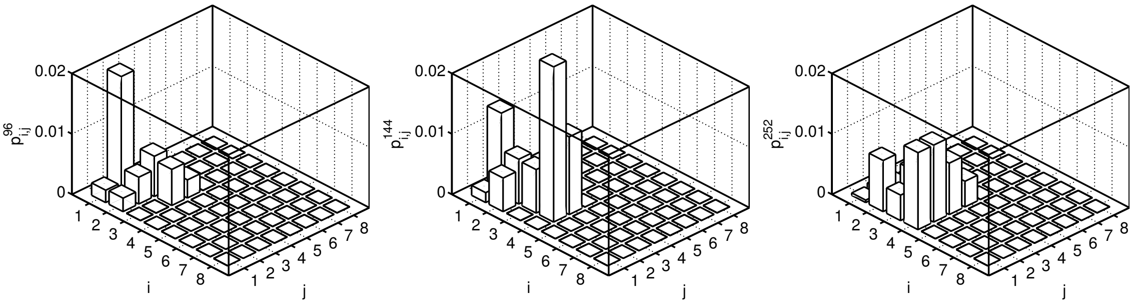
Figure 5.2: Example of a transition sequence. User's movement in one day-sequence is captured in a space-time cube. Dots represent user-cell association, solid lines stand for transitions between cells, and dotted line represents orthogonal projection on coordinate plane. Rectangles with labels 1 and 2 enclose sets of cells which represent *places*—the user visited two places (1 and 2) during the day. There are in total three transitions between the places during a day: T1 and T3 between places 1 and 2, and T2 between places 2 and 1.

cells, and the mobile phone connects to them even when the user is not moving (so called *cell jitter*). We use time-based clustering to recognize user's places, because the GPS trace covers only 32% of all cells in the cell trace which precludes spatio-temporal clustering. We define a *place* as a set of neighboring cells in which the user cumulatively spends more than 60 minutes anytime during a day. A *transition* between places is the act of leaving a place and visiting another, or the same place subsequently at least 4 minutes later.

5.2.4 Inferring Model Parameters

In this section we explain how transition sequences TS, obtained from the dataset, were used to infer model parameters.

We express the model features F1 and F2 by mining the transition probabilities between places, $p_{i,j}^{\widetilde{t}}$, depending on time of day. We simplify the structure of time by quantizing the day into $T = 288$ 5-minute equidistant time slots $\widetilde{t} \in \{1, \ldots, T\}$. So, for example $\widetilde{t} = 2$ represents a time period from 12:05 a.m. to 12:10 a.m. Then, $p_{i,j}^{\widetilde{t}}$ is estimated from transition sequences TS as $n_{i,j}/(\sum_{j=1}^{L} n_{i,j} + n_{i,\oslash})$, where $n_{i,j}$ denotes the number of occurrences of transitions from a place i to a place j in all transition sequences, and $n_{i,\oslash}$ denote the number of cases when no transition starts when at place i during time period $\widetilde{t}$. Figure 5.3 shows the transition probability $p_{i,j}^{\widetilde{t}}$, derived from the data, on



(a) *08:00–08:05* **(b)** *12:00–12:05* **(c)** *21:00–21:05*

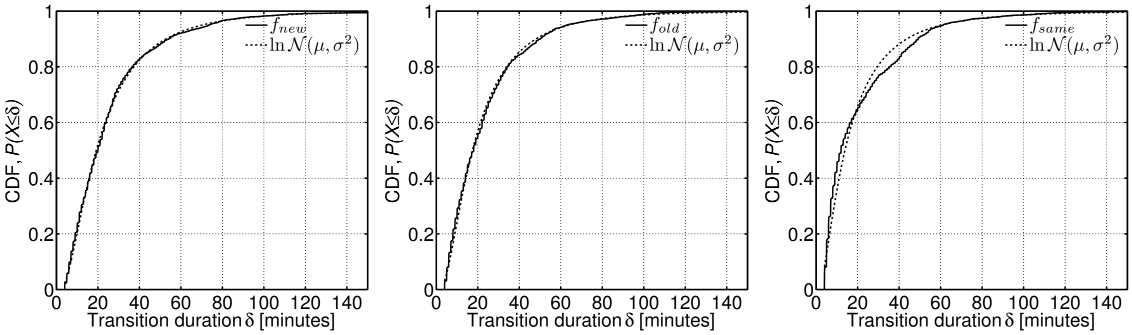
Figure 5.3: Transition probability $p^{\widetilde{t}}_{i,j}$ at different time of the day. The y-axis contains labels i for places of transition origin, the x-axis contains labels j for places of transition destination, height of each bar denotes the transition probability $p^{\widetilde{t}}_{i,j}$ from a place i to a place j during the time period $\widetilde{t}$. In the morning **(a)**, the probability of visiting new places (with higher place label) is noticeably high. Around noon **(b)** the probabilities are concentrated on the main diagonal, which means that users leave a place and return to the same place again later. In the evening **(c)**, users tend to return to previously visited places, which is indicated by transition probabilities concentrated under the main diagonal.

three examples at different times of day. The number of data points needed for computing each row of the transition matrix can be computed using Thompson's formula [187] for estimating sample size n in multinomial proportions:

$$n = \operatorname*{argmin}_{c \geq 1} \left\{ \frac{\frac{1}{c}(1 - \frac{1}{c})\chi^2_{[1,\alpha/c]}}{d^2} \right\}, \tag{5.3}$$

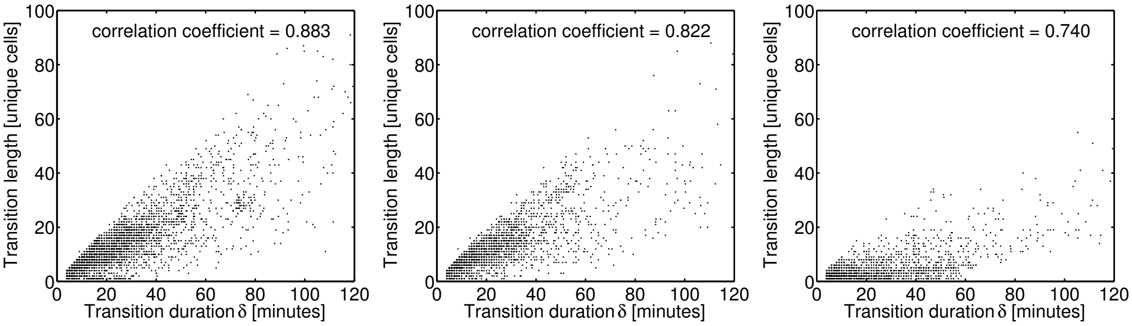
where α is the type I error, and where d denotes the prescribed accuracy, i.e., the maximal difference allowed between any element of the theoretical probability distribution and its empirical estimation. Thomson has also computed the so called "surreal numbers" d^2n for various values of α and showed that n depends only on α and d; for $\alpha = 0.05$ the surreal number $d^2n = 1.2736$. Having $\alpha = 0.05$, obtaining precision $d \leq 0.05$ for $p^{\widetilde{t}}_{i,*}$ would have required at least $n = 1.27359/(0.05^2) = 510$ users who were during the time interval $\widetilde{t}$ at place i. Here, the drawback of the small number of users in the dataset we have used becomes obvious. On the other hand, since it is extremely rare for a user to visit a large number of distinct places during one day (see Figure 5.11 at page 86), one can hardly expect to gather a dataset of enough size for high-precision estimations for the whole transition matrix.

Duration of a transition (model feature F3) is estimated from the transition sequences by probability density functions $f_{\text{new}}, f_{\text{same}}, f_{\text{old}}$. Figure 5.4 shows the distributions of transition duration in each of the transition classes. The figure provides a useful



(a) *Transition class* new **(b)** *Transition class* old **(c)** *Transition class* same

Figure 5.4: Probability density function of transition duration. The dotted lines represent a log-normal fit. Parameters μ and σ with 95% confidence bounds are provided in Table 5.3.



(a) *Transition class* new **(b)** *Transition class* old **(c)** *Transition class* same

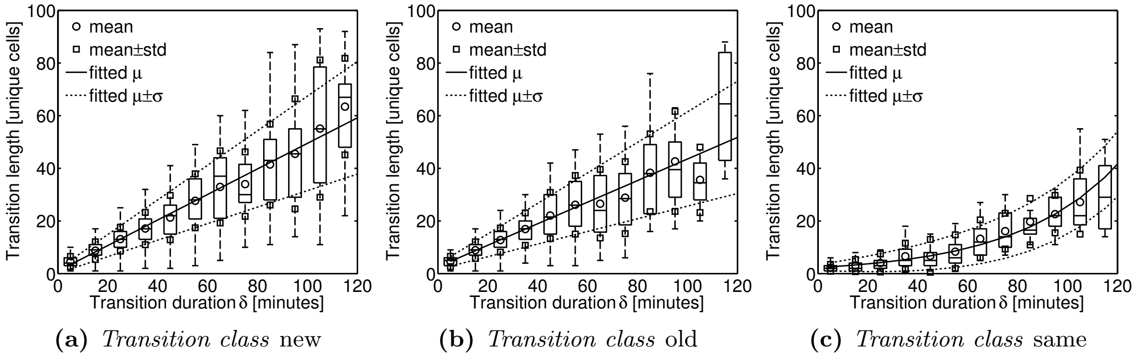
Figure 5.5: Correlation of transition duration and transition length.

insight: while distributions for classes *new* and *old* are quite similar, there are more short transitions in class *same*. As depicted, transition duration can be approximated by a log-normal distribution, although the Kolmogorov-Smirnov test (KS-test) did not reject the null hypothesis that the empirically observed distributions come from the distribution found as its log-normal fit. Table 5.3 provides fit parameters with 95% confidence bounds, KS statistics and critical value at $\alpha = 0.05$ significance level.

Length of transitions (number of unique cells) is positively correlated with duration of transitions (see Figure 5.5), however, there is a significant variation between transition lengths having the same transition duration. One may attribute this to the various means of transport or to the fact that the density of cells varies widely in the mobile

	μ	σ	KS stat	KS critical
$f_{\text{new}} \sim \ln\mathcal{N}(\mu,\sigma)$	-1.1380 ± 0.0007	0.7879 ± 0.0005	0.0340	0.0077
$f_{\text{old}} \sim \ln\mathcal{N}(\mu,\sigma)$	-1.2353 ± 0.0009	0.7906 ± 0.0007	0.0480	0.0094
$f_{\text{same}} \sim \ln\mathcal{N}(\mu,\sigma)$	-1.4635 ± 0.0011	0.8940 ± 0.0008	0.1004	0.0101

Table 5.3: Transition-duration distribution parameters



(a) *Transition class* new **(b)** *Transition class* old **(c)** *Transition class* same

Figure 5.6: Estimation of parameters for transition-length distributions.

network between urban and rural areas, thus making it possible to discover a different number of cells during the same period. Unfortunately, the data do not provide sufficient information for supporting these hypotheses: the GPS trace does not cover the whole cell trace and the network topology can not be reconstructed from the sparse cell trace precisely enough. We approximate the transition lengths by a truncated[2] Normal distribution $\mathcal{N}(\mu, \sigma)$ with the mean and standard deviation parameters dependent on the duration of the transition δ and its class. We denote these distributions by $g_{\text{new}}(\delta)$, $g_{\text{same}}(\delta)$, and $g_{\text{old}}(\delta)$. Figure 5.6 depicts how we estimated the dependence of μ and σ on transition duration in transition length distribution: transition lengths are summarized in boxplots, each boxplot corresponds to transition durations from one of equally spaced bins (width = 10 minutes). Each boxplot then represents a Normal distribution[3]. Parameters μ and σ in distributions $g_{\text{new}}(\delta)$ and $g_{\text{old}}(\delta)$ are linear functions of δ whereas in $g_{\text{same}}(\delta)$ the dependence is expressed by an exponential function for parameter μ and a linear function for parameter σ. Table 5.4 provides fit parameters with 95% confidence bounds and R^2 statistics of the goodness of fit.

[2] Because transition length must be non-negative, integer value, we round the value towards positive infinity and take the maximum from the sampled value and zero.

[3] Our intuition for approximating the transition lengths by a Normal distribution is based on graphical testing for normality (Q-Q plots). In addition to that, we have performed Lilliefors test [138] of the default null hypothesis that the transition lengths in the equally spaced bins come from a distribution in the normal family, against the alternative that it does not come from a normal distribution. At the 5% significance level, one third of the bins passed the Lilliefors test.

	μ			σ		
	a	b	R^2	c	d	R^2
$g_{\text{new}}(\delta) \sim \mathcal{N}(a\delta + b, c\delta + d)$	29.13 ± 3.04	3.38 ± 3.29	0.98	10.31 ± 1.41	1.64 ± 1.53	0.97
$g_{\text{old}}(\delta) \sim \mathcal{N}(a\delta + b, c\delta + d)$	24.59 ± 2.09	4.62 ± 2.06	0.98	10.19 ± 1.43	1.79 ± 1.41	0.96
$g_{\text{same}}(\delta) \sim \mathcal{N}(ae^{b\delta}, c\delta + d)$	2.56 ± 0.82	1.45 ± 0.22	0.97	5.64 ± 1.56	1.55 ± 1.54	0.86

Table 5.4: Transition-length distribution parameters

Algorithm 1 Generation of a synthetic transition sequence TS

```
Input: $p_{i,j}^{\widetilde{t}}$ for all $i, j \in L$ and $\widetilde{t} \in \{1, \ldots, T\}$,
  $f_{\text{new}}, f_{\text{same}}, f_{\text{old}}, g_{\text{new}}(\delta), g_{\text{same}}(\delta), g_{\text{old}}(\delta)$
Output: $TS = \{(O_i, D_i, t_i, \delta_i, l_i)\}_{i=1}^{N}$
  $\widetilde{t} \leftarrow 1$, $i \leftarrow 1$
  $D_0 \leftarrow 1$                                          ▷ first place is 1
  $L' \leftarrow \{1, 2\}$                                    ▷ set of reachable places
  while $\widetilde{t} < T$ do
    if no transition (prob. $1 - \sum_{j \in L} p_{D_{i-1},j}^{\widetilde{t}}$) then
      $\widetilde{t} \leftarrow \widetilde{t} + 1$
    else
      $O_i \leftarrow D_{i-1}$,
      $D_i \leftarrow d \in L'$ with probability $\frac{p_{O_i,d}^{\widetilde{t}}}{\sum_{j \in L'} p_{O_i,j}^{\widetilde{t}} + (1 - \sum_{j \in L} p_{O_i,j}^{\widetilde{t}})}$
      if $D_i = max(L')$ then                                 ▷ new place has been visited
        $L' \leftarrow L' \cup max(L') + 1$     ▷ in the next time period new place can be visited
      end if
      $t_i \leftarrow$ uniformly sampled from interval $\widetilde{t}$
      $\delta_i \leftarrow$ sampled from $f_*$, $*$ denotes transition class
      $\hat{l}_i \leftarrow$ sampled from $g_*(\delta_i)$, $*$ denotes transition class
      $l_i \leftarrow max(0, \lceil \hat{l}_i \rceil)$
      $\widetilde{t} \leftarrow$ nearest time period after $t_i + \delta_i$
      $i \leftarrow i + 1$
    end if
  end while
```

Generation of a new, synthetic transition sequence from the above parameters works according to Algorithm 1.

5.2.5 Model Validation

In this section we show by comparing the features F1–F3 that the synthetic traces from the model correspond to the NRC-Lausanne dataset. We compare the original transition sequences and 6,600 synthetic transition sequences generated by the model.

Users' daily patterns, the model feature F1, represent the number and ordering of different places visited by a user during a day. For example, a typical daily pattern for the majority of users would be 1—2—1, where 1 could represent "home", 2 could stand for "work" and dash (—) denotes transition between places. Figure 5.7 compares the most frequent patterns in the dataset with the synthetic traces generated from the model, showing a high correspondence in the frequency. Figure 5.8 shows that the distribution of users' daily patterns is heavy-tailed, i.e., a small number of patterns occur often while numerous patterns are rare. As depicted, the model captures this daily pattern distribution.

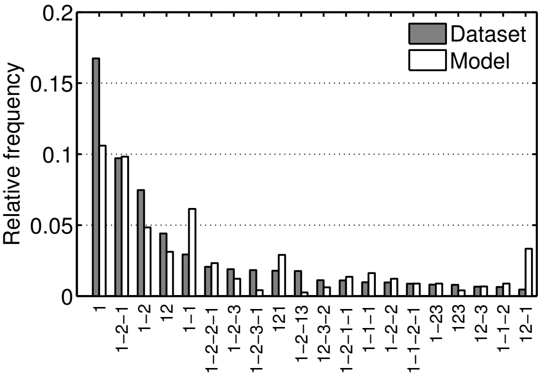


Figure 5.7: The example of the 20 most frequent daily patterns. Interestingly, the most frequent transition pattern represents staying at a place for the whole day.

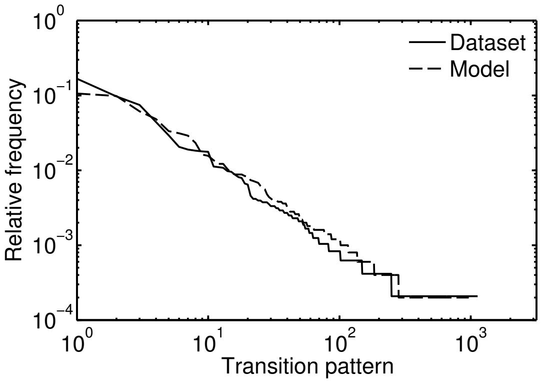
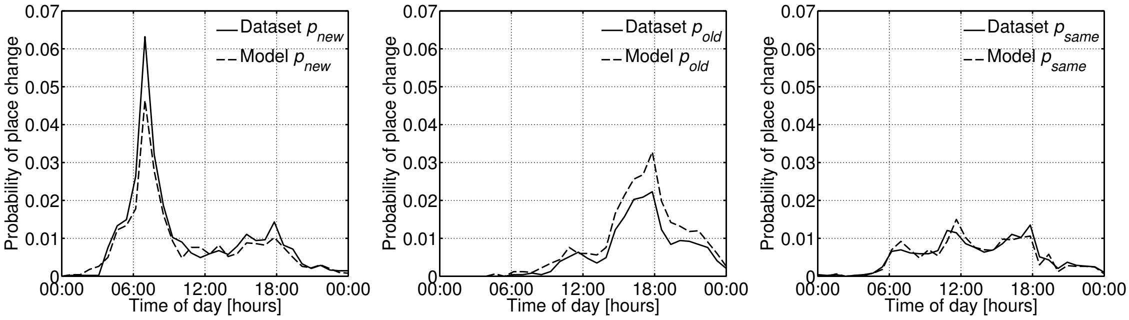


Figure 5.8: Heavy-tail distribution of different daily patterns.



(a) *Transition class* new

(b) *Transition class* old

(c) *Transition class* same

Figure 5.9: Probability of transitions during day.

The fine-grained temporal characteristics of human movement during one day, the model feature F2, are depicted in Figure 5.9. It compares transition probabilities $p_{i,j}^{\tilde{t}}$ during a day per each of the three transition classes. It shows that in the morning users commute to new, previously not visited places (p_{new} in Figure 5.9a), in the afternoon they return to previously visited places (p_{old} in Figure 5.9b), while during the day they tend to leave the place and return to the same place later (p_{same} in Figure 5.9c).

Daily user-cell associations, the model feature F3, are depicted in Figure 5.10. Clearly, the model quantifies the total number of cells visited during a day well. A two-sample Kolmogorov-Smirnov test for the goodness of fit passed—at $\alpha = 0.05$ significance level the hypothesis that empirical and synthetic data are from the same distribution could not be rejected (p-value 0.1873). This is achieved by the fact that the synthetic transition sequences approximately follow similar distributions of total number of the places visited during a day (Figure 5.11) and the total number of transitions between the places (Figure 5.12) from the dataset.

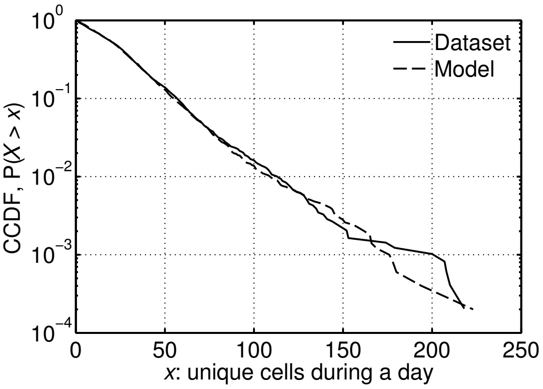


Figure 5.10: Distribution of the total unique cells during a day.

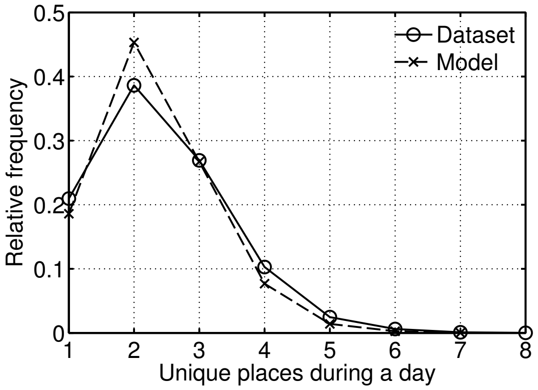


Figure 5.11: Distribution of the total places during a day. Almost half of the transition sequences contain only two places.

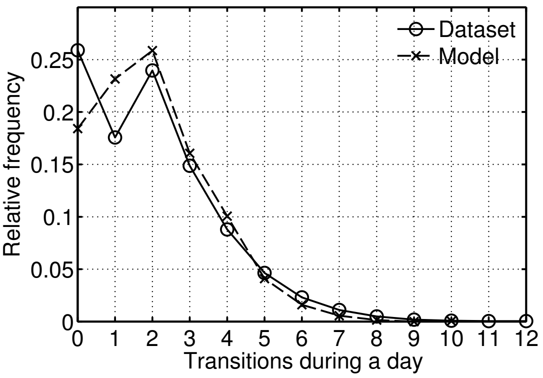


Figure 5.12: Distribution of the total transitions during a day.

5.3 Simulating User-Cell Association

5.3.1 Mobile Network Representation

We model a mobile network with c cells as a Voronoi tessellation [27] of a unit square simulation area where a spatial Poisson process of constant intensity represents the cells' positions[4]. Connectivity between cells (and thus possible handovers) is expressed by the Delaunay triangulation $DT = (V, E)$ [27], the dual of Voronoi tessellation: if cells $u, v \in V$ are neighbors, there exists an edge $(u, v) \in E$.

[4] To obtain a more realistic network, we can use a non-homogeneous Poisson process in which higher density of cells corresponds to the cities, or even apply any relevant background knowledge, such as population density, transportation networks (roads, railways) and commuting patterns in the studied region.

5.3.2 User Movement

We simulate the movement of a set of users $U = \{1, \dots, n\}$ in the network as a discrete-time walk on the DT graph, with probabilities of transitions between nodes and their selection given by transition sequences $TS_u = \{O_i, D_i, t_i, \delta_i, l_i\}_{i=1}^{N}$. These are sampled independently from the mobility model, one transition sequence for each user $u \in U$. As the users traverse the DT graph, the number of distinct nodes visited corresponds to the number of cells they have been associated with.

Let $p_{u,v,k} = \{u = c_0, c_1, c_2, ..., c_{k-1}, c_k = v\}$ denote a simple path (without cycles) between nodes $u, v \in V$ of length $k \in \mathbb{N}$ in the $DT = (V, E)$ graph, i.e. $\forall i, i = 0, 1, \dots k : (c_i, c_{i+1}) \in E$. By the length of k we mean that the simple path has exactly k edges and $k - 1$ unique nodes in the path excluding u and v. And, let $S \in \mathbb{N}^{|V| \times |V|}$ be an all-pairs shortest path matrix for the DT graph with unit edge-weights.

The simulation consists of four steps, explained below by the example of a single user.

Step 1—Selecting places. Let us consider a user's transition sequence $TS_u = \{O_i, D_i, t_i, \delta_i, l_i\}_{i=1}^{N}$. The user's place labels are mapped to DT nodes by a randomly selected one-to-one function

$$m : L \to v, \tag{5.4}$$

where $L = \{O_1\} \cup \{D_i | i = 1, \dots, N\}$ and $v \subset V$. Function m should be found with respect to the lengths of the paths between places in the user's TS, such that

$$\forall i \exists p_{u,v,k} : u = m(O_i) \wedge v = m(D_i) \wedge k = l_i + 1. \tag{5.5}$$

However, finding a simple, k-length path $p_{u,v,k}$ is known to be NP-hard (can be reduced to the Hamiltonian Cycle problem), and even probabilistic algorithms [25] are too slow on large graphs. Therefore, we relax the requirement on path length k in this step and select the function m such that

$$\forall i \exists p_{u,v,k} : u = m(O_i) \wedge v = m(D_i) \wedge k \geq S_{u,v} \wedge k \leq \min(l_i + 1, \max_{w \in V}(S_{u,w})). \tag{5.6}$$

Such selection of path length k guarantees triangle inequality with respect to all path lengths between places for a simple method of construction of the function m: The function m can be found by selecting the first node that corresponds to O_1 uniformly, by random, and then by traversing the TS_u from $i = 1$ to $i = N$ and randomly selecting vertices u and v in corresponding distance k using the pre-computed all-pairs shortest path matrix S.

Step 2—Finding transition cells. A randomly selected path $p^i_{u,v,k}$, such that $u = m(O_i)$, $v = m(D_i)$ and $k = \max(S_{u,v}, l_i + 1)$, is considered to be a sequence of nodes (cells) the user visits between places O_i and D_i during the i-th transition. For the same reasons as above (finding k-length simple path is NP-hard problem) we simplify this task and look for a path $\widetilde{p}^i_{u,v,k} = \{u, \dots, w, \dots, v\}$ with a maximal number of unique nodes and the desired length, i.e., it holds

$$w \in \operatorname*{argmin}_{x \in V}(p_{u,x,S_{u,x}} \cap p_{x,v,S_{x,v}}) \quad \text{and} \quad S_{u,w} + S_{w,v} = k. \tag{5.7}$$

Using the pre-computed all-pairs shortest path matrix S to find the node w is fast, although it can result in a non-simple path.

Step 3—Processing handovers. We express the user-cell association during the i-th transition $(O_i, D_i, t_i, \delta_i, l_i)$ as a sequence

$$A^i = \{(\tau_j, c_j)\}_{j=0}^{l_i+1}, \tag{5.8}$$

where τ_j denotes the time of the change of association to the j-th cell c_j in the path $p^i_{u,v,k} = \{u = c_0, c_1, c_2, ..., c_{k-1}, c_k = v\}$ between O_i and D_i. Assuming that the speed of the user is constant on the whole path, then the user-cell association in time changes proportionally to the distance between the nodes in the path. Further, we assume that a user-cell association change happens on the boundary between cells. Since the shortest distance between nuclei of two adjacent Voronoi cells to their common boundary is equal (by definition of the Voronoi tessellation), the cell-association changes when the user is in the middle of the Delaunay triangulation edge between the cells' nuclei. Thus, the elements of the sequence A^i are as follows:

$$\tau_0 = t_i \tag{5.9}$$

$$\tau_j = t_i + \delta_i \left(\frac{\sum_{l=1}^{j-1} d_l + d_j/2}{\sum_{l=1}^{n} d_l} \right), \tag{5.10}$$

where d_l is the Euclidean distance between nodes c_{l-1} and c_l on the l-th edge (c_{l-1}, c_l).

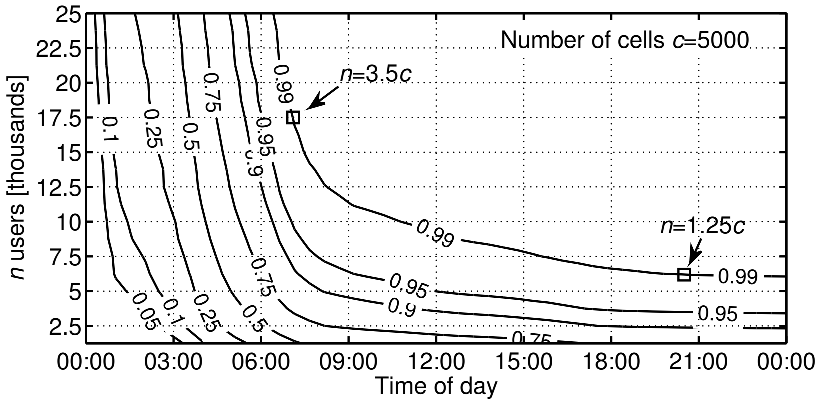


Figure 5.13: Ratio of cells observed during a day.

Step 4—Results. Let C_u denote the set of all cells a user u visited during all transitions in the transition sequence, i.e.,

$$C_u = \{v \in V | \exists i, j, A_u^i = \{(\tau_j, c_j)\} : \tau_j \leq T \wedge v = c_j\}. \tag{5.11}$$

The number of unique cells a user u visits by the time T is the cardinality of the set C_u. A higher number of users in the simulation at one time is handled independently, so the total number of cells visited by n users is simply $|\bigcup_{u=1}^{n} C_u|$.

5.4 Simulation Results and Discussion

Assume we have a mobile network that consists of c cells, there are n users involved in crowdsensing, and the crowdsensing campaign starts at midnight. How long does it take the users to visit $x\%$ of all cells in the network? We have simulated the movement of users in a network with $c = 5{,}000$ cells and number of users $n = \{0.25c, 0.5c, \ldots, 5c\}$.

Figure 5.13 shows a relationship between the ratio of cells observed during a day and the number of users involved in crowdsensing. We can see that at least $n = 1.25c = 6{,}250$ users are needed to observe at least 99% of all cells by the end of the day. Additionally, there is a significant difference between the times to visit all cells as the number of users increases from $n = 1.25c$ to $n = 3.5c$. Because of a low number of users who travel during the early morning, about $n = 3.5c$ users are needed to visit 99% of the cells by 07:00. Markedly, having more than $3.5c$ users yields only minor improvements.

The presented results pose several issues. First, it is questionable whether any third-party can persuade a user-pool of at least three-times the number of cells in the network to participate in crowdsensing. Consider the Czech Republic with 10.5 million inhabitants living in approx. 78,000 km^2 $\cong$ 30,500 sq mi. Each of the three biggest mobile operators has about $c = 14{,}000$ cells in the network [90], so about $n = 3c = 42{,}000$ users are needed to visit 99% of the cells by 8 a.m. Let us assume the sensing software is built on the Android platform, and the smartphone penetration (50%) and Android share on the smartphone market (48%) are similar to the U.S. [153]. Then a calculation ($10{,}500{,}000 \times 1/3 \times 0.5 \times 0.48 = 840{,}000$) shows that every *twentieth* user of an Android smartphone (per each operator) should participate in crowdsensing, making it seem viable.

Other issues are related to the model proposed and the process of simulating user-cell associations.

The model presented in this study may seem limited by the lack of any spatial relation to a real geographic background. However, because it captures user's movement in a mobile network in terms of cell transitions without conditioning on the real-world cell tower locations, the model is area-independent. As such, it is not limited to the area covered by the dataset and can be applied to any arbitrary cellular network topology—either a real one or an artificial one. Nevertheless, the parameters of the model may change with different network technology (GSM, UMTS and LTE) or with a larger and more representative user-pool.

Apparently, working with a real-world mobile network topology would bring more concerns about the simulation settings. First, the number of cells that constitute user's places varies from one to about 10 cells per place, depending o cell jitter in cell-dense areas. The number of cells per place should be correlated with cell density in the studied region and then reflected in the simulation. This change can shorten the time to observe a particular ratio of cells in the network. Second, commuting patterns in the studied region should be considered and mined from the data [84] to obtain more realistic user-places distribution than the random one we use. Finally, any background information on transportation networks such as roads or railways would positively affect not only the distribution of user's places but also the selection of a path between places. Frequent commuting patterns and similar paths taken by users between places would result in smaller number of cells visited by the users. Our approach to the simulation in this work is minimalistic, mainly for the reason that the dataset is too sparse to support any of the simulation enhancements above.

5.5 Case Study: Dealing with the Dynamic Cell-ID Method

In this section we provide detailed analysis of simulation results to demonstrate the ability of crowdsensing to deal with obfuscation of the network topology by a particular mechanism, called the Dynamic Cell-ID method.

Today, user location is delivered for free by third parties (e.g., My Location service by Google) exploiting the fact that the Cell-ID assignment is static and the signal covers a given geographic area. The *Dynamic Cell-ID* mechanism, described in three recent patents [30, 62, 201] and allegedly considered for deployment by China Mobile [141], may alter the situation. The key idea is to mask part of the static Cell Global Identities by providing different, dynamically generated Cell-IDs. A new, *dynamic Cell-ID* is calculated by the base station and is transmitted to the mobile device, while the original, *static Cell-ID* remains intact in the core network (see Figure 5.14). From time to time (patent [201] suggests once a day), all dynamic Cell-IDs are permuted among the network cells. This process is achieved by an unspecified, time-dependent, invertible function that maps static Cell-IDs to the dynamic ones and vice versa. Such mapping function can be arbitrarily complicated, or it may even represent a simple random permutation, so one can assume that the mapping function can not be discovered by simply observing the changes of dynamic Cell-IDs over time. Dynamic Cell-ID thus represents network topology obfuscation and makes geographical conversion extremely hard for an outside party. Only the mobile operator knows (and stores) the present static to dynamic Cell-ID mapping and the GPS coordinates of the cells. With frequent, for example daily changes, third-party Cell-ID databases such as [52, 18] would have difficulty maintaining the correct Cell-ID information, enabling network operators to commercialize the mapping of dynamic Cell-IDs to geographical coordinates.

Deploying Dynamic Cell-ID would have consequences. GPS-less devices, still a majority of all mobile phones (62% in 2011 [32]), rely on network-based (Cell-ID) positioning. Third-parties would fail to provide free localization applications, unless they paid operators for the Dynamic Cell-ID mapping, influencing customer's end price. A-GPS-enabled phones would be affected by having a longer time-to-first-fix (order of minutes), as commercial Secure User Plane Location (SUPL) servers would not be able to advise the A-GPS receiver on the approximate satellite positions (based on the current Cell-ID of the user) because SUPL-server databases would become outdated every time the dynamic

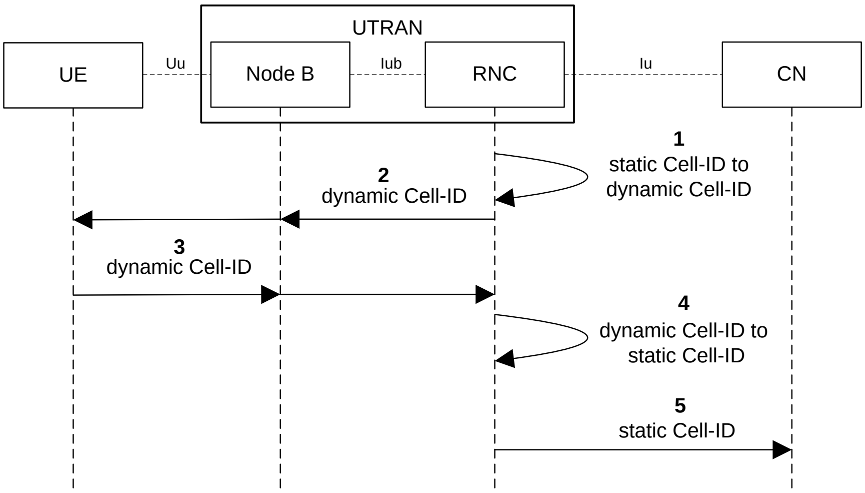


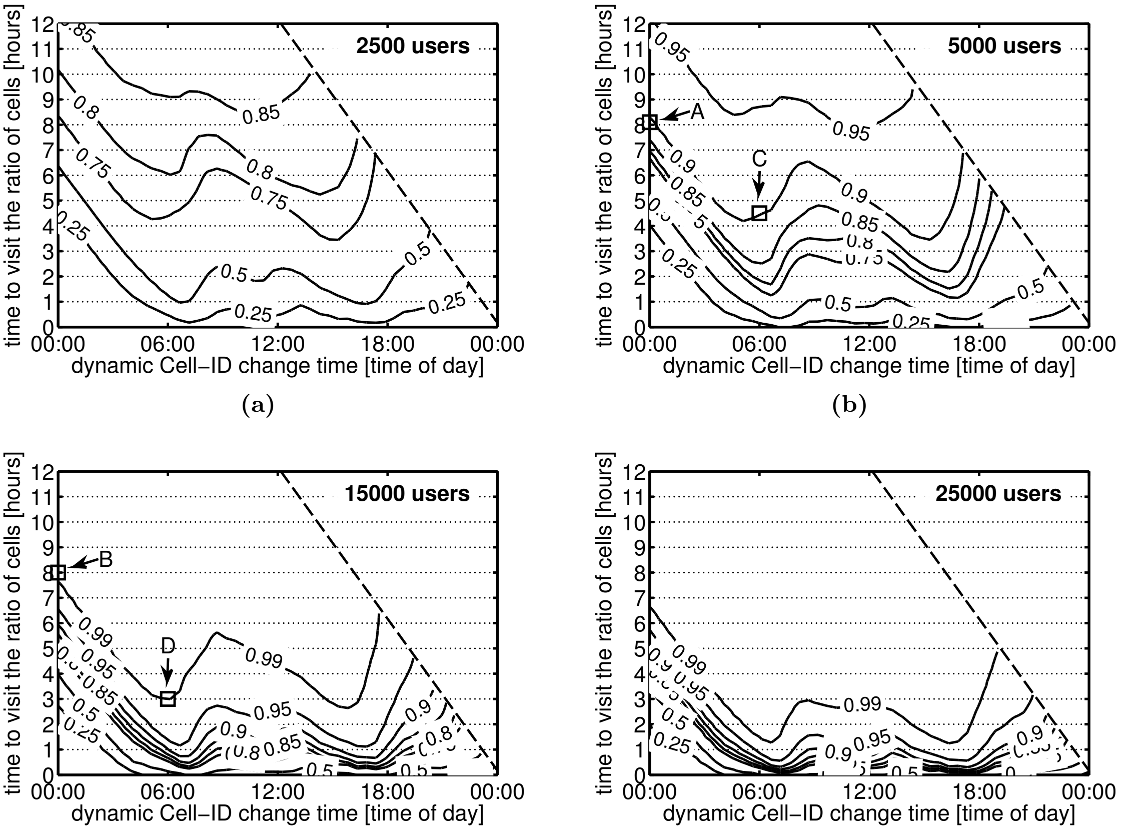
Figure 5.14: Principle of dynamic Cell-ID in an UMTS network according to [62]. The Cell-ID (specifically, Cell-ID, LAC or both these parts of the CGI) are changed at the Radio Network Controller (RNC) level (1). Node B works normally, i.e., it passes the dynamic Cell-ID (instead of the static one) to the User Equipment (UE) (2). The UE communicates with the Node B using the dynamic Cell-ID (3). Finally, in RNC the dynamic Cell-ID is transformed back to its corresponding static value (4). The static Cell-ID is then used within the Core Network (CN) (5).

Cell-IDs are changed. Finally, various cell-fingerprinting methods [200, 159], popular in research and academia as cheap and reliable positioning methods, would be rendered inoperative.

There are two principal ways to deal with Dynamic Cell-ID. Either the mapping may be bought from the mobile operator, which might be costly or the operator might not be willing to sell it. Or, third-parties can assign coordinates to Cell-IDs by wardriving[5] or crowdsensing methods. While wardriving is time-consuming and limited in resources (vehicles, drivers), crowdsensing is advantageous in time and coverage, especially when many mobile users are involved. Nevertheless, two questions arise: How many crowdsensing participants are needed? And, how long does it take them to scan the entire network?

We use the mobility model from Section 5.2 and run the simulation described in Section 5.3 to obtain an estimation of Cell-ID-crowdsensing performance. Figure 5.15 shows the results for $c = \{2{,}500, 5{,}000, 15{,}000, 25{,}000\}$ users in a network that consists of 5,000 cells. By comparing Figures 5.15a–5.15d we see, unsurprisingly, that collecting network Cell-IDs takes a shorter time when more users are involved. However, users' mobility during a day significantly affects the duration of the network scan: in the morning

[5] Wardriving is the act of searching for WiFi hotspots and other information, such as mobile network Cell-IDs, at particular locations by driving around.



(a)

(b)

(c)

(d)

Figure 5.15: Impact of dynamic Cell-ID renumbering time on cell discovery in a network with 5,000 cells. Example A shows that in case the dynamic Cell-ID renumbering occurs at midnight, 5,000 users visit 90% of all cells at 08:00 (after 8 hours from midnight). Higher number of users results in shorter discovery time and more cells discovered: under similar conditions, three-times more users visits in the same period 99% of all cells (example B). Examples C and D show the Cell-ID-crowdsensing performance in the time of the day with high user mobility. If the dynamic Cell-ID renumbering occurs at 06:00, then 5,000 users discovers 90% of all cells in 4 hours (at 10:00) , whereas with 15,000 users 99% of all cells are discovered in 3 hours (at 09:00).

and afternoon users commute and travel more, resulting in a shorter scan time. On the contrary, it takes longer to scan the network during the night and around noon, as mobility of users is low. This may be a clue for *when* operators should renumber the dynamic Cell-IDs to strike at the heart of third-party Cell-ID databases the most.

Crowdsensing as a fight-back method against the Dynamic Cell-ID method is quite a powerful tool, but as our results show it significantly depends on the user-pool size. Apparently, having the network scanned within couple of hours anytime during a day is possible, but with an almost unrealistic number of users. As a result, with Dynamic Cell-ID adopted, a third-party service relying on a crowdsensed Cell-ID database may suffer from bad localization performance for hours-long period.

5.6 Conclusion

In this chapter, we have presented a method of obtaining the limits of crowdsensing in discovering the mapping of mobile network Cell-IDs to geographic locations.

Based on the NRC-Lausanne dataset, we have built a mobility model which describes user-cell associations in a mobile network over a day. Using the model we have generated thousands of artificial yet realistic traces of user movement, which show high similarity with the original dataset in three key features: users' daily patterns (visited places and transitions between them), temporal characteristics of varying human mobility during one day, and quantification of daily user-cell associations in terms of unique cells visited.

We applied the model-based traces of user movement by a large-scale simulation to an approximation of a mobile network topology. The results show that crowdsensing is quite a powerful tool: for example only 25% more users than cells suffices to map 99% cells of a mobile network to geographic locations over a day. With 3 times more users than cells it is possible to map 99% cells in a couple of hours anytime during a day.

The application of crowdsensing in discovering the mapping of mobile network Cell-IDs to geographic locations may be particularly interesting for Cell-ID-based location-services providers. These may want to use the "power of the crowd" to build, maintain and repair their Cell-ID databases—a matter essential to their business.

Chapter 6

Conclusions

In this thesis, we have investigated the principal limits of tracking methods in mobile networks. Tracking data, simply a timestamped history of mobile users' positions in the network, is a sought-after and still scarce source of information for research studies in telecommunications, human and time geography, transportation, urban studies, network design, cloud computing, and other fields of science.

The main problem with tracking methods in mobile network is that their suitability for large-scale tracking, network-wide application, and their technological limitations are often not discussed or remain unknown. Three principal methods of tracking can be considered. *Network-based active tracking* enables to collect data selectively in desired continuity and extent, at the price of possible impact of extra load on network performance. *Network-based passive tracking*, based on Call Data Records, represents a source of user mobility-related data at unprecedented scales, but is limited by communication frequencies of an individual. *Terminal-based tracking* may deliver content-rich data together with positioning information, but it can only be deployed to cooperating users of mobile phones and its network-wide span is in question.

We have addressed the following topics related to the limits of tracking methods in mobile networks: (1) technological limits of network-based active tracking; (2) user mobility-characteristics between communication places and extensions beyond the low temporal granularity of passive tracking data; and (3) the scaling limits of cooperative terminal-based tracking in terms of mapping a mobile-network infrastructure to geographic locations.

6.1 Limits of Network-based Active Tracking

We have studied a particular method of active tracking in a mobile network, the SMS-based one, in a live GSM network. Using our own existing tracking solution (implemented before our Ph.D. studies [63, 66]) we have conducted a tracking measurement and described performance characteristics of the tracking platform and the network nodes.

We have modeled and simulated the tracking process to obtain possible combinations of the number of users to be tracked simultaneously and the corresponding tracking interval. We have shown that even with the minimal interconnection configuration, the baseline implementation is capable of tracking thousands of users periodically on the scale of minutes.

We have estimated the limits of SMS-based active tracking in terms of minimal tracking interval, showing that the minimum time between two consecutive Cell-ID retrievals is limited to an approximate value of 9 seconds in a GSM network. The constraints of the Location Server come primarily from its interconnection to the mobile network, which limits the quantitative performance to about 500 location retrievals per minute. Compared to other state-of-the-art positioning methods, SMS-based Cell-ID positioning performs better then all contemporary methods but the Cell-ID+TA, which however needs a dedicated protocol to work with.

We have examined and discussed the limitations of the core and radio access networks. Whereas core network does not seem limited by the additional tracking load, radio access network can deliver only a limited number of SMSes per each cell over a fixed time interval, on the order of tens of SMSes.

To learn about scalability of the method, we have carried out a detailed calculation of the impact of tracking on the GSM network-infrastructure capacity at different levels—network cells and location areas. The impact on a single cell capacity is significant: the number of tracked users should be only a fraction of all users in the cell. Higher number of tracked users causes a fast rise of GoS on signaling channels, above an acceptable level. Impact on the location area depends on paging capacity of the serving Base Station Controller—at certain level the BSC is the bottleneck because its capacity is exceeded before the limiting value of GoS at location-area cells is reached.

We have shown that mobility of the tracked users represents a significant problem when they meet at the same cell—even after a couple of minutes the cell may be rendered inoperable. A simple solution to prevent such situation, a token bucket mechanism that would spread the positioning requests in time, could spare the signaling capacity of the cell, but since the arriving users would bring additional voice traffic load, the capacity of the traffic channels would become the limiting factor anyway.

Finally, we have demonstrated a practical use-case of active tracking—roaming optimization in mobile networks—and proposed a metric to assess weak cells in terms of roaming traffic. A Gaussian kernel-density estimator with the cell-weakness metrics as a re-weighting function represents a simple yet powerful visualization of weak places in the network.

6.2 Extending Utility of Passive Tracking Data

Because network-based active tracking is still not widely adopted by mobile operators, passive tracking data represent a significant basis of contemporary research in many areas.

To provide new insights into mobility of users, based on the basis of Call Data Records, we have proposed to compare the coarse-grained CDR trajectories with some corresponding ground-truth trajectories representing a continuous trace with user's positions. We have used existing publicly available data, the Reality Mining Dataset, from which we have derived a substitution for both CDRs and a finer trajectory of user-cell associations. To achieve this, we have significantly extended the dataset with spatial coordinates of cell towers by using the Google Location API and our novel LAC-clustering algorithm for spatial outlier detection from GSM cell-tower data.

We have provided a detailed analysis of user inter-call behavior, i.e., user movement characteristics between two consecutive communication locations. We have found three key inter-call mobility attributes: (1) unskewed behavior in spatial dimension, (2) straight course between calls, and (3) staying behavior at call places. We have shown that these findings are in strong contrast with the assumptions of the existing modeling methods, the linear weighted interpolation and the probabilistic extension of space-time prism, and therefore these models cannot be used for precise posterior CDR analysis.

To improve the accuracy of CDR-based deduction of user's presence at call places over time, we have formulated a new probabilistic Inter-Call Mobility (ICM) model, spatio-temporally fitting the aggregated mobility behavior of the Reality Mining Dataset users.

Using a finite Gaussian mixture model, the ICM model approximates the ground-truth of users' inter-call movement: from a single timestamped call coordinate to a probabilistic distribution of user's position between calls. The ICM model is expressed analytically and thus reusable and general enough for practical application to *any* CDR traces. Moreover, the ICM model allows for description of user's position at a particular time in between calls and, vice versa, given geographical coordinates, probability of user's presence at a particular position over time can be derived.

Finally, we have shown the ICM model applicability on the example of user proximity probability. This is largely applicable for example in studies of virus spreading [197].

6.3 Exploring Limits of Crowdsensing

We have studied the limits of terminal-based tracking from the point of view of its potential for mapping of mobile-network cells to geographic locations.

Since there exists no mobility model which would count user-cell associations over a time period, we have proposed a trace-based mobility model to quantify the users' ability to detect a number of cells in the network. We have used the NRC-Lausanne dataset to extract the information about users' places, i.e., significant locations in terms of time spent at the corresponding cells, and the transitions between places during one day. We have quantified (1) the total number and ordering of places as they are visited by a user during a day, (2) the start time of all user's transitions between places in the day, and (3) the duration of transitions and their length, measured in the number of unique cells visited during the transition. The trace-drive model uses the statistics above to generate artificial traces, in which the following features are similar to the original dataset: users' daily patterns (visited places and transitions between them), temporal characteristics of varying human mobility during one day, and quantification of daily user-cell associations in terms of unique cells visited.

Using a large-scale simulation, we have applied the model-based traces of user movement to an approximation of a mobile network topology. By this we have determined the fraction of mobile-network cells visited by a fixed number of users over a time interval. Under the assumption that the user's mobile phone reports its GPS coordinates and the Cell-ID of a serving cell, the results show that crowdsensing is quite a powerful tool: for example with only 25% more users than cells sufficing to map 99% cells of a mobile network to geographic locations over a day. With 3 times more users than cells it is possible to map 99% cells in a couple of hours anytime during a day.

We have discussed the applicability of crowdsensing as a fight-back method against obfuscation of the network topology by the Dynamic Cell-ID method, which is based on periodic changes of Cell-IDs in the network to prevent unauthorized geo-location services offered by third-parties. We have found that users' mobility during a day significantly affects the duration of the network scan: in the morning and afternoon users commute and travel more, resulting in a shorter scan time. On the contrary, it takes longer to scan the network during the night and around noon, as mobility of users is low. Our study on crowdsensing limits may be particularly interesting for contemporary and new Cell-ID-based location-services providers. At the same time, it can serve to the mobile network providers as a hint whether they should try to monetize their costly network infrastructure and apply the network-infrastructure obfuscating mechanisms.

6.4 Future Research Directions and Open Issues

The SMS-based active tracking method, examined in this thesis, does not scale well in GSM networks because their per-transceiver limit of paging requests and deliverable SMSes cannot be increased beyond a relatively low technology-specific value. Strategies to prevent overloading of signaling channels exist [71], but yield only a minor improvement in terms of cell capacity, and represent a trade-off between the count of communication and traffic channels. A prediction of future user-cell association such as [64] could be used to adjust the rate of positioning requests into the high-loaded cells, thus making active tracking less demanding for the network. Contemporary mobile network such as UMTS and LTE perform better with the SMS-based active tracking, mainly because radio-access technology is more efficient. In UMTS networks, the Wideband Code Division Multiple Access (W-CDMA) radio access technology is employed to better utilize the frequency spectrum and yield higher bandwidth. A complex study that analytically estimates the maximum number of simultaneous signaling services, such as SMS or Paging, is available in [166]. The authors calculate that more than 700 text messages can be delivered in one second to one cell (42,000 SMS/min), which is about 230 times more than in GSM networks. LTE networks, with their data-oriented architecture based on IP Multimedia Subsystem (IMS) [43], are expected to be deployed in a transition scenario where Circuit-

Switched (CS) legacy networks live side by side with LTE, enabling the provisioning of Voice and SMS services through reuse of the legacy networks, called CS-fallback [8]. In this case, the SMS-based active tracking limits are the same as for GSM networks. However, future LTE-enabled terminals will be mainly data-oriented and will eventually use the voice and SMS in a fully *all-IP* manner over the SIP protocol [134], making the SMS-based active tracking antiquated by less demanding alternatives [87]. Apparently, the trend in positioning towards convergence of radio access technologies and hybrid methods seems inevitable [73], but their complexity is in direct contradiction to our approach to active tracking. We believe that SMS-based active tracking with Cell-ID positioning remains important for two reasons. First, although being substituted with evolving networks, GSM covers more than 85% of world's population and still is the most-utilized technology in the majority of countries. And second, in terms of revenue, 9 out of the top 20 telecommunication markets are in developing countries [111], thus a cheap and easy-to-deploy solution may be in their interest.

In our analysis of passive-tracking data, we have considered only calls and text messages to be the communication events in Call Data Records. Similar methodology to express the inter-call mobility could be applied to CDRs with data sessions, which are finer-grained in the temporal dimension. However, because most of the subsequent events would happen at the same cell, a GPS track would serve better then a user-cell associations we used. A related aspect to study is to overcome the spatial limitation of the ICM model. In this thesis we have considered places of communication events at least three kilometers apart, mainly because of the limited accuracy of the Cell-ID-based movement description. This constraint may be removed when a more accurate movement trace is available. Finally, since the ICM model gives out *one* position of a user between places of its communication activity, other approaches based on fitting the CDR-based call trajectory with different interpolation methods could deliver a continuous, approximate trajectory. A strong enhancement of our model would be the application of background information, such as road networks, similar to the work on space-time prisms using uncertainty [123]. The spectrum of potential applications of the ICM model is not limited to the proximity probability of two users, as demonstrated in this work, but can be useful for any a-posteriori interpretation of CDRs.

Future extension of our study on the limits of crowdsensing may relax the assumptions posed in our work and utilize the information about cells surrounding the actual serving cell during the crowdsensing process. This may have a strong effect on estimating the actual accuracy of the mapping of Cell-IDs to geographic locations, in terms of geographical distance, which we were not focused on in this work. In addition to that, using a different, state-of-the-art mobility model with an overlaying mobile-network topology may be particularly interesting to compare with our results. However, a mobility model which would consider different scales (urban, rural areas), mobility properties (visiting time, return time) and temporal and spatial mobility patterns concurrently, is still awaited to be discovered. In this work, we did not aim to make our trace-based model parsimonious—an elegant way to overcome this has been recently proposed in [179] by describing motifs of human movement and expressing only user's probability of not being at a "home" place. Finally, applying any relevant background knowledge, such as population density, transportation networks (roads, railways) and commuting patterns in the studied region would bring the simulation results even closer to reality.

Open issues in the area of user tracking in mobile networks are related to the utilization of tracking data. The spectrum of potential applications, based on the tracking data, is wide: from network energy-efficient performance, roaming-customer retention, to the design of content-distribution strategies. Network-based active tracking may acquire better insight in network migration in the geographical, commercial or technological sense. Emerging virtual operators may benefit from the knowledge about users' preferences in visiting different access networks over time, roaming in different countries or selecting particular network technology (UMTS, LTE, WiFi). Self-adaptive pricing policy, based on network usage and predicted user-behavior may be interesting for all network providers.

6.5 Research Contributions

The topics covered in this thesis were published in several conference and journal papers. The details of our original research contributions and the relevant publication record are as follows.

Chapter 3: Active Tracking in Mobile Networks

- A model describing SMS-based active tracking in mobile networks has been presented, its parameters come from a large-scale measurement on a real tracking platform, connected to a live mobile network. Limits of the implementation have been simulated using a discrete-time simulation of the tracking process.

- Limits of SMS-based active tracking have been deduced and described: basic constraints of the method, constraints of the location server and of the mobile network. Principal limitations of the GSM radio access network have been derived.
- A metric of cell-weakness in terms of roaming traffic has been proposed and used for roaming optimization with active tracking.

Works related to these results are:

- Ficek, M.; Pop, T. & Kencl, L. Active Tracking in Mobile Networks: An In-depth View. *The International Journal of Computer and Telecommunications Networking*, Elsevier, 2013, *in press*. [Online]. Available: `<http://www.sciencedirect.com/science/article/pii/S1389128613000996>`
- Ficek, M.; Pop, T.; Vláčil, P.; Dufková, K.; Kencl, L. & Tomek, M. Performance Study of Active Tracking in a Cellular Network Using a Modular Signaling Platform. In *Proceedings of the 8th international conference on Mobile systems, applications, and services (MobiSys '10)*, ACM, 2010, pages 239–254.
- Ficek, M. & Kencl, L. Improving roamer retention by exposing weak locations in GSM networks. In *Co-Next Student Workshop '09: Proceedings of the 5th international student workshop on Emerging networking experiments and technologies*, ACM, 2009, pages 17–18.

Chapter 4: Extending Utility of Passive Tracking Data

- A heuristic approach to spatial outlier detection from GSM mobility data has been presented. The existing dataset has been spatially extended to enable a comparative study between trajectories from Call Data Records and from user-cell associations.
- Detailed analysis of users' mobility behavior in between their communication places (calls, text messages) has been presented. It shows that the nature of human mobility between communication events is in strong contrast with the assumptions of the existing modeling methods.
- A probabilistic spatio-temporal refinement of Call Data Records, the *Inter-Call Mobility* (ICM) model, has been presented. The model is analytically tractable and can be used for practical application to any CDR traces to describe the spatio-temporal mobility behavior of mobile-network users between their communication events.

Works related to these results are:

- Ficek, M. & Kencl, L. Inter-Call Mobility Model: A Spatio-temporal Refinement of Call Data Records Using a Gaussian Mixture Model. In *Proceedings of IEEE INFOCOM 2012*, IEEE, 2012, pages 469–477.
- Ficek, M. & Kencl, L. Spatial extension of the Reality Mining Dataset. In *Mobile Adhoc and Sensor Systems (MASS), 2010 IEEE 7th International Conference on*, IEEE, 2010, pages 666–673.

Chapter 5: Exploring the Limits of Crowdsensing

- A data-driven mobility model has been proposed to express the number of unique mobile-network cells a user is capable of visiting during one day. The model describes users' daily patterns, captures fine-grained temporal characteristics of human movement during a day, and quantifies daily user-cell associations.
- A large-scale simulation of users' movement in an approximation of a mobile network has been conducted to assess the coverage capabilities of a user pool for the time period in an area.
- Limits of crowdsensing in discovering the mapping of mobile-network cell identifiers to geographic locations as a method against mobile-network-topology obfuscation has been presented.

Works related to these results are:

- Ficek, M.; Clark, N. & Kencl, L. Can Crowdsensing Beat Dynamic Cell-ID? In *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones (SenSys '12)*, ACM, 2012, pages 10:1–10:5.

Bibliography

[1] 3GPP, "TS 03.071: Location services (LCS); Functional description; Stage 2." 2004, v7.11.0.

[2] 3GPP, "TS 30.90: Unstructured Supplementary Service Data (USSD) - Stage 2," 1999, v7.0.0.

[3] 3GPP, "TS 05.10: Radio subsystem synchronization," 1999, v8.12.0.

[4] 3GPP, "TS 09.78: Customised Applications for Mobile network Enhanced Logic (CAMEL); CAMEL Application Part (CAP) specification," 2000, v7.1.0.

[5] 3GPP, "TS 22.071:Location services (LCS); Service description; Stage 1." 2011, v10.0.0.

[6] 3GPP, "TS 23.038: Alphabets and language-specific information," 2008, v7.0.0.

[7] 3GPP, "TS 23.271: Functional stage 2 description of Location Services (LCS)," 2010, v10.0.1.

[8] 3GPP, "TS 23.272: Circuit Switched Fallback in Evolved Packet System," 2008, v8.0.0.

[9] 3GPP, "TS 24.011: Point-to-point (PP) short message service (SMS) support on mobile radio interface," 2011, v10.0.0.

[10] 3GPP, "TS 29.002: Mobile Application Part (MAP) specification," 2002, v3.13.0.

[11] 3GPP, "TS 36.304: Evolved universal terrestrial radio access (E-UTRA); User equipment (UE) procedures in idle mode," 2011, v10.3.0.

[12] 3GPP, "TS 36.355 : LTE; Evolved universal terrestrial radio access (E-UTRA); LTE positioning protocol (LPP)," 2011, v9.4.0.

[13] 3GPP, "TS 43.059: Functional stage 2 description of Location Services (LCS) in GERAN," 2007, v7.3.0.

[14] 3GPP, "TS 44.035: Broadcast network assistance for Enhanced Observed Time Difference (E-OTD) and Global positioning system (GPS) positioning methods," 2009, v9.0.0.

[15] 3GPP, "TS 45.010: Radio access network; Radio subsystem synchronization," 2011, v9.0.0.

[16] 3GPP, "TS 45.811: Feasibility study on Uplink TDOA in GSM and GPRS," 2002, v6.0.0.

[17] 3GPP, "TS 49.031: Location services (LCS); BSSAP-LE," 2011, v10.0.0.

[18] 8motions, "OpenCellId," 2010. [Online]. Available: <http://www.opencellid.org/> [Accessed: 2013.05.02]

[19] Acoustic, "Learn how to find GPS location on any smartphone, and then make it relevant," 2008. [Online]. Available: <http://www.codeproject.com/KB/windows/DeepCast.aspx> [Accessed: 2011.08.03]

[20] R. Ahas, A. Aasa, S. Silm *et al.*, "Mobile positioning in space-time behaviour studies: Social positioning method experiments in estonia," *Cartography and Geographic Information Science*, vol. 34, no. 34, pp. 259–273, 2007.

[21] R. Ahas, A. Aasa, A. Roose *et al.*, "Evaluating passive mobile positioning data for tourism surveys: An Estonian case study," *Tourism Management*, vol. 29, no. 3, pp. 469–486, 2008.

[22] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[23] M. Allamanis, S. Scellato, and C. Mascolo, "Evolution of a location-based online social network: Analysis and models," in *Proc. of IMC*. ACM, 2012, pp. 145–158.

[24] N. Alon, C. Avin, M. Koucky *et al.*, "Many random walks are faster than one," in *Proc. of SPAA*. ACM, 2008, pp. 119–128.

[25] N. Alon *et al.*, "Color-coding," *Journal of the ACM*, vol. 42, no. 4, pp. 844–856, 1995.

[26] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 450–461, 2007.

[27] F. Aurenhammer, "Voronoi diagrams - a survey of a fundamental geometric data structure," *ACM Computing Surveys*, vol. 23, no. 3, pp. 345–405, 1991.

[28] F. Bai, N. Sadagopan, and A. Helmy, "IMPORTANT: A framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks," in *Proc. of INFOCOM*. IEEE, 2003, pp. 825–835.

[29] Behavio, "funf: Open Sensing Platform," 2010. [Online]. Available: <http://funf.org/> [Accessed: 2013.03.19]

[30] M. Bensimon, E. Lebomin, C. Giraud-sauveur *et al.*, "Method to mask cell identifiers or location area codes of a mobile network with respect to a mobile terminal," U.S. Patent 20 090 170 477, 2009.

[31] A. Berg, "Method for providing location information," U.S. Patent 20 020 160 789, 2002.

[32] Berg Insights, "Handset Connectivity Technologies 3rd Ed." 2012. [Online]. Available: <http://www.berginsight.com> [Accessed: 2013.03.18]

[33] K. Biermann, "Betrayed by our own data," Zeit Online, 2011. [Online]. Available: <http://www.zeit.de/digital/datenschutz/2011-03/data-protection-malte-spitz> [Accessed: 2013.04.07]

[34] V. D. Blondel, M. Esch, C. Chan *et al.*, "Data for Development: the D4D Challenge on Mobile Phone Data," *ACM Computing Research Repository*, vol. abs/1210.0137, pp. 1–10, 2012.

[35] Broadcom Corporation, "AGPS Server and Worldwide Reference Network," 2007.

[36] J. Burke, D. Estrin, M. Hansen *et al.*, "Participatory sensing," in *Proc. of WSW.* ACM, 2006, pp. 117–134.

[37] K. Burnham and D. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*, 2nd ed. Springer, 2002, ISBN 0387953647.

[38] J. P. Buzen, "Fundamental laws of computer system performance," in *Proc. of SIGMETRICS.* ACM, 1976, pp. 200–210.

[39] N. Caceres, J. P. Wideberg, and F. G. Benitez, "Deriving origin destination data from a mobile phone network," *IET Intelligent Transport Systems*, vol. 1, no. 1, pp. 15–26, 2007.

[40] F. Calabrese, F. C. Pereira, G. Di Lorenzo *et al.*, "The geography of taste: analyzing cell-phone mobility and social events," in *Proc. of PerComp.* Springer-Verlag, 2010, pp. 22–37. [Online]. Available: `<http://dx.doi.org/10.1007/978-3-642-12654-3_2>` [Accessed: 2013.04.28]

[41] California Center for Innovative Transportation, "Mobile millenium," 2013. [Online]. Available: `<http://traffic.berkeley.edu/>` [Accessed: 2013.04.30]

[42] S. Calinon, F. Guenter, and A. Billard, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 2, pp. 286–298, 2007.

[43] G. Camarillo and M.-A. Garcia-Martin, *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds, Second Edition.* John Wiley & Sons, 2006, ISBN 0470018186.

[44] A. Chao, "Non-parametric estimation of the number of classes in a population," *Scandinavian Journal of Statistics*, vol. 11, pp. 265–270, 1984.

[45] A. Chao and S.-M. Lee, "Estimating the number of classes via sample coverage," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 210–217, 1992.

[46] A. Chao, R. K. Colwell, C.-W. Lin *et al.*, "Sufficient sampling for asymptotic minimum species richness estimators," *Ecology*, vol. 90, no. 4, pp. 1125–1133, 2009.

[47] M. Y. Chen, T. Sohn, D. Chmelev *et al.*, "Practical metropolitan-scale positioning for GSM phones," in *Proc. of Ubicomp.* Springer-Verlag, 2006, pp. 225–242.

[48] Y. Chon, H. Shin, E. Talipov *et al.*, "Evaluating mobility models for temporal prediction with high-granularity mobility data," in *Proc. of PerCom.* IEEE, 2012, pp. 206–212.

[49] M. Chu, "New magical blue circle on your map," 2007. [Online]. Available: `<http://googlemobile.blogspot.com/2007/11/new-magical-blue-circle-on-your-map.html>` [Accessed: 2011.06.15]

[50] A. Chukarin, N. Pershakov, and K. Samouylov, “Performance of Sigtran-based signaling links deployed in mobile networks,” in *Proc. of ConTel*. IEEE, 2007, pp. 163–166.

[51] R. Colwell and J. Coddington, “Estimating terrestrial biodiversity through extrapolation,” *Philosophical Transaction of the Royal Society of London B*, vol. 345, pp. 101–118, 1994.

[52] Combain Mobile AB, “Location-API,” 2009. [Online]. Available: `<http://location-api.com/>` [Accessed: 2011.06.15]

[53] I. Constandache, R. Choudhury, and I. Rhee, “Towards mobile phone localization without war-driving,” in *Proc. of INFOCOM*. IEEE, 2010, pp. 1–9.

[54] C. Cooper, A. Frieze, and T. Radzik, “Multiple random walks in random regular graphs,” *SIAM Journal on Discrete Mathematics*, vol. 23, no. 4, pp. 1738–1761, November 2009.

[55] N. Deligiannis, S. Louvros, and S. Kotsopoulos, “Mobile positioning based on existing signalling messages in GSM networks,” 2007. [Online]. Available: `<http://sfhmmy.ntua.gr/2007/papers/paper10.pdf>` [Accessed: 2013.04.07]

[56] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[57] Dialogic, “Adding location-based services to existing architectures,” 2007.

[58] Dialogic, “SPCI4 user manual,” 2007.

[59] Dialogic, “Dialogic DSI SS7 protocol stack MAP programmer’s manual,” 2009.

[60] Dialogic, “Dialogic DSI signaling interface unit based on dialogic DSI SS7G3x signaling servers,” 2009.

[61] J. A. Downs, M. W. Horner, and A. D. Tucker, “Time-geographic density estimation for home range analysis.” *Annals of GIS*, vol. 17, no. 3, pp. 163–171, 2011.

[62] Z. Duan, X. Ren, and J. Qu, “Method and device for configuring a cell identity,” Chinese Patent WO/2009/140 914, 2009.

[63] K. Dufková, J. Danihelka, M. Ficek *et al.*, “Can active tracking of inroamer location optimise a live GSM network?” in *Proc. of CoNEXT Student Workshop*. ACM, 2007, pp. 1–2.

[64] K. Dufkova, J.-Y. L. Boudec, L. Kencl *et al.*, “Predicting user-cell association in cellular networks from tracked data,” in *Proc. of MELT*. Springer-Verlag, 2009, pp. 19–33.

[65] K. Dufkova, M. Bjelica, B. Moon *et al.*, “Energy savings for cellular network with evaluation of impact on data traffic performance,” in *Proc. of EW*, 2010, pp. 1–9.

[66] K. Dufková, M. Ficek, L. Kencl *et al.*, “Active GSM cell-id tracking: ”Where did you disappear?”,” in *Proc. of MELT*. ACM, 2008, pp. 7–12.

[67] P. Dutta, P. M. Aoki, N. Kumar *et al.*, “Common sense: participatory urban sensing using a network of handheld air quality monitors,” in *Proc. of SenSys.* ACM, 2009, pp. 349–350.

[68] N. Eagle, A. S. Pentland, and D. Lazer, “Inferring friendship network structure by using mobile phone data,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15 274–15 278, 2009.

[69] S. B. Eisenman, E. Miluzzo, N. D. Lane *et al.*, “The bikenet mobile sensing system for cyclist experience mapping,” in *Proc. of SenSys.* ACM, 2007, pp. 87–101.

[70] Ericsson, “Location area dimensioning guideline,” 2002, 19/10056-HSC 103 12/4 Rev C 2005-06-09.

[71] Ericsson, “SDCCH dimensioning guideline the Ericsson GSM system,” 2005, 14/100 56-HSC 103 12/4 Uen B2 2005-08-23.

[72] Ericsson, “Mobile Positioning System,” 2007. [Online]. Available: <http://www.ericsson.com/ourportfolio/products/mobile-positioning-system?nav=fgb_101_743|fgb_101_746> [Accessed: 2013.04.02]

[73] Ericsson, “Positioning with LTE – Maximizing performance through integrated solutions,” 2011.

[74] Ericsson, “Ericsson mobility report,” 2011. [Online]. Available: <http://www.ericsson.com/res/docs/2012/ericsson-mobility-report-november-2012.pdf> [Accessed: 2013.04.09]

[75] U. Feige, “A tight lower bound on the cover time for random walks on graphs,” *Random Structures & Algorithms*, vol. 6, no. 4, pp. 433–438, 1995.

[76] M. Ferrante and N. Frigo, “A note on the coupon - collector’s problem with multiple arrivals and the random sampling,” *ArXiv e-prints*, vol. 1209, no. 2667, 2012.

[77] M. Ficek and L. Kencl, “Spatial extension of the Reality Mining Dataset,” in *Proc. of MASS*, 2010, pp. 666–673.

[78] M. Ficek, “CRAWDAD data set ctu/personal (v. 2012-01-11),” 2012. [Online]. Available: <http://crawdad.org/ctu/personal> [Accessed: 2013.04.07]

[79] M. Ficek and L. Kencl, “Improving roamer retention by exposing weak locations in GSM networks,” in *Proc. of CoNEXT Student Workshop.* ACM, 2009, pp. 17–18.

[80] M. Ficek, T. Pop, P. Vláčil *et al.*, “Performance study of active tracking in a cellular network using a modular signaling platform,” in *Proc. of MobiSys.* ACM, 2010, pp. 239–254.

[81] A. Fischer, “CellTrack,” 2006. [Online]. Available: <http://www.afischer-online.de/sos/celltrack/> [Accessed: 2011.06.15]

[82] P. Flajolet, D. Gardy, and L. Thimonier, “Birthday paradox, coupon collectors, caching algorithms and self-organizing search,” *Discrete Applied Mathematics*, vol. 39, no. 3, pp. 207–229, 1992.

[83] Foursquare Labs, Inc., “Foursquare,” 2012. [Online]. Available: <https://foursquare.com/> [Accessed: 2013.04.02]

[84] V. Frias-Martinez, C. Soguero, and E. Frias-Martinez, "Estimation of urban commuting patterns using cellphone network data," in *Proc. of SIGKDD UrbComp Workshop.* ACM, 2012, pp. 9–16.

[85] R. Ganti *et al.*, "Mobile Crowdsensing: Current State and Future Challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[86] L. Gao, M. Li, A. Bonti *et al.*, "Multi-dimensional routing protocol in human associated delay-tolerant networks," *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1–1, 2012.

[87] C. Gessner and O. Gerlach, "Voice and SMS in LTE," 2011.

[88] M. C. González, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[89] Google Inc., "Google Latitude," 2010. [Online]. Available: `<http://www.google.com/latitude/>` [Accessed: 2011.06.15]

[90] GSMweb, "GSMweb," 2010. [Online]. Available: `<http://gsmweb.cz/>` [Accessed: 2013.03.19]

[91] Y. J. Guo, *Advances in Mobile Radio Access Networks.* Artech House, Inc., 2004, ISBN 1580537278.

[92] R. H. Güting and M. Schneider, *Moving Objects Databases.* Morgan Kaufmann, 2005, ISBN 0-12-088799-1.

[93] P. J. Haas, J. F. Naughton, S. Seshadri *et al.*, "Sampling-based estimation of the number of distinct values of an attribute," in *Proc. of VLDB.* Morgan Kaufmann Publishers Inc., 1995, pp. 311–322.

[94] T. Hägerstrand, "What about people in regional science?" *Papers in Regional Science*, vol. 24, no. 1, pp. 6–21, 1970.

[95] E. Halepovic and C. Williamson, "Characterizing and modeling user mobility in a cellular data network," in *Proc. of PE-WASUN.* ACM, 2005, pp. 71–78.

[96] G. Heine and M. Horrer, *GSM Networks: Protocols, Terminology, and Implementation.* Artech House, 1999, ISBN 0890064717.

[97] A. Heinemann, J. Kangasharju, and M. Muehlhaeuser, "Opportunistic data dissemination using real-world user mobility traces," in *Proc. of AINAW.* IEEE, 2008, pp. 1715–1720.

[98] N. Henze, "A multivariate two-sample test based on the number of nearest neighbor type coincidences," *The Annals of Statistics*, vol. 16, no. 2, pp. 772–783, 1988.

[99] H. Hohwald, E. Frías-Martínez, and N. Oliver, "User modeling for telecommunication applications: Experiences and practical implications," in *Proc. of UMAP.* Springer-Verlag, 2010, pp. 327–338.

[100] HP, "HP OpenCall Universal Signaling Platform data sheet," 2006. [Online]. Available: `<http://h20208.www2.hp.com/opencall/library/products/signal/universal_signaling_platform_ds.pdf>` [Accessed: 2013.04.04]

[101] W.-J. Hsu, T. Spyropoulos, K. Psounis *et al.*, "Modeling spatial and temporal dependencies of user mobility in wireless mobile networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1564–1577, 2009.

[102] R. Huerta and L. S. Tsimring, "Contact tracing and epidemics control in social networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 66, no. 5, p. 056115, 2002.

[103] B. Hull, V. Bychkovsky, Y. Zhang *et al.*, "Cartel: a distributed mobile sensor computing system," in *Proc. of SenSys.* ACM, 2006, pp. 125–138.

[104] E. Hyytia, P. Lassila, and J. Virtamo, "A markovian waypoint mobility model with application to hotspot modeling," in *Proc. of ICC.* IEEE, 2006, pp. 979–986.

[105] Intel, "Modular, multiservices carrier platform: Improve application flexibility and lower operating costs," 2004.

[106] Intersec, "Multi channel gateway - next generation SMS-C," 2012. [Online]. Available: `<http://intersec-group.net/frsms-center_23.php?PHPSESSID=4346eacaa64ae9d4b272d288ef870e3b>` [Accessed: 2013.04.02]

[107] S. Isaacman, R. Becker, R. Cáceres *et al.*, "Identifying important places in people's lives from cellular network data," in *Pervasive Computing.* Springer, 2011, vol. 6696, ch. 9, pp. 133–151.

[108] S. Isaacman, R. Becker, R. Cáceres *et al.*, "Human mobility modeling at metropolitan scales," in *Proc. of MobiSys.* ACM, 2012, pp. 239–252.

[109] ITU-T, "Introduction to CCITT Signalling System No.7," ITU-T Recommendation Q.700, 1993.

[110] ITU-T, "Methods for dimensioning resources in Signalling System No.7 networks," ITU-T Recommendation E.733, 1998.

[111] ITU-T, "Measuring the information society 2012," 2012. [Online]. Available: `<http://www.itu.int/ITU-D/ict/publications/idi/material/2012/MIS2012_without_Annex_4.pdf>` [Accessed: 2013.04.09]

[112] A. Jagoe, *Mobile Location Services: The Definitive Guide.* Prentice Hall, 2003, ISBN 978-0-13-008456-9.

[113] A. Jardosh, E. M. Belding-Royer, K. C. Almeroth *et al.*, "Towards realistic mobility models for mobile ad hoc networks," in *Proc. of MobiCom.* ACM, 2003, pp. 217–229.

[114] E. D. Kaplan and C. Hegarty, *Understanding GPS: Principles and Applications*, 2nd ed. Artech House Publishers, 2005, ISBN 1580538940.

[115] D. Karamshuk *et al.*, "Human Mobility Models for Opportunistic Networks," *IEEE Communications Magazine*, vol. 49, no. 12, pp. 157–165, 2011.

[116] F. Kattan, "Dynamic cell-ID: Clever way to block Google, but will it backfire?" 2010. [Online]. Available: `<http://franciscokattan.com/2010/02/06/dynamic-cell-id-clever-way-to-block-google-but-will-it-backfire/>` [Accessed: 2011.08.08]

[117] S. Keshav, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the Telephone Network.* Addison-Wesley, 1997, ISBN 0201634422.

[118] K.-H. Kim, A. Min, D. Gupta *et al.*, “Improving energy efficiency of Wi-Fi sensing on smartphones,” in *Proc. of INFOCOM.* IEEE, 2011, pp. 2930–2938.

[119] M. Kim, D. Kotz, and S. Kim, “Extracting a mobility model from real user traces,” in *Proc. of INFOCOM.* IEEE, 2006, pp. 1–13.

[120] Z. Koppanyi, T. Lovas, A. Barsi *et al.*, “Tracking vehicle in gsm network to support intelligent transportation systems,” in *Proc. of ISPRS*, 2012, pp. XXXIX–B2:139–144.

[121] D. Kotz and T. Henderson, “Crawdad,” 2010. [Online]. Available: `<http://www.crawdad.org/index.php>` [Accessed: 2013.01.07]

[122] B. Kuijpers and W. Othman, “Trajectory databases: Data models, uncertainty and complete query languages,” *Journal of Computer and System Sciences*, vol. 76, no. 7, pp. 538–560, 2010.

[123] B. Kuijpers, B. Moelans, W. Othman *et al.*, “Analyzing trajectories using uncertainty and background information,” in *Proc. of SSTD.* Springer-Verlag, 2009.

[124] A. Kupper, *Location-Based Services: Fundamentals and Operation.* Wiley, 2005, ISBN 0470092319.

[125] K. Laasonen, “Mining cell transition data,” Ph.D. dissertation, University of Helsinki, Finland, 2009.

[126] A. LaMarca, Y. Chawathe, S. Consolvo *et al.*, “Place Lab: Device Positioning Using Radio Beacons in the Wild,” in *Proc. of Pervasive.* Springer-Verlag, 2005, pp. 116–133.

[127] N. Lane, E. Miluzzo, H. Lu *et al.*, “A survey of mobile phone sensing,” *IEEE Communications Magazine*, vol. 48, no. 9, pp. 140–150, 2010.

[128] J. E. Larsen and K. Jensen, “Mobile context toolbox: an extensible context framework for S60 mobile phones,” in *Proc. of EuroSSC.* Springer-Verlag, 2009, pp. 193–206.

[129] J. K. Laurila, D. Gatica-Perez, I. Aad *et al.*, “The Mobile Data Challenge: Big Data for Mobile Computing Research,” in *Proc. of Pervasive.* Springer-Verlag, 2012.

[130] J.-Y. Le Boudec and M. Vojnovic, “The random trip model: Stability, stationary regime, and perfect simulation,” *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1153–1166, 2006.

[131] K. Lee, S. Hong, S. J. Kim *et al.*, “SLAW: A mobility model for human walks,” in *Proc. of INFOCOM.* IEEE, 2009, pp. 855–863.

[132] W.-M. Lee, “Location-based services using Cellid in Windows Mobile,” 2008. [Online]. Available: `<http://www.devx.com/wireless/Article/39709>` [Accessed: 2011.08.03]

[133] Y. Lee, Y. Ju, C. Min *et al.*, “Comon: cooperative ambience monitoring platform with continuity and benefit awareness,” in *Proceedings of the 10th international conference on Mobile systems, applications, and services (MobiSys’12).* ACM, 2012, pp. 43–56.

[134] F. Leitao, S. Freire, and S. Lima, “Sms over lte: Interoperability between legacy and next generation networks,” in *Proc. of ISCC.* IEEE, 2010, pp. 634–639.

[135] J. J. Li and B. Faltings, “Incentive Schemes for Community Sensing,” in *Proc. of CompSust*, 2012.

[136] K. Li and T. C. Du, “Building a targeted mobile advertising system for location-based services,” *Decision Support Systems*, vol. 54, no. 1, pp. 1–8, 2012.

[137] B. Liang and Z. J. Haas, “Predictive distance-based mobility management for multi-dimensional PCS networks,” *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 718–732, 2003.

[138] H. W. Lilliefors, “On the kolmogorov-smirnov test for normality with mean and variance unknown,” *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[139] S. Lim, C. Yu, and C. Das, “Clustered mobility model for scale-free wireless networks,” in *Proc. of LCN.* IEEE, 2006, pp. 231–238.

[140] K. Lin, A. Kansal, D. Lymberopoulos *et al.*, “Energy-accuracy trade-off for continuous mobile device location,” in *Proc. of MobiSys.* ACM, 2010, pp. 285–298.

[141] G. Lu, “China Mobile Is To Implement Dynamic Cell-ID, LBS Services Need Pay for the Location,” 2011. [Online]. Available: <http://technode.com/2011/01/06/china-mobile-to-implement-dynamic-cell-id/> [Accessed: 2013.03.18]

[142] H. Lu, J. Yang, Z. Liu *et al.*, “The jigsaw continuous sensing engine for mobile phone applications,” in *Proc. of SenSys.* ACM, 2010, pp. 71–84.

[143] B. Ly, “Mobile Data Challenge 2012: Unlocking the secrets of smartphone data,” Conversations by Nokia, 2012. [Online]. Available: <http://conversations.nokia.com/2012/06/20/mobile-data-challenge-2012-unlocking-the-secrets-of-smartphone-data/> [Accessed: 2013.03.28]

[144] S. Mathur, T. Jin, N. Kasturirangan *et al.*, “Parknet: drive-by sensing of road-side parking statistics,” in *Proc. of MobiSys.* ACM, 2010, pp. 123–136.

[145] G. Mclachlan and D. Peel, *Finite Mixture Models*, 1st ed. Wiley-Interscience, 2000, ISBN 0471006262.

[146] M. Mcnett and G. M. Voelker, “Access and mobility of wireless pda users,” *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 9, no. 2, pp. 40–55, 2005.

[147] P. Mohan, V. N. Padmanabhan, and R. Ramjee, “Nericell: rich monitoring of road and traffic conditions using mobile smartphones,” in *Proc. of SenSys.* ACM, 2008, pp. 323–336.

[148] A. Montilla Bravo, J. Moreno, and I. Soto, "Advanced positioning and location based services in 4g mobile-ip radio access networks," in *Proc. of PIMRC*, vol. 2. IEEE, 2004, pp. 1085–1089.

[149] R. Montoliu and D. Gatica-Perez, "Discovering Human Places of Interest from Multimodal Mobile Phone Data," in *Proc. of MUM.* ACM, 2010, pp. 12:1–12:10.

[150] M. Mun, S. Reddy, K. Shilton *et al.*, "PEIR, the personal environmental impact report, as a platform for participatory sensing systems research," in *Proc. of MobiSys.* ACM, 2009, pp. 55–68.

[151] S. Nath, "Ace: exploiting correlation for energy-efficient and continuous context sensing," in *Proc. of MobiSys.* ACM, 2012, pp. 29–42.

[152] National Communications System, "SMS over SS7," 2003.

[153] Nielsen, "Smartphones Account for Half of all Mobile Phones," 2012. [Online]. Available: <http://blog.nielsen.com/nielsenwire> [Accessed: 2013.03.19]

[154] R. Noldus, *CAMEL: Intelligent Networks for the GSM, GPRS and UMTS Network.* John Wiley & Sons, 2006, ISBN 0470016949.

[155] J. Novak, "Lokalizacni data mobilnich telefonu: Moznosti vyuziti v geografickem vyzkumu," Ph.D. dissertation, Charles University in Prague, 2010.

[156] E. Oliver and S. Keshav, "Data driven smartphone energy level prediction," 2010. [Online]. Available: <https://cs.uwaterloo.ca/research/tr/2010/CS-2010-06.pdf> [Accessed: 2011.08.08]

[157] Open Mobile Alliance (OMA), "OMA Mobile Location Service V1.2," 2011. [Online]. Available: <http://technical.openmobilealliance.org/Technical/release_program/mls_v1_2.aspx> [Accessed: 2013.04.04]

[158] Orange, "D4D Challenge," 2012. [Online]. Available: <http://www.d4d.orange.com/home> [Accessed: 2013.04.07]

[159] J. Paek *et al.*, "Energy-efficient Positioning for Smartphones Using Cell-ID Sequence Matching," in *Proc. of MobiSys.* ACM, 2011, pp. 293–306.

[160] C. Palm, *Table of the Erlang Loss Formula.* C.E. Fritzes Hovbokhandel, 1947.

[161] Parizene, "Netmonitor," 2010. [Online]. Available: <http://androlib.com/android.application.com-parizene-netmonitor-iCEn.aspx> [Accessed: 2013.04.02]

[162] M. Pettersen, R. Eckhoff, P. Lehne *et al.*, "An experimental evaluation of network-based methods for mobile station positioning," in *Proc. of PIMRC.* IEEE, 2002, pp. 2287–2291.

[163] D. Pfoser and C. Jensen, "Capturing the uncertainty of moving-object representations," in *Advances in Spatial Databases.* Springer, 1999, vol. 1651, pp. 111–131.

[164] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo *et al.*, "Activity-aware map: identifying human daily activity pattern using mobile phone data," in *Proc. of HBU.* Springer-Verlag, 2010, pp. 14–25.

[165] Plusminus, "Poor mans GPS - Cell(tower)ID / Location Area Code - lookup," 2007. [Online]. Available: `<http://www.anddev.org/map-tutorials-f18/poor-mans-gps-cell-tower-id-location-area-code-lookup-t257.html>` [Accessed: 2011.08.03]

[166] Y. Qian, D. Tipper, and S. Sasanus, "Impact of signaling load on the UMTS call blocking/dropping," in *Proc. of VTC.* IEEE, 2008, pp. 2507–2511.

[167] QUALCOMM Company, "Location technologies for GSM, GPRS and UMTS networks," 2003.

[168] A. Quigley, B. Ward, C. Ottrey *et al.*, "BlueStar, a Privacy Centric Location Aware System," in *Proc. of PLANS.* IEEE, 2004, pp. 684–689.

[169] M.-R. Ra, B. Liu, T. F. La Porta *et al.*, "Medusa: a programming framework for crowd-sensing applications," in *Proc. of MobiSys.* ACM, 2012, pp. 337–350.

[170] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[171] K. Raja, W. Buchanan, and J. Munoz, "We Know Where You Are," *Communications Engineer*, vol. 2, no. 3, pp. 34–39, 2004.

[172] G. Ranjan, H. Zang, Z.-L. Zhang *et al.*, "Are call detail records biased for sampling human mobility?" *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, no. 3, pp. 33–44, 2012.

[173] C. Ratti, A. Sevtsuk, S. Huang *et al.*, "Mobile landscapes: Graz in real time." in *Location Based Services and TeleCartography.* Springer-Verlag, 2007, pp. 433–444, ISBN 978-3-540-36727-7.

[174] S. Reddy, A. Parker, J. Hyman *et al.*, "Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype," in *Proc. of EmNets.* ACM, 2007, pp. 13–17.

[175] I. Rhee, M. Shin, S. Hong *et al.*, "On the Levy-walk nature of human mobility," in *Proc. of INFOCOM.* IEEE, 2008, pp. 924–932.

[176] S. Scellato, M. Musolesi, C. Mascolo *et al.*, "Nextplace: a spatio-temporal prediction framework for pervasive systems," in *Proc. of Pervasive.* Springer-Verlag, 2011, pp. 152–169.

[177] H. V. Schelling, "Coupon collecting for uneqal probabilities," *The American Mathematical Monthly*, vol. 61, no. 5, pp. 306–311, 1954.

[178] P. Schmitz and A. Cooper, "Using mobile phone data records to determine criminal activity space," in *Proc. of IQPC*, 2007, pp. 1–29.

[179] C. Schneider, V. Belik, T. Couronne *et al.*, "Unraveling daily human mobility motifs," *Journal of the Royal Society Interface*, 2013, under review. [Online]. Available: `<http://humnet.scripts.mit.edu/wordpress/wp-content/uploads/2011/04/Interface_paper_2013.pdf>` [Accessed: 2013.04.07]

[180] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[181] L. Shi and T. Wigren, "AECID fingerprinting positioning performance," in *Proc. of GLOBECOM*. IEEE, 2009, pp. 1–6.

[182] C. Song, T. Koren, P. Wang *et al.*, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[183] M. Stasiak, M. Glaabowski, A. Wisniewski *et al.*, *Modeling and Dimensioning of Mobile Networks: From GSM to LTE*. John Wiley & Sons, Ltd, 2010, ISBN 9780470976036.

[184] C. M. Takenga and K. Kyamakya, "Robust Positioning System Based on Fingerprint Approach," in *Proc. of MobiWac*. ACM, 2007, pp. 1–8.

[185] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., 2005, ISBN 0321321367.

[186] A. S. Tanenbaum, *Computer Networks (4th Edition)*. Prentice Hall PTR, 2002, ISBN 0130661023.

[187] S. K. Thompson, "Sample size for estimating multinomial proportions," */The American Statistician(*, vol. 41, no. 1, pp. 42–46, 1987.

[188] P. Traynor, W. Enck, P. McDaniel *et al.*, "Mitigating attacks on open functionality in SMS-capable cellular networks," *IEEE/ACM Transactions on Networking*, vol. 17, no. 17, pp. 182–193, 2009.

[189] E. Trevisani and A. Vitaletti, "Cell-ID location technique, limits and benefits: an experimental study," in *Proc. of WMCSA*. IEEE, 2004, pp. 51–60.

[190] TruePosition, Inc., "U-TDOA: Enabling new location-based safety and security solutions," 2008.

[191] K. Trushkowsky, T. Kraska, M. Franklin *et al.*, "Crowdsourced enumeration queries," in *Proc. of ICDE*, 2013.

[192] M. Turner, S. Love, and M. Howell, "Understanding emotions experienced when using a mobile phone in public: The social usability of mobile (cellular) telephones," *Telematics and Informatics* , vol. 25, no. 3, pp. 201–215, 2008.

[193] P. Vacek, "SS7 Box - project in Research & Development Centre at FEE in Prague," in *Proc. of RTT*, 2003, pp. 200–202.

[194] M. Vieira, V. Frias-Martinez, N. Oliver *et al.*, "Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics," in *Proc. of SocialCom*. IEEE, 2010, pp. 241–248.

[195] M. P. Wand and M. C. Jones, *Kernel Smoothing (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1994, ISBN 0412552701.

[196] H. Wang, F. Calabrese, G. Di Lorenzo *et al.*, "Transportation mode inference from anonymized and aggregated mobile phone call detail records," in *Proc. of ITSC*. IEEE, 2010, pp. 318–323.

[197] P. Wang, M. C. González, C. A. Hidalgo *et al.*, "Understanding the spreading patterns of mobile phone viruses," *Science*, vol. 324, no. 5930, pp. 1071–1076, 2009.

[198] Z. Wang, D. Zhang, D. Yang *et al.*, "Detecting overlapping communities in location-based social networks," in *Proc. of SocInfo.* Springer-Verlag, 2012, pp. 110–123.

[199] S. Weixiong, Z. Yanfeng, Z. Jin *et al.*, "Collecting and analyzing mobility data from mobile network," in *Proc. of IC-BNMT.* IEEE, 2009, pp. 810–815.

[200] T. Wigren, "Adaptive Enhanced Cell-ID Fingerprinting Localization by Clustering of Precise Position Measurements," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 3199–3209, 2007.

[201] P. Willars, J. Bolin, and T. Wigren, "Methods and systems for obscuring network topologies," U.S. Patent 20 110 116 632, 2011.

[202] D. Willkomm, S. Machiraju, J. Bolot *et al.*, "Primary users in cellular networks: A large-scale measurement study," in *Proc. of DySPAN.* IEEE, 2008, pp. 1–11.

[203] S. Winter and Z.-C. Yin, "Directed movements in probabilistic time geography," *International Journal of Geographical Information Science*, vol. 24, no. 24, pp. 1349–1365, 2010.

[204] P. Wu, J. Zhu, and J. Y. Zhang, "Mobisens: A versatile mobile sensing platform for real-world applications," *Mobile Networks and Applications*, vol. 18, no. 1, pp. 60–80, 2013.

[205] Y. Xiao, P. Simoens, P. Pillai *et al.*, "Lowering the barriers to large-scale mobile crowdsensing," in *Proc. of HotMobile.* ACM, 2013, pp. 9:1–9:6.

[206] K. Yadav, V. Kumar, S. Jairath *et al.*, "Poster: cloud-enabled content search and sharing system for mobile phones (MobiShare)," in *Proc. of MobiSys.* ACM, 2012, pp. 521–522.

[207] D. Yang, G. Xue, X. Fang *et al.*, "Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing," in *Proc. of Mobicom.* ACM, 2012, pp. 173–184.

[208] G. Yang, "Discovering significant places from mobile phones: a mass market solution," in *Proc. of MELT.* Springer-Verlag, 2009, pp. 34–49.

[209] J. Yang, A. Varshavsky, H. Liu *et al.*, "Accuracy characterization of cell tower localization," in *Proc. of Ubicomp.* ACM, 2010, pp. 223–226.

[210] M. Youssef, A. Youssef, C. Rieger *et al.*, "PinPoint: An Asynchronous Time-Based Location Determination System," in *Proc. of MobiSys.* ACM, 2006, pp. 165–176.

[211] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proc. of MobiCom.* ACM, 2007, pp. 123–134.

[212] H. Zang, F. Baccelli, and J. Bolot, "Bayesian inference for localization in cellular networks," in *Proc. of INFOCOM.* IEEE, 2010, pp. 1963–1971.

[213] H. Zhang and R. Dantu, "Predicting social ties in mobile phone networks," in *Proc. of ISI*, 2010, pp. 25–30.

[214] H. Zhang, R. Dantu, and J. Cangussu, "Change point detection based on call detail records," in *Proc. of ISI.* IEEE, 2009, pp. 55–60.

[215] Y. Zheng, “Tutorial on location-based social networks,” in *Proc. of WWW.* ACM, 2012, pp. 1–4.

[216] Z. Zhuang, K.-H. Kim, and J. P. Singh, “Improving energy efficiency of location sensing on smartphones,” in *Proc. of MobiSys.* ACM, 2010, pp. 315–330.

Appendix A

Spatial Outlier Detection from GSM Mobility Data

In [77] we published a method that spatially extends the Reality Mining Dataset, i.e., it pairs the Cell-IDs in the dataset, to which the users were attached, with their geographical positions. The method is based on querying the Google Location API that returns an approximate mobile phone location, even when GPS in the device is not enabled [49]. The Location API may be requested not only from the mobile phone, but also from a plain computer over the Internet, as demonstrated by many authors [19, 132, 165]. We examined the spatial accuracy of the cell-tower positions from the Google Location API by comparing a cell tower database, obtained from our cooperating mobile provider, and positions retrieved from Google. We observed that the locations from Google are unbiased, and so they are suitable for approximate location estimation with Cell-ID granularity.

The main obstacle in spatial extension of the Reality Mining Dataset is that the dataset has been recorded in 2005 and we retrieved locations for its Cell-IDs in 2009. This time period represents four years of mobile networks evolution, change, cell renumbering, and it induces a number of wrong or missing values in the Google Cell-ID database. That is the main reason why only 46.75% of all unique cell locations from the Reality Mining Dataset were retrieved with geographical coordinates.

Figure A.1 depicts all geographical positions retrieved from the Google Location API for the Reality Mining Dataset. Contrary to our expectations, the locations are distributed all around the world and are not related only to Boston, USA, where the user-pool comes from. We observed that dense areas of visited locations are present not

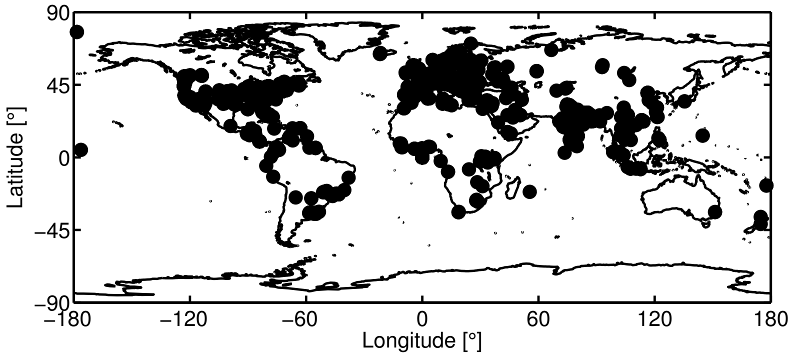


Figure A.1: Cell locations of the Reality Mining Dataset retrieved from the Google Location API.

only within the USA (San Francisco, New York, Chicago, San Diego), but also in Europe (London, Helsinki, Milan, Paris, Budapest) and Asia.

Nevertheless, many locations in Figure A.1 are placed in an unlikely or even impossible position on the map. For this reason we proposed a novel algorithm, called LAC-clustering, that detects and removes outliers in such data. LAC-clustering is a heuristic enhancement of general *agglomerative hierarchical clustering* [185]. It is based on the observation that cells with the same Location Area Code, i.e., belonging to the same location area, must be close enough to each other to ensure seamless communication handover.

LAC-clustering removes spatial outliers from cells that belong to the same location area in steps as follows:

Algorithm 2 LAC-clustering

```
1: Select cells with the same LAC.               ▷ Process for each Location Area Code
2: Compute proximity matrix.
3: Let each cell location be a cluster.
4: repeat                                          ▷ Create hierarchical cluster tree
5:     Merge the two closest clusters.
6:     Update the proximity matrix.
7: until only one cluster remains
8: Use distance criterion for forming clusters.    ▷ Hierarchical cluster tree pruning
9: Select one Location Area cluster representative.
```

First, a proximity matrix based on the Euclidean distance metric is computed (line 2) for all cells with the same Location Area Code (LAC). Second, a hierarchical cluster tree is created (lines 4–7) with linkages based on the *single linkage* method (also called *nearest neighbor*), where proximity of two clusters is defined as the smallest distance between two objects in the two different clusters. Next, the hierarchical cluster tree is pruned to partition the cell coordinates into clusters with the clustering criterion being

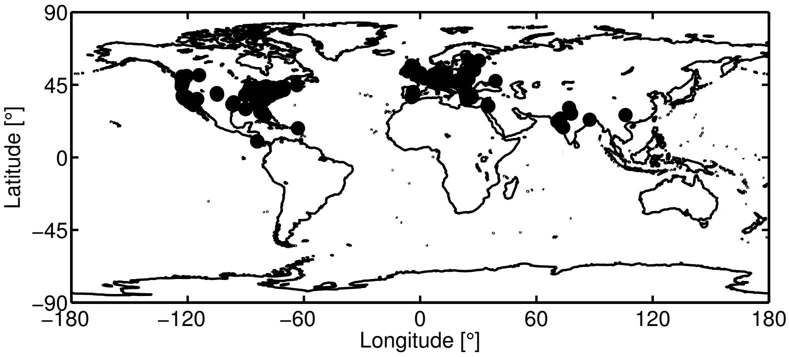


Figure A.2: Cell locations of the Reality Mining Dataset clustered according to Location Area Code.

the distance between cells (line 8). Finally, a single cluster with the highest number of locations is selected as a representative of the location area.

We selected the Euclidean distance metrics as a planar approximation of a real distance between geographical coordinates. The key factor in cluster construction is the appropriate distance selection. We choose the 35 km distance, which is the technological limit in GSM networks for successful communication between a mobile station and a cell tower. Therefore, all cell locations in one single cluster are simply points with a maximum distance of 35 km from each other. LAC-clustering is discussed in a detailed manner in [77], including correctness in terms of false positives and false negatives and the quality of clustering measured in the Location Area shape similarity.

Figure A.2 shows the result of applying LAC-clustering to the locations in the RMD retrieved from the Google Location API.

We have shared the spatially extended Reality Mining Dataset with several researchers to support their work on energy-efficient continuous context sensing [151], routing protocols in delay-tolerant networks [86] and opportunistic content sharing [206].

Appendix B

List of Personal Publications

Related to the Thesis

Journal papers

- M. Ficek, T. Pop, and L. Kencl, "Active tracking in mobile networks: An in-depth view," *Computer Networks*, 2013, *in press.* [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1389128613000996>

Conference papers

- M. Ficek, N. Clark, and L. Kencl, "Can crowdsensing beat Dynamic Cell-ID?" in *Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones (PhoneSense '12).* ACM, 2012, pp. 10:1–10:5.
- M. Ficek and L. Kencl, "Inter-Call mobility model: A spatio-temporal refinement of call data records using a Gaussian mixture model," in *Proceedings of The 31st Annual IEEE International Conference on Computer Communications (IEEE INFOCOM '12).* IEEE, 2012, pp. 469–477.
- M. Ficek, T. Pop, P. Vláčil, K. Dufková, L. Kencl, and M. Tomek, "Performance study of active tracking in a cellular network using a modular signaling platform," in *Proceedings of the 8th international conference on Mobile systems, applications, and services (MobiSys '10).* ACM, 2010, pp. 239–254.
- M. Ficek and L. Kencl, "Spatial extension of the Reality Mining Dataset," in *IEEE 7th International Conference on Mobile Adhoc and Sensor Systems (MASS '10)*, 2010, pp. 666–673.

- M. Ficek and L. Kencl, “Improving roamer retention by exposing weak locations in GSM networks,” in *Proceedings of the 5th international student workshop on Emerging networking experiments and technologies (Co-Next Student Workshop ’09)*. ACM, 2009, pp. 17–18.

- K. Dufková, M. Ficek, L. Kencl, J. Novak, J. Kouba, I. Gregor, and J. Danihelka, “Active GSM cell-id tracking: ”Where did you disappear?”,” in *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments (MELT ’08)*. ACM, 2008, pp. 7–12.

- K. Dufková, J. Danihelka, M. Ficek, I. Gregor, and J. Kouba, “Can active tracking of inroamer location optimise a live GSM network?” in *Proceedings of the 2007 ACM CoNEXT conference (CoNEXT ’07)*, no. 42. ACM, 2007, pp. 1–2.

Others

- M. Ficek, “CRAWDAD data set ctu/personal (v. 2012-01-11),” 2012. [Online]. Available: `<http://crawdad.org/ctu/personal>`

Non-related to the Thesis

Conference papers

- E. B. Martinez, M. Ficek, and L. Kencl, “Mobility data anonymization by obfuscating the cellular network topology graph,” in *Proceedings of the IEEE International Conference on Communications (ICC) ’13*, 2013, *to appear.*