KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Privacy in Location-Based Services

Michael Herrmann

Supervisors:
Prof. dr. Claudia Diaz
Prof. dr. ir. Bart Preneel

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science (PhD): Electrical Engineering

August 2016

Privacy in Location-Based Services

Michael HERRMANN

Examination committee:
Prof. dr. ir. Jean Berlamont, chair
Prof. dr. Claudia Diaz, supervisor
Prof. dr. ir. Bart Preneel, supervisor
Prof. dr. Mireille Hildebrandt
Prof. dr. ir. Frank Piessens
Prof. dr. ir. Vincent Rijmen
Dr. Carmela Troncoso
(IMDEA Software Institute)
Dr. Kévin Huguenin
(Laboratory for Analysis and Architecture of Systems (LAAS-CNRS))

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science (PhD): Electrical Engineering

August 2016

Uitgegeven in eigen beheer, Michael Herrmann, Kasteelpark Arenberg 10, bus 2452, B-3001 Leuven (Belgium)

Acknowledgements

I would like to thank my promoter Prof. Claudia Diaz and Prof. Bart Preneel for their support and guidance that they gave me. I am grateful they provided me the opportunity to do my research in such an excellent environment.

I would like to express my gratitude to Prof. Bart Preneel and to the members of my PhD committee - Prof. Mireille Hildebrandt, Dr. Carmela Troncoso, Dr. Kévin Huguenin, Prof. Frank Piessens, Prof. Vincent Rijmen - for reviewing my thesis manuscript and for providing me with insightful and valuable feedback. I thank also Prof. Jean Berlamont for chairing my defense.

Throughout my PhD I had the privilege to work with exceptional researchers from whom I have learned a lot. My biggest thanks go to Carmela Troncoso. Not only did she teach me so many things and had so many discussions with me, but she also taught me how to do research and how to do high quality work. Your enthusiasm and passion is truly inspiring. I also thank Mireille Hildebrandt who invited me to do work at the intersection of engineering, law and ethics. She was never tired to answer all my questions and patiently waited for me to understand. The interdisciplinary work with her turned out to be one of the most rewarding parts of my thesis. My special thanks go also to my coauthors Alfredo Rial, Gunes Acar, Ren Zhang, Eline Vanrykel, Kai Ning, Fernando Perez-Gonzalez, Laura Tielemans. I also thank the students Kilian Rausch, Roger Ribas, Kai Ning and Eline Vanrykel for doing their thesis with me and for their dedicated work.

My work would not have been so much fun without the nice atmosphere at COSIC and I thank Ero Balsa, Gunes Acar, Marc Juarez, Fatemeh Shirazi, Tariq Elahi, Rafael Galvez, Eduard Marín, Sara Cleemput, Iraklis Symeonidis, Cristina Pérez, Seda Guerses all all my colleagues for making my time in Leuven great. I cannot thank enough Péla Noë, Elsy Vermoesen and Wim Devroye for helping me with all the administrative overhead and for being always kind and helpful.

I was twice a visitor in Vigo and I thank Carmela Troncoso and Fernando Pérez-González for making my stays fruitful and fun. I also thank Mireille Hildebrandt for inviting me to Vrije Universiteit Brussel.

I cannot thank enough my wife for her endless love, support and patience. I owe my deepest gratitude to my grandparents. I also thank my father, my parents in law and my friends for their support.

Finally I would like to acknowledge the Research Foundation Flanders (FWO) for funding my research.

Michael Herrmann
Leuven, August 2016

Abstract

The Internet has fundamentally changed our society, the way we live, work and express ourselves. Improvements in information processing and communication technology have created mobile devices that allow the usage of smart mobile services, i.e. *apps*, that keep us constantly connected and that deliver digital content. Millions of such mobile service are available today. While being inexpensive in terms of actual money, user data has become the actual currency, rewarding the company that knows the most about their users. This has created an eco-system which, invisible during normal usage, records virtually every action online. The recorded data fuels sophisticated artificial algorithms that learn our interests, desires and secrets for purposes such as behavioral advertisement and surveillance.

We focus on location-based services (LBSs) to address the privacy concerns of the mobile-device eco-system. Currently deployed LBSs are designed in a privacy-invasive way, because the service provider and other third parties learn accurate location information on their users. This is a significant threat to the user's privacy, because entities with access to accurate location data are able to infer sensitive information, such as users' home/work address, religious beliefs, sexual orientation and income level.

We understand LBSs as a new socio-technical practice and assess its implications on privacy from an interdisciplinary perspective, including the engineering, ethics, social and legal domain. We employ the concept of contextual integrity in order to evaluate how the changes of information flow, imposed by the new socio-technical practice, affect society. Furthermore, we propose a framework that allows to quantify threats towards an adversary that is able to observe app traffic. We show that this poses a threat to the user's privacy and in particular to the user's location privacy.

In order to address the privacy concerns of LBSs, we study technical solutions that allow users to keep their data confidential. Therefore, we study the design

of private protocols for service providers that want to provide guarantees that they cannot learn user location information. Our design accounts for the current business model of most mobile services to monetize their investments with data that they learn about their users. Particularly, the overhead on the service provider's infrastructure is minimal allowing for a low cost maintenance of the service. Furthermore, the service provider is able to learn, in a privacy-preserving way, statistics about the locations that the users share among each other. This may serve as a form of monetization.

Since most service providers, due to their business model, are reluctant to implement privacy-friendly protocols, we study the design and analysis of obfuscation-based protection mechanisms. These location-privacy preserving mechanisms (LPPMs) allow users to protect their whereabouts when engaging in privacy-invasive protocols. We propose a framework that allows the computation of the optimal LPPM for users who engage sporadically in LBSs and that is tailored to a user's mobility profile and the user's constraints. We furthermore propose a framework that allows for a first-order location privacy approximation for users that employ LPPMs.

Finally, we conclude the research results of this thesis and outline paths of future work. Here we put particular emphasis on studying quantification frameworks that take a middle ground between complexity and simplicity. Furthermore, we suggest to study how our framework can be applied for the purpose of privacy visualization. This may serve as a practical tool for the visualization of both the user's current location privacy as well as an assessment on how the privacy level changes for the next query to the LBS and thus contribute for users to better understand the privacy implications of their actions. Finally, possibilistic thinking should also be applied to the design of LPPMs.

Beknopte samenvatting

Het internet heeft de manier waarop we leven, werken, en onszelf uitdrukken fundamenteel veranderd. Verbeteringen in informatie- en communicatietechnologieën hebben geleid tot draagbare apparaten die *smart* mobiele diensten ter beschikking stellen die ons constant geconnecteerd houden en die digitaal content leveren. Millioenen zulke mobiele diensten zijn beschikbaar vandaag. Deze diensten zijn goedkoop, omdat gebruikersdata het eigenlijke betaalmiddel is geworden, waardoor bedrijven met de meeste gebruikerskennis beloond worden. Dit heeft een onzichtbaar ecosysteem gecreëerd dat bijna elke actie online opneemt. De opgenomen data wordt gebruikt in geavanceerde *artificial intelligence* algoritmes die onze interesses, wensen, en geheimen leren voor gedrags-georienteerde reclame en surveillance.

Wij focusen op *location-based serivces* (LBSs) om de privacy-zorgen van het mobiel-apparaat-ecosysteem aan te pakken. LBSs die in de praktijk gebruikt worden zijn niet ontwikkeld met privacy in gedachten, omdat de dienstverlener en andere derde partijen nauwkeurige locatie-informatie leren over hun gebruikers. Dit vormt een belangrijk gevaar voor de gebruiker's privacy, omdat toegang tot nauwkeurige locatie-data gevoelige informatie kan lekken, zoals thuis- of werk-adressen, godsdienstige overtuigingen, seksuele orientatie, en inkomen.

Wij zien LBSen als een nieuwe socio-technisch praktijk en bepalen de implicaties op privacy vanuit een interdisciplinair perspectief, waaronder de ingenieurs-, ethiek-, sociaal- en legaal-domein. We gebruiken het concept van contextueel integriteit om te begrijpen hoe veranderingen van een informatie-stroom, opgelegd door de nieuwe socio-technisch praktijk, de maatschappij beïnvloeden. Verder stellen we een raamwerk voor dat het gevaar van tegenstanders die applicatie-verkeer kunnen zien kwantificeert. We tonen aan dat dit een gevaar vormt voor de gebruiker's privacy in het algemeen, maar ook voor de gebruiker's locatie-privacy.

Om de privacy-zorgen van LBSs aan te pakken, bestuderen we technische

oplossingen die gebruikers toestaan om hun data vertrouwelijk te houden. Dus, we bestuderen het ontwerp van private-protocollen voor dienstverleners die locatie-privacy willen aanbieden. Ons ontwerp neemt het huidige businessmodel van de meeste mobiele diensten in acht, om hun belegginging aan te munten met data dat ze over hun gebruikers leren. In het bijzonder, de kosten bovenop de dienstverlener's infrastructuur is minimaal, waardoor de kost voor het onderhoud van de dienst laag blijft. Bovendien, kan de dienstverlener stastistieken over de locaties die de gebruikers met elkaar delen in een privacy-vriendelijke manier berekenen. Dit kan een vorm van monetisatie zijn.

Vanwege hun businessmodel zijn de meeste dienstverleners onwillig om privacy-vriendelijke protocollen te implementeren, waardoor we het ontwerp en de analyse van obfuscatie-gebaseerde beveiligings-mechanismes bestuderen. Deze *location-privacy preserving mechanisms* (LPPMs) geven gebruikers de mogelijkheid om hun locatie te verbergen tegen privacy-brekende protocollen. We stellen een raamwerk voor waarin de optimale LPPM voor gebruikers die sporadisch gebruik maken van LBSs berekend kan worden, en dat ook op maat is gemaakt voor de gebruiker's mobiliteitsprofiel en beperkingen. Verder stellen we een raamwerk voor dat eerste-order locatie-privacy approximate toelaat voor gebruikers die LPPMs gebruiken.

Ten laatste concluderen we de onderzoeksresultaten van deze thesis, en schetsen mogelijke toekomstig werk. Hier leggen we vooral nadruk op het verbeteren van kwantificatie-raamwerken die een goede evenwicht behouden tussen complexiteit en eenvoud. Verder suggereren we hoe ons raamwerk gebruikt kan worden voor het visualizeren van privacy. Dit zou als een praktisch instrument gebuikt kunnen worden voor de visualizatie van de gebruiker's huidig locatie-privacy en een beoordeling van hoe de privacy-niveau verandert voor de volgende verzoek naar de LBS en kan dus gebruikt worden om gebruikers een beter begrip te geven over hoe hun acties effect hebben op privacy. Ten slotte, *possibilistic thinking* zou ook toegepast kunnen worden op het ontwerp van LPPMs.

Contents

List of Figures

List of Tables

Abbreviations

AA	Ad and Analytics.
AdSDK	Advertisement Software Development Kit.
AES	Advanced Encryption Standard.
APEC	Asia-Pacific Economic Cooperation.
app	Mobile Application.
BGN	Boneh-Goh-Nissim.
CI	Contextual Integrity.
COPPA	Children's Online Privacy Protection Act.
DHT	Distributed Hash Table.
DPD	Data Protection Directive (95/46/EC).
DRD	Data Retention Directive (Directive 2006/24/EC).
ECHR	European Convention on Human Rights.
ePrivacy Directive	ePrivacy Directive (2002/58/EC, as amended by 2009/136/EC).
EU	European Union.
FACTA	Fair and Accurate Credit Transactions Act.
FRVP	Fair Rendez-Vous Point.
FTC	Federal Trade Commission.
GAID	Google Advertising ID.
GDPR	General Data Protection Regulation (2016/679).
GPS	Global Positioning System.
GSM	Global System for Mobile Communications.
GSN	Geo-Social Network.
HIPPA	Health Insurance Portability and Accountability Act.

HTML	Hypertext Markup Language.
HTTP	Hyptertext Transfer Protocol.
HTTPS	Hyptertext Transfer Protocol Secure.
IMEI	International Mobile Equipment Identity.
IMS	Identity Management Systems.
IMSI	International Mobile Subscriber Identity.
IP	Internet Protocol.
IPC	Inter-Process Communication.
LBS	Location-Based Service.
LPPM	Location Privacy Protection Mechanism.
LSBS	Location-Sharing-based Service.
MAC address	Media-Access-Control address.
MOS	Mobile Operating System.
MRAID	Mobile Rich Media Advertisement Interface Definitions.
NAT	Network Address Translation.
OBA	Online Behavioral Advertisement.
OECD	Organisation for Economic Co-operation and Development.
OPPA	Online Privacy Protection Act.
OS	Operating System.
P2P	Peer-to-Peer.
PEqT	Private Equality Testing.
PET	Privacy Enhancing Technology.
PIPEDA	Personal Information Protection and Electronic Documents Act.
PIR	Private Information Retrieval.
PKI	Public Key Infrastructure.
PO	Platform Operator.
POI	Point-of-Interest.
PTSI	Private Threshold Set Intersection.
QL	Quality Loss.
QR	Quick Response.
SDK	Software Development Kit.
SM	Smoothing Module.
SSO	Single Sign-On.

TCP	Transmission Control Protocol.
TTP	Trusted Third Party.
UDHR	Universal Declaration of Human Rights.
WP	Article 29 Data Protection Working Party.
WWW	World-Wide-Web.

Part I

Privacy-Preserving Location-Based Services

Chapter 1

Introduction

The Internet has changed the way mankind lives and is ubiquitous in the daily routine of billions of people. The development of the Internet started in the late 1960s and gained shape in the 1980s with development of protocols such as TCP/IP that allowed to connect various autonomous networks. While still being a technology that did not concern the average citizen, starting with the 1990s the mass commercialization of the Internet took place. Telecommunication providers offered services that allowed end-users to be connected to the Internet in order to browse the World-Wide-Web (WWW) or to send email.

At the beginning of the 21st century the Internet fundamentally changed. Instead of simply consuming content, users started to create the content themselves and to use interactive and collaborative services. This development, commonly referred to as *Web 2.0*, has significantly influenced our society. To name only a few: people spend a substantial amount of their time in online social networks communicating, socializing, engaging in their hobbies or expressing their opinions. Events all around the globe are almost instantly reported via instant messengers and hundred of thousands collaboratively create and maintain large online encyclopedias.

During the last decade, mobile devices, such as smartphones and tablet computers, have further increased the ubiquity of online services in our lives. We not only engage in online communications or purchase items online when we are at home at our desktop computer, but do this during any instant of our live. Furthermore, we employ sensors of these mobile devices, such as Global Positioning System (GPS) sensors. This does not only allow us to engage in context aware services, but also increases the types of information that we create, transmit and store at remote machines.

A consequence of these developments is that billions of people create an unprecedented amount of data that either includes or allows the inference of highly sensitive information. Either way, it turns out that the entity that *learns* such information can use it in an extremely profitable way. The services that know more information about people are able to optimize their services and thus attract more users. Furthermore, services that know more about the interests of their users are able to present them behavioral advertisement. In this particular form of online advertisement, online advertisers present targeted ads that relates to the user's online activities. In a simple example, a user who was recently shopping online for cloths would receive matching advertisement where other users would see different advertisement.

Today, a multi-billion dollar economy exists whose main driving factor is user data. Users are typically unaware of the exact functioning of this industry. They are not aware of, for example, the multiple trackers that they download when visiting webpages or the Ad and Analytics (AA) software that they run on their mobile devices. They are also not aware of how their data is collected, aggregated and processed in economic transactions.

1.1 Goals of the Thesis

We aim at investigating the threats on information privacy that emerge when people share information while engaging in online mobile services. Information privacy has gained considerable attention in the last decades due to constant advancements in information technology. People constantly share information that is analyzed by intelligent algorithms that aim at inferring as much as possible about Internet users [98, 197, 211]. This may have significant consequences for individuals as they have virtually no control about how their data is being used and how inferences about them take effect when, for example, shopping online [153].

For the scope of this thesis we focus on Location-Based Service (LBS). In these services the user's location is processed in order to deliver a service, such as finding a nearby Point-of-Interest (POI) or getting directions. Due to their usefulness, LBS are being used by most mobile device users [6,212]. Some services require the users to reveal their location either *sporadically* or *continuously*. The difference between the two is that in the sporadic case two consecutive queries to the LBS of the same user happen with sufficient time in between so that they do not correlate. We shall understand LBSs from a broader prospect than only from a technology perspective. Instead, we consider processing location data as a new socio-technical practice, having impact on technology, society

and legislation. We adopt an interdisciplinary approach in order to understand the threats on privacy from engineering, ethical and legal perspectives. We employ the ethical as well as engineering approach in order to analyze how the socio-technical practice of location sharing alters established information flow. Changes in information flow may disturb the common sense of privacy and this indicates that the new socio-technical practice may have a significant impact on society. We weigh these impacts on privacy of the socio-technical practice against the advantages that it brings. Furthermore, we undertake an analysis of the new socio-technical practice from a legal perspective.

Mobile devices are the main platform that people use in order to engage in LBSs. A main part of the privacy problem of LBSs is that the software of these platforms is not designed with privacy in mind. This thesis contributes to the understanding of information that users leak by engaging in LBSs.

In this thesis we develop and analyze technologies that allow users to hide their location data while engaging in LBSs. We tackle the privacy issues of LBS from a data hiding perspective, i.e. confidentiality perspective. Note that, especially from an interdisciplinary perspective, privacy can be achieved in other ways than via confidentiality. From a legal perspective, for example, privacy can be achieved although other entities receive the user's data but the legal regime forbids or limits its processing.

We first take the perspective of a LBS provider that wants to provide its service in a privacy-preserving, i.e. *private*, way. Therefore, we employ state-of-the-art cryptographic mechanisms that allow the user to hide her location from the service provider. Since this makes it impossible to monetize user data, the most predominant business model of LBSs, we investigate services that require little computational overhead from the service provider and thus are cost efficient. Furthermore, we investigate mechanisms that allow the service provider to compute privacy-preserving statistics on location data that can be used to monetize the investments of developing and maintaining the service. This may help to create a competitive market around privacy-preserving LBSs.

Since most service providers are reluctant to employ privacy-preserving protocols, this thesis also investigates the design and analysis of Location Privacy Protection Mechanisms (LPPMs) as a second way for the user to hide her location data. LPPMs allow users to obfuscate their whereabouts with various strategies when engaging in LBSs. While a wide series of such LPPMs are described in the literature, we investigate the design of optimal obfuscation strategies that are tailored towards a particular user profile and preferences. We also investigate ways to quantify the level of location privacy that users enjoy when protecting their whereabouts with LPPMs. This is a crucial part of obfuscation-based location privacy protection. Even if LPPMs allow some

protection of location privacy, the user still leaks information on her whereabouts. Furthermore, a framework for location privacy quantification allows for the unified evaluation of the different LPPMs that exist in the literature.

1.2 Contribution of the Thesis

We employ the framework of Contextual Integrity (CI) in order to analyze how the socio-technical practice of location sharing impacts privacy. This is combined with a technical analysis of LBSs. Furthermore, we analyze LBSs from the legal perspective of *purpose limitation*. The relevant work is described in [105].

We propose a framework that allows for a large-scale, automated evaluation of the information leaks on Mobile Application (app) traffic. This framework analyzes the extent to which mobile apps enable third party surveillance by sending unique identifiers over unencrypted connections. It also contributes to the analysis of how user information, such as identifier and location information, is being used by different software modules running in apps. Furthermore, we analyze the effect that mobile ad-blocking tools have on the data that is being transmitted in the clear by apps. The relevant work is described in [195].

For the design of private LBSs we address LBSs that allow users to share their location. We employ two schemes based on identity-based broadcast encryption. For the first protocol we assume the user to reveal the friends who should receive her location updates. This allows for a particular efficient design of the service. For the second protocol we employ anonymous identity-based broadcast encryption. Both of our protocols have the advantage that the LBS provider does not need to perform computationally expensive operations. Furthermore, we extend the first protocol in order to allow the service provider to collect privacy-preserving statistics on the locations shared among the users. While this extension requires the LBS provider to perform additional computations, the obtained statistics could be monetized to compensate for this additional overhead. The relevant work is described in [107].

For the domain of obfuscation-based location privacy protection we first propose a framework that allows the computation of optimal LPPMs for users who sporadically engage in LBSs. Our framework accounts for resource constraints of mobile devices with which LBSs are typically being accessed. This makes our framework suitable to compute optimal dummy and precision-based obfuscation strategies. An analysis of several optimal obfuscation strategies further shows the trilateral trade-off of privacy, Quality Loss (QL) and bandwidth overhead

that a user makes when setting her parameters for an LPPM. The relevant work is described in [108].

Finally, we propose the novel notion of *possibilistic location privacy* and compare it with the existing probabilistic model. This comparison shows that although the probabilistic model is in principle able to quantify privacy of very complex scenarios, its computational complexity renders such evaluations impossible in realistic scenarios. We provide a framework that enables the possibilistic privacy evaluation of LPPMs and show that our framework can quantify privacy in an extremely efficient manner. The relevant work is described in [106].

1.3 Structure of the Thesis

This thesis consists of two parts. Part I provides the introduction of this thesis: it introduces the privacy concept, provides a description of how location data is being used in current mobile device platforms, outlines existing work on private LBSs and obfuscation-based LPPMs. In this chapter we also explain what our contributions are.

Chapter 2. This chapter discusses the concept of privacy and its importance. We address in detail the concept of CI due to its importance to this thesis. We provide an overview on the development and the current privacy legislation. The last part of the chapter presents the foundations of privacy enhancing technologies.

Chapter 3. In this chapter we address the usage of location data. We provide a description of the common mobile platforms on which users engage in LBSs. This includes a description of mobile devices, the apps and the business model of app developers. We also address privacy concerns of mobile platforms in general and concerns on location privacy in particular. We outline in this chapter the legal protection of location data. Furthermore, we provide an overview on the literature that reports about user perception of LBSs, which includes a description on users' attitudes towards LBSs and the users' privacy preferences. Finally, we summarize the inferences any entity with access to location information can make about an individual.

Chapter 4. Here we outline the design of private LBSs. We categorize existing proposals into four types of services: geo-social networks, friend-nearby notification, point of interest finder, and traffic monitoring.

Chapter 5. This chapter outlines existing work on obfuscation-based protection mechanisms, i.e. LPPMs. We outline work on the quantification of

these LPPMs. This includes a discussion on commonly used location privacy metrics as well as the state of the art quantification framework. Furthermore, we summarize existing proposals and present them along the four obfuscation strategies: hiding, reducing-precision, perturbation and dummies.

Chapter 6. In this chapter we draw the conclusions of this thesis and present an overview of future work.

Part II contains the publications related to this doctoral thesis.

- Our interdisciplinary analysis of LBSs [105].
- Our framework for the quantification of information leakage in mobile devices [195].
- The design of our LBS that allows users to privately share location and location related information [107].
- Our model for the computation of the optimal LPPM in the sporadic setting [108].
- Our quantification framework for the possibilistic evaluation of location privacy [106].

The complete list of publications can be found on page 92.

Chapter 2

Privacy

Privacy is a concept that, according to Smith [178], is extensively studied in the philosophical, psychological, sociological, legal and engineering discipline for more than 100 years. Concerns regarding privacy have typically been raised whenever technological advancements have allowed to alter how information can be gathered, accessed or used.

There are many definitions and aspects of privacy [152], but *information privacy* has received considerable attention during the last decades. Information privacy studies the field between the massive data dissemination/collection/processing and the legal/political/technological issues surrounding them. It has significant impact on the way that we live in the information age that is, especially since the 1960s, dominated by constant progress in information technology. Furthermore, as Solove noted [179], the question of how we shape our society in the information age is extremely challenging. One of the first discourses on information privacy is Warren's and Brandeis's [170] work on *The Right to Privacy*. Concerned of the technology that allows instant photographs and its meaning to the privacy of society, Warren and Brandeis have raised the question how the law is supposed to protect the peoples' privacy. They argue that people have *the right to be left alone*, i.e. the freedom from interference. From a philosophical perspective, this is a *negative right*, because it obliges inaction. There are, however, other scholars that consider privacy to be a *positive right*, i.e. a right that obliges action. A common example for a work defining privacy as positive right is Westin [202], who defines privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others".

While the described technological advancements have brought society numerous advantages, their threats to information privacy cannot be doubted. Nowadays, billions of people constantly create data while they browse the WWW, use online social networks or employ their mobile devices. The latter are typically equipped with several sensors allowing to process not only virtual activities, such as clicks in apps, but also to process physical location, audio, video and so on. It is easy to see that processing such data is a very delicate matter as it allows to gain very sensitive insights in the private lives of the users.

The 21st century has shown us that there is a very large interest in user data. Multi-billion companies exist whose core business model is to gather data, aggregate it and monetize the results. Recent revelations by Snowden, see Landau [131], have shown that intelligence agencies all around the world use any possibility to massively collect data and to surveil people in the name of national security. While many practices are scary as they are, little hope exists that the situation will improve in the near future. With an estimated 50 billion devices on the Internet in 2020, we are stepping into a world in which virtually every electronic device is able to communicate to any other device. In this world, enough data is being created that algorithms developed by artificial intelligence researchers have the potential to analyze every single aspect of our lives. Users, tempted by smart and useful services, are typically unaware or overwhelmed to understand how and by whom their data is being processed.

It is crucial to note that in our society information necessarily needs to be exchanged. People need to reveal their income to tax authorities, cannot subscribe to gym memberships without revealing their address and bank account information, tell their friends about their weekend plans and many more. Information privacy is therefore a relational concept that depends on the entities being involved, such as close friends, family, employers, service providers and government agencies. These entities *process* information and this may also be in the sense of the citizen/consumer who provided it. Therefore an entity that processes user data, be it a human or a computer algorithm, cannot necessarily be seen to be adversarial as they may act in the very interest of the citizen/consumer. However, for the design and analysis of Privacy Enhancing Technologies (PETs) (cf.C̃hapter 4 and 5) we adopt a perspective in in which a user wants to share information only with her intended receivers but not with third parties. Thus, entities such as the service provider, third party software developers or government agencies that may attempt to learn user data are considered to be adversarial.

In Section 2.1 we explain how several scholars argue that privacy is essential for both individuals and for the whole society. Furthermore, we discuss the philosophical concept of CI that connects privacy issues with social norms and contexts that exist while information is exchanged. The arguments and concepts

invite for a discussion on what is at stake when information is transmitted. However, they do not have any legal power. In Section 2.2 we discuss the legal protection of privacy and data protection for several countries.

2.1 Importance of Privacy

There is a broad consensus that privacy is valuable and beneficial at an individual, group, organizational and societal level, where the individual and the societal level have received most attention [178]. For Westin [202] privacy is related to informational self-determination. It is essential for an individual to decide which information about herself in which situations should be revealed. Reiman [167] marks that the autonomy over personal information is key to intimate relationships. Intimacy is constituted and signalled by the information that people choose to share with another. Likewise, Gerstein [94] noted that people create different levels of intimacy and trust depending on the personal information they disclose. For Kang [120] people create intimacy on the basis of what experiences they share with each other; he notes that this would not be possible without privacy. For Kupfer [129] privacy is necessary for autonomy and personal growth. Without privacy people would neither be able to develop "...a purposeful, self-determining, responsible agent..." nor would they engage in self-discovery and self-criticism. Phillips [160] argues that privacy is essential to build a sphere of autonomy without intrusions from the state or any pressure of social norms. For Altman [5], people need personal space, and thus privacy, in order to separate themselves from others and to define the self. Removing that space leads to hostility and unease. Kang [120] enumerates two further values of privacy. First, preventing misuse of personal information since the individuals are able to control their own information. Second, avoiding embarrassment for actions that are perfectly normal, yet, when exposed to the public, are considered to be embarrassing. Wagner DeCew [66] names a series of other values in privacy that are all centered around the control of one's own information. Particularly, she names: freedom from scrutiny, judgment, prejudice, pressure to conform and exploitation. While Gavison [90] also promotes individual values of privacy, such as autonomy, selfhood and human relations, she further argues that privacy is necessary for promoting liberty and establishing a free society. Similarly, Solove [179] adds s psychological well-being, individuality and creativity, but also finds that privacy is necessary for freedom and democracy as individual values of privacy.

2.1.1 Contextual Integrity (CI)

Nissenbaum introduces the concept of CI in order to understand what is at stake with privacy and to illustrate the issues that arise when data is being shared. Key to CI is that there is a common sense, a cultural practice, on how information is supposed to be shared. The sense of privacy is always disturbed when established informational norms are disturbed. Nissenbaum posits two types of informational norms. Norms of appropriateness and norms of distribution. While the former shapes what information is appropriate to disclose, the latter shapes how information may be distributed. If a a certain information flow violates any of the two informational norms it is considered to be inappropriate. This could have several reasons and depends on the type of information that is being transmitted, the persons to whom the data relates, between which parties this data is being transmitted, and, finally, the conditions or constraints under which this transmission is being made. For example, a patient who is visiting her doctor communicates, i.e., transfers, sensitive information about her health condition to her doctor. This is a perfectly normal and acceptable transmission of information. However, if that doctor communicates the shared information to a friend of the patient, this information flow would clearly be considered inappropriate, because the doctor would have violated the social norm, and even the law, of doctor-patient-confidentiality. However, if the doctor would communicate the data to a colleague, then this is within our expectation of doctor-patient-confidentiality. Numerous other examples exist to illustrate Nissenbaum's concept of CI. Information shared within a lawyer and her client underlies more restrictions to be shared among others than, for example, information shared among colleagues. All these examples appear to be rather intuitive, illustrating the common sense that exists when information is being shared.

While CI is merely a concept for understanding the concept of privacy, Nissenbaum introduces the CI *heuristic* as a more practical tool to determine whether a new technology violates CI. More specifically, her heuristic allows the assessment of a *socio-technical practice* with respect to its implications on CI. While a technology is merely a technical possibility, the respective socio-technical practice is the technology being used in particular environments that are shaped by economic, social and political factors. For example, determining the location with the help of a GPS receiver and a set of orbiting satellites is a technology, the socio-technical practice of LBSs is also influenced by business incentives, peoples' desire to share locations among each other and the respective legislation on how location data is to be protected.

Applying the CI heuristic to a particular scenario is a process that involves nine different steps. The steps one till five serve to shape the scenario of the

new socio-technical practice. This includes the definition of *sender*, *receiver* and *information type* that is being transferred. Furthermore, they include the analysis of the *prevailing context*, *transmission principles* and *information norms* of the new socio-technical practice. The context, such as health or education, defines the setting in which the information flow happens and shapes the setting and the rules under which the information flow takes place. For example, a fraud detection system that considers the location where a credit card payment is being made happens clearly in a financial context, which already shows the sensitivity of the socio-technical practice and the information that is being shared. The transmission principles and the informational norms define the rules that govern the information exchange and how the information may be processed further. Information exchanged between banks and their clients is clearly confidential and thus the bank is expected not to share this data with other entities.

The sixth step of the CI heuristic is a *prima facie* assessment of whether CI has been breached by observing whether entrenched informational norms have been violated. If this step comes to a negative assessment, then the CI heuristic stops at this point. However, in the case of a positive assessment, the heuristic continues with the steps seven till nine. Step seven and eight assess the severity of the CI violation. Particularly, step seven investigates the harms, threats to autonomy, freedom, power structures, justice, fairness, equality, social hierarchy and democracy that are due to the new socio-technical practices. In step eight the practices that directly impinge on values, goals and ends of the identified context are analyzed. In the ninth step the CI heuristic concludes and analyzes whether the advantages of the new socio-technical practice compensates for the breach of CI. This reflects that the CI heuristic may leave room for informational norms to change if this is overall beneficial for society.

2.2 Privacy Legislation

The right of privacy was first declared by the Universal Declaration of Human Rights (UDHR) of 1948 [147] in Article 12: “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.” In 1950 the European Convention on Human Rights (ECHR) [150] defined the human rights and fundamental freedoms in Europe. In Article 8 the ECHR provides the right to “respect for his private and family life, his home and his correspondence”. This right shall only be limited if three conditions are fulfilled. First, there must be a legitimate goal for the infringement, such as in the interests of national security and public

safety. Second, the infringing measure must be in accordance with the law. This includes i) the infringement must be foreseeable; ii) the relevant provision needs to be accessible; and iii) sufficient safeguards must be contained in the relevant provision that limit the infringing measure and enable redress with an independent authority. Third, the infringing measure needs to be necessary in a democratic society, i.e. the measure is appropriate for the legitimate aim, there is no less infringing measure and the seriousness of the infringement is proportional to the legitimate aim. The Organisation for Economic Co-operation and Development (OECD) issued its principles for protection of personal data in 1980 [149]. These include: Collection Limitation Principle, Data Quality Principle, Purpose Specification Principle, Use Limitation Principle, Security Safeguards Principle, Openness Principle and Individual Participation Principle.

Declarations such as the UDHR or the OECD are not legally binding for its member states and thus have no legal effect. Instead, such treaties are used as customary international law or a tool to apply diplomatic pressure. One of the first countries to adopt information privacy protection was the United States with the Privacy Act of 1974 [181] that "... establishes a code of fair information practices that governs the collection, maintenance, use, and dissemination of information about individuals that is maintained in systems of records by federal agencies". The Unites States Privacy Act had significant influence on the definition of the OECD principles for protection of personal data in [149]. This was followed by Australia with its Privacy Act of 1988 [19]. In 1995 the European Union (EU) approved the Data Protection Directive (95/46/EC) (DPD) for which the ECHR and the OECD principles served as a blueprint. Since the DPD is a directive, member states of the EU have some room for interpretation on how the DPD is being implemented in national law. On May 4th 2016, the EU published the General Data Protection Regulation (2016/679) (GDPR) [192]; it will become effective in 2018. This regulation illustrates the increased importance of data protection within the EU legislation. A more modern and unified legislation may allow the EU to meet the challenges of data protection in the 21st century. Furthermore, it will create a unified legislation within all EU member states, because, as a regulation (rather than a directive) it is a binding legislative act. Several other countries have implemented a data protection legislation, such as Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) of 2000 [39] and Japan's Personal Information Protection Law of 2003 [116]. In 2005 the Asia-Pacific Economic Cooperation (APEC) [13] implemented a privacy framework for cross-border privacy protection in the Asia-Pacific region.

The U.S. constitution does not mention a right to privacy. However, legal experts agree that several amendments to the constitution defend various aspects of privacy [148]. The first amendment defends freedom of speech, religion, and

association; the third amendment restricts quartering of soldiers in homes and prohibits it during peace time; the fourth amendment prohibits unreasonable searches and seizures; the fifth amendment prohibits self-incrimination; the ninth amendment protects rights not enumerated by the constitution; finally, the fourteenth protects personal liberty versus state actions. Furthermore, the U.S. issued several other acts on specific grounds or contexts that partially regulate information flow. These acts include the Children's Online Privacy Protection Act (COPPA) of 1998, the Health Insurance Portability and Accountability Act (HIPPA) of 1996 and the Fair and Accurate Credit Transactions Act (FACTA) of 2003. The only notable exception with respect to data protection legislation is the state of California that has issued in 2003 the Online Privacy Protection Act (OPPA). This act requires, for example, websites to post a privacy policy. The Federal Trade Commission (FTC) proposes fair information practices in electronic marketplaces [89] that, among others, promote user notice/awareness and choice/consent for personal data processing. While they are widely accepted, they are only guidelines and not enforceable by law.

2.3 Privacy Protection

Research on privacy technologies started in the late 1970s and at the beginning of the 1980s. One of the first papers on this topic was authored by Chaum [47] who proposed a method to communicate when all parties are under surveillance. Confidentiality in online communications, i.e. hiding who talked with whom, when and about what over electronic communication networks, was among the most commonly studied topics of privacy research in the 1980s. In the 1990s many new topics were studied including electronic cash [45] and credential systems [32, 38]. Powerful solutions for both topics could be proposed thanks to breakthroughs in cryptography, such as zero-knowledge proofs [83]. In the 2000, the scope of privacy widened and its focus also included to study the privacy of the entire system instead of only focusing on the mathematical core.

Danezis and Gürses [62] separate *Privacy Technologies* into PETs and *Privacy Preserving Data Publishing and Mining*. The former includes technologies such as *communications anonymity and anonymizers* that have the goal to make users of online communications systems not identifiable within a set of others, i.e. the anonymity set [159]. Typically anonymizing networks are being distinguished on the basis of the latency between user request and system response. Tor [68] is probably the most widely used representative of a low-latency anonymizer. There is a wide variety of high-latency solutions, such as [61,145], that are all variations of Chaum mixes [47]. More recently, Chaum proposed *cMix* [46], a mix network that aims at solving the performance overheads of its predecessors. Privacy

preserving data publishing and mining includes technologies that allow personal information databases to be analyzed in a privacy-preserving manner. This is achieved for example by suppressing and generalizing identifiers [140, 184] or by employing cryptographic primitives [3, 194]. Differential Privacy [71] proposed by Dwork achieves a similar goal. The key difference to tabular data analysis is that a differentially private mechanism adds noise to the query result in such a way that the result from a query which a specific individual's data is included is almost identical from the result of a query in which it is left out.

According to Danezis and Gürses s [62] PETs can be categorized as: privacy as confidentiality, privacy as control and privacy as practice.

- Privacy as confidentiality refers to a user's ability to prevent information from becoming public. There are many technologies that help to achieve confidentiality in several scenarios and systems. The technology that is suitable for a specific scenario depends on the property that needs to be private and the adversary model that is considered. An encryption algorithm, such as the Advanced Encryption Standard (AES) [163], can be used to provide confidentiality of the message content. While Tor provides confidentiality of the message content, it also provides the user with unlinkability of her different HTTP sessions to adversaries that are not able to observe network traffic at certain positions.

- Privacy as control is closely related to informational self-determination (cf. Section 2.1). In practice Identity Management Systems (IMS) and Single Sign-On (SSO) systems offer users the possibility to control which entity is allowed to learn which information. Every major IT company, such as Microsoft, Apple, Google or Facebook, offers a service that allows users to log in with their own credentials to a variety of their own services or even to third-party services. This provides usability and security advantages, because users do not have to maintain numerous credentials. Furthermore, those services may be used to attest attributes, such as the user's age, to third parties. Of course this requires users to first prove certain attributes to the service provider. This exactly shows the privacy problem with such services, because the operator does not only learn all the actions that the user performs, but also the user's attributes. To solve the privacy problems with IMS and SSO systems, several anonymous credential systems, such as [36, 55, 125], have been proposed. Those systems, however, are not yet deployed in practice.

- Privacy as practice is an extension to privacy as control. The user does not only have the choice about what information is revealed to which service, but she is also able to understand how information is being transmitted,

aggregated and used. Services offer feedback to the users to help them understanding the privacy of their actions (e.g., [132]). Other services, such as *BlueKai Registry* proposed in [157] allow users to inspect the profiles online advertisers store about them and additionally provide ways to correct and modify the profile.

Usability is a major problem for privacy technologies. Users who wish to employ privacy preserving mechanisms typically have to sacrifice performance. For example, as measured by Wendolsky et al. [201] browsing the web via Tor can be significantly slower than browsing the web with a regular Internet browser or, as noted by Leon et al. [133], privacy enhancing technologies reduce or break the functionality of systems. Another major challenge for privacy technologies is that the typical user has insufficient understanding on how technical systems work [142]. Therefore many researchers propose *nudges*, such as [198,203], which are paternalistic interventions that are designed for users to employ systems in a more private way.

Related to privacy as confidentiality is Brunton's and Nissenbaum's work in [34]. They elaborate that it can be justified to apply obfuscation as protection mechanism against data collection and analysis. Obfuscation is a technique to protect user privacy; it can be applied in many systems. Unlike encryption algorithms, that allow users to keep data confidential, in obfuscation users reveal the data but hide it under a set of other, false information or make the data less precise. This happens with the goal to make it harder to collect sensitive information by producing misleading, false or ambiguous data in order to make data gathering less reliable and therefore less valuable. For Brunton and Nissenbaum obfuscation is the only means for users to level a playing field, in which users of information systems face two major asymmetries. First, users face a *power asymmetry*, because they cannot choose to be not surveilled or have their data recorded, respectively. Second, they face an *epistemic asymmetry* since they are often not fully aware of what happens when they engage in information systems and have no control over what happens with their data. In many cases, obfuscation is the last resort for users to protect their privacy. Opt-out mechanisms shift the responsibility to the user whose data is being collected; the law operates slowly in the presence of a quickly changing technological landscape; and PETs require trade-offs in utility and are typically only known by experts who already behave much more privacy aware than the average user.

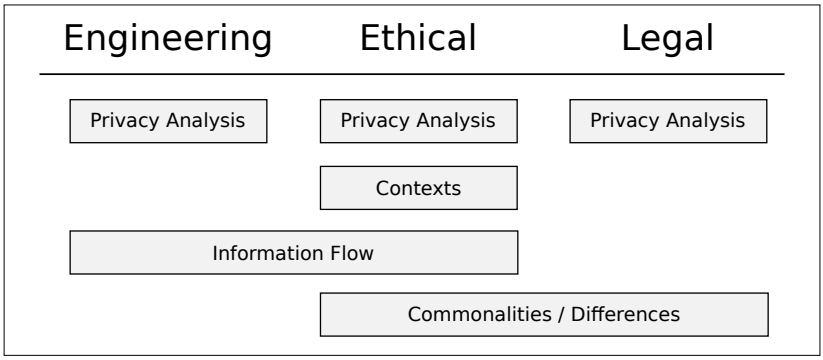


Figure 2.1: Contributions of our interdisciplinary work

2.4 Contribution

Figure 2.1 outlines the main contributions of our work. We conduct a privacy assessment of LBSs from an engineering, an ethical and a legal perspective. Scholars from every of the involved disciplines obtain an added value for their own field and additionally obtain an insight in how the topic is addressed in other disciplines. For example, the detail provided by the legal perspective serves as a valuable reference for legal scholars. However, scholars from other disciplines may find the provided information especially valuable, because works in their discipline typically lack this kind of legal detail. Also, the combination of the engineering, ethical and legal discipline to assess privacy implications in LBS contributes towards a better understanding of privacy issues from a broader perspective.

We apply the CI heuristic in a novel way. Typically, the CI heuristic is used to analyze how a new socio-technical practice impacts the CI in very specific situations. For example, consider a library that provides a paper-based repository of all available books that can be used by visitors to locate the books in which they are interested. If the library would replace this paper-based system by a computer-based search system, then the situation for the visitors would fundamentally change. While consulting the paper-based repository leaves no digital trace, any action on the computer-based system can be recorded. Finding the prevailing context of the CI heuristic is in this scenario relatively straight forward, because visitors use the system only in the library. This is different with LBSs as they are employed in numerous situations and contexts. In our work we analyze the impact of the new socio-technical practice on CI independent of a particular scenario. Instead, we employ the context of traveling for our analysis. This allows us to analyze a socio-technical practice, such as

LBS, that is used in many different scenarios.

Our analysis also shows that the technical detail from an engineering perspective provides a substantive added value in connection with the ethical as well as the legal discipline. Furthermore, we identify the special relation between the ethical and the legal discipline, i.e. the commonalities and differences between the concept of CI and purpose limitation.

In summary, our interdisciplinary work in [105] contributes to all parts of this chapter except Section 2.3. As we apply the CI heuristic to analyze the socio-technical practice of LBSs, we particularly contribute Section 2.1.1. With our discussion of the legal concept of purpose limitation itself, we contribute to the legal assessment of LBSs and thus our work also contributes to Section 2.2.

Chapter 3

Location Data

In this chapter we present the necessary background on location data. We provide a detailed explanation on apps that are the most important tools for people to utilize the capabilities of their mobile devices, such as their location data. This also includes a description of the prevailing business model of the app-eco system and its privacy issues, as well as the users' attitude towards this business model. We continue with the explanation of the legal protection of location data and outline users' attitude towards LBSs. Finally, we provide a summary of the threats to user privacy when location data is misused.

3.1 Mobile Device Eco-System

Mobile devices are nowadays ubiquitous companions. More than two billion people worldwide are using smartphones and more than 1.2 billion people are using tablet computers. Smartphones are intuitive to use and they are equipped with a wide variety of communication modules, such as WiFi, Bluetooth and 4G/LTE, allowing their users to be connected in ways never possible before. Another reason for the success of smartphones is their platform, i.e. the Mobile Operating System (MOS) and the app-store that are maintained by the Platform Operator (PO). More traditional mobile phones usually run a MOS developed by the phone manufacturer and offered only a very limited number of additional applications, which typically have also been implemented by the phone manufacturer. The software of modern mobile devices, however, is much more sophisticated. The MOS developer offers a Software Development Kit (SDK) that allows third-party developers to implement apps for their

respective MOS. Furthermore, the MOS developer runs markets that allows these developers to upload, offer and sell their apps. These software libraries and app markets are the major reasons that helped to create a multi-billion dollar market for companies to develop their apps. Tablet computers are equally intuitive to use. They employ the same software, have larger screens and overall superior hardware. This makes them in many scenarios an ideal replacement to traditional computers and laptops.

We focus in the following on the two most successful MOS developers: Google with their Android and Apple with their *iOS* Operating Systems (OSs). Note that other other companies and organizations, such as Microsoft (Windows 10), Research in Motion (BlackBerry 10) and Mozilla (Firefox OS), offer a very similar infrastructure.

3.1.1 Mobile Applications

The software design of mobile devices is, in the definition of Zittrain [213], a *generative technology*. Third-party software developers are able to develop apps that run on the manufacturer's OS. This design is key to the success of current mobile device infrastructures. Hundred of thousand of third-party software developers provide a wide variety of apps for almost any possible purpose. This makes mobile devices very useful and has created a huge market around those apps as we see in the next section. While the advantages of apps cannot be doubted, the fact that users run third-party software on their phones raises security and privacy concerns. POs implement several measures to provide a secure and private platform.

In order to prevent one app to interfere with another app, MOS typically implement some sort of *application sandboxing*. Android builds on the Linux user separation to employ sandboxing [9]. The Linux kernel prevents an app from interfering with another app; it also restricts the app's memory, CPU usage and restricts access to device sensors, such as the GPS receiver. Since Apple does not provide any official information on their sandboxing, the only information available has been obtained via reverse-engineering the code [27, 60].

Another major security and privacy concern is the presence of malicious apps. Such apps are comparable to malware on ordinary computers. They typically have a benign-looking functionality, but actually run malicious code on a user's mobile device, such as sending text messages to premium numbers or monitoring the sensors of the mobile device [166]. The most important countermeasure that the PO implements towards this threat is policy enforcement in their own app-market. The PO defines and enforces rules that an app needs to comply with in order to be offered via the respective app-market. App-market providers

analyze the apps that developers want to upload to their store. This process is called *application vetting*. Although the vetting process of Google's PlayStore and Apple's app-store are similar, there are some differences between them. The most important difference is that the *Google Bouncer*, Google's tool for application vetting, permits apps the dynamically load code during the runtime of the app, while Apple forbids such operations.

Sandboxing and vetting protect the users against malicious apps, such as sending unauthorized text messages or accessing private data of the user. However, apps may have proper reasons for accessing such resources and data. For example, a navigation app needs access to the user's location in order to provide directions from her current location or a messenger app may need access to the address book of the user. To account for legitimate access, the MOS usually implements a permission control that allows the users to specify which resources an app may access.

Android separates between normal and dangerous permissions [10]. All dangerous permissions are organized in *permission groups*. Starting with Android 6.0, which was released on Oct. 5th 2015, apps need to ask permission the first time they want to access a certain resource. This allows the user to learn how the phone's resources are being used. If the user grants a privilege of a certain category, then the app automatically is allowed to access all the resources of the category. For example if an app requests the ACCESS_COARSE_LOCATION privilege in order to get the coarse location from the user, then the MOS displays to the user that the app needs access to the phone's location. If this request is being granted and the app requests ACCESS_FINE_LOCATION at a later time, then the MOS automatically grants the request. Previous versions of Android request the user to approve all possible permissions that the app requests at install time. The app is only installed if the user grants all the requested permissions. In May 2016, only 7.5% of all Android devices were running this latest Android version [11].

Apple separates in iOS between *Entitlements* [14] and the permission dialog system. Entitlements are general permissions of the app that cannot be influenced by the user. They give an app-specific capabilities or security permissions, such as accessing cloud storage or enabling the app to send the user messages even if the user is not actively using the mobile device, i.e. push notifications. The permission dialog system works similar to the one of Android 6.0. If an app requires a privilege, then it prompts the user for approval. The user may change certain permissions of the app at any time.

As mentioned in Section 3.1, the PO provides SDKs to ease the development process of apps. Another reason that makes the development of apps relatively easy is the huge market of third-party software libraries that can be embedded

in an app. These libraries offer a broad range of different functionalities, such as advertisement, networking and databases. As a result, most apps do not only contain code from the app developer, but also from many different software developers. Android as well as iOS grants the permissions of the app the app as a whole, including code from the app developer and any third-party software libraries.

3.1.2 Business Model

One of the reasons that such a broad variety of mobile apps exists is that the Google Play and Apple market provide many ways for developers to run their app development as a business and to monetize their apps. There are three main ways of monetization: selling the app via the app-store, offering in-app purchases, and advertisement.

Selling an app is a straightforward business model. In this case the app developer asks a certain price for the app that every user needs to pay in order to download and run the app. Note that the app developer does not obtain the full price, as a certain fraction of the turnover is kept by the app-market provider. The second possibility to monetize an app is to offer so called *in-app purchases*. These allow the user to buy additional features from within the app. Similar to the case before, the app-market provider themselves keep a fraction of the in-app purchase turnover. The third possibility for monetization is via displaying advertisement to the user.

There are several large mobile AA networks, such as AdMob, MoPub, AirPush and AdMarvel, that provide third-party libraries that can easily be integrated in an app. With embedding an AA library, the app-developer enables the AA network to display behavioral advertisement that the AA network considers relevant to the user. The app-developer then typically earns a fixed price depending on whether the ad was only being displayed to the user or whether the user has clicked on the ad. Additionally to the advertisement functionality of AA libraries, they also provide analytics functionality that provide the app developer with an overview on how the app is being used and on the revenue of their app.

Although it is not directly a business model, it has been shown that many app developers hope to eventually get acquired by another tech company. While there are many reasons for companies to acquire other companies [84], recent acquisition by major Internet companies show that there are two prevailing reasons. Either a company possesses a technological advantage over the acquirer and/or has a large and vital user base. Take for example Facebook's acquisitions [57] of Gowalla and Instagram or Alphabet's acquisitions [56]

of YouTube, AdMob and DoubleClick. Certainly, compared to the number of companies developing apps, the number of acquisitions is relatively small; however, an acquisition can be extremely profitable, e.g. Facebook has acquired WhatsApp for US$ 19 billion.

In the following we provide more information on how ad and analytic networks operate, because their presence in the mobile device eco-system has a significant impact on the user's privacy (see also Section 3.1.3). There are three main entities in mobile advertisement [180]: publishers, advertisers and the AA network. Publishers are the entities that want to have advertisement displayed to relevant users and therefore contact the advertiser and the AA network. Advertisers create the actual ad that is displayed to the user, the *ad creative*. The ad-network provides the Advertisement Software Development Kit (AdSDK) that connects advertisers and app-developers. While the main purpose of the AdSDK is to display ads in an app, it may also modify the creative. For example, AdMob adds buttons to let users turn off interest-based advertisement or to report offensive ads and may add trackers to check whether the ad was being displayed on the user side.

AA networks are faced with two major challenges. First, being able to determine the most relevant ad to an app user and, second, being able to re-identify, i.e. to track, users, for example to count user clicks on ads. For the former, the more information the AA network possesses about the user, the better it is able to determine what is the most relevant target audience for an ad. The AdSDK typically requires access to the phone's location in order to serve geo-localized advertisement. The latter challenge turns out to be difficult in the mobile device eco-system. In the conventional web ecosystem, third-party cookies can be used to track users [141]. Recently, more sophisticated techniques for user tracking, such as *canvas fingerprinting* and *evercookies*, have been developed [1]. Mobile advertisement that is integrated in apps is presented in WebView instances. Since there are no cookies in such instances, AdSDK developers rely on persistent identifiers [114]. According to the Google Play developer program, Android apps are required to use the Google Advertising ID (GAID), which is available on Android phones with Google Play services and which can be reset by the user. On phones without Google Play Services, most apps need to use permanent identifiers, such as Android ID (created when the device is installed), International Mobile Equipment Identity (IMEI) (permanent), International Mobile Subscriber Identity (IMSI) (permanent) and Media-Access-Control address (MAC address) (permanent). The Internet Protocol (IP) address of a phone does not serve as a good identifier as it periodically changes and many mobile devices share the same IP via Network Address Translation (NAT).

3.1.3 Privacy

A mobile device contains a lot of *private data* of its user. This data is either added by the user herself, such as the address book or calendar entries; data of the device itself, such as the device's IMEI; or data of the device's sensors, respectively, such as the user's location. In the following we shall refer to such data as *sensitive data*, because it can typically be used to learn information about a mobile device user. For example, the IMEI can be used to re-identify users in online sessions (see Section 3.1.2) and GPS data can reveal various information about the respective individual (see Section 3.2). Note that this definition is not sensitive data in a legal meaning. In law, sensitive data is a well defined category for that a particular protection applies. Different countries have different definitions of sensitive data, but common examples include health information, ethnic or racial origin, religious beliefs, political opinions and sexual preferences [190, 192].

Naturally, some apps need to access and process sensitive data of the user. For example, an app for finding nearby friends needs to process location data of the user, a calendar app stores the user's calendar entries or a video-chat app needs to access the phone's microphone or video. However, the reality of the mobile device eco-system shows that apps may process and transmit sensitive data indiscriminately. This raises serious privacy concerns, because most apps gain enough location data to be able to profile app users [78]. This problem also exists because Android does not allow for fine grained privacy settings to give the user more control, such as allowing an app only to access location when not being at home. Around 50% of the most popular apps request location data [78].

Mobile Application Analysis

Reverse-engineering the app's source code followed by static analysis is a well-established method for analyzing how apps process and transmit user data. While this may constitute a violation of copyright, it is a common practice on researchers, nevertheless. Felt et al. [80] and Vides et al. [196] showed that apps, as well as third party libraries, request unnecessary privileges. Such over-privilege grants the app access to more sensitive data and thus leads to a potentially more severe privacy problem. The phone's identifiers, such as IMEI and IMSI, are among the most commonly used sensitive data that apps access as shown by Enck et al. [75]. However, apps also transmit other data, such as the user's location [136, 169], call logs [101], account information [101] and the user's contact list [136].

A further analysis of the apps source code shows that not all types of third-party libraries equally collect sensitive data. Instead, AA libraries turn out to be a major collector of sensitive data [30, 81, 182]. They either probe for certain permissions during runtime of the app [75], which allows them to access sensitive data, or, as Book et al. [30] showed, AA libraries require over time more and more privileges themselves in order to work within an app. Hence, it is also possible that an app requests location information only due to the AA library that it includes, but not because the actual app needs it for its service. It is not very surprising that AA libraries process so much sensitive data. The more data a mobile AA network is able to process, the better it is able to predict user interests and thus becomes more attractive for advertisers. Note that this is not different to behavioral ads in the WWW.

Analyzing the app's source code alone does not suffice to quantify to what extent sensitive data is being accessed. Dynamic analysis tools actually run the apps and thus may be identify such dynamic behavior. TaintDroid [74] by Enck et al. uses taint analysis for the analysis and tracking of sensitive information flows within the app. Among 30 popular Android apps they found many cases in which the personal information of users, such as phone number or location, is collected and transmitted to remote destinations. Other studies using dynamic analysis tools have shown that a substantial amount of app-traffic is unencrypted and contains sensitive information such as users' location or real identities [59, 182, 205]. Wei et al. [200] found that apps send unencrypted traffic to many remote destinations. Furthermore they found that the remote destinations commonly belong to third parties. In the near future we can expect that the amount of unencrypted traffic being sent by apps declines as the use of Hyptertext Transfer Protocol Secure (HTTPS) increases. For example, Apple has decided to make the deployment of the HTTPS protocol mandatory as app communication protocol [52].

Dynamic app analysis is also capable of identifying more sophisticated techniques of apps to access sensitive data. Apps may conceal their accessing of sensitive data by not doing it in their own code, but by leaving this task to code that is dynamically loaded [51]. Furthermore, Feng et al. [82] have shown that apps may circumvent sandboxing by using Inter-Process Communication (IPC) calls or SD-card memory to exchange data. This way, apps may combine their privileges to learn data about the app users. For example, considering two apps where one is able to access the Internet and the other one the user's accurate location. Neither of them alone may leak sensitive information, but they could combine their privileges to transmit location data of the user to a remote destination. Sooel et al. [180] have shown that ad creatives themselves may access sensitive data from the user. This is possible because AA networks give ad advertisers access to Mobile Rich Media Advertisement Interface Definitions (MRAID), an

interface that allows ad creatives to be written in Hypertext Markup Language (HTML) and to call a limited set of JavaScript functions. This allows the ad creative to track the user's whereabouts by having the AA network send the user's current location along with a unique identifier to a remote machine under control of the advertiser. The authors further show how MRAID can be exploited to learn a user's medication from the *GoodRx* app; the gender preference from *POF Free Dating App*; and the user's browser history of the web browser *Dolphin*.

Feng et al. [82] revealed that 96% of the top free apps request Internet access and at least one persistent identifier. This allows a third party to link two random apps with high probability. This is especially a problem in the case of AA libraries, because persistent identifiers allow them to aggregate different app sessions and thus get an insight in the user's life that she is unaware of and from which she has no means to opt out.

Android Location Privacy

Fawaz and Shin [79] propose *LP-Guardian*, a modified Android version that is specifically designed for protecting the user's location data. Whenever an app requests location data, LP-Guardian obfuscates the user's current location in the background. The type of obfuscation depends on two main factors. First, how sensitive the location is that the user visits. Therefore, LP-Guardian is able to determine sensitive places, such as the user's home location; it depends on user input to set the sensitivity level of the current location. Second, LP-Guardian takes into account the accuracy of the user's location that is needed by the app. The authors find that the city level suffices for 68% of the apps to provide their service. Another key feature of LP-Guardian is that it is able to generate fake routes for apps that require continuous location tracking. For certain apps, such as fitness apps, fake routes that resemble the actual speed of the user are sufficient to provide valuable service quality. LP-Guardian also recognizes location requests from AA libraries and ensures that the same AA library obtains the same obfuscated location irrespective of the app in which it is running. This way LP-Guardian guarantees that no AA library is able to obtain a better location estimate on the user due to tracking in multiple apps.

Fawaz et al. [78] propose *LP-Doctor*, a system that also aims at restricting the app's access to location data while maintaining the app's required functionality. The key feature of LP-Doctor is that it is not necessary that the user installs a modified version of Android, as required by LP-Guardian. Instead, LP-Doctor is a modified version of an *app-launcher* that an Android user may install in a similar way as an ordinary app. The task of an app-launcher is to start the

app after the user clicked on the app's icon. LP-Doctor takes advantage of this and obfuscates the user's location with Laplacian noise [7] if the user starts an app that requests location data and if it is necessary according to the internal privacy metric. In order to decide whether the current app call poses a privacy threat, LP-Doctor considers the user's input at the installation time of the app and, additionally, tracks the user's movements. The latter enables LP-Doctor to reuse pseudo-locations if the user starts the same app at a privacy sensitive place. This guarantees that an app, that the user starts frequently from the same, sensitive place, is not able to average out the Laplacian noise over time as described in [7]. The main disadvantage of LP-Doctor is that it fails to protect the user against apps that request the user's location while running in the background.

3.1.4 User Studies

Although behavioral advertisement is a well-established business in apps, there is not much research on how users feel about having their location data being used for behavioral advertisement. However, studies on behavioral advertisement for regular WWW sessions suggest that the majority of people reject such technologies. It is not only that users oppose that their sensitive data is processed for marketing reasons, they also oppose being tracked while surfing the WWW. Tracking is a key technology for Online Behavioral Advertisement (OBA) as it enables so called *trackers* to re-identify users in different WWW sessions and thus to learn as much as possible about their activities and their interests, respectively. The most traditional way to re-identify users is with the help of Hyptertext Transfer Protocol (HTTP)-cookies. Since WWW browsers allow informed users to delete HTTP-cookies, they have some sort of control over this form of re-identification. However, more recent technologies, such as cookie respawning, Flash cookies or fingerprinting [1, 112], leave the user with no countermeasure in order to escape tracking. Even worse, these technologies are hardly noticeable.

Considering that tracking is a technology that happens invisible in the background, it is not surprising that users have a poor understanding of OBA and that they are very concerned about how their data is being used within OBA. This is shown by a study of Ur et al. [193]. They demonstrated that users confuse anti-malware software with a tool to protect against OBA or the users believe that there exists a browser setting that protects them against the data collection due to OBA. Even the notice and choice mechanisms that the ad industry provides are misleading to the users and are poorly understood. If an OBA provider offers an opt-out mechanism, users are typically not aware of it. The poor understanding of OBA is not limited to typical users, who tend to

have little knowledge about information technology. Instead, even technically savvy people do have limited knowledge of third party tracking techniques as a user study by Agarwal et al. has shown [2].

Besides the lack of transparency and understandability of OBA, a study by Fawaz et al. [78] revealed that the behavioral user profiles have poor accuracy and are not, as promised by the companies, anonymous. In a first step, Fawaz et al. conducted an in-person interview to inspect together with the users their own behavioral profile that was created by some OBA companies. These interviews revealed that OBA companies store wrong data about their users. In a second phase, Fawaz et al. identified that users are mostly concerned about: how their data may be used, the level of detail of their data, the aggregation and the amount of sensitive data that is being collected.

Most people oppose that their private data is collected and processed for marketing purposes. In the study of Turow et al. [189], 66% of the users dislike private data analysis for advertisement. Interestingly, when the authors informed the study participants about common ways that marketers gather data, rejection increases to 73% - 86%, depending on the particular way that data was being collected. In their study, Turow et al. also did not find much support for tracking anonymously as it is also rejected by the majority (68%). Most participants of a study conducted by McDonald et al. [142] consider OBA to be privacy invasive and thus believe that it violates their fundamental rights. The study of Ur et al. [193] found that it depends on the context whether people accept being tracked. More participants tend to accept tracking practices, when they are planning a vacation or when they are shopping for a car. However, almost none of the participants accept any form of tracking when they are engaging in rather sensitive behavior, such as researching sexually transmitted diseases. Another, more practical, concern has been studied by Agarwal et al. [2]. They show that people are also afraid of embarrassing ads being displayed on their screen when they are next to other people. Similar to the study of Ur et al. [193], their study participants state that their concerns regarding tracking for OBA purposes depend on the topic, i.e. the context.

Some works also investigate possible ways to improve the situation of OBA and study how it could be designed in a fairer way. Kelley et al. [121] realize that their study participants have complex and unique privacy preferences. Furthermore, they find that choice is key for the users. The more users are able to express their privacy preferences, the more they are willing to share data and the more they are willing to have their data being processed. This is in line with the study of Agarwal et al. [2] where the participants stated that they wish to be able to block certain OBA ad topics. Fawaz et al. [78] find that, although users in general appreciate the opportunity to investigate their own OBA profile, the effect of editing needs to be clearer to the users. Furthermore

OBA companies should elaborate on how the data was collected or inferred, respectively, and they should also elaborate on how the collected and inferred data is being used [78, 165].

3.2 Threats

As we have seen in Section 3.1.3, the design of the mobile device eco-system gives many parties access to user location data. This raises significant privacy concerns with respect to the LBS provider and third-party software developers, such as AA network providers. However, there are further entities that learn location data which may raise separate privacy concerns. In many LBSs the user shares her location with other users of the LBS. Furthermore, operators of the communication infrastructure can also learn user location data if it is being transmitted unencrypted. Likewise, surveillance agencies that gain access to the communication infrastructure may be able to learn user location data.

Clearly, the scenario in which an entity learns user location data may differ and, along with that, also the necessary countermeasures. For example, if users are able to set within the app of the LBS the other users with whom they are willing to share their location data, then this may solve privacy concerns with respect to the other users in the system. However, this would still not solve privacy concerns towards all the other entities that may learn user location data. Instead of providing detail on particular scenarios where entities may learn user location data, in the following we provide an overview why the protection of location data is difficult and an overview about the threats to a user's privacy once an entity has gained access to location information.

Several researchers observed that anonymization of location traces is hardly possible, because the traces of users are unique. As Bettini et al. [24] showed, the location traces of a person is a spatio-temporal pattern that is almost unique to a user. Therefore, pseudo-anonymization does not offer any protection of user traces, because the pseudonym allows to reconstruct user traces. This may still allow to infer home/work addresses of the user or, if location traces that are not pseudo-anonymized are available, both data sets can easily be correlated. De Mulder et al. [65] showed that the adversary is able to compute a Markov model from location traces that serves as a mobility profile. This mobility profile is sufficiently unique that it allows to recognize pseudo-anonymized location traces. Ma et al. [139] employ a maximum likelihood estimator, minimum square approach and a weighted exponential approach to re-identify pseudo-anonymized traces even if the traces are obfuscated with Gaussian or Normal noise. De Montjoye et al. [64] analyze the movement of users among Global System for

Mobile Communications (GSM) base stations and find that movement patterns are highly unique as four location and time points suffice to uniquely identify 95% of all people.

Analyzing location traces provides a lot of sensitive information about the respective people. The user's home and work address are among the most easily identified locations. This is particularly concerning, because, as shown by Golle et al. [100], knowing the home/work region reduces the anonymity set of a user dramatically. Hoh et al. [110] show that it is relatively easy to infer the user's home and other places of interest when the user resides in traffic monitoring applications. Ashbrook et al. [18] show that by analyzing the GPS data of user movements it is possible to infer meaningful places for a user and to predict the user's movement. Since GPS data is noisy, they employ a variant of the k-means algorithm to cluster the GPS data points. Every cluster must be a place where the user resided for some time and thus be of importance to the user. Additionally, this data reveals how much time the user spends in a certain building, potentially revealing additional information about the purpose of the user's stay. Furthermore, Ashbook et al. build a Markov model where the nodes are a meaningful location and the transition probabilities are retrieved from the user's movements between meaningful locations. This model allows to predict user movements. Freudiger et al. [88] propose an algorithm that allows to identify users and to reveal their interests when having access to their GPS data. In the first stage of the algorithm, they employ, similarly to Ashbrook et al., a k-means algorithm. In the second stage of the algorithm they label the GPS cluster according to a pre-defined set of rules. For example, clusters between 9 am and 5 pm are likely to indicate the working place and places where the user stayed overnight are likely to be the home address. Their analysis shows that an entity having access to accurate location information is able to infer sensitive information about the users. In particular, they investigate the success probability of identifying home, work, home/work places depending on how frequent the user queries the LBS. Furthermore, they show that the anonymity set for a home, work and home/work locations is rather low and that the 10 most meaningful locations for the user can easily be revealed. Krumm et al. [127] show that even if the location trace of the user was created by a device that the user does not carry around, such as a car tracker, the home address of the user can be identified with several heuristics.

The above inferences are also possible even if the adversary has only pieces of the location trace at hand, i.e. an incomplete location trace. As shown by Hoh et al. [109] Kalman filters can be used to track the user with high accuracy [109]. Liao et al. [135] also employ Kalman filters in order to identify, beside home/work locations, the user's transportation type and when the user deviates from the daily routine. Olteanu et al. [154] show that even if the user

does not actively reveal her location, friends may reveal her current whereabouts using geo-localized tags in an online social network.

Polakis et al. [161] investigate the location-privacy guarantees of popular LBSs. Particularly, they analyze applications, such as *Facebook* and *Grindr*, that allow for approximate nearby information. Modifying the client app allows an adversary to localize other users with an accuracy of a few meters. While this is more difficult, even users who are traveling can be localized with an accuracy of a few hundred meters. Riederer et al. [168] propose *FindYou*, a system that collects user location data directly from the LBSs, such as Foursquare and Instagram, and confronts the user with inferences that the system makes about the user. Possible inferences include: home location, race and income level. Theodorakopoulos et al. [185] show that even if the user employs a location protection mechanism, the level of protection degrades the more often the user obfuscates the same location

3.3 Legal Protection of Location Data

There are several directives in place for the protection of location data of which we discuss in detail the DPD, the ePrivacy Directive (2002/58/EC, as amended by 2009/136/EC) (ePrivacy Directive) directive and the recently published GDPR.

The Data Protection Directive (95/46/EC) (DPD) [190] defines the legal rules for processing personal data. The term *processing* is very broad and can be anything from recording, handling and deleting data. The term *personal data* refers to "any information relating to an identified or identifiable natural person (data subject)", whereas the identification can be direct or indirect. Direct means to identify an individual without third party data sources. For example, the address of an LBS user typically directly identifies the individual and is thus personal data. Third party data sources are necessary for an indirect identification. This typically applies to all kinds of pseudonyms, such as employee numbers or IP addresses. Although an employee number alone does not necessarily identify an individual, in combination with the employer's database that maps employee numbers to names and addresses, an individual can be identified. The DPD is applicable if the data controller either has an establishment in one of the EU Member States or makes use of equipment situated on the territory of a Member State. Hence, the DPD also applies if the data processor resides outside the EU, but offers its service through equipment that is situated within the EU. Requirements for data processors with respect to the data quality are also defined in the DPD. Data must be processed fairly and lawfully and data may

only be collected for "specified, explicit and legitimate purpose". Furthermore, the data needs to be "adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed". Furthermore, the data controller must implement sound security measures in order to protect data from being corrupted, destoryed, or disclosed in an unauthorized way. The DPD grants the user considerable rights with respect to learn how her personal data is being used. Particularly, the user is "... to be informed that processing is taking place, to consult the data, to request corrections and even to object to processing in certain circumstances". Since third parties may process the user's data, the DPD states that the data subject is to be informed "... when the data are first disclosed to a third party". In any case, the data subjects have the right to access, rectification, erasure, blocking and the right to object. If a data processor engages in unlawful processing of personal data, then the DPD states that the member states of the EU have to put in place effective sanctioning mechanisms. The DPD obliges the data processor to apply protection of personal data as long as it is not anonymized.

The ePrivacy Directive [191] defines additional rules to the DPD. While the DPD is solely concerned with protecting the personal data of natural persons, the ePrivacy Directive is part of a larger set of directives regulating the electronic-communications sector and applies to the sector of electronic communications as well as to natural persons. The ePrivacy Directive [191] defines a regime of different rules depending on whether location or traffic data is processed. Within the ePrivacy Directive location data is any data that indicates "... the geographical location of a terminal equipment of a user..."; and traffic data refers to "... any data processed for the purpose of the conveyance of a communication ... or for the billing thereof". Whenever traffic and location data is considered to be also personal data, then both regimes, the ePrivacy Directive as well as the DPD, apply. In this case the ePrivacy Directive prevails over the DPD. Regardless of the service, the ePrivacy Directive lays down a provision with regard to the processing of location data. The data needs to be made anonymous and the user needs to give her prior and informed consent to the processing of her data. The user should be able to withdraw her consent temporarily with simple means and free of charge. Furthermore, the data processor may only process the data as long as necessary for the provision of the value added service.

The Article 29 Data Protection Working Party (WP) [63] is an independent body with advisory status under the Article 29 of the DPD. While the WP does not have any legislative power, its opinions and recommendations on data protection have considerable impact. The members of the WP are representatives from each EU Member State, representatives of the EU institutions and one representative of the European Commission. The WP has issued two opinions

and recommendations that are directly related to LBS that are very helpful to understand the European data protection regime on location data.

In opinion WP 115 [16] the WP concludes that location data always relates to an identified or identifiable person and thus location data shall fall within the scope of the DPD. Hence, providers of LBSs fall under the DPD regime. The opinion WP 185 [17] establishes that the ePrivacy Directive applies only to a certain type of providers: providers of *public electronic communication services and networks*, which are telecommunication providers [58], fall under the scope of the ePrivacy Directive, while providers of LBSs on mobile devices do not. Telecommunication providers fall under the regulation of the ePrivacy Directive regardless if they only process location data that is base station data or if they also process location data that is WiFi or GPS data (WP 185, p. 8). The WP particularly excludes LBSs on mobile devices, such as Google Maps or Foursquare, from the ePrivacy Directive and considered them to be *information society services*. This decision has been criticized by Cuijpers and Pakárek [58] because it puts the telecommunication provider under a stronger regime than LBS even if the same data is being processed.

Besides the issue of which service provider falls under which regime, the WP also provides clarification on the matter of anonymization and consent. The WP does further discuss several additional issues relating location data. Although this does not directly address information society services, it is rather likely that these statements also apply to them. First, the WP acknowledges that that "true anonymisation is increasingly hard to realise and that the combined location data might still lead to identification" (WP 185, p. 18). This has considerable impact, because under the DPD the data processor needs to implement protection for non-anonymous data. Furthermore, the WP elaborates on legitimate grounds for the processing of location data. These include the following two main points (WP 185, p. 19). First, the consent, necessary for the processing of location data, cannot be obtained through general terms and conditions. The consent needs to be specific, it should by default only be valid for a certain period and it needs to be very easy for the data subject to withdraw their consent. Second, LBSs should be turned off by default.

As already mentioned in Section 2.2, the new GDPR was approved in May 2016 and will replace the DPD in May 2018. The GDPR was initialized in 2012; the GDPR intends to address current and future information technologies and thus make Europe fit for the digital age. The GDPR adds additional elements, such as the transparency principle, the clarification of the data minimization principle and the establishment of a comprehensive responsibility and liability of the data controller. It, furthermore, provides additional information to the data subject, including the storage period. Also, the data subject has several additional rights: the right to lodge a complaint in relation to international

transfers, a *right to be forgotten*, a right not to be subject to a measure based on profiling, and a right to have her data transferred from one electronic processing system to another one. Data processors have to employ data protection by design, by default and to carry out a data protection impact assessment prior to risky processing operations. In the case of infringements, this stipulates judicial remedy obliging the supervisory authority to act on a complaint and it includes considerable penalties and sanctions. Additionally, the GDPR includes that there are data processors that are beyond the controller's instructions and addresses their obligations as a *joint controller*.

The DPD, the ePrivacy Directive directive and the GDPR all have in common that they also define exceptions of their protection. Typical cases in which the legislation is restricted is for measures to safeguard national security, defense, public security, important economic interests and law enforcement. This is related to the Data Retention Directive (Directive 2006/24/EC) (DRD) that accounts for the value of traffic and location data (in the sense of the ePrivacy Directive) for the investigation, detection, and prosecution of criminal offenses. The DRD was supposed to regulate the duration of which telecommunication providers have to store traffic and location data of their customers, which they have to provide on request of police and security agencies. On 8 April 2014 the Court of Justice of the European Union declared the DRD adversely affects "... the essence of the fundamental rights to respect for private life and to the protection of personal data" [151] and thus nullified the DRD. However, it is not clear what this means in practice. For example, the United Kingdom has passed in 2014 a new Data Retention and Investigatory Powers Act [124].

3.4 User Perception of LBS

As we saw in Section 3.1.4, users mostly oppose the current practice of Online Behavioral Advertisement (OBA). In this section we report on whether users appreciate LBSs in general. A series of user studies has investigated the attitude of users towards LBS and the respective threats to their privacy. Overall we can draw the conclusion that users appreciate LBS, but that they are concerned about their privacy and would like to stay in control of their location data. Table 3.1 provides information on the participants of the available studies. While overall the studies provide some evidence on the users' preferences, the so called privacy-paradox [126] also needs to be considered. The privacy paradox is widely studied and describes the gap between the privacy attitude of users and their actual privacy behavior. Typically, when being asked, users express rather large concerns regarding their privacy but, in practice, rarely hesitate to share private information for a small compensation, such as a small monetary

Table 3.1: Recruitment information of user studies

Work	# Users	Profession	Age
Barkhuus et al. [21]	16	University	mean 23.7
Tsai et al. [188]	587	Various	> 18
Bilogrevic et al. [25]	35	University	20 - 38
Egelman et al. [72]	493	Various	> 18
Toch et al. [186]	25	Various	20 - 40
Consolvo et al. [53]	16	Non-technical	unknown
Iachello et al. [115]	11	Various	2 families (52 - 14)
Anthony et al. [12]	30	Undergraduates	mean 20
Brush et al. [35]	32	Various	12 households
Tsai et al. [187]	123	University	unknown
Almuhimedi [4]	23	Various	18 - 44

reward [41,104]. Kokolakis finds that there are four main reasons for the privacy paradox: i) users perform calculus between the expected loss of privacy and the potential gain of disclosure; ii) users have no choice but revealing their data in order to participate in online social life; iii) users are subject to cognitive biases and heuristics, such as the optimism-bias, overconfidence and perceived benefits; iv) users have no ability to make informed judgment about trade-offs in privacy decisions. Zafeiropoulou et al. [209] find in their study that in LBS there is no strong correlation between users' privacy preferences and their actual behavior. This shows that the privacy paradox also exists in LBS.

3.4.1 Overview on User Studies

Several works have opted for a questionnaire as the main body of their user study: Barkhuus et al. [21] interviewed users about the perceived usefulness of hypothetical LBSs, such as friend nearby notification; Tsai et al. [188] conduct an online survey of American Internet users to investigate when they would be willing to share their locations; Bilogrevic et al. [25] use a questionnaire in order to learn user opinion about their private LBS. Egelman et al. [72] show users pictures about different LBS that have the same functionality but differ in their privilege requirements in order to learn the users' preferences; Toch et al. [186] recruit people directly via a people nearby LBS and conduct an interview on how they use those applications.

Other studies equipped users with some sort of device that allows to reach the users in their daily routines. The main point of these studies is that the study participants can be reached in real situations at random times, asking them

about their whereabouts and whether they would like to share their location with family, friends, colleagues or strangers. Consolvo et al. [53] provide the users PalmOS devices, Iachello et al. [115] people receive questionnaires on mobile phones and Anthony et al. [12] provide users with beepers.

Brush et al. [35] collect GPS traces of the study participants from 12 households. After the collection phase the authors show the participants the data and ask them whether they would be willing to share the location data and whether protecting the location data with certain obfuscation methods would increase the participant's willingness to share the location data.

In a later study, Tsai et al. [187] are using a Facebook location-sharing application that the study participants can use on their own laptops. The app allows the participants to define time-based rules that determine when they are willing to share locations among their friends.

In the study of Almuhimedi [4] the participants have an app directly installed on their smartphone. The app monitors how frequently the other apps, which are already installed on the participants' device, request sensitive information, such as the participants' location. In the next phase, the authors send nudges to the users to have them limit the privileges that are granted to the other app.

3.4.2 User Attitude Towards LBS

Users mostly embrace LBS as a technology and consider it to be useful. Furthermore, users are concerned about their location data in general, but privacy concerns are bigger when location data is being used for commercial services [121]. In the case of Barkhuus et al. [21], the participants of the questionnaire consider especially services, such as friend nearby notification or turning the phone in silent mode when being for example in the cinema, as useful. The study of Consoolvo et al. [53] and Egelman et al. [72] show that users appreciate LBS for sharing location data with their friends and that they are willing to share accurate location data among each other.

Tsai et al. [188] provide a more comprehensive overview on the services considered useful by the study participants. These include: finding people in an emergency, keeping track of children and family, point-of-interest finder, finding nearby friends, carpooling, allowing users to keep track of their own activities and finding new people.

Toch et al. [186] show in their study that location is a driver for users to understand other users, to gain trust, to convey trust, to feel secure and to filter

other users. Note that location does not mean local. Knowing their location helps to increase trust, even if the other people are far away.

3.4.3 Privacy Preferences of Users

Users are concerned about what happens when they reveal location data. Tsai et al. [188] show that users are afraid that they may expose their home addresses or may enable other people to stalk them when using LBS. Also, the study of Brush et al. [35] reveals that the participants associate certain risks with a service that shares location data. However, the study finds that user data obfuscation could reduce the perceived risk.

The fact that users are concerned about their privacy is illustrated in the studies of Bilogrevic et al. [25] and Egelman et al. [72]. Bilogrevic et al. show on the example of an application, which finds a fair meeting point among a group of people, that the participants prefer if the group members do not learn the location data of all the other users. Egelman et al. let people choose between different price/permission variants of apps where the app is the more expensive the less permissions it requires. Their results indicate that people are willing to pay more for a mobile application if it requires less privileges. In particular, a quarter of the participants is willing to pay the highest premium of $1.50 for the app that only requires the Internet permission and 40% of the participants are willing to pay a premium for an app that does not request access to their location.

In the 2003 study of Barkhuus et al. [21] the majority of people has no concerns regarding revealing location data. The study may have this outlying result due to a small number of participants or perhaps, given that the study was conducted prior to the time at which privacy incidents happened on a regular basis, the study simply reflects that people have been less privacy concerned in the past. In any case, the participants in Barkhuus et al. study are more comfortable with LBS if the location data does not leave their phone.

Iachello et al. [115] show that participants would not like their location data being transmitted in an automatic fashion. Instead, the participants like to be in control of their location data and they would like to decide every time their location data is being requested. Furthermore, users do not want a third party or strangers to receive their location data, but only the intended receiver. If the participants are in control of their location data and get to decide to share their location data in case-by-case scenarios, then they see little use in deception or lying about their whereabouts.

The study conducted by Consolvo et al. [53] investigates the factors

that determine the participants' location sharing behavior in more detail. Unsurprisingly, the main drivers in this behavior include the person with whom the location is to be shared, the place where the participant is located and the participants' current mood. Furthermore, if the request does not match the current context, for example if colleagues request location information when the user is in a private context, then people will likely reject the request for location information. This is in line with the results of Anthony et al. [12] and Tsai et al. [187]. The former shows that that people have different sharing behavior depending on their current place and social context. In the latter study the participants would like to define location-based and group-based sharing rules and they are generally unwilling to share their location with strangers. Only the participants in [35] indicate that they are also willing to share location with the service provider in order to have their service delivered.

Almuhimedi et al. [4] investigate the usability aspect of mobile applications. They reveal that permission managers are essential for users to adjust their preferences and that nudges can increase the efficiency of the permission managers.

3.5 Contribution

Our work in [195] contributes towards an analysis of how apps process sensitive information, such as device identifiers and user locations. In this sense our work mainly contributes to Section 3.1.3 of this chapter. In particular, we analyze the extent to which apps enable global surveillance. Snowden has revealed the extent of global surveillance by revealing a large number of programs run by the intelligence agencies from the US and the UK. The Tempora program [103] analyzes global Internet traffic combined with many other sources. The Badass program [156] analyzes unencrypted app traffic.

Our framework downloads Android apps and executes them in an automated fashion and records the traffic that is being sent by the apps. Subsequently, our framework analyzes the recorded traffic as it would be done by a surveilling adversary. While the framework employs several tools, such as the Android UI/Application exerciser *Monkey* [8] and the network traffic dumping tool *dumpcap* [204], we developed the core functionalities of the framework from scratch. The framework can be separated into the *setup* and the *analysis* part. The setup part consists of software to parse and download apps from Google Play, to automatically execute apps and to record the app traffic. In the analysis part of our framework we developed software modules that extract relevant information from the recorded traffic and software that evaluates the

success of a surveilling adversary. The latter includes a similar technique to the one proposed by Engelhardt et al. [76] as well as a novel technique based on Transmission Control Protocol (TCP) timestamps. Overall, our framework is the first to quantify the success of a surveilling adversary on mobile app traffic.

We employ our framework in an analysis of 1260 Android apps. Our analysis shows that up 57% of the unencrypted mobile app traffic is linkable. Third party software, such as AA software, turns out to be the most frequent sender of unique identifiers. Furthermore, our analysis shows that passive network fingerprinting techniques, such as processing TCP timestamps, increases the efficiency of the adversary, because TCP timestamps can be used to link unencrypted traffic even if no unique identifier is transmitted. Furthermore, we evaluate the extent to which ad-blocking tools hamper the adversary's ability to link user app sessions. While these tools have not been designed to protect against a surveilling adversary, they seem to be the best protection mechanism available, since they prevent traffic from AA libraries. Our analysis shows, however, that the two tools that we analyze, *Adblock Plus for Android* [99] and *Disconnect Malvertising* [54], have only a very limited effect on the efficiency of the surveilling adversary.

3.6 Conclusion

Mobile devices and their smart applications have become ubiquitous. One of the most popular mobile services are LBSs. They have proven to be useful and studies have shown that users appreciate the functionality such as location-sharing or Point-of-Interest (POI) finder. Unfortunately, it turns out that most LBS are privacy invasive. This is mostly because the current market rewards companies that collect as much user information as possible, as this data can be used for behavioral advertisement or could be monetized in another way. However, privacy issues also exist due to platforms that have not been designed with privacy in mind.

Overall, the current practice of LBS has resulted in a massive collection of private user data in general and location data in particular. While, location data is commonly considered to be personal data, it is subject to protection under the scope of the DPD. Therefore, any processing of location data needs a legal ground according to the DPD and the location data needs to be deleted or anonymized if there is no longer a legal ground or the purpose has been exhausted. However, several works have shown that it is difficult to anonymize location data, because movement patterns of individuals tend to be unique. This

leaves the user in the inconvenient situation of leaking accurate location data while engaging in LBSs that allows various entities to infer sensitive information.

Chapter 4

Design of Private Location-Based Services

Most LBSs for mobile devices are privacy invasive, because the service provider learns the user's location data. A variety of services have been proposed that make this privacy invasion impossible. Such services employ cryptographic primitives that allow them to provide the necessary privacy guarantees. One of the most commonly used cryptographic primitives (see for example [25,162,210]) is homomorphic encryption that allows computations to be carried out on ciphertexts. There exists several cryptosystems that provide this homomorphic property. For example, the *Paillier* [158] or the *Boneh-Goh-Nissim* [29] cryptosystem possess the *additive homomorphic property*. Given two plaintexts m_1 and m_2 and their respective encryptions $E(m_1)$ and $E(m_2)$, the following equation holds:

$$E.(m_1) \odot E.(m_2) = E(m_1 + m_2)\,,$$

where $\odot$ is an arithmetic operation in the encrypted domain that corresponds to the integer sum of operation on the plaintexts.

While providing strong protection, private LBSs are typically tailored to a very specific use-case and thus cannot be reused for other services. Furthermore, most of the designs for private LBSs assume the service provider cannot be trusted and users do not share their location information with the service provider. Instead, they assume the service provider to be *honest-but-curious*. This assumption accounts for the commercial interest of service providers in practical scenarios that makes them honestly follow the private protocol. However, the service provider is still interested in learning as much as possible about its users.

Table 4.1: Comparison of private GSN.

Work	Building Block	Workload	Comments
Freudiger et al. [86]	Hybrid encryption	User	No TTP when using a DHT
Dong et al. [69]	Proxy re-encryption [28]	LBS	—
Carbunar et al. [40]	Dedicated protocols	Both	Requires extra hardware
Puttaswamy et al. [164]	Coordinate Transformation	Both	Requires obfuscation
Herrmann et al. [107]	Broadcast Encryption	User	Private statistics

Therefore, he may infer patterns in user behavior or employ *side information* that he obtained, such as historical data about the users [174] or information from the user's friends [154].

In the following we outline private alternatives for the most commonly used LBSs. Geo-Social Networks (GSNs) combine functionality of LBSs and online social networks. Another very popular LBS is friend-nearby notification in that users are automatically notified if a friend of them is in close proximity. Point-of-Interest (POI) finder allow users to find interesting places, such as sights or restaurants. Finally, traffic-monitoring is designed to provide intelligence about the traffic situation and thus route traffic in a more efficient way.

4.1 Geo-Social Networks

Since GSNs, such as Foursquare, are particularly popular among the users of LBSs, a substantial effort has been invested to develop privacy-preserving counterparts. Works, such as Freudiger et al. [86] and Dong et al. [69], propose private location-sharing services, which is the core functionality of any GSNs. In such a service two users, Alice and Bob, wish to either mutually or unidirectionally share their locations. Carbunar et al. [40] and Puttaswamy et al. [164] propose services that provide more GSN functionality, such as checking-in or leaving recommendations of venues. Table 4.1 provides an overview of these systems. For comparison, the table also includes our work in Herrmann et al. [107] as location sharing is a fundamental building block of a GSN.

Freudiger et al. [86] proposes a private location-sharing service and employ hybrid encryption to protect the user's privacy. Every user is assumed to

have a public/private key pair and users can exchange their public keys either out-of-band or with the help of the service provider. If Alice wants to inform Bob about her current location, she chooses a secret key, encrypts her location with the secret key, encrypts the secret key with the public key of Bob and sends both to Bob. The receiver Bob uses his private key to decrypt the symmetric key, enabling him to decrypt Alice's location. The data being sent between Alice and Bob could be transferred with the help of a central service provider, but if the service provider is reluctant to store and forward encrypted data, Alice and Bob could exchange their information with the help of a Distributed Hash Table (DHT), such as [183].

Dong et al. [69] argue that a solution using pairwise secrets or a public/private key infrastructure is not practical. Their location-sharing service is designed such that users only store their own keys in their devices and no other keys are needed to participate in the service. Furthermore, their protocol is designed to be lightweight on the user device. The proposed solution is based on *proxy re-encryption* [28]. In such a scheme a proxy function is employed to convert a ciphertext for a particular key into a ciphertext for another key without revealing any information on the key or the plaintext. If Alice wishes to share locations with Bob, then Alice retrieves Bob's public key from the service provider and issues a proxy re-encryption key for Bob and stores it at the service provider. The actual location exchange is done with the service provider as a third party that performs the proxy re-encryption. Bob, wishing to obtain an update on Alice's location, queries the service provider for her location. The service provider checks if Bob is allowed by Alice to learn her location, performs the proxy re-encryption and sends the ciphertext to Bob who is able to decrypt. The proposed scheme allows Alice to specify the precision of the GPS coordinates and thus the granularity of the location that Bob learns. Using the service provider as third party also comes with the advantage that it can enforce revocation. If a user revokes a friend's privilege to learn her location, the service provider no longer performs the re-encryption. The proposed service can run on low end smartphones, because the computationally most expensive operation, the re-encryption, is carried out by the service provider.

Carbunar et al. [40] propose a private alternative to a GSN, which provides similar functionalities to Foursquare. The service provider does not learn location information, but only ensures correctness of the protocol. Every communication between the users and the service provider happens via an anonymizing network, such as Tor. The core of the system is a system called `SPOTR` that allows the GSN provider to certify the location of the users. `SPOTR` relies on a dedicated piece of hardware at a venue that is able to perform simple computations and to display a Quick Response (QR) code. This QR code encodes a signature of the venue and a time of validity. Users scan this QR code and send it to

the GSN provider. If the sent QR code contains a valid signature and correct validity period, the GSN provider confirms the user's check-in with a signature. A protocol named *GeoBadge* extends SPOTR to achieve a functionality allowing users to prove that they have been present at the same venue a certain number of times. In GeoBadge the user sends anonymously with every check-in a nonce that is signed by the GSN provider. When the user wishes to verify her k check-ins, she presents the k signatures to the GSN provider who replies with the respective signed token. The protocol *GeoM* extends GeoBadge to facilitate the *mayor-functionality* of Foursquare. The key difference is that check-ins happen now in discrete time intervals, i.e. epochs, and that users prove to the GSN provider that they have checked in at a certain epoch. The proof is zero-knowledge in the sense that the user proves that she knows values that the service provider issued during a check-in in a specific epoch. The user who can prove the most check-ins receives a signed *mayor-token*. Finally, *MPBadge* extends GeoBadge to implement a functionality allowing a group of l users to prove they have checked-in at the same time at a certain venue. In this protocol, every of the l users checks-in independently and obtains an additional signed value. The GSN provider issues a MPBadge only if one user in the group is able to present a value being computed as the result of all l additional signed values. The implementation and analysis of the protocols shows that the proposed schemes have reasonable computational demands. While indeed protocols such as the GeoBadge and MPBadge provide similar functionality to services such as Foursqure, the system by Carbunar et al. does not allow users to privately share their location among each other.

Puttaswamy et al. [164] propose another privacy-preserving GSN that is very efficient: the only operations that users have to employ are encryptions and coordinate transformations. Every user in the system is assumed to have a secret which she shares with her friends. A user who wishes to store data for a particular place at the GSN provider, such as a check-in or a venue rating, first transforms the coordinates of the venue and stores them at the service provider. A friend who is interested in messages of her friends to a particular venue, transforms the venue's coordinates according the friends' secret and queries the service provider. Once the data is retrieved, the user is able to decrypt the friends' messages with the shared secret. While storing data at the GSN is a very efficient operation, retrieving data from friends is less efficient and mechanisms need to be implemented to avoid that the service provider learns additional information, such as a the set of transformed locations belonging to the same venue or friend relations. Puttaswamy et al. [164] propose several mechanisms based on obfuscation that offer a privacy/performance trade-off.

Table 4.2: Comparison of private Friend-Nearby LBSs. All systems require a TTP.

Work	Cryptosystem	Protection Lying	Location Probing
Zhong et al. [210]	Paillier [158]	None	Yes
Narayanan et al. [146]	PTSI [146]	Location Tags	No
Lin et al. [137]	PEqT [85]	Location Tags	No
Bilogrevic et al. [25]	Paillier [158] BGN [73]	None	No
Herrmann et al. [107]	Several: [22, 93, 125]	None	No

4.2 Friend-Nearby Notification

In friend-nearby notification services, users wish to be informed or to learn whether they are in close proximity. Similar to services that allow the exchange of location information, Alice and Bob engage in a private protocol and either one user or both learn at the end if they are nearby or, in some proposals, they also learn the exact location of each other. If they use a protocol that does not allow them to learn the exact location of each other, they could engage in a protocol such as the ones of Freudiger et al. [86] and Dong et al. [69]. A challenge of every friend-nearby notification service is that Alice or Bob may lie about their location and thus learn the other party's location without revealing their own location. This could lead to privacy violations, because Alice or Bob may only be willing to let people know their location if they are also in the same area. The proposed protocols are not equally suited to detect or prevent such misbehavior. We provide an overview on privacy-preserving friend-nearby notification LBSs in Table 4.2. Again, this table includes for comparison our work in Herrmann et al. [107] as location sharing can be used for a friend-nearby notification service.

One of the first private protocols for friend-nearby notification was designed by Zhong et al. [210]. They introduce three protocols, named *Louis*, *Lester* and *Pierre*, that are based on homomorphic encryption [158]. The Louis protocol requires a Trusted Third Party (TTP) for Alice and Bob to learn if they are nearby. This TTP does not, however, learn the locations of Alice and Bob. Furthermore, the TTP never communicates with Bob, but Alice relays the messages for Bob. The protocol either stops after the first phase where Alice learns if Bob is nearby or, if both engage in the optional second phase of the protocol, Alice and Bob learn their respective locations. In the Louis protocol, both users are able to lie about their location if the TTP collaborates with

them. Furthermore, Alice is able to lie about her location if the second phase of the protocol is not run. In the Louis protocol, both users are able to lie about their location and the only protection against misbehavior is that users can physically verify the result of the protocol run. If the Louis protocol returns that the users are nearby, but one of the users cannot see the other, then she knows that Bob or the TTP must have misbehaved. Likewise, if one of the users spots the other user although the protocol returned that the users are not nearby, then misbehavior is revealed. The Lester protocol works without a TTP. However, while Alice learns location information about Bob, Bob does not learn location information about Alice in return. Furthermore, Alice does not learn exact location information, but only if Bob is within a certain radius of her location. Unfortunately, the Lester protocol does not allow Alice or Bob to learn whether the other lies about their current location. Even worse, Alice can use a single response of Bob to probe his location for several radii. The only protection the Lester protocol provides is that both Alice and Bob can participate in the protocol and change their inputs in such a way that the result of the protocol is that they are not nearby without the other user being able to detect this behavior. In the Pierre protocol, Alice and Bob will, as in the Lester protocol, not learn exact locations, but only a coarse location area that they are within. While cheating about their location is still possible in the Pierre protocol, it comes with the significant advantage that Alice cannot probe a single response of Bob for multiple guesses. Instead, Alice is only able to check if Bob is nearby to the location that she has provided in the protocol run. While of course Alice could initiate multiple, consecutive location runs, Bob is able to detect that kind of probing and stop participating in the protocol.

Narayanan et al. [146] propose three protocols for friend-nearby notification. Two protocols are based on private set intersection [85] and assume a TTP. The key distribution of the system is done via Facebook instead of traditional Public Key Infrastructure (PKI) solutions. In the first protocol, both Alice and Bob need to be simultaneously online. The second protocol supports asynchronous private proximity testing. The first protocol is more efficient but insecure against collaboration between TTP and users. To prevent cheating in the protocol run, i.e. lying about one's location, the paper suggests to use *location tags*. A location tag is a nonce derived from electromagnetic signals present at the respective location. Using a location tag as input to the protocol comes with the advantage that lying about the location is much harder, because a cheating user would have to find the location tag for a location where she is not present. However, it is unlikely that Alice and Bob will both measure the exact same electromagnetic signals even if they are in close proximity. Therefore, the authors propose Private Threshold Set Intersection (PTSI) [77], which allows the system to determine proximity if the location tags are only similar but not equal. This comes, however, at the cost of efficiency.

The efficiency problem of Narayanan et al.'s proposal is solved by Lin et al. [137]. They employ a shingling technique [33] that allows them to use the more efficient Private Equality Testing (PEqT) in such a way that the protocol finds Alice and Bob nearby even if their location tags are similar but not strictly identical. Furthermore, they propose to compute location tags based on messages received from GSM base stations on their *paging channel*, which is used to ping mobile devices in the area of the base station. Since GSM has a much higher signal strength, users can receive the same signals even if they are further apart. Signals on the paging channel turn out to be a valid source for location tags since the data on the paging channel includes a unique identifier, which is assigned to the base station, and the paging requests turn out to be unique and random.

Bilogrevic et al. [25] propose a scheme that allows users to find a Fair Rendez-Vous Point (FRVP) in a privacy-preserving way. This is a meeting point that is both fair, i.e. the maximum distance everyone has to travel to the meeting point is minimized, and private, i.e. everyone only gets to know the final meeting point and no user or the service provider gets to learn private location information of the other users. Bilogrevic et al. find a FRVP by solving the k-center problem with $k = 1$. The solution requires the service provider to compute distances. The authors employ homomorphic encryption to do this in a privacy preserving way and propose two different protocols. The first protocol computes the distances with the help of the Boneh-Goh-Nissim (BGN) [73] cryptosystem and the second solution computes it with the ElGamal [73] and Paillier [158] cryptosystems. Their analysis showed that the ElGamal/Paillier system is more efficient on both the user's device as well as the LDS, because the BGN schemes makes use of rather expensive bilinear mappings. However, even on nowadays outdated Nokia N810 mobile devices (ARM 400 MHz CPU) the operations of both cryptosystems can be performed efficiently.

4.3 POI Finder

One of the most commonly used LBSs allows users to find Point-of-Interest (POI) around their location. Therefore, the user submits a query to the service provider along with her location and, optionally, some additional information on what kind of POIs the user is interested in.

Private Information Retrieval (PIR) [48, 130] is suited to implement a POI finder in a privacy-preserving way. PIR is a mechanism that allows users to query a database without the database server learning what information the user requested. Ghinita et al. [95] propose two protocols based on PIR.

The first protocol works with a single PIR request at the cost of providing only approximate results. The second protocol has a higher computational and communication overhead but provides more accurate query results. Both protocols are built on a data structure based on Hilbert curves and search trees that convert the map coordinates of POIs into 1-dimensional coordinates preserving the proximity of POIs. This allows to apply PIR on originally two-dimensional data.

The drawback of the Ghinita et al.'s solution is that PIR is computationally heavy on the service provider. To overcome this issue, Olumofin et al. [155] propose a POI finder that combines PIR with a generalization technique called cloaking. The user reveals in what larger area she is located and the service provider runs the PIR scheme only in this larger area. This has the advantage that the PIR runs are more efficient, because it is run only on the subset of the entire database including all POIs. However, the solution comes with the drawback that the user has to reveal at least some information on her location. This effectively allows the user to engage in a privacy/performance trade-off. The larger the area the user reveals, the more uncertainty the service provider has about the user's whereabouts at the cost of longer protocol runtime.

4.4 Traffic Monitoring

The idea of traffic monitoring is that cars on the road are equipped with tracking devices and report to a LBS additional data, such as their current speed. This allows the LBS to compute statistics, such as current road utilization, that can be used to navigate cars in a more efficient way. Clearly, such a system can provide considerable advantages, including the detection of traffic jams and the subsequent redirection of other cars to a faster route. However, if cars constantly reveal their locations, any observer would be able to learn private information about the users as described in Section 3.2.

Hoh et al. [110] employ basic cryptographic operations such as encryption and hash functions in order to build a private and secure traffic monitoring infrastructure. Cars send the LBS provider encrypted messages and only the LBS knows the respective decryption keys. While this provides some privacy against entities that observe traffic data, the LBS needs to be trusted and is able to learn all the location information of the users.

A privacy-preserving alternative called *PrivStat* is proposed by Popa et al. [162]. Their system allows to compute privacy-preserving location statistics, such as traffic monitoring, that protects against any kind of side information. The authors design their system such that the LBS provider does not need to be

trusted, since it may try to infer user movement. Furthermore, the clients are neither trusted as they may try to bias aggregates in their favor. Central to the protocol is the so-called Smoothing Module (SM) that creates the keys of a Paillier cryptosytem [158] and ensures that clients upload the right number of measurement values needed for the execution of the protocol. The SM may be a dedicated entity on a measurement point or may be implemented in a distributed fashion running on the users' devices. In both cases, PrivStat is designed to detect possible data corruptions of the SM. During the runtime of the system, the clients encrypts every measurement point, using the public key from the SM, and send it to the LBS provider. In return, the LBS provider is able to compute the aggregate of all received measurement data thanks to the homomorphic property of the Paillier cryptosystem. Subsequently, the LBS provider sends the encrypted aggregate to the SM that is able to decrypt the aggregate. Finally, the LBS provider verifies that the SM has not corrupted the decrypted value using the trapdoor permutation of the Paillier cryptosystem. Since clients upload their measurement values anonymously, PrivStat implements accountability functionality. By employing e-cash schemes, in a similar way as demonstrated by Camenisch et al. [37], PrivStat can ensure that users are only able to upload a maximum number of measurement points; employing interval zero-knowledge proofs ensures that every measurement point is within a pre-defined range.

4.5 Contribution

Our work in [107] proposes a novel Location-Sharing-based Service (LSBS). Location sharing is a key functionality of GSNs and friend-nearby notification services. Thus, our work mainly contributes to Section 4.1 and Section 4.2, respectively.

Similar to existing work, such as [40, 69, 164], the LBS provider does not learn the location of the users in our work. However, we put particular attention to the business model of the LBS. Learning the users' locations is of fundamental importance to the business model of most LBSs. For example, knowing the users' locations allows them to present them ads of nearby shops or attractions. In our work, we address the absence of this form of monetization in two ways. First, we keep the overhead of operations on the LBS provider side minimal. This allows to maintain the service at the lowest possible costs. Second, an extension of our solutions allows the service provider to learn statistics of the locations that users have shared among each other in a privacy-preserving way.

We propose two protocols that are based on identity-based broadcast encryption [93]. Users encrypt their location data and use the LBS provider

only as a communication channel. Consequently, the LBS provider only has to forward ciphertexts and does not engage in any computationally demanding operations unlike prior work. While both protocols allow a user to protect her location data, the user reveals her friends graph in the first protocol. This allows for a particular efficient service. The second protocol offers a trade-off between performance and location privacy. Since users cannot reveal their friends, they define a larger region of interest for the locations they want to share with their friends (*share-area*) and another region of interest for the locations they receive (*receive-area*). The service provider then forwards location update only to those users that are currently in the share-area and have set their receive-area such that it includes the share-area.

We extend the first protocol in order to allow the service provider to learn aggregated data on the number of times a user visited a location. This may be interesting for venues that want their customers to check-in into their venue and to share this information with their friends. Every time a user checks-in a location, the user shares the venue's location with her friends and increases a committed counter for that location. While this committed counter is hidden from the service provider, the user can choose to disclose the counter of a particular location. Doing so may entitle her to certain discounts or prizes of that venue.

We have implemented both of our protocols on a Samsung S III mini smart phone that runs a 1 GHz dual core processor. Our experiments have shown that even rather complex cryptographic operations, such as bilinear mappings, impose an insignificant computational overhead, even on this rather outdated mobile device.

4.6 Conclusion

As this chapter has shown, it is possible to apply cryptographic primitives in order to design private LBSs. In such a setting the service provider is usually considered to be honest-but-curious, meaning that he follows the protocol honestly but may try to learn information about his users. Unfortunately, as the current practice shows, such proposals for private LBS are currently merely of academic interest, as currently deployed LBSs are implemented in a privacy-invasive way. There is almost no hard evidence shedding light on why this is the case, but several reasons seem plausible. Mainly the business interests of the service provider seem to hinder a wide adoption of private LBSs. Furthermore, as we have seen in Section 3.1.3, many different entities, such as third party software libraries, may require access to the user's location data. Consequently,

even if the LBS provider would be open to implement a private service, the lack of private third party software libraries may make this impossible. Another possibility for the lack of private LBSs is the competitiveness of the market. The adopted cryptographic primitives require a substantial knowledge on information security and their implementation requires considerable effort. Both requires the investment of additional resources for which the market does not provide an immediate reward. Finally, developing an industry product from a proof of concept implementation, common in academic works, requires substantial resources. Especially factors such as usability and marketing are difficult to get right.

Chapter 5

Obfuscation-based Location Privacy

Most LBSs do not employ mechanisms as described in the previous chapter to protect user location data. Instead, they are designed in a way that they learn user location data and they leak location data to third parties. Obfuscation can be a practical way for users to protect their location when engaging in these, privacy-invasive LBSs. Instead of using their accurate location x, users employ a Location Privacy Protection Mechanism (LPPM) that computes a pseudo-location z and then transmits this pseudo-location to the service provider. As a result, the LBS provider and third parties receive only altered or approximate location instead of accurate location information. There are four main types of obfuscation strategies that have been extensively studied in the literature: hiding location data, perturbation, reducing precision, and dummies. Which obfuscation strategy is best suitable for a given scenario depends on the setting. A particular difficult case for all obfuscation-based LPPMs is to protect the users in a setting where they continuously query the LBS. In this case, subsequent queries become linkable and patterns tend to persist. However, in a case where the users query the LBS just once, i.e. *sporadic*, obfuscation can offer a much better level of protection.

An adversary that observes obfuscated location data tries to invert the operations of the LPPMs, i.e. to find x when observing z. For this task the adversary is assumed to be *strategic*, i.e. to know: the functioning of the LPPMs, the user's obfuscation parameters and additional information, such as the terrain or the user's regular movements. One of the key parameters for the user to choose is the Quality Loss (QL) that she is willing to tolerate in order to protect her

location privacy. The reason for QL is that the user alters her location and thus the LBS provider is only able to compute the response to imprecise data. While there is QL with most LBSs when obfuscation is employed, some services can naturally better tolerate obfuscation. For example, a weather app works similarly fine with accurate location information and location information on the granularity of a city.

5.1 Quantification of Location Privacy

Choosing the right way to measure the effectiveness of LPPMs, i.e. choosing an appropriate *metric*, is crucial when evaluating the level of protection level that they provide. Otherwise, assessments of their protection may not correspond to their actual level. Furthermore, different LPPMs can only be compared in a meaningful and fair way if they are evaluated under the same, meaningful metric. Authors of LPPMs have adopted a wide variety of different metrics when they have evaluated their LPPM proposals. In most of the cases, the chosen metrics originate from other areas of privacy research, such as anonymous communication, where they have proven to be useful. However, sometimes authors have invented ad hoc metrics to evaluate the efficiency of their proposals.

Shokri et al. [174] propose a framework for the unified evaluation of LPPMs and argue that the adversary's error in reconstructing the user's location or trace is the right metric for the quantification of location privacy. The proposed framework relies on a Hidden Markov model and applies a series of well-established statistical methods to quantify location privacy. The evaluation of a real-world data set shows the shortcoming of two other metrics that are commonly used to evaluate LPPMs. The error of the adversary is now considered to be the accepted metric for location privacy evaluations and has been used in several other works, such as [7, 31, 44, 175, 176].

Prior to Shokri's work, entropy and k-anonymity were among the most commonly used metrics. Entropy [173] has been widely used among the different types of LPPMs, such as [23, 113, 118, 134, 143, 199]. In privacy research, this metric was originally applied by Diaz et al. [67] and Serjantov and Danezis [171] to measure user privacy in anonymous communication systems. In a location privacy setting the adversary assigns a probability p_{ij} that corresponds his observations of the system. This could be, for example, the probability of two LBS queries i and j to be the consecutive queries by the same user; the probability of two consecutive queries i, j to belong to the same user; or the probability that two users i, j have switched their pseudonym. Applying Shannon's entropy measure yields the uncertainty of the adversary:

$$h_j = -\sum_i p_{ij} \cdot \log(p_{ij}) \tag{5.1}$$

An LPPM is then considered to increase the privacy of the user if it increases the uncertainty of the adversary's. Another widely used metric is k-anonymity, which was the natural choice as metric of many precision-based LPPMs, such as [24, 49, 92, 97, 102, 119], as their functioning is inspired by k-anonymity in a database setting [184]. The assumption of the metric is that if the user submits an area where there are more possible users that may have issued the query, i.e. a higher k value, then the user enjoys a higher level of location privacy. Despite its widespread use, researchers noticed that k-anonymity is not suitable to measure location privacy and may even be detrimental to the user's location privacy [177].

5.1.1 Contribution

We propose the novel notion of possibilistic location privacy in [106]. This notion provides a first-order estimation of the protection provided by LPPMs. The main difference to the probabilistic approach by Shokri et al. is its simplicity. Particularly, given a obfuscated location, we compute the area where the user is with a certain probability and consider this as the user's possibilistic area, i.e. the area where the user can possibly be. For a subsequent LBS query, we again compute a possibilistic area and intersect it with the previous possibilistic area considering possible user movements between the consecutive queries. This allows us to further narrow down the possible region of the user for every received LBS query.

In principle the probabilistic approach by Shokri et al. can evaluate arbitrarily complex situations with a great variety of user behavior, inference attacks, and adversarial prior knowledge. However, we show that the computational needs soon become prohibitively expensive for meaningful evaluations, because the complexity of the framework grows quadratically in the number of considered locations. In our case, the complexity of our possibilistic approach is constant in the number of locations. Due to this much more advantageous complexity, it turns out that we achieve with our possibilistic approach an assessment of user privacy that is much more fine grained than with the probabilistic approach in a feasible time.

We use our possibilistic framework to provide the first evaluation of the privacy level provided by the geo-indistinguishability mechanism proposed by Andrés et al. [7] and Chatzikokolakis et al. [44]. In particular, we evaluated the privacy

level of those two LPPMs when the user exposes locations with high enough frequency such that the consecutive locations become linkable by the adversary.

5.2 Obfuscation-based Protection Schemes

In the following we outline the existing proposals for LPPMs and categorize them along the four main types of obfuscation strategies. Most of the LPPMs assume a TTP in place that computes a user's pseudo-location z. While this may help to achieve a better obfuscation, the TTP becomes the single point of trust and single point of failure. As a result, some LPPMs run only on the user's device or communicate only with other users in order to compute the pseudo-location.

5.2.1 Hiding Events

With this obfuscation strategy the user stops engaging in the LBS for a certain time or at a certain place. While the user, obviously, achieves location privacy when she is not querying the LBS, the key benefit in this strategy is that the user may change or exchange the pseudonym that she uses when contacting the LBS.

Beresford and Stajano [23] propose the concept of mix zones. When a user enters such a zone, she remains silent until she leaves the zone. While being inside the mix zone, the user changes the pseudonym with which she queries the LBS. She either creates a new pseudo-random pseudonym or exchanges the pseudonym with another user. The strategic adversary tries then to infer the mapping of pseudonym exchanges and its success can be measured with the entropy metric. Similarly, Jiang et al. [117] study how silent periods in WiFi networks can be used to exchange MAC addresses. Huang et al. [113] further formalize mix zones and generalize them to a Mix-based model used in anonymous communication, such as [47]. Freudiger et al. [87] study the optimal placement of mix zones. Therefore they introduce a privacy metric that captures the adversary's probability of wrongly mapping a pseudonym exchange. They formalize then the optimal Mix network that maximizes the adversary's probability to perform wrong mappings.

Another approach is proposed by Li et al. [134]. They do not rely on fixed mix zones, but have the users exchange their pseudonyms when they are nearby. They also measure the privacy of the users with entropy. Hoh et al. [111] propose a similar approach but realize that the entropy metric does not provide

a meaningful privacy estimate, because it does not express how long users can be tracked. Therefore, they propose the *mean-time to confusion* metric that captures the mean during which the adversary is able to track a user.

5.2.2 Reducing Precision

One of the most widely used LPPM strategies is to reduce the precision. Instead of using accurate locations, the user provides a cloaking region to the LBS provider. For example, a user provides as location the region of a city instead of her accurate GPS coordinates. Reducing precision typically assumes that the service provider supports queries with cloaks [20]. This obfuscation strategy comes with the advantage for the user that the provider has then some level of uncertainty about the user's actual location and whether it is within the cloaking region, respectively. The privacy of the user is typically expressed in the k-anonymity metric, but some works also employ entropy as a privacy metric [118, 199].

A series of works assume a TTP that is aware of all the user's locations to compute the cloaking area. The algorithm to compute the area applies the popular concept of *k-anonymity* from databases [184] to the area of location privacy. The first LPPM of this type was proposed by Gruteser and Grunwald [102]. Users choose a value k and the LPPM computes a cloaking area that covers $k-1$ other users. The users also share with the TTP spatial and temporal constraints. These constraints express possible user preferences on how large the area at least/at most has to be and for how long the user is willing to wait until $k-1$ users are sufficiently close to compute the region. The TTP organizes the current user positions in a quad-tree in order to be able to efficiently compute the obfuscated areas.

Several follow-up works and improvements on Gruteser and Grunwald's work exist. All have in common that they assume a TTP that knows the user's actual locations in order to compute the cloaking region. Bettini et al. [24] observe that if every single query is protected using k-anonymity, subsequent queries may not be. They therefore introduce the concept of *historical k-anonymity* where a trace of k-anonymous queries itself is k-anonymous. If historical k-anonymity cannot be ensured, the user exchanges her pseudonym using ad hoc mix zones, such as [134]. Wang et al. [199] propose four additional heuristics to achieve historical k-anonymity. Gedik et al. [91] show how quasi-identifiers can be used to break k-anonymity-based LPPMs and outline possible ways to protect against this threat. In another work, Gedik et al. [92] propose a mechanism that employs k-anonymity, but finds areas including k users with the help of a more efficient algorithm that is based on the graph-theoretic concept of cliques.

Mokbel et al. [144] also find the cloaking area with the help of quad-trees, but they can find smaller cloaking areas that also cover k users. Kalnis et al. [119] propose two mechanisms to build the cloaking area. In the first, the cloaking area includes the $k-1$ nearest users of the user who wants to query the LBS. The authors note that the $k-1$ other users covered by the cloaking area may themselves compute cloaks that cover different users, because other users are closer to them. This allows the adversary to rule out some of the k users as being the initiator. The second mechanism does not have this drawback, because it chooses $k-1$ other users based on a Hilbert curve.

Users may not wish to share their data with a TTP when they obfuscate their whereabouts. Chow et al. [49] propose a distributed protocol to find a cloak of k-nearest neighbors. Users try to find $k-1$ other users who follow the protocol in their surroundings. Once they found enough users, they compute the cloak and use it for their LBS. The most severe drawback of their solution is that the initiator of the cloak tends to be localized in the center of the cloak. Ghinita et al. [97] propose *PRIVE* that relies on distributed B+-trees. This solution has the advantage that the initiator's location under the cloak is more equally distributed and that, similarly to the LPPM of Kalnis et al. [119], all k users in the cloak compute the same cloak. Mobihide [96] uses a DHT [183] and Hilbert curves to solve some performance issues of PRIVE that are due to the B+-trees. Ju and Shin [118] propose an LPPM that creates a cloak covering k POIs instead of k users.

Duckham and Kulik [70] propose a cloaking mechanism that is tailored towards POI queries and that is not built on k-anonymity. Instead, the user chooses an arbitrary region as cloaking area and sends this along her query. The service provider computes an answer that consists of average distance information of all nearby POIs to all possible locations within the cloaking area. The user has then the option to either refine her query, i.e. reduce the cloaking area, or to choose the POI that is on average closest with respect to all the locations within the cloaking area.

5.2.3 Perturbation

A user may perturb her location, i.e. use the library as her current location instead of her true location the hospital, in order to protect her location privacy. Perturbation has the significant advantage that the obfuscation is a single location and thus can be used with most of the LBSs as they require a single location as input.

Hoh and Gruteser [109] propose to perturb the location of a user onto the trajectory of another nearby user. For the adversary it then appears as if

the trajectories of two users have crossed and cannot distinguish which user went in which direction. To determine the obfuscation points, the authors formulate the system as a nonlinear optimization problem that maximizes the error of the adversary. Ardagna et al. [15] observe that localizing a user, even when no LPPM is in place, cannot be done accurately. The inaccuracy of the GPS sensor adds noise to the user's location. This effect is regardless of the technology that is being used to determine the location. Ardagna et al. propose to meet the privacy requirements of the user by perturbing x far away such that an adversary observing z is unlikely to choose an x' such that $x = x'$. Shokri et al. [176] formalize an LPPM as a function that considers: i) the user's privacy requirement; ii) the user's maximal tolerable QL; and iii) the adversary's knowledge. They propose a framework that allows to find the user's optimal perturbation LPPM that, first, maximizes the adversary's expected error when employing the optimal inference attack and, second, respects the user's maximal tolerable QL. The authors formulate the problem of finding the optimal protection mechanism and the optimal inference attack as a zero-sum Bayesian Stackelberg game. Theodorakopoulos et al. [185] propose to sample the pseudo-location z from Gaussian noise as this distribution has the highest entropy compared to other distributions.

Andrés et al. [7] propose a perturbation-based LPPM based on the Laplace distribution. Their LPPM provides *geo-indistinguishability*, a notion of location privacy that is also introduced by Andrés et al. and based on a generalized version of differential privacy from Chatzikokolakis et al. [42]. An LPPM provides geo-indistinguishability if the probability of reporting an obfuscated location z is similar for two close locations x and and x' and the more different the further x and x' are apart. The mechanism from Andrés et al. has several advantages over that of Shokri et al. First, the geo-indistinguishable LPPM operates on continuous locations while Shokri's LPPM operates on discrete regions. Second, the mechanism abstracts from the adversary's knowledge. A series of other geo-indistinguishable mechanisms have been proposed. Bordenabe et al. [31] propose a geo-indistinguishable LPPM that perturbs the user's location such that the QL is minimal. The geo-indistinguishable mechanisms of Chatzikokolakis et al. [43] reuses z if the user's location x is still in close proximity. This allows them to be more economic about the noise that is needed to obfuscate a series of queries to the LBS. In another proposal, Chatzikokolakis et al. [44] propose a geo-indistinguishable solution that adjusts the noise added to pseudo-locations z depending on the density in the current area. For example, the mechanism adds more noise in rural areas than in cities. Xiao et al. [206] argue that applying a differentially private mechanisms, such as the geo-indistinguishability mechanism from Andrés et al. [7], to protect location privacy comes with new challenges, because the subsequent locations of a trajectory contain temporal correlations that leak additional information. As a solution the authors propose

a perturbation LPPM that applies the noise depending on the next possible movements of the user.

5.2.4 Dummies

A dummy-based LPPM queries the LBS with a series of fake locations that may include the user's actual location x. If the LBS accepts dummy queries, the LPPM can send the set of locations directly to the LBS and receives in return an answer to every location in z. However, if the LBS only accepts a single location as input, then the LPPM could still send subsequent queries for every location in z to the LBS. While the idea of dummies is rather intuitive, Krumm [128] shows that the creation of realistic dummies is a quite difficult task. By analyzing a real-world data set, Krumm shows that key parameters for the creation of realistic dummy trajectories, such as the user's driving speed, the randomness of routes, start/end locations and the GPS noise, are difficult to create artificially.

Kido et al. [122, 123] proposes two LPPMs that create fake movement data. The basic mechanism computes random trajectories and a more sophisticated version considers the distribution of other users in the map and high density regions when computing trajectories that lead to high user privacy. You et al. [208] propose an LPPM that computes dummy trajectories such that they intersect with real trajectories as well as other dummy trajectories. The authors argue that this increases user privacy, because more places appear possible as the current location of the user. In the LPPM of Lu et al. [138] the user defines her privacy requirement as a tuple of two values (k, s). The LPPM then creates k dummy locations in an area of size s. Meyerowitz and Choudhury et al. [143] propose to also cache the replies of true and dummy requests to the LBS. In a setting with a TTP, the TTP caches the replies. A user first queries the TTP that answers the query itself if the respective reply is within its cache and queries the LBS if it is not. If there is no TTP, the authors suggest to have the cache on the user devices in a Peer-to-Peer (P2P) fashion using a DHT. An LPPM that creates dummy locations according to the ℓ-diversity paradigm [140] is proposed by Xue et al. [207]. Particularly, the LPPM computes the dummies in such a way that at least ℓ different location types, such as library, hospital or museum, are among the dummy locations. Furthermore, a more sophisticated LPPM also considers the probability distribution of the location types in order to create more realistic dummy locations.

Chow and Golle [50] as well as Shanker et al. [172] create dummy routes with the help of other LBSs. For example, the user may query Google Maps for a route between two places and alter this route with adding some noise to find

a realistic driving route. The advantage of both proposals is that they do not need a TTP in order to function. Instead, all the computations can happen on the user's device.

Bindschaedler and Shokri [26] propose a framework for the construction of synthetic, yet plausible dummy traces. Therefore the framework takes real location traces as seeds in order to compute the dummy traces. They further leverage the fact that the location traces of the users are semantically similar. Most users commute between a work place and their home and further visit several other places, such as a friend's place or their favorite restaurant.

5.2.5 Contribution

Our contribution in [108] computes a user's optimal LPPM for sporadically querying the LBS. Our work extends a framework proposed by Shokri et al. [176] for the computation of optimal perturbation LPPMs. Similar to Shokri et al.'s framework, we consider a strategic adversary, i.e. an adversary that knows and exploits the LPPM algorithm, the user's movement profile and the user's maximal tolerable QL.

The key feature of our framework is that it considers constraints on resources, such as the bandwidth, which are typically of concern on resource-constrained mobile devices. Considering such constraints allows us to compute the optimal dummy-based and precision-based LPPMs. Our evaluation provides two additional insights. We show that the user's maximal level of privacy can be achieved by either sufficiently relaxing the QL constraint, sufficiently relaxing the bandwidth constraint or with an adequate relaxation of both. Second, if the user can tolerate communication overhead, then dummy as well as precision-based LPPMs can provide a better protection than the perturbation-based LPPM.

5.3 Conclusion

Most available LBSs are designed such that the LBS provider and third parties learn the location data of the users. LPPMs are a commonly proposed mechanism for users to protect their privacy when engaging in such privacy-invasive LBSs. Key to the performance assessment of LPPMs is a unified evaluation framework. The state-of-the-art framework is based on a Markov model. Furthermore, the framework employs the distortion between the user's actual location and the adversary's estimate as the proper metric to quantify

location privacy. This framework allows to evaluate the different LPPMs. In the literature the commonly used obfuscation strategies are: location hiding, reducing precision, perturbation and dummies. While indeed LPPMs can be used with most LBSs, they impose a QL for the user, because the LBSs provider needs to operate on imprecise data. The extent of this QL depends on several factors, such as the user's privacy requirements.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In Part I of this thesis we have explained what privacy is and outlined why it needs to be protected. The focus of this thesis is on privacy in mobile services and to allow users to hide their location data. We have outlined the current mobile device eco-system and that the data of users engaging in mobile services is being processed indiscriminately. This practice mainly emerged for two reasons. First, for economical reasons as the current business model rewards entities knowing most about their users. Second, the mobile device eco-system has not been designed with privacy in mind and and thus allowing for indiscriminate data collection. It turned out that the massive data collection is a major threat to the privacy for users of mobile devices. In the following, we have focused on location privacy, a topic that became increasingly popular along the widespread use of mobile Location-Based Services (LBSs). We have outlined that in existing LBSs the user cannot keep her location data confidential, because the LBS provider along with several other entities learn accurate location information of the user.

In [105] we have conducted an interdisciplinary analysis and considered LBSs as a new socio-technical practice instead of just a technology. We have employed the concept of Contextual Integrity (CI) in order to show that the current functioning of LBSs violate privacy. Our analysis also included a legal assessment showing that current LBSs are not in line with the legal concept of *purpose limitation*. Furthermore, we presented in [195] a framework that is able to quantify information leaks on mobile devices. Thus, our work provides further

evidence that the current mobile eco-system does not properly protect sensitive user privacy, such as unique identifiers and location data.

With respect to the protection of location privacy, we have first presented in [107] a private LBS that allows users to exchange location data and location-based information such that no other entity is able to read the user's information. We acknowledged that many LBS providers need to monetize their service and addressed this in two ways. First, we designed our service such that it imposes minimal working overhead on the service provider's infrastructure. Second, we extended our protocols such that the LBS provider is able to learn private statistics on the user movements. This may serve as some sort of revenue for the LBS provider to monetize its investments. We note that while statistics that are created in a privacy-preserving may not constitute a privacy problem for the individual, they way may still be invasive. For example, a service provider may be able to use statistics to discriminate users if their behavior deviates from the average behavior.

Since most LBS providers are reluctant to run privacy-friendly services, location obfuscation is often the last resort of users who wish to protect their location privacy. In [108] we have studied the design of optimal Location Privacy Protection Mechanisms (LPPMs) that allow users to trade off their privacy with resources they are willing to invest and a loss in service quality that they are willing to tolerate. Finally, we have proposed in [106] our novel notion of possibilistic location privacy. This first-order estimation of location privacy is very efficient such that it allows the quantification of user location privacy in real-world scenarios.

6.2 Future Work

While our interdisciplinary analysis shows that the way LBSs process user location data violates CI, it does not investigate CI when protection mechanisms are in place. This may include private LBSs, which we have outlined in Chapter 4 together with our proposal of a private location-sharing service in [107], as well as obfuscation-based mechanisms, which we have summarized in Chapter 5. The investigation of this impact is an interesting path of future work of the interdisciplinary part of this thesis. Furthermore, since we have taken a more general perspective with respect to the concept of contexts in order to provide a more general analysis, future work could include the analysis of specific and concrete scenarios. This will allow to gain more specific insights in how the processing of location data may violate CI.

Our framework for the quantification of information leaks, described in [195], should be extended such that it also analyzes encrypted messages. This is particularly important as more and more Mobile Application (app) providers and third party software employ HTTPS as communication protocol. The extended version of our current framework, which was designed to only consider plaintext messages, may be able to more accurately assess the information leakage on mobile platforms. Another interesting line of research is a further analysis of possible countermeasures. In a first step, this should include an in-depth analysis on how mobile ad-blockers impact the adversary's efficiency. Our current analysis only shows that ad-blockers have a limited effect on the adversary's efficiency without going into detail why this is the case. For example, this could be because either ad-blockers do actually not block the parts of the ad traffic that allow the adversary to link user app sessions or because third party libraries other than AA libraries send sufficiently often identifiers for the adversary to achieve the attack efficiency that we describe in our work. In the second step, after the reasons for the efficiency of the adversary's attack have been investigated in detail, future work could include the development of protection mechanisms.

Our proposal for private LBSs has the drawback that users can lie about their location. This means that they can claim to be at a particular place, while they are actually elsewhere. This has significant disadvantages especially for reward-based check-in services, such as Foursquare, because users can check-in into a venue arbitrarily often. The only protection that our protocols provide against this threat is that always when a user checks-in, she also has to share this location with her friends. However, cheating users may simply choose to share with some of their friends who do not mind to receive false location information or they may create fake profiles with whom they share their false location information. A solution that prevents this kind of misuse could include dedicated hardware at the venue with whom the user engages in a private protocol via nearby communication technologies, such as WiFi or Bluetooth.

There are several further lines of future work on our contributions on LPPMs. Our possibilistic approach in [106] needs to be further compared to the state of the art framework that is based on a Markovian setting. The first line of future work is to analyze the impact of the user's movement profiles. The probabilistic framework proposed by Shokri et al. [174] requires prior knowledge on the users' mobility profile. In current evaluations the adversary was assumed to possess accurate prior knowledge, but imprecise or false knowledge was not evaluated. Although assuming the existence of accurate location information provides an upper bound of the adversary's attack, in more practical scenarios the adversary possesses only inaccurate prior knowledge. Furthermore, changes in the daily routine or one-time events may not be captured by adversaries that

too heavily rely on prior knowledge. Since our possibilistic approach does not rely on any prior knowledge, a comprehensive evaluation will reveal further interesting insights on the practicability of the framework of Shokri et al..

A second line of future work is the investigation of quantification frameworks that take a middle approach between simplicity and complexity. Our possibilistic framework is a first-order estimation and thus takes an extreme position on simplicity of the model that yields a practical but simplified analysis. Its competitor the probabilistic approach that, in theory, is capable to quantify location privacy in the highest detail possible, but that has a complexity that makes any reasonable analysis prohibitively expensive. Opting for a quantification that is not as simplistic as our approach, yet not as complex as the current framework, may provide additional quantification accuracy at reasonable costs. Second, the possibilistic area of a user is a rather intuitive representation of location privacy. This approach may, therefore, be suitable to design tools that visualize the user's current level of location privacy or the impact on their privacy level when they query the LBS another time. Third, while our possibilistic approach allows for practical privacy quantification, it is unclear how the insights of the possibilistic evaluation provides on the design of LPPMs.

Future work should investigate whether LPPMs can be designed to offer strong protection against possibilistic strategies. In a subsequent step, if such LPPMs exist, future work should include the investigation whether such LPPMs also provide better protection against more complex attack strategies, such as the Markovian approach. Furthermore, with respect to the design of LPPMs, we often find a similar situation than with quantification frameworks. Our framework in [108] provides optimal protection, but is computationally very demanding. Investigating how the possibilistic approach can be applied to the design of LPPMs may also solve the problem of finding strong LPPMs with reasonable computational overhead that also are not tailored towards a particular mobility profile, but allow for deviations from the daily routine.

Bibliography

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security (CCS)*, pages 674–689. ACM, 2014.

[2] Lalit Agarwal, Nisheeth Shrivastava, Sharad Jaiswal, and Saurabh Panjwani. Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising. In *Proceedings of the 9th ACM Symposium on Usable Privacy and Security (SOUPS)*, pages 8:1–8:13. ACM, 2013.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-Preserving Data Mining. In *ACM Sigmod Record*, volume 29, pages 439–450. ACM, 2000.

[4] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location has been Shared 5,398 Times!: A Field Study on Mobile App Privacy Nudging. In *Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems (CHI)*, pages 787–796. ACM, 2015.

[5] Irwin Altman. *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. ERIC, 1975.

[6] Monica Anderson. More Americans Using Smartphones for Getting Directions, Streaming TV. `http://www.pewresearch.org/fact-tank/2016/01/29/us-smartphone-use/`. Accessed: 2016-08-16.

[7] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM Conference on Computer and Communications Security (CCS)*, pages 901–914. ACM, 2013.

[8] Android. UI/Application Exerciser Monkey. https://developer.android.com/studio/test/monkey.html. Accessed: 2016-07-25.

[9] Android. System and kernel security. https://source.android.com/security/overview/kernel-security.html, 2015.

[10] Android. System permissions. https://developer.android.com/guide/topics/security/permissions.html, 2015.

[11] Android. Dashboards. https://developer.android.com/about/dashboards/index.html, 2016.

[12] Denise Anthony, Tristan Henderson, and David Kotz. Privacy in Location-Aware Computing Environments. *IEEE Pervasive Computing*, 6(4):64–72, 2007.

[13] Asia-Pacific Economic Cooperation (APEC). APEC Privacy Framework. http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~/media/Files/Groups/ECSG/05_ecsg_privacyframewk.ashx. Accessed: 2016-04-18.

[14] Apple. Entitlements. https://developer.apple.com/library/ios/documentation/Miscellaneous/Reference/EntitlementKeyReference/Chapters/AboutEntitlements.html, 2016.

[15] Claudio Agostino Ardagna, Marco Cremonini, Ernesto Damiani, S De Capitani Di Vimercati, and Pierangela Samarati. Location Privacy Protection Through Obfuscation-Based Techniques. In *21st Annual Working Conference on Data and Applications Security IFIP*, volume 4602 of *Lecture Notes in Computer Science*, pages 47–60. Springer Berlin Heidelberg, 2007.

[16] Article 29 Data Protection Working Party. Opinion 115. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2005/wp115_en.pdf. Accessed: 2016-04-18.

[17] Article 29 Data Protection Working Party. Opinion 185. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2011/wp185_en.pdf. Accessed: 2016-04-18.

[18] Daniel Ashbrook and Thad Starner. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.

[19] Australia. Australia's Privacy Act. `https://www.oaic.gov.au/privacy-law/privacy-act/`. Accessed: 2016-04-18.

[20] Bhuvan Bamba, Ling Liu, Peter Pesti, and Ting Wang. Supporting Anonymous Location Queries in Mobile Environments With PrivacyGrid. In *Proceedings of the 17th International Conference on World Wide Web (WWW)*, pages 237–246. ACM, 2008.

[21] Louise Barkhuus and Anind K Dey. Location-Based Services for Mobile Telephony: a Study of Users' Privacy Concerns. In *Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction (INTERACT)*, volume 3, pages 702–712. Citeseer, 2003.

[22] Adam Barth, Dan Boneh, and Brent Waters. Privacy in Encrypted Content Distribution Using Private Broadcast Encryption. In *10th International Conference Financial Cryptography and Data Security (FC)*, volume 4107 of *Lecture Notes in Computer Science*, pages 52–64, 2006.

[23] A.R. Beresford and F. Stajano. Mix Zones: User Privacy in Location-Aware Services. In *Proceedings of the 2nd Annual Conference on Pervasive Computing and Communications Workshops*, pages 127–131. IEEE, 2004.

[24] Claudio Bettini, X. Sean Wang, and Sushil Jajodia. Protecting Privacy Against Location-Based Personal Identification. In *Second VLDB Workshop on Secure Data Management (SDM*, volume 3674 of *Lecture Notes in Computer Science*, pages 185–199. Springer Berlin Heidelberg, 2005.

[25] Igor Bilogrevic, Murtuza Jadliwala, Vishal Joneja, Kübra Kalkan, Jean-Pierre Hubaux, and Imad Aad. Privacy-Preserving Optimal Meeting Location Determination on Mobile Devices. volume 9, pages 1141–1156. 2014.

[26] Vincent Bindschaedler and Reza Shokri. Synthesizing Plausible Privacy-Preserving Location Traces. In *Proceedings of the 2016 IEEE Symposium on Security and Privacy, (S&P)*, pages 247–262. IEEE, 2016.

[27] Dionysus Blazakis. The Apple Sandbox. *securityevaluators.com*, 2011.

[28] Matt Blaze, Gerrit Bleumer, and Martin Strauss. Divertible Protocols and Atomic Proxy Cryptography. In *Advances in Cryptology—EUROCRYPT*, volume 1403 of *Lecture Notes in Computer Science*, pages 127–144. Springer Berlin Heidelberg, 1998.

[29] Dan Boneh, Eu-Jin Goh, and Kobbi Nissim. Evaluating 2-DNF Formulas on Ciphertexts. In *Second Theory of Cryptography Conference*, volume

3378 of *Lecture Notes in Computer Science*, pages 325–341. Springer Berlin Heidelberg, 2005.

[30] Theodore Book, Adam Pridgen, and Dan S Wallach. Longitudinal Analysis of Android Ad Library Permissions. *arXiv preprint arXiv:1303.0857*, pages 1–9, 2013.

[31] Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Optimal Geo-Indistinguishable Mechanisms for Location Privacy. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security (CCS)*, pages 251–262. ACM, 2014.

[32] Stefan Brands. Rapid Demonstration of Linear Relations Connected by Boolean Operators. In *Advances in Cryptology—EUROCRYPT*, volume 1233 of *Lecture Notes in Computer Science*, pages 318–333. Springer Berlin Heidelberg, 1997.

[33] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. Syntactic clustering of the web. *Computer Networks*, 29(8-13):1157–1166, 1997.

[34] Finn Brunton and Helen Nissenbaum. Vernacular Resistance to Data Collection and Analysis: A Political Theory of Obfuscation. *First Monday*, 16(5), 2011.

[35] AJ Brush, John Krumm, and James Scott. Exploring End User Preferences for Location Obfuscation, Location-Based Services, and the Value of Location. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp)*, pages 95–104. ACM, 2010.

[36] Jan Camenisch, Maria Dubovitskaya, and Gregory Neven. Oblivious Transfer With Access Control. In *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS)*, pages 131–140. ACM, 2009.

[37] Jan Camenisch, Susan Hohenberger, and Anna Lysyanskaya. Balancing Accountability and Privacy Using E-Cash. In *5th International Conference on Security and Cryptography for Networks (SCN)*, volume 4116 of *Lecture Notes in Computer Science*, pages 141–155. Springer Berlin Heidelberg, 2006.

[38] Jan Camenisch and Anna Lysyanskaya. An Efficient System for Non-transferable Anonymous Credentials with Optional Anonymity Revocation. In *Advances in Cryptology—EUROCRYPT*, volume 2045 of *Lecture Notes in Computer Science*, pages 93–118. Springer Berlin Heidelberg, 2001.

[39] Canada. The Personal Information Protection and Electronic Documents Act (PIPEDA). `https://www.priv.gc.ca/leg_c/r_o_p_e.asp`. Accessed: 2016-04-18.

[40] Bogdan Carbunar, Radu Sion, Rahul Potharaju, and Moussa Ehsan. The Shy Mayor: Private Badges in GeoSocial Networks. In *10th International Conference on Applied Cryptography and Network Security (ACNS)*, volume 7341 of *Lecture Notes in Computer Science*, pages 436–454. Springer Berlin Heidelberg, 2012.

[41] Juan Pablo Carrascal, Christopher J. Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. Your Browsing Behavior for a Big Mac: Economics of Personal Information Online. In *22nd ACM International World Wide Web Conference (WWW)*, pages 189–200. ACM, 2013.

[42] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the Scope of Differential Privacy Using Metrics. In *13th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 7981 of *Lecture Notes in Computer Science*, pages 82–102. Springer Berlin Heidelberg, 2013.

[43] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. A Predictive Differentially-Private Mechanism for Mobility Traces. In *14th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 8555 of *Lecture Notes in Computer Science*, pages 21–41. Springer Berlin Heidelberg, 2014.

[44] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. Constructing Elastic Distinguishability Metrics for Location Privacy. *Proceedings on Privacy Enhancing Technologies (PoPETS)*, 2015(2):156–170, 2015.

[45] David Chaum. Blind Signature System. In *Advances in Cryptology—CRYPTO*, pages 153–153. Plenum Press, New York, 1984.

[46] David Chaum, Farid Javani, Aniket Kate, Anna Krasnova, Joeri de Ruiter, Alan T Sherman, and Debajyoti Das. cMix: Anonymization by High-performance Scalable Mixing. Technical report, 2016.

[47] David L Chaum. Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. *Communications of the ACM*, 24(2):84–90, 1981.

[48] Benny Chor, Eyal Kushilevitz, Oded Goldreich, and Madhu Sudan. Private Information Retrieval. *Journal of the ACM (JACM)*, 45(6):965–981, 1998.

[49] Chi-Yin Chow, Mohamed F Mokbel, and Xuan Liu. A Peer-to-Peer Spatial Cloaking Algorithm for Anonymous Location-Based Service. In *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS)*, pages 171–178. ACM, 2006.

[50] Richard Chow and Philippe Golle. Faking Contextual Data for Fun, Profit, and Privacy. In *Proceedings of the 8th ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 105–108. ACM, 2009.

[51] Graham Cluley. The Hacking Team Android Malware App That Waltzed Past Google Play's Security Checks. `https://heatsoftware.com/security-blog/10368/the-hacking-team-android-malware-app-that-waltzed-past-google-plays-security-checks/`, July 2015.

[52] Kate Conger. Apple will require HTTPS connections for iOS apps by the end of 2016. `https://techcrunch.com/2016/06/14/apple-will-require-https-connections-for-ios-apps-by-the-end-of-2016/`. Accessed: 2016-07-19.

[53] Sunny Consolvo, Ian E Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. Location Disclosure to Social Relations: Why, When, & What People Want to Share. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 81–90. ACM, 2005.

[54] Disconnect Corporation. Disconnect Malvertising for Android. `https://disconnect.me/mobile/disconnect-malvertising/sideload`. Accessed: 2016-06-07.

[55] Scott Coull, Matthew Green, and Susan Hohenberger. Controlling Access to an Oblivious Database Using Stateful Anonymous Credentials. In *12th International Conference on Public Key Cryptography (PKC)*, volume 5443 of *Lecture Notes in Computer Science*, pages 501–520. Springer Berlin Heidelberg, 2009.

[56] Crunchbase. Acquisitions Alphabet. `https://www.crunchbase.com/organization/google/acquisitions`. Accessed: 2016-07-19.

[57] Crunchbase. Acquisitions Facebook. `https://www.crunchbase.com/organization/facebook/acquisitions`. Accessed: 2016-07-19.

[58] Colette Cuijpers and Martin Pekárek. The Regulation of Location-Based Services: Challenges to the European Union Data Protection Regime. *Journal of Location Based Services*, 5(3-4):223–241, 2011.

[59] Shuaifu Dai, Alok Tongaonkar, Xiaoyin Wang, Antonio Nucci, and Dong Song. NetworkProfiler: Towards Automatic Fingerprinting of Android Apps. In *Proceedings of the International Conference on Computer Communications (INFOCOM)*, pages 809–817. IEEE, 2013.

[60] Dino A Dai Zovi. Apple iOS 4 Security Evaluation. *Black Hat USA*, 2011.

[61] George Danezis, Roger Dingledine, and Nick Mathewson. Mixminion: Design of a Type III Anonymous Remailer Protocol. In *Proceedings of the 2003 IEEE Symposium on Security and Privacy (S&P)*, pages 2–15. IEEE, 2003.

[62] George Danezis and Seda Gürses. A Critical Review of 10 Years of Privacy Technology. *Proceedings of Surveillance Cultures: A Global Surveillance Society*, pages 1–16, 2010.

[63] Article 29 Data Protection Working Party. `http://ec.europa.eu/justice/data-protection/article-29/index_en.htm`. Accessed: 2016-04-18.

[64] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific Reports*, 3, 2013.

[65] Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. Identification via Location-Profiling in GSM Networks. In *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 23–32. ACM, 2008.

[66] Judith Wagner DeCew. *In Pursuit of Privacy: Law, Ethics, and the Rise of Technology*. Cornell University Press, 1997.

[67] Claudia Diaz, Stefaan Seys, Joris Claessens, and Bart Preneel. Towards Measuring Anonymity. In *Second International Workshop on Privacy Enhancing Technologies (PET)*, volume 2482 of *Lecture Notes in Computer Science*, pages 54–68. Springer Berlin Heidelberg, 2002.

[68] Roger Dingledine, Nick Mathewson, and Paul F. Syverson. Tor: The Second-Generation Onion Router. In *Proceedings of the 13th USENIX Security Symposium (USENIX Security)*, pages 303–320, 2004.

[69] Changyu Dong and Naranker Dulay. Longitude: A Privacy-Preserving Location Sharing Protocol for Mobile Applications. In *5th IFIP WG 11.11 International Conference on Trust Management (IFIPTM)*, pages 133–148. Springer Berlin Heidelberg, 2011.

[70] Matt Duckham and Lars Kulik. A Formal Model of Obfuscation and Negotiation for Location Privacy. In *Third International Conference on Pervasive Computing (PERVASIVE)*, volume 3468 of *Lecture Notes in Computer Science*, pages 152–170. Springer Berlin Heidelberg, 2005.

[71] Cynthia Dwork. Differential Privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2006.

[72] Serge Egelman, Adrienne Porter Felt, and David Wagner. Choice Architecture and Smartphone Privacy: There's a Price for That. In *The Economics of Information Security and Privacy*, pages 211–236. Springer Berlin Heidelberg, 2013.

[73] Taher ElGamal. A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms. In *Advances in Cryptology—CRYPTO*, volume 196 of *Lecture Notes in Computer Science*, pages 10–18. Springer Berlin Heidelberg, 1984.

[74] William Enck, Peter Gilbert, Seungyeop Han, Vasant Tendulkar, Byung-Gon Chun, Landon P Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N Sheth. TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones. *ACM Transactions on Computer Systems (TOCS)*, 32(2):5:1–5:29, 2014.

[75] William Enck, Damien Octeau, Patrick McDaniel, and Swarat Chaudhuri. A Study of Android Application Security. In *Proceedings of the 20th USENIX Security Symposium (USENIX Security)*, pages 1–16, 2011.

[76] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th ACM International Conference on World Wide Web (WWW)*, pages 289–299, 2015.

[77] Ronald Fagin, Moni Naor, and Peter Winkler. Comparing Information Without Leaking It. *Communications of the ACM*, 39(5):77–85, 1996.

[78] Kassem Fawaz, Huan Feng, and Kang G Shin. Anatomization and Protection of Mobile Apps' Location Privacy Threats. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security)*, pages 753–768, 2015.

[79] Kassem Fawaz and Kang G Shin. Location Privacy Protection for Smartphone Users. In *Proceedings of the 2014 ACM Conference on*

Computer and Communications Security (CCS), pages 239–250. ACM, 2014.

[80] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android Permissions Demystified. In *Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS)*, pages 627–638. ACM, 2011.

[81] Adrienne Porter Felt, Kate Greenwood, and David Wagner. The Effectiveness of Application Permissions. In *Proceedings of the 2nd USENIX Conference on Web Application Development*, pages 75–86, 2011.

[82] Huan Feng, Kassem Fawaz, and Kang G Shin. LinkDroid: Reducing Unregulated Aggregation of App Usage Behaviors. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security)*, pages 769–783, 2015.

[83] Amos Fiat and Adi Shamir. How to Prove Yourself: Practical Solutions to Identification and Signature Problems. In *Advances in Cryptology—CRYPTO*, volume 263 of *Lecture Notes in Computer Science*, pages 186–194. Springer Berlin Heidelberg, 1986.

[84] Michael ES Frankel. *Mergers and Acquisitions Basics: The Key Steps of Acquisitions, Divestitures, and Investments*. John Wiley & Sons, 2005.

[85] Michael J Freedman, Kobbi Nissim, and Benny Pinkas. Efficient Private Matching and Set Intersection. In *Advances in Cryptology-EUROCRYPT*, volume 3027 of *Lecture Notes in Computer Science*, pages 1–19. Springer Berlin Heidelberg, 2004.

[86] Julien Freudiger, Raoul Neu, and Jean-Pierre Hubaux. Private Sharing of User Location Over Online Social Networks. In *HotPETs*, pages 1–12, 2010.

[87] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. On the Optimal Placement of Mix Zones. In *7th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 4776 of *Lecture Notes in Computer Science*, pages 216–234. Springer Berlin Heidelberg, 2009.

[88] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. Evaluating the Privacy Risk of Location-Based Services. In *15th International Conference on Financial Cryptography and Data Security (FC)*, volume 7035 of *Lecture Notes in Computer Science*, pages 31–46. Springer Berlin Heidelberg, 2012.

[89] Federal Trade Commission (FTC). Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress. `https://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission`. Accessed: 2016-05-03.

[90] Ruth Gavison. Privacy and the Limits of Law. *The Yale Law Journal*, 89(3):421–471, 1980.

[91] B. Gedik and Ling Liu. Location Privacy in Mobile Systems: A Personalized Anonymization Model. In *25th International Conference on Distributed Computing Systems (ICDCS)*, pages 620–629, 2005.

[92] Bugra Gedik and Ling Liu. A Customizable k-Anonymity Model for Protecting Location Privacy. In *24th International Conference on Distributed Computing Systems (ICDCS)*, pages 1–12. IEEE, 2004.

[93] Craig Gentry and Brent Waters. Adaptive Security in Broadcast Encryption Systems (With Short Ciphertexts). In *Advances in Cryptology-EUROCRYPT*, volume 5479, pages 171–188. Springer, 2009.

[94] Robert S Gerstein. Intimacy and Privacy. *Ethics*, 89(1):76–81, 1978.

[95] Gabriel Ghinita, Panos Kalnis, Ali Khoshgozaran, Cyrus Shahabi, and Kian-Lee Tan. Private Queries in Location Based Services: Anonymizers are Not Necessary. In *Proceedings of the 2008 ACM International Conference on Management of Data (SIGMOD*, pages 121–132. ACM, 2008.

[96] Gabriel Ghinita, Panos Kalnis, and Spiros Skiadopoulos. MOBIHIDE: A Mobile Peer-to-Peer System for Anonymous Location-Based Queries. In *10th International Symposium on Advances in Spatial and Temporal Databases (SSTD)*, volume 4605 of *Lecture Notes in Computer Science*, pages 221–238. Springer Berlin Heidelberg, 2007.

[97] Gabriel Ghinita, Panos Kalnis, and Spiros Skiadopoulos. PRIVE: Anonymous Location-Based Queries in Distributed Mobile Systems. In *Proceedings of the 16th ACM International Conference on World Wide Web (WWW)*, pages 371–380. ACM, 2007.

[98] O. Glickman. Web Page Analysis System for Computerized Derivation of Webpage Audience Characteristics, April 15 2014. US Patent 8,700,543.

[99] Eyeo GmbH. About Adblock Plus for Android. `https://adblockplus.org/en/`. Accessed: 2016-06-07.

[100] Philippe Golle and Kurt Partridge. On the Anonymity of Home/Work Location Pairs. In *7th International Conference on Pervasive Computing*, volume 5538 of *Lecture Notes in Computer Science*, pages 390–397. Springer Berlin Heidelberg, 2009.

[101] Michael C Grace, Wu Zhou, Xuxian Jiang, and Ahmad-Reza Sadeghi. Unsafe Exposure Analysis of Mobile In-App Advertisements. In *Proceedings of the 5th ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, pages 101–112. ACM, 2012.

[102] Marco Gruteser and Dirk Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st International Conference on Mobile systems, Applications and Services (MobiSys)*, pages 31–42. ACM, 2003.

[103] The Guardian. GCHQ Taps Fibre-Optic Cables for Secret Access to World's Communications. `https://www.theguardian.com/uk/2013/jun/21/gchq-cables-secret-world-communications-nsa`. Accessed: 2016-06-07.

[104] Il-Horn Hann, Kai-Lung Hui, Sang-Yong Tom Lee, and Ivan PL Png. Overcoming Online Information Privacy Concerns: An Information-Processing Theory Approach. *Journal of Management Information Systems*, 24(2):13–42, 2007.

[105] Michael Herrmann, Mireille Hildebrandt, Laura Tielemans, and Claudia Diaz. Privacy in Location-Based Services: An Interdisciplinary Approach. *SCRIPTed*, 13(2):25, 2016.

[106] Michael Herrmann, Fernando Pérez-González, Carmela Troncoso, and Bart Preneel. Possibilistic Location Privacy. Technical report, COSIC/ESAT, KU Leuven, 2016.

[107] Michael Herrmann, Alfredo Rial, Claudia Diaz, and Bart Preneel. Practical Privacy-Preserving Location-Sharing Based Services with Aggregate Statistics. In *Proceedings of the 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks (WiSec)*, pages 87–98. ACM, 2014.

[108] Michael Herrmann, Carmela Troncoso, Claudia Diaz, and Bart Preneel. Optimal Sporadic Location Privacy Preserving Systems in Presence of Bandwidth Constraints. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 167–178. ACM, 2013.

[109] Baik Hoh and Marco Gruteser. Protecting Location Privacy Through Path Confusion. In *First International Conference on Security and Privacy*

for Emerging Areas in Communications Networks (SecureComm), pages 194–205. IEEE, 2005.

[110] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Enhancing Security and Privacy in Traffic-Monitoring Systems. *IEEE Pervasive Computing*, 5(4):38–46, 2006.

[111] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Preserving Privacy in GPS Traces via Uncertainty-Aware Path Cloaking. In *Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS)*, pages 161–171. ACM, 2007.

[112] Chris Jay Hoofnagle, Ashkan Soltani, Nathan Good, Dietrich James Wambach, and Mika D Ayenson. Behavioral Advertising: The Offer You Cannot Refuse. *Harvard Law and Policy Review*, pages 1–24, 2012.

[113] Leping Huang, Hiroshi Yamane, Kanta Matsuura, and Kaoru Sezaki. Towards Modeling Wireless Location Privacy. In *5th International Workshop on Privacy Enhancing Technologies (PET)*, pages 59–77. Springer Berlin Heidelberg, 2006.

[114] Interactive Advertising Bureau (IAB). Understanding Mobile Cookies. `http://www.iab.net/media/file/IABDigitalSimplifiedMobileCookies.pdf`.

[115] Giovanni Iachello, Ian Smith, Sunny Consolvo, Gregory D Abowd, Jeff Hughes, James Howard, Fred Potter, James Scott, Timothy Sohn, Jeffrey Hightower, et al. Control, Deception, and Communication: Evaluating the Deployment of a Location-Enhanced Messaging Service. In *7th International Conference on Ubiquitous Computing (UbiComp)*, volume 3660 of *Lecture Notes in Computer Science*, pages 213–231. Springer Berlin Heidelberg, 2005.

[116] Japan. Japan's Act on Protection of Personal Information (Act No. 57 of 2003). `http://www.cas.go.jp/jp/seisaku/hourei/data/APPI.pdf`. Accessed: 2016-04-18.

[117] Tao Jiang, Helen J Wang, and Yih-Chun Hu. Preserving Location Privacy in Wireless LANs. In *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services (MobiSys)*, pages 246–257. ACM, 2007.

[118] Xiaoen Ju and Kang G. Shin. Location Privacy Protection for Smartphone Users Using Quadtree Entropy Maps. *Journal of Information Privacy and Security*, 11(2):62–79, 2015.

[119] Panos Kalnis, Gabriel Ghinita, Kyriakos Mouratidis, and Dimitris Papadias. Preventing Location-Based Identity Inference in Anonymous Spatial Queries. *IEEE Transactions on Knowledge and Data Engineering*, 19(12):1719–1733, 2007.

[120] Jerry Kang. Information Privacy in Cyberspace Transactions. *Stanford Law Review*, 50(4):1193–1294, 1998.

[121] Patrick Gage Kelley, Michael Benisch, Lorrie Faith Cranor, and Norman Sadeh. When Are Users Comfortable Sharing Locations with Advertisers? In *Proceedings of the ACM Conference on Human Factors in Computing Systems (SIGCHI)*, pages 2449–2452. ACM, 2011.

[122] Hidetoshi Kido, Yutaka Yanagisawa, and Tetsuji Satoh. An Anonymous Communication Technique Using Dummies for Location-Based Services. In *Proceedings on International Conference on Pervasive Services (ICPS)*, pages 88–97. IEEE, 2005.

[123] Hidetoshi Kido, Yutaka Yanagisawa, and Tetsuji Satoh. Protection of Location Privacy using Dummies for Location-based Services. In *Proceedings of the 21st International Conference on Data Engineering Workshops (ICDE)*, pages 1248–1248. IEEE, 2005.

[124] United Kingdom. Data Retention and Investigatory Powers Act 2014. `http://www.legislation.gov.uk/ukpga/2014/27/section/8/enacted`. Accessed: 2016-07-26.

[125] Markulf Kohlweiss and Alfredo Rial. Optimally Private Access Control. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society (WPES)*, pages 37–48. ACM, 2013.

[126] Spyros Kokolakis. Privacy Attitudes and Privacy Behaviour: A Review of Current Research on the Privacy Paradox Phenomenon. *Computers & Security*, pages –, 2015.

[127] John Krumm. Inference Attacks on Location Tracks. In *5th International Conference on Pervasive Computing*, volume 4480 of *Lecture Notes in Computer Science*, pages 127–143. Springer Berlin Heidelberg, 2007.

[128] John Krumm. Realistic Driving Trips For Location Privacy. In *7th International Conference on Pervasive Computing*, volume 5538 of *Lecture Notes in Computer Science*, pages 25–41. Springer Berlin Heidelberg, 2009.

[129] Joseph Kupfer. Privacy, Autonomy, and Self-Concept. *American Philosophical Quarterly*, 24(1):81–89, 1987.

[130] Eyal Kushilevitz and Rafail Ostrovsky. Replication is not Needed: Single Database, Computationally-Private Information Retrieval. In *38th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 364–373. IEEE, 1997.

[131] Susan Landau. Making Sense from Snowden: What's Significant in the NSA Surveillance Revelations. *IEEE Security & Privacy*, 11(4):54–63, 2013.

[132] Scott Lederer, Jason I Hong, Anind K Dey, and James A Landay. Personal Privacy Through Understanding and Action: Five Pitfalls for Designers. *Personal and Ubiquitous Computing*, 8(6):440–454, 2004.

[133] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising. In *Proceedings of the 2012 ACM Conference on Human Factors in Computing Systems (CHI)*, pages 589–598. ACM, 2012.

[134] Mingyan Li, Krishna Sampigethaya, Leping Huang, and Radha Poovendran. Swing & Swap: User-centric Approaches Towards Maximizing Location Privacy. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society (WPES)*, pages 19–28. ACM, 2006.

[135] Lin Liao, Donald J. Patterson, Dieter Fox, and Henry A. Kautz. Learning and Inferring Transportation Routines. *Artificial Intelligence*, 171(5-6):311–331, 2007.

[136] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp)*, pages 501–510. ACM, 2012.

[137] Zi Lin, Denis Foo Kune, and Nicholas Hopper. Efficient Private Proximity Testing with GSM Location Sketches. In *16th International Conference on Financial Cryptography and Data Security (FC)*, volume 7397 of *Lecture Notes in Computer Science*, pages 73–88. Springer Berlin Heidelberg, 2012.

[138] Hua Lu, Christian S. Jensen, and Man Lung Yiu. PAD: Privacy-Area Aware, Dummy-Based Location Privacy in Mobile Services. In *Proceedings of the 7th ACM International Workshop on Data Engineering for Wireless and Mobile Access (Mobide)*, pages 16–23. ACM, 2008.

[139] Chris YT Ma, David KY Yau, Nung Kwan Yip, and Nageswara SV Rao. Privacy Vulnerability of Published Anonymous Mobility Traces. *Transactions on Networking*, 21(3):720–733, 2013.

[140] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-Diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):1–52, 2007.

[141] Jonathan R Mayer and John C Mitchell. Third-Party Web Tracking: Policy and Technology. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (S&P)*, pages 413–427. IEEE, 2012.

[142] Aleecia McDonald and Lorrie Faith Cranor. Beliefs and Behaviors: Internet Users' Understanding of Behavioral Advertising. pages 1–31. TPRC, 2010.

[143] Joseph T. Meyerowitz and Romit Roy Choudhury. Hiding Stars with Fireworks: Location Privacy through Camouflage. In *15th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pages 345–356. ACM, 2009.

[144] Mohamed F. Mokbel, Chi-Yin Chow, and Walid G. Aref. The New Casper: Query Processing for Location Services Without Compromising Privacy, 2006.

[145] Ulf Möller, Lance Cottrell, Peter Palfrader, and Len Sassaman. Mixmaster Protocol—Version 2. *Draft, July*, 2003.

[146] Arvind Narayanan, Narendran Thiagarajan, Mugdha Lakhani, Michael Hamburg, and Dan Boneh. Location Privacy via Private Proximity Testing. In *Proceedings of the Network & Distributed System Security Symposium (NDSS)*, pages 1–17. Internet Society, 2011.

[147] United Nations. The Universal Declaration of Human Rights of the United Nations. `http://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf`. Accessed: 2016-04-18.

[148] Helen Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79:101–139, 2004.

[149] OECD. OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. `https://www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf`. Accessed: 2016-04-22.

[150] Council of Europe. European Convention on Human Rights. http://www.echr.coe.int/Documents/Convention_ENG.pdf. Accessed: 2016-04-18.

[151] The Court of Justice of the European Union. The Court of Justice Declares the Data Retention Directive to be Invalid. http://curia.europa.eu/jcms/upload/docs/application/pdf/2014-04/cp140054en.pdf. Accessed: 2016-07-13.

[152] Stanford Encyclopedia of Philosophy. Privacy. http://plato.stanford.edu/entries/privacy/. Accessed: 2016-04-18.

[153] Lukasz Olejnik, Minh-Dung Tran, and Claude Castelluccia. Selling off User Privacy at Auction. In *Proceedings of the Network & Distributed System Security Symposium (NDSS)*, 2014.

[154] Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, and Jean-Pierre Hubaux. Quantifying Interdependent Privacy Risks with Location Data. *IEEE Transactions on Mobile Computing*, PP(99):1–1, 2016.

[155] Femi G. Olumofin, Piotr K. Tysowski, Ian Goldberg, and Urs Hengartner. Achieving Efficient Query Privacy for Location Based Services. In *10th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 6205 of *Lecture Notes in Computer Science*, pages 93–110. Springer Berlin Heidelberg, 2010.

[156] SPIEGEL Online. Mobile Apps Doubleheader: BADASS Angry Birds. http://www.spiegel.de/media/media-35670.pdf. Accessed: 2016-06-07.

[157] Oracle. The BlueKai Registry. http://www.bluekai.com/registry/. Accessed: 2016-07-19.

[158] Pascal Paillier. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. In *Advances in Cryptology—EUROCRYPT*, volume 1592 of *Lecture Notes in Computer Science*, pages 223–238. Springer Berlin Heidelberg, 1999.

[159] Andreas Pfitzmann and Marit Hansen. A Terminology for Talking About Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf, August 2010. v0.34.

[160] David J Phillips. Privacy policy and PETs The Influence of Policy Regimes on the Development and Social Implications of Privacy Enhancing Technologies. *New Media & Society*, 6(6):691–706, 2004.

[161] Iasonas Polakis, George Argyros, Theofilos Petsios, Suphannee Sivakorn, and Angelos D Keromytis. Where's Wally?: Precise User Discovery Attacks in Location Proximity Services. In *Proceedings of the 22nd ACM Conference on Computer and Communications Security (CCS)*, pages 817–828. ACM, 2015.

[162] Raluca Ada Popa, Andrew J Blumberg, Hari Balakrishnan, and Frank H Li. Privacy and Accountability for Location-based Aggregate Statistics. In *Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS)*, pages 653–666. ACM, 2011.

[163] NIST FIPS Pub. 197: Advanced Encryption Standard (AES). *Federal Information Processing Standards Publication*, 197:441–0311, 2001.

[164] Krishna PN Puttaswamy, Shiyuan Wang, Troy Steinbauer, Deepak Agrawal, Amr El Abbadi, Christopher Kruegel, and Ben Y Zhao. Preserving Location Privacy in Geosocial Applications. *IEEE Transactions on Mobile Computing*, 13(1):159–173, 2014.

[165] Ashwini Rao, Florian Schaub, and Norman M. Sadeh. What do They Know About Me? Contents and Concerns of Online Behavioral Profiles. *CoRR*, abs/1506.01675, 2015.

[166] Vaibhav Rastogi, Yan Chen, and William Enck. AppsPlayground: Automatic Security Analysis of Smartphone Applications. In *Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy (CODASPY)*, pages 209–220. ACM, 2013.

[167] Jeffrey H Reiman. Privacy, Intimacy, and Personhood. *Philosophy & Public Affairs*, 6(1):26–44, 1976.

[168] Christopher J. Riederer, Daniel Echickson, Stephanie Huang, and Augustin Chaintreau. FindYou: A Personal Location Privacy Auditing Tool. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 243–246, 2016.

[169] Sanae Rosen, Zhiyun Qian, and Z Morely Mao. AppProfiler: A Flexible Method of Exposing Privacy-Related Behavior in Android Applications to End Users. In *Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy (CODASPY)*, pages 221–232. ACM, 2013.

[170] Louis D. Brandeis Samuel D. Warren. The Right to Privacy. *Harvard Law Review*, 4(5):193–220, 1890.

[171] Andrei Serjantov and George Danezis. Towards an Information Theoretic Metric for Anonymity. In *Second International Workshop on Privacy Enhancing Technologies (PET)*, volume 2482 of *Lecture Notes in Computer Science*, pages 41–53. Springer Berlin Heidelberg, 2002.

[172] Pravin Shankar, Vinod Ganapathy, and Liviu Iftode. Privately Querying Location-Based Services With SybilQuery. In *Proceedings of the 11th International Conference on Ubiquitous Computing (UbiComp)*, pages 31–40. ACM, 2009.

[173] Claude Elwood Shannon. A Mathematical Theory of Communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[174] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying Location Privacy. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy (S&P)*, pages 247–262. IEEE, 2011.

[175] Reza Shokri, George Theodorakopoulos, George Danezis, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Quantifying Location Privacy: The Case of Sporadic Location Exposure. In *11th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 6794 of *Lecture Notes in Computer Science*, pages 57–76. Springer Berlin Heidelberg, 2011.

[176] Reza Shokri, George Theodorakopoulos, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Protecting Location Privacy: Optimal Strategy Against Localization Attacks. In *Proceedings of the 19th ACM Conference on Computer and Communications Security (CCS)*, pages 617–627. ACM, 2012.

[177] Reza Shokri, Carmela Troncoso, Claudia Diaz, Julien Freudiger, and Jean-Pierre Hubaux. Unraveling an Old Cloak: k-Anonymity for Location Privacy. In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 115–118. ACM, 2010.

[178] H Jeff Smith, Tamara Dinev, and Heng Xu. Information Privacy Research: An Interdisciplinary Review. *Management Information Systems Quarterly*, 35(4):989–1016, 2011.

[179] Daniel J Solove. *Understanding Privacy*. Harvard University Press, May, 2008.

[180] Sooel Son, Daehyeok Kim, and Vitaly Shmatikov. What Mobile Ads Know About Mobile Users. In *Proceedings of the Network & Distributed System Security Symposium (NDSS)*, pages 1–15. Internet Society, 2016.

[181] United States. Privacy Act of 1974, 5 U.S.C. § 552a. https://www.justice.gov/opcl/privacy-act-1974. Accessed: 2016-07-12.

[182] Ryan Stevens, Clint Gibler, Jon Crussell, Jeremy Erickson, and Hao Chen. Investigating User Privacy in Android Ad Libraries. In *Workshop on Mobile Security Technologies (MoST)*, pages 1–10. IEEE, 2012.

[183] Ion Stoica, Robert Morris, David Karger, M Frans Kaashoek, and Hari Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. *ACM SIGCOMM Computer Communication Review*, 31(4):149–160, 2001.

[184] Latanya Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[185] George Theodorakopoulos. The Same-Origin Attack Against Location Privacy. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 49–53. ACM, 2015.

[186] Eran Toch and Inbal Levi. Locality and Privacy in People-Nearby Applications. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 539–548. ACM, 2013.

[187] Janice Y Tsai, Patrick Kelley, Paul Drielsma, Lorrie Faith Cranor, Jason Hong, and Norman Sadeh. Who's Viewed You? The Impact of Feedback in a Mobile Location-Sharing Application. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 2003–2012. ACM, 2009.

[188] Janice Y Tsai, Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Location-Sharing Technologies: Privacy Risks and Controls. *Journal of Law and Policy for the Information Society*, 6:119, 2010.

[189] Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley, and Michael Hennessy. Americans Reject Tailored Advertising and Three Activities That Enable It. *Available at SSRN 1478214*, 2009.

[190] European Union. Data Protection Directive (95/46/EC). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML. Accessed: 2016-04-18.

[191] European Union. ePrivacy Directive (2002/58/EC, as amended by 2009/136/EC). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:en:HTML. Accessed: 2016-04-18.

[192] European Union. General Data Protection Regulation. `http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf`. Accessed: 2016-06-27.

[193] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proceedings of the 8th ACM Symposium on Usable Privacy and Security (SOUPS)*, pages 4:1–4:15. ACM, 2012.

[194] Jaideep Vaidya and Chris Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 639–644. ACM, 2002.

[195] Eline Vanrykel, Gunes Acar, Michael Herrmann, and Claudia Diaz. Leaky Birds: Exploiting Mobile Application Traffic for Surveillance. In *20th International Conference on Financial Cryptography and Data Security (FC)*, pages 1–18. Springer Berlin Heidelberg, 2016.

[196] Timothy Vidas, Nicolas Christin, and Lorrie Cranor. Curbing Android Permission Creep. In *Proceedings of the Web*, volume 2, pages 91–96, 2011.

[197] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. Multi-task Representation Learning for Demographic Prediction. In *Proceedings of the 38th European Conference on Advances in Information Retrieval (IR)*, pages 88–99, 2016.

[198] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. A Field Trial of Privacy Nudges for Facebook. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pages 2367–2376. ACM, 2014.

[199] Yu Wang, Dingbang Xu, Xiao He, Chao Zhang, Fan Li, and Bin Xu. L2P2: Location-aware Location Privacy Protection for Location-based Services. In *Proceedings of the 2012 International Conference on Computer Communications (INFOCOMM)*, pages 1996–2004. IEEE, 2012.

[200] Xuetao Wei, Lorenzo Gomez, Iulian Neamtiu, and Michalis Faloutsos. ProfileDroid: Multi-Layer Profiling of Android Applications. In *Proceedings of the 18th ACM International Conference on Mobile Computing and Networking (MobiCom)*, pages 137–148. ACM, 2012.

[201] Rolf Wendolsky, Dominik Herrmann, and Hannes Federrath. Performance Comparison of Low-Latency Anonymisation Services from a User Perspective. In *7th International Symposium on Privacy Enhancing*

Technologies Symposium (PETS), volume 4776 of *Lecture Notes in Computer Science*, pages 233–253. Springer Berlin Heidelberg, 2007.

[202] Alan F Westin. Privacy and Freedom. *New York: Atheneum*, 1967.

[203] Shomir Wilson, Justin Cranshaw, Norman Sadeh, Alessandro Acquisti, Lorrie Faith Cranor, Jay Springfield, Sae Young Jeong, and Arun Balasubramanian. Privacy Manipulation and Acclimation in a Location Sharing Application. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 549–558. ACM, 2013.

[204] Wireshark. dumpcap - Dump network traffic. `https://www.wireshark.org/docs/man-pages/dumpcap.html`. Accessed: 2016-07-25.

[205] Ning Xia, Han Hee Song, Yong Liao, Marios Iliofotou, Antonio Nucci, Zhi-Li Zhang, and Aleksandar Kuzmanovic. Mosaic: Quantifying Privacy Leakage in Mobile Networks. In *ACM SIGCOMM Computer Communication Review*, volume 43, pages 279–290. ACM, 2013.

[206] Yonghui Xiao and Li Xiong. Protecting Locations With Differential Privacy Under Temporal Correlations. In *Proceedings of the 2015 ACM Conference on Computer and Communications Security (CCS)*, pages 1298–1309. ACM, 2015.

[207] Mingqiang Xue, Panos Kalnis, and Hung Keng Pung. Location Diversity: Enhanced Privacy Protection in Location Based Services. In *4th International Symposium on Location and Context Awareness (LoCA)*, volume 5561 of *Lecture Notes in Computer Science*, pages 70–87. Springer Berlin Heidelberg, 2009.

[208] Tun-Hao You, Wen-Chih Peng, and Wang-Chien Lee. Protecting Moving Trajectories with Dummies. In *8th International Conference on Mobile Data Management (MDM)*, pages 278–282. IEEE, 2007.

[209] Aristea M. Zafeiropoulou, David E. Millard, Craig Webber, and Kieron O'Hara. Unpicking the Privacy Paradox: Can Structuration Theory Help to Explain Location-based Privacy Decisions? In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci)*, pages 463–472. ACM, 2013.

[210] Ge Zhong, Ian Goldberg, and Urs Hengartner. Louis, Lester and Pierre: Three Protocols for Location Privacy. In *7th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 4776 of *Lecture Notes in Computer Science*, pages 62–76. Springer Berlin Heidelberg, 2007.

[211] Yuan Zhong, Nicholas Jing Yuan, Wen Zhong, Fuzheng Zhang, and Xing Xie. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 295–304, 2015.

[212] Kathryn Zickuhr. Location-Based Services. `http://www.pewinternet.org/2013/09/12/location-based-services/`. Accessed: 2016-08-16.

[213] Jonathan L. Zittrain. The Generative Internet. *Harvard Law Review*, 119(7):1974–2040, 2006.

Part II

Publications

List of Publications

Conferences and Workshops with Proceedings

1. Michael Herrmann, Carmela Troncoso, Claudia Diaz, and Bart Preneel. Optimal Sporadic Location Privacy Preserving Systems in Presence of Bandwidth Constraints. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 167–178. ACM, 2013.

   – See p. 187.

2. Michael Herrmann, Alfredo Rial, Claudia Diaz, and Bart Preneel. Practical Privacy-Preserving Location-Sharing Based Services with Aggregate Statistics. In *Proceedings of the 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks (WiSec)*, pages 87–98. ACM, 2014.

   – See p. 155.

3. Eline Vanrykel, Gunes Acar, Michael Herrmann, and Claudia Diaz. Leaky Birds: Exploiting Mobile Application Traffic for Surveillance. In *20th International Conference on Financial Cryptography and Data Security (FC)*, pages 1–18. Springer Berlin Heidelberg, 2016.

   – See p. 129.

4. Michael Herrmann, Mireille Hildebrandt, Laura Tielemans, and Claudia Diaz. Privacy in Location-Based Services: An Interdisciplinary Approach. *SCRIPTed*, 13(2):25, 2016.

   – See p. 95.

5. Michael Herrmann and Christian Grothoff. Privacy-Implications of Performance-Based Peer Selection by Onion-Routers: A Real-World Case Study Using I2P. In *11th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 6794 of *Lecture Notes in Computer Science*, pages 155–174. Springer Berlin Heidelberg, 2011.

6. Michael Herrmann, Ren Zhang, Kai-Chun Ning, Claudia Diaz, and Bart Preneel. Censorship-Resistant and Privacy-Preserving Distributed Web Search. In *14th IEEE International Conference on Peer-to-Peer Computing (P2P)*, pages 1–10. IEEE, 2014.

Technical Reports

1. Michael Herrmann, Fernando Pérez-González, Carmela Troncoso, and Bart Preneel. Possibilistic Location Privacy. Technical report, COSIC/ESAT, KU Leuven, 2016.

   – See p. 219.

2. Eline Vanrykel, Gunes Acar, Michael Herrmann, and Claudia Diaz. Exploiting Unencrypted Mobile Application Traffic for Surveillance. Technical report, COSIC/ESAT, KU Leuven, 2016.

3. Michael Herrmann, Fernando Pérez-González, Carmela Troncoso, and Bart Preneel. Description of the YaCy Distributed Web Search Engine. Technical report, COSIC/ESAT, KU Leuven, 2014.

Publication

Privacy in Location-Based Services: An Interdisciplinary Approach

Publication Data

Michael Herrmann, Mireille Hildebrandt, Laura Tielemans, and Claudia Diaz. Privacy in Location-Based Services: An Interdisciplinary Approach. *SCRIPTed*, 13(2):25, 2016.

Contributions

- Main author for Section 1-4 and Section 7.

Privacy in Location-Based Services: An Interdisciplinary Approach

Michael Herrmann[1], Laura Tielemans[2], Mireille Hildebrandt[2], and Claudia Diaz[1]

[1] KU Leuven ESAT/COSIC, iMinds, Leuven, Belgium
{name.surname}@esat.kuleuven.be
[2] Law, Science, Technology and Society, Vrije Universiteit Brussel
{name.surname}@vub.ac.be

Abstract. There exists a wide variety of location-based services (LBSs) that simplify our daily life. While engaging with LBSs, we disseminate accurate location data to remote machines and thus lose control over our data. It is well known that this raises significant privacy concerns as access to accurate location data may reveal sensitive information about an individual. In this work, we investigate the privacy implications of LBSs from a joint perspective of engineering, legal and ethical disciplines. We first outline from a technical perspective how user location data is potentially being dissiminated. Second, we employ the Contextual Integrity (CI) heuristic, an ethical approach developed by Helen Nissenbaum, to establish whether and if so, how, the dissemination of location data breaches the users' privacy. Third, we show how the concept of purpose limitation (PL) helps to clarify the restrictions on the dissemination of location data from a legal perspective. Our interdisciplinary approach allows us to highlight the privacy issues of LBSs in a more comprehensive manner than singular disciplinary exercises afford, and it enables us to contribute towards a better understanding among the relevant disciplines. Additionally, our case study allows us to provide two further contributions that are of separate interest. We address the problem of competing prevailing contexts without suggesting that the ensuing incompatability of informational norms can be resolved theoretically, even though it must be resolved in practice. This ties in with the difference between a legal approach that has to align justice with legal certainty and an ethics approach that aims to align prevailing social norms with moral reasoning. In the end, our interdisciplinary research shows how CI and PL are in many ways complementary.

1 Introduction

During the last decades, interdisciplinary research has gained significant popularity. While scholars typically work in their own self-contained and isolated domain within their community of experts, we refer to interdisciplinary research as that which brings together approaches of at least two different disciplines.

In this work, we report on our interdisciplinary research on the protection of location data. We tackle the problem from an engineering, legal, and ethical disciplinary perspective. While there are several reasons why interdisciplinary research can be fruitful,[3] we think interdisciplinary research is particularly useful for matters regarding data protection. In order to understand how data is created, transmitted and processed, one needs an understanding of technical systems, i.e. from the perspective of engineering. Yet, data processing is not merely a matter of technical possibilities, but also one of legal regulation. Hence one needs knowledge from the legal domain. Finally, we use Helen Nissenbaum's Contextual Integrity (CI) heuristic,[4] based on an ethical approach, as a middle ground between legal and technical assessments of privacy violations.

People use their smart devices, with their positioning capabilities, to engage in a wide variety of location-based services (LBSs). These services have in common that users must share their current whereabouts with a service provider to, for example, find nearby points of interest, share location data with friends, or get directions. It is well known that the ensuing mass dissemination of location data generates significant privacy concerns because location data reveals information about users that is potentially sensitive, difficult to anonymise,[5] and entities with access to accurate location data are able to make inferences about, for example, home/work address, income level, religious beliefs, sexual preferences or health issues.[6] To make things worse, behind the scenes, users share their location data with many more entities than they may be aware of and their location data may be used for purposes that they would never anticipate. This is mainly due to the current business model of many LBSs. In the case of free

[3]For a more complete assessment on the positive impact of interdisciplinary research, please refer to M Nissani, "Ten Cheers for Interdisciplinarity: The Case for Interdisciplinary Knowledge and Research" (1997) 34 *The Social Science Journal* 201-216.

[4]H Nissenbaum Privacy in Context: Technology, Policy, and the Integrity of Social Life (*Stanford University Press*, 2009); H Nissenbaum, "Respecting Context to Protect Privacy: Why Meaning Matters" (2015) *Science and Engineering Ethics* 1–22.

[5]P Golle and K Partridge, "On the Anonymity of Home/Work Location Pairs" (2009) *Pervasive Computing* 390-397.

[6]M Gasson et al., "Normality Mining: Privacy Implications of Behavioral Profiles Drawn from GPS Enabled Mobile Phones" (2011) 41 *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Review* 251-261.

services, service providers finance their service by either adding third party advertisers to their applications or by selling user data to data brokers.[7] Note that LBSs thus collect location data that is not necessary to deliver their service.[8]

This work is, to the best of our knowledge, the first that tackles the protection of location data from an engineering, ethical and legal perspective. From an interdisciplinary perspective, our article has four main contributions: first, the technical detail from an engineering perspective provides a substantive added value in connection with the ethical as well as the legal discipline. Second, our article serves as a reference for scholars of the involved disciplines to learn how the issue is addressed in the other two disciplines. Third, we identify a special relation between the ethical and the legal discipline, i.e. the connection between the concept of contextual integrity and purpose limitation. Fourth, our article serves as a case study on how to do interdisciplinary investigations of a data privacy matter. Additional to these interdisciplinary contributions, our work also provides valuable contributions to the CI heuristic and to the connection between the CI heuristic and data protection law. We show how the CI heuristic can be applied in a way that sensitises readers (and users) to what is at stake, and clarifies what the heuristic adds to the commonly stated opinion that location data can be sensitive data. Finally, our article discusses the legal concept of purpose limitation with respect to location data and argues its added value compared to contextual integrity.

Many of the terms used in this work have a precise meaning within one discipline, while evoking less precise connotations within the "other" discipline. For instance, in legal terms sensitive data refers to a specific category of data, summed up in art. 8 of the Data Protection Directive (DPD) and in art. 9 of the General Data Protection Regulation that will replace the DPR from May 2018.[9] This concerns personal data revealing e.g. ethnic or racial origin,

[7] C Timberg, "Brokers use 'billions' of data points to profile Americans" (2014) available at http://www.washingtonpost.com/business/technology/brokers-use-billions-of-data-points-to-profile-americans/2014/05/27/b4207b96-e5b2-11e3-a86b-362fd5443d19_story.html (accessed 7 July 2016). Under EU law this would probably be prohibited, though admittedly enforcement is lacking. For more information on the practice of data brokers: Federal Trade Commission, "A Call for Transparency and Accountability" available at http://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf (accessed 7 July 2016).

[8] Under EU law this is prohibited ex art. 6 DPD, unless the data are processed for a compatible purpose and a valid legal ground is applicable. Under US law such general restrictions do not apply, therefore US companies are less concerned about re-use of personal data.

[9] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281/31 (Data Protection Directive (DPD)), and

political belief. Being qualified as such has legal effect, since the processing of such data is by default prohibited. Location data is not sensitive data in this sense, though individuals may perceive their location to reveal sensitive information in a more general sense, and when correlated with other data location data may indeed result in data that is 'sensitive' in the sense of EU data protection law. Since this article is co-authored by computer engineers and lawyers, we refer to sensitive data in the general sense of the term and will specify when we use the term in the legal sense. Other examples of potential misunderstandings may arise, for instance, when engineers speak of "users", "clients", and "service providers", whereas lawyers speak of "citi"data subjects" and "data controllers. This is important because legal terms have legal effect and must therefore be used with precision. By specifying the legal meaning whenever relevant, we thus o contribute to the necessary dialogue between both disciplines on the challenges and solutions regarding the proliferation of location data.

2 Location-based Services

2.1 Overview

With LBSs, we commonly refer to services that take the user's current or past location as input to provide a service. Enabled by the mass usage of mobile devices with positioning capacities, LBSs have become very popular over the last decade.[10] Today, there is a wide variety of LBSs. Google Maps is arguably the most popular and most commonly used LBS. It allows users to get directions to almost any possible place and thus became a companion on most smartphones. Other highly popular LBSs are services such as geo-social networks (GSNs), where users share information about their current whereabouts in the form of a check-in and as a way to maintain their social network. Foursquare, one of the best-known GSNs, allows users to check into venues, leave comments, share their activity with their friends, and obtain rewards for their system usage. Although they also have a strong focus on social interaction, applications such as Highlight are different from applications such as Foursquare since their

Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation (GDPR)). Art. 99.2 GDPR stipulates that it will apply in the Member States of the EU from 18th May 2018.

[10] 49% of all users possessing a smart phone use LBS according to M Duggan, "Cell Phone Activities 2013" available at `http://pewinternet.org/Reports/2013/Cell-Activities.aspx` (accessed 8 July 2016)

focus is rather on displaying publicly available information on people in close proximity.[11] Other examples of LBSs include applications for directions in a city's public transport[12] and even Twitter, which has a functionality for adding location tags to one's tweets.

2.2 The Business Model of Location-based Services

Many LBSs are free to use[13] and although user data monetisation is not necessarily limited to free applications, such applications aim to monetise the information their providers gain about the users. This can be either information the users intentionally or unintentionally share while engaging with the LBS, or data the LBS learns by observing the user's activities. More importantly, monetisation may depend on inferences drawn from the data, possibly combined with data from other sources (Facebook, for instance, has contracts with large data brokers).[14]

2.3 Involved Parties

In the technical literature, a so-called "adversary" is everyone that may observe the user's location data. In the following, as a preparation for applying the CI heuristic, we elaborate on the most common entities that may observe user location data. With user we refer to the human engaging with an LBS by means of a mobile device (MD). The mobile device may be any piece of hardware manufactured by a hardware manufacturer (HM), such as a smart phone or tablet computer, which is able to determine its current location. On the software side, the mobile device may run software developed by multiple parties. The operating system (OS) is the software that is loaded on the mobile device when powered on. The operating system manufacturer (OSM) may be different from the HM of the mobile device. However, the HM may have modified the operating system in order to run its own services on the MD. Usually, the OS allows the user to download and install mobile applications (MA), which have been developed by a mobile application developer (MAD). The MA usually consists of the program written by the MAD - the core-application - but may also

[11] Other such services are for example: Sonar, Banjo and Kismet.

[12] Examples include OBB Scotty, Visit Paris by Metro, Tube Map London Underground.

[13] Exceptions are for example: C2G (carpooling), Caches (Geo-Caching), MeetupGroup (group meeting).

[14] K Hill, "Facebook Joins Forces With Data Brokers To Gather More Intel About Users For Ads" available at `http://www.forbes.com/sites/kashmirhill/2013/02/27/facebook-joins-forces-with-data-brokers-to-gather-more-intel-about-users-for-ads/` (accessed 8 July 2016).

include third party software (TPS) written by a third party software developer (TPSD). To summarise, a user's mobile device may run software provided by the following parties: the HM, OSM, MAD and TPSD. Along with common terminology, we use the term LBS to refer to the combined software of the MAD and the TPSD, and use the terms core-application and TPS only if we need to refer to this separately. The data that is sent and received by a mobile device is usually transferred by one or several network operators (NO). Depending on how the MA or TPS is implemented, the NO has access to the user's data in encrypted or unencrypted form. Finally, Government entities (Gv), such as law enforcement agencies, tax authorities or intelligence agencies, may be able to obtain access to the user's data by means of warrants that allow for eavesdropping or hacking.[15]

3 Introducing the Contextual Integrity (CI) Heuristic

Contextual integrity (CI) is a concept introduced by Helen Nissenbaum to better understand what is at stake with privacy and to uncover the issues that can arise when sharing data. In her book, Privacy in Context,[16] Nissenbaum introduces the CI decision heuristic as a tool to determine whether a new socio-technical practice violates informational norms and thereby infringes privacy. The CI heuristic considers the interplay between context, roles, actors, attributes, values, informational norms, and transmission principles. The key idea of this framework is that information flows between people, and between people and other entities, occur in a specific context, taking note that this context implies specific informational norms and transmission principles. Such norms and principles may, for instance, determine how the exchanged information can be further disseminated. Specifically, user privacy is breached if information is shared in disregard for a transmission principle implied in the context where the information was first shared. For example, in the context of professional advice, where a client might share information with her lawyer, one of the transmission principles will be that the information is confidential. If the lawyer shares this information with the client's colleagues, the contextual integrity would be violated, because the transmission principle under which the information was first exchanged does not foresee a further information flow from the lawyer to such others.

[15] A Landau "Making sense from Snowden: What's significant in the NSA surveillance revelations" (2013) 4 *IEEE Security & Privacy* 54-63.

[16] H Nissenbaum *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Palo Alto: Stanford University Press, 2009).

Table 1: Summary of the parties involved

Entity	Acronym	Description
Mobile Device	MD	Smart phone or tablet computer with positioning capabilities
Hardware Manufacturer	HM	Company that manufactures MDs such as Samsung or Apple
Operating System	OS	Software which enables the usage of the hardware by MAs
Operating System Manufacturer	OSM	Company that developed the OS
Mobile Application	MA	Software that runs on top of the OS
Mobile Application Developer core-application	MAD	Company that developed the MA Software developed by the MAD
Third-Party Software	TPS	Any additional software that is integrated with the MA
Third-Party Software Developer	TPSD	Company that developed the TPS
Location-based Service	LBS	Service that utilizes the geo location of users consisting of core-application and TPS.
Location-based Service Provider	LBSP	Legal and technical entity running the LBS
Network Operator	NO	Company that runs the physical communication infrastructure
Government	GV	Any governmental institution with legal right to access companies databases or communication infrastructure

The framework of contextual integrity refines the predictability of informational flows and privacy expectations; it is not so much the average or the "general" behaviour of people that will create these expectations, but rather the normative framework that shapes what is considered as "reasonably expectable". CI is not about the regularity of behaviour but about legitimate expectations. The importance attached to existing informational norms and transmission principles may be qualified as conservative. However, the point of CI is not that informational norms should not change, but that such change should be determined by those sharing a context and not imposed by socio-technical innovations. So, even if the CI heuristic can be qualified as conservative in some respects, it acknowledges that rapid technological developments may have advantages for society that justify change. This entails that emerging socio-technical practices may allow for new informational norms that challenge and reconfigure contextual integrity. However, not anything goes, and such reconfiguration requires careful deliberation. Based on this, those concerned may come to the conclusion that these new informational norms actually benefit society. This, however, necessitates appropriate safeguards to prevent, reduce and/or compensate the potential harm for those that suffer the consequences of newly distributed information flows.

The CI heuristic includes nine steps that we briefly outline in the following:

1. Describe the new socio-technical practice in terms of information flows.
2. Identify the prevailing context of the new practice.
3. Add general information about the sender, receiver and referent that are part of the practice.
4. Identify the transmission principles of the prevailing context.
5. Identify applicable entrenched informational norms.
6. Conduct a prima facie assessment on whether contextual integrity is breached by analysing the results of the previous steps and by observing whether entrenched norms have been violated.
7. Investigate what harms or threats to autonomy, freedom, power structures, justice, fairness, equality, social hierarchy and democracy are implicated by the new socio-technical practices.
8. Analyse how the practices directly impinge on values, goals, and ends of the identified context.
9. Conclude the CI heuristic and analyse whether the violation of entrenched norms is justified by the benefits of the new practice for the society,

considering the fact that harms and benefits may be distributed in ways that disadvantage parties that are already disadvantaged.

4 Applying the CI Heuristic

4.1 Choosing a Context: Gateway or Vanishing Point

In the following, we employ Nissenbaum's contextual integrity (CI) heuristic for the socio-technical practice of LBSs. A key challenge is the identification of the prevailing context, required in step 2. In a sense the choice of the relevant context is the gateway as well as the vanishing point of the entire exercise. The idea behind contextual integrity is that privacy cannot be determined in general, but depends on the context. In her book, Nissenbaum discusses a series of specific contexts, which enable an investigation of relevant information flows in a great level of detail and a concrete evaluation of specified scenarios.

The usage of MDs and MAs has become ubiquitous in our daily life and therefore users engage with LBSs in many different situations. Each situation involves various contexts, e.g. checking into a favourite café during a work break (leisure, work); getting directions during weekend trips (holidays, leisure); searching for a restaurant while being in an unfamiliar city during a business trip (business, leisure). Instead of discussing all the possible contexts, we identify the context of travel as the prevailing context of people engaging with LBSs. Since "travel" can be subdivided into business (or other work-related) travel, leisure travel (holiday) and migration (including illegal immigration), we will focus on business and leisure, as migraation entails a very different set of informational norms and transmission principles.

Our analysis can be seen as a case study that should sensitise readers (and users) to the privacy risks of LBSs, showing what the heuristic adds to general statements that qualify location data as sensitive data.

4.2 Socio-technical Practice in Terms of Information Flow

The first step of the CI heuristic requires us to explain how a new socio-technical practice impacts information flows by either changing existing ones or by generating new ones. We will analyse the information flows of LBSs under the third step, when discussing the participants that exchange information. Here we discuss information flows in terms of three types of personal data, as described by the World Economy Forum (WEF): as volunteered, observed or

inferred data. Volunteered data is data that an individual deliberately shares and transmits over a communication network, for example postings, credit card details or photographs. In the process of sharing volunteered data, additional data are captured by service providers and third parties, often without the user being aware of that collection, even if she has provided consent for this by, for example, agreeing to the terms of service. This additional data is referred to as observed data. Observed data often consists of behavioural data, such as clickstream or location data. Finally, inferred data is the output of data analysis, which can be based on either volunteered or observed data or both.

For a proper understanding of inferred data, we refer to Hildebrandt, who introduces the notion of Big Data Space (BDS) to explain the complexity of the influences of inferred data, due to the fact that "[...] databases are fused or matched, while the knowledge that is inferred can be stored, sold and re-used in other databases".[17] When third parties that the user is unaware of observe volunteered data, that volunteered data may also be considered as observed data as will behavioural data, which may be similarly observed by parties the user is not at all aware of. We therefore emphasise that a crucial aspect of the concept of inferred data is that it is difficult, if not impossible, to foresee how volunteered and observed data are used to create inferred data. In line with that, we note that data that is observed while a person is using the LBS may be combined with data from other sources, for example, with data revealed in one of her online social network profiles or data stored with data brokers. Indeed, inferred information usually serves to learn more details about a user. This may be for the purpose of personalisation or advertising and we thus argue that inferred information typically includes interests, habits and, for instance, income level or health risks of an individual person, based on how her behaviour matches inferred profiles mined from Big Data Space.

4.3 Identifying Prevailing Context

We observe that the context of travel is naturally related to the usage of LBSs, because traveling includes a person's journey, stay and departure from certain locations, which may be known, anticipated or inferred by other people, companies, governments and computing systems. Furthermore, choosing the context of travel nicely illustrates the fact that users employ LBSs while being busy with all kinds of activities, which implies that they are probably engaged in different contexts at the same time. For example, a check-in during a business

[17] M Hildebrandt, "Location Data, Purpose Binding and Contextual Integrity: What's the Message?" (2014) in L Floridi (ed) *Protection of Information and the Right to Privacy-A New Equilibrium?* Law, Governance and Technology Series, vol. 17, (Springer; Dordrecht) 31-62, at 35.

meeting is part of a business context, but the check-in is also definitely an action that is engaged in the user's travel context, as a person must travel to reach the location of the meeting.

In the Western world, people tend to take freedom of movement within and between countries for granted. EU citizens, for instance, probably assume that one does not require permission for moving from one place to another within the EU, and expect that no questions will be asked when doing so. Similarly, they may think that they have a right to travel unmonitored from surveillance. This shows what Western people expect in terms of informational norms and transmission principles. Arguably, specific forms of transport, such as traveling by air, are monitored more closely than other types of transport. This is related to the potential for terrorist attacks or illegal immigration. Within the EU, the expectations around freedom of movement are tested in times of terrorist suicide attacks and mass immigration following the crisis in the Middle East. Although we focus on the subcontext of leisure (and business travel), informational norms and transmission principles, in the context of such travel, will be challenged by threats to public security. This implies that the context of public security may overlap with that of travel, which highlights that the choice of the prevailing context has major implications for the outcome of the heuristic. In section 5 we will return to this point.

4.4 Identifying Sender, Receiver and Referent

We consider two cases for the identification of the sender, receiver and referent of information flows in LBSs. The first case illustrates the sender, receiver and referent from the perspective of a normal user, whereas the second case illustrates what is going on "behind the scenes". The key difference is that in the first case, the Location-based Service Provider (LBSP) is seen as a data processor that operates to serve the user, whereas in the second case, the LBSP, along with other entities, turns out to be a data controller that is processing the data for their own benefit.

Figure 1 shows the functionality of an LBS from the normal user's perspective. First, user A sends a request to the LBS, revealing information that is necessary to obtain the service and additional implicit information. For LBSs such as Google Maps, the LBSP replies to the request, while storing and analysing the received (and observed) data "to further improve its service". For LBSs such as Foursquare or Highlight, the LBSP may additionally send location data of user A to other users of the LBS. In such a case, the user typically defines the set of intended receivers.

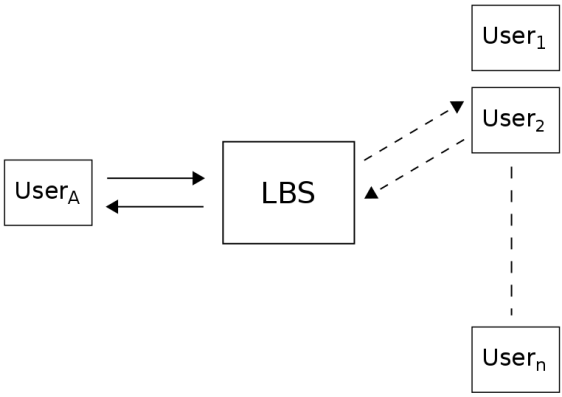


Figure 1: LBSP from a normal user's perspective (user is data controller).

Figure 2 illustrates the more comprehensive setting of how location data is disseminated in LBSs. There are two major differences to the previous case: First, we now understand that, besides the LBSP, all other entities that run software on the MD, such as the TPSD or the OSM, may access the user's location data.[18] Second, we note that location data is not only transmitted when the user actively uses the LBS, such as during a Foursquare check-in, but may also be accessed while the LBS is running in the background. In the latter case as the information flow is not intentionally sent by the user, the receiving party can be seen as the sender (it actually sends the information from the sender's device or application to itself).

While there is enough evidence that sensitive data is transmitted from smart devices to remote destinations,[19] there is little knowledge on how the different entities depicted in Figure 2 utilise the user's location data. The LBSP, TPSD, OSM and HM may use the data (possibly (pseudo)anonymised) to optimise and personalise their services, to combine it with other data from their services,[20] or to sell the data to other third parties such as data brokers or advertising networks. Data brokers collect, aggregate and infer information from big data,

[18] T Book, Theodore, Adam Pridgen, and Dan S. Wallach. "Longitudinal analysis of android ad library permissions." arXiv preprint arXiv:1303.0857 (2013).

[19] X Wei et al, "ProfileDroid: Multi-Layer Profiling of Android Applications" (2012) *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*.

[20] From a legal perspective any such processing requires of a legal ground and a specific, explicit and legitimate purpose that restricts the use of personal data. However, before discussing the legal constraints, we check what is technically possible and feasible in view of current business models.

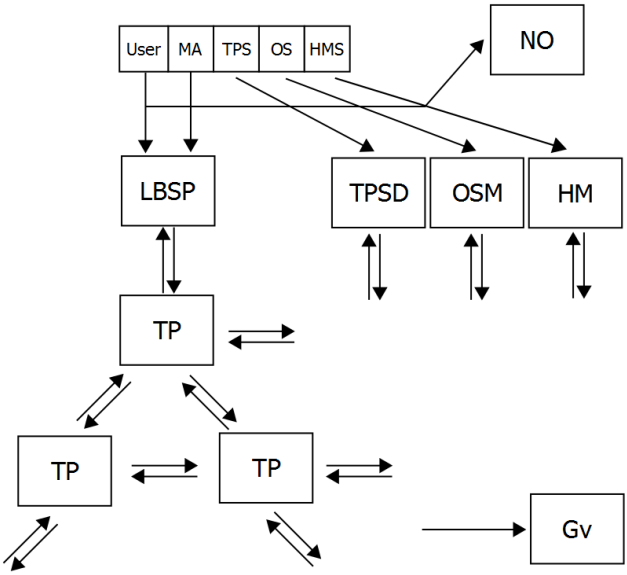


Figure 2: Overview of entities potentially learning a user's location data (user is not data controller).

while online advertising networks monetise users' behavioural data via targeted advertisements[21]. Two further entities may receive location data of users engaging with LBSs. First, the network provider (NP) since it is the provider of the communication infrastructure. Secondly, the government (Gv) since it may gain access to the data via traffic eavesdropping,[22] or by obtaining access to the LBSP's databases through the application of a warrant, or by hacking, or even simply by using the service themselves.[23]

Finally, Figure 3 illustrates that there is an actual information flow back to the user. This may be in the form of online advertisements or service personalisation that are based on the profile that has been created about the user. Location data is one of the few cases where we have evidence that it is being used for

[21] L Olejnik et al, "Selling off Privacy at Auction" (2014) *Network & Distributed System Symposium* 1-15

[22] M Lee, "Secret 'BADASS' Intelligence Program Spied on Smartphones" available at `https://theintercept.com/2015/01/26/secret-badass-spy-program/` (accessed 8 July 2016)

[23] C Paton, "Grindr urges LGBT community to hide their identities as Egypt persecutes nation's gay community" available at `http://www.independent.co.uk/news/world/africa/grindr-urges-lgbt-community-to-hide-their-identities-as-egypt-persecutes-nations-gay-community-9757652.html` (accessed 8 July 2016). Note that in most jurisdictions warrants are required for remote access to computing systems.

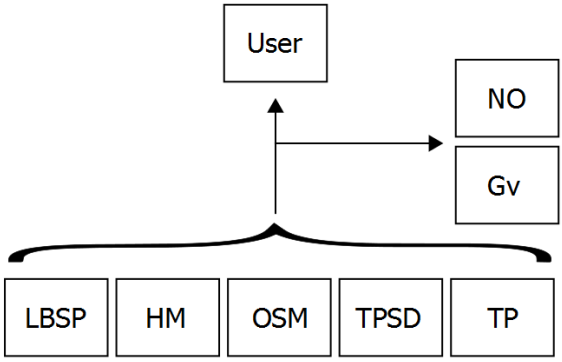


Figure 3: Information flow back to the user in form of personalization.

the creation of such a user profile.[24] But these profiles do not necessarily imply an information flow back to the user as they may also be used to exclude an individual from a service, premium, from employment, education or simply from access to a building or compound. Similarly, police intelligence increasingly employs location data for crime mapping, thus targeting individuals based on their location at so-called "hot spots".

4.5 Identifying Principles of Transmission

Although there are many reasons for people to travel, most types of offline travel activities - i.e. travel without using online LBSs - include the following information flows: a company, such as a railway company or hotel, sells tickets or rents accommodation that indicate the traveller's current, past or future location at certain times. Due to space constraints, we elaborate on the transmission principle in relation to the examples of a train company and a hotel but note that our reasoning can be applied to many other scenarios, such as any kind of travel activity that requires a reservation. Activities that require no reservation, like buying museum tickets with cash, may entail fewer or less extensive information flows in an offline scenario.

A leisure trip usually includes booking a means of transportation, booking hotels and telling family or friends about the trip. For the means of transportation, let us consider the example of a train ride. The information exchanged with the train company depends on the way the ticket was purchased. In the most

[24] Joseph Turow et al, "Americans Reject Tailored Advertising and three Activities that Enable it." (2009) *SSRN* 1478214.

anonymous case, the ticket is purchased at a counter or machine and paid with cash. In this case no information is exchanged but that an unidentified passenger intends to travel a certain route. If, however, a credit card is used, then more information is included in the sale, allowing the train company to link travels booked with the same credit card.[25] Furthermore, the credit card company will record the location of the user along with the purchased product. While the credit card company may have legitimate reasons for collecting this information, such as fraud detection, we note that this practice may not be part of users' legitimate expectations and would justify a CI analysis of its own. Finally, if the purchase was completed using some other account information, such as through a membership in a bonus program or via an online account, the information flow between the traveller and a railway company will include additional details on the traveller such as age, billing address and details of the journey such as time/place of departure/arrival. Regardless of how the booking was completed, the flow of information is governed under several transmission principles: reciprocity and consent as the information is necessary for the monetary exchange for a service; necessity, accuracy and completeness of data because without the correct information, the railway company is not able to issue the correct train ticket; purpose binding because the user has a reasonable expectation that the railway company will only use this information for selling the ticket; and confidentiality because the traveller expects to keep records about her journeys private.

The information a traveller needs to reveal to the hotel is governed by the same transmission principles. With respect to the traveller's family and friends, a traveller may inform them with various details about her journey. The involved transmission principles are reciprocity since it is common that friends and colleagues inform each other of their travel plans; or confidentiality when it is clear that the traveller does not want a friend to pass on information about her travel plans. We note that anyone with information on someone's travel could infer more or less accurate real-time whereabouts of this person. For instance, a person knowing that someone stays in a hotel for a certain period is able to infer that this person is likely to be in the hotel during night. A similar argument can be made based on knowing that someone takes a train at a specific time.

[25] EU data protection law may prohibit this, but the technology does allow for such linking of different data. Due to the opacity of actual data flows, caused by complexity as well as trade secrets or IP rights, it is difficult to establish which data are actually linked. Precisely for that reason it is pertinent to take into account what is technically feasible. When composing legislation that prohibits such behaviour it seems wise to focus on preventing that such exchanges are feasible (technical and economic disincentives), e.g. by imposing data protection by default coupled with adequate enforcement mechanisms (administrative fines and tort liability). This is precisely what the upcoming GDPR does (art. 23 GDPR imposes a legal obligation to develop data protection by default and design; art. 83.6 determines that fines of up to 4% of global turnover can be imposed by the supervisory authority).

In an offline context, however, such inferences are always based on guesses and have a limited accuracy.

We note that – especially in data-driven scenario's - the outlined primary information flows may result in subsequent information flows directed to tax authorities (for fraud detection), justice and police (for criminal investigations) and intelligence services (for the prevention of threats to national security or the gathering of intelligence information). In this section we focus on the primary information flows, generated in a commercial context. We also note that in the subcontexts of work-related travel and immigration extra information flows may occur, as employers or immigration services may either tap into existing information flows or establish their own infrastructure to keep check of the location of their employees or potentially illegal immigrants.

4.6 Locating Applicable Entrenched Informational Norms and Points of Departure

As in the case of our reasoning about transmission principles, we argue that the entrenched informational norms and points of departure are similar for most types of leisure travel, though it should be clear that in the sub-contexts of work-related travel and immigration other informational norms will be at stake. In the following we distinguish between two cases of information collected in offline traveling activities. First, we address the expectations of how a service provider will use the location data. Second, we examine the circumstances under which a company is allowed to share information on the traveller's location with other entities.

A train traveller or hotel guest can expect that the information she provided to the railway or the hotel is in some sense used for an internal review of their business processes. The traveller or hotel guest should not expect an unknown business model based on the monetisation of her location data. For example, every company needs to aggregate basic statistics about its customers, but a railway company sending advertisements to their customers that include detailed information about the travel behaviour of a particular customer would not be in line with entrenched informational norms. A judgment on whether these norms are violated if the customer data is transferred to another organisation depends, amongst others, on the type of organisation. For example, selling customer information to a data broker can be problematic even if they are pseudonymised, because this implies that data brokers learn about the traveling behaviour of people with whom they never had any business. However, access to customer data based on a legal obligation, such as tax law, may be acceptable to the extent that tax authorities must be capable of inspecting a company's

accounting system. Further problems arise if law enforcement, immigration services and intelligence services gain unrestricted or mostly invisible access to these types of data. Several years ago, the Dutch navigator TomTom found its reputation damaged when users found out that their aggregated data were used to predict traffic violations at certain road tracks.[26] Such examples clarify that in a travel context, people do not expect their location data (and what can be inferred from them) to end up with third parties whose access they cannot foresee. Some would claim that even if they would – cynically – expect this, it would still violate their reasonable expectation of privacy concerning this travel data.

4.7 Prima Facie Assessment

The purpose of the prima facie assessment is to determine whether there are red flags attached to changes or violations of entrenched informational norms due to a new socio-technical practice or if an entirely new information flow is being created. It is easy to see that both are the case. LBSs change the described information flows, because they generate, store and mine not merely volunteered data (as was always the case) but also massive flows of behavioural data that were simply non-existent before the massive employment of web services. Buying a train ticket hardly gave the train company any insight on the buyer's location, besides perhaps the location of the purchase. Offering the purchase of tickets via an app, however, may allow the train company to constantly track the user.[27] Entirely new information flows are being generated since numerous companies, such as LBSP, TPSD, OSM, HM and third parties, potentially gain direct access to a traveller's location data whenever she is using LBSs.

4.8 Evaluation I

In the first part of the evaluation, we assess how the new socio-technical practice generates threats to autonomy, potential harms, how it affects the existence of freedom, power structures, justice, fairness, equality, social hierarchy, and democracy.

[26] A Preuschat and H Luttikhedde, "TomTom to Bar Police Data Use" available at `http://www.wsj.com/articles/SB10001424052748703922804576300390277973646` (accessed 8 July 2016).

[27] JP Achara et al, "Detecting Privacy Leaks in the RATP App: How we Proceeded and what we Found" (2014) 10 *Journal of Computer Virology and Hacking Techniques* 229-238.

As we have laid out in Section 2.3 and Section 4.4, various entities may gain access to accurate location data. This may be either because the user herself transmits location data (volunteered data) or because some software, running on the user's MD, is covertly transmitting location data to second or third parties (observed data, whether with or without consent, whether legal or illegal). From a privacy perspective this is highly problematic, because the collection and analysis of accurate location information allows for the inference of highly sensitive information about an individual. Users of LBSs may be confronted with these inferences either immediately or in the future, where the 'future' could be minutes, days, weeks and even up to years and decades ahead since location data, once available, may persist in Big Data Space.[28]

The power structure and fairness in the market is threatened, because the user may be disadvantaged when negotiating with a company that has access to her location data. For example, an LBS user may find herself in a situation where she has to pay a higher health insurance premium or is even rejected for insurance. If an LBSP, or any TPSD whose software is embedded in the LBS, tracks the user's whereabouts it may sell this location data to a data broker specialised in health risk assessment. The analysis of this data broker may be based on the user being relatively more often in health related facilities, which may result in assigning this user a higher health risk. An insurance company that consults the data broker may then decide that it either increases premiums or refuses insurance all together. Note that in this scenario the user would have no idea how the insurance company reached its decision, nor would she be able to correct false predictions[29] about her health. One could think of numerous other examples of how the user would be disadvantaged when negotiating with an entity that knows her whereabouts. The traveller's autonomy is at stake, because she is unable to inspect and correct the inferences made in the BDS. This may have severe consequences, including unwarranted or even prohibited discrimination.

There are further implications in recent revelations which have shown that intelligence agencies have a strong interest in location data to enhance their surveillance systems.[30]. While one can only speculate on how this data is being

[28] On the threats that emerge in the combination of location and other types of tracking see e.g. JH Ziegeldorf et al, "Privacy in the Internet of Things: Threats and Challenges" (2014) 7 *Security and Communication Networks* 2728–42.

[29] Inferences made by behavioural advertisers have shown to be wrong: R Ashwini et al, "What do They Know About me? Contents and Concerns of Online Behavioral Profiles" (2015) *arXiv preprint arXiv:1506.01675*.

[30] B Gellman and A Soltani, "NSA Tracking Cellphone Locations Worldwide, Snowden Documents Show" available at `http://www.washingtonpost.com/world/national-security/nsa-tracking-cellphone-locations-worldwide-snowden-documents-show/2013/12/04/5492873a-5cf2-11e3-bc56-c6ca94801fac_story.html` and J Ball, "Angry Birds and 'Leaky' Phone Apps Targeted by NSA and GCHQ for User Data" available at

used, the very existence of such programs threatens society in several ways. A government with access to location data of a large majority of the population is able to determine which people have contacts with refugees or protesters and could thus effectively prohibit or even preempt demonstrations or free speech.[31] These threats to democracy and justice are also simultaneously threats to freedom and equality since authoritarian states may imprison or discriminate people based on their location data (mobility) and the inferences drawn from them. Finally, the US government uses location data to coordinate drone strikes to kill people qualified as enemy combatants without due process and with disastrous results for those standing around (qualified as collateral damage). Though this concerns issues of extraterritorial jurisdiction and international relations, which fall outside the scope of this article, we note that it has a major impact on attempts to establish democracy and the rule of law at the global level.

4.9 Evaluation II

In the second part of the evaluation, we elaborate on how the socio-technical practice of LBS directly impinge on the values, goals, and ends of the context of travel. This particular context is highly relevant for the freedom of movement, and for the fundamental right to be left alone, which entails that by default, it is no one's business where a person travels. However, the usage of MDs and LBSs enables the collection of accurate location data that provides evidence on where a traveller went, for how long and how often. Therefore, the freedom and anonymity of traveling no longer exists as they did, because unforeseeable commercial enterprises as well as governmental agencies may gain access to this information, aggregate it and make it machine searchable. Furthermore, whereas our journeys usually had no impact on other contexts of our life such as business or healthcare, due to the business model of LBSs and increasing governmental surveillance activities, inferences made from location data may reach parties outside the context of travel, with unforeseeable consequences in other contexts, such as employment, insurance and law enforcement.

`http://www.theguardian.com/world/2014/jan/27/nsa-gchq-smartphone-app-angry-birds-personal-data` (both accessed 8 July 2016)

[31] On the use of location data by government agencies see e.g. SJ Nouwt et al "Power and Privacy: the Use of LBS in Dutch Public Administration" in B Van Loenen, JWJ Besemer and JA Zevenberger (eds) SDI Converngence. *Research, Emerging Trends, and Critical Assessment* (Optima Rotterdam; Graphic Communication 2009) 75–87.

4.10 Conclusion

Our evaluation shows that the new socio-technical practice of LBSs comes with significant threats for their users. Organisations may gain access to accurate location information, which enables them to infer sensitive information. These inferences may be used for unknown purposes, and transferred or sold by other, unforeseeable organisations. This may lead to an individual being disadvantaged due to the knowledge these organisations have about her. Even worse, people have no practical means to realise or escape this situation since the collection, transfer and processing of their location data is entirely opaque. Furthermore, individuals are mostly unaware of and unable to correct false inferences, as well as unable to object to unfair targeting. Finally, as we have argued, the new socio-technical practice introduces threats not only to individuals, but also to society at large.

We acknowledge that LBSs are undoubtedly useful and may enhance our daily lives. Solutions like PocketFinder may help to prevent harm and danger to children and elderly. Location data may serve to identify credit card fraud or to find stolen vehicles. If the collection of location data happens in a transparent way with obvious ways for an individual to opt-out and to control inferences being made from her location data, we would accept that the new informational norms may be beneficial and should be embraced. However, the hidden collection of location data by numerous entities capable of using this data in unforeseeable ways is not justifiable and clearly violates the integrity of the context of leisure travel. As this implies a violation of the reasonable expectation that people have concerning their privacy, this should be considered an unjustified breach of privacy.

5 The Complexities of Intertwined Contexts

LBSs are ubiquitous in their nature and becoming increasingly omnipresent. As we have seen above, a crucial step in Nissenbaum's decision heuristic is defining the prevailing context. This raises the question of what counts as a context and how one can identify the prevailing context. A context can be seen as an overarching social institution, an architecture that attributes roles and mutual expectations, which guide the interactions between people. Institutions – in the sociological sense of the term – determine how behaviours are "read" and which actions are deemed appropriate. A context in this sense entails the institutional environment that hosts more specific institutions, such as for instance marriage, church, school or corporate enterprise, that are part of, respectively, the contexts of family life, religion, education or economics.

Depending on the context, different informational norms and transmission principles will apply and prescribe the appropriateness of the information flows. As indicated in the previous section, this entails that the choice of the prevailing context is both the gateway and the vanishing point of the CI decision heuristic.[32]

The assessment of contextual integrity and its violation becomes complex when the assessor has to deal with two or more (sub)contexts. Multi-layered and overlapping contexts make it rather difficult to pinpoint only one "right" context as the prevailing one. Not only will a service rendered to the subject trigger sharing of location data, but, as we have seen above, location data is shared on many levels simultaneously and subsequently. Such sharing may involve various contexts, for instance, a commercial context (behavioural advertising based on shared location data captured by the LBS), a health context (LBS activated when traveling to a hospital) or a work-related context (LBS activated when commuting or travelling for one's professional occupation). This renders the analysis of an isolated information flow inadequate as a criterion to decide on the prevailing context. Therefore, we have opted for the travel context, which is at stake in each of the scenarios where LBSs are employed.

Even so, one could argue that in each of the scenarios the overarching context is that of economics, which would imply competing prevailing contexts. With an eye to advancing "the state of privacy", Nissenbaum has revisited her theory to defend it against misconceptions by policy makers. Recalling her definition of context as a social domain or social sphere that provides organising principles for legitimate privacy expectations, she differentiates her understanding of context from three others. First, from context as a technology system or platform, second from context as a business model or business practice, and, third from context as sector or industry.[33]We are not sure that this resolves the problem of competing prevailing contexts, since commerce is itself a social domain that penetrates many other social domains. The problem may be that social domains are (no longer) mutually exclusive. Since choosing a different prevailing context might lead to a completely different outcome, as entrenched transmission principles and informational norms will differ. This leads to the question of who gets to decide on the choice of the prevailing context: the service provider, the data subject, the people, the social scientist, the ethicist? Who is the assessor?

If we take into account that different contexts lead to different results, qualifying

[32]In that sense context seems to be given rather than the result of struggles and reconfigurations, see the praise and the critique of R Bellanova "Waiting for the barbarians or shaping new societies? A review of Helen Nissenbaum's Privacy In Context (Stanford: Stanford University Press, 2010) (2011) 16 *Information Polity* 391-395, notably at 394.

[33]H Nissenbaum (2015), see note 6 above.

a context as prevailing has far-reaching implications. One could therefore conclude that the choice of context is not only the gateway to the CI decision heuristic, but also its vanishing point. Once the heuristic has progressed beyond the second step, potentially conflicting norms and principles that pertain to other contexts have become invisible. In our case, to come to a sustainable conclusion as to violations of CI in the case of LBSs, we would need to develop the decision-heuristic for a number of relevant contexts, such as e.g. economics, travel, health and education, depending on the situation or scenario. This could lead to conflicting transmission principles and informational norms and would basically require deciding which context should be qualified as the primary or overruling context amongst several prevailing contexts. It is not obvious that this decision can be made at a general level for each instance where LBSs are employed. If that means that we must decide per situation which context is primary, the heuristic no longer provides clear guidelines to evaluate the impact of LBS on contextual integrity. We believe, however, that this rather tricky challenge is not something we should resolve. On the contrary, it sensitises us to the fact that location is no longer equivalent with context, as it perhaps used to be (with separate spaces for work, family life, religious worship and leisure time). It also means that the principle of purpose limitation may be a more apt criterion to decide on the legitimacy of an information flow (as well as other types of processing), taking into account the context(s) on a case-by-case basis.

6 Contextual Integrity and Purpose Limitation

6.1 The legal obligation of purpose limitation (PL

Contextual integrity and the legal obligation of purpose limitation (PL) share some common ground. Both require for the flow and distribution of personal data to be appropriate, assuming that both collecting and further processing of personal data should be limited. Both look beyond collection, though PL regards any form of personal data processing (including analysis) while contextual integrity seems to be restricted to transmission of personal data. Also, the CI decision heuristic concerns an ethical inquiry, whereas PL is a legal obligation within the jurisdiction of the European Union (EU). Before analysing the CI decision heuristic from the legal perspective, we will first explain the background and content of the legal obligation of PL.

The legal obligation of purpose limitation derives from the OECD Fair Information Principles, as formulated in 1980.[34] Within the context of EU data

[34]OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, 23 September 1980, updated in 2013, available at `http://www.oecd.org/sti/ieconomy/`

protection law, it seems to be one of the most crucial characteristics of the protection of personal data (as specified in art. 8 of the Charter of Fundamental Rights of the EU, which defines the fundamental right to data protection). We will now discuss five points to be made for a proper understanding of PL within the legal framework of the EU.

First of all, the 1995 EU Data Protection Directive (DPD), as well as the upcoming General Data Protection Regulation (GDPR),[35] require that processing of personal data is based on a specific and legitimate purpose made explicit by the data controller. The data controller is defined as the entity that decides the purposes and the means of personal data processing, and is held liable for compliance with the DPD (art. 6).[36] "Purpose" thus serves, first, to define who is responsible for personal data processing, while also, second, providing the person whose data are processed (the data subject) with a clear idea of what they can be used for. Purpose thus determines the relationship between sender and receiver of the data, and includes the case of behavioural data, where the receiver can be seen as actually sending the data to itself. PL, then, seals the relationship between data subject and data controller, and forces the controller to somehow make sure that the data subject "knows" how her data may be used. Depending on the circumstances, the purpose can for instance be specified in a statute (e.g. if the tax administration requires LBS data to determine fraud), or be announced in the terms and conditions of a web service (e.g. specifying that location data will be used for marketing purposes). In the latter case, we may doubt whether the users of the service are aware of the purpose, considering the lengthy privacy policies that hide such information. Especially if the legal ground for the processing concerns "the legitimate interest of the data controller" (as in the case of Google's search engine), the legitimacy of the processing will have to be evaluated on a case-by-case basis to check to what extent the rights and interests of the data subject are being respected.[37]

oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm (accessed 14 August 2016).

[35] See note 12 above

[36] Art. 6 DPD (and art. 5 GDPR) stipulates purpose limitation and holds the data controller liable. Art. 2(d) DPD (and art. 4.7 GDPR) specifies that the data controller is the entity that determines the purpose and means of the processing of personal data.

[37] This relates to the so-called f-ground for legitimate processing of personal data (art. 7.f DPD, art. 6.1.f GDPR), which allows for processing on the ground of necessity "for the purposes of the legitimate interests pursued by the controller of by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection (. . .)." In CJEU 13th May 214, C-131/12 (Google Spain v Costeja González), the European Court of Justice of the EU decided that Google Spain processes the personal data of those listed in the search results on the basis of the legal ground of art. 7.f and found that the economic interest of the search engine provider can - in general - not overrule the fundamental rights of the data subject.

Second, art. 6.1(a) of the upcoming General Data Protection Regulation states that processing is allowed if "the data subject has given consent to the processing of his or her personal data for one or more specific purposes." In the case of LBSs, this requirement is often sidestepped by demanding consent for the re-use of the location data for the purpose of marketing or monetisation (though the latter is hardly ever made explicit). If such additional consent is refused, the service provider usually simply refuses to contract. Art. 7.4 of the GDPR, however, stipulates that "[w]hen assessing whether consent is freely given, utmost account shall be taken of whether, inter alia, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract." This would basically mean that PL cannot be eroded by forcing consent for secondary purposes on those who wish to access an LBS. When this stipulation comes into force (in May 2018), LBSs may have to change their business model by, for instance, charging a fee instead of relying upon the monetisation of location data.

Third, the current DPD requires that data controllers do not process personal data for purposes that are incompatible with the purpose they have explicitly specified when initiating the processing of the data (at the time of first collection). Though it seems that the OECD Guidelines allowed for a person to give consent to process her data for other or even any purposes,[38] the DPD clearly stipulates that purpose binding holds, even in the case of consent.[39] This means that even under current law, consent to process one's data for whatever purpose is not valid. It also entails that reuse of personal data for incompatible purposes is only lawful after the data have been fully anonymised, or after informed and freely given consent has been obtained for the new purpose (taking into account art. 7.4 as discussed above).[40] Together with the proportionality principle, European law thus prevents excessive processing of personal data. The proportionality principle refers to art. 9 DPD and stipulates that the processing of personal data needs to be adequate, relevant, and not excessive in relation to its purpose.

Fourth, to the extent that data processing infringes the privacy of a person, the

[38] Art. 9 of the OECD Guidelines formulates the purpose specification principle; art. 10 the use limitation principle. According to art. 10, however, data can be used for other purposes "with the consent of the data subject or by authority of law". This can be read as meaning that one can consent to waive one's right to use limitation.

[39] EU law requires that processing of personal data is based on one of six legal grounds that legitimate the processing of personal data (art. 7 DPD, art. 6 GDPR); consent is one of these legal grounds. On top of this requirement, the purpose must be explicitly specified and processing must be restricted to the specified purpose or one that is compatible (art. 6 DPD, art. 5 GDPR). This requirement cannot be waived.

[40] See, however, Art. 29 WP Opinion 5/2014, WP216 on anonymisation techniqes, that seems to pratically rule out effective full anonymisation. This entails that such "anonymisation" must be qualified as pseudonymisation from the perspective of the law.

proportionality principle also relates to the issue of whether such processing is proportional, considering the legitimate aim it serves. This links the DPD with art. 8 of the European Convention of Human Rights (ECHR) that enshrines the right to privacy. Art. 8 determines that whenever a measure infringes the privacy of a person, the measure that constitutes the interference must be justified on the basis of a triple test. First, the infringement must be directed to a legitimate aim. Such aims are limitatively summed up in art. 8.2, but formulated in such a broad manner that this is seldom a problem. Second, the infringement must be "in accordance with the law", meaning that the infringement is based on a legal norm that is accessible and foreseeable while incorporating the necessary safeguards for the person whose privacy is infringed. Relevant safeguards may regard the limitation of the scope and the duration of the infringement, a right to object and the need to obtain a warrant or permission from an independent authority.[41] Third, the infringement must be necessary in a democratic society, which implies – according to the case law of the European Court of Human Rights – that there is a pressing social need for the infringing measure.[42] This – third – part of the triple test, is interpreted in terms of proportionality; the infringement must be proportional to the aim served, meaning that the measure is appropriate to achieve the aim and not excessive in relation to the aim. The proportionality principle of art. 8.2 of the ECHR regards the processing of personal data only insofar as it infringes one's privacy. Though processing credit card data for the sale of LBSs falls within the scope of the DPD, it is not an infringement of one's privacy. The processing of location data by the provider of an LBS may, however, constitute an infringement if it violates a person's reasonable expectation of privacy, notably when data are shared with third parties that link the data with other information to gain a more complete insight into a person's social network or life style. If that third party is a government agency art. 8 will definitely apply. Since the ECHR provides a person within the jurisdiction of the Council of Europe with rights against their governments, it is not obvious that art. 8 always applies to the processing of personal data by other third parties. The so-called direct and indirect horizontal effect of art. 8 may extend the protection to privacy infringements by private parties. Direct horizontal effect refers to e.g. tort law, that could render an LBSP liable for harm caused if it infringes the privacy of its customers (e.g. by sharing location data with parties capable of building invasive profiles with additional data).[43] Indirect horizontal effect refers to a positive obligation of the state to protect its

[41] Cf. e.g. ECrHR, 29 June 2006, Weber and Saravia v Germany, Appl. Nr. 54934/00, which sums up the safeguards relevant for a measure being qualified as 'in accordance with the law', § 95.

[42] Cf. e.g. ECtHR, 26 April 1979, The Sunday Times v UK (Series A no 30), § 62.

[43] Cf. e.g. E Frantziou, "The Horizontal Effect of the Charter of Fundamental Rights of the EU: Rediscovering the Reasons for Horizontality" (2015) 21 *European Law Journal* 657-679, at 666. Frantziou distinguishes between direct and indirect horizontal effect and positive obligations. For the sake of space constraints, we will not enter this discussion, but we note

citizens against privacy infringements by private parties. Considering the fact that the fundamental right to data protection explicitly requires Member States of the EU to implement this right in relation to non-state actors, such positive obligations may indeed give rise to liability of the state on the nexus of privacy and data protection.Indirect horizontal effect refers to a positive obligation of the state to protect its citizens against privacy infringements by private parties. Considering the fact that the fundamental right to data protection explicitly requires Member States of the EU to implement this right in relation to non-state actors, such positive obligations may indeed give rise to liability of the state on the nexus of privacy and data protection.[44]

To determine the meaning of a legal text that has force of law, such as the DPD or the upcoming GDPR, lawyers will usually refer to relevant case law, which also has "force of law". Based on art. 29 of the DPD, however, a Working Party was established to advise on the interpretation of the DPD. Though itssuch interpretations isare not legally binding, it hasthey have considerable authority and its Opinions must be considered when deciding the meaning of element of the DPD. In 2013, the Art. 29 Working Party issued an Opinion to clarify the meaning of the principle of purpose limitation that further elucidates how this legal obligation operates and how data controllers should implement purpose binding in their personal data life cycle management.[45] The Working Party notably concerns itself with the interpretation of what constitutes a compatible or incompatible purpose, since this determines when processing is no longer lawful. The Working Party notably concerns itself with the interpretation of what constitutes a compatible or incompatible purpose, since this determines when processing is no longer lawful.[46] A change of purpose will usually be relevant when personal data are reused or recycled in relation to other business models or by third parties that may provide entirely different services that cannot be understood as serving a compatible purpose. For instance, if an LBS stores location data for longer than strictly necessary to fulfil the contract (to provide friends with one's location, to offer promotions of nearby shops, to calculate the invoice), the LBS needs to check whether this is still compatible with the specified purpose it explicitly expressed when to offer promotions

her argument (at 672-3) that horizontal effect also relates to the fact that fundamental rights are not merely individual interests but also collective goods.

[44]Also, art. 82.1 GDPR requires that "Any person who has suffered material or non-material damage as a result of an infringement of this Regulation shall have the right to receive compensation from the controller or processor for the damage suffered." Art. 82.2 states that "Any controller involved in processing shall be liable for the damage caused by processing which infringes this Regulation. A processor shall be liable for the damage caused by processing only where it has not complied with obligations of this Regulation specifically directed to processors or where it has acted outside or contrary to lawful instructions of the controller."

[45]Art. 29 WP, Opinion 03/2014, WP 203 on purpose limitation.

[46]Ibid, at 20-36.

of nearby shops, to calculate the invoice), the LBS needs to check whether this is still compatible with the specified purpose it explicitly expressed when first processing the data. The Opinion clarifies that compatibility as to the original purpose must be decided on a case to case basis, taking into account the following key factors:[47]

1. The relationships between the purposes for which the personal data have been collected and the purposes of further processing;
2. The context in which the personal data have been collected and the reasonable expectations of the data subjects as to their further use;
3. The nature of the personal data and the impact of the further processing on the data subjects;
4. The safeguards adopted by the controller to ensure fair processing and to prevent any undue impact on the data subjects.

As to the further use of location data, this entails, first, that reuse for an entirely unrelated purpose (e.g. tax fraud detection) is problematic. Second, it implies that the fact that sensitive data may be inferred from location data indicates the need to be cautious about concluding that a purpose is compatible, notably when such data or their inferences can have adverse effects on a data subject (rejection for an insurance or a job interview, detailed monitoring by law enforcement agencies based on risk assessments that rely in part on location data, paying a higher price for consumer goods due to having visited shops on the high end of the market). Third, much will depend on the question of whether adequate safeguards have been implemented, for instance, pseudonymisation of energy usage data that include location data. Interestingly, the Working Party finds that the context in which the data have been collected is a key factor to determine the compatibility of the purpose. This links with the idea of contextual integrity, though we note that context is only one of the key factors, rather than the sole criterion to decide the legitimacy of personal data processing.

6.2 Interfacing CI and PL

Some authors believe that the PL obligation is a consecration of Nissenbaum's CI theory.[48] Others believe that the CI theory is not very relevant for non-US jurisdictions that have a framework of general data protection law that

[47] Ibid, at 23-27.

[48] F Dumortier, "Facebook and Risks of 'De-contextualization' of Information" in S Gutwirth, Y Poullet and P De Hart (eds) *Data Protection In a Profiled World* (Springer; 2010) 119-137.

incorporates PL.[49] As a point of departure we note that, though her theory may be used to add clarity to the legal framework, CI covers only the flow of information and depends entirely on the original context of disclosure.

As such, firstly, the scope of PL within the legal framework is broader than merely the transmission and distribution of personal data within an identifiable context, as it includes any processing operations performed on personal data that are stored (where no data flow is at stake). We note that EU data protection law holds specific protection for location data, even if it cannot be qualified as personal data.[50]

Second, the scope of PL is broader in that it considers the processing of personal data in other contexts without assuming that such processing is necessarily illegitimate whenever it takes place in another context. PL depends on the declared intent of – and the usage by - the data controller, taking into account the context of collection.

Third, the scope of protection generated by PL may be eroded if data controllers specify a broad purpose or a whole range of purposes when they collect the data. In that case, PL may provide little protection other than forcing data controllers to determine and make explicit the relevant purposes before they start processing. The requirement of specificity should actually prevent overly broad purpose determinations, but so far both private and public entities often resort to either very broad formulations of purposes (e.g. "to improve the provision of services"), or to whole series of specific purposes (e.g. to detect tax fraud and social security fraud, or to achieve compliance with the principle of 'collect only once' in the context of e-government). As a matter of fact, the lack the protection caused by broad definitions or multiple disparate purposes is due a lack of enforcement; if data protection authorities were to have the competence to impose adequate fines, the requirement of explicit specification should be sufficient to prevent erosion of the obligation. A similar lack of the protection is inherent in the notion of context. Notably, policy makers, courts or data controllers may decide that in the case of LBSs the prevailing context is commercial, whatever other context is at stake. The same goes for the prevailing context of national and public security, which often overrules any context, even the commercial one. The lack of the protection offered by CI due to either the overruling context of commerce or that of public security is inherent in the

[49] CJ Bennet, "Book Review: Nissenbaum, Helen (2010) Privacy in Context: Policy and the Integrity of Social Life. Stanford: Stanford University Press" (2011) 8 *Surveillance and Society* 541-543, at 542-543.

[50] See notably art. 9 ePrivacy Directive 2002/58/EC, concerning the processing of personal data and the protection of privacy in the electronic communications sector. On the intricacies concerning the legal regime for location data, see C Cuijpers and BJ Koops, "How Fragmentation in European Law Undermines Consumer Protection: the Case of Location-Based Services" (2008) 33 European Law Review 880–897.

difficulty of identifying the prevailing context, especially when more than one prevailing context is at stake, depending on one's perspective.

Fourth, PL is a legal obligation that creates civil and administrative liability for data controllers, whereas CI is an ethical theory that invites citizens, policy makers and business undertakings to consider what should count as appropriate data flows and appropriate distribution of personal data. Though a law that does not even aim to achieve justice cannot be qualified as law in a constitutional democracy, the law is both more and less than ethics. Law also aims for legal certainty and purposefulness. This implies that its scope is both more restricted than ethics (for instance, if achieving justice is at odds with legal certainty, notably where people disagree about what is just in a particular context) and broader (for instance, where no agreement can be found on what constitutes an appropriate and properly distributed information flow, the law will provide for legal certainty and decide on such issues in line with PL and proportionality).

Taking account of these differences, we shall now see how context fits into the decision on the lawfulness of reuse of personal data, by referring to the second key factor for determining whether its purpose is compatible with the initial, explicitly specified purpose. As discussed above, this key factor concerns "the context in which the personal data have been collected and the reasonable expectations of the data subjects as to their further use." Indeed, this implies two things. First, it implies that processing personal data should be aligned with the reasonable expectations that come with the context where they were first collected. This points to the relevant transmission principles and informational norms of the CI heuristic. Second, it implies that processing such data in another context is not prohibited, but that to determine the legitimacy of cross-contextual processing, the expectations raised in the original context are critical. In a sense, this key factor seems to integrate the CI heuristic into the determination of the compatibility of the purpose. At the same time, it does not make the CI heuristic decisive as other key factors must be taken into consideration. As a consequence, the original context does not over-determine the judgement on whether or not the processing of location data is legitimate, though it plays a key role.

In art. 35 of the upcoming GDPR, a new legal obligation is established that requires data controllers that wish to employ new technologies to assess whether these technologies generate high risks for rights and freedom of individuals. If this is the case, the controller should perform a data protection impact assessment (DPIA).[51] We note that the CI decision heuristic provides an

[51] Art. 35.1 GDPR: "Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations

interesting framework for such an assessment, while the DPIA has the added advantage of requiring an assessment of mitigating measures,[52] thus integrating the notion of data protection by design and default into the assessment.[53] It is pivotal that users of LBS have access to ways and means to protect themselves against persistent tracking and tracing.[54] However, by imposing DPbD, the GDPR obliges providers to develop architectures on their side that protect location privacy, which seems crucial to incentivise the market for DPbD.[55] We therefore believe that the CI heuristic and the DPIA should inspire and test each other, notably with respect to the assessment of the risks to the rights and freedoms of individuals. This would enable a more stringent assessment of how PL and CI contribute to reducing such risks, while providing users with a better grasp of how they may be targeted based on the location data they leak. On top of that, the force of law that "backs" PL, in combination with legal obligations to conduct a DPIA and to implement DPbD, should create a market for socio-technical design solutions that support PL.

7 Conclusion

In this work, we investigated the privacy implications for users engaging in LBSs and the sharing of their location data with remote services. We used Nissenbaum's CI heuristic as a framework to perform the assessment in a structured way. We applied CI to our case and we modeled all involved parties that may get access to location data and further modeled the relevant information flows. This revealed that users rarely share their location data with only the LBSP but end up sharing the data with a series of other entities, such as TPSD, OSM, HM or the Gv.

The context in which a user revealed her location is key in Nissenbaum's heuristic when evaluating whether the user's privacy has been breached. Unfortunately, LBSs are used in such a ubiquitous manner that it is impossible to conduct the

on the protection of personal data. A single assessment may address a set of similar processing operations that present similar high risks."

[52] Art. 35.7.d GDPR.

[53] In art. 25 GDPR a new legal obligation requires 'data protection by default and by design'. See M Hildebrandt and L Tielemans, "Data Protection by Design and Technology Neutral Law" (2013) 29 Computer Law & Security Review 509–521.

[54] See e.g. F Brunton and H Nissenbaum, Obfuscation: A User's Guide for Privacy and Protest (2015) *The MIT Press*. M Herrmann et al, "Optimal Sporadic Location Privacy Preserving Systems in Presence of Bandwidth Constraints" (2013) *12th ACM Workshop on Workshop on Privacy in the Electronic Society* 167-178.

[55] See e.g. M Herrmann et al, "Practical Privacy-Preserving Location-Sharing based services with aggregate statistics" (2014) *Proceedings of the 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks* 87-98.

CI heuristic for every possible context in which users may find themselves when engaging with LBSs. This is not a drawback of the CI heuristic, but inherent in the use of mobile devices that result in the integration and overlapping of different contexts at the same time and the same place. In the case of LBSs, we found that the prevailing context is that of travel, though one can argue that it will often coincide with the prevailing contexts of commerce and public security. Focusing on the context of travel, we show that LBSs create new information flows and alter existing ones with the result that numerous parties obtain users' accurate location, resulting in threats to the user's fair treatment, her autonomy, and other fundamental rights and freedoms. This is because location data exposed via mobile applications may be recorded by LBSs for an unlimited period of time and can be exchanged with a wide variety of parties (even though this would be unlawful under EU law). A user's location data may thus be combined with lifelong movement profiles, which are machine searchable and could be used in many unforeseeable ways to draw inferences about the individual, even far into the future. This in particular is an issue since the user has few effective means to access, let alone control, the information flows that instigate personalised targeting. To assess the extent to which control is lost, and whether this violates reasonable expectations of privacy, we considered the concept of contextual integrity in relation to the principle of purpose limitation.

From a legal perspective, the processing of personal data within the scope of the DPD (and the upcoming GDPR) requires a legitimate, specific and explicit purpose, which restricts the proliferation of location data and the inferences it affords. Currently, however, the requirements of legitimacy, specificity and explicitness are often circumvented by LBSP by formulating very broad terms of use hidden in lengthy terms of service or privacy policies, or by trading free services for extensive invisible profiling. We hope and expect that the upcoming GDPR will enable a more effective enforcement that makes potential usage of location data foreseeable and enables data subjects to object to the processing of excessive or irrelevant data. In its Opinion on purpose limitation the Art. 29 Working Party has clarified that the context is key when assessing whether the purpose of processing is compatible with the original purpose. This context argument nicely links PL to Nissenbaum's CI heuristic and we elaborate on the differences and similarities of the two: Firstly, PL is a broader concept because CI determines privacy violations only if data is being transmitted, whereas PL concerns all types of processing, including analytics. PL is also broader because a violation of PL does not necessarily depend on the context; processing the data in the same context for another purpose may be a violation of PL, where it may be acceptable in terms of CI (depending on whether it violates an informational norm of that context). Secondly, whereas PL may be circumvented by entities stating very general purposes or a long series of specific purposes allowing for almost any processing, CI may be circumvented by defining the

prevailing context in a way that enables it to overrule informational norms and transmission principles of overlapping prevailing contexts, notably those of commerce and public security. In that sense much depends on enforcement in the case of PL and the perspective taken by the assessor in the case of CI. Thirdly, CI is an ethical theory that offers a structured approach to reflect on and assess privacy as contextual integrity, whereas PL is a legal obligation, that has legal effect such as liability and the right to object.

Context is everything, but not everything is context. Purpose limitation enables both foreseeability and holding LBSP to account in a court of law, if the law is enforced. We conclude that both concepts are pivotal for a sustainable and responsible employment of location data, noting that the CI decision heuristic should inform the templates of the Data Protection Impact Assessment that will soon be a legal obligation within EU jurisdiction.

Publication

Leaky Birds: Exploiting Mobile Application Traffic for Surveillance

Publication Data

Eline Vanrykel, Gunes Acar, Michael Herrmann, and Claudia Diaz. Leaky Birds: Exploiting Mobile Application Traffic for Surveillance. In *20th International Conference on Financial Cryptography and Data Security (FC)*, pages 1–18. Springer Berlin Heidelberg, 2016.

Contributions

- Conceptual work shared with Gunes Acar. Responsible for developing tools to download apps.

Leaky Birds: Exploiting Mobile Application Traffic for Surveillance

Eline Vanrykel[1], Gunes Acar[2], Michael Herrmann[2], and Claudia Diaz[2]

[1] KU Leuven, Leuven, Belgium
eline.vanrykel@gmail.com
[2] KU Leuven ESAT/COSIC, iMinds, Leuven, Belgium
{name.surname}@esat.kuleuven.be

Abstract. Over the last decade, mobile devices and mobile applications have become pervasive in their usage. Although many privacy risks associated with mobile applications have been investigated, prior work mainly focuses on the collection of user information by application developers and advertisers. Inspired by the Snowden revelations, we study the ways mobile applications enable mass surveillance by sending unique identifiers over unencrypted connections. Applying passive network fingerprinting, we show how a passive network adversary can improve his ability to target mobile users' traffic.

Our results are based on a large-scale automated study of mobile application network traffic. The framework we developed for this study downloads and runs mobile applications, captures their network traffic and automatically detects identifiers that are sent in the clear. Our findings show that a global adversary can link 57% of a user's unencrypted mobile traffic. Evaluating two countermeasures available to privacy aware mobile users, we find their effectiveness to be very limited against identifier leakage.

1 Introduction

Documents that have been revealed by the former NSA contractor Edward Snowden shed light on the massive surveillance capabilities of the USA and UK intelligence agencies. One particular document released by the German newspaper *Der Spiegel* describes the ways in which traffic of mobile applications (apps) is exploited for surveillance [15]. The document, which reads

"Exploring and Exploiting Leaky Mobile Apps With BADASS," provides a unique opportunity to understand the capabilities of powerful network adversaries. Furthermore, the document reveals that identifiers sent over unencrypted channels are being used to distinguish the traffic of individual mobile users with the help of so-called *selectors*. Similar revelations about the use of Google cookies to target individuals imply that BADASS is not an isolated incident [11, 34].

While it is known that a substantial amount of mobile app traffic is unencrypted and contains sensitive information such as users' location or real identities [23, 35, 43], the opportunities that mobile traffic offers to surveillance agencies may still be greatly underestimated. Identifiers that are being sent in the clear, may allow the adversary to link app sessions of users and thus to learn more information about the surveilled users than he could without. The purpose of this study is to evaluate this risk and to quantify the extent to that it is possible to track mobile app users based on unencrypted app traffic.

To this end we present a novel framework to quantify the threat that a surveillance adversary poses to smartphone users. The framework automates the collection and analysis of mobile app traffic: it downloads and installs Android apps, runs them using Android's *The Monkey* [17] tool, captures the network traffic on cloud-based VPN servers, and finally analyzes the traffic to detect unique and persistent identifiers. Our framework allows large-scale evaluation of mobile apps in an automated fashion, which is demonstrated by the evaluation of 1260 apps. We choose the apps among all possible categories of the Google Play store and of different popularity levels.

Our study is inspired by a recent work by Englehardt *et al.* [25]. They studied the surveillance implications of cookie-based tracking by combining web and network measurements. The evaluation method they use boils down to measuring the success of the adversary by the ratio of user traffic he can cluster together. We take a similar approach for automated identifier detection but we extend their work to capture non-cookie-based tracking methods that are suitable for user tracking. Moreover, we show how TCP timestamp-based passive network fingerprinting can be used to improve the clustering of the traffic and may allow to detect the boot time of Android devices.

1.1 Contributions

Large-scale, automated study on surveillance implications of mobile apps. We present an automated analysis of 1260 Android apps from 42 app categories and show how mobile apps enable third party surveillance by sending unique identifiers over unencrypted connections.

Table 1: Unique smartphone identifiers present on Android, an overview.

Name	Persistence	Permission
Android ID	until a factory reset	None
MAC Address	lifetime of the device	`ACCESS_WIFI_STATE`
IMEI	lifetime of the device	`READ_PHONE_STATE`
IMSI	lifetime of the SIM card	`READ_PHONE_STATE`
Serial number	lifetime of the device	None [41]
SIM serial number	lifetime of the SIM card	`READ_PHONE_STATE`
Phone number	lifetime of the SIM card	`READ_PHONE_STATE`
Google Advertising ID	until reset by the user	`ACCESS_NETWORK_STATE`, `INTERNET`

Applying passive network fingerprinting for mobile app traffic exploitation. We show how a passive network adversary can use TCP timestamps to significantly improve the amount of traffic he can cluster. This allows us to present a more realistic assessment of the threat imposed by a passive adversary. Further, we show how an adversary can guess the boot time of an Android device and link users' traffic even if they switch from WiFi to 3G or vice versa.

Evaluation of the available defenses for privacy aware users. We analyze the efficacy of two mobile ad-blocking tools: Adblock Plus for Android [12] and Disconnect Malvertising [13]. Our analysis shows that these tools have a limited effect preventing mobile apps from leaking identifiers.

2 Background and Related Work

Android apps and identifiers. Android apps and third-parties can access common identifiers present on the smartphone, such as MAC address, Google Advertising ID or IMEI number. We call these identifiers *smartphone IDs*. An overview of the Android smartphone IDs can be found in Table 1. Developers may also choose to assign IDs to users (instead of using smartphone IDs), for identifying individual app installations or simply to avoid asking for additional permissions [10]. We refer to such identifiers as *app assigned IDs*.

Privacy implications of mobile apps. Although privacy implications of Android apps have been extensively studied in the literature [24, 27, 29], prior

work has mainly focused on the sensitive information that is collected and transmitted to remote servers. Xia et al. showed that up to 50% of the traffic can be attributed to the real names of users [43]. Enck et al. developed TaintDroid [24], a system-wide taint analysis system that allows runtime analysis and tracking of sensitive information flows. While it would be possible to use TaintDroid in our study, we opted to keep the phone modifications minimal and collect data at external VPN servers. This allows us to have a more realistic assessment of application behavior and adversary capabilities.

Our work differs from these studies, by quantifying the threat posed by a passive network adversary who exploits mobile app traffic for surveillance purposes. We also show how the adversary can automatically discover user identifiers and use passive network fingerprinting techniques to improve his attack.

Passive network monitoring and surveillance. Englehardt et al. [25] show how third-party cookies sent over unencrypted connections can be used to *cluster* the traffic of individual users for surveillance. They found that reconstructing 62-73% of the user browsing history is possible by passively observing network traffic.

In addition to using identifiers to track smartphones, an eavesdropping adversary can use passive network fingerprinting techniques to distinguish traffic from different physical devices. Prior work showed that clock skew [31, 33, 44], TCP timestamps [22, 42] and IP ID fields [20] can be used to remotely identify hosts or count hosts behind a NAT. In this study, we use TCP timestamps to improve the linking of users' mobile traffic in short time intervals. We assume the adversary to exploit TCP timestamps to distinguish traffic of users who are behind a NAT. Moreover, we demonstrate how an adversary can discover the boot time of an Android device by exploiting TCP timestamps.

3 Threat Model

In this paper we consider passive network adversaries whose goal is to link app traffic of smartphone users. The adversaries observe unique identifiers that are being transmitted from mobile apps in the clear and apply network fingerprinting techniques. We consider that the adversaries cannot break cryptography or launch MITM attacks such as SSLstrip [32].

We distinguish between two types of passive adversaries: A *global passive adversary*, who can intercept all Internet traffic at all time; and a *restricted passive adversary* who can only observe a limited part of the network traffic.

Both adversaries have the capability to collect bulk data. This may be achieved in various ways, such as tapping into undersea fiber-optic cables; hacking routers or switches; intercepting traffic at major Internet Service Providers (ISP) or Internet Exchange Points (IXP) [3].

There can be several models in which an adversary may have limited access to the user's traffic. In this study we evaluate adversaries whose limitation is either *host-based* or *packet-based*. The host-based adversary is only able to see traffic bound to certain hosts; for example, because the adversary is only able to obtain warrants for intercepting traffic within its own jurisdiction. The packet-based adversary may only have access to a certain point in the Internet backbone and thus miss traffic that is being sent along other routes. For both adversaries, we evaluate the success based on different levels of network coverage (Section 7.2). We simulate partial network coverage by randomly selecting hosts or packets to be analyzed depending on the model. For instance, for the host-based model with 0.25 network coverage, we randomly pick one-fourth of the hosts and exclude the traffic bound to remaining hosts from the analysis.

4 Data Collection Methodology

4.1 Experimental Setup

We present the experimental setup[4] that is used for this paper in Fig. 1. It includes a controller PC, two smartphones and two VPN servers for traffic capture. The main building blocks of our framework are as follows:

Controller PC. The Controller PC runs the software that orchestrates the experiments and the analysis. It has three main tasks: 1) installing apps on the smartphones and ensuring that the experiment runs smoothly, e.g. checking the phone's WiFi and VPN connections, 2) sending SSH commands to the remote VPN servers to start, stop and download the traffic capture, 3) analyzing the collected data.

Smartphones. We conducted our experiments with two *Samsung Galaxy SIII Mini* smartphones running Android version 4.1.2. We *rooted* the phones to address issues such as storage and uninstallation problems. Although we

[3] All these methods are feasible, as illustrated by the Snowden revelations [7, 28].

[4] The source code of the framework, as well as the collected data will be made available to researchers upon request.

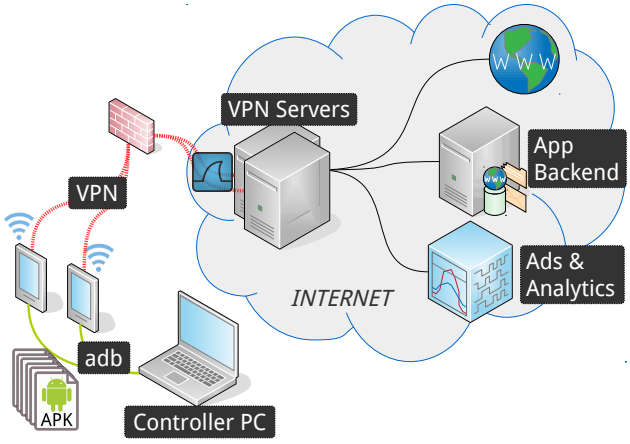


Figure 1: Our setup in this study consists of a Controller PC that manages the experiments, two Android phones that run apps, and two VPN servers that capture the network traffic.

considered using the Android emulator as in other works [23, 36, 38], our preliminary tests [39] showed that the number of transmitted identifiers is significantly less in the emulator compared to the same setting with a real smartphone and the emulator lacks certain identifiers, such as the WiFi MAC address. We also chose not to intercept system API calls or instrument the operating system, such as in [24, 26], since we preferred a simpler and more portable solution.

The Monkey. We use *The Monkey* [17] tool to automate the experiments and simulate the user interaction at large scale. The Monkey generates a pseudo-random event stream that includes touch, motion and keyboard events.

Traffic Capture. The network traffic is captured by two remote VPN servers, using the *dumpcap* [5] command line tool. Using VPN servers, we could capture *all* the network traffic and not only HTTP traffic, which would be the case with an HTTP proxy. Also, since we record the traffic on remote machines, we ensure that there is no packet drop due to lack of buffer space on resource constrained devices [14]. However, we captured traffic locally on the phone during the evaluation of ad-blockers for Android. These tools use a proxy or VPN themselves to block ads. Since Android does not allow simultaneous VPN connections, we captured the traffic locally by running *tcpdump* on the smartphones. To ensure comparability, we exclude all the captures where we observed packet drops from the analysis (20% of the cases, 171 apps in two

experiments).

Traffic parser. For parsing the captured network traffic, we developed a Python script based on the `dpkt` [3] packet parsing library. The script allows us to decode IPv4 and IPv6 datagrams, reassemble TCP streams, decompress compressed HTTP bodies and to parse GRE and PPTP encapsulation used by the VPN. We extract HTTP headers and bodies, packet timestamps, IP addresses and port numbers from the packets for later use. Since it is outside of the scope of this study, we did not decrypt SSL/TLS records. However, for the TCP timestamp analysis described in Section 6 it is beneficial, yet not necessary, to extract TCP timestamps from all TCP packets, including the ones from encrypted HTTPS traffic. Note that this is within our adversary model, because TCP headers are sent in the clear and thus available to a passive adversary.

Having described the main building blocks of the experimental setup, now we outline the different modes and steps of the experiments:

Experiment modes. We run experiments in two different modes to evaluate the difference in identifier transmission; i) if the app is simply opened and ii) if the user actually interacts with the app. We refer to the former as *startscreen experiment* and to the latter as *interactive experiment*. The Monkey is used to simulate user interaction in the interactive experiments.

Evaluation of ad-blocker apps. We evaluate the effect of apps that block ads and trackers. While those apps are not specifically designed to prevent identifier leakage, they may still reduce the number of identifiers being sent in the clear. Specifically, we repeated the experiment of the top-popularity apps after we installed and activated the ad-blocker apps *Adblock Plus for Android* [12] and *Disconnect Malvertising* [13].

Steps of the experiment. Our framework executes the steps of the experiments in an entirely automated fashion. The Controller PC connects the smartphone to the VPN server by running a Python based *AndroidViewClient* [4] script that emulates the touch events necessary to start the VPN connection on the smartphone. Since installing all the apps at once is not possible due to storage constraints, our framework conducts the experiment in cycles. In each cycle we install 20 apps and then run them sequentially[5]. The apps for each cycle are randomly chosen from the entire set of apps, with the condition that each app is only picked once. Before running an app, the Controller PC kills the process of the previous app. This way we are able to prevent the traffic of the previously tested app mistakenly being recorded for the subsequent app. After finished running the 20 apps, the Controller PC runs all 20 apps a second time

[5]We chose 20 since this was the maximum number of apps that can be installed on an Android emulator at once, which we used in the preliminary stages of the study.

in the same order. Running each app twice enables the automated detection of identifiers outlined in Section 5.1. Finally, the Controller PC completes the current cycle by uninstalling all 20 apps.

4.2 Obtaining Android Applications

To obtain the Android apps, we developed scripts for crawling the Google Play store and, subsequently, to download APK files. Our scripts are based on the Python Selenium [16] library, the `APK downloader` browser extension and webpages [1]. Using this software, we crawled the entire Play Store and obtained information on $1,003,701$ different Android apps. For every app we collected information such as number of downloads, rating scores and app category. This allows us to rank the apps of every category according to their popularity.

For every app category we choose 10 apps from three different popularity levels: *top-popularity*, *mid-popularity* and *low-popularity*. While we use the most popular apps for the top-popularity category, we sample the mid-popularity and low-popularity apps from the 25th and 50th percentiles from each category. At the time we conducted the crawl, there were 42 different app categories and we therefore obtained a total of 1260 ($42 \times 10 \times 3$) apps. The average time for evaluating one app is 64 seconds.

5 Analysis Methodology

In the following we show how an adversary is able to extract identifiers from network traffic and then use these identifiers to cluster data streams, i.e. linking data streams as belonging to the same user. This is the same that an adversary with the goal of surveilling Internet traffic would do, i.e. extracting and applying a set of *selectors* that match unique and persistent mobile app identifiers.

5.1 Identifier Detection

Suitable identifiers for tracking need to be persistent and unique, i.e. the same ID cannot appear on different phones and IDs need to be observable over

multiple sessions. Our framework automatically detects such unique identifiers in unencrypted mobile app traffic. While the overall approach is similar to the one in [18, 25] we extend the cookie-based identifier detection technique to cover mobile app traffic. We assume that the smartphone IDs (such as Android ID or MAC address) are not known a priori to the adversary. The adversary has to extract IDs based on the traffic traces only. Yet, we use smartphone IDs as the ground truth to improve our automated ID detection method by tuning its parameters.

For finding identifiers, we process HTTP request headers, bodies and URL parameters. Specifically, the steps of the unique identifier detection are as follows:

- Split URLs, headers, cookie contents and message bodies using common delimiters, such as "=", "&", ":", to extract key-value pairs. Decode JSON encoded strings in HTTP message bodies.
- Filter out cookies with expiry times shorter than three months. A tracking cookie is expected to have a longer expiry period [25].
- For each key-value pair, we construct an *identifying rule set* and add it to the potential identifier list. This is the tuple (host, position, key), where *host* is extracted from the HTTP message and *position* indicates whether the key was extracted from a cookie, header or URL.
- Compare values of the same key between runs of two smartphones.
- – Eliminate values if they are not the same length.
  - Eliminate values that are not observed in two runs of the same app on the same smartphone.
  - Eliminate values that are shorter than 10 or longer than 100 characters.
  - Eliminate values that are more than 70% similar according to the Ratcliff-Obershelp similarity measure [21].
- Add (host, position, key) to the rule set.

We tuned similarity (70%) and length limits (10, 100) according to two criteria: minimizing false positives and detecting all the smartphone identifiers (Table 1) with the extracted rule set. We experimented with different limit values and picked the values that gave us the best results based on these criteria. A more thorough evaluation of these limits is omitted due to space constraints, but interested readers can refer to [18,25] for the main principles of the methodology.

5.2 Clustering of App Traffic

While the ultimate goal of the adversary is to link different app sessions of the same user by exploiting unique identifiers transmitted in app traffic, the first challenge of the adversary is to identify the traffic of one app. An app may open multiple TCP connections to different servers and linking these connections can be challenging. The user's public IP address is not a good identifier: several users may share the same public IP via a NAT. Moreover, IP addresses of mobile phones are known to change frequently [19].

In this work we consider two different clustering strategies. In the *TCP stream based linking*, the attacker can only link IP packets based on their TCP stream. The adversary can simply monitor creation and tear down of TCP streams and ensure that the packets being sent within one stream are originating from the same phone. The second, more sophisticated strategy employs passive network fingerprinting techniques to link IP packets of the same app session. We will refer this technique as *app session based linking* and outline it in Section 6.

Following Englehardt et al. [25] we present linking of the user traffic as a graph building process. We use the term *node* to refer to a list of packets that the adversary is certain that they belong to the same user. As explained above, this is either a TCP stream or an app session. For every node the adversary extracts the identifying rule set (host, position, key) as described in Section 5.1. Starting from these nodes, the adversary inspects the content of the traffic and then tries to link nodes together to so-called *components*.

Therefore, the attacker will try to match a node's identifiers to the identifiers of the existing components. We account for the fact that some developers do not use the smartphone ID right away as identifier, but derive an identifier from it by hashing or encoding. Thus the clustering algorithm will also try to match the SHA-1, SHA-256, MD5 and murmur3 hashes and base64 encoded form of the identifiers. For every node, there exist three possibilities when comparing the node's identifiers to a existing component's identifiers:

1. **The node's value (or its derivative) matches the identifiers of an existing component:** The node will be added to the component and the respective identifiers are being merged, i.e. the newly added node may include identifiers not yet included in the component.

2. **None of the node's identifiers or their derivatives can be matched to an existing component:** The node creates its own component which is disconnected from all other components.

3. **The node shares identifiers with multiple components:** These components are merged together and the node is added to this component.

For the remainder of this paper, we refer to the component that contains the highest number of nodes as the *Giant Connected Component* (GCC). Furthermore, we define the ratio of number of nodes in GCC to the number of nodes in the whole graph as the *GCC ratio*. The *GCC ratio* serves as a metric for measuring the adversary's success for linking users' traffic based on the amount of traffic he observes.

5.3 Background Traffic Detection

The Android operating system itself also generates network traffic, for example to check updates or sync user accounts. Although we find in our experiments that the Android OS does not send any identifiers in the clear, we still take measures to avoid that this traffic pollutes our experiment data. Particularly, we captured the network traffic of two smartphones for several hours multiple times without running any app. A complete overview of all HTTP queries made during such captures can be found in [40]. We excluded all the HTTP requests to these domains during the analysis stage. Although we excluded background traffic from our analysis, the adversary may try to exploit the background traffic in a real-world attack.

6 Linking Mobile App Traffic with TCP Timestamps

In this section we elaborate on the adversary's ability to employ passive fingerprinting techniques to link different IP packets originating from the same smartphone. As mentioned in Section 5.2, this gives a significant advantage to the adversary when clustering the user traffic. In particular, the adversary is able to analyze TCP timestamps for this task as they are commonly allowed by the firewalls [33].

TCP timestamps are an optional field in TCP packets that include a 32-bit monotonically increasing counter. They are used to improve the protocol performance and protect against old segments that may corrupt

TCP connections [30]. While the exact usage of TCP timestamps is platform dependent, our inspection of the Android source code and capture files from our experiments revealed that Android initializes the TCP timestamp to a fixed value after boot and uses 100Hz as the timestamp increment frequency [2]. Thus, at any time t, TCP timestamp of a previously observed device can be estimated as follows: $TS_t = TS_{prev} + 100 \times (t - t_{prev})$, where TS_{prev} is the timestamp observed at t_{prev} and $(t - t_{prev})$ is the elapsed time. The adversary can therefore link different visits from the same device by comparing the observed TCP timestamps to his estimate. Prior studies have shown that distinguishing devices behind a NAT using TCP timestamps can be done in an efficient and scalable manner [22, 37, 42].

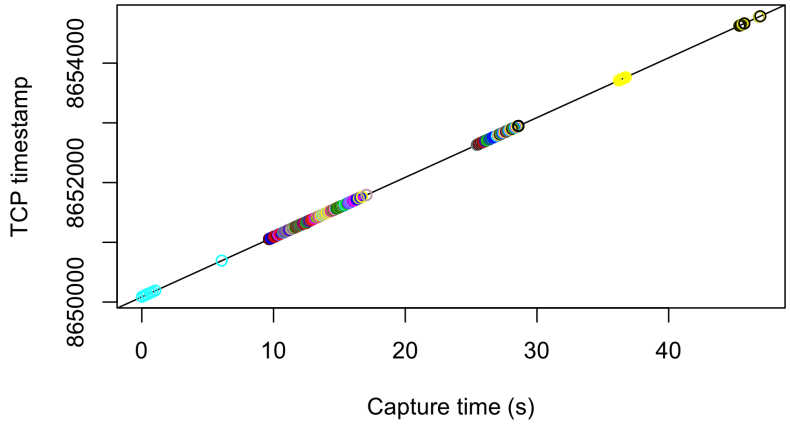


Figure 2: TCP timestamp vs. capture time plot of Angry Birds Space app follows a line with a slope of 100, which is the timestamp resolution used by Android. Different TCP sessions, indicated by different colors, can be linked together by exploiting the linearity of the TCP timestamp values.

Fig. 2 demonstrates the linear increase of the TCP timestamps of a phone running the "Angry Bird Space" app. To demonstrate the linkability of TCP streams, every point in Fig. 2 is colored based on its TCP source and destination port. The straight line shows that the adversary can easily link different TCP streams of the same device by exploiting the linearity of the timestamps. The adversary is also able to consider TCP timestamps of encrypted communications, because TCP timestamps are sent unencrypted in the packet headers. This

allows adversaries within our threat model to further increase the success of the linking. Furthermore, TCP timestamps can be used to link traffic even if the user switches from WiFi to mobile data connection or vice versa [40]. Finally, the linking is still feasible even if the adversary misses some packets, for instance, due to partial coverage of the network.

Limitations. During the background traffic detection experiments, we observed cases in which TCP timestamps are not incremented linearly. Consulting the Android System Clock Documentation, we determined that the CPU and certain system timers stop when the device enters the *deep sleep* state [9]. This power saving mechanism is triggered only when the screen is off and the device is not connected to the power outlet or USB port. Therefore, the phone will never go into deep sleep when a user is interacting with an app and the TCP timestamps will be incremented in a predictable way, allowing the linking of the traffic by app sessions.

Implications for traffic linking. We will assume the adversary can use TCP timestamps to cluster packets generated during the use of an app (app session), as the phone never enters *deep sleep* mode when it is in active use. As mentioned in Section 5.2, we will refer to this as *app session based linking*.

Android boot time detection. In addition to linking packets from different TCP streams, TCP timestamps can also be used to guess the boot time of remote devices [6]. Among other things, boot time can be used to determine if the device is patched with critical updates that require a reboot. Since it is not directly related to traffic linking attack considered in the study, we explain the boot time detection methodology in the unabridged version of this paper [40].

7 Results

7.1 Identifier Detection Rules

We present in Table 2 an overview of the identifying rule set that we detected by the methodology explained in Section 5.1. Recall that identifying rules correspond to "selectors" in the surveillance jargon, which allow an adversary to target a user's network traffic. In total we found 1597 rules with our method, of which 1127 (71%) correspond to a smartphone ID or its derivative. Our results show that the Android ID and Google Advertising ID are the most frequently transmitted smartphone IDs, accounting for 72% (812/1127) of the

Table 2: The extracted ID detection rules and corresponding smartphone IDs. *SID*: Smartphone ID, *AAID*: App Assigned ID.

Exp. Mode	**App popularity**	**Android ID**	**Google Ad ID**	**IMEI**	**MAC**	**Other SIDs**	**AAIDs**	**Total ID Rules**
Interactive	top	165	111	63	29	16	193	577
Startscreen	top	115	56	45	19	11	91	337
Interactive	mid	56	28	20	6	5	60	175
Startscreen	mid	48	28	16	5	4	40	141
Interactive	low	73	61	22	15	8	53	232
Startscreen	low	47	24	16	7	8	33	135
Total		504	308	182	81	52	470	1597

Table 3: Examples rules found in the constructed identifying rule set. The values are modified to prevent the disclosure of real identifiers of the phones used in the study.

Host	**Position**	**Key**	**ID**	**Value**
`data.flurry.com`	Body	offset60	Android ID	AND9f20d23388...
`apps.ad-x.co.uk`	URI	custom_data meta_udid	Unknown	19E5B4CEE6F5...
`apps.ad-x.co.uk`	URI	macAddress	WiFi MAC	D0:C4:F7:58:6C:12
`alog.umeng.com`	Body	header, device_id	IMEI	354917158514924
`d.applovin.com`	Body	device_info, idfa	Google Ad ID	0e5f5a7d-a3e4-...

total. We group the least commonly transmitted smartphone IDs under the *Other Smartphone IDs* column, which include the following: device serial number, IMSI, SIM serial number and registered Google account email. Furthermore, we found 29% of the extracted rules to be app-assigned IDs.

Analyzing the extracted rules for the top-popularity, interactive experiments, we found that 50% of the identifiers are sent in the URI of the HTTP requests (291 rules). In 39% (225) of the rules, the IDs are sent in the HTTP request body, using the POST method. Only 3% (18) of the cases, the identifier was sent in a cookie. The average identifier length in our rule set is 26.4 characters. A sample of identifying rules is given in Table 3.

After extracting identifier detection rules, we apply them to the traffic captured

Table 4: The most common third-party hosts found to collect at least an identifier over unencrypted connections. The listed hosts are contacted by the highest number of apps (based on 420 top-popularity apps, interactive experiment).

Host	# Apps	Collected IDs
data.flurry.com	43	Android ID
ads.mopub.com	32	Google advertising ID
apps.ad-x.co.uk	22	Google advertising ID, IMEI, Serial number, Android ID
alog.umeng.com	16	IMEI
a.applovin.com	16	Google advertising ID

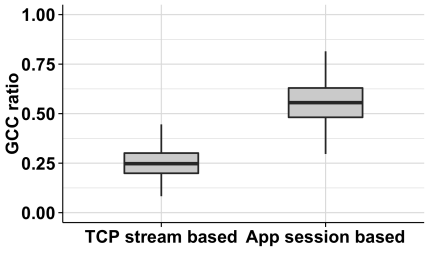
during the experiments. Due to space constraints we present the detailed results on the transmitted IDs in the unabridged version of this paper [40].

Moreover, analyzing the traffic captures of the top-popularity apps, we found that certain apps send precise location information (29 apps), email address (7 apps) and phone number (2 apps) in the clear. Together with the linking attack presented in this paper, this allows an adversary to link significantly more traffic to real-life identities.

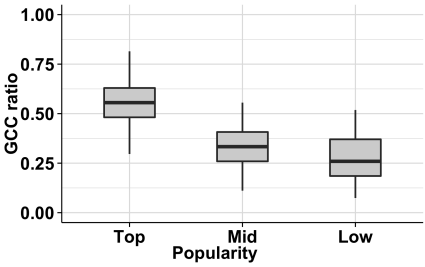
We found that 1076 different hosts were contacted over unencrypted connections during the experiments for the top-popularity apps in the interactive mode. The *data.flurry.com* domain is the most popular third-party domain collecting Android ID from 43 different apps (Table 4). Note that *data.flurry.com* received a notable mention in the slides of the BADASS program [15] for its identifier leakage.

7.2 Traffic Clustering

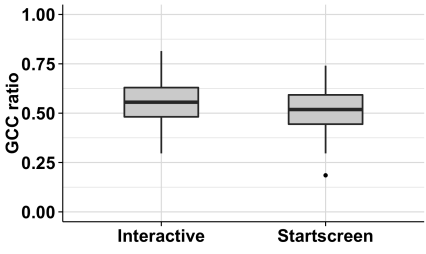
Here we evaluate the adversary's success in terms of unencrypted app traffic ratio (GCC ratio) that he can link together in different settings. We follow the analysis methodology explained in Section 5.2 and present clustering results for 100 randomly selected combinations of 27 apps. We pick 27 apps since it is the average number of apps used per month according to a recent survey [8]. Running 100 iterations with a different combination of (27) apps allowed us to reduce the variance between different runs and account for all the studied apps. We only consider apps that send at least one HTTP request and calculate the GCC ratio based on the unencrypted traffic. For the top-popular apps in interactive mode, these account for 69% of the apps. For simplicity, we will



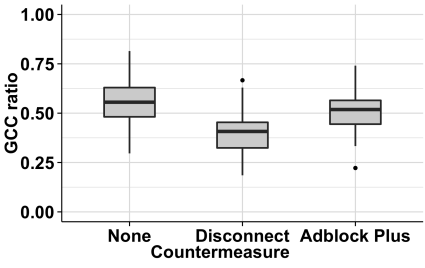
(a) GCC ratio for the top-popularity apps, shown for TCP stream and app session based linking.



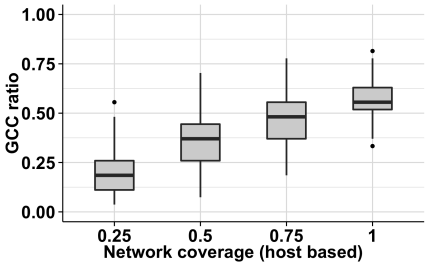
(b) GCC ratio for apps of different popularity levels for interaction mode.



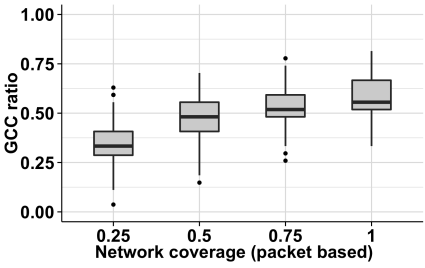
(c) GCC ratio for top-popularity apps, shown for interaction and startscreen mode.



(d) GCC ratio for the top-popularity apps, shown while using different privacy enhancing tools.



(e) GCC ratio for the top-popularity apps, shown for different network coverage levels of a *host based* restricted adversary.



(f) GCC ratio for the top-popularity apps, shown for different network coverage levels of a packet based restricted adversary.

present the clustering results for only one phone and a single run of each app. The results from two phones did not have any significant difference.

Effect of using TCP timestamps for traffic linking. The left boxplot in Fig. 3a, shows that when the adversary does not take TCP timestamps into account (TCP stream based linking), he can cluster 25% of users' unencrypted traffic. However, by exploiting TCP timestamps he can increase the GCC ratio to 57%.

Effect of app popularity Fig. 3b shows that popularity has a significant impact on the linking success of the adversary. The most popular apps allow the adversary to cluster 57% of the unencrypted traffic, while the apps from the mid-popular and low-popular level result in a GCC ratio of 32% and 28%, respectively.

Due to space constraints, we will only present results for the apps from the top-popularity level in the rest of this section.

Effect of user interaction. Fig. 3c shows the GCC ratio for two different experiment modes, *interaction* and *startscreen*. Although, the number of identifiers sent in two modes are significantly different (577 vs. 337), the graph shows a similar GCC ratio around 53% for two modes. A possible explanation is that the identifiers that enable linking are already sent as soon as the app is started.

Effect of countermeasures. Fig. 3d shows that both ad-blocking apps provide a limited protection against linking of the app traffic. Using Adblock Plus leads to an average linking of 50%. Disconnect Malvertising performs better, with a GCC rate of 38%, reduced from 57%.

Restricted adversary. Fig. 3c shows that an adversary that can only intercept traffic to 50% of the hosts can link up to 38% of the unencrypted mobile app sessions. For the *packet based* restricted adversary model, we observe that an adversary with a limited coverage of 25% of the user's packets can still link 37% of all app sessions together (Fig. 3d). In both models restricted adversary's success grows almost linear with his network coverage. Note that packet based restricted adversary can link significantly more traffic than the host-based model for the same network coverage ratio. This may be due to being able to observe packets from more hosts which will allow to link apps across sessions.

8 Limitations

Some apps may not be fully discovered by *The Monkey*, leading to an incomplete view of the network traffic. Also, apps that require user logins may not be sufficiently analyzed by our automated methodology. For those reasons, our results should be taken as lower bounds.

While we assume that the smartphones can be distinguished by their TCP timestamps, some middleboxes may interfere with user traffic. Firewalls, proxies or cache servers may terminate outgoing HTTP or TCP connections and open a new connection to the outside servers. Furthermore, end-user NAT devices may have various configurations and hence behave differently compared to enterprise NATs. In such cases, the adversary's ability to link traffic by TCP timestamps may be reduced.

We used *rooted* Android phones in our experiments. Although rooting the phones may introduce changes in the observed traffic, we assumed the changes to be minimal.

9 Conclusion

The revealed slides of the BADASS program have shown that unencrypted mobile app traffic is exploited for mass surveillance. Identifiers sent in the clear by the mobile applications allow targeting mobile users, linking of their traffic and building a database of their online activities.

In this study, we evaluated the surveillance threat posed by a passive network adversary who exploits mobile app traffic for surveillance purposes. We presented a novel framework that automates the analysis of mobile app network traffic. Our framework and methodology is designed to be flexible and can be used in other mobile privacy studies with slight modifications.

Our results show that using TCP timestamps and unique identifiers sent in the unencrypted HTTP traffic, a global adversary can cluster 57% of users' unencrypted mobile app sessions. We demonstrated that a passive adversary can automatically build a rule set that extracts unique identifiers in the observed traffic, which serves as a "selector" list for targeting users.

Our results suggest that popular apps leak significantly more identifiers than the less popular apps. Furthermore, while interacting with the app increases

the number of leaked identifiers, solely starting an app amounts to the same attack effectiveness.

We evaluated two countermeasures designed to block mobile ads and found that they provide a limited protection against linking of the user traffic. Encrypting mobile app traffic can effectively protect against passive network adversaries. Moreover, a countermeasure similar to HTTPS Everywhere browser extension can be developed to replace insecure HTTP connections of mobile apps with secure HTTPS connections on the fly.

Acknowledgment

Thanks to Yves Tavernier for sharing his valuable insights about middleboxes, and anonymous reviewers for their helpful and constructive feedback. This work was supported by the Flemish Government FWO G.0360.11N Location Privacy, FWO G.068611N Data mining and by the European Commission through H2020-DS-2014-653497 PANORAMIX and H2020-ICT-2014-644371 WITDOM.

References

[1] APK Downloader [Latest] Download Directly | Chrome Extension v3 (Evozi Official). http://apps.evozi.com/apk-downloader/.

[2] Cross Reference: /external/kernel-headers/original/asm-arm/param.h. http://androidxref.com/4.1.2/xref/external/kernel-headers/original/asm-arm/param.h#18.

[3] dpkt 1.8.6.2 : Python Package Index. https://pypi.python.org/pypi/dpkt.

[4] dtmilano/AndroidViewClient. https://github.com/dtmilano/AndroidViewClient/.

[5] dumpcap - The Wireshark Network Analyzer 1.12.2. https://www.wireshark.org/docs/man-pages/dumpcap.html.

[6] Nmap Network Scanning - Remote OS Detection - Usage and Examples. http://nmap.org/book/osdetect-usage.html.

[7] NSA Prism program taps in to user data of Apple, Google and others. http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data.

[8] Smartphones: So many apps, so much time. http://www.nielsen.com/us/en/insights/news/2014/smartphones-so-many-apps--so-much-time.html.

[9] SystemClock | Android Developers. http://developer.android.com/reference/android/os/SystemClock.html.

[10] Identifying App Installations | Android Developers Blog. http://android-developers.blogspot.be/2011/03/identifying-app-installations.html, 2011.

[11] 'Tor Stinks' presentation. http://www.theguardian.com/world/interactive/2013/oct/04/tor-stinks-nsa-presentation-document, 2013.

[12] About Adblock Plus for Android. https://adblockplus.org/android-about, 2015.

[13] Disconnect Malvertising for Android. https://disconnect.me/mobile/disconnect-malvertising/sideload, 2015.

[14] Manpage of TCPDUMP. http://www.tcpdump.org/tcpdump_man.html, 2015.

[15] Mobile apps doubleheader: BADASS Angry Birds. http://www.spiegel.de/media/media-35670.pdf, 2015.

[16] Selenium - Web Browser Automation. http://docs.seleniumhq.org/, 2015.

[17] UI/Application Exerciser Monkey | Android Developers. http://developer.android.com/tools/help/monkey.html, 2015.

[18] Gunes Acar, Christian Eubank, and Steven Englehardt. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2014.

[19] Mahesh Balakrishnan. Where's That Phone?: Geolocating IP Addresses on 3G Networks. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference (IMC)*, pages 294–300, 2009.

[20] Steven M. Bellovin. A Technique for Counting NATted Hosts. *Proceedings of the second ACM SIGCOMM Workshop on Internet measurment (IMW)*, page 267, 2002.

[21] Paul E. Black. Ratcliff/Obershelp Pattern Recognition. `http://xlinux.nist.gov/dads/HTML/ratcliffObershelp.html`, December 2004.

[22] Elie Bursztein. Time has Something to Tell us About Network Address Translation. In *Proceedings of NordSec*, 2007.

[23] Shuaifu Dai, Alok Tongaonkar, Xiaoyin Wang, Antonio Nucci, and Dawn Song. NetworkProfiler: Towards Automatic Fingerprinting of Android Apps. *2013 Proceedings IEEE International Conference on Computer Communications (INFOCOM)*, pages 809–817, April 2013.

[24] William Enck, Landon P Cox, Peter Gilbert, and Patrick Mcdaniel. TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones. *OSDI'10 Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, 2010.

[25] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W Felten. Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pages 289–299, 2015.

[26] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android Permissions Demystified. *Proceedings of the 18th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, page 627, 2011.

[27] MC Grace, Wu Zhou, X Jiang, and AR Sadeghi. Unsafe Exposure Analysis of Mobile In-App Advertisements. *Proceedings of the fifth ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, 067(Section 2), 2012.

[28] The Guardian. GCHQ Taps Fibre-Optic Cables for Secret Access to World's Communications. `https://www.theguardian.com/uk/2013/jun/21/gchq-cables-secret-world-communications-nsa`.

[29] Peter Hornyack, Seungyeop Han, Jaeyeon Jung, Stuart Schechter, and David Wetherall. These Aren't the Droids You're Looking for: Retrofitting Android to Protect Data From Imperious Applications. In *Proceedings of the 18th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 639–652. ACM, 2011.

[30] Van Jacobson, Robert Braden, Dave Borman, M Satyanarayanan, JJ Kistler, LB Mummert, and MR Ebling. RFC 1323: TCP Extensions for High Performance, 1992.

[31] Tadayoshi Kohno, Andre Broido, and Kimberly C Claffy. Remote Physical Device Fingerprinting. *IEEE Transactions on Dependable and Secure Computing*, 2(2):93–108, 2005.

[32] Moxie Marlinspike. New Tricks for Defeating SSL in Practice. *BlackHat DC, February*, 2009.

[33] Steven J Murdoch. Hot or Not: Revealing Hidden Services by Their Clock Skew. In *Proceedings of the 13th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 27–36. ACM, 2006.

[34] Ashkan Soltani, Andrea Peterson, and Barton Gellman. NSA uses Google Cookies to Pinpoint Targets for Hacking. `https://www.washingtonpost.com/news/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking/`, 2013.

[35] Ryan Stevens, Clint Gibler, and Jon Crussell. Investigating User Privacy in Android Ad Libraries. *IEEE Mobile Security Technologies (MoST)*, page 10, 2012.

[36] Guillermo Suarez-Tangil, Mauro Conti, Juan E Tapiador, and Pedro Peris-Lopez. Detecting Targeted Smartphone Malware With Behavior-triggering Stochastic Models. In *19th European Symposium on Research in Computer Security (ESORICS)*, pages 183–201. Springer, 2014.

[37] Ali Tekeoglu, Nihat Altiparmak, and Ali Şaman Tosun. Approximating the number of active nodes behind a NAT device. In *2011 Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN)*, pages 1–7. IEEE, 2011.

[38] Alok Tongaonkar, Shuaifu Dai, Antonio Nucci, and Dawn Song. Understanding Mobile App Usage Patterns Using In-app Advertisements. In *14th International Conference on Passive and Active Measurement (PAM)*, volume 7799 of *Lecture Notes in Computer Science*, pages 63–72. Springer, 2013.

[39] Eline Vanrykel. Passive Network Attacks on Mobile Applications. Master's thesis, Katholieke Universiteit Leuven, 2015.

[40] Eline Vanrykel, Gunes Acar, Michael Herrmann, and Claudia Diaz. Exploiting Unencrypted Mobile Application Traffic for Surveillance (Technical Report). `https://securewww.esat.kuleuven.be/cosic/publications/article-2602.pdf`, 2016.

[41] David Weinstein. Leaking Android Hardware Serial Number to Unprivileged Apps. `http://insitusec.blogspot.be/2013/01/leaking-android-hardware-serial-number.html`, 2013.

[42] Georg Wicherski, Florian Weingarten, and Ulrike Meyer. IP Agnostic Real-time Traffic Filtering and Host Identification Using TCP Timestamps. In *38th IEEE Conference on Local Computer Networks (LCN)*, pages 647–654. IEEE, 2013.

[43] Ning Xia, Han Hee Song, Yong Liao, and Marios Iliofotou. Mosaic: Quantifying Privacy Leakage in Mobile Networks. *Proceedings of the 2013 ACM SIGCOMM Conference*, (ii):279–290, 2013.

[44] Sebastian Zander and Steven J Murdoch. An Improved Clock-skew Measurement Technique for Revealing Hidden Services. In *USENIX Security Symposium*, pages 211–226, 2008.

Publication

Practical Privacy-Preserving Location-Sharing Based Services with Aggregate Statistics

Publication Data

Michael Herrmann, Alfredo Rial, Claudia Diaz, and Bart Preneel. Practical Privacy-Preserving Location-Sharing Based Services with Aggregate Statistics. In *Proceedings of the 2014 ACM Conference on Security and Privacy in Wireless & Mobile Networks (WiSec)*, pages 87–98. ACM, 2014.

Contributions

- Main author for Sections 1, 2, 5, 7 and 8.

Practical Privacy-Preserving Location-Sharing Based Services with Aggregate Statistics

Michael Herrmann[1], Alfredo Rial[2], Claudia Diaz[1], and Bart Preneel[1]

[1] KU Leuven ESAT/COSIC, iMinds, Leuven, Belgium
{name.surname}@esat.kuleuven.be
[2] IBM Research, Rüschlikon, Switzerland
lia@zurich.ibm.com

Abstract. Location-sharing-based services (LSBSs) allow users to share their location with their friends in a sporadic manner. In currently deployed LSBSs users must disclose their location to the service provider in order to share it with their friends. This default disclosure of location data introduces privacy risks. We define the security properties that a privacy-preserving LSBS should fulfill and propose two constructions. First, a construction based on identity based broadcast encryption (IBBE) in which the service provider does not learn the user's location, but learns which other users are allowed to receive a location update. Second, a construction based on anonymous IBBE in which the service provider does not learn the latter either. As advantages with respect to previous work, in our schemes the LSBS provider does not need to perform any operations to compute the reply to a location data request, but only needs to forward IBBE ciphertexts to the receivers. We implement both constructions and present a performance analysis that shows their practicality. Furthermore, we extend our schemes such that the service provider, performing some verification work, is able to collect privacy-preserving aggregate statistics on the locations users share with each other.

1 Introduction

The emergence of mobile electronic devices with positioning capabilities (e.g. through the Global Positioning System, GPS), such as smartphones and tablet computers, has fostered the appearance of a wide variety of Location Based Services (LBSs). With these services, users can find nearby places of

interest, share their location with friends, and obtain information about their surroundings.

Location-sharing-based services (LSBSs) permit users to share their current location or activity with other people. The shared location data may be in the form of GPS coordinates, although in GeoSocial Networks (GSN), such as *Foursquare* and *Facebook-check-in*, it is common that users announce their location in a more socially meaningful way by providing the venue (e.g., name of the restaurant) at which they are currently present. The action is commonly referred to as *check-in*. Every day millions of users enjoy GSN and share millions of locations with each other.[3]

While LSBSs are indeed useful, the disclosure of location data raises significant privacy concerns. Service providers and other third parties with access to accurate location data can infer private user information, such as their movement patterns, home address, lifestyle and interests [26]. Further, making these inferences is even easier if users share the venue rather than just submitting coordinates, as any uncertainties introduced by possible inaccuracies in the GPS coordinates are removed. We note that, although GSNs offer configurable privacy settings [27], they are still privacy invasive, as the LSBS provider learns the users' location regardless of the privacy settings.

Location Privacy Preserving Mechanisms (LPPMs) that implement obfuscation strategies, such as adding dummy locations [35] or reducing precision [31], are unsuitable for LSBS. This is because, when transmitting an obfuscated location to the service provider, the service provider naturally is only able to forward this obfuscated location to the user's friends. This conflicts with the main functionality of LSBSs. Therefore, LPPMs have been proposed in which users share keys allowing them to encrypt and decrypt their location data [23, 44]. In those solutions, the LSBS provider stores encrypted location data and computes the reply for a user requesting location data of her friends. A provider offering such an LSBS is unable to learn statistics about its users' whereabouts. Consequently, this renders the common business model of offering a free service in exchange for the users' data impossible. An alternative is to offer a paid service, but this might only be feasible if the fees are sufficiently low.

In this paper we propose two schemes based on identity-based broadcast encryption [21]. The first protocol reveals the identities of the friends that should receive location information to the LSBS provider and also to the other receivers of that location information. In the second protocol, those identities are hidden towards the service provider as well as towards other users (including

[3] `https://foursquare.com/about/`
`http://www.socialbakers.com/blog/167-interesting-facebook-places-numbers`

the receivers of the location update), thanks to the use of anonymous identity-based broadcast encryption. The advantage over existing work is that in our schemes the LSBS provider does not need to perform complex operations in order to compute a reply for a location data request, but only needs to forward data. This reduces the cost of an LSBS provider that is then able to offer its service for a lower price if pursuing a subscription-based business model. Furthermore, we extend our schemes such that the service provider is able to collect privacy-preserving statistics on the locations shared by the users. This extension does require the LSBS provider to perform additional computations. The obtained statistics could be monetized to compensate for this additional overhead or to facilitate a free service.

We have implemented both schemes on a *Samsung S III mini* smartphone and provide results on the computation time, bandwidth overhead and energy consumption. Our evaluation shows that the performance of the first scheme is independent of the number of users in the system. Furthermore, it imposes minimal computational and bandwidth overhead on both, the LSBS provider and the users' mobile devices. In the second scheme a user is able to choose a trade-off between privacy, computation and bandwidth overhead. We study this trade-off and provide recommendations to increase the level of privacy for the same amount of resources.

The remainder of this paper is structured as follows. In Section 2, we review previous work on privacy-preserving location-based services and argue that none of the existing approaches is suitable for implementing privacy-preserving LSBSs. We define privacy-preserving LSBS in Section 3. In Section 4, we introduce our two schemes. We provide a detailed performance analysis showing the feasibility of our approach in Section 5. In Section 6, we extend our schemes to allow for aggregate statistics collection. We discuss our approach and results in Section 7. Finally, in Section 8 we conclude our work.

2 Related Work

In this section we review obfuscation-based LPPMs and argue that they are not suitable for protecting location privacy in LSBS. Subsequently, we review LPPMs that rely on cryptographic primitives. Some of them have been deliberately designed for protecting location privacy in LSBS; others have a more general purpose. Our evaluation shows that obfuscation-based LPPMs are not suitable for privacy-preserving LSBS and that our system has several advantages over existing privacy-preserving LSBS.

Other works have examined location privacy in GSNs considering a different threat model. Gambs et al. [27] and Vicente et al. [47] review several GSNs and analyze their privacy issues. However, in their privacy evaluation, they do not consider it to be a privacy breach if the service provider learns the user locations. An analysis of the inferences that can be made about users based on where they check-in while using Foursquare is provided by Pontes et al. [43].

2.1 Obfuscation-based LPPMs

While works like [23,44] have already noted that LPPMs based on anonymization and precision-reduction are not suitable to protect location privacy in LSBS, we provide here a more detailed evaluation. We therefore follow the categorization in [46] which distinguishes between these four obfuscation strategies: *location hiding*, *perturbation*, *adding dummy regions*, and *reducing precision*. In the following we argue that none of these obfuscation-based LPPMs are applicable to protect location privacy in LSBSs. We therefore consider the following LSBS application scenario: A user A is currently at one of her favorite locations and wishes to share this information with her friends. This could be either because A simply wants to inform her friends, or to enable them to join her at this place.

The *location hiding* strategy [5] consists of not sending the location data to the LBS and is thus impractical. In this case user A would not be able to share her location with her friends. Some LPPMs propose to change pseudonyms after a period of location hiding [32, 37]. However, this is also impractical, because the check-in is supposed to be received by the same set of friends and therefore identifies user A. LPPMs that rely on *perturbation* submit a location different from A's actual location [34]. As a rather inconvenient result, the user's friends learn a wrong place and if they decide to join their friend, they would realize upon arriving that A is actually not present there. The *adding dummy regions* strategy [35] consists of submitting fake locations along with the user's actual location. In this case A would check-in at several places and her friends could not distinguish real from fake check-ins. Finally, with the *reducing precision* strategy [31] A would send a *cloak region* that contains her current location, but she would not reveal her precise whereabouts, making it extremely difficult for her friends to find her. We note that with all obfuscation strategies, users could rely on personal contact in order to obtain A's precise location after learning the obfuscated location. However, such a system would have significant usability issues. The key limitation of these techniques is that they do not make a distinction between information revealed to friends and to the service provider. Thus, in order to protect their location information towards the service provider, users must lower the quality of location information shared with their friends.

2.2 Cryptographic LPPMs

Freudiger [25] proposes that users should use symmetric or asymmetric encryption and use the service provider solely as a rendez-vous point to exchange encrypted data. Longitude by Dong and Dulay [23] propose to use proxy-encryption, which guarantees that the service provider is not able to learn the location update and, furthermore, that the ciphertext can be modified by the service provider such that only intended receivers are able to successfully decrypt. Puttaswamy et al. [44] propose a scheme which combines encryption with location transformation in order to build a location-based application, such as privacy-preserving LSBS. As already mentioned in Section 1 the proposed LPPMs impose a computational overhead at the LSBS provider, which makes offering such a service more expensive. In our schemes the cost for the LSBS provider is kept at a minimum in order to make running such a service as cheap as possible. Furthermore, the service provider can decide to engage in additional computation overhead and therefore obtain statistics about its users' whereabouts. We note that this overhead is kept low since the service provider only needs to forward data and verify zero-knowledge proofs, whose cost can be reduced using batch verification. Note however that our scheme, in contrast to the works mentioned above, introduces a trusted key generation center.

Carbunar et al. [17] propose privacy preserving protocols for badge and mayor applications in GSNs. While this is closely related to our work, their scheme does not allow users to exchange their locations.

In privacy-preserving friend nearby notification, users can privately learn whether a friend is in close proximity. Such a service can be realized by homomorphic encryption [49], private equality testing [41] and private threshold set intersection [39]. These protocols are different from our solution, because in privacy-preserving location sharing protocols, location updates are sent to friends regardless of their current location, i.e. regardless of how close they are.

Bilogrevic et al. [6] propose two protocols to allow users to compute a fair rendez-vous point in a privacy preserving manner. This differs from our work in that we focus on location sharing, and not on deciding on where to meet after a group of users has deliberately decided to do so.

Popa et al. [162] propose a privacy-preserving protocol to compute aggregate statistics from users' location updates. However, in this protocol users do not share their location with other users.

Finally, some works employ Private Information Retrieval (PIR) so that the users retrieve information (e.g., points of interest) related to their surroundings [30,42]. PIR could in principle be employed to build privacy-preserving LSBS. However,

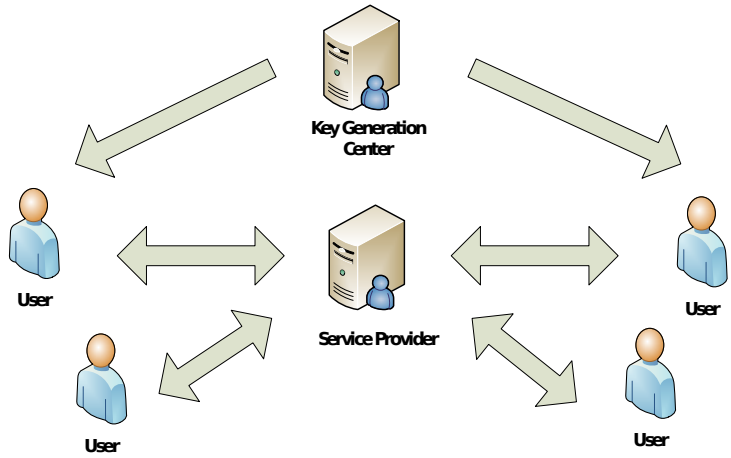


Figure 1: System Model of a privacy-preserving LSBS. The key generation center sets up the public parameters and provides users with secret keys. The service provider transfers messages between users.

PIR operations are rather costly at the service provider side which we again argue that introduces intolerable overhead for a service provider.

3 Definition of Privacy for LSBS

Our LSBS involves the following parties: a key generation center, a service provider $\mathcal{P}$ and a set of users $\mathcal{U}_i$ for $i = 1$ to n. Figure 1 shows the parties in the system.

A privacy-invasive protocol that realizes the desired functionality works as follows. A user $\mathcal{U}_i$ sends a message to the service provider that indicates the place *loc* that $\mathcal{U}_i$ wishes to share, and the set $S \subseteq [1, n]$ of users $\mathcal{U}_j$ $(j \in S)$ that should learn that $\mathcal{U}_i$ shares *loc*. Then the service provider forwards to users $\mathcal{U}_j \in S$ the message $(\mathcal{U}_i, loc)$ to inform them that $\mathcal{U}_i$ has shared *loc*. As can be seen, this protocol is privacy-invasive. The service provider learns the location *loc* that $\mathcal{U}_i$ shares, and the identities of the users $\mathcal{U}_j$ $(j \in S)$. The privacy properties that our LSBS should fulfill are the following:

Sender Privacy. No coalition of users $\mathcal{U} \notin S$ and service provider $\mathcal{P}$ should learn any information on *loc*.

Receiver Privacy. No coalition of users $\mathcal{U}_j$ such that $j \neq \{i, j^*\}$ and service provider $\mathcal{P}$ should learn any information on whether $j^* \in S$ or $j^* \notin S$.

Our schemes in Section 4 are secure against active adversaries, i.e., adversaries that deviate in any possible way form the protocol execution. The security of our schemes relies on the security of identity based broadcast encryption. The key generation center is trusted, which is an assumption that every identity-based cryptographic scheme makes.

4 Constructions of LSBS

Our schemes are based on identity-based broadcast encryption, which we describe in Section 4.1. In Section 4.2, we describe the *sender private* scheme, which fulfills the sender privacy property. In Section 4.3, we describe the *fully private* scheme, which fulfills both the sender privacy and the receiver privacy properties.

4.1 Identity-Based Broadcast Encryption

Broadcast encryption allows a sender to encrypt a message m to a set of receivers $S \in [1, n]$, so that no coalition of receivers not in S can decrypt. A broadcast encryption scheme consists of the following algorithms:

$\mathsf{Setup}(1^\lambda, n, \ell)$**.** On input the number of users n, the security parameter 1^λ, and the maximum size $\ell \leq n$ of a broadcast recipient group, output the public key pk and the secret key sk.

$\mathsf{KeyGen}(i, sk)$**.** On input an index i and the secret key sk, output a secret key d_i for user $\mathcal{U}_i$.

$\mathsf{Enc}(pk, S, m)$**.** On input the recipient group $S \in [1, n]$, the public key pk and the message m, output the ciphertext c.

$\mathsf{Dec}(pk, d_i, c)$**.** On input the public key pk, the secret key d_i of user $\mathcal{U}_i$ and a ciphertext c, output m if $i \in S$ or else the failure symbol $\perp$.

In IBBE, a trusted key generation center $\mathcal{KGC}$ creates the parameters and computes the secret keys of the receivers. Note that the secret key sk allows the decryption of every ciphertext. If ciphertexts c do not reveal the set of receivers S, the broadcast encryption scheme is anonymous.

4.2 Sender-Private LSBS

Our sender-private LSBS (SPLS) uses an IBBE scheme that is not anonymous. In such a scheme, ciphertexts c contain a description of the recipient group S. Our scheme works as follows:

Setup Phase. $\mathcal{KGC}$ executes the setup algorithm $\mathsf{Setup}(1^\lambda, n, \ell)$ on input the security parameter 1^λ, the number of users n and the maximum size $\ell \leq n$, publishes the public key pk and stores the secret key sk. Users obtain pk.

Registration Phase. Each user $\mathcal{U}_i$ registers with the service provider by sending the index i. Additionally, $\mathcal{U}_i$ receives the key d_i from $\mathcal{KGC}$, which runs $\mathsf{KeyGen}(i, sk)$.

Main Phase. To share a location loc, a user $\mathcal{U}_i$ runs $c \leftarrow \mathsf{Enc}(pk, S, i||loc)$ and sends c to the service provider $\mathcal{P}$. $\mathcal{P}$ gets S from c and sends c to the users $\mathcal{U}_j$ $(j \in S)$. Each user $\mathcal{U}_j$ runs $\mathsf{Dec}(pk, d_j, c)$ to output the message $i||loc$.

We note that the registration and main phases can be interleaved, i.e., users can join our SPLS dynamically.

Our scheme fulfills the sender privacy property. The IBBE scheme ensures that no coalition of service provider $\mathcal{P}$ and users $\mathcal{U} \notin S$ can decrypt a ciphertext c computed on input S. However, this scheme does not fulfill the receiver privacy property. Since the IBBE scheme is not anonymous, the ciphertext c reveals the identity of the receivers $\mathcal{U}_j$ $(j \in S)$.

Construction of IBBE

In Section 5, we instantiate our SPLS with a broadcast encryption scheme in order to implement it and evaluate its performance. Broadcast encryption was first formalized by Fiat and Naor [24]. The first fully collusion secure broadcast schemes, i.e., schemes where security holds even if all the users not in the recipient group S collude, were described in [40]. The first public key broadcast encryption scheme was proposed in [22].

In the schemes mentioned above, the size of the ciphertext grows linearly with the size of the recipient group. Boneh, Gentry and Waters [9] proposed the first schemes with constant size ciphertexts. Their schemes have selective security, i.e., the adversary should decide the target recipient group to be attacked before the setup phase. Identity-based broadcast encryption was proposed in [21], which describes also selectively secure schemes.

Broadcast encryption and identity-based broadcast encryption with adaptive security was first achieved in [28]. These schemes achieve constant size ciphertexts in the random oracle model and under q-based assumptions. In [48] and [36], broadcast encryption schemes secure under static assumptions are proposed. In [36], an identity-based broadcast encryption scheme secure under static assumptions is also proposed, but it only achieves selective security.

The main property of identity-based broadcast encryption is that the scheme is efficient when the total number of users n is exponential in the security parameter 1^λ. Since our SPLS could have millions of users, schemes that, despite having constant size ciphertexts, have public key or user secret keys of size that grows linearly with n are less suitable. Therefore, we instantiate our SPLS with the adaptively secure identity-based broadcast encryption in [28]. This scheme, which is secure in the random oracle model, in addition to constant size ciphertexts, has a public key of size independent of n that grows linearly with ℓ and user decryption keys d_i of constant size. The scheme in [28] employs bilinear maps.

Bilinear maps Let $\mathbb{G}$, $\tilde{\mathbb{G}}$ and $\mathbb{G}_t$ be groups of prime order p. A map $e : \mathbb{G} \times \tilde{\mathbb{G}} \to \mathbb{G}_t$ must satisfy bilinearity, i.e., $e(g^x, \tilde{g}^y) = e(g, \tilde{g})^{xy}$; non-degeneracy, i.e., for all generators $g \in \mathbb{G}$ and $\tilde{g} \in \tilde{\mathbb{G}}$, $e(g, \tilde{g})$ generates $\mathbb{G}_t$; and efficiency, i.e., there exists an efficient algorithm $\mathsf{BMGen}(1^k)$ that outputs the pairing group setup $(p, \mathbb{G}, \tilde{\mathbb{G}}, \mathbb{G}_t, e, g, \tilde{g})$ and an efficient algorithm to compute $e(a, b)$ for any $a \in \mathbb{G}$, $b \in \tilde{\mathbb{G}}$. If $\mathbb{G} = \tilde{\mathbb{G}}$ the map is symmetric and otherwise asymmetric.

Let (E, D) be a secure symmetric key encryption scheme. The scheme in [28] works as follows:

$\mathsf{Setup}(1^\lambda, n, \ell)$**.** On input the number of users n, the security parameter 1^λ, and the maximum size $\ell \leq n$ of a broadcast recipient group, run $(p, \mathbb{G}, \mathbb{G}_t, e) \leftarrow \mathsf{BMGen}(1^\lambda)$. Set $g_1, g_2 \leftarrow \mathbb{G}$. Set $\alpha, \beta, \gamma \leftarrow \mathbb{Z}_p$. Set $\hat{g}_1 \leftarrow g_1^\beta$ and $\hat{g}_2 \leftarrow g_2^\beta$. Set $pk = (p, \mathbb{G}, \mathbb{G}_t, e, n, \ell, g_1^\gamma, g_1^{\gamma \cdot \alpha}, \{g_1^{\alpha^j}, \hat{g}_1^{\alpha^j}, g_2^{\alpha^k}, \hat{g}_2^{\alpha^k} : j \in [0, \ell], k \in [0, \ell - 2]\})$. Output pk and the secret key $sk = (\alpha, \gamma)$.

$\mathsf{KeyGen}(i, sk)$**.** On input an index i and the secret key sk, pick random $r_i \leftarrow \mathbb{Z}_p$ and output

$$d_i = (i, r_i, h_i = g_2^{\frac{\gamma - r_i}{\alpha - i}})$$

$\mathsf{Enc}(pk, S, m)$**.** On input the recipient group $S \in [1, n]$, the public key pk and the message m, set $\tau \leftarrow \{0,1\}^{\mathcal{O}(\lambda)}$. Set $F(x)$ as the $(\ell - 1)$-degree polynomial that interpolates $F(i) = H(\tau, i)$ for $i \in S$ and $F(i) = 1$ for $i \in [n + j]$ with $j \in [k + 1, \ell]$, where $H : \{0,1\}^{\mathcal{O}(\lambda)} \times [1, n] \to \mathbb{Z}_p$ is a

hash function modelled as a random oracle. Set $k = |S|$ and parse S as $\{i_1, \ldots, i_k\}$. Set $i_j \leftarrow n + j$ for $j \in [k+1, \ell]$. Set $P(x) = \prod_{j=1}^{\ell}(x - i_j)$. Set $t \leftarrow \mathbb{Z}_p$ and set $K \leftarrow e(g_1, \hat{g}_2)^{\gamma \cdot \alpha^{\ell-1} \cdot t}$. Set $\text{Hdr} \leftarrow \langle C_1, \ldots, C_4 \rangle = \langle \hat{g}_1^{P(\alpha)\cdot t}, g_1^{\gamma \cdot t}, g_1^{F(\alpha)\cdot t}, e(g_1, \hat{g}_2)^{\alpha^{\ell-1} \cdot F(\alpha) \cdot t} \rangle$. Compute $C \leftarrow \mathsf{E}(K, m)$ and output $c = (\tau, \text{Hdr}, C, S)$.

$\mathsf{Dec}(pk, d_i, c)$**.** On input the public key pk, the secret key d_i and a ciphertext c, parse d_i as $\langle i, r_i, h_i \rangle$, c as (τ, Hdr, C, S) and Hdr as $\langle C_1, \ldots, C_4 \rangle$. Define $P(x)$ as above and compute $F(x)$ from τ as above. Set

$$P_i(x) = x^{\ell-1} - \frac{P(x)}{(x-i)},\ F_i(x) = \frac{F(x) - F(i)}{(x-i)},$$

$$\text{and } e_i = -\frac{r_i}{F(i)}$$

and

$$K \leftarrow e(C_1, h_i \cdot g_2^{e_i \cdot F_i(\alpha)}) \cdot e(C_2 \cdot C_3^{e_i}, \hat{g}_2^{P_i(\alpha)}) / C_4^{e_i}.$$

Output $m \leftarrow \mathrm{D}(K, C)$.

We note that a user of LSBS usually shares her location with the same recipient group, i.e., with her friends. Therefore, broadcast encryption is used to share a symmetric key with that recipient group, and messages are encrypted using an efficient symmetric key encryption scheme. Broadcast encryption is used again only when the recipient group changes or when the symmetric key should be renewed for security reasons.

4.3 Fully-Private LSBS

Our fully-private LSBS (FPLS) uses an anonymous IBBE scheme. In such scheme, ciphertexts c do not reveal the recipient group. The setup and registration phases of this scheme work as the ones described in Section 4.2. The main phase works as follows:

Main Phase. To share a location loc, a user $\mathcal{U}_i$ runs $c \leftarrow \mathsf{Enc}(pk, S, i||loc)$ and sends c to the service provider $\mathcal{P}$. $\mathcal{P}$ forwards c to every user $\mathcal{U}_j$ such that $j \neq i$. Each user $\mathcal{U}_j$ runs $\mathsf{Dec}(pk, d_j, c)$ to get either the message $i||loc$ or $\perp$.

As in the construction in Section 4.2, this scheme fulfills the sender-private property. Additionally, this scheme fulfills the receiver privacy property. Since

the IBBE scheme is anonymous, the ciphertext c does not reveal the identity of the receivers $\mathcal{U}_j$ $(j \in S)$.

This construction requires location updates to be broadcast to all users. Therefore, we propose a variant that allows to trade-off communication efficiency and location-privacy. In this variant, the map is divided into regions $\text{reg}_1, \ldots, \text{reg}_r$ and users reveal to the service provider the region where they are located and the region from where they would like to receive location updates.

Region Phase. A user $\mathcal{U}_i$ sends to the service provider the region to which location updates she wishes to receive should be associated.

Main Phase. To share a location $loc \in \text{reg}$, a user $\mathcal{U}_i$ runs $c \leftarrow \mathsf{Enc}(pk, S, i||loc)$ and sends (c, reg) to the service provider $\mathcal{P}$. $\mathcal{P}$ forwards c to every user $\mathcal{U}_j$ such that $j \neq i$ and reg equals the region sent by $\mathcal{U}_j$ in the Region Phase. Each user $\mathcal{U}_j$ runs $\mathsf{Dec}(pk, d_j, c)$ to get either the message $i||loc$ or $\perp$.

Construction of Anonymous IBBE

In Section 5, we instantiate our FPLS with an anonymous broadcast encryption scheme in order to implement it and evaluate its performance. Barth et al. [4] propose an anonymous broadcast encryption scheme secure in the random oracle model where the ciphertext size is $\mathcal{O}(S)$. The public key size is $\mathcal{O}(n)$, while user secret keys and decryption time are constant. Libert et al. [38] proposed a scheme with the same asymptotic performance but secure in the standard model.

Recently, a scheme with public key size $\mathcal{O}(n)$, secret key size $\mathcal{O}(n)$, ciphertext size $\mathcal{O}(r\mathsf{log}(\frac{n}{r}))$, where r is the set $n - S$, and constant decryption time was proposed in [1]. Despite the fact that in this scheme ciphertexts do not grow linearly with n, actually $\mathcal{O}(r\mathsf{log}(\frac{n}{r}))$ is asymptotically larger than $\mathcal{O}(n - r)$ for large values of r, which are likely in our FPLS. Furthermore, the scheme in [1] does not provide anonymity with respect to users who are able to decrypt, i.e., those users can find out the identity of the other users who can decrypt.

We modify the scheme in [4] so that it employs as building block an anonymous identity-based encryption scheme instead of a key-private public key encryption scheme. This allows users to employ their identities as public keys. Such modification was suggested in Barth et al. [4] and security follows from the security of the original scheme.

An identity-based encryption (IBE) scheme consists of the algorithms ($\mathsf{IBESetup}$, $\mathsf{IBEExtract}, \mathsf{IBEEnc}, \mathsf{IBEDec}$). The algorithm $\mathsf{IBESetup}(1^\lambda)$ outputs parameters

$params$ and master secret key msk. $\mathsf{IBEExtract}(params, msk, id)$ outputs the secret key sk_{id} for identity id. $\mathsf{IBEEnc}(params, id, m)$ outputs ct encrypting m under id. $\mathsf{IBEDec}(params, sk_{id}, ct)$ outputs message m encrypted in ct.

An IBE scheme is *anonymous* [2] if it is not possible to associate the identity used to encrypt a message m with the resulting ciphertext. We employ the scheme in [8], which is anonymous [10], to implement the anonymous broadcast encryption scheme.

Another building block of the anonymous IBBE scheme is a strongly existentially unforgeable signature scheme. A signature scheme consists of algorithms (Kg, $\mathsf{Sign}, \mathsf{Vf}$). $\mathsf{Kg}(1^\lambda)$ outputs a key pair (ssk, vsk). $\mathsf{Sign}(ssk, m)$ outputs a signature s on message m. $\mathsf{Vf}(vsk, s, m)$ outputs accept if s is a valid signature on m and reject otherwise. We employ the scheme secure in the random oracle model proposed in [7].

The remaining building block is a secure symmetric key encryption scheme (E, D). We employ the advanced encryption standard [20]. The anonymous IBBE scheme works as follows.

$\mathsf{Setup}(1^\lambda, n, \ell)$**.** Choose a group $\mathbb{G}$ of primer order p where CDH is hard and DDH is easy and pick a generator $g \in \mathbb{G}$. Choose a hash function $H : \mathbb{G} \to \{0,1\}^\lambda$ which is modeled as a random oracle. Compute $params$ and msk via $\mathsf{IBESetup}(1^\lambda)$. For $i = 1$ to n, pick random $a_i \leftarrow \mathbb{Z}_p$. Output Output $pk = (\mathbb{G}, g, g^{a_1}, \ldots, g^{a_n}, H, params)$ and $sk = (msk, a_1, \ldots, a_n)$.

$\mathsf{KeyGen}(i, sk)$**.** On input an index i and the secret key sk, compute a secret key $sk_i \leftarrow \mathsf{IBEExtract}(params, msk, i)$. Output $d_i = (sk_i, a_i)$.

$\mathsf{Enc}(pk, S, m)$**.** On input the recipient group $S \in [1, n]$, the public key pk and the message m, execute the following steps.

1. Compute $(ssk, vsk) \leftarrow \mathsf{Kg}(1^\lambda)$.
2. Pick a random symmetric key K.
3. Pick random $r \leftarrow \mathbb{Z}_p$ and set $T = g^r$.
4. For each tuple $(i, g^{a_i}) \in S$, set the ciphertext $c_i \leftarrow H(g^{a_i r})||\mathsf{IBEEnc}(params, i, vsk||g^{a_i r}||K)$.
5. C_1 is the concatenation of all c_i ordered by the values of $H(g^{a_i r})$.
6. Compute $C_2 = \mathsf{E}_K(m)$.
7. Sign $s \leftarrow \mathsf{Sign}(ssk, T||C_1||C_2)$.
8. Return the ciphertext $C = s||T||C_1||C_2$.

$\mathsf{Dec}(pk, d_j, c)$**.** On input the public key pk, the secret key d_j and a ciphertext c, execute the following steps.

1. Calculate $l = H(T^{a_j})$.
2. Find c_j such that $c_j = l||c$ for some c_j in C_1.
3. Calculate $p \leftarrow \mathsf{IBEDec}(params, sk_j, c_j)$.
4. If p is $\perp$, return $\perp$.
5. Parse p as $vsk||x||K$.
6. If $x \neq T^{a_j}$, return $\perp$.
7. If $\mathsf{Vf}(vsk, s, T||C_1||C_2)$ outputs accept, then output $m = \mathsf{D}_K(C_2)$ else $\perp$.

We remark that, if the user is not in the recipient group and therefore she cannot decrypt, the decryption algorithm only requires the computation of a hash function.

5 Performance Analysis

Location-sharing-based applications are usually run on a mobile device, such as a smartphone or a tablet computer. Therefore, the available resources at the client side are limited in terms of computational power and bandwidth when on mobile connection. Furthermore, mobile devices usually have a rather low battery capacity. Thus an application must use the CPU or mobile communication interfaces, such as WiFi or 3G, as moderate as possible in order not to drain the battery too much. In order to evaluate the overhead of our schemes, we implemented them in the C programming language using the *Pairing-Based Cryptography*[4] (PBC) library. Subsequently, we imported the schemes within Android application using Android's Native Development Kit[5] (NDK) and tested the application on a Samsung S III mini (1 GHz dual-core CPU) which runs the CyanogenMod[6] operating system.

In the following we provide the *additional* overhead imposed by our schemes. This overhead is due to the computation of cryptographic operations and due to the transmission of key material and ciphertexts. Mobile applications for LSBSs, such as Foursquare, are used on a large user base and the overhead imposed by these services is accepted in practice.

[4] `http://crypto.stanford.edu/pbc/`
[5] `https://developer.android.com/tools/sdk/ndk/index.html`
[6] `http://www.cyanogenmod.org/`

For the energy consumption of our schemes, we measure the different current consumption of the phone when the CPU is idle and when the CPU load of one core is at the maximum. We found that the difference is 150 mA at 3.8 V and thus that the power consumption is 570 mW if one CPU core is at full load. Please note that we could not use PowerTutor[7] to estimate the energy consumption of our schemes, because PowerTutor was designed for a Nexus 1 mobile phone. Although PowerTutor does also run on our Samsung S III mini, the energy measurements are likely to be inaccurate, because both phones have a different CPU and we found that PowerTutor is unable to read the traffic sent and received on our phone. For the runtime estimation of our schemes, we executed our protocols 50 times and computed the average. Multiplying the runtime with the power consumption equals to the energy consumption of our schemes.

The energy consumption of data transmission via a mobile interface, such as WiFi and 3G, turns out to be significantly more difficult. This is because the actual energy consumption for sending and receiving data depends on many factors, such as amount of data, time between two consecutive data transmissions and network reception. We therefore use Balasubramanian et al. [3] work to outline ways to minimize the energy consumed by our schemes.

5.1 Evaluation of the Sender-Private LSBS

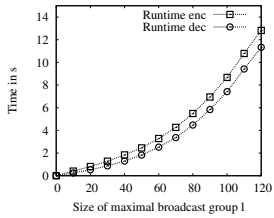
The complexity of the Gentry and Waters [28] scheme that we employ for our SPLS only depends on the size of the maximal broadcast group l. This means that for a given l, the computational and bandwidth overhead for computing the symmetric key stays constant, regardless of n (the number of users in the system) or k (actual receiver of a broadcast message). We therefore limit our evaluation to the system parameter l.

As we can see in Figure 2a the time for creating the symmetric key increases polynomially with l. For a reasonably large l, such as 100, our phone needs 8.66 seconds to compute the symmetric key in the encryption protocol and 7.42 in the decryption protocol. While this appears to be rather high, we must stress that a user is able to reuse a symmetric key until the broadcast group changes or the key got compromised. Therefore, a single symmetric key can be used for thousands of location shares. Furthermore, as we can see in Figure 2b, the actual energy consumed for computing a symmetric key is 4.94 Joule for the encryption protocol and 4.23 Joule for the decryption protocol, which is very low. The capacity of the standard battery of a Samsung S III mini (3.8 V and

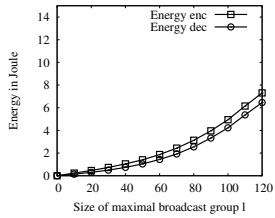
[7] `http://powertutor.org/`

1500 mAh) is 20520 Joule and therefore computing even dozens of symmetric keys per day would not drain the battery too much.

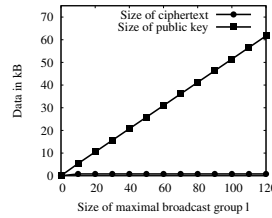
We show the bandwidth overhead of the FPLS in Figure 2c. For creating a new symmetric key a user needs to send 788 byte of data to the receiver of the broadcast group. Please note that even for $k > 1$ the user only needs to send 788 byte instead of $k \times 788$ byte, because the service provider forwards the traffic to the intended receivers. The public key of our scheme grows linearly in l. However, please note that the public key only needs to be sent very rarely. This is when a user signs up for the service, the new user receives the public key from the service provider, and if the service provider decides to increase/decrease the size of the maximal broadcast group and thus changes the parameter l. In those cases, all the users in the system receive the new public key.



(a) Overview of runtimes for encryption and decryption.



(b) Overview of energy consumption for encryption and decryption.



(c) Overview of bandwidth overhead.

Figure 2: Runtime, energy consumption and bandwidth overhead of the SPLS for increasing l and fixed $n = 1000$ and $k = 5$.

5.2 Evaluation of the Fully-Private LSBS

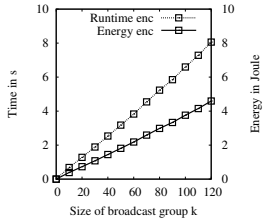
In the following we will first show that the computational overhead that is imposed by the FPLS is feasible for current mobile devices. Subsequently we will show that the FPLS imposes a significant bandwidth overhead. This is a problem for two reasons. Firstly, data plans usually include a fixed data volume to be transmitted before either the speed of the connection gets throttled or additional costs incur. Secondly, using mobile communication interfaces, such as 3G or WiFi, is expensive in terms of energy and therefore sending and receiving higher amounts of data additionally drains the battery. However, we make several suggestions on how our protocol should be deployed to significantly reduce the energy consumption, although the transferred data volume remains the same. Furthermore, as introduced in Section 4.3, the concept of big regions

greatly helps in making the FPLS feasible. We therefore consider the FPLS as a protocol which allows a very broad-ranged trade-off between location privacy and performance for users that do not wish to reveal their friendship graph. At the one extreme a very high level of location privacy is possible if the user is willing to spend the necessary resources. On the other extreme a user reveals accurate location information towards the service provider and thereby decreases the amount of data that is received.

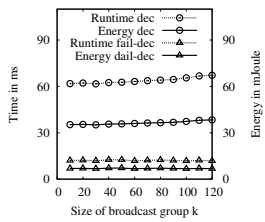
As we can see in Figure 3a, the computation and energy overhead for encryption grows only linearly in k and is therefore lower than in the SPLS. Figure 3b shows that the overhead for decryption is about two magnitudes lower than in the SPLS. Furthermore, Figure 3b shows the computation and energy overhead for fail-decryptions. These are the resources a user needs to spend in order to determine that a location update is actually not for her. Recalling the battery capacity of 20520 Joule, we can see that in terms of computational overhead the FPLS allows for sharing multiple locations per day and receiving thousands of location updates.

As already mentioned the main drawback of the FPLS is the data a user receives due to flooding. Figure 3c shows the lengths of a ciphertext as k increases. In [17], the authors state that the very majority of users has less than 100 friends. Therefore, we can assume most ciphertexts to have a length of at most 60 kB. If an LSBS would have 1 million users, this would result in approximately $10^6 \times 60\text{kB} = 57.22$ GB of data every user receives per day. Clearly, this is not feasible. However, choosing a big region such that only 0.01% of all the world-wide location updates are received results in a bandwidth overhead of 5.86 MB per day.

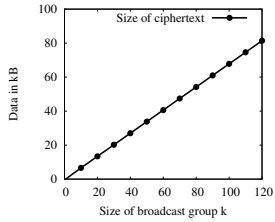
The energy that is consumed for receiving data can be optimized with two measures. Firstly, receiving data via the WiFi interface consumes significantly less energy than receiving the same amount of data via 3G interface [3] (§ 3.6). For example receiving 500 kB of data via 3G consumes around 19 Joule and only 5 Joule via WiFi. Secondly, as noted in [3] (§ 3.6.1), the energy consumption strongly depends on the inter-transfer time between downloads. For example, receiving 50 kB transmissions with inter-transfer time 1 second consumes around 5 Joule, while it consumes 10 Joule if the inter-transfer time is 9 seconds. We therefore suggest: (i) to receive most of the traffic when a WiFi connection is available; (ii) that the service provider caches location updates for a certain period and sends them all at once in order to have few but large data transmissions to the users.



(a) Overview of runtime and energy consumption for encryption.



(b) Overview of runtime and energy consumption for decryption and fail-decryption.



(c) Overview of bandwidth overhead.

Figure 3: Runtime, energy consumption and bandwidth overhead of the FPLS for increasing k and fixed $n = 1000$.

6 LSBS with Aggregate Statistics Collection

In our scheme, the service provider acts as a channel between users. The service provider relays messages between the sender and the receivers, but learns nothing about the content of the messages sent. The business model of currently deployed LSBS relies on learning user check-ins. Service providers use that information in order to, e.g., give recommendations on most visited places and give discounts to users that check-in a number of times at a particular location.

We describe how our scheme can be extended to allow the service provider to collect aggregated data on how many times users visit each of the locations. In this extension, each time a user checks-in, the user increases a committed counter for that location. This committed counter is hidden from the service provider. However, after a number of check-ins, the user can choose to disclose the counter of one of the locations in order to, e.g., get a discount. The commitment ensures that the user cannot cheat and open the committed counter to a different value.

We note that, in currently deployed LSBS, users can check-in at a certain location without being present at that location. The countermeasure against that is that the wrong location disclosed to the service provider is also disclosed to the user's friends, which can cause annoyance.

Our extension provides the same countermeasure. Using zero-knowledge proofs, the user proves to the service provider that the location shared with her friends equals the location whose committed counter is increased. In order to do that, we make the following modification in our scheme. Instead of employing symmetric key encryption to encrypt the location message, we employ public

key encryption. The key output when decrypting the broadcast encryption ciphertext is an ElGamal private key, while the corresponding public key is transmitted along with the broadcast encryption ciphertext.

In [17], a scheme that employs hardware devices at the physical location in order to ensure that users can check-in only if they are at the location is proposed. It is also possible to extend our scheme with hardware devices to achieve that property.

We note that the total number of locations where a user can check-in is usually large. The user should maintain a committed counter for each of the locations and, at each check-in, increment it without disclosing the location or the value of the counter, yet proving that the location equals the location shared with her friends. If we employ Pedersen commitments, this operation would have a cost linear on the total number of locations, which would make it impractical. In order to have a cost independent of the total number of locations, we employ P-commitments [33], which are based on vector commitments.

6.1 Cryptographic Building Blocks

We recall the notation for zero-knowledge proofs and the definitions of P-commitments in [33].

Zero-Knowledge Proofs of Knowledge

We employ of classical results for efficient zero-knowledge proofs of knowledge (ZKPK) for discrete logarithm relations [11, 12, 16, 18, 19, 45]. In the notation of [14], a protocol proving knowledge of exponents $w_1, \ldots, w_n$ satisfying the formula $\phi(w_1, \ldots, w_n)$ is described as

$$\text{ℵ}\, w_1, \ldots, w_n : \phi(w_1, \ldots, w_n) \tag{1}$$

Here, we use the symbol "ℵ" instead of "$\exists$" to indicate that we are proving "knowledge" of the exponents, rather than just its existence. The formula $\phi(w_1, \ldots, w_n)$ consists of conjunctions and disjunctions of "atoms". An atom expresses *group relations*, such as $\prod_{j=1}^{k} g_j^{\mathcal{F}_j} = 1$, where the g_j are elements of prime order groups and the $\mathcal{F}_j$'s are polynomials in the variables $w_1, \ldots, w_n$.

There exist practical zero-knowledge proofs for the types of relations required in our protocols. We refer to Camenisch et al. [13, 15] for details.

Extended zero-knowledge formulas A proof system for (1) can be transformed into a proof system for the following more expressive statements about secret exponents $(w_i)_i = \mathit{sexps}$ and secret bases $(g_i)_i = \mathit{sbases}$:

$$\nexists\, \mathit{sexps}, \mathit{sbases} : \phi(\mathit{sexps}, \mathit{bases} \cup \mathit{sbases}) \tag{2}$$

The transformation uses a blinded base $g'_i = g_i h^{\rho_i}$ for every g_i. It adds h and all g'_i to the public *bases*, ρ_i to the secret *sexps*, and rewrites $g_i^{\mathcal{F}_j}$ into $g_i'^{\mathcal{F}_j} h^{-\mathcal{F}_j \rho_i}$ for all i, j. Finally, we observe that the proof system supports pairing product equations $\prod_{j=1}^{k} e(g_j, \tilde{g}_j)^{\mathcal{F}_j} = 1$ in groups of prime order $|\mathbb{G}|$ with a bilinear map $e : \mathbb{G} \times \tilde{\mathbb{G}} \to \mathbb{G}_t$, by treating the target group as the group of the proof system—we focus on the special case of $i = j$ for simplicity: the embedding for secret bases is unchanged, except for the case in which both bases in a pairing are secret. In the latter case, $e(g_j, \tilde{g}_j)^{\mathcal{F}_j}$ needs to be transformed into $e(g'_j, \tilde{g}'_j)^{\mathcal{F}_j} e(g'_j, \tilde{h})^{-\mathcal{F}_j \rho_j} e(h, \tilde{g}'_j)^{-\mathcal{F}_j \tilde{\rho}_j} e(h, \tilde{h})^{-\mathcal{F}_j \rho_j \tilde{\rho}_j}$.

Macro notation for zero-knowledge statements We use a macro language to specify named proof components that can be reused as sub-components of larger proofs. For example, we may define a proof macro for the long division of a by q as follows: $\mathbf{Div}(a, q) \mapsto (b) \equiv \nexists\, a, q, b, r : a = b * q + r \wedge r < b \wedge 0 \leq a, b, q, r \leq \sqrt{|\mathbb{G}|} \wedge b > 1$. Semantically, the function **Div** states that the division of a by q gives b with remainder r. Secret a is interpreted as an initial value and secret b as a new value. In terms of cryptography, it is simply syntactic sugar, but important sugar as demonstrated by the long list of side conditions to guarantee a unique positive solution modulo the order of $\mathbb{G}$. Proving these inequalities is itself non-trivial and could be further expanded using macros.

Named proof components can be used in further higher-level proofs without their details being made explicit. For example, the proof $\nexists \ldots, a, q, b : \cdots \wedge\ b = \mathbf{Div}(a, q)$ can from now on be used in proof expressions instead of the complex restrictions above. All variables within the component declaration (e.g., variables a, q, b in $\mathbf{Div}(a, q) \mapsto (b)$) can be re-used in the high level proof. Variables whose knowledge is proved, but that are not in the declaration, are considered inaccessible to the higher-level proof.

P-commitments

A vector commitment scheme allows Alice to succinctly commit to a vector $\mathbf{x} = \langle x_1, \ldots, x_n \rangle \in \mathcal{M}^n$ such that she can compute an opening w to x_i and can replace x_i by a new value x'_i by updating her commitment, such that both w and the update value is of size independent of i and n. A vector commitment scheme consists of the following algorithms.

$\mathsf{Setup}(1^k, \ell)$**.** On input the security parameter 1^k and an upper bound ℓ on the size of the vector, generate the parameters of the commitment scheme par, which include a description of the message space $\mathcal{M}$ and a description of the randomness space $\mathcal{R}$.

$\mathsf{Commit}(par, \mathbf{x}, r)$**.** On input a vector $\mathbf{x} \in \mathcal{M}^n$ $(n \leq \ell)$ and $r \in \mathcal{R}$, output a commitment com to $\mathbf{x}$.

$\mathsf{Prove}(par, i, \mathbf{x}, r)$**.** Compute a witness w for x_i.

$\mathsf{Verify}(par, com, x, i, w)$**.** Output accept if w is a valid witness for x being at position i and reject otherwise.

$\mathsf{Update}(par, com, i, x, r, x', r')$**.** On input a commitment com with value x at position i and randomness r, output a commitment com' with value x' at position i and randomness r'. The other positions remain unchanged.

A commitment scheme must be hiding and binding. The hiding property requires that any probabilistic polynomial time (p.p.t.) adversary $\mathcal{A}$ has negligible advantage in guessing which of two vectors $\mathbf{x}$ of values of its choice has been used to compute a challenge commitment. The binding property requires that any p.p.t. adversary $\mathcal{A}$ has negligible advantage in computing a vector $\mathbf{x}$ of length n, randomness r, a value x, a position $i \in [1..n]$ and a witness w such that $\mathbf{x}[i] \neq x$ but the commitment $com \leftarrow \mathsf{Commit}(par, \mathbf{x}, r)$ can be opened to x at position i using w.

A P-commitment scheme is a secure vector commitment scheme that supports three ZKPKs.

Create. A proof of correct commitment generation that proves knowledge of $(\mathbf{x}, r)$ such that $\mathsf{Commit}(par, \mathbf{x}, r) = com$. We call the proof macro $\mathbf{Create}(\mathbf{x}) \rightarrow (com, r, (w_i)_i)$ as it proves that a P-commitment was correctly initialized to the vector $\mathbf{x}$. The prover can then use this commitment in subsequent proof steps. To simplify our macro notation, we use $M = (com, r, (w_i)_i)$ as a shorthand for the collection of com, r, and different w_i values and refer to it as the memory of a P-commitment proof.

Get. A proof of a witness w that a value x was committed to in com at position i.

$$\mathbf{Get}(M, i) \rightarrow (x) \equiv$$

$$\nexists\, x, i, w :$$

$$\mathsf{Verify}(par, com, x, i, w) = \mathsf{accept} \ \wedge i \in [1, n]$$

Set. A proof that a commitment com' is an update of commitment com at position i. This proof is slightly more involved because it requires the prover to prove knowledge of the old vector value for the updated position to bind the old and the new commitment together:

$$\begin{aligned}&\mathbf{Set}(M, i, x') \to (M') \equiv \\ &\quad \nvDash \mathbf{x}[i], x', i, r, r', w : \\ &\quad com' = \mathsf{Update}(par, com, i, \mathbf{x}[i], r, x', r') \wedge \\ &\quad \mathsf{Verify}(par, com, \mathbf{x}[i], i, w) = \mathsf{accept} \wedge i \in [1, n]\end{aligned}$$

6.2 Construction

As mentioned above, in this extension the location message m is encrypted using an ElGamal encryption $c = (c_1, c_2) = (y^r \cdot m, g^r)$, where $y = g^x$ is the public key and x is the secret key. The secret key is encrypted in the broadcast encryption ciphertext, while the public key is transmitted along with the broadcast encryption ciphertext. We employ a zero-knowledge proof of knowledge of m encrypted to in c:

$$\nvDash m, t : e(c_1, g) = e(m, g) \cdot e(t, g) \wedge e(t, g) = e(y, c_2)$$

In our scheme, the indices $(i_1, \ldots, i_n)$ of the committed vector will be the locations, and n is the total number of locations. We note that the schemes proposed in [33] to implement P-commitments employ a structure preserving signature (SPS) scheme to sign together an index i with the generator g_i for position i in the parameters of the commitment scheme. SPS sign group elements, and therefore we can prove in zero-knowledge equality between the location message m encrypted in c and the index i of the P-commitment. Sender and receivers employ a table or a hash function to map a location to an element of group $\mathbb{G}$.

In the registration phase, the service provider executes $\mathsf{Setup}(1^k, \ell)$, where ℓ is the number of locations, and sends par to the users. Then, each user creates a vector $\mathbf{x} = (0, \ldots, 0)$ of size ℓ, picks $r \in \mathcal{R}$ and runs $\mathsf{Commit}(par, \mathbf{x}, r)$ to get com. The user sends com to the service provider, along with a proof

$$\nvDash \mathbf{x} : \mathbf{Create}(\mathbf{x}) \to (M) \wedge \mathbf{x} = (0, \ldots, 0)$$

where $M = (com, r, (w_i)_i)$. This proof initializes the counters for each of the locations to 0 and can be done very efficiently in the P-commitment schemes in [33].

In the main phase, when a user sends a broadcast ciphertext to the service provider for location i encrypted in ciphertext (c_1, c_2), the user sets $\mathbf{x}[i]' = \mathbf{x}[i] + 1$, picks random $r' \in \mathcal{R}$ and runs $\mathsf{Update}(par, com, i, \mathbf{x}[i], r, \mathbf{x}[i]', r')$ to get com'. The user sends com' to the service provider, along with a proof

$$\not\exists\, i, t, \mathbf{x}[i], \mathbf{x}[i]' :$$

$$e(c_1, g) = e(i, g) \cdot e(t, g) \wedge e(t, g) = e(y, c_2) \wedge$$

$$\mathbf{Set}(M, i, \mathbf{x}[i]') \rightarrow (M') \wedge \mathbf{x}[i]' = \mathbf{x}[i] + 1$$

The user proves that she increments the committed counter for the same location in the message encrypted in c. We recall that the cost of this proof is independent of the number of locations. When the service provider receives the broadcast encryption ciphertext, the ElGamal ciphertex and public key, the commitment com' and the proof, the provider verifies the proof. If it is correct, then the provider replaces the stored com by com' and sends the broadcast encryption ciphertext and the ElGamal ciphertext and public key to the receivers. The receivers decrypt first the broadcast encryption ciphertex to get the ElGamal secret key, which is used to decrypt the ElGamal ciphertext and get the sender's location.

When a user wishes to open the counter corresponding to one of the locations, she can use algorithm $\mathsf{Prove}(par, i, \mathbf{x}, r)$ to compute a witness w for location i and send $(\mathbf{x}[i], i, w)$ to the provider. The service provider runs $\mathsf{Verify}(par, com, \mathbf{x}[i], i, w)$ and accepts the value of the counter $\mathbf{x}[i]$ if the verification is successful. Alternatively, the user can also prove statements about the committed counter in zero-knowledge, e.g., prove that the counter has surpassed a threshold that entitles her to receive a discount. The proof **Get** is employed for this purpose.

The security of this extension relies on the security of P-commitments. The hiding property, along with the zero-knowledge property of proofs of knowledge, ensures that the service provider does not learn the values of the committed counters or the locations whose counters are increased. Additionally, the binding property of P-commitments and the extractability of proofs of knowledge ensure that the committed counters are updated correctly and that they cannot be opened to a wrong value.

7 Discussion

The computation overhead of the SPLS is mainly due to the creation of a symmetric key. An actual location sharing operation is then encrypted using a fast encryption scheme, such as AES. While we note that the symmetric

key of the SPLS can be reused for many location sharing operations, we argue that even computing several symmetric keys per day is feasible. Firstly, as our evaluation shows computing symmetric keys consumes very little energy and can thus be done several times without draining the battery. Secondly, since modern smartphones have multi-core processors embedded, one core can be occupied for creating a symmetric key while the phone is still usable for other operations, such as email checking or surfing the web. All in all we thus argue that on the user side our scheme imposes negligible overhead to the user's device.

The FPLS on the other hand imposes a significant bandwidth overhead. However, we note that it is to the best of our knowledge the only scheme that allows a user to hide her friends without relying on proxies, such as in [44]. It does so by offering a privacy/performance trade-off, which has been proposed before in schemes for privacy-preserving LSBS [44]. Note that the FPLS is not vulnerable to velocity-based attacks [29] for two reasons. Firstly location updates happen sporadically and not continuously and hence big regions are much harder to correlate. Secondly, and more importantly, the big regions are much bigger than in k-anonymity schemes, such as [31].

We note that our schemes are also suitable to implement other services, such as social recommendation applications. This is because in practice users can share arbitrary information in the ciphertext. Instead of encrypting location information, users could share their reviews, such as how they like the food in a particular restaurant. Furthermore, the low overhead of the SPLS and the strategy of reusing symmetric keys would allow to use the scheme for location tracking applications. Such applications require rather frequent location updates instead of sporadic ones, which is usually the case for check-in applications. Furthermore, we note that our schemes are more efficient than a unicast solution in which every user sends an encrypted location update to each of her friends. This is because in our schemes the user only needs to transmit the encrypted location update to the service provider that is then forwarding it to the recipients or all users of the service, respectively. This consumes significantly less bandwidth and also less energy than in the unicast solution.

Although in the setup routine of the SPLS, as well as the FPLS, the service provider initially needs to commit to a maximal number of users n, we note that even if there are more than n users in the service, the service provider does not need to re-initialize the service. In the SPLS, n is only used for checking that $l \leq n$. This condition, however, is maintained if it was true before and n increases. In the FPLS, the scheme's public key can be extended, because the g^{a_i} with $i > n$ can be computed when necessary and distributed among all users of the service.

Besides location-sharing, badge and mayorship protocols are another main functionality of a GSN. For the latter privacy preserving protocols have been proposed in [17]. We note that it would be possible to combine both approaches to build a privacy-preserving GSN.

8 Conclusions

We have defined the privacy properties that an LSBS should provide and we have proposed two LSBS based on identity-based broadcast encryption. Both constructions allow a user to share her location with her friends without disclosing it to the service provider. The first construction discloses to the service provider the receivers of a location update, while the second does not. As advantages from previous work, in our schemes the LSBS provider does not need to perform complex operations in order to compute a reply for a location data request, but only needs to forward IBBE ciphertexts to the receivers. This allows to run a privacy-preserving LSBS at significantly lower costs. We implement both constructions and present a performance analysis that shows their practicality. Furthermore, we extend our schemes such that the service provider, performing some verification work, is able to collect privacy-preserving statistics about the places users share among each other. This could be a way to monetize the privacy-preserving LSBS.

Acknowledgments We thank the anonymous reviewers for their valuable comments. We also thank Carmela Troncoso for insightful discussions and helpful feedback. This research was supported in part by the projects: IWT SBO SPION, FWO G.0360.11N, FWO G.0686.11N, GOA TENSE (GOA/11/007), iMinds SoLoMIDEM, and KU Leuven OT project Traffic Analysis Resistant Privacy Enhancing Technologies.

References

[1] Outsider-anonymous broadcast encryption with sublinear ciphertexts.

[2] Michel Abdalla, Mihir Bellare, Dario Catalano, Eike Kiltz, Tadayoshi Kohno, Tanja Lange, John Malone-Lee, Gregory Neven, Pascal Paillier, and Haixia Shi. Searchable Encryption Revisited: Consistency Properties, Relation to Anonymous IBE, and Extensions. In *Advances in Cryptology–CRYPTO 2005*, pages 205–222. Springer, 2005.

[3] Niranjan Balasubramanian, Aruna Balasubramanian, and Arun Venkataramani. Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications. In *Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 280–293, 2009.

[4] Adam Barth, Dan Boneh, and Brent Waters. Privacy in Encrypted Content Distribution Using Private Broadcast Encryption. volume 4107 of *Lecture Notes in Computer Science*, pages 52–64. Springer, 2006.

[5] A.R. Beresford and F. Stajano. Mix Zones: User Privacy in Location-Aware Services. In *Proceedings of the 2nd Annual Conference on Pervasive Computing and Communications Workshops*, pages 127–131. IEEE, 2004.

[6] Igor Bilogrevic, Murtuza Jadliwala, Kübra Kalkan, Jean-Pierre Hubaux, and Imad Aad. Privacy in Mobile Computing for Location-Sharing-Based Services. In *Privacy Enhancing Technologies - 11th International Symposium (PETS)*, pages 77–96, 2011.

[7] Dan Boneh and Xavier Boyen. Short Signatures Without Random Oracles. In *Advances in Cryptology-EUROCRYPT 2004*, pages 56–73. Springer, 2004.

[8] Dan Boneh and Matt Franklin. Identity-based encryption from the weil pairing. In *Advances in Cryptology-CRYPTO 2001*, pages 213–229. Springer, 2001.

[9] Dan Boneh, Craig Gentry, and Brent Waters. Collusion Resistant Broadcast Encryption With Short Ciphertexts and Private Keys. In *Advances in Cryptology–CRYPTO 2005*, pages 258–275. Springer, 2005.

[10] Xavier Boyen and Brent Waters. Anonymous Hierarchical Identity-based Encryption (Without Random Oracles). In *Advances in Cryptology-CRYPTO 2006*, pages 290–307. Springer, 2006.

[11] Stefan Brands. Rapid Demonstration of Linear Relations Connected by Boolean Operators. In Walter Fumy, editor, *Advances in Cryptology — EUROCRYPT '97*, volume 1233 of *LNCS*, pages 318–333. Springer Verlag, 1997.

[12] Jan Camenisch. *Group Signature Schemes and Payment Systems Based on the Discrete Logarithm Problem.* PhD thesis, ETH Zürich, 1998.

[13] Jan Camenisch, Nathalie Casati, Thomas Groß, and Victor Shoup. Credential authenticated identification and key exchange. In *CRYPTO*, pages 255–276, 2010.

[14] Jan Camenisch, Stephan Krenn, and Victor Shoup. A Framework for Practical Universally Composable Zero-Knowledge Protocols. In *ASIACRYPT*, pages 449–467, 2011.

[15] Jan Camenisch, Stephan Krenn, and Victor Shoup. A Framework for Practical Universally Composable Zero-Knowledge Protocols. Cryptology ePrint Archive, Report 2011/228, 2011. `http://eprint.iacr.org/`.

[16] Jan Camenisch and Markus Michels. Proving in Zero-Knowledge That a Number n is the Product of Two Safe Primes. In Jacques Stern, editor, *Advances in Cryptology — EUROCRYPT '99*, volume 1592 of *LNCS*, pages 107–122. Springer Verlag, 1999.

[17] Bogdan Carbunar, Radu Sion, Rahul Potharaju, and Moussa Ehsan. The Shy Mayor: Private Badges in GeoSocial Networks. In *10th International Conference on Applied Cryptography and Network Security (ACNS)*, volume 7341 of *Lecture Notes in Computer Science*, pages 436–454. Springer Berlin Heidelberg, 2012.

[18] D. Chaum and T. Pedersen. Wallet databases with observers. In *CRYPTO '92*, volume 740 of *LNCS*, pages 89–105, 1993.

[19] R. Cramer, I. Damgård, and B. Schoenmakers. Proofs of Partial Knowledge and Simplified Design of Witness Hiding Protocols. In *CRYPTO*, pages 174–187, 1994.

[20] Joan Daemen and Vincent Rijmen. *The Design of Rijndael: AES-The Advanced Encryption Standard*. Springer, 2002.

[21] Cécile Delerablée. Identity-based Broadcast Encryption With Constant Size Ciphertexts and Private Keys. *Advances in Cryptology–ASIACRYPT 2007*, pages 200–215, 2007.

[22] Yevgeniy Dodis and Nelly Fazio. Public Key Trace and Revoke Scheme Secure Against Adaptive Chosen Ciphertext Attack. *Public Key Cryptography—PKC 2003*, pages 100–115, 2002.

[23] Changyu Dong and Naranker Dulay. Longitude: A Privacy-Preserving Location Sharing Protocol for Mobile Applications. In *5th IFIP WG 11.11 International Conference on Trust Management (IFIPTM)*, pages 133–148. Springer Berlin Heidelberg, 2011.

[24] Amos Fiat and Moni Naor. Broadcast Encryption. In *Advances in Cryptology—Crypto'93*, pages 480–491. Springer, 1994.

[25] Julien Freudiger, Raoul Neu, and Jean-Pierre Hubaux. Private Sharing of User Location Over Online Social Networks. In *HotPETs*, pages 1–12, 2010.

[26] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. Evaluating the Privacy Risk of Location-Based Services. In *15th International Conference on Financial Cryptography and Data Security (FC)*, volume 7035 of *Lecture Notes in Computer Science*, pages 31–46. Springer Berlin Heidelberg, 2012.

[27] Sébastien Gambs, Olivier Heen, and Christophe Potin. A Comparative Privacy Analysis of Geosocial Networks. In *Proceedings of the 4th ACM International Workshop on Security and Privacy in GIS and LBS (SPRINGL)*, pages 33–40. ACM, 2011.

[28] Craig Gentry and Brent Waters. Adaptive Security in Broadcast Encryption Systems (With Short Ciphertexts). *Advances in Cryptology-EUROCRYPT 2009*, pages 171–188, 2009.

[29] Gabriel Ghinita, Maria Luisa Damiani, Claudio Silvestri, and Elisa Bertino. Preventing Velocity-based Linkage Attacks in Location-aware Applications. In *17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems (ACM-GIS)*, pages 246–255, 2009.

[30] Gabriel Ghinita, Panos Kalnis, Ali Khoshgozaran, Cyrus Shahabi, and Kian-Lee Tan. Private queries in location based services: anonymizers are not necessary. SIGMOD '08, pages 121–132, New York, NY, USA, 2008. ACM.

[31] Marco Gruteser and Dirk Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st International Conference on Mobile systems, Applications and Services (MobiSys)*, pages 31–42. ACM, 2003.

[32] Leping Huang, Hiroshi Yamane, Kanta Matsuura, and Kaoru Sezaki. Towards Modeling Wireless Location Privacy. In *5th International Workshop on Privacy Enhancing Technologies (PET)*, pages 59–77. Springer Berlin Heidelberg, 2006.

[33] Markulf Kohlweiss and Alfredo Rial. Optimally Private Access Control. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 37–48. ACM, 2013.

[34] John Krumm. Inference Attacks on Location Tracks. In *5th International Conference on Pervasive Computing*, volume 4480 of *Lecture Notes in Computer Science*, pages 127–143. Springer Berlin Heidelberg, 2007.

[35] John Krumm. Realistic Driving Trips For Location Privacy. In *7th International Conference on Pervasive Computing*, volume 5538 of *Lecture Notes in Computer Science*, pages 25–41. Springer Berlin Heidelberg, 2009.

[36] Allison Lewko, Amit Sahai, and Brent Waters. Revocation Systems With Very Small Private Keys. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 273–285. IEEE, 2010.

[37] Mingyan Li, Krishna Sampigethaya, Leping Huang, and Radha Poovendran. Swing & Swap: User-centric Approaches Towards Maximizing Location Privacy. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society (WPES)*, pages 19–28. ACM, 2006.

[38] Benoît Libert, Kenneth Paterson, and Elizabeth Quaglia. Anonymous Broadcast Encryption: Adaptive Security and Efficient Constructions in the Standard Model. *Public Key Cryptography–PKC 2012*, pages 206–224, 2012.

[39] Zi Lin, Denis Foo Kune, and Nicholas Hopper. Efficient Private Proximity Testing with GSM Location Sketches. In *16th International Conference on Financial Cryptography and Data Security (FC)*, volume 7397 of *Lecture Notes in Computer Science*, pages 73–88. Springer Berlin Heidelberg, 2012.

[40] Dalit Naor, Moni Naor, and Jeff Lotspiech. Revocation and Tracing Schemes for Stateless Receivers. In *Advances in Cryptology-CRYPTO 2001*, pages 41–62. Springer, 2001.

[41] Arvind Narayanan, Narendran Thiagarajan, Mugdha Lakhani, Michael Hamburg, and Dan Boneh. Location Privacy via Private Proximity Testing. In *Proceedings of the Network & Distributed System Security Symposium (NDSS)*, pages 1–17. Internet Society, 2011.

[42] Femi Olumofin, Piotr Tysowski, Ian Goldberg, and Urs Hengartner. Achieving Efficient Query Privacy for Location Based Services. In *Privacy Enhancing Technologies*, pages 93–110. Springer, 2010.

[43] Tatiana Pontes, Marisa Vasconcelos, Jussara Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. We Know Where You Live: Privacy Characterization of Foursquare Behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp)*, pages 898–905. ACM, 2012.

[44] Krishna PN Puttaswamy, Shiyuan Wang, Troy Steinbauer, Deepak Agrawal, Amr El Abbadi, Christopher Kruegel, and Ben Y Zhao. Preserving Location Privacy in Geosocial Applications. *IEEE Transactions on Mobile Computing*, 13(1):159–173, 2014.

[45] C. Schnorr. Efficient signature generation for smart cards. *Journal of Cryptology*, 4(3):239–252, 1991.

[46] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying Location Privacy. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy (S&P)*, pages 247–262. IEEE, 2011.

[47] C.R. Vicente, D. Freni, C. Bettini, and Christian S. Jensen. Location-Related Privacy in Geo-Social Networks. *IEEE Internet Computing*, 15(3):20–27, 2011.

[48] Brent Waters. Dual System Encryption: Realizing Fully Secure IBE and HIBE Under Simple Assumptions. *Advances in Cryptology-CRYPTO 2009*, pages 619–636, 2009.

[49] Ge Zhong, Ian Goldberg, and Urs Hengartner. Louis, Lester and Pierre: Three Protocols for Location Privacy. In *7th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 4776 of *Lecture Notes in Computer Science*, pages 62–76. Springer Berlin Heidelberg, 2007.

Publication

Optimal Sporadic Location Privacy Preserving Systems in Presence of Bandwidth Constraints

Publication Data

Michael Herrmann, Carmela Troncoso, Claudia Diaz, and Bart Preneel. Optimal Sporadic Location Privacy Preserving Systems in Presence of Bandwidth Constraints. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 167–178. ACM, 2013.

Contributions

- Main author for all sections.

Optimal Sporadic Location Privacy Preserving Systems in Presence of Bandwidth Constraints

Michael Herrmann[1], Carmela Troncoso[2], Claudia Diaz[1], and Bart Preneel[1]

[1] KU Leuven ESAT/COSIC, iMinds, Leuven, Belgium
{name.surname}@esat.kuleuven.be
[2] Gradiant, Vigo, Spain
ctroncoso@gradiant.org

Abstract. Various Location Privacy-Preserving Mechanisms (LPPMs) have been proposed in the literature to address the privacy risks derived from the exposure of user locations through the use of Location Based Services (LBSs). LPPMs obfuscate the locations disclosed to the LBS provider using a variety of strategies, which come at a cost either in terms of quality of service, or of resource consumption, or both. Shokri *et al.* propose an LPPM design framework that outputs optimal LPPM parameters considering a strategic adversary that knows the algorithm implemented by the LPPM, and has prior knowledge on the users' mobility profiles [23]. The framework allows users to set a constraint on the tolerable loss quality of service due to perturbations in the locations exposed by the LPPM. We observe that this constraint does not capture the fact that some LPPMs rely on techniques that augment the level of privacy by increasing resource consumption.

In this work we extend Shokri *et al.*'s framework to account for constraints on bandwidth consumption. This allows us to evaluate and compare LPPMs that generate dummies queries or that decrease the precision of the disclosed locations. We study the trilateral trade-off between privacy, quality of service, and bandwidth, using real mobility data. Our results show that dummy-based LPPMs offer the best protection for a given combination of quality and bandwidth constraints, and that, as soon as communication overhead is permitted, both dummy-based and precision-based LPPMs outperform LPPMs that only perturb the exposed locations. We also observe that the maximum value of privacy a user can enjoy can be reached by either sufficiently relaxing the quality loss or the bandwidth constraints, or by choosing an adequate combination of both constraints. Our results

contribute to a better understanding of the effectiveness of location privacy protection strategies, and to the design of LPPMs with constrained resource consumption.

1 Introduction

Location Based Services (LBSs) enable users to, among others, let their friends know where they are, find nearby points of interest, or obtain contextual information about their surroundings. The typical LBS implementation is such that user locations are by default disclosed to the LBS provider. This raises privacy concerns, as location information is known to reveal potentially sensitive private information (e.g., visiting the mosque, church, or temple reveals religious beliefs). A variety of Location Privacy-Preserving Mechanisms (LPPMs), e.g., [5, 7, 17], have been proposed in prior research to mitigate these concerns. To do so, these mechanisms obfuscate user locations before sending them to the LBS provider.

The great majority of LPPMs in the literature are designed considering a non-strategic adversary. This assumes that the adversary is unaware of the LPPM obfuscation algorithm, and that he has no prior knowledge on the users' mobility profiles. However, both the LPPM's internal algorithm and the user mobility patterns leak information that can be exploited by the adversary to reduce her estimation error when locating users [21]. Hence, designs and evaluations that neglect such information overestimate the level of privacy protection offered by the LPPM.

Shokri *et al.* [23] proposed a framework to design LPPM with optimal parameters considering an adversary that has (and exploits) information on: i) the LPPM algorithm implemented; and ii) the mobility profile of the user. This framework facilitates the design of LPPMs that maximize the location estimation error of strategic adversaries. Furthermore, the framework allows users to establish a maximum tolerated quality of service loss stemming from the use of the LPPM. The framework is suitable to model LBSs in which users only reveal their location *sporadically*, i.e. subsequent location exposures of the same user are assumed to be sufficiently apart in time that it is not possible to link them as related to the same individual. Examples of applications in which location revelations are sporadic include check-in services [1], or services for finding nearby points of interest [2].

The problem statement in Shokri's framework [23] does not consider constraints on resources utilization (e.g., bandwidth, battery consumption). These are however likely to be a concern for users in reality, since LBSs are mostly accessed

from resource-constrained mobile devices. Our first contribution is to extend the framework to account for resource limitations.

Prior research has only applied the framework to the design of perturbation-based mechanisms, i.e., LPPMs that modify the location that is disclosed to the LBS provider. As second contribution, we model two other popular privacy-preserving strategies in the context of the framework. Both types of mechanisms increase the adversary's uncertainty on the user's actual position by raising the number of locations from which the user could have issued a query. In dummy-based mechanisms [14, 16, 26] the LPPM sends fake locations to the LBS server along with the actual user requests. In precision-based mechanisms [9, 11, 25] the LLPM decreases the precision of the disclosed location sent to the LBS provider, so that there is a bigger geographical region in which they user might be located.

Contrary to the perturbation-based LPPMs considered by Shokri *et al.* [23], dummy-based and precision-based LPPMs may consume more resources (e.g., bandwidth and battery) in order to conceal the user's location. Our third contribution is a study of the trilateral trade-off between quality of service, bandwidth consumption, and privacy using these LPPMs as case study. We find that for the considered LPPMs both quality loss and bandwidth constraints can be traded for privacy. In fact, the maximum achievable level of privacy can be reached either when the quality loss constraint is sufficiently loose (as in [23]), when sufficient bandwidth is allowed, or when an adequate combination of both is allowed. Our simulations show that, for given bandwidth and quality constraints, dummy-based LPPMs offer better protection than precision-based LPPMs. This is because dummy-based LPPMs have more degrees of freedom than precision-based LPPMs in choosing the cover locations, and hence can better exploit the available resources. Finally, both dummy-based and precision-based offer a better privacy level than just perturbation for the same quality of service, provided that the system can tolerate the introduction of a communication overhead.

The rest of this paper is organized as follows: the next section gives an overview of the state of the art in location privacy-preserving systems design. Section 3 describes the system and adversarial models, as well as the constraints imposed on the design; and Section 4 revisits Shokri *et al.*'s framework. We describe the linear programs to compute different classes of resource-consuming LPPMs in Section 5, and validate them against real data in Section 6. Finally, we conclude in Section 7.

2 Related Work

It is widely accepted that the disclosure of location data entails a privacy risk: Hoh et al. show that car driving traces enable the inference of the drivers' home addresses [13]; this information by itself, or together with the driver's work place, can be used to re-identify anonymous traces [10, 15]. Furthermore, Freudiger *et al.* point out that a people's mobility patterns are persistent and unique [8]. Therefore, users are identifiable by the LBS even if they only share their location during a short period of time. Once location data is identifiable, it may reveal a detailed picture of the person's habits, lifestyle, and preferences [3]. To counter this threat various obfuscation-based Location Privacy Preserving Mechanisms (LPPMs) been proposed in the literature. These mechanisms obfuscate the revealed locations and thus prevent (or at least limit) the possible inferences that could be made from the data.

Following the categorization proposed by Shokri *et al.* [21] we briefly introduce the existing obfuscation strategies and refer the reader to [20] for a more detailed review. *Pertur-bation-based* LPPMs [12, 17] modify a user's reported location such that at least two users might be associated to a location. *Pseudonymization-based* LPPMs regularly change the identity with which users identify themselves to the LBS provider, in order to prevent the linkage of two subsequent user locations, thus preventing the adversary from reconstructing the trajectories followed by the users of the system. These LPPMs can be combined with *hiding-based* LPPMs, which allow users to sometimes hide their location [6], further decreasing the adversary's capability to link location exposures. *Precision-based* LPPMs [4, 9, 11, 25] reduce the granularity of the location data revealed to the provider, so that it is not possible to pinpoint the exact location of a user within a geographical region. Finally, *dummy-based* LPPMs [14, 16, 26] automatically generate queries with fake position data that are indistinguishable from the users' real queries. Here the adversary is unable to determine whether the location associated with a query corresponds to the user's actual position, or is a decoy.

Shokri *et al.* have proposed methods to quantify and systematically evaluate the level of privacy provided by obfuscation-based LPPMs [21, 22]. They formalize the obfuscation process performed by the LPPM, as well as the attack strategies that an adversary can use to invert the location transformations made by the LPPM. They measure privacy as the expected error of a strategic adversary when estimating the actual location of a user. This quantitative approach is a cornerstone of their LPPM design framework, where they propose a systematic method to design LPPMs that are optimal with respect to strategic adversaries, who are aware of the LPPM's internal operation and the users' mobility profiles [23].

This framework allows users to indicate the maximum quality loss (derived from the use of the LPPM) that they are willing to tolerate. The design framework then outputs a set of parameters for the LPPM that maximize the error of the adversary when attempting to locate users. Our work builds on this framework and extends it to account for not only quality loss, but also for limitations on bandwidth consumption.

Finally, we note that there are other approaches to building location privacy systems that are not based on obfuscation strategies and are thus out of the scope of this paper. This includes cryptographic approaches such as those based on Private Information Retrieval protocols [18].

3 System Model

In this paper, we extend the framework by Shokri *et al.* [23] to account for bandwidth constraints in Location Privacy Preserving Mechanisms (LPPMs). Therefore, we follow the framework's system model and definitions and augment them when needed to account for bandwidth constraints. The focus of the framework is on user-centric mechanisms, in which the configuration of the LPPM is decided on independently by each user, without knowledge about other users in the system. Thus, without loss of generality, we restrict our model and analysis to a single user. We note that cloaking mechanisms, in which the geographical region disclosed is chosen taking into account the positioning of a set of users [11], can also be modeled as user-centric mechanisms because their privacy guarantees depend only on the size of the region [24].

User model: Similarly to prior work [23] we consider that the user moves around in a finite geographical area that is divided into M discrete regions $\mathcal{R} = \{r_1, r_2, \ldots, r_M\}$. Users only expose their location $r \in \mathcal{R}$ *sporadically* to an LBS provider in order to obtain a service. A user's LBS usage pattern is described by her *mobility profile* $\psi(r)$, $\sum_r \psi(r) = 1$, a probability distribution describing her likelihood of being at location r when querying the LBS. We make no particular assumption on the users' mobility patterns, i.e., we impose no restrictions on the profiles $\psi(r)$. As usage is sporadic, the locations from which the user accesses the service at different time instants are independent from each other. Therefore, the mobility profiles only describe the frequency with which users' visit locations, and does not contain information about transitions between regions.

Location privacy-preserving mechanism: The user runs in her personal device an LPPM that transforms her real location $r \in \mathcal{R}$ into a pseudo-location $r' \in \mathcal{R}'$. This transformation is made according to a probability distribution

$f(r'|r) = \Pr(r'|r)$. The pseudo-location r' is exposed to the LBS provider instead of her actual location r. Shokri *et al.* [23] consider that $\mathcal{R}' = \mathcal{R}$. In this work we extend $\mathcal{R}'$ to be the powerset of $\mathcal{R}$ except the empty set; i.e., $\mathcal{R}' = \mathcal{P}(\mathcal{R}) - \{\emptyset\}$. Hence, i) pseudo-locations r' may or may not contain the real location r; and ii) differently from prior work [23], in which pseudo-locations r' are formed by one region in $\mathcal{R}$, here r' may be formed by one or more regions r_i in $\mathcal{R}$.

Adversary model: We consider that the user wants to protect her real location towards a passive adversary that has access to the locations exposed to the LBS. This adversary could be the LBS provider, an eavesdropper of the user-provider communication, or other LBS subscribers with which exposed locations are shared. We assume that the adversary knows the users' profiles $\psi(r)$, which can be inferred, for instance, using existing learning techniques [21].

Following prior work [23] we model the adversary's strategy as a probability distribution $h(\hat{r}|r') = \Pr(\hat{r}|r')$. This distribution describes the probability that, given an exposed location r', the estimated location $\hat{r}$ corresponds to the user's real position r. We measure the privacy loss as the adversary's expected error in this estimation $\hat{r}$ given that the real location is r. We model the adversarial error as a function $d_p(\hat{r}, r)$ that depends on both the user's privacy criteria and on the semantics of the location [23]; as well as on the transformation function $f(r'|r)$ implemented by the LPPM. (We provide examples of functions $d_p(\cdot)$ that are adequate for particular LPPMs in Section 5.)

Quality of service: Users expect to obtain relevant information from their queries to the LBS. Because the response of the LBS to a query depends on the observed location r', and not on the real location r, the information contained in the response may be of less utility to the user than that contained in a response to a query in which r is exposed. Given an LPPM $f(\cdot)$, the *expected quality loss* suffered by the user can be computed as:

$$\mathrm{E}[\mathrm{Q}_{\mathrm{loss}}(\psi, f, d_q)] = \sum_{r,r'} \psi(r) f(r|r') d_q(r', r)\,. \tag{1}$$

In this formula $\psi(r)$ represents the prior probability of the user accessing the LBS from location r (i.e., according to her mobility profile); $f(r'|r)$ represents the probability of exposing r' given that the user is at r; and the function $d_q(r', r)$ represents the quality loss resulting from exposing r' instead of r to the LBS provider. (We provide examples of $d_q(\cdot)$ functions adequate for particular LPPMs in Section 5.) In layman words, $\mathrm{E}[\mathrm{Q}_{\mathrm{loss}}(\psi, f, d_q)]$ reflects the average discontent experienced by users when utilizing an LPPM.

We assume that the user imposes a maximum tolerable service quality loss $\mathrm{Q}_{\mathrm{loss}}^{\max}$. The LPPM output must satisfy the constraint $\mathrm{E}[\mathrm{Q}_{\mathrm{loss}}(\psi, f, d_q)] < \mathrm{Q}_{\mathrm{loss}}^{\max}$.

Bandwidth constraints: The fact that Shokri *et al.* consider $\mathcal{R}' = \mathcal{R}$ implies that the LPPM never incurs in communication overhead when sending r' instead of r. Since we have set $\mathcal{R}' = \mathcal{P}(\mathcal{R}) - \{\emptyset\}$, sending r' may require more bandwidth than sending r (e.g., if r' is composed by several regions in $\mathcal{R}$). LBSs are mostly accessed from mobile devices which in general have restricted connectivity and limited resources, and hence users may want to limit the overhead introduced by the LPPM. We extend the existing model [23] to account for this fact by defining the *expected bandwidth overhead* incurred by LPPM $f(\cdot)$ as:

$$\mathrm{B}_{\mathrm{cost}}(\psi, f, d_b) = \sum_{r,r'} \psi(r) f(r|r') d_b(r', r)\,, \tag{2}$$

In this formula $\psi(r)$ and $f(r'|r)$ have the same role as in Eq. (1). The function $d_b(r', r)$ represents the additional cost in terms of bandwidth derived from exposing r' instead of r. (We provide examples of $d_b(\cdot)$ functions adequate for particular LPPMs in Section 5.)

We assume that the user imposes a maximum tolerable bandwidth $\mathrm{B}_{\mathrm{cost}}^{\max}$. As with quality loss constraints, the LPPM must satisfy $\mathrm{B}_{\mathrm{cost}} < \mathrm{B}_{\mathrm{cost}}^{\max}$.

We note that, although we only consider limitations on communication overhead, the function $d_b(\cdot)$ can model other constraints related to resource consumption resulting from exposed pseudo-locations that may be formed by several regions, e.g., the increase in battery consumption needed to send more packets, or to process more responses.

Privacy: The level of privacy enjoyed by users depends on the attack strategy deployed by the adversary. Following the definition by Shokri *et al.* [21, 23] we measure the *expected privacy* of the user as:

$$\mathrm{Privacy}(\psi, f, h, d_p) = \sum_{r,r',\hat{r}} \psi(r) f(r'|r) h(\hat{r}|r') d_p(\hat{r}, r)\,. \tag{3}$$

Each summand in this equation represents the probability that the user obtains a privacy level $d_p(\hat{r}, r)$, when she accesses the LBS from location location r, exposes pseudo-location r', and the adversary estimates $\hat{r}$ given the observation.

Figure 1 illustrates the relationships between the different elements of this model. Note that we consider that the defense (resp., the attack) takes into account the attack (resp., the defense) implemented by the adversary (resp., the user), as well as the user's mobility profile and her constraints in terms of bandwidth and quality of service.

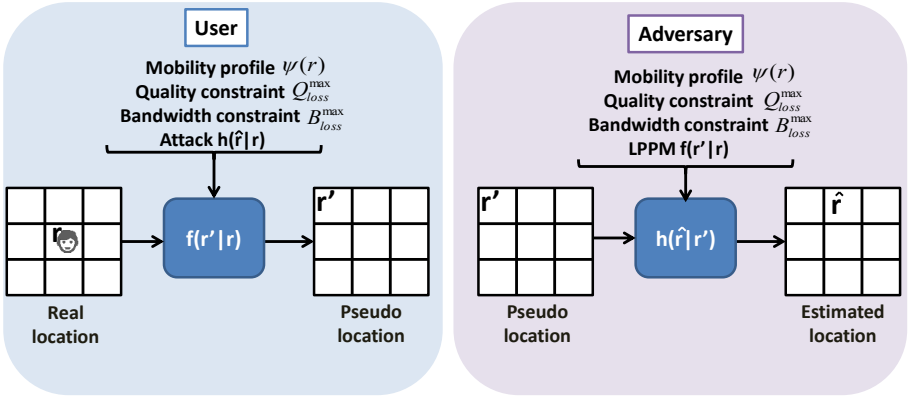


Figure 1: System model

4 Game Theory in Location Privacy

In this section we revisit the design methodology proposed by Shokri *et al.* in prior work [23]. This method allows the user to choose optimal parameters for the LPPM $f(\cdot)$, given an adversary that implements the optimal attack $h(\cdot)$ against this defense. Given a user mobility profile $\psi(r)$ and quality of service constraint $\mathrm{Q}_{\mathrm{loss}}^{\max}$, the method models the design of the optimal LPPM as an instance of a zero-sum Bayesian Stackelberg game.

The Stackelberg competition in the context of location privacy is stated as follows: a *leader* (the user), commits first to an LPPM $f(\cdot)$ that satisfies the quality constraint $\mathrm{Q}_{\mathrm{loss}}^{\max}$. For this purpose the LPPM takes the user's actual location r as input, and outputs a pseudo-location r'. Upon observing the exposed location, a *follower* (the adversary), estimates the real location through the attack $h(\cdot)$, taking into account both the user's profile $\psi(r)$ and the LPPM $f(\cdot)$ chosen by the user. The adversary 'pays' an amount $d_p(\hat{r}, r)$ to the user that represents the estimation error from the adversary's perspective, and the location privacy gain from the user's perspective.

Both players aim at maximizing their payoffs: the adversary tries to minimize the amount to pay (i.e., minimize her estimation error), while the user tries to maximize this amount (i.e., maximize her location privacy). The game is zero-sum, as the adversary's information gain equals the privacy lost by the user, and vice-versa. It is also a Bayesian game since the adversary only has access to probabilistic data about the user's real location; i.e., her information on the user is incomplete.

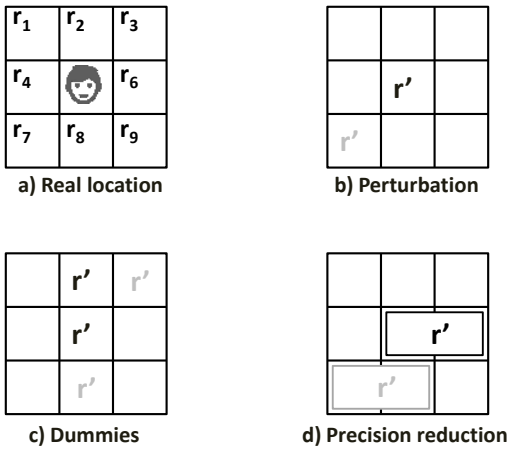


Figure 2: Toy example ($\mathcal{R} = \{r_1, \ldots, r_9\}$): a) Real user location; b) Perturbation-based LPPM $\mathcal{R}' = \mathcal{R}$; c) Dummy-based LPPM; d) Reducing precision-based LPPM.

4.1 Perturbation-based LPPM

Shokri *et al.* validate their framework by applying it to the design of perturbation-based strategies. In this scenario $\mathcal{R}' = \mathcal{R}$, and hence the pseudo-locations r' output by the LPPM are formed by one region $r_i \in \mathcal{R}$, which may or may not be equal to the real location r. Let us consider the toy example in Fig. 2a, in which the area $\mathcal{R}$ is formed by 9 regions, and where the user queries the LBS provider from location r_5. Two possible pseudo-locations r' are shown in Fig. 2b, depicted in black and gray. Note that the black **r'** coincides with the real user location $r = r_5$, while the grey pseudo-location r'$= r_7$ does not.

Solution: We now present the linear programs developed in prior work [23] to compute the optimal perturbation and attack strategies $f(\cdot)$ and $h(\cdot)$. These linear programs compute the theoretic equilibrium of the game described above.

The user runs the following linear program to find the optimal parameters for her perturbation-based LPPM:

$$\textbf{Choose } f(r'|r), x_{r'}, \forall r, r' \text{ that}$$

$$\textbf{maximize } \sum_{r'} x_{r'} \tag{4}$$

subject to

$$x_{r'} \leq \sum_{r} \psi(r) f(r'|r) d_p(\hat{r}, r), \forall \hat{r}, r' \tag{5}$$

$$\sum_{r} \psi(r) \sum_{r'} f(r'|r) d_q(r', r) \leq \mathrm{Q}_{\mathrm{loss}}^{\max} \tag{6}$$

$$\sum_{r'} f(r'|r) = 1, \forall r \tag{7}$$

$$f(r'|r) \geq 0, \forall r, r' \tag{8}$$

The decision variable $f(r'|r)$ represents the LPPM algorithm, while $x_{r'}$ represents the expected privacy of the user (see Appendix A). The inequalities defined by Eq. (5) express the privacy constraint, ensuring that $f(r'|r)$ is chosen to maximize $x_{r'}$; while the inequalities defined by Eq. (6) express the quality constraint, ensuring that the expected quality of service loss is at most $\mathrm{Q}_{\mathrm{loss}}^{\max}$. Finally Eq. (7) and (8) ensure that $f(\cdot)$ is a proper probability distribution.

On the other hand, the adversary runs the following linear program to obtain the optimal attack function $h(\hat{r}|r')$, which minimizes privacy when the user implements a perturbation-based LPPM $f(r'|r)$:

$$\textbf{Choose } h(\hat{r}|r'), y_r, \forall r, r', \hat{r}, \text{and} z_q \in [0, \infty) \text{ that}$$

$$\textbf{minimize } \sum_r \psi(r)\, y_r + z_q \mathrm{Q}_{\mathrm{loss}}^{\max} \tag{9}$$

subject to

$$y_r \geq \sum_{\hat{r}} h(\hat{r}|r') d_p(\hat{r}, r) + z_q d_q(r', r), \forall r, r' \tag{10}$$

$$\sum_{\hat{r}} h(\hat{r}|r') = 1, \forall r' \tag{11}$$

$$h(\hat{r}|r') \geq 0, \forall r', \hat{r} \tag{12}$$

$$z_q \geq 0 \tag{13}$$

The decision variable $h(\hat{r}|r')$ represents the adversary's attack strategy on the LPPM algorithm, and y_r the expected privacy of the user (see Appendix A). The variable z_q acts as *shadow price* for the quality. It expresses the loss (gain) in privacy when the maximum tolerated expected quality loss $\mathrm{Q}_{\mathrm{loss}}^{\max}$ decreases (increases) by one unit. We refer the reader to Shokri's prior work for more details on the meaning of this variable [23]. The inequalities defined by Eq. (10) represent constraints on privacy, ensuring that $h(\hat{r}|r')$ is chosen to minimize privacy given the quality constraints; and Eqs (11) and (12) ensure that $h(\cdot)$ is a proper probability distribution. Finally Eq. (13) ensures that the trade-off between quality and privacy expressed by z_q is non-negative.

Quality, Bandwidth, and privacy constraints: Perturbation-based LPPMs output one-sized regions $r' \in \mathcal{R}' = \mathcal{R}$. This determines the functions used to model the constraints imposed by the user. Since pseudo-locations and real locations have the same size, there is no communication overhead in the model. Therefore, the bandwidth constraint $\mathrm{B}_{\mathrm{cost}}^{\max}$ does not affect the optimization and does not appear in the linear programs.

Furthermore, in this setting both the quality and the privacy constraints can be expressed in terms of the distance between the exposed location r' (resp., the inferred location $\hat{r}$) and the actual user location r [23]. For the sake of simplicity, in our experiments for perturbation-based LPPMs we model both $d_q(r', r)$ and $d_p(\hat{r}, r)$ as the Manhattan distance between the two locations (e.g., $d_p(\hat{r}, r) = \|\hat{r} - r\|_1$).

5 Bandwidth-consuming LPPMs

In this section we model two popular families of Location Privacy-Preserving Mechanisms (LPPMs) in the literature that consume extra bandwidth to increase users privacy: dummy-based LPPMs, and precision-based LPPMs. To model these strategies we extend the game-theoretic approach outlined in the previous section to also account for bandwidth constraints. We describe two linear programs that output the user's optimal LPPM $f(\cdot)$ and the adversary's optimal attack $h(\cdot)$, while respecting the quality and bandwidth constraints.

5.1 Dummy-based LPPM

Dummy-based LPPMs automatically generate dummy queries that are sent to the LBS provider along with the user's real queries [14, 16, 26]). The dummy queries contain fake locations and their goal is to increase the adversary's estimation error on the user's real location, since for the adversary all received locations are equally likely to correspond to the user's actual position.

A dummy-based LPPM $f(r'|r)$ outputs pseudo-locations r' from $\mathcal{R}' = \mathcal{P}(\mathcal{R}') - \{\emptyset\}$ formed by one or more *non-contiguous* regions $r_i \in \mathcal{R}$, which may or may not contain the real location r. In the toy example shown in Fig. 2c we can see two possible outputs r' when the user sends one dummy query formed by two regions. The black pseudo-location **r'**$= \{r_2, r_5\}$ contains the real location $r = r_5$, while the grey pseudolocation r'$= \{r_3, r_8\}$ does not. In the latter case the LPPM no only generates decoy locations, but also perturbs the user's position.

Solution: The linear program to compute the optimal dummy-based LPPM is similar to the perturbation-based case, with one important difference: it includes a set of inequalities to ensure that the expected communication overhead associated to the use of dummies does not exceed the maximum expected

bandwidth consumption $\mathrm{B}_{\mathrm{cost}}^{\max}$:

$$\textbf{Choose } f(r'|r), x_{r'}, \forall r, r' \text{ that}$$

$$\textbf{maximize } \sum_{r'} x_{r'} \tag{14}$$

subject to

$$x_{r'} \leq \sum_{r} \psi(r) f(r'|r) d_p(\hat{r}, r), \ \forall \hat{r}, r' \tag{15}$$

$$\sum_{r} \psi(r) \sum_{r'} f(r'|r) d_q(r', r) \leq \mathrm{Q}_{\mathrm{loss}}^{\max} \tag{16}$$

$$\sum_{r} \psi(r) \sum_{r'} f(r'|r) d_b(r', r) \leq \mathrm{B}_{\mathrm{cost}}^{\max} \tag{17}$$

$$\sum_{r'} f(r'|r) = 1, \ \forall r \tag{18}$$

$$f(r'|r) \geq 0, \ \forall r, r' \tag{19}$$

The inequalities defined by Eqs (15), (16), (18), and (19) have the same role as in the perturbation-based case. Eq. (17) adds the bandwidth constraint, so that the expected bandwidth consumption does not exceed $\mathrm{B}_{\mathrm{cost}}^{\max}$.

From the adversary's point of view, the linear program used to compute the optimal attack $h(\hat{r}|r)$ differs from the perturbation-based case in that we introduce a new shadow price z_1 in Eq. (25). This new decision variable models the relation between privacy and bandwidth in the same manner as z_q models the relation privacy between privacy and quality. We obtain:

Choose $h(\hat{r}|r'), y_r, \forall r, r', \hat{r}, z_q \in [0,\infty), z_b \in [0,\infty)$ to

$$\textbf{minimize} \quad \sum_r \psi(r)\, y_r + z_q \mathrm{Q}_{\mathrm{loss}}^{\max} + z_b \mathrm{B}_{\mathrm{cost}}^{\max} \tag{20}$$

subject to

$$y_r \geq \sum_{\hat{r}} h(\hat{r}|r') d_p(\hat{r}, r) + z_q d_q(r', r) + z_b d_b(r', r), \forall r, r' \tag{21}$$

$$\sum_{\hat{r}} h(\hat{r}|r') = 1, \forall r' \tag{22}$$

$$h(\hat{r}|r') \geq 0, \forall r', \hat{r} \tag{23}$$

$$z_q \geq 0 \tag{24}$$

$$z_b \geq 0 \tag{25}$$

Quality, bandwidth, and privacy constraints: As the dummy-based LPPM transmits dummy locations to the LBS provider, the functions $d_q(r', r)$ and $d_b(r', r)$, which express the constraints on quality and bandwidth, need to take into account that pseudo-locations r' can be composed by several regions.

With respect to the quality of service function $d_q(r', r)$ we distinguish two cases. If the actual location r is among the regions contained in the pseudo-location r', then the quality loss is zero, as the user receives a response corresponding to her real location r. Formally, $d_q(r', r) = 0$, $\forall r' : r \in r'$. If on the other hand the real location is not within the exposed pseudo-location, we assume that the response for the nearest location will provide the most useful response to the user, and thus measure the quality loss as the minimum of the distances between the real location and each of the locations r_i contained in r'. For instance, considering the Manhattan distance, $d_q(r', r) = \min_{r_i \in r'} \|r_i - r\|_1$, $\forall r' : r \notin r'$.

The bandwidth function $d_b(r', r)$ takes into account that the system sends and receives more traffic when dummies are implemented. This extra bandwidth consumption may be due to an increase in the length of the query if all dummies are sent in one request; or to an increase in the number of queries if dummies are sent in separate requests. In this paper we consider that each dummy increases the bandwidth overhead by 2 units: one unit for uploading and one unit for downloading. Formally: $d_b(r', r) = (\sum_{r_i \in r'} 2) - 2$.

As in the perturbation-based case, the privacy function $d_p(\hat{r}, r)$ considers the locations $\hat{r}, r \in \mathcal{R}$ and hence this function does not need to be modified.

5.2 Precision-based LPPM

Precision-based LPPMs reduce the precision of the location exposed by disclosing a larger region [9, 11, 25]. This makes it hard for the adversary to pinpoint the exact location of the user. As in the previous case, the LPPM $f(r'|r)$ outputs pseudo-locations r' from $\mathcal{R}' = \mathcal{P}(\mathcal{R}') - \{\emptyset\}$, but in this case r' is formed by a set of one or more *contiguous* regions $r_i \in \mathcal{R}$ that may or may not contain the real location r. The locations contained in r' form the region that is sent to the LBS provider. In the toy example shown in Fig. 2d, we can see two possible outputs r' when the precision is halved by exposing two regions. The black pseudo-location $\mathbf{r'} = \{r_5\} \bigcup \{r_6\}$ contains the real location $r = r_5$, while the grey pseudo-location $\mathbf{r'} = \{r_7\} \bigcup \{r_8\}$ does not. In the latter case the LPPM no only exposes decoy locations, but also perturbs the user's position.

Solution: The bandwidth consumed by a precision-based LPPM strongly depends on the type of information required by the LBS. Let us consider an LBS that returns nearby points of interest. When the user issues a request for a large pseudo-location r' (i.e., with reduced precision), the response contains more points than when the pseudo-location is small, requiring more bandwidth. This is similar to the dummy-based case but has different quality loss and communication overhead, as explained below. Hence, the optimal defense can be computed using the appropriate functions $d_q(\cdot)$ and $d_b(\cdot)$ in the linear program (Eqs (14)-(19)). We refer to this type of systems as *nearby precision-based LPPMs*.

Now consider an LBS in which the provider returns the value of interest (e.g., traffic congestion) for a representative location within r'. In this case the LBS response contains just one value independently of the size of the region, and hence diminishing the precision does not increase the bandwidth consumption. This is similar to the perturbation-based case, where there LPPM does not incur in a communication overhead, but has different quality loss as explained below. The optimal LPPM parameters can be computed using the appropriate function $d_q(\cdot)$ in the linear program (Eqs (4)-(8)). We denote these systems as *aggregated precision-based LPPMs*.

Quality, Bandwidth, and privacy constraints: The quality loss introduced by precision-based LPPMs depends on the type of system. For nearby precision-based LPPMs there is no quality loss when the user's actual location r is included in r', because the response includes the points of interest nearest to this location, and thus $d_q(r', r) = 0$, $\forall r' : r \in r'$. Otherwise, we measure the quality loss as

the minimum distance between the user location r and the locations contained in r' ($d_q(r', r) = \min_{r_i \in r'} \|r_i - r\|_1$, $\forall r' : r \notin r'$). For aggregated precision-based LPPMs, in which the response is one representative value, larger regions r' reduce the expected quality of service. In our experiments we measure the quality loss as the average distance from the user location r to the regions $r_i \in \mathcal{R}$ in r', i.e., $d_q(r', r) = \sum_{r_i \in r'} \|r_i - r\|_1 / N$, being N the number of regions in r'.

The bandwidth consumption only increases for nearby precision-based LPPMs. We define the function $d_b(\cdot)$, that describes the communication cost, as $d_b(r', r) = (\sum_{r_i \in r'} 1) - 1$, and add one unit of bandwidth for each extra region r_i included in r'.

The estimation of the adversary is a location $\hat{r} \in \mathcal{R}$, and thus the privacy constraint does not need to be modified.

6 Evaluation

The linear programs presented in the previous section output optimal LPPM parameters. In this section we evaluate the trade-off between location privacy, service quality, and communication overhead in different types of LPPMs. For this purpose we measure the expected Privacy(ψ, f, h, d_p) offered by an LPPM for a given mobility profile $\psi(r)$, using different combinations of maximum tolerable expected quality loss $\mathrm{Q}_{\mathrm{loss}}^{\max}$ and expected bandwidth consumption $\mathrm{B}_{\mathrm{cost}}^{\max}$. These constraints are modeled depending on the strategy followed by the LPPMs as described in Sections 4.1, 5.1, and 5.2. For precision-based LPPMs we distinguish between *nearby precision-based LPPMs*, which incur in communication overhead but no quality loss; and *aggregated precision-based LPPMs*, which do not consume extra bandwidth but reduce the quality of service.

Existing Dummy-based LPPMs [14, 16, 26]: In these schemes the LPPM algorithm selects a fixed number of requests b_u containing dummy locations. These dummy locations, which are sent to the LBS provider along with the real request, are chosen depending on the user's mobility profile. The real location may be perturbed or not. We model existing dummy-based LPPMs as follows: the user sets a value for the bandwidth consumption b_u that establishes the allowed communication overhead. Then r' is chosen according to the user's mobility profile from all possible pseudo-locations that contain b_u dummies. We note that, in some proposed systems, dummies are chosen also depending on previous exposures in order to resemble realistic movements. However, since we limit our analysis to sporadic LBSs, in which the locations from which the

user makes subsequent requests are not correlated, we do not consider past exposures when selecting dummy locations.

Existing Precision-based LPPMs [9, 11, 25]: In these schemes the user sets a parameter that defines the precision of the exposed location. The real location may be perturbed or not. We model existing precision-based LPPMs as follows: Given that the user chooses a maximal precision reduction s_u, the LPPM selects r' from all pseudo-locations containing s_u contiguous regions $r_i \in \mathcal{R}$, such that the following condition holds: $\forall r_i \in r' : \|r_i - r\|_1 \leq s_u$, considering the Manhattan distance as quality loss function.

Existing attacks: Similarly to prior work [23] we evaluate LPPMs with respect to Bayesian inference attacks [22]. This attack inverts the algorithm implemented by the LPPM using the posterior probability distribution over all locations given the user's profile.

Optimal attacks: We also evaluate the different LPPMs against optimal attacks. We test the performance of the optimal LPPM towards the optimal attack output by the framework; and the performance of existing defenses against the optimal attack against described in prior work which we repeat here for convenience [23]:

$$\textbf{Minimize} \sum_{\hat{r},r',r} \psi(r) f(r'|r) h(\hat{r}|r') d_p(\hat{r}, r) \tag{26}$$

$$\textbf{subject to} \sum_{\hat{r}} h(\hat{r}|r') = 1, \forall r', \text{and } h(\hat{r}|r') \geq 0, \forall \hat{r}, r' \tag{27}$$

6.1 Experimental Setup

We use real mobility profiles obtained from the CRAWDAD dataset epfl/-mobility [19] to evaluate the LPPMs' performance. This dataset contains GPS coordinates of approximately 500 taxis collected over 30 days in the San Francisco Bay Area.

The level of privacy offered by the LPPMs depends on the size of the area of interest, as well as on the number of regions M in which the area is divided. These parameters define the size of the regions r_i, and hence influence the accuracy with which the adversary estimates the user location. When the choice of parameters results in small regions r_i, the adversary can locate the user with more precision than when regions are big (e.g., a large region of interest divided

in few regions). In the following we justify our choices for the size of the area of interest and the number of regions used in our experiments.

Number of regions. The number of regions has a strong impact on the running time of the optimization because the number of possible real locations, pseudo-locations, and estimated locations define the number of inequalities involved in the linear programs. In our evaluation we need to run a large number of linear programs to test a significant sample of quality/bandwidth constraint combinations. Hence, we need to choose an appropriate number of regions in the area of interest to be able to run our experiments in reasonable time.

Let us consider that the area of interest is divided with a grid of $M = \alpha \times \beta$ regions, with no particular restriction on the regions' shape or size. In the strategies considered in this paper, the number of real and estimated locations (r and $\hat{r}$) is the same, and equal to the cardinality of $\mathcal{R}$, i.e., $M = \text{card}(\mathcal{R}) = \alpha \cdot \beta$. However, the number of possible pseudo-locations depends on the strategy implemented by the LPPM. The perturbation-based LPPM transforms real locations into one-region pseudo-locations, hence $\text{card}(\mathcal{R}') = \text{card}(\mathcal{R})$. The dummy-based strategy allows pseudo-locations to contain any combination of non-contiguous locations, and we can compute the number of possibilities for r' as $\text{card}(\mathcal{R}') = \sum_{i=1}^{M} \binom{M}{i}$. Finally, in the precision-based mechanisms pseudo-locations contain combinations of contiguous locations. For simplicity in our experiments for precision-based LPPMs we limit $\mathcal{R}'$ to rectangular pseudo-locations (this would make the pseudo-location $r' = \{r_4\} \bigcup \{r_7\} \bigcup \{r_8\}$ in Figure 2 ineligible). Therefore, the number of pseudo-locations is $\text{card}(\mathcal{R}') = \sum_{i=0}^{\alpha-1} \sum_{j=0}^{\beta-1} (\alpha - i)(\beta - j)$.

We run the linear programs on an HP ProLiant DL980 G7 server with 512 GB RAM and 8 processors Intel E7 2860 with 10 cores each (total 80 cores) using MATLAB's `linprog()` function, and MATLAB's parallel computing capabilities. Table 1 shows the amount of time needed to compute an LPPM function $f(r'|r)$ for different grid sizes αxβ, averaged over combinations of quality and bandwidth restrictions. As expected, the linear program running time grows slower for perturbation-based LPPMs than for preci-sion-based LPPMs, and dummy-based LPPMs quickly become intractable (in fact, we could not compute any LPPM for a 5x5 grid).

While running the experiments we also noticed that when the size of the grid increases MATLAB's linear program solver could not find a solution for some of the optimization problems. The percentage of successful optimizations for each scenario is shown in the third column of Table 1. We note that other linear program solvers could improve this percentage, as well as reduce the running time of the optimization.

Table 1: Performance times for different grid sizes

Perturbation-based			
Grid size	**Mean**	**Std**	**% finished**
2x2	0.22s (0.00 h)	0.26s	100.00
3x3	0.28s (0.00 h)	0.36s	100.00
4x4	0.39s (0.00 h)	0.34s	100.00
5x5	2.30s (0.00 h)	0.64s	100.00
6x6	16.21s (0.00 h)	5.20s	100.00
7x7	211.42s (0.06 h)	128.48s	100.00
8x8	679.58s (0.19 h)	336.75s	100.00
9x9	3437.09s (0.95 h)	1450.49s	100.00
10x10	13199.39s (3.67 h)	6660.02s	100.00
Dummy-based			
Grid size	**Mean**	**Std**	**% finished**
2x2	0.22s (0.00h)	0.18s	100.00
3x3	0.82s (0.00h)	0.33s	100.00
4x4	6710.29s (1.86h)	32653.84s	78.82
Precision-based			
Grid size	**Mean**	**Std**	**% finished**
2x2	0.29s (0.00 h)	0.10s	100.00
3x3	0.26s (0.00 h)	0.18s	100.00
4x4	0.84s (0.00 h)	0.35s	100.00
5x5	6.47s (0.00 h)	2.26s	100.00
6x6	68.51s (0.02 h)	39.74s	100.00
7x7	470.37s (0.13 h)	292.18s	96.88
8x8	1772.80s (0.49 h)	546.09s	72.84
9x9	7056.62s (1.96 h)	1570.97s	68.00
10x10	26223.24s (7.28 h)	6080.76s	63.64

For performance reasons, in our experiments we choose a grid size of 8x6 for perturbation-based and precision-based LPPMs, and 4x3 for dummy-based LPPMs. However, we must stress that a user only needs to run the linear program optimization *once* to compute her optimal protection strategy, and that the mobile device can outsource this operation to a trusted server via a adequately secured connection. Therefore, in reality a larger number of regions can be considered.

Area of interest size. Given a number of regions, the size of the area of interest defines the adversary's inference accuracy. Consider an area of 100 km^2 divided by a 10x10 cartesian grid. The adversary can narrow his estimation of the users' location to at most 1 km^2. If on the other hand the area is only 1

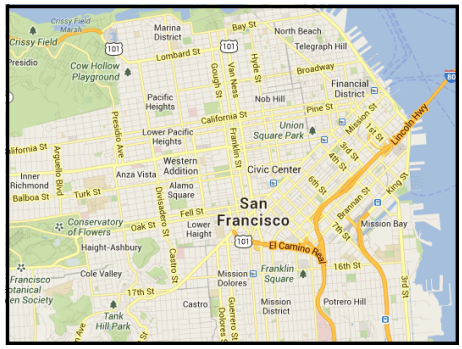


Figure 3: Considered area in San Francisco.

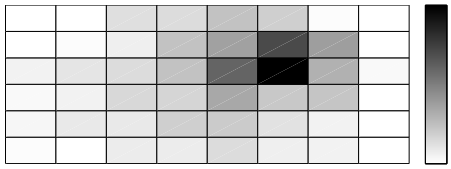


Figure 4: User profile. The darker the region the higher the probability that the user accesses the LBS from this location.

km^2 the the adversary can tighten his estimation to 0.01 km^2.

In order to make our experiments meaningful we select an area of 8×6 km $=$ 48km^2 in Downtown San Francisco which we show in Fig. 3. We divide the area in regions using a cartesian grid of 8x6 or 4x3, depending on the experiment. These grid sizes allow the adversary to infer (with more or less accuracy) the neighborhoods visited by the user. We note that in San Francisco frequent visits to a neighborhood may reveal sensitive information, such as sexual orientation (Castro district), financial status (Financial district), and cultural preferences (Haight-Ashbury).

In [23] Shokri *et al.* demonstrate that the trade-offs between privacy and quality constraints have the same tendency for different users, and that the maximum level of privacy achievable by the LPPM depends on the user's mobility profile. We have run experiments for many individuals in the dataset and confirmed these results. Therefore, without loss of generality, we only show results for one user. We choose as target user the one for which more data is available in the dataset, to have a good estimation of the user's mobility profile. The target user's mobility profile, computed using 36 295 location exposures inside Downtown San Francisco, is shown in Figure 4.

6.2 Results

We separate our evaluation in three steps. First, we show that the optimal dummy-based and precision-based LPPMs designed using the framework are superior to state of the art LPPMs. Second, we evaluate the impact of quality loss and bandwidth overhead constraints on the privacy provided by optimal LPPMs. Finally, we compare the optimal dummy LPPM with the nearby precision based LPPM in terms of privacy, bandwidth consumption and quality loss.

We note that few points are missing in the figures. This is because MATLAB's optimization algorithm was not able to find the solution for these particular combinations of constraints.

Perturbation-based LPPM

For the sake of completeness we make a performance analysis of the perturbation-based LPPM used in prior work using our dataset [23]. The results are shown in Fig. 5, where we compare the privacy offered by the optimal perturbation-based LPPM towards the optimal attack output by the linear program, for different expected quality loss constraints. Confirming previous results [23], we observe that when the service quality constraint is loosened sufficiently the level of privacy provided by the LPPM maxes out. This is because these loose constraints allow the LPPM to choose pseudo-locations that do not leak information that is useful for the attack. Therefore the best estimation of the adversary is only dependent on his prior knowledge, i.e., the user's mobility profile. Once quality constraints are sufficiently loosened, the linear program does can output parameters that do not fulfill tightly the quality constraint. As a consequence the average expected quality loss grows slowly and stabilizes around an optimal value that can be much smaller than the maximum tolerated expected quality loss $\mathrm{Q}_{\mathrm{loss}}^{\mathrm{max}}$.

Bandwith-consuming Optimal LPPMs vs. Existing LPPMs

Let us consider a case in which the quality loss allows the LPPMs to perturb the real location; i.e., $\mathrm{Q}_{\mathrm{loss}}^{\mathrm{max}} > 0$, and thus r' does not necessarily contain r. Given the considered grid sizes, we observe that as soon as some communication overhead is allowed both optimal and existing LPPMs reach the maximum level of privacy achievable.

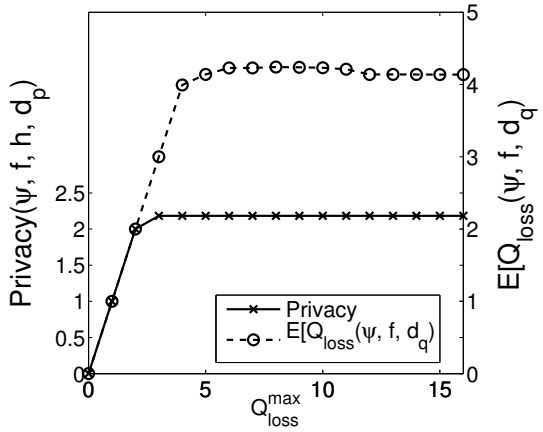
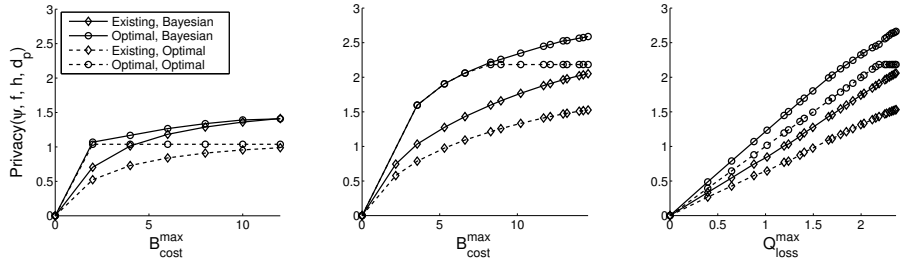


Figure 5: Perturbation-based LPPM: privacy level against the optimal attack; and average expected quality loss.

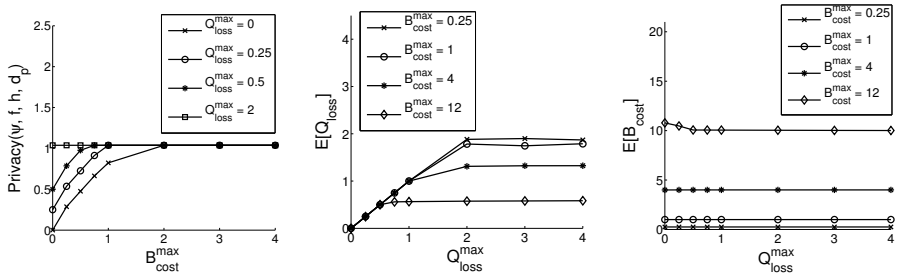


(a) Dummy-based LPPMs (4x3 grid). $Q_{loss}^{max} = 0$; no perturbation.

(b) Nearby precision-based LPPMs (8x6 grid). $Q_{loss}^{max} = 0$; no perturbation.

(c) Aggregated precision-based LPPMs (8x6 grid). $B_{cost}^{max} = 0$; no communication overhead.

Figure 6: Comparison of Optimal and existing LPPMs and attacks.



(a) Privacy (b) Expected quality loss (c) Expected bandwidth

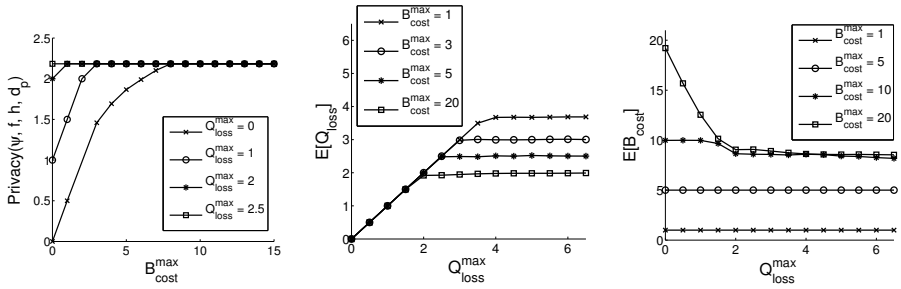
Figure 7: Dummy-based LPPM. (4x3 grid)

Hence, our analysis focuses on the case where the quality constraint does not allow for perturbation, i.e., $\mathrm{Q}_{\mathrm{loss}}^{\max} = 0$. In order to fairly compare optimal and existing algorithms for every possible user constraint b_u (resp., s_u), we construct an existing dummy-based LPPM (resp., precision-based) as described above, and evaluate its quality loss and bandwidth overhead. These values are used as constraints in the linear programs described in Section 5, which output optimal LPPM parameters that meet the same requirements than their corresponding existing counterparts.

Figure 6 shows the results of the comparison depending on the bandwidth constraint $\mathrm{B}_{\mathrm{cost}}^{\max}$. We observe that both the optimal defense and attack perform better than their existing counterparts. Like with quality loss, if the bandwidth constraint is sufficiently loosened the level of privacy maxes out. Note that due to the running time of the algorithms the dummy-based strategy is tested on a smaller grid, and hence the maximum privacy achievable, given by the mobility profile, is lower than in the precision-based case. Finally, the aggregate precision-based LPPM does not impose any bandwidth overhead (see Section 5) and therefore the evaluation in in Fig. 6c considers different values for the quality constraint $\mathrm{Q}_{\mathrm{loss}}^{\max}$.

Trilateral Privacy, Quality, Bandwidth Trade-off

We now study the trade-off between privacy, quality, and bandwidth consumption for dummy- and nearby precision-based LPPMs. We note that the aggregate precision-based LPPM does not impose a bandwidth overhead, and hence its performance is similar to that of the perturbation-based mechanism shown in Figure 5, with a slight difference in the expected quality of service loss.



(a) Privacy. (b) Expected quality loss. (c) Expected bandwidth.

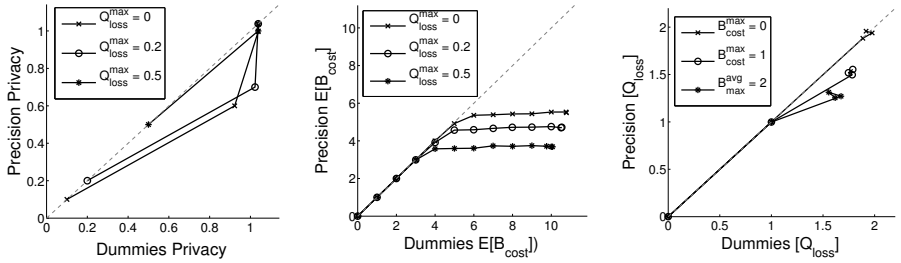
Figure 8: Precision-based LPPM. (8x6 grid)

Figures 7a and 8a show the impact of quality loss and bandwidth constraints on privacy for the optimal dummy- and nearby precision-based LPPMs. As expected, when no extra bandwidth consumption is allowed ($B_{cost}^{max} = 0$) privacy increases with the amount of perturbation allowed by the quality constraint. For a given tolerable expected quality loss Q_{loss}^{max}, relaxing the bandwidth constraint increases the level of privacy achievable until it maxes out. Similarly, loosening the quality constraint increases the level of privacy for a given communication overhead.

Next we examine the trade-off between the expected quality loss $E[Q_{loss}]$ and expected bandwidth overhead $E[B_{cost}]$ for given combinations of Q_{loss}^{max} and B_{cost}^{max}. Recall that when privacy maxes out, further loosening the quality constraint slows the growth of the average expected quality loss. Similarly, the more bandwidth is allowed the less expected quality loss needs to be traded-off for privacy (see Figures 7b and 8b); and the more quality loss is allowed, the less bandwidth needs to be used on average (see Figures 7c and 8c).

Dummy vs. Nearby Precision LPPMs

Finally, we compare dummy-based and nearby precision-based LPPMs in a 4x3 grid. Figure 9a shows the privacy level obtained by both algorithms for different quality and bandwidth constraints (the former showed in the legend, and the latter increased one unit at a time until privacy maxes out). Unsurprisingly, in Fig. 9a we see that for the same combination on constraints, the dummy LPPM performs better in terms of its achieved level of privacy. This is because the optimal nearby precision-based LPPM is restricted to choose $r' \in \mathcal{R}'$ that contain contiguous regions, while the optimal dummy-based LPPM has no



(a) Privacy. (b) Bandwidth. (c) Quality of service loss.

Figure 9: Comparison of optimal dummy-pased LPPM vs. nearby precision-based LPPM.

such contiguity restriction and can make the most of the allowed bandwidth consumption.

With respect to bandwidth overhead, we can see in Fig. 9b that the expected bandwidth consumption $E[B_{cost}]$ of both algorithms is the same until $E[Q_{loss}]$ stabilizes (i.e., when privacy maxes out). Once privacy has maxed out, the expected bandwidth consumption stabilizes for the nearby precision-based LPPM, but continues growing for the dummy-based LPPM. This is because we consider rectangular contiguous pseudo-locations in the precision-based case and therefore there are less eligible regions than in the dummy-based case, where there is no such restriction. For instance, in a 3x3 grid precision-based pseudo-locations can only be formed by 1, 2, 4, 6, and 9 contiguous regions in $\mathcal{R}$, while dummy-based LPPMs can output pseudo-locations containing any combination of 1 to 9 regions. Hence, even if the bandwidth constraint is loosened, the precision-based LPPM has fewer large pseudo-locations to choose from, and thus consumes less bandwidth than the dummy-based strategy, which can select more expensive alternatives.

In terms of quality loss, the dummy-based LPPM suffers more quality degradation than the precision-based LPPM (see Fig. 9c). This is due to the freedom of the dummy-based strategy to select any combination of locations. This allows dummy-based LPPMs to squeeze the quality constraint more efficiently than the precision-based strategy, which is limited to choosing contiguous locations. The clusters at the end of the lines in the figure reflect that the values $E[Q_{loss}]$ and $E[B_{cost}]$ fluctuate slightly once they have stabilized (Fig. 9b).

7 Conclusions

Location Privacy-Preserving Mechanisms (LPPMs) mitigate privacy risks derived from the disclosure of location data when using Location Based Services (LBSs). Shokri *et al.* proposed in prior work a framework to design optimal LPPMs towards strategic adversaries, aware of the LPPM algorithm and the users' mobility patterns [23], for applications in which users only reveal their location sporadically. The proposed framework allows users to set a limit on the maximum tolerated quality loss incurred by the LPPM, but it fails to capture constraints on the resource consumption (e.g., bandwidth) introduced by some LPPM strategies, such as sending dummies, or decreasing the precision of exposed locations.

In this work we have extended Shokri *et al.*'s framework to allow the user to specify a bandwidth constraint. Furthermore, we have modeled two popular strategies to trade-off bandwidth for privacy: a scheme based on sending dummy locations to the LBS, and a scheme based on reducing the precision of the location sent to the LBS.

We have evaluated the performance of LPPMs that consume bandwidth using the CRAWDAD taxi dataset. Our results show that the optimal dummy- and precision- based LPPMs provide more privacy than their respective naive counterparts. Furthermore, both LPPMs perform better than perturbation-based strategies if communication overhead is allowed by the user, with dummy-based LPPMs being the the best choice for a given combination of quality and bandwidth constraints. Furthermore, the results of our simulations show that users can achieve the maximum privacy allowed by their mobility profiles by either permitting a sufficiently large quality of service loss, or bandwidth consumption, or an adequate combination of both.

Acknowledgments We thank Reza Shokri for sharing his optimization code. This research was supported in part by the European Union under project LIFTGATE (Grant Agreement Number 285901) and the European Regional Development Fund (ERDF); and by the projects: IWT SBO SPION, FWO G.0360.11N, FWO G.0686.11N, and GOA TENSE (GOA/11/007).

References

[1] Foursquare. https://foursquare.com/.

[2] Google maps. https://maps.google.com/.

[3] Zeit Online: Betrayed by Our Own Data. http://www.zeit.de/digital/datenschutz/2011-03/data-protection-malte-spitz.

[4] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. pages 901–914, 2013.

[5] Alastair R. Beresford and Frank Stajano. Location Privacy in Pervasive Computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.

[6] A.R. Beresford and F. Stajano. Mix Zones: User Privacy in Location-Aware Services. In *Proceedings of the 2nd Annual Conference on Pervasive Computing and Communications Workshops*, pages 127–131. IEEE, 2004.

[7] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. On the Optimal Placement of Mix Zones. In *7th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 4776 of *Lecture Notes in Computer Science*, pages 216–234. Springer Berlin Heidelberg, 2009.

[8] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. Evaluating the Privacy Risk of Location-Based Services. In *15th International Conference on Financial Cryptography and Data Security (FC)*, volume 7035 of *Lecture Notes in Computer Science*, pages 31–46. Springer Berlin Heidelberg, 2012.

[9] B. Gedik and Ling Liu. Location Privacy in Mobile Systems: A Personalized Anonymization Model. In *25th International Conference on Distributed Computing Systems (ICDCS)*, pages 620–629, 2005.

[10] Philippe Golle and Kurt Partridge. On the Anonymity of Home/Work Location Pairs. In *7th International Conference on Pervasive Computing*, volume 5538 of *Lecture Notes in Computer Science*, pages 390–397. Springer Berlin Heidelberg, 2009.

[11] Marco Gruteser and Dirk Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st International Conference on Mobile systems, Applications and Services (MobiSys)*, pages 31–42. ACM, 2003.

[12] Baik Hoh and Marco Gruteser. Protecting Location Privacy Through Path Confusion. In *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SecureComm)*, pages 194–205. IEEE, 2005.

[13] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. Enhancing Security and Privacy in Traffic-Monitoring Systems. *IEEE Pervasive Computing*, 5(4):38–46, 2006.

[14] Hidetoshi Kido, Yutaka Yanagisawa, and Tetsuji Satoh. An Anonymous Communication Technique Using Dummies for Location-Based Services. In *Proceedings on International Conference on Pervasive Services (ICPS)*, pages 88–97. IEEE, 2005.

[15] John Krumm. Inference Attacks on Location Tracks. In *5th International Conference on Pervasive Computing*, volume 4480 of *Lecture Notes in Computer Science*, pages 127–143. Springer Berlin Heidelberg, 2007.

[16] Hua Lu, Christian S. Jensen, and Man Lung Yiu. PAD: Privacy-Area Aware, Dummy-Based Location Privacy in Mobile Services. In *Proceedings of the 7th ACM International Workshop on Data Engineering for Wireless and Mobile Access (Mobide)*, pages 16–23. ACM, 2008.

[17] Joseph T. Meyerowitz and Romit Roy Choudhury. Hiding Stars with Fireworks: Location Privacy through Camouflage. In *15th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pages 345–356. ACM, 2009.

[18] Femi G. Olumofin, Piotr K. Tysowski, Ian Goldberg, and Urs Hengartner. Achieving Efficient Query Privacy for Location Based Services. In *10th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 6205 of *Lecture Notes in Computer Science*, pages 93–110. Springer Berlin Heidelberg, 2010.

[19] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. CRAWDAD Data Set EPFL/Mobility (v. 2009-02-24). Downloaded from http://crawdad.cs.dartmouth.edu/epfl/mobility, February 2009.

[20] R. Shokri, J. Freudiger, and J. Hubaux. A Unified Framework for Location Privacy. In *HotPETs*, pages 1–12, 2010.

[21] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying Location Privacy. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy (S&P)*, pages 247–262. IEEE, 2011.

[22] Reza Shokri, George Theodorakopoulos, George Danezis, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Quantifying Location Privacy: The Case of Sporadic Location Exposure. In *11th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 6794 of *Lecture Notes in Computer Science*, pages 57–76. Springer Berlin Heidelberg, 2011.

[23] Reza Shokri, George Theodorakopoulos, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Protecting Location Privacy: Optimal

Strategy Against Localization Attacks. In *Proceedings of the 19th ACM Conference on Computer and Communications Security (CCS)*, pages 617–627. ACM, 2012.

[24] Reza Shokri, Carmela Troncoso, Claudia Diaz, Julien Freudiger, and Jean-Pierre Hubaux. Unraveling an Old Cloak: k-Anonymity for Location Privacy. In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 115–118. ACM, 2010.

[25] Yu Wang, Dingbang Xu, Xiao He, Chao Zhang, Fan Li, and Bin Xu. L2P2: Location-aware Location Privacy Protection for Location-based Services. In *Proceedings of the 2012 International Conference on Computer Communications (INFOCOMM)*, pages 1996–2004. IEEE, 2012.

[26] Tun-Hao You, Wen-Chih Peng, and Wang-Chien Lee. Protecting Moving Trajectories with Dummies. In *8th International Conference on Mobile Data Management (MDM)*, pages 278–282. IEEE, 2007.

A Privacy decision variables

In this section we sketch the derivation of the privacy decision variables used in the linear programs in Sections 4.1 and 5. We refer the reader to [23] for more details on the linear programs derivation.

Recall that in the Stackelberg approach the adversary knows the user's choice of LPPM $f(\cdot)$, as well as the user's profile $\psi(r)$. Hence, the adversary can compute the posterior probability $\Pr(r|r')$ that the user being at r when the exposed pseudo-location is r', as well as the probability $\Pr(r)$ of observing r' as follows:

$$\Pr(r|r') = \frac{\Pr(r,r)}{\Pr(r')} = \frac{f(r'|r)\psi(r)}{\sum_r f(r'|r)\psi(r)}\,, \tag{28}$$

$$\Pr(r') = \sum_r \psi(r) f(r'|r)\,. \tag{29}$$

The goal of the adversary is to choose the estimated location $\hat{r}$ that minimizes the expected privacy of the user conditioned to the exposed location being r':

$$\min_{\hat{r}} \sum_r \Pr(r|r') d_p(\hat{r}, r) \tag{30}$$

Combining Eqs (28), (29), and (30), we can express the unconditional expected privacy that the user aims at maximizing as:

$$\sum_{r'} x_{r'}\,, \tag{31}$$

where we have defined

$$x_{r'} \doteq \min_{\hat{r}} \sum_{r} \psi(r) f(r'|r) d_p(\hat{r}, r') \,. \tag{32}$$

Shokri *et al.* note that $x_{r'}$ can be transformed as a series of linear constraints $x_{r'} \leq \min_{\hat{r}} \sum_r \psi(r) f(r'|r) d_p(\hat{r}, r'), \forall r$ and hence $x_{r'}$ can be use as decision variable representing the privacy offered by an LPPM.

Similarly, if we consider the attack $h(\hat{r}, r')$ given that a true location is r and corresponding exposed pseudo-location r', the conditional expected user privacy is:

$$\sum_{\hat{r}} h(\hat{r}|r') d_p(\hat{r}, r) \,. \tag{33}$$

Taking into account the prior knowledge of the adversary on the user's profile $\psi(r)$ the unconditional expected user privacy can be written as:

$$\sum_{r} \psi(r)\, y_r \,, \tag{34}$$

where

$$y_r \doteq \max_{r'} \sum_{\hat{r}} h(\hat{r}|r') d_p(\hat{r}, r) \,. \tag{35}$$

Shokri *et al.* note that y_r can be transformed as a series of linear constraints $y_r \geq \sum_{\hat{r}} h(\hat{r}|r') d_p(\hat{r}, r), \forall r'$ and hence y_r can be use as decision variable representing the privacy obtained against an attack $h(\hat{r}, r)$.

Publication

Possibilistic Location Privacy

Publication Data

Michael Herrmann, Fernando Pérez-González, Carmela Troncoso, and Bart Preneel. Possibilistic Location Privacy. Technical report, COSIC/ESAT, KU Leuven, 2016.

Contributions

- Main author for all sections but Section 2.

Possibilistic Location Privacy

Michael Herrmann[1], Fernando Pérez-González[2], Carmela Troncoso[3], and Bart Preneel[1]

[1] KU Leuven ESAT/COSIC, iMinds, Leuven, Belgium
{name.surname}@esat.kuleuven.be
[2] University of Vigo, Department of Signal Theory and Communications, Vigo, Spain
fperez@gts.uvigo.es
[3] The IMDEA Software Institute, Madrid, Spain
carmela.troncoso@imdea.org

Abstract. The most comprehensive framework for quantifying location privacy up to date relies on a Markovian approach whose computational complexity grows heavily with the number of considered regions/time slots. This hampers its use in practice: it can only evaluate movements in a small surface divided in fine regions, or in a large surface divided in coarse regions which are not informative. We introduce a novel notion for quantifying privacy, possibilistic location privacy, that trades off expensive probabilistic reasoning for simple intersection operations. This gain in performance allows to quantify location privacy on large surfaces with significant more precision than the Markovian approach. We describe two algorithms to compute possibilistic location privacy and use them to quantify the protection offered by two geo-indistinguishability mechanisms when users frequently expose their location. Our experiments, performed on real data, confirm that the computational requirements of our approach are extremely low while it obtains accurate results that would be infeasible to compute from a Markovian perspective.

1 Introduction

Quantifying location privacy is key in order to evaluate and compare location privacy-preserving mechanisms [1, 5, 9, 14]. A comprehensive solution to this problem was proposed by Shokri et al. in [19] where they propose a framework for tackling the quantification problem, and a realization of the framework as a tool to measure location privacy [18]. This tool implements a Markovian approach to compute location privacy, i.e., it models users' movements as a

Markov chain, it models location privacy and inference attacks as a probability density functions that operate on real and obfuscated locations, and it measures privacy as the adversary's expected estimation error.

While Shokri et al.'s approach is very promising in terms of flexibility, its underlying probabilistic reasoning is very expensive and thus it does not scale to realistic scenarios (its computational complexity grows quadratically with the number of considered regions, and linearly with the number of time slots [19]). In fact, in order to be feasible in practice it requires to considerably simplify the scenario being evaluated by either limiting the targeted surface [1], decreasing the number of considered discrete regions (e.g., quantizing the map into coarse regions [19], or considering only the top likely locations [3]), considering large time slots [19], or pruning the possible states to keep the complexity of the hidden Markov model manageable [20, 21]. The need for such simplifications effectively means that the Markovian approach is only suitable to study small surfaces, or to produce non-informative coarse-grained results on large region.

In this paper we introduce a novel notion for quantifying privacy in LBS, *possibilistic location privacy*, that provides a first-order estimation of the privacy protection provided by obfuscation-based mechanisms. The key feature of our approach is simplicity: instead of performing expensive probabilistic computations, it only performs intersections between sets of possible user locations at different instants in time. Thus, as opposed to the Markovian approach, it can efficiently handle any surface size and any frequency of location exposure. We propose algorithms to compute possibilistic location privacy and show with real-world location data that its computational complexity is extremely low, i.e., suitable to perform real-time location privacy evaluation on commodity hardware.

We are not the first ones to observe that possibilistic reasoning may be adequate to study location privacy. This approach has been shown to be capable of undermining the protection provided by precision reduction mechanisms [9] by Ghinita *et al.*. [8]. Yet, their work is limited to proposing mechanisms that do not leak any information from a possibilistic perspective, and does not provide means to quantify location privacy when perfect protection is not possible. On the contrary, we focus on the potential of this approach as a quantification tool useful to evaluate and compare the effectiveness of obfuscation-based privacy-preserving location.

We use possibilistic location privacy to quantify, for the first time, the location privacy offered by two protection mechanisms based on geo-indistinguishability [1,5] in scenarios where users frequently expose their location. Prior work only evaluated these mechanisms for sporadic exposures [1, 3], provided a theoretical analysis of the privacy loss trend when correlated points

are exposed to the adversary [1, 5], or observed that trajectories can be inferred from exposed locations [22]; but none of these papers provides means to compute the concrete privacy level achieved by the mechanism. Thanks to our possibilistic approach we provide a first order quantification of location privacy for these mechanisms and uncover tradeoffs in their configuration not studied in previous works. We note that these mechanisms cannot be easily analyzed from the Markovian perspective since geo-indistinguishability operates on continuous locations and time and thus a meaningful quantized representation of users' movements would incur in a prohibitive cost.

Our contributions can be summarized as follows:

- We introduce *possibilistic location privacy*, a new notion to quantify the privacy protection provided by obfuscation-based mechanisms, and we show its relation to the probabilistic model underlying previously proposed Markovian approaches.
- We provide algorithms to compute possibilistic privacy that are extremely efficient, i.e. require less than 1 second (the highest GPS update frequency) per location, and suitable for accurately quantifying location privacy for obfuscation-based mechanisms at large scale, e.g., we consider a target surface of $6.5M\text{km}^2$ in our experiments.
- We show that the computational needs of Markovian approaches become prohibitive when computing meaningful accurate results in large surfaces.
- We quantify, for the first time, the privacy offered by two geo-indistinguishability mechanisms when more than one location is exposed to the adversary. Our experiments uncover new trade-offs in one of these mechanisms not considered by previous work.

The rest of this paper is organized as follows: the next section presents our possibilistic model for quantification and shows its connection to probabilistic approaches. Section 3 proposes two algorithms to efficiently compute possibilistic regions. We compare our approach to the Markovian evaluation from Shokri et al. [19] in Section 4 and illustrate its capabilities quantifying the privacy offered by two geo-indistinguishability-based schemes in Section 5. We conclude the paper in Section 6.

2 Quantifying Location Privacy: From Probabilistic to Possibilistic

Quantifying location privacy consists on evaluating the performance of inference attacks against users' private location information. In this section we introduce our possibilistic quantification approach, showing that i) it is a special case of a probabilistic model equivalent to the underlying Markovian approach in the framework of Shokri *et al.*, and ii) it can be integrated in this framework to enable the quantification of location privacy with high accuracy in larger regions than the Markovian approach, as we show in Sect. 5.

Before diving into the description we introduce the following notation to model user movements and obfuscation mechanisms, which is summarized in Table 1:

- **Probability density function**: denoted by $f(\cdot)$.
- **User real location**: $\mathbf{x}_i$ is the column vector containing the real (x, y) coordinates of a user at instant i. $\mathbf{X}_i$ is a matrix containing all user locations up to instant i, where each column contains a tuple of (x, y) coordinates. The set of all possible locations is denoted by $\mathcal{X} \subset \mathbb{R}^2$.
- **User movements**: $\mathbf{\Psi}_i$ is the vector that contains the parameters that probabilistically determine $\mathbf{x}_{i+1}$ from $\mathbf{x}_i$. This vector may contain information related to movement patterns such as the time elapsed between instants i and $i+1$, the average velocity in this period, etc; or related to terrain information such as possible turns, existence of walls, lakes, etc. For simplicity we assume that $\mathbf{\Psi}_i$ is independent of the actual location $\mathbf{x}_i$, but we note that the extension to a more general case that considers dependencies is straightforward. We denote by $f(\mathbf{x}_{i+1}|\mathbf{\Psi}_i, \mathbf{x}_i)$ the probability density function describing users' mobility patterns.
- **Observed obfuscated locations**: $\mathbf{z}_i$ is the column vector containing the obfuscated coordinates exposed by a user at instant i, i.e., the adversary's observation at instant i. $\mathbf{Z}_i$ is a matrix containing all the obfuscated observations up to instant i, and we denote by $f(\mathbf{Z}_i)$ the probability density function describing the probability of observing $\mathbf{Z}_i$. $\mathcal{Z}$ denotes the set of all possible obfuscated locations which are coordinates in our analysis, though we note that our formulation can be extended to account for other type of obfuscated locations, such as a center and radius of a circle, n coordinates defining a polygon, etc.
- **Obfuscation mechanism**: We denote as $f(\mathbf{z}_i|\mathbf{x}_i)$ a probability density function describing the probability of exposing obfuscated location $\mathbf{z}_i$ when

the real location is $\mathbf{x}_i$, i.e., $f(\mathbf{z}_i|\mathbf{x}_i)$ models the operation of a obfuscation mechanism.

We make the following assumptions regarding the statistical description of the involved variables: i) $f(\mathbf{x}_{i+1}|\mathbf{\Psi}_i, \mathbf{X}_i) = f(\mathbf{x}_{i+1}|\mathbf{\Psi}_i, \mathbf{x}_i)$. This means that the next location only depends on both the current location and the parameter vector $\mathbf{\Psi}_i$. Notice that depending on the trajectory this will not be true in general, but we make this assumption here to simplify the presentation of the analysis; and ii) $f(\mathbf{z}_i|\mathbf{X}_i) = f(\mathbf{z}_i|\mathbf{x}_i)$. This means that the obfuscation mechanism is memoryless, that is, obfuscation is a (probabilistic or deterministic) function of the current location. We assume a probabilistic function for its generality. While carrying out the analysis for memoryless mechanisms may seem very restrictive, we show in Sect. 5 that it does not prevent our model from being useful in presence of obfuscation mechanisms that have memory.

Finally, we assume that the adversary is *causal*, i.e., she constructs an estimate $\hat{\mathbf{x}}_i$ of the user's true location at instant i from the vector of observations $\mathbf{Z}_i$ (notice that the adversary could construct non-causal estimates by using later observations, e.g., $\mathbf{z}_{i+1}$). We leave the general case for future work.

2.1 Probabilistic Location Privacy

This section illustrates the adversary's approach from a probabilistic perspective, that is equivalent to the Markovian approach used by Shokri *et al.*. For the sake of example we focus on localization attacks in which the adversary tries to find the location of a user $\mathbf{x}_i$ at a given time i. In this case, a reasonable choice for the adversary is to use the *Maximum Likelihood* principle to estimate the real location of the user as:

$$\hat{\mathbf{x}}_i = \arg\max f(\mathbf{x}_i|\mathbf{Z}_i, \mathbf{\Psi}_i), \tag{1}$$

where the conditioning to $\mathbf{\Psi}_i$ represents the prior information that the adversary may have about user movements or terrain information. Since it does not influence the analysis, we drop this parameter from the notation to improve readability but note that it is important in the computation of probabilities.

To carry out the above optimization the adversary needs to calculate the probability density function $f(\mathbf{x}_i|\mathbf{Z}_i)$ which, applying the probability chain rule,

can be done recursively as follows:

$$\begin{aligned} f(\mathbf{x}_i|\mathbf{Z}_i) &= \frac{f(\mathbf{x}_i,\mathbf{Z}_i)}{f(\mathbf{Z}_i)} = \frac{f(\mathbf{z}_i|\mathbf{x}_i,\mathbf{Z}_{i-1})\cdot f(\mathbf{x}_i,\mathbf{Z}_{i-1})}{f(\mathbf{Z}_i)} \\ &= \frac{f(\mathbf{z}_i|\mathbf{x}_i,\mathbf{Z}_{i-1})\cdot f(\mathbf{x}_i|\mathbf{Z}_{i-1})\cdot f(\mathbf{Z}_{i-1})}{f(\mathbf{Z}_i)} \\ &= f(\mathbf{Z}_i)^{-1}\cdot f(\mathbf{z}_i|\mathbf{x}_i,\mathbf{Z}_{i-1})\cdot f(\mathbf{Z}_{i-1})\cdot \\ &\quad \sum_{\mathbf{x}_{i-1}\in\mathcal{X}} f(\mathbf{x}_i|\mathbf{Z}_{i-1},\mathbf{x}_{i-1}) f(\mathbf{x}_{i-1}|\mathbf{Z}_{i-1})\,. \end{aligned} \tag{2}$$

Now we make use of our assumptions: since the obfuscation algorithm is memoryless we can write $f(\mathbf{z}_i|\mathbf{x}_i,\mathbf{Z}_{i-1}) = f(\mathbf{z}_i|\mathbf{x}_i)$, and since user movements do not depend on the obfuscation mechanism we can write $f(\mathbf{x}_i|\mathbf{Z}_{i-1},\mathbf{x}_{i-1}) = f(\mathbf{x}_i|\mathbf{x}_{i-1})$. Then, (2) becomes

$$\begin{aligned} f(\mathbf{x}_i|\mathbf{Z}_i) &= f(\mathbf{Z}_i)^{-1}\cdot f(\mathbf{z}_i|\mathbf{x}_i)\cdot f(\mathbf{Z}_{i-1})\cdot \\ &\quad \sum_{\mathbf{x}_{i-1}\in\mathcal{X}} f(\mathbf{x}_i|\mathbf{x}_{i-1})\cdot f(\mathbf{x}_{i-1}|\mathbf{Z}_{i-1}) \\ &= g(\mathbf{Z}_i)\cdot f(\mathbf{z}_i|\mathbf{x}_i)\cdot \\ &\quad \sum_{\mathbf{x}_{i-1}\in\mathcal{X}} f(\mathbf{x}_i|\mathbf{x}_{i-1})\cdot f(\mathbf{x}_{i-1}|\mathbf{Z}_{i-1})\,, \end{aligned} \tag{3}$$

where $g(\mathbf{Z}_i) = f(\mathbf{Z}_i)^{-1}\cdot f(\mathbf{Z}_{i-1})$ only depends on the observations, so it can be regarded to as a normalization factor with no effect on the optimization to be carried out by the adversary. This equation defines a forward recursion that at instant i contains three ingredients: 1) the obfuscation mechanism: $f(\mathbf{z}_i|\mathbf{x}_i)$; 2) the location evolution: $f(\mathbf{x}_i|\mathbf{x}_{i-1})$, and 3) the conditional probability of the user's real location given the adversary's observation at instant $i-1$: $f(\mathbf{x}_{i-1}|\mathbf{Z}_{i-1})$.

Let us define $\alpha(\mathbf{x}_i) \doteq f(\mathbf{x}_i|\mathbf{Z}_i)$, where the dependence with the vector of observations is implicit. Then, (3) is equivalent to[4]

[4]We note that if the possible locations would be defined over a continuous space, then the expression in (4) would be valid after replacing the sum by an integral.

$$\alpha(\mathbf{x}_i) = g(\mathbf{Z}_i) \cdot f(\mathbf{z}_i|\mathbf{x}_i) \cdot \sum_{\mathbf{x}_{i-1} \in \mathcal{X}} f(\mathbf{x}_i|\mathbf{x}_{i-1}) \cdot \alpha(\mathbf{x}_{i-1}) \,. \tag{4}$$

A considerable simplification to (4) is afforded when the distribution $f(\mathbf{x}_i|\mathbf{x}_{i-1})$ can be written as

$$f(\mathbf{x}_i|\mathbf{x}_{i-1}) = \varphi(\mathbf{x}_i - \mathbf{x}_{i-1}) \,, \tag{5}$$

where φ is a probability density function. Equation (5) implies that the random variable $\mathbf{x}_i$ can be written as $\mathbf{x}_i = \mathbf{x}_{i-1} + \mathbf{v}_i$, where $\mathbf{v}_i$ is independent of $\mathbf{x}_{i-1}$. In other words, $\mathbf{v}_i$ models the movement of the user from $\mathbf{x}_{i-1}$ to $\mathbf{x}_i$ according to the parameters in $\mathbf{\Psi}$. The assumption in (5) allows us to write the sum in (4) as a convolution:

$$\alpha(\mathbf{x}_i) = g(\mathbf{Z}_i) \cdot f(\mathbf{z}_i|\mathbf{x}_i) \cdot (\varphi(\mathbf{x}) * \alpha(\mathbf{x})) \mid_{\mathbf{x}=\mathbf{x}_i} \tag{6}$$

where $*$ denotes convolution. The interpretation of (6) is the following: given $\alpha(\mathbf{x}_{i-1})$ we first convolve it with the probability density function of $\mathbf{v}_i$ to determine the *prior* distribution of $\mathbf{x}_i$; then we update it with the conditional probability $f(\mathbf{z}_i|\mathbf{x}_i)$ to produce the *posterior* distribution after observing $\mathbf{z}_i$. This produces $\alpha(\mathbf{x}_i)$ and the process can be repeated every time a new observation is available.

Recall that $\alpha(\mathbf{x}_i) \doteq f(\mathbf{x}_i|\mathbf{Z}_i)$. Thus, (6) computes the sought probability distribution over real locations given the observation, which allows to carry out the localization attack.

2.2 Possibilistic Location Privacy

In the following we present the possibilistic approach for estimating locations. In contrast to the Markovian approach from the previous section that considers probability distributions, our possibilistic approach is designed along a set-membership formalization. In this sense, we only consider a set of possible locations, i.e. areas in that the user can *possibly* be. Note that in the following whenever we write $\mathcal{O} + \mathcal{P}$ this sum is meant for a set of locations and a set of movement vectors in the set-theoretic sense.

Let us assume that the user's location $\mathbf{x}_{i-1}$ is known to be contained in a compact set of locations $\mathcal{S}_{i-1} \subset \mathbb{R}^2$. Then, suppose that, given a location $\mathbf{x}_{i-1}$, we know that the next location $\mathbf{x}_i$ is such that $\mathbf{x}_i - \mathbf{x}_{i-1} \in \mathcal{V}_i$, for some compact

set $\mathcal{V}_i$. Making the same assumption as in (5), $\mathcal{V}_i$ can be seen as the region which the user can reach given $\mathbf{v}_i$ that models user movements according to the parameters in $\mathbf{\Psi}$. Therefore, *prior* to observing $\mathbf{z}_i$, it is easy to see that the set of possible locations at instant i is $\mathcal{S}_{i-1} + \mathcal{V}_i$. Intuitively, this means that the user, who was known to be in $\mathcal{S}_{i-1}$ may stay at any of the locations in this set, or travel along a vector in $\mathcal{V}_i$ to a new location.

While this allows us to capture possible movements of the user, we can further take into account that the the adversary observes the user's obfuscated location $\mathbf{z}_i$, and knows that the distance between $\mathbf{x}_i$ and $\mathbf{z}_i$ is usually bounded in order to provide some utility for the user. Thus, we can obtain $\mathbf{z}_i - \mathbf{x}_i \subset \mathcal{Z}_i$, for some compact set $\mathcal{Z}_i$ according to the obfuscation mechanism parameters. Therefore, the set of possible locations based on observing $\mathbf{z}_i$ is given by $\mathbf{z}_i + \mathcal{Z}_i$, so the *posterior* feasible set can be written as

$$\mathcal{S}_i = (\mathbf{z}_i + \mathcal{Z}_i) \cap (\mathcal{V}_i + \mathcal{S}_{i-1})\,. \tag{7}$$

Equation (7) can be seen as the analog of (6) with *possibilities* instead of *probabilities*. We can regard $\mathcal{S}_i$ as the feasible set associated to $\mathbf{x}_i$, $\mathcal{V}_i$ is the feasible set corresponding to $\mathbf{v}_i$, and $\mathcal{Z}_i$ is the feasible set associated to $\mathbf{z}_i|\mathbf{x}_i$.

To see the connection between (7) and (6) more clearly, let us *binarize* the range of the probability density functions involved in (6). To this end, it is convenient to use the indicator function $\mathbb{1}$, such that, given a set $\mathcal{S} \subset \mathbb{R}^2$, $\mathbb{1}_{\mathcal{S}}(\mathbf{x}) = 1$ if $\mathbf{x} \in \mathcal{S}$, and is zero otherwise. We further denote by $|\mathcal{R}|$ the *area* of $\mathcal{R}$. If we write

$$\begin{aligned} \alpha(\mathbf{x}_i) &= |\mathcal{S}_i|^{-1}\mathbb{1}_{\mathcal{S}_i}\,, \\ \varphi(\mathbf{x}) &= |\mathcal{V}_i|^{-1}\mathbb{1}_{\mathcal{V}_i}\,, \\ f(\mathbf{z}_i|\mathbf{x}_i) &= |\mathcal{Z}_i|^{-1}\mathbb{1}_{z_i+\mathcal{Z}_i}\,, \end{aligned} \tag{8}$$

then (7) and (6) are identical (save for a normalization factor) *as long as* the convolution output $\alpha(\mathbf{x}) * \varphi(\mathbf{x})$ is binarized by a function B such that $B(\alpha(\mathbf{x}) * \varphi(\mathbf{x})) = 1$ if $\alpha(\mathbf{x}) * \varphi(\mathbf{x}) > 0$ and $B(\alpha(\mathbf{x}) * \varphi(\mathbf{x})) = 0$ if $\alpha(\mathbf{x}) * \varphi(\mathbf{x}) = 0$. This binarization is reminiscent of the *dilation* operator that is commonly used in morphological image processing [16]. Starting with a set $\mathcal{S}_{i-1}$ we *dilate* it using $\mathcal{V}_i$ as kernel (i.e., gradually enlarge the boundaries of $\mathcal{S}_{i-1}$ according to the structure defined by $\mathcal{V}_i$). This gives the a priori feasible set. Then, an intersection with $\mathbf{z}_i + \mathcal{Z}_i$ is carried out to yield the a posteriori feasible set $\mathcal{S}_i$. See (7).

Table 1: Notation

$f()$	Denotes a probability density function
$\mathbf{x}_i$	The user's actual location at time i
$\mathbf{X}_i$	Vector of all real locations up to instant i
$\mathbf{z}_i$	The user's observed obfuscated location at time i
$\mathbf{Z}_i$	Vector of all observed obfuscated locations up to instant i
$\mathbf{\Psi}_i$	Vector of parameters that probabilistically determine $\mathbf{x}_{i+1}$ from $\mathbf{x}_i$
$\mathbf{v}_i$	Vector modelling user movement from $\mathbf{x}_{i-1}$ to $\mathbf{x}_i$ according to the parameters in $bt\Psi$
$\mathcal{Z}_i$	Movement vectors associated to $\mathbf{z}_i \vert \mathbf{x}_i$.
$\mathcal{V}_i$	Movement vectors associated to $\mathbf{v}_i$.
$\mathcal{E}_i$	**Expanded region**: Feasible set associated to $\mathbf{S}_{i-1} + \mathcal{V}_i$.
$\mathcal{B}_i$	**Obfuscated region**: Feasible set associated to $\mathbf{z}_i + \mathcal{Z}_i$.
$\mathcal{S}_i$	**Possibilistic region**: Feasible set associated to $\mathbf{x}_i$.
$\vert\mathcal{R}\vert$	Area of a region $\mathcal{R}$

As a conclusion, we see that the set-membership approach is a simple (binary, so to speak) way to carry out the update in (6). Notice that since the set-membership approach implicitly quantizes the probability values, a maximum likelihood estimate as in (1) would not generally give a unique solution: all values in $\mathcal{S}_i$ are equally feasible (in probabilistic terms, they all have the same likelihood).

In the following we call $\mathcal{S}_i$ *Possibilistic Region*, $\mathcal{E}_i = \mathcal{S}_{i-1} + \mathcal{V}_i$ *Expanded Region*, and $\mathcal{B}_i = \mathbf{z}_i + \mathcal{Z}_i$ *Obfuscated Region*.

2.3 Possibilistic Location Privacy in Prior Work

The framework proposed by Shokri et al. [19] consists on the following elements: $\langle \mathcal{U}, \mathcal{A}, \text{LPPM}, \mathcal{O}, \text{ADV}, \text{METRIC} \rangle$.

The set $\mathcal{U}$ represents the users in the system that may expose locations. For the sake of simplicity our analysis is centered on one user but, as we show in Sect. 4, the efficiency of the possibilistic approach allows it to handle large populations. The set $\mathcal{A}$ models the actual user traces, represented in our notation by $\mathbf{X}_i = [\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_i]$; LPPM is the obfuscation algorithm that transforms actual locations $\mathbf{x}_i$ in obfuscated observations $\mathbf{z}_i$, represented in our notation by $f(\mathbf{z}_i|\mathbf{x}_i)$; and the set of observable obfuscated user traces $\mathcal{O}$ is represented in our notation by $\mathbf{Z}_i = [\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_i]$. The adversary ADV is defined as an entity that implements some inference attack to learn information

about the user locations given the observed obfuscated trace, her knowledge of the LPPM, and her knowledge on the users mobility model (represented by $\mathbf{\Psi}$ in our notation). The metric METRIC captures the location privacy of the user quantified as the adversary's expected error after the attack.

The integration of our approach in Shokri et al.'s framework means that operation of the adversary ADV changes from probabilistic to possibilistic. In our setting, the adversary ADV deploys the possibilistic approach to find the *Possibilistic Region* $\mathcal{S}_i$ in order to localize the user, though we note that possibilistic reasoning could be used to launch other attacks formalized in [19], e.g., it could be used to launch a tracking attack by finding possible trajectories (i.e., subsequent feasible possibilistic regions); or to find meeting points by intersecting possibilistic regions of different users.

While the change in the adversary's approach may seem irrelevant, it has great impact on the scalability of the framework to quantify location privacy. In Shokri et al.'s work quantification relies on expensive probabilistic algorithms for which performance depends on the number of regions and time instant to be evaluated. To reduce the computational load the framework assumes that users move within an area that is partitioned into M distinct discrete regions, and that time is discrete being the set of time instants when the users may be observed $\mathcal{T} = \{1, \ldots, T\}$. This assumption reduces the precision of the framework, since the maximum accuracy achievable is limited to the size of the discrete regions, and still it is prohibitely expensive if precision is kept as a reasonable level, see in Sect. 4. The possibilistic approach, on the other hand, relies on a cheap intersection operation and hence can take the quantification to large scale ($6.5Mkm^2$ in our experiments) without trading off precision, analyzing scenarios in which the Markovian approach would be infeasible.

Possibilistic Location Privacy Metrics

Because our approach is not probabilistic, we cannot use any of the options for METRIC proposed in [19]. We now define adequate metrics to capture the user's location privacy (i.e., the performance of the adversary) for the possibilistic case. Following Shokri *et al.*'s insights on the dimensions of privacy we provide metrics to quantify accuracy, certainty and correctness.

To quantify the *accuracy* of the possibilistic approach we use the ***possibilistic area size***, denoted as ϕ:

$$\phi_i = |\mathcal{S}_i| \,. \tag{9}$$

This metric captures the precision with which the adversary can pinpoint an estimated user location, measured as the size of the area of possible locations

for the user. Since in the possibilistic approach all locations in the area are equally likely to be the real location of the user, the area size also captures the *uncertainty* of the adversary about the user's location – considering the definition of uncertainty in [19]: "the ambiguity of the posterior distribution of the possible user locations given the observed exposed region with respect to finding a unique answer – that unique answer need not be the correct one".

One of the most important features of a quantification framework is to allow for meaningful comparisons between protection mechanisms. The possibilistic area size provides a good way to compare in terms of absolute location privacy. However, there are other interesting features that one may want to evaluate such as the gain in certainty obtained when correlated obfuscated locations are exposed to the adversary. To this end, we define the ***certainty gain***, denoted as ρ_i, as:

$$\rho_i = 1 - \frac{|\mathcal{S}_i|}{|\mathcal{B}_i|}\,. \tag{10}$$

This metric captures the gain in certainty the adversary experiences from the case where she observes an isolated obfuscated location, to the case where subsequent locations are observed. In the former case uncertainty is given by $|\mathcal{B}_i| = |\mathbf{z}_i + \mathcal{Z}_i|$ since there is no prior/posterior observation to intersect with; while in the latter case uncertainty is given by the size of the possibilistic area $\mathcal{S}_i$, see Equation (7).

Neither the area size ϕ, nor the certainty gain ρ capture whether the user's true location is actually inside $\mathcal{S}_i$, i.e., the *correctness* of the attack. Even if $\mathcal{S}_i$ tends to be very small, it may be of limited use to the adversary if the user's true location is unlikely to be included. We define the ***adversary success***, denoted as σ, that captures whether the adversary finds the correct answer in her attack as:

$$\sigma_i = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathcal{S}_i \\ 0, & \text{otherwise}\,. \end{cases} \tag{11}$$

3 Practical Algorithms for Computing Possibilistic Regions

We now provide two algorithms to compute possibilistic regions $\mathcal{S}_i$ that either use past observed obfuscated locations to compute the region $\mathcal{S}_i$ at time i, or use the current observed location at time i to further reduce regions $\mathcal{S}_j$ at past time instants j, $j < i$.

Computing the current Possibilistic Region from past observations. Given an obfuscated observed location $\mathbf{z}_i$ at time i, and the previous n_{bwd} possibilistic regions, the current possibilistic region $\mathcal{S}_i$ corresponding to the real location $\mathbf{x}_i$ can be computed according to Algorithm 1. The parameter $n_{\text{bwd}} \in [1, \ldots, i-1]$ models the amount of past information to be taken into account. Figure 1 illustrates the operation of Algorithm 1 for $n_{\text{bwd}} = 2$.

Algorithm 1 Compute current $\mathcal{S}_i$ given n_{bwd} past possibilistic regions.

1: $\mathcal{B}_i = \text{collect}(\mathbf{z}_i, p_{\text{mass}})$
2: **for** $j = [1, 2, \ldots, n_{\text{bwd}}]$ **do**
3: $\quad \mathcal{E}_j = \text{expand}(\mathcal{S}_{i-j}, \mathcal{V}_{i-j+1})$
4: **end for**
5: $\mathcal{S}_i = \text{intersect}(\mathcal{B}_i, \mathcal{E}_1, \ldots, \mathcal{E}_{n_{\text{bwd}}})$
6: **if** $\mathcal{S}_i = \emptyset$ **then** $\mathcal{S}_i = \mathcal{B}_i$ **end if**
7: **Return:** $\mathcal{S}_i$

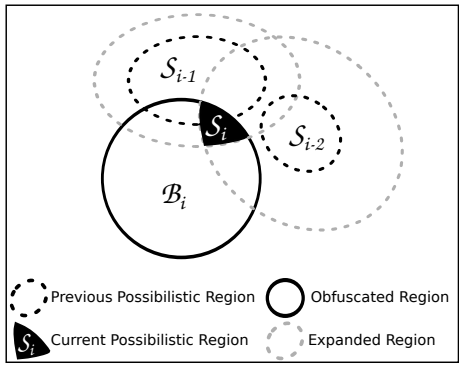


Figure 1: Graphical representation of Algorithm 1 for $n_{\text{bwd}} = 2$. $\mathcal{B}_i$ (black thick line) is the obfuscated region where $\mathbf{x}_i$ can be, given $(\mathbf{z}_i + \mathcal{Z}_i)$. $\mathcal{S}_{i-1}$ and $\mathcal{S}_{i-2}$ (black dashed lines) are the last possibilistic regions, $\mathcal{E}_i$ and $\mathcal{E}_{i-1}$ (grey dashed lines) are the expanded regions corresponding to $\mathcal{S}_{i-1}$ and $\mathcal{S}_{i-2}$. The intersection of $\mathcal{B}_i$, $\mathcal{E}_i$, and $\mathcal{E}_{i-1}$ is the current possibilistic region $\mathcal{S}_i$.

First, (line 1), the function `collect` constructs the obfuscated region $\mathcal{B}_i$ from the observed obfuscated location $\mathbf{z}_i$, i.e., a region containing the set of all possible locations that could have generated $\mathbf{z}_i$ according to the obfuscation mechanism $f(\mathbf{z}_i|\mathbf{x}_i)$ (see Sect. 2). We parametrize `collect` with the parameter $p_{\text{mass}} \in [0, 1]$ that limits the returned obfuscated region $\mathcal{B}_i$ to contain only p_{mass} of the probability density of $f(\mathbf{x}_i|\mathbf{z}_i)$, i.e., $\mathcal{B}_i$ is chosen so that $f(x_i|z_i)$ summed over $\mathcal{B}_i$ is p_{mass}. This is to avoid considering large non-informative $\mathcal{B}_i$s, since

for some obfuscation mechanisms this region can be arbitrarily large but with a large fraction containing the real location with negligible probability.

In the for loop (line 2-4) the function `expand` expands each of previous possibilistic region according to the user movements indicated by $\mathbf{v}_{i-j+1}$, i.e., finds the expanded region $\mathcal{E}_j = \mathcal{S}_{i-j} + \mathcal{V}_i$ where the user location may possibly be after she moves from $\mathbf{x}_{i-j}$ to $\mathbf{x}_{i-j+1}$ (e.g., determining the Minkowski sum [2] around $\mathcal{S}_{i-j}$, as explained in [8]). Once all expanded regions are constructed, the function `intersect` (line 5) finds the intersection of the expanded regions with the obfuscated region $\mathcal{B}_i$ computed in line 1 to obtain the current possibilistic region $\mathcal{S}_i$. If `intersect` returns $\emptyset$ (i.e., no intersection between $\mathcal{B}_i$ and any $\mathcal{E}_j$), then $\mathcal{S}_i = \mathcal{B}_i$ (line 6). This means that past observations are of no use to reduce the obfuscated region $\mathcal{B}_i$. We provide details on our implementation of the functions `expand` and `intersect` in Appendix A.

Using the current observation to update past Possibilistic Regions. We now present a second algorithm, Algorithm 2, where the adversary uses the current obfuscated region to update previous possibilistic locations, i.e., the previous possibilistic regions $\mathcal{S}_{i-1}$ to $\mathcal{S}_{i-n_{\text{fwd}}}$, narrowing down the region in which the users' real location could possibly be. The parameter $n_{\text{fwd}} \in [1, \ldots, i-1]$ represents the number of previous possibilistic regions to be updated. Figure 2 illustrates the operation of Algorithm 2 for $n_{\text{fwd}} = 2$.

Algorithm 2 Update previous possibilistic areas $[\mathcal{S}_{i-n_{\text{fwd}}}, \ldots, \mathcal{S}_{i-1}]$ given current observation.

1: $\mathcal{B}_i = \text{collect}(\mathbf{z}_i, p_{\text{mass}})$
2: **for** $j = [1, 2, \ldots, n_{\text{fwd}}]$ **do**
3: $\quad \mathcal{E}'_j = \text{expand}(\mathcal{S}_i, \mathcal{V}_{i-j+1})$
4: $\quad \text{tmp} = \text{intersection}(\mathcal{E}'_j, \mathcal{S}_{i-j})$
5: $\quad$ **if** $\text{tmp} \neq \emptyset$ **then** $\mathcal{S}_{i-j} = \text{tmp}$ **end if**
6: **end for**
7: **Return:** $[\mathcal{S}_{i-n_{\text{bwd}}}, \ldots, \mathcal{S}_{i-1}]$

Similarly to Algorithm 1 the function `collect` collects p_{mass} of the obfuscated region $\mathcal{B}_i$ from the observed obfuscated location $\mathbf{z}_i$ (line 1). The for loop (line 2-4) traverses the n_{fwd} past possibilistic regions to update. For each past instant $j \in 1, 2, \ldots, n_{\text{fwd}}$ the `expand` function (line 3) expands the obfuscated region $\mathcal{B}_i$ to cover all possible past locations since time $i - j + 1$ that could have lead to the exposure of $\mathbf{z_i}$ at time i, obtaining the expanded region $\mathcal{E}'_j$. The function `intersect` obtains the intersection between this expanded region and $\mathcal{S}_{i-j}$ (line 4), and updates $\mathcal{S}_{i-j}$ if the intersection is not empty (line 5).

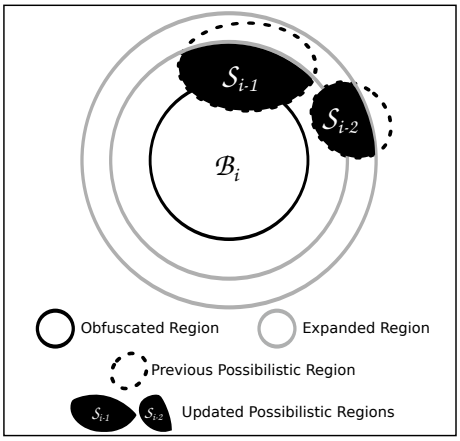


Figure 2: Graphical representation of Algorithm 2 for $n_{\text{fwd}} = 2$. $\mathcal{B}_i$ (black thick line) is the obfuscated region where $\mathbf{x}_i$ can be, given obfuscated location $\mathbf{z}_i$. $\mathcal{S}_{i-1}$ and $\mathcal{S}_{i-2}$ (black dashed lines) are the last possibilistic regions. $\mathcal{E}'_i$ and $\mathcal{E}'_{i-1}$ (grey lines) are the expanded regions where the user could possibly have been in the past, given $\mathcal{B}_i$. Updated possibilistic regions are obtained intersecting $\mathcal{S}_{i-1}$, resp. $\mathcal{S}_{i-2}$, with $\mathcal{E}'_i$, resp. $\mathcal{E}'_{i-1}$.

4 Comparison with Markovian-based Quantification

The main difference between the possibilistic and Markovian [19] approaches lies on the core operation used to infer the users' real location from the obfuscated observations. The possibilistic approach only requires cheap set intersection operations performed on few or several past locations, whereas the probabilistic reasoning behind the Markovian approach relies on expensive statistical operations that require computations over the full surface and all considered time instants. As a consequence of these expensive operations, to be feasible the Markovian approach requires to considerably simplify the scenario being evaluated. This is reflected in the assumptions taken by Shokri *et al* in their framework that, as we discuss below, introduce a trade-off between the quantification accuracy and its runtime.

Assumptions: While both possibilistic and Markovian approaches are similar in that they assume that the adversary has knowledge of the obfuscation algorithm (LPPM) operation, the prior knowledge required to quantify location privacy differs in two key assumptions.

First, in order to be feasible the Markovian approach requires the surface in which location privacy is evaluated to be quantized, and thus this surface needs to be pre-defined. As a result, Markovian-based quantification can only handle locations within such pre-defined "world" and if users visit places that are outside of this world these locations cannot be captured by the quantification framework. The possibilistic approach, on the contrary, does not require the targeted surface to be defined beforehand. Thus, its scope is not limited to any set of pre-defined locations and it can quantify privacy for any location that the user visits.

Second, the Markovian approach requires knowledge of probability distributions that describe users' behavior, which in most cases is hard (if not impossible) to obtain, and the quantification output heavily depends on the quality on this information. On the other hand, the possibilistic approach operates solely on the exposed obfuscated locations provided by the user without the need of any prior knowledge or computation, providing a first-order approximation of the location privacy provided by an obfuscation algorithm without relying on comprehensive behavior models. Since the analysis in Sect. 2 considers knowledge about users' speed v, it is important to note that this parameter does not need to be known beforehand but it could be inferred from the observations. We evaluate in Sect. 5.3 the potential impact of inaccurate speed estimation on the output of possibilistic privacy quantification.

Tradeoff Accuracy vs. Runtime: The need for dividing a surface in discrete regions to perform quantification introduces a tradeoff between the accuracy and the runtime of the Markovian approach. For instance, according to Shokri et al. [19] a localization attack operates in $\mathcal{O}(TM^2)$, where M is the number of regions in which a surface is divided and T the number of time instances in a trace. Thus, to perform highly accurate quantification where the analyst has to consider small regions (resp. time intervals) the surface needs to be divided in a large number of regions incurring a high computational cost. If cost is to be kept small, it is necessary to reduce the number of regions or time intervals. Then, regions become large providing coarse results of reduced utility. The possibilistic approach, on the contrary, works on a continuous domain and its accuracy does not depend on the targeted surface but on the parameters of the location privacy-preserving mechanism under study. Furthermore, the computation time of the two practical algorithms we propose in Sect. 3 is constant for every measurement ($\mathcal{O}(n_{\text{bwd}})$, respectively $\mathcal{O}(n_{\text{fwd}})$). Even if both algorithms are combined to compute the current possibilistic area, and to update past possibilistic areas, the computation complexity would still be constant ($\mathcal{O}(n_{\text{bwd}} + n_{\text{fwd}})$).

We now provide an empirical comparison of the computation time required by

both algorithms. To study the Markovian approach runtime we employ Shokri et al's tool Location-Privacy and Mobility Meter [18]. For our experiment we quantify the location privacy of against a localization attack [19] for every location in 15 random trajectories from the dataset described in Sect. 5.2. We limit the scope of the map to a $20 \times 25\text{km} = 500\text{km}^2$ surface that contains the locations visited in these traces, we set $T = 100$ (comparable to $T = 94$ used by Shokri *et al.*), and we vary M to study its effect on the trade-off. Given that the size of the surface is fixed, the size of the discrete regions is defined by $500\text{km}^2/M$, which in turn determines the accuracy with which the Markovian approach can quantify privacy.

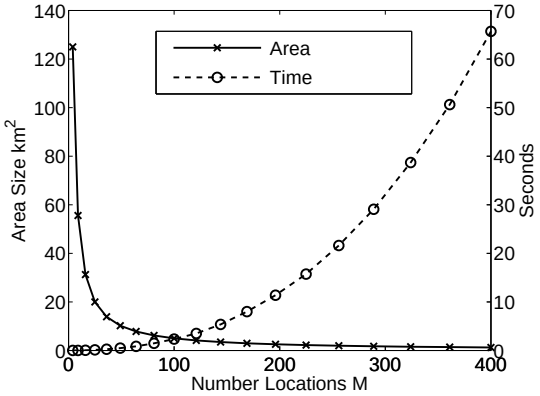


Figure 3: Accuracy, area of a region in km^2, (left y-axis, dashed lines) and Runtime, in seconds, (right y-axis, thick line) vs. number of regions M.

Figure 3 shows the trade-off between runtime and accuracy as the number of regions M increases. As expected, accuracy improves significantly when the surface is divided in more, thus smaller, regions. However, this improvement comes at a high impact in runtime that grows quadratically as predicted by Shokri et al. [19]. These timings are in contrast with those of the possibilistic approach, shown in Table 2, where quantification is performed combining Algorithms 1 and 2 in Sect. 3 implemented using Matlab R2012b.[5] In the best case ($n_{\text{bwd}} = 1$, $n_{\text{fwd}} = 0$) quantification can be done in 0.21 seconds on average, and even for the most requiring combination (20 intersections for $n_{\text{bwd}} = n_{\text{fwd}} = 10$) one localization takes only 0.315 seconds on average. Overall, we see that the possibilistic quantification runtime is very low and can be used for real time computation. More importantly these runtimes are constant, i.e.,

[5]We note that implementing the possibilistic approach in C++, as the LPM^2 tool, would make the runtimes even more competitive.

Table 2: Runtime for one location privacy quantification with the possibilistic approach. (mean [min,max] *in miliseconds*)

		$\boldsymbol{n}_{\text{bwd}} = 1$	$\boldsymbol{n}_{\text{bwd}} = 5$	$\boldsymbol{n}_{\text{bwd}} = 10$
$\boldsymbol{n}_{\text{fwd}}$	0	21 [2, 123]	93 [2, 461]	226 [2, 1419]
	1	40 [3, 192]	113 [4, 517]	250 [4, 1511]
	5	91 [4, 453]	162 [4, 669]	294 [4, 1632]
	10	137 [4, 694]	210 [4, 952]	315 [4, 1542]

they do not depend on the size of the surface in which location privacy is quantified.

The highest accuracy represented in Fig. 3 is 1.25km^2, when $M = 400$. Even for such poor precision, the Markovian quantification requires more than 60 seconds. On the other hand, the possibilistic approach can quantify location with an accuracy of 0.191km^2 and 0.254km^2 for the two obfuscation algorithms we evaluate in the next section at low cost as indicated in Table 2. In order to obtain this accuracy employing the Markovian approach one would have to quantize the area of interest in $M = 2618$, respectively $M = 1969$, regions. This number of regions, two orders of magnitude higher than the largest M in Fig. 3, clearly leads to prohibitive computation times.

Alternatively, one could reduce the targeted surface so as to limit the number of regions to be considered and obtain high accuracy at a reasonable cost. For instance, if $M = 64$ the Markovian approach can compute a localization in 0.8 seconds and can cover a total surface of aprox 12km^2 for a precision of 0.191km^2, or a surface 16km^2 if precision is relaxed to 0.254km^2. Note that if we aim at computation times comparable to those of the posibilistic approach the Markovian approach can only consider $M = 16$ regions, which cover scarcely 4km^2 to obtain 0.254km^2 accuracy. These surfaces are not only small, but limit the range of obfuscation mechanisms that can be studied since it is likely (as in the case of the schemes we study in the next sections) that obfuscated locations fall outside of the considered area. Thus, it is not possible to find an scenario in which a fair comparison between the possibilistic and Markovian approaches in terms of accuracy is feasible in reasonable time.

5 Location Privacy for Geo-Indistinguishability

In this section we provide a case study where we use the possibilistic approach to quantify location privacy. We choose two Geo-Indistinguishability-based approaches as target obfuscation mechanisms, which are an interesting use case

for two reasons. First, they do not rely on discrete regions to operate and hence their evaluation with Markovian-based quantification is cumbersome. In fact, even in simple sporadic cases where only one exposed location has to be evaluated [1, 3] approximations are needed to quantify privacy using Shokri *et al.*'s approach [19, 20]. Either the evaluation is done on a very small target surface (a grid covering approximately $1km^2$), such as Andrés *et al.* in [1]; or is done considering just the top visited regions, such as Bordenabe *et al.* in [3], where the final analysis only covers approximately a 5% of the initial area of interest (50 out of the initially considered 900 locations).

Second, to the best of our knowledge, there exists no quantification of the privacy loss incurred by these mechanisms when users expose correlated geo-indistinguishable locations. Andrés et al. [1] prove that the location privacy provided by a set of correlated geo-indistinguishable points decreases linearly with the number of points; and Chatzikokolakis et al. [5] provide an alternative method that allows to disclose correlated points with sublinear loss of privacy. Yet, the loss incurred with each exposure has not been quantified so far.

The goal of our analysis is twofold: to show the potential of the possibilistic approach as method for quantifying location privacy, and to show its potential as comparison tool for obfuscation-based mechanisms. This comparison complements the results of Chatzikokolakis *et al.* [5], providing a quantitative evaluation of the effectiveness of their method compared to the original mechanism proposed by Andrés *et al.* [1].

5.1 Geo-Indistinguishable Location Obfuscation Mechanisms

Geo-indistinguishability [1] is a location privacy notion based on the extension of the differential privacy concept [7] to arbitrary metrics proposed by Chatzikokolakis et al. [4]. An obfuscation mechanism provides geo-indistinguishability if the probability of reporting obfuscated location $\mathbf{z}$ is similar for two close locations $\mathbf{x}$ and $\mathbf{x}'$, i.e., observing $\mathbf{z}$ does not provide much information to the adversary about which is the actual location. More formally, a mechanism $f(\mathbf{z}|\mathbf{x})$ satisfies ϵ-geo-indistinguishability iff $f(\mathbf{z}|\mathbf{x}) \leq e^{d_2(\mathbf{x},\mathbf{x}')} f(\mathbf{z}|\mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}$, where d_2 is the euclidean distance.

Independent Geo-Indistinguishability

The first mechanism we evaluate is the one proposed in [1], based on the Laplace distribution. At time i, this mechanism draws the exposed obfuscated location $\mathbf{z}_i$ from the polar Laplacian centered at the user real location $\mathbf{x}_i$:

$D_\epsilon(r, \Theta) = \frac{\epsilon^2}{2\pi} r e^{-\epsilon r}$, where r is the distance between $\mathbf{z}_i$ and $\mathbf{x}_i$, and Θ is the angle that the line $\mathbf{z}_i\mathbf{x}_i$ forms with respect to the horizontal axis of the Cartesian system. We note that D_ϵ has rotational symmetry and thus produces circular equiprobable contours.

Predictive Geo-Indistinguishability

The second mechanism we evaluate is the one proposed by Chatzikokolakis *et al.* [5] in which the user may reuse the previous obfuscated location $\mathbf{z}_i$ if her current location is sufficiently close to it instead of drawing a new independent obfuscated location every time instant. The intuition is that re-using locations diminishes the information leakage inherent to subsequent exposures, while there is no quality of service penalty for the user because the user is still near the obfuscated location.

A predictive geo-indistinguishable mechanism consists of three components: i) a prediction function $\Omega : \mathcal{X} \to \mathcal{Z}$, that takes as input the previous locations $\mathbf{X}_i$ and outputs a prediction $\tilde{z}$ for the next obfuscated location; ii) a test function $\Theta(\epsilon_\theta, l_\theta, \tilde{z}) : \mathcal{X} \to \{0, 1\}$ that takes as input the current location $\mathbf{x}_i$ and returns 1 if the prediction is accurate and can be re-used (i.e., $d_2(\mathbf{x}_i, \tilde{z}) \leq l + Lap(\epsilon_\Theta)$, being l and ϵ_Θ parameters that represent utility), 0 otherwise; and iii) a noise function $N(\varepsilon_N) : \mathcal{X} \to \mathcal{Z}$ that is used if a new pseudo-location needs to be computed, parameterized by the budget ϵ_N. When the budget is exhausted the predictive mechanism stops exposing obfuscated locations.

Other Differentially-Private Mechanisms

Xiao *et al.* [22] note that by observing a series of geo-indistinguishable correlated locations, an adversary may be able to infer the user's trajectory or the user's destination. As a solution, they propose the *Planar Independent Mechanism* that provides differential privacy on a so-called δ*-location set*, that reflects a set of probable locations that accumulate at least 1-δ of the probability that a user might appear given his last location and a Markovian mobility model (i.e., leaves out locations with low probability of being the real one). This mechanism obtains geo-indistinguishability offering better utility for the user, but like [19] relies on a Markovian approach to obtain the probabilities that are used in the construction of the δ-location, i.e., the protection mechanism is constrained in the surface it can cover and its computational complexity grows as shown in Sect. 4. Thus, obtaining obfuscated regions becomes too expensive to include this approach in our quantification experiments.

Liu *et al.* [13] note that the protection of differential privacy may be reduced if the obfuscated data points contain correlations. They propose the notion of *dependent differential privacy* in order to formalize probabilistic dependence constraints and provide a mechanism, the Dependent Perturbation Mechanism that achieves it. While their use case is related to location data, and they use the same data set as we do, their goal is to protect the social graph and hence their mechanism does not provide location privacy.

5.2 Experimental Setup

We use the GeoLife GPS[6] data set [23–25] for our evaluation. This data set includes 18 655 trajectories from 182 users in the urban area of Beijing, China collected during a period of over three years (from April 2007 to August 2012). The GeoLife data set illustrates many types of mobility, from daily routines to leisure activities, and contains traces collected in different means of transportation. Throughout our experiments we use the actual users' speed to expand possible regions, and study the effect of an incorrect speed estimation in Sect. 5.3.

Trajectories in the GeoLife data set are recorded at a variety of sampling rates. In order to obtain a regular data set that allows us to fairly test the influence of different parameters on the possibilistic quantification algorithms, we apply the following pre-processing. First, we remove trajectories containing large gaps between locations exposures caused, for example, by the user entering a subway, or by deliberate/accidental halts in the operation the GPS receiver. Thend, we only conserve trajectories that contain at least 100 consecutive locations, and where the average time between exposure of two consecutive locations is less than 2 minutes. Larger intervals would cause the intersection between subsequent exposures to be the empty set, and the possibilistic approach would trivially outputs the size of the observed obfuscated area, i.e., $\phi = |\mathcal{S}| = |\mathcal{B}|$. We are left with trajectories from 37 different users.

Second, for each of these users we select 10 trajectories at random and we quantize time to the minute (much higher exposure frequency than Shokri *et al.* [19]). While the possibilistic approach could handle more frequent locations, this quantization ensures that in most cases the user moves between subsequent exposures providing information that can be quantified in our evaluation. For each trajectory we select location samples that are separated at least $60 + \delta$ seconds, where δ is a small gaussian noise. We repeat the process 10 times per considered trajectory, ending up with a total of 33,681 locations with on

[6] `http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/`

average 91 locations per trajectory. The surface covered by these trajectories is $6.5M\text{km}^2$

To ensure that we do not present a favorable case for our approach regarding frequency of exposures we introduce the parameter p_{jump} to create controlled gaps in the traces. This parameter allows us to evaluate the influence of sporadic vs. continuous location exposures on the possibilistic location privacy quantification. Similarly to [5] the parameter p_{jump} denotes the probability that the user performs a "jump", i.e., that the user does not expose her location within an hour. Naturally, the higher this probability, the more sporadic the user's location exposures are. We evaluate the impact of sporadicity in exposures in Sect. 5.3.

Geo-Indistinguishability setup for the experiments. The main parameter to be chosen for both mechanisms is ϵ. We select $\epsilon = \{0.1, 0.01, 0.001\}$ which results in the real location of the user being in a circular area of radius $r = \frac{\ln(2)}{\epsilon} = \{7m, 70m, 700m\}$ centered at in the obfuscated location [6]. We believe that these choices adequately represent user preferences for low, medium and high privacy protection.

For the predictive mechanism we adopt the functions proposed in [5]: the test and noise functions described above; and the *parrot function* as prediction function Ω, that always predicts the last reported location as the next obfuscated location. For the sake of simplicity we choose to disregard the details of the budget managers in [5], i.e., we consider that the user always has enough budget to produce a new estimated location. This allows us to consider the effect of long trajectories in our experiments without the user stopping using the service because of lack of budget. Finally, we choose $l_\theta = 250m$, to decide when a prediction is *too far away* from the real location.

5.3 Results

The Influence of p_{mass} on Possibilistic Location Privacy Quantification

We first study the effect of p_{mass}, defined in Sect. 3, on the quantification of location privacy. This parameter determines how much of the obfuscated region described by $\mathbf{z}_i + \mathcal{Z}_i$ will be taken into account by our implementation of the possibilistic approach, i.e., how large $\mathcal{B}_i$ is. Since the two considered obfuscation schemes draw their noise from a Laplace distribution (CDF: $C_\epsilon(r') = 1 - (1+\epsilon)e^{-\epsilon r}$), for a target collected p_{mass} the obfuscated area $\mathcal{B}_i$ shall be a

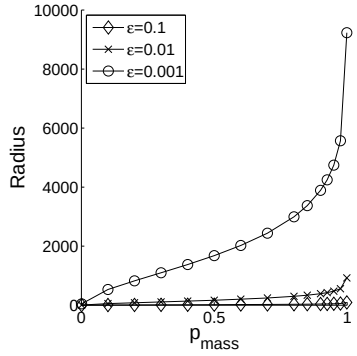
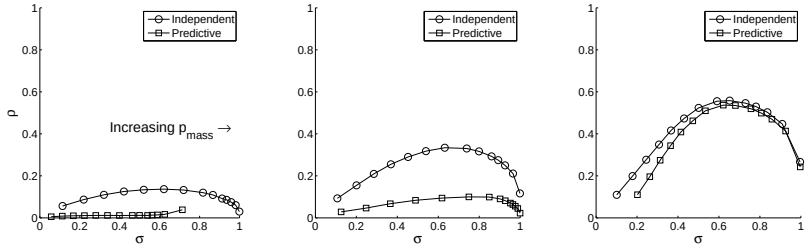


Figure 4: Radius (in meters) of the circular obfuscated area $\mathcal{B}_i$ depending on p_{mass}.



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.01$ (c) $\epsilon = 0.001$

Figure 5: Certainty gain and adversary success with respect to p_{mass} and privacy level ϵ. (p_{jump}=0.1, $n_{\text{bwd}} = 1$, $n_{\text{fwd}} = 0$, $l_\theta = 250m$)

circle of radius $r' = (ln(1+\epsilon) - ln(1-p_{\text{mass}}))/\epsilon$. We plot in Figure 4 the effect of p_{mass} on r' for the three privacy levels ϵ considered in our experiments.

Figure 5 shows the impact of p_{mass} on both the certainty gain ρ, and the adversary success σ for different privacy levels ϵ. Each point represents a different p_{mass}, and ρ and σ are averaged over all trajectories in the dataset truncated at location number 100. We choose to represent ρ in all figures in the results section since combined with Fig. 4 it allows to easily estimate the possibilistic area size ϕ to quantify absolute privacy, and is more convenient for the comparison between the two geo-indistinguishability approaches.

Since it determines the area covered by $\mathcal{B}_i$, the parameter p_{mass} has an impact on the three aspects of privacy captured by our metrics (see Sect. 2.3). Unsurprisingly, increasing p_{mass} monotonically increases the success σ (the

larger the region $\mathcal{B}_i$, the more chances $\mathcal{S}_i$ contains the true location). The remarkable low success for the Predictive mechanism when $\epsilon = 0.1$ stems from our choice of l_θ, as we clarify in the next section.

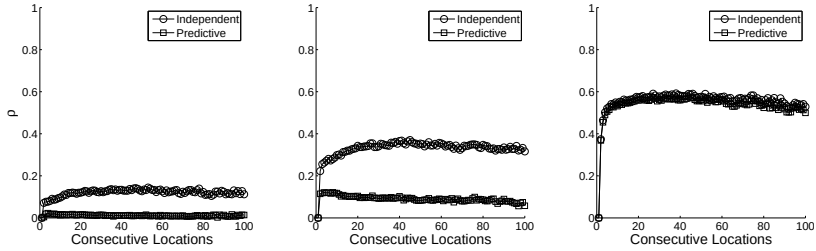
With respect the certainty gain ρ and the possibilistic area size ϕ, very small p_{mass} values lead to (most of the times) no intersection between $\mathcal{B}_i$ and previous expanded regions $\mathcal{E}$ resulting in $\phi = |\mathcal{B}_i|$ and little certainty gain on average. As p_{mass} increases so does $\mathcal{B}_i$, and it creates a non-empy intersection with $\mathcal{E}$ increasing the certainty gain. At some point, $\mathcal{B}_i$ becomes so large that its intersection with $\mathcal{E}_{i-1}$ grows so much that the certainty gains decreases again. This effect is more visible for small values of ϵ that produce larget obfuscated regions $\mathcal{B}_i$, see Fig. 5c.

The best result is obtained when p_{mass} is between 0.7 and 0.8. For the Independent geo-indistinguishability mechanism the certainty gain is around $18\%, 37\%, 58\%$ obtaining average accuracy of $\phi = 0.0021, 0.15, 10.72$ km^2 for $\epsilon = 0.1, 0.01, 0.001$, respectively; and for the Predictive geo-indistinguishability mechanism the gain is reduced to $8\%, 12\%, 56\%$ and $\phi = 0.0023, 0.21, 11.11$ km^2 on average. As expected, larger ϵ's (i.e., lower privacy) result in smaller certainty gain than smaller ϵ's (i.e., higher privacy), because when ϵ is large obfuscated regions are very small, and thus there is not much room for improvement.

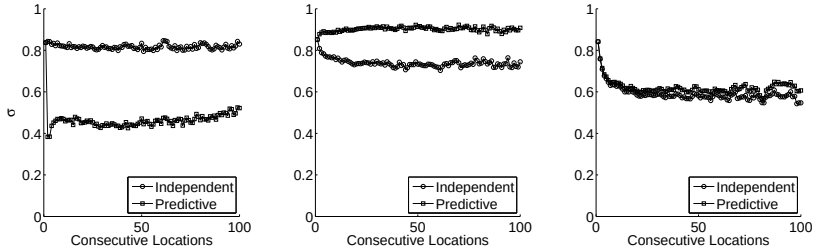
We observe that the Predictive scheme generally provides better protection than the Independent method in terms of certainty gain, since by repeating obfuscated locations it avoids leaking information that can be used by the adversary to improve her inferences about the user whereabouts. However, this advantage is reduced as the privacy level is higher. When ϵ decreases, the distance r' between the real and obfuscated locations increases (see Fig. 4) and, since we configure the Predictive mechanism to consider locations far away when they are further apart more than $l_\theta = 250$m, increasingly often the test function returns false and the mechanism produces a new obfuscated location. In other words, the smaller ϵ the less obfuscated locations are reused, and the more similar are the Predictive and Independent mechanisms. We note that this effect only happens because we allow a large budget for the predictive mechanism, otherwise after few exposures the Predictive mechanism would stop producing new obfuscations offering high privacy but poor quality of service for the users.

Privacy Evolution as the User Moves

The results in the previous section represent an average privacy quantification for the first 100 exposed obfuscated locations. We now show the evolution of privacy while the user is traveling (averaged over all users for each location).



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.01$ (c) $\epsilon = 0.001$



(d) $\epsilon = 0.1$ (e) $\epsilon = 0.01$ (f) $\epsilon = 0.001$

Figure 6: Evolution of certainty gain ρ and success σ as the user moves for different privacy levels (ϵ). ($p_{\text{jump}} = 0.1$, p_{mass}=0.8, $n_{\text{bwd}} = 1$, $n_{\text{fwd}} = 0$, $l_\theta = 250m$)

Figure 6, top, shows the evolution of the certainty gain. We see that in the beginning of the trace there is a significant growth in certainty gain, e.g., for $\epsilon = 0.01$ ϕ is reduced from 0.2817km^2 to 0.1860km^2 after 30 exposures, and after this point the reduction follows a diminishing returns trend. Also, as in the previous experiment, we observe that in terms of gain the Predictive defense performs better than the Independent scheme, with the advantage being reduced as ϵ decreases.

Regarding success, we observe that for the predictive mechanism ϵ seems to have great influence: an adversary would experience particularly good results in terms of correctness for $\epsilon = 0.01$, and much worse results for the other choices of privacy parameter. The reason is the choice of $l_\theta = 250m$, which means that obfuscated locations will be reused until the user moves more than 250m away from her location. This has the following effects:

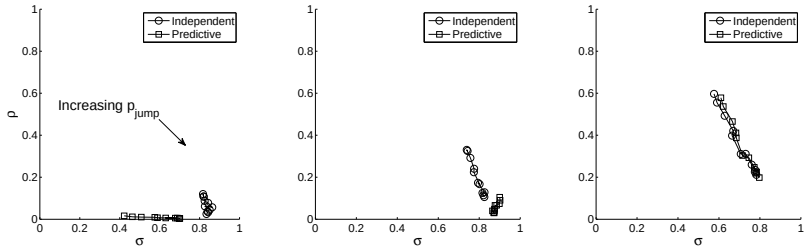
When $\epsilon = 0.1$ the obfuscated region radius is 29.94m (see Fig. 4). Since an obfuscated location will be reused until the user traverses 250m it is easy to see that in this case the real location will soon fall outside the reused exposed

location $\mathcal{B}_i$ and hence also outside the possibilistic region $\mathcal{S}_i$, resulting in low success.

When $\epsilon = 0.01$, the radius of the obfuscated region is 299.43m, very similar to l_θ. This means that either the user reuses her location because she did not move 250m, and naturally her real location is inside $\mathcal{B}_i$; or the user moves producing a new obfuscated location. In the latter case, due to the operation of the predictive mechanism, it is very likely that the intersection of the new obfuscated region and the last possibilistic region ($\mathcal{B}_i \bigcap \mathcal{S}_{i-1}$) contains the real location causing the average high success rate.

Finally, when $\epsilon = 0.001$ the Predictive becomes like the Independent scheme (i.e., it produces new obfuscated locations even if the previous obfuscated location would be useful) reducing the likelihood that the real location is in the intersection, and thus decreasing the adversary's success.

In this case the possibilistic approach reveals a property of the Predictive mechanism not considered in [5]. While in principle $\epsilon = 0.1$ provides little privacy in terms of accuracy and certainty, if exposures are very frequent and l_θ is chosen wisely the adversary's inference is rarely correct and thus privacy increases. Hence, this mechanism may be very useful to hide movements along nearby locations, e.g., inside a shopping mall, and at the same time have good utility. Similarly, we see that large ϵ not only provides high privacy in terms of adversary's accuracy and uncertainty, but also in that it provides low adversary's success. We also uncover the fact that choosing l_θ close to $r' = (ln(1+\epsilon) - ln(1-p_{\text{mass}}))/\epsilon$ has a bad effect on privacy because the adversary almost always chooses a region containing the user´s real position.



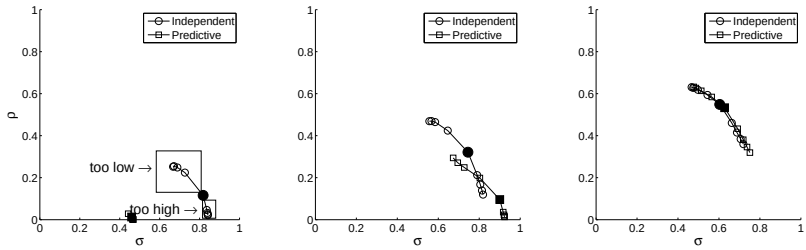
(a) $\epsilon = 0.1$ (b) $\epsilon = 0.01$ (c) $\epsilon = 0.001$

Figure 7: Certainty gain and adversary success with respect to p_{jump} and privacy level, ϵ. (p_{mass}=0.8, $n_{\text{bwd}} = 1$, $n_{\text{fwd}} = 0$, $l_\theta = 250m$)

Influence of Sporadicity on Privacy

In the previous sections we assumed users to have a rather continuous LBS querying behavior. We now study the effect of query frequency on the the possibilistic quantification of privacy by varying the parameter p_{jump} from 0 (very frequent usage of LBS) to 1 (sporadic usage of LBS). Figure 7 shows the results of the experiment. Unsurprisingly, the more sporadic the usage pattern, the more privacy the users enjoy both in terms of accuracy/certainty gain and adversary success. We observe that the larger p_{jump} the smaller the certainty gain: every time there is a silence it is highly likely that the intersection $\mathcal{S}_i = \emptyset$, and hence no reduction on the obfuscated area is achieved. Also, since the obfuscated region $\mathcal{B}_i$ is rarely reduced, the real location often falls inside the possibilistic region $\mathcal{S}_i$ and the adversary's success increases stabilizing around 0.8 since this is the collected probability mass of $\mathcal{B}_i$.

Influence of Speed Estimation on Privacy



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.01$ (c) $\epsilon = 0.001$

Figure 8: Certainty gain and adversary success with respect to speed and privacy level, ϵ. (p_{mass}=0.8, p_{jump}=0.1, $n_{\text{bwd}} = 1$, $n_{\text{fwd}} = 0$, $l_\theta = 250m$)

A key parameter for the possibilistic approach is the user's speed estimation, which is used by the `expand` function in Algorithms 1 and 2. Intuitively, underestimating speed results in smaller $\mathcal{E}_i$, which in turns reduces the size of the intersection with $\mathcal{B}_i$, thus increasing accuracy but decreasing success since it becomes less likely that the user's actual location is included in S_i. Conversely, overestimating speed results in larger $\mathcal{S}_i$, increasing the likelihood that the real location falls within this region but decreasing accuracy. Figure 8 confirms that this is the case for both obfuscation mechanisms. This figure also reinforces the findings in previous sections: When $\epsilon = 0.1$ the Predictive mechanism offers great accuracy but very low correctness because of the common

reuse of small obfuscated locations, for $\epsilon = 0.01$ the Predictive location offers worse performance in terms of correctness than the Independent because the obfuscated region has similar radius as l_θ, and when ϵ is very small both algorithms perform similarly.

Influence of Considered Locations on Privacy

As discussed in Section 3 one can use multiple possibilistic areas using the parameters n_{bwd} and n_{fwd} at the cost of increasing the computational overhead of the algorithm, Sect. 4. As it turns out, varying these parameters leads to a trade off between certainty gain (and hence accuracy) and correctness: increasing either n_{bwd} or n_{fwd} on average reduces the size of the possibilistic region $\mathcal{S}_i$, which in turn increases the certainty gain. However, we find that considering more information also leads to a decrease in the adversary's success σ due to two reasons: i) reducing $\mathcal{S}_i$ decreases the likelihood that the real location is in the possibilistic region; and ii) considering more regions entails the use of more region expansions computed assuming a speed estimation (see function `expand` in Section 3), and the more assumptions, the easier it is that the possibilistic region deviates with respect to the actual location of the user.

In the extreme case, for for $\epsilon = 0.001$ and $n_{\text{bwd}} = n_{\text{fwd}} = 10$, the certainty gain ρ can be increased up to 81%, resp. 79%, at the cost of reducing success to 36%, resp. 40% for the Independent and Predictive schemes. While this is a high penalty, there are more convenient combinations, e.g., $n_{\text{bwd}} = n_{\text{fwd}} = 1$ provide a certainty gain of 41%, while only reducing success to 71%, for the Independent mechanism when $\epsilon = 0.01$. More combinations can be found in in Table 3 in Appendix B.

6 Conclusion and Future work

Shokri et al. state-of-the art framework [19] for quantifying location privacy employs a Markovian approach. This approach can capture a great variety of user behaviors, inference attacks, and allows to easily integrate adversarial knowledge in the quantification of privacy. On the downside, it relies on expensive probabilistic reasoning which requires huge computational power that grows quadratically with the number of locations to be considered. As a result, it can either be used to quantify privacy on limited surfaces, or to produce coarse quantification results reducing the utility of the evaluation.

In this paper we have proposed a possibilistic approach for location privacy quantification. This approach is based on finding the possible region where

the user can be located given subsequent obfuscated exposures by using set intersection operations. The simplicity of the approach results in constant computational complexity, and allows to obtain much more accurate results than the Markovian approach at a very low cost.

We have illustrated the capability of the possibilistic approach as quantification mechanism by evaluating the privacy offered by two geo-indistinguishability based schemes [1,5] using real data. Our evaluation provides the first quantitative estimation of the privacy offered by these mechanisms when they are continuously used, and uncovers unknown tradeoffs regarding the configuration of the mechanism proposed in [5]. Furthermore, the efficiency of the possibilistic approach allows us to carry our evaluation in a much larger surface than any prior location privacy evaluation, 6.5M km^2, demonstrating that the possibilistic analysis is suitable for large-scale location privacy evaluation.

Our work has demonstrated that the possibilistic approach can quantify location privacy at a low cost. This opens new research directions regarding the evaluation and design of location privacy-preserving mechanisms. First, while the possibilistic approach can provide reasonable accuracy, it is clear that its binary nature (possible or not possible) disregards any probabilistic information that may be available to the adversary and that could refine the quantification result. Thus, one future line of research is the development of methods that use the possibilistic approach as means to reduce the surface in which probabilistic operations need to be performed such that the latter becomes feasible in realistic scenarios.

Second, it has been shown that the Markovian approach can be used in the design optimal protection mechanisms against strategic adversaries [11,17,20,21]. However, the computational cost of these methods is very high mainly due to the probability computations, and thus they have only been evaluated in small scenarios. Hence, another interesting research direction is to explore ways to integrate possibilistic reasoning in the optimization processes underlying these design strategies to reduce their computational needs so that they can scale to large surfaces.

References

[1] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM Conference on Computer and Communications Security (CCS)*, pages 901–914. ACM, 2013.

[2] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag TELOS, 3rd ed. edition, 2008.

[3] Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Optimal Geo-Indistinguishable Mechanisms for Location Privacy. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security (CCS)*, pages 251–262. ACM, 2014.

[4] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the Scope of Differential Privacy Using Metrics. In *13th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 7981 of *Lecture Notes in Computer Science*, pages 82–102. Springer Berlin Heidelberg, 2013.

[5] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. A Predictive Differentially-Private Mechanism for Mobility Traces. In *14th International Symposium on Privacy Enhancing Technologies Symposium (PETS)*, volume 8555 of *Lecture Notes in Computer Science*, pages 21–41. Springer Berlin Heidelberg, 2014.

[6] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. Constructing Elastic Distinguishability Metrics for Location Privacy. volume 2015, pages 156–170, 2015.

[7] Cynthia Dwork. Differential Privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *33rd International Colloquium on Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[8] Gabriel Ghinita, Maria Luisa Damiani, Claudio Silvestri, and Elisa Bertino. Preventing Velocity-based Linkage Attacks in Location-aware Applications. In *17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems (ACM-GIS)*, pages 246–255, 2009.

[9] Marco Gruteser and Dirk Grunwald. Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In *Proceedings of the 1st International Conference on Mobile systems, Applications and Services (MobiSys)*, pages 31–42. ACM, 2003.

[10] D. Henrion, S. Tarbouriech, and D. Arzelier. LMI Approximations for the Radius of the Intersection of Ellipsoids: Survey. *Journal of Optimization Theory and Applications*, 108(1):1–28, 2001.

[11] Michael Herrmann, Carmela Troncoso, Claudia Diaz, and Bart Preneel. Optimal Sporadic Location Privacy Preserving Systems in Presence of

Bandwidth Constraints. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society (WPES)*, pages 167–178. ACM, 2013.

[12] Leonid G Khachiyan. Rounding of Polytopes in the Real Number Model of Computation. *Mathematics of Operations Research*, 21(2):307–320, 1996.

[13] Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples. In *Proceedings of the Network & Distributed System Security Symposium (NDSS)*, pages 1–15. Internet Society, 2016.

[14] Joseph T. Meyerowitz and Romit Roy Choudhury. Hiding Stars with Fireworks: Location Privacy through Camouflage. In *15th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pages 345–356. ACM, 2009.

[15] Nima Moshtagh. Minimum Volume Enclosing Ellipsoid. *Convex Optimization*, 111:112.

[16] Jean Serra. *Image Analysis and Mathematical Morphology*. Academic Press, Inc., 1983.

[17] Reza Shokri. Privacy games: Optimal user-centric data obfuscation. *PoPETs*, 2015(2):299–315, 2015.

[18] Reza Shokri, Vincent Bindschaedler, George Theodorakopoulos, George Danezis, Jean-Pierre Hubaux, and Jean-Yves Le BoudecNathan Carter. Location-Privacy and Mobility Meter (LPM^2): A Tool to Model Human Mobility and Quantify Location Privacy. `http://icapeople.epfl.ch/rshokri/lpm/doc/`, 2011. [Online – Last accessed 29 February 2016].

[19] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying Location Privacy. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy (S&P)*, pages 247–262. IEEE, 2011.

[20] Reza Shokri, George Theodorakopoulos, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Protecting Location Privacy: Optimal Strategy Against Localization Attacks. In *Proceedings of the 19th ACM Conference on Computer and Communications Security (CCS)*, pages 617–627. ACM, 2012.

[21] George Theodorakopoulos, Reza Shokri, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. Prolonging the Hide-and-Seek Game: Optimal Trajectory Privacy for Location-Based Services. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES)*, pages 73–82. ACM, 2014.

[22] Yonghui Xiao and Li Xiong. Protecting Locations with Differential Privacy under Temporal Correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1298–1309, 2015.

[23] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding Mobility Based on GPS Data. In *10th International Conference on Ubiquitous Computing (UbiComp)*, pages 312–321, 2008.

[24] Yu Zheng, Xing Xie, and Wei-Ying Ma. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[25] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining Interesting Locations and Travel Sequences From GPS Trajectories. In *Proceedings of the 18th International Conference on World Wide Web (WWW)*, pages 791–800, 2009.

A Implementation Considerations

Next we describe the implementation of the functions `intersect` and `expand` of the Algorithms 1 and 2 that we use for our evaluation in the following sections. Describing analytically and accurately the intersection between two or more regions is not feasible, since in most cases the resulting structure is expected to be an arbitrary figure. To ensure reasonable computation time and easy maintenance we represent the intersection as the minimal overbounding ellipse. Finding the minimal bounding ellipsoid is an NP-hard problem [10], hence we use *Khachiyan's algorithm* [12] to efficiently find the minimum bounding ellipse enclosing a set of points.

Taking this into account we represent the ellipse as tuple $[A, C]$ [15], where A is a two by two matrix that encodes the ellipse's radii lengths a, b and rotation angle ϕ of the ellipse (with respect to the considered reference axes), and C are the coordinates of the ellipse. See Appendix A.1 for more information about the ellipse encoding and how to compute it from particular ellipse's radii and rotation angle.

The function `expand` can be implemented as follows: we obtain the radii a, b of a region $\mathcal{B}_i$ (or $\mathcal{S}_i$) from the matrix A and extend them according to $v \cdot (t_i - t_{i-j+1})$, where v is the user's estimated speed. Then, we invert the transformation to go back to the matrix representation.

To illustrate our implementation of the algorithm `intersect` we provide an example for the intersection of two regions but note that the general case is straightforward. We represent the corresponding obfuscated region $\mathcal{B}_i$ and the expanded regions $\mathcal{E}_i$ with two ellipses given by $[A_\mathcal{B}, C_\mathcal{B}]$ and $[A_\mathcal{E}, C_\mathcal{E}]$. $A_\mathcal{B}$ and $A_\mathcal{E}$ are given by the user's choice of the privacy parameter for the chosen obfuscation algorithm and the parameter p_{mass}, and $C_\mathcal{B}$ and $C_\mathcal{E}$ are given by $\mathbf{z}_1, \mathbf{z}_2$. To compute the intersection, we approximate both ellipses selecting a varying number of points in its countour (i.e., points that fulfil the equality $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$). We choose the number of points depending on the size of the region (at most $1,500$). After the approximation, we rotate all the points according to $\phi_\mathcal{B}$, $\phi_\mathcal{E}$ and move the points according to $C_\mathcal{B}$ and $C_\mathcal{E}$ and keep only those points that are inside both ellipses. Finally we use Khachiyan's algorithm to find the minimum bounding ellipse containing these points.

The memory consumption of the `expand` and `intersect` functions is very low. Our implementation of the `expand` algorithm consumes 32 floats and thus a memory of 256 byte. The memory consumption of the `intersect` algorithm depends on the number of ellipses to intersect, i.e. n_{bwd} for Algorithm 1 and two for Algorithm 2. Our implementation of the `intersect` algorithm consumes for two ellipses 71 kB (6 times the maximum number of approximation points, i.e. $1,500$ floats) and 24 kB (2 times $1,500$ floats) for every additional ellipse.

A.1 Matrix Representation

In the following we will provide the formulas for computing the matrix representation A of an ellipse in 2D coordinates given the ellipse's radii a, b and its rotation angle ϕ. The computation of A is the multiplication of three matrices U, D, V, every of dimension 2x2:

$$U = V = \begin{pmatrix} -\cos\phi & \sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$$

$$D = \begin{pmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{pmatrix}$$

$$A = U \times D \times V$$

For obtaining the ellipse's radii a, b and rotation angle ϕ from A we need to apply the singular value decomposition (svd) of A. This gives us the three

matrices U, D, V and the values for a, b, ϕ can then be computed as follows:

$$[U, D, V] = \text{svd}(A)$$

$$a = \frac{1}{\sqrt{D_{11}}}, \quad b = \frac{1}{\sqrt{D_{22}}}$$

$$\phi = \begin{cases} \pi - \arcsin(U_{12}) & \text{for } U_{22} < 0 \\ \arcsin(U_{12}) & \text{otherwise} \end{cases}$$

B Influence of Number of Considered Locations on Privacy

Table 3 shows the evolution of the accuracy and success of the possibilistic approach as more past and/or future observations are considered. Note that $n_{\text{bwd}} = n_{\text{fwd}} = 0$ represents the case where no prior/posterior observation is intersected with region $\mathcal{B}_i = \text{collect}(\mathbf{z}_i, p_{\text{mass}} = 0.8)$, i.e., it represents the maximum protection provided by the obfuscation mechanism that serves as basis for the computation of the certainty gain ρ. For this case, obviously the certainty gain is 0, and the success is around 80% since it is given by the proportion of total feasible locations collected given p_{mass}.

As explained in Sect. 5.3, increasing the number of considered possibilistic regions trades-off accuracy for correctness. An exception is the case of the Predictive algorithm when $\epsilon = 0.01$ where, due to to the choice of l_θ, success increases with respect to the baseline case (see Sect. 5.3 for more details). However we observe that, even in this case success decreases as either n_{bwd} or $n_{\text{fwd}} = 0$ increases.

Table 3: Improvement with n_{bwd} and n_{fwd}. Each element of the table represents a tuple (ρ/σ). ($p_{\text{mass}} = 0.8$, p_{jump}=0.1).

Independent		n_{fwd}			
		0	**1**	**5**	**10**
$\epsilon = 0.1$	n_{bwd} **0**	0.00 / 80.00	9.59 / 82.89	12.83 / 81.53	13.06 / 81.32
	n_{bwd} **1**	12.01 / 81.72	17.72 / 80.34	20.12 / 79.14	20.29 / 78.97
	n_{bwd} **5**	13.84 / 80.94	19.46 / 79.53	21.70 / 78.39	21.87 / 78.21
	n_{bwd} **10**	14.38 / 80.73	19.94 / 79.32	22.15 / 78.15	22.31 / 77.98
$\epsilon = 0.01$	n_{bwd} **0**	0.00 / 80.00	22.72 / 79.87	35.05 / 73.57	36.94 / 71.82
	n_{bwd} **1**	33.04 / 73.92	41.24 / 71.19	48.15 / 66.54	49.28 / 65.26
	n_{bwd} **5**	38.99 / 69.96	46.26 / 67.28	46.26 / 67.28	52.85 / 61.95
	n_{bwd} **10**	40.31 / 69.09	47.28 / 66.43	52.69 / 62.32	54.82 / 60.83
$\epsilon = 0.001$	n_{bwd} **0**	0.00 / 80.00	34.46 / 76.54	58.01 / 62.36	63.18 / 55.91
	n_{bwd} **1**	55.53 / 59.07	65.04 / 55.70	74.50 / 48.08	76.99 / 44.29
	n_{bwd} **5**	64.67 / 49.83	72.03 / 47.03	78.94 / 41.26	80.84 / 38.43
	n_{bwd} **10**	67.82 / 46.46	74.15 / 43.76	80.30 / 38.57	81.97 / 36.04

Predictive		n_{fwd}			
		0	**1**	**5**	**10**
$\epsilon = 0.1$	n_{bwd} **0**	0.00 / 80.00	0.98 / 46.96	4.16 / 46.87	4.93 / 46.83
	n_{bwd} **1**	1.16 / 46.34	2.15 / 46.29	5.32 / 46.19	6.10 / 46.17
	n_{bwd} **5**	4.75 / 44.57	5.71 / 44.53	8.72 / 44.42	9.47 / 44.40
	n_{bwd} **10**	5.82 / 44.08	6.75 / 44.04	9.70 / 43.96	10.44 / 43.94
$\epsilon = 0.01$	n_{bwd} **0**	0.00 / 80.00	5.20 / 92.22	8.39 / 91.14	8.95 / 90.85
	n_{bwd} **1**	9.03 / 90.18	12.64 / 89.70	15.73 / 88.68	16.24 / 88.41
	n_{bwd} **5**	11.04 / 89.12	14.63 / 88.64	17.61 / 87.68	18.10 / 87.43
	n_{bwd} **10**	11.50 / 88.90	15.11 / 88.45	18.09 / 87.49	18.57 / 87.24
$\epsilon = 0.001$	n_{bwd} **0**	0.00 / 80.00	29.12 / 79.87	49.07 / 68.12	54.02 / 62.81
	n_{bwd} **1**	53.67 / 62.12	62.14 / 59.01	70.73 / 52.19	73.09 / 48.99
	n_{bwd} **5**	62.68 / 53.67	69.36 / 51.19	75.82 / 45.52	77.70 / 43.01
	n_{bwd} **10**	65.47 / 50.90	71.45 / 48.28	77.38 / 43.08	79.04 / 40.64

Curriculum

Michael Herrmann was born on July 16th, 1982 in Bensheim, Germany. In September 2008, he received his Bachelor's degree in Computer Science from the Technische Universität Darmstadt. In March 2011, he received his Master's degree also in Computer Science from the Technische Universität München (TUM). His master thesis was written under the supervision of Christian Grothoff at the Emmy-Noether research group at the chair of Network Architectures and Services at TUM.

After working for six months as research assistant at the chair of Network Architectures and Services at TUM, he started a PhD program at the Computer Security and Industrial Cryptography (COSIC) group of the Department of Electrical Engineering (ESAT) of KU Leuven. His PhD research was sponsored by the Fonds Wetenschappelijk Onderzoek (FWO) as a member of the project *Contextual Privacy and the Proliferation of Location Data*.

He visited the Department of Signal Theory and Communications of the University of Vigo in Spain from February 2013 till May 2013 and a second time from January 2015 - April 2015, where he collaborated with Professor Fernando Pérez-González and Dr. Carmela Troncoso. He visited the Law, Science, Technology & Society (LSTS) research group of the Vrije Universiteit Brussel under the supervision of Professor Mireille Hildebrandt. He was also a intern at Inria Saclay - Ile de France from May 2015 - October 2015 under the supervision of Dr. Aline Carneiro Viana.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
COSIC
Kasteelpark Arenberg 10, bus 2452
B-3001 Leuven
Michael.Herrmann@esat.kuleuven.be
http://www.esat.kuleuven.be/cosic/