Developing Advanced Privacy Protection Mechanisms for Connected Automotive User Experiences

by

Huaxin Li

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
(Computer and Information Science)
in the University of Michigan-Dearborn
2018

Master's Thesis Committee:

Associate Professor Di Ma, Chair
Associate Professor Brahim Medjahed
Professor Qiang Zhu

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Di Ma for the continuous support of my master study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my master study.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Brahim Medjahed, and Dr. Qiang Zhu for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

Last but not least, my deepest gratitude goes to my family for their never ending support and love.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **ii**

LIST OF FIGURES **v**

LIST OF TABLES **vi**

ABSTRACT **viii**

Chapter 1 Introduction **1**

1.1 Thesis Contribution 3

1.2 Thesis Organization 4

Chapter 2 Related Works **5**

Chapter 3 Preliminaries **7**

3.1 Connected Vehicle Landscape 7

3.2 Data Collection in Connected Vehicles 10

3.3 Data Sharing in Connected Vehicles 11

Chapter 4 Privacy Analyses and Requirements in Connected Vehicles Applications **13**

4.1 Case Study of Privacy Leakage 13

4.2 Privacy Risks and Implications in Connected Vehicles 15

4.2.1 Personal Identity Leakage 15

4.2.2 Location Privacy 16

4.2.3 Abundant Information Inference 17

4.2.4 Privacy Leakage through Data Aggregation 17

4.3 Traditional Privacy Protection Method 18

4.3.1 Protecting Personal Identifiable Information 19

4.3.2 Protecting Personal Publishable Information 20

Chapter 5 Secure and Privacy-preserving Data Collection Mechanisms for Connected Vehicles 22

5.1 Design Principles and Goals 22

5.2 Technical Approach 23

5.2.1 Characterize Multi-modal Data 23

5.2.2 Privacy Preserving Data Uploading 28

5.3 Prototype Implementation 30

5.3.1 Implementation of Data collection 30

5.3.2 Implementation of Data Uploading 32

Chapter 6 Privacy Preserving Data Sharing in Connected Vehicle Services 35

6.1 Attacker Model 37

6.2 Privacy Risk of Top Locations Inference 38

6.2.1 Problem Modeling 39

6.2.2 Inference Algorithm 40

6.3 Privacy risk by de-anonymization 41

6.4 Countermeasures Methodology 42

6.4.1 Defending Inference Attacks in Location Sharing 42

6.4.2 Defending Fingerprinting Attacks 45

6.5 Experimental Results 47

6.5.1 Dataset 48

6.5.2 Top Location Inference and Preservation 48

6.5.3 Fingerprint Attacks and Defense 51

Chapter 7 Conclusion 57

Bibliography **57**

LIST OF FIGURES

3.1 Generic data flow for connected vehicle services 8

4.1 Data usage and risk in usage based insurance (UBI) 14

4.2 Data usage and risk in drowsy detection 15

5.1 Architecture overview 30

5.2 Fine-grained uploading policies 32

5.3 Context-aware uploading policies 33

6.1 Privacy risk by information inference 37

6.2 Privacy risk by de-anonymization 38

6.3 Top locations inference using daily data 45

6.4 Top locations inference using weekly data 46

6.5 Cumulative distribution function (CDF) of average inference error 49

6.6 Experiments for two-end data removing policy 50

6.7 Experiments for context-aware data sharing policy 52

6.8 Results under different amount of data 55

6.9 Results after adding Laplace noise 56

LIST OF TABLES

3.1 Data summary in connected vehicle services 11

5.1 Privacy characterization of data in connected vehicle applications 26

6.1 Evaluations of fingerprinting attacks 53

6.2 Evaluations of countermeasures for fingerprinting attacks 55

ABSTRACT

The transportation industry is experiencing an unprecedented revolution. This revolution is being led by the rapid development of connected and automated vehicle (CAV) technologies together with cloud-based mobility services featured with huge amount of data being generated, collected, and utilized. This big data trend provides not only business opportunities but also challenges. One of the challenges is data privacy which is inherently unavoidable due to the information sharing nature of such mobility services and the advancement in data analytics. In this thesis, privacy issues and corresponding countermeasure that related to connected vehicle landscape are comprehensively studied. First of all, an overview of the landscape of emerging mobility services is provided and several typical connected vehicle services are introduced. Then we analyze and characterize data that can be collected and shared in these services and point out potential privacy risks. In order to protect user privacy while ensuring service functionality, we develop novel privacy protection mechanisms for connected automotive user experiences. Specifically, we consider the whole life cycle of data collection and sharing. To support privacy preserving data collection, we design fine-grained and privacy-aware data uploading policies that ensure the balance between enforcing privacy requirements and keeping data utility, and implement a prototype that collects data from vehicle, smartphone, and smartwatch securely. To support privacy preserving data sharing, we demonstrate two kinds of risks, additional individual information inference and user de-anonymization, during data sharing through concrete attack designs. We also propose corresponding countermeasures to defend against such attacks and minimize user privacy risks. The feasibility of such attacks and our defense strategies are evaluated with real world vehicular data.

CHAPTER 1

Introduction

In current days, the transportation industry is experiencing the biggest growth in the domain of mobile and connected devices with recent advances in Internet of Things (IoT) technologies. In order to enhance stability, safety, and engagement, modern vehicles have been designed into smart and intelligent cyber-physical platforms built on abundant sensing and computing systems. At the same time, data generated by in-vehicle sensors and computing systems are expected to be collected to fuel more services provided by car manufacturers, third party companies, and the road infrastructure. So vehicles have also gained the capability to connect and interact with external sources through channels such as ODB-II and wireless communications. Most recently, the spreading of mobile and wearable devices, such as smart phones and watches, makes it possible to collect additional dimensions of data from mobile devices in order to achieve goals such as profiling drivers and reconstructing driving scenarios[HOO+14; LMM+17]. Since it has been observed that various kinds of application domains, from real-time sensor data collections for usage based insurance to multiple dimension data integration for personalized mobile services (e.g., FordPass provides services including online payment, parking lots searching, and vehicle localization), rely on real-time vehicular data, we can envision a future of an increasing amount of streaming data will be collected from multiple sources, such as in-vehicle systems and drivers' mobile devices, and integrated for providing valuable services to the users. Data collection at the endpoint is an important task for today's automotive industry.

The consequence in the case of loss of privacy is expected to be disastrous, as witnessed by many business cases over the last decade [KFS+15]. In many applications, users are often required to transmit personally identifiable and sensitive information, such as name, address, phone number, e-mail, credit card number, or date of birth, that can be associated to a particular individual. With regards to sensitive data, greater care must be taken in handling, for example, financial information or real-time GPS locations. Meanwhile, due to multi-source data integration and large amount of data storage, there are also concerns on other kinds of information, such as data streams from sensors deployed on vehicles and mobile devices [GFS+14; ETK+16].

With increasing utility of data for valuable services and the same increasing concern on user privacy, we are facing the dilemma as utility and privacy always become trade-off. As users are willing to reveal their personal information to selected parties for enjoying benefits brought by the big data analysis, the major challenge to enhance data privacy is how to achieve both the data usability and privacy. And the challenge exists at both endpoint data collection and server side data utilization and sharing mentioned above. Note that data privacy management identifies the ways in which both organizations and individuals can control how personal or sensitive data are collected, used, and shared. Well-designed mechanisms for controlling the compliance of privacy preserving policies and regulations would help organizations to improve the trust towards their customers. Such policies/regulations may involve the application of specific laws, national legislations, standards, specific processes or ad hoc restrictions.

In general, users are probably unaware of the diverse influences and implications on their data uploading, sharing and privacy decisions (e.g., access authorization, sharing with third parties). The main issues include: i) direct information without proper management, through data sharing to third parties. For example, sharing personally identifiable information (PII) information to third parties without anonymity. ii) Privacy leakage through data analysis and inference. Given large amount of data from a single user, information about his/her driving patterns, daily activities, personal behaviors can be inferred by an adversary. iii) Privacy issue due to correlation and linkage, and integration from multiple sources. Since a driver might enjoy multiple services from more than one agency, his/her information is possibly integrated from multiple sources, which might cause much severer privacy issues.

Hence, the innovation brought by emerging connected mobility services and applications leads to the growth of the usage of various data, resulting the urgent need for mechanisms to enforce privacy protection. Such a research topic for security researchers is becoming even more important, as very little effort has been done in this field yet, apart from several ad hoc research solutions.

In this thesis, privacy protection mechanisms and algorithms are designed and implemented for connected automotive applications to meet the requirement of privacy preserving while considering user experiences. The scope covers both data collection and uploading from mobile clients/endpoints (vehicle, smartphone, wearable devices, etc.) and data sharing from the data server. For the client side, a holistic framework that securely collects data from multiple sources (e.g., vehicle and brought-in devices) and integrates them in the cloud to enable next-generation connected applications and services with guaranteed user privacy protection is designed. For the server side, potential privacy leakage due to data sharing is studied and corresponding solutions are proposed to mitigate the privacy leakage.

1.1 Thesis Contribution

Contributions of this thesis include:

- We study the connected vehicle landscape and analyze for several representative services and typical data collection and sharing cases.

- We characterize multi-modal data from their privacy requirements standpoint in order to understand privacy implications of each data item individually as well as all possible combinatorial datasets that could be generated.

- We discuss and propose several solutions for privacy preserving data collection and uploading. This discussion provides insights for developing secure data collection and fine-grained, context-aware data uploading approaches to address the trade-off between privacy and data utility.

- We also point out potential and significant privacy issues during data sharing under the background of emerging connected vehicle applications. Especially we illustrate such potential privacy risks by applying concrete attacks on the sanitized data collected by the NHTSA SafetyPilot project [BS14].
- For each attack, we propose corresponding countermeasures considering both the life cycle of the data and also how it is used. Our experiments demonstrate the practicality of such countermeasures.

1.2 Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 discuss related works. Chapter 3 describes the connected automotive scenarios and preliminaries of data collection and sharing. Chapter 4 presents the privacy implication and leakage risk that can happen in connected automotive application scenarios and transitional protection mechanisms. Chapter 5 provides secure and data privacy preserving approaches in data collection mechanisms for connected vehicles. Chapter 6 presents privacy preserving data sharing in connected vehicle services through two kinds of information leakage, while Chapter 7 concludes the paper.

CHAPTER 2

Related Works

Recently, increasing concerns towards data privacy have been brought when it comes to car connectivity [Jer14; DRK16], especially given the prospect of multi-source data collection and integration and large amount of data sharing that is foreseen by some recent works [PGS17; LMM+17]. In order to draw more public attentions, [Jer14] provided an overview of the currently enrolled technologies in modern vehicles and presented emerging data collection, types of collected data, and purposes for collecting these data.

Under the background of connected vehicles, privacy leakage due to data collection and sharing were studied by recent works. One study showed that consumers have willingness to enroll in current usage based car insurance and share data to car insurers if given a minor financial compensation [DRK16]. However, recent study showed that driving data collected in applications, such as usage based insurance, might cause serious privacy concerns [GFS+14; ZCL+17; NVB+16]. For examples, several research works found out mobile sensor data, such as acceleration, can be used to recover driving paths and infer drivers' locations [GFS+14; ZCL+17]. Even daily driving distance are possible to reflect drivers' financial status [VJ13]. A study proposed a method to compute insurance cost locally on a device connected to the vehicle so that GPS data are not necessarily to be shared to insurers' servers [TDK+11], but it requires extra cost to produce and promote such devices to drivers.

Some research works in other areas are also related to our problem. Zang et al. analyzed large-scale phone call records to present the risk of re-identification attacks with published location data [ZB11]. Their results reflect the location data should be coarse in time domain (e.g., short period of data) or space domain (e.g., lower granularity) to avoid privacy leakage. Bindschaedler and Shokri published the first paper to systematically generate plausible synthetic location traces for protecting the user location privacy in location based services. The synthetic traces were designed to resist inference attacks aiming at identifying the true location of users, while greatly maintaining the data utility [BS16]. The studies in this thesis will also address some location privacy issues in connected vehicles by proposing data management policies to reduce the possibility of sensitive information inference (top locations) while considering data utility for data consumers.

Fingerprinting targeted objects through data analyses has been studied in different areas [VVN16; TSC+16; ETK+16; CCS17; CDP+17]. Vehicular senors data might not be linkable to PII information directly. However, previous works have shown that it is possible to fingerprint drivers through in-vehicle sensor data collected from ODB-II [ETK+16] or IMU data from drivers' smartphones [CCS17]. Investigations in this thesis will show that the risk of de-anonymization also exists in the connected vehicle scenarios by evaluating fingerprint attack on BSM data from 43 drivers.

With the revolution of connected vehicle network technology, huge amount of vehicular data can be generated, stored, and consumed for service providers. Security and privacy protection techniques in data collection [AHM+15], storage [RKS+16; KSL+14; KSL+17], querying [PS14], content sharing [DDM+17], access control [ASM+14], should also be studied in the connected vehicle landscape. To the best of our knowledge, this thesis work is the first to comprehensively study the connected vehicle landscape and its privacy issues in data sharing. And we also take an initial step towards privacy preserving connected vehicle data sharing, which is compatible to topics mentioned above.

CHAPTER 3

Preliminaries

3.1 Connected Vehicle Landscape

In earlier vehicle networks, telematics services such as vehicle tracking and navigation require only limited amount of data to be exchanged in the vehicle service networks. Local traffic information used to be uni-directionally broadcast over satellite radio channels and GPS coordinates of current vehicle's position were sent to a service server only upon requests or intermittently via cellular communication channels due to low bandwidth and high cost.

Recent advances in cellular communication technology and popularity of mobile services end up leading a rapid deployment of vehicle connectivity to Internet cloud services. More advanced generation of cellular technology is being integrated to modern vehicles and Wi-Fi adapters embedded in some vehicles are opportunistically looking for available Wi-Fi networks for Internet connectivity. Even vehicles without equipping such wireless connectivity at a factory setting are able to be tethered to a mobile device or use aftermarket OBD (on-board diagnosis) port dongles connected to the Internet. Additionally, technologies like DSRC and 5G enable vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications, thus connecting the car to a broader, smarter Internet of Things (IoT) ecosystem.

As a result of this ubiquitous vehicular connectivity, a number of powerful, data-driven applications are being developed for enhancing driving and customer experiences. Among many, we introduce some examples below. It is important to note that the goal and the data collection

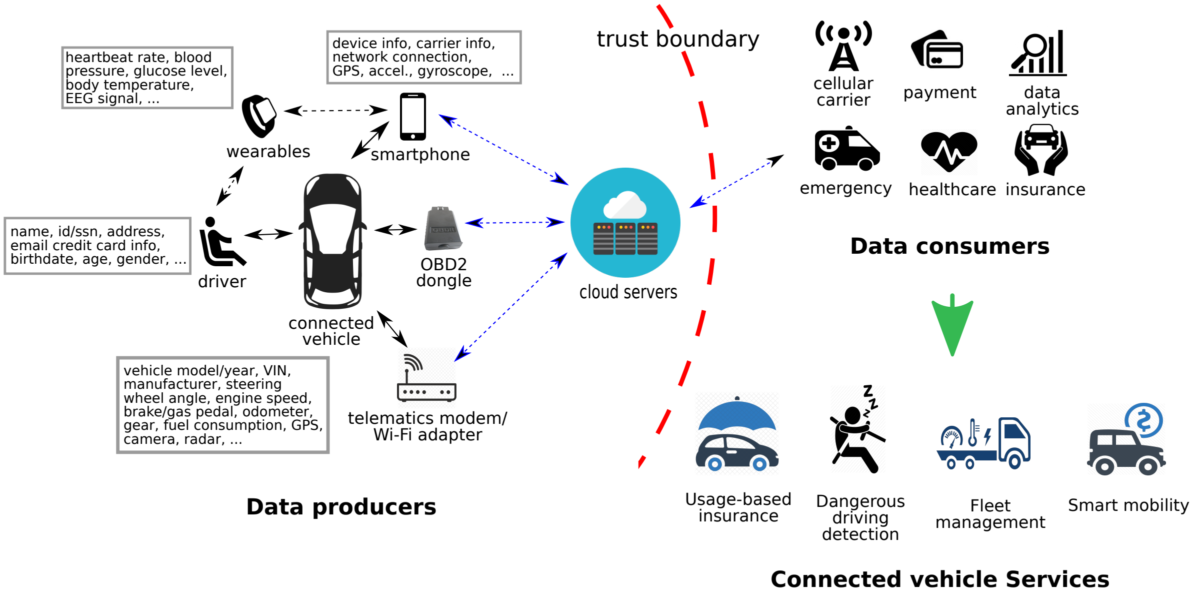

heartbeat rate, blood pressure, glucose level, body temperature, EEG signal, ...
device info, carrier info, network connection, GPS, accel., gyroscope, ...
trust boundary
cellular carrier
payment
data analytics
emergency
healthcare
insurance
wearables
smartphone
name, id/ssn, address, email credit card info, birthdate, age, gender, ...
driver
OBD2 dongle
cloud servers
Data consumers
connected vehicle
vehicle model/year, VIN, manufacturer, steering wheel angle, engine speed, brake/gas pedal, odometer, gear, fuel consumption, GPS, camera, radar, ...
telematics modem/ Wi-Fi adapter
Data producers
Usage-based insurance
Dangerous driving detection
Fleet management
Smart mobility
Connected vehicle Services

Figure 3.1: Generic data flow for connected vehicle services

path vary by application, but it is identical for all of them that the amount of exchanged data, the communication frequency, the number of data modality are incomparably large compared to the initial vehicle networks. More importantly, the data collected from individual vehicles/drivers/in-cabin devices are not just consumed within the data source origins or vehicle clouds, but can be extensively shared with other stakeholders.

Usage-based insurance (UBI): Today, drivers have the opportunity to opt-in to UBI for reduced premiums by allowing companies to monitor their driving behavior. Rather than relying only on drivers' accident or traffic violation history, UBI can give a more reasonable insurance premium by observing the current driving behaviors on road. Taking advantage of the V2X communication and data collection techniques, insurance companies offer consumers insurance policies with the option of installing application on smartphone to collect sensors data (e.g. acceleration, gyroscope, magnetometer) or installing a monitoring device into their vehicle to collect driving data (e.g. mileage, speed, hard braking, GPS) [MCH00; Comom; HPF+15]. Then a risk score is computed based on these data to recalculate insurance premium.

Dangerous driving detection: To enhance driver's safety, it is crucial to detect any dangerous driving behavior and take a necessary action. Drowsiness detection is an example of dangerous driving detection, as driver drowsiness has been reported as a major cause of mortality in road accidents [LLC15]. To detect driver drowsiness, sensor based [SSM12], face expression based [AR11], and physiological signals based [LLC15] approaches, are proposed. To avoid privacy concern in using video capture of drivers, wearable devices, such as smart watch, can still collect comprehensive data, such as electroencephalographic (EEG) signal and heart rate.

Fleet management: Fleet vehicles and their carrying freights are very important assets, so that telematics services have been adopted much earlier in this business. Rapid advances in wireless technology enable real-time vehicle/freight tracking services. For more fine-grained monitoring and even predictive path planning, collected are not only the location information and vehicle status, but also driver information and surrounding environmental information such as weather, traffic, road condition.

Smart mobility: As seamless mobile user experiences are essential for connected or autonomous vehicles, many smart mobility services are deeply integrated into modern vehicles. For instance, automotive manufacturers have provided a smartphone app, with which can monitor the current status of vehicle (e.g. physical access, electrical battery remaining) and remotely control a vehicle (e.g. door locking/unlocking, engine warm-up). To better support customers' urban mobility experience, more features are being integrated into the app. Examples include in-vehicle payment and quick navigation for gas, parking, drive-in shops.

Fig. 3.1 represents generally what types of data are produced and how they can be consumed to enable the connected vehicle services. These data generated from connected vehicles, drivers, and their smart devices are gathered in cloud servers. These data are shared with multiple businesses to enable emerging connected vehicle services. It is important to note that even though the data is collected and stored in a secure manner, it can be potentially shared with malicious parties. Generally, the whole procedure from data generation to data consumption can be divided into two parts with separately research focus: data collection and data sharing.

3.2 Data Collection in Connected Vehicles

In summary, Fig. 3.1 presents data related to the above scenarios according to data sources. Personal information is usually data that are collected when a user enrolls into applications (e.g., UBI, car sharing) or registers an account (e.g., mobile payment). This kind of information is really sensitive to personal privacy and should be well protected by the companies. However, the recent example of the Equifax data breach, where 143 million American consumers' sensitive personal information was leaked due to hacker intrusion [BTP+kh], shows that without strong security mechanism, these sensitive data can be stolen.

With the development of automotive and V2X techniques, driving data are able to be collected from vehicles and uploaded to cloud servers. Vehicle data contains abundant information of both vehicle itself and driving status, which can be used to various applications, including not only car insurance or car sharing mentioned above, but also connected vehicle networks and autonomous driving.

Meanwhile, mobile devices and wearable devices provide more dimensions of data sources and convenient data access. Smart phones are able to record a set of sensor data, which can be used to calculate the usage of services, estimate driving movement, tracking driving path, etc. Wearable devices can be used to collect biological data of drivers, which directly benefits applications like drowsy/misbehavior detection.

Table 3.1 summarizes representative data and explanations from different data sources. These data can be collected and uploaded to cloud server for further use.

Table 3.1: Data summary in connected vehicle services

Data sources	Data items	Explanations/examples
Personal information	User Name	These information is usually collected at enrollment or registration information
	ID/Social security number	
	Home address	
	Email address	
	Credit card	
	Birth date, age, gender, ..	
Smart Phone	Device info.	Phone model, brand, version
	Network info.	Carrier information, network connection (2G/3G/LTE/Wi-Fi)
	Location data	GPS (high accuracy), network (low accuracy)
	Motion sensors	Accelerometer, gyroscope, rotation vector, step counter, step detector
	Position sensors	Magnetic field, orientation, proximity
	Environment sensors	Temperature, light, pressure, relative humidity
Wearable devices	Human body data	Heart rate, skin temperature, electroencephalographic (EEG) signal
Vehicle	Vehicle info.	Model, VIN, make, etc.
	In-vehicle sensors	Steering wheel angle, engine speed, brake pedal, odometer, gear, fuel consumption, accelerator, etc.
	Car embedded system	GPS, camera, etc.

3.3 Data Sharing in Connected Vehicles

Another important session for applications in connected vesicle network is the data sharing. Service providers usually require abundant data for analyzing, so they should be able to query necessary data from the cloud server to support their services. Based on the above application examples, we identify two categories of data sharing applications. One is individual data sharing for personalized mobility service. In this category, service providers require information or data from an individual user so that they can provide personalized service. For example, UBI and dangerous driving detection belong to this category. The other category is group data sharing for group mobility services, where anonymized data from a group of users are required and the service providers are not necessarily aware of the ownership information of these data. For example, fleet management belongs to this category. Smart mobility applications usually involve both types of data sharing.

CHAPTER 4

Privacy Analyses and Requirements in Connected Vehicles Applications

In this chapter, we provide privacy analyses and survey for typical connected vehicle applications and related scenarios to understand what privacy issues might occur and why we need to protect user privacy.

As we introduced in Chapter 3.1, there are various applications in the connected vehicle landscape. To fully support these applications, multiple dimensions of user data are collected by companies, which can bring a lot of user information leakage and sensitive information sharing (e.g., PII). In this chapter, we first give two case studies to show how data are used in services and what risks can occur from these data.

4.1 Case Study of Privacy Leakage

Fig. 4.1 provides a case study of usage based auto insurance. In order to estimate personalized premium, insurance company may collect some data from drivers to evaluate the drivers' driving behaviors. For example, in-vehicle sensor data can provide insights to know whether a driver is an aggressive driver (e.g., whether hard brakes very often or whether the driver use turning signals before turning). And the vehicular GPS signals reflect how long a driver drives every day, which may reflect the possibility of potential accidents, because long distance driving might cause fatigue and distractions. Besides, To be enrolled into insurance policies, drivers need to provide

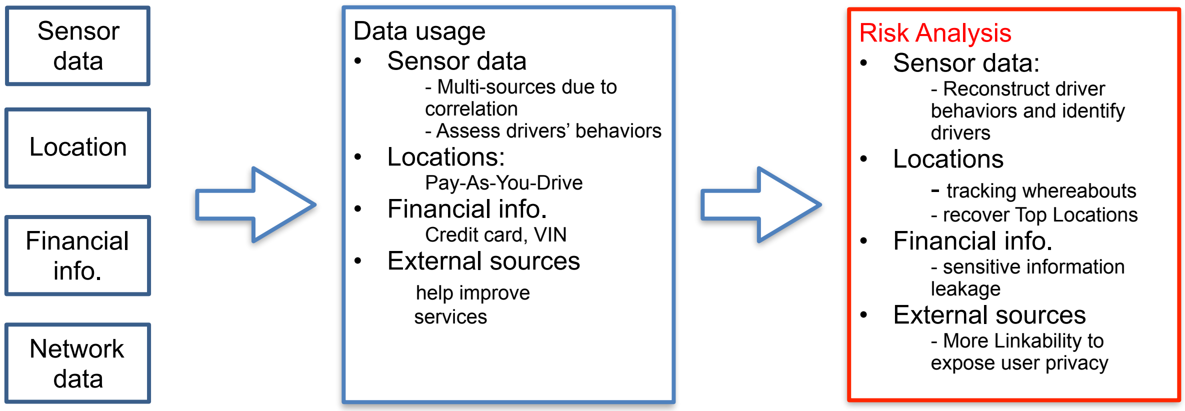


Figure 4.1: Data usage and risk in usage based insurance (UBI)

some financial information such as credit cards, VIN, income. However, sharing such kinds of data will cause direct information leakage, such as being tracked through real-time sensor data and GPS data, or direct financial information leakage if the insurance company is untrusted.

Fig. 4.2 shows the case study of drowsy detection, which is an emerging application as a safety feature in today's connected vehicle applications. We can see some collected data such as sensor data are in common with usage based insurance, and similarly, used to identify whether a driver's driving behavior is normal or not. But some other data types are different because the goal of the service is different. Vehicular camera will collect photos of a driver's face when he/she is driving, so that fatigue detection can be done through the facial expression analysis. And biometric data, such EGG, heart rate, would also be collected through to check body/condition of the driver. However, sharing of these data could cause serious privacy concerns because the face is an important identity of a driver, and biometric data can indicate health status of drivers. So without proper data management and privacy protection, data sharing might lead to serious consequences in privacy leakage.

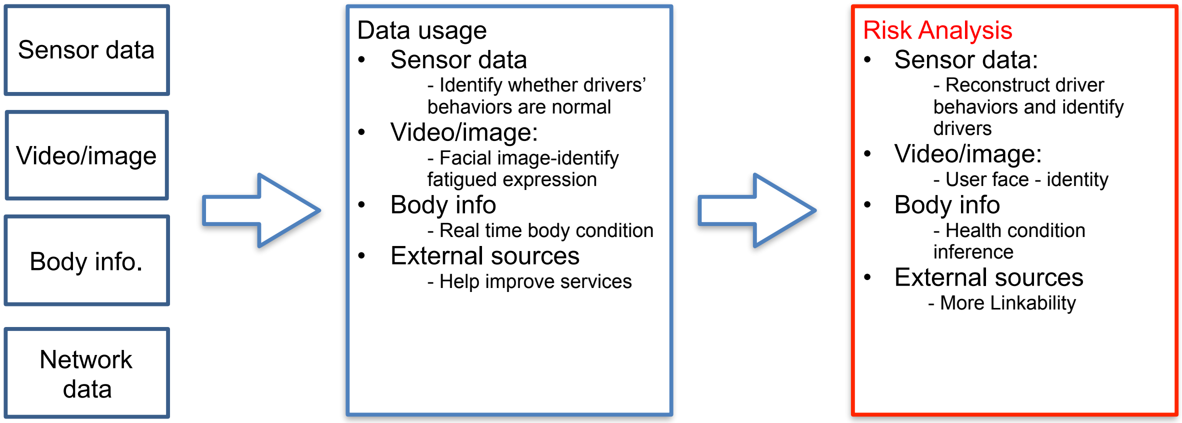


Figure 4.2: Data usage and risk in drowsy detection

4.2 Privacy Risks and Implications in Connected Vehicles

In some scenarios, simple anonymization techniques, such as removing ID or k-anonymity can be used to preserve user privacy. However, it cannot well prevent privacy leakage in more general cases, due to the following reasons: i) in some cases, individuals' information has to be shared to service providers to enjoy services; ii) through data analysis, PII (e.g., user identity) can be inferred by some seemingly non-PII information. So in this section, we will discuss such privacy risks by investigating potential attacks on real-world connected vehicle data through data collection and sharing.

4.2.1 Personal Identity Leakage

Personal information will not be directly leaked through driving data or smartphone/watch data. However, it is not difficult to deduce some information through the data that is collected. For some static data such as brand, model, and configuration of vehicles and device model, carrier of smartphone/watch the users' financial status can be roughly inferred. Meanwhile, cascading this device information is likely to uniquely identify a user of the narrow the user into a small number of candidates. A similar example is shown in reference [DBC16].

The driving data can be used to uniquely identify or fingerprint the users, which may cause a long time tracking to leak more privacy. By using 16 sensors collected from ODB II data (see Table 1), it is possible to differentiate 15 drivers with 100% accuracy when training with all of the available sensors using 90% of driving data from each person. Furthermore, it is possible to reach high identification rates using less than 8 minutes of training data [ETK+16]. By using gyroscope, accelerometer readings from smartphone, a smartphone user can also be fingerprinted by more than 90% F1-score in experiment environment [DBC16]

4.2.2 Location Privacy

Except from directly reading GPS locations from device, some sensor data can also be exploited to recover users' locations or even traces, which breaches location privacy. Sensor data that can be collected from vehicle/phone sensors, such as accelerometer, gyroscope, and magnetometer. They are usually viewed as statistics that don't have semantic meanings and will not reflect privacy. However, existing work has proven that a zero permission Android application is able to estimate a driver driving path and destinations, using only IMUs from smartphone. By modeling roads as a graph, an attacker is able to exploit the IMUs data to reconstruct the vehicle speed, tuning, and steering. As a result, the attacker can recover driving path by fitting estimated movements to the roads' speed limit/intersections/directions of roads on a map [NVB+16].

More surprisingly, by simply reading the phone's aggregate power consumption over a period of a few minutes, an application can learn information about the user's locations [MSV+15], the path can be recovered. The insight is that signal strength profiles measured on two different days are stable. So the power consumption in different path also has fixed patterns. Based on prior knowledge and machine learning techniques, the path can be recovered. So, without GPS location, the users' location privacy can also be threatened if no protection is provided on data collection.

4.2.3 Abundant Information Inference

In many cases, more user information can be inferred through non-PII information. Since future connected user dynamic resources can also be extracted from indirect risks they may associate with, which further depend on the variable time and frequency of access. For example, if a driver's GPS information is only accessed once or infrequently, a malicious party cannot derive the movement profile of a user (which may tell where the user is living, working, shopping, etc.). However, if a malicious party can access GPS information of a user regularly such as every 30 minutes, seven days a week, she is capable of deriving user location profile. Correspondingly, users' living places, working places, frequently visiting shopping malls are not difficult to be inferred, given daily driving data. The inferred information can be used to identify the users financial status, habits and recover daily activities, which are serious personal information leakage [LZD+16]. Furthermore, even the demographics of users (gender, education level, etc.) can be recovered in a large scale using supervised machine learning techniques, based on the observation that similar characters of users are likely to visit similar places [LZD+16].

4.2.4 Privacy Leakage through Data Aggregation

A potential privacy risk comes from the various categories of data related to modern connected vehicle scenarios. Collecting multiple dimensional of data will benefit services undeniably, but it will also bring more privacy concerns as well. In the past decades, we have witnessed multiple serious privacy breach issues due to information aggregation from different sources [KFS+15; LCZ+17b; LCZ+17a]. In this paper, we identify potential privacy risks through multiple sources data, and we can briefly summarized them into homogeneous information aggregation and heterogeneous information aggregation.

Homogeneous information means different data items from the same category of data. As we will show in Chapter 6.5, we fingerprinted drivers using nine sensor data from either in-vehicle sensors or vehicular cameras. Besides, from smartphone or smartwatch, we can collect more IMU sensor data that are homogeneous to vehicular sensors. These data are called homogeneous information. And aggregating homogeneous information will potentially make privacy be leaked easily, as we will show that aggregating all sensor data achieve the best accuracy for fingerprinting.

Heterogeneous information means information from different categories of data. Since a service may require different categories of data, this makes privacy protection more difficult. For example, if using an Android smartphone to collect driving data, the Android app is able to easily collect the OS version, API level, device, model, product information of the smartphone in the same time by stealthily invoking “System.getProperty("os.version")”, “android.os.Build.VERSION.SDK”, “android.os.Build.DEVICE”, “android.os.Build.MODEL”, “android.os.Build.PRODUCT”. With these device information, it is not difficult to identify a unique user or limit candidates from a large set of users. And it will be a complement to de-anonymize the drivers using sensor data, and bring more privacy concerns. What’s worse, equipped with de-anonymization techniques [KFS+15], an attacker may aggregate information of an individual from multiple services/scenarios. These facts show that user information can be directly leaked if attackers can collect as much data as they can.

4.3 Traditional Privacy Protection Method

Depending on data types, their use cases and contexts, there will be fine-grained privacy enhancement data upload policies that optimize privacy requirements and resource consumption (e.g., processing, network bandwidth). In the following, we discuss different techniques and strategies to be explored in this task to achieve resource efficient and privacy preserving uploading under different contexts.

We differentiate data collected by connected vehicles and devices attached to them into two major categories: personal identifiable data, and personal publishable data.

4.3.1 Protecting Personal Identifiable Information

Personal identifiable data refers to the identifier, photos and messages, which are highly sensitive, and thus usually not disclosed to any third party. Especially considering the recently reported iCloud data breach issue, it is important to have corresponding mechanisms to secure user privacy. Usually, two ways are used: (1) simply removing such information from data to be uploaded or (2) exploiting cryptography based encryption are two ways to enhance the data confidentiality.

Removing sensitive information is the most straightforward way to protect user privacy and it also helps reducing the data size and achieves better bandwidth efficiency. However, lacking such information may prevent a driver from receiving valuable services. For example, a photo of the interior of the car taken by a camera inside the car during an accident may help 911 customer service to accurately accessing the severity of the accident and take appropriate actions.

On the other hand, to ensure confidentiality and prevent unauthorized reading in cloud and beyond, sensitive data can be encrypted before they are uploaded to the cloud. However, encryption makes the utilization, such as keyword search over plaintext, of the data difficult. Homomorphic encryption and searchable encryption are two special encryption techniques that are useful to make encrypted data utilizable (to certain extent) [BDO+04]. Depending on how uploaded data is to be processed in the cloud, we need to investigate whether these techniques can be utilized and then develop corresponding secure communication protocols for data uploading.

- Partially homomorphic encryption. Homomorphic encryption is a form of encryption where a specific algebraic operation performed on the plaintext is equivalent to another (possibly different) algebraic operation performed on the ciphertext [DPS+12]. Thus, it allows computations to run directly on encryptions of their inputs and produce encryptions of outputs. Since the input, intermediate result and the output of a computation can all be in encrypted form, homomorphic encryption based computation can be run by an untrusted party without revealing its inputs, outputs and internal states. Fully homomorphic encryptions supporting both addition and multiplication are far more powerful as they allow any computation be homomorphically evaluated. Unfortunately, existing fully homomorphic encryption schemes are still very expensive to be used in practice. Partially homomorphic encryptions (supporting computation of only one operation, either addition or multiplication), such as unpadded RSA, ElGamal, and Pailier are very efficient. If only simple data statistical computation (e.g., SUM, AVERAGE, MAX) to be performed on uploaded data, we can utilize partially homomorphic encryption schemes to provide additional protection of user data stored in the cloud.

- Searchable encryption. Searchable encryption is an encryption technique which allows users to search over encrypted data via keyword(s) without decrypting the data first [BBO07]. There are both symmetric and asymmetric searchable encryption schemes. Symmetric searchable encryption is much more efficient while asymmetric or public searchable encryption support a richer set of search predicates, such as conjunctive keyword search, subset, range queries, and graph structure. Fuzzy key word searchable encryption has recently been proposed to accommodate various typos and representation inconsistencies in different user query inputs.

4.3.2 Protecting Personal Publishable Information

Personal publishable data including locations (which can still leak user privacy) can be partially released to the service providers or advertisement providers to allow faster and targeted services and advertisement. For this kind of data, cloaking, generalization or mix zone are typical approaches to enhance the privacy of the publishable data to achieve certain privacy properties such as k-anonymity, t-closeness or i-identity. Alternatively, computational encryption schemes such as efficient homomorphic encryption and searchable encryption mentioned previously can be utilized if the data is to be used statistically. We also notice that many evidences have suggested that privacy is context-dependent. For instance, existing studies on actual disclosure behavior of online social networks show that many users increase the amount of personal information revealed to their friends while simultaneously decrease the amounts of data revealed to strangers. On mobile devices, different apps, user preferences and even physical environments (locations) play different factors to affect the privacy level that the user tends to choose. Therefore, for personal publishable data, we also propose to develop context-based privacy management module to provide an automatic way to translate the various context factors to a specific privacy level. The level of privacy will be used to determine the frequency of data uploading.

However, when it comes to connected vehicle scenarios, we should consider not only protecting privacy but also maintaining normal service usability. Since privacy protection often comes with a cost of reducing data utility, the balance between them should be addressed. This means that we need to tailor suitable privacy preserving mechanisms or solutions for connected vehicles applications. In the following sections, we will present our solutions at both data collection and data sharing respectively.

CHAPTER 5

Secure and Privacy-preserving Data Collection Mechanisms for Connected Vehicles

In this chapter, we introduce a secure and privacy-preserving data collection framework for connected vehicles.

5.1 Design Principles and Goals

In order to design a privacy preserving data collection framework, it is important to understand the privacy implications and requirement. Based on analyses of failures and defect of existing systems, we list the following key insights as a guideline for designing privacy preserving framework.

- *Always keeping usability into consideration.* The most important insight for designing real-world connected vehicle applications is keeping usability into consideration. Privacy usually comes with a cost, in order to minimize the cost of user operation, we propose to conduct rigorous user studies and design controllable solutions.

- *Providing user-centric privacy control.* Based on the idea of keeping usability in mind, we propose to allow different users to have different privacy preserving settings. Mechanisms that allow users to set their privacy preference rather than default data access settings and policies are required to provide personalized privacy protections.

- *Understanding multi-modal data privacy.* Multiple sources data collection modal will benefit constructing better profiles of users. However, it would threaten user privacy as well. Different kinds of data might have different impacts on the user privacy, according to the role who owns the data and the properties of the specific data. Some kinds of data may have higher risk impacts than others if they are more related to the user's personal identity. Meanwhile, if a kind of data are accessed multiple times or over a long time, it would be more sensitive than being accessed once. As a result, to understand privacy impacting factors of different kinds of data is necessary.

- *Categorizing risk types*. Different types of security and privacy risks (e.g., personal information leakage, location privacy, financial risks etc.) can occur under the background of connected mobility services. Comprehensive risk analysis and categorization should be studied with considering the roles of users, trusted data collectors, and third-party service providers.
- *Providing resilience against data breach*. Collecting data and storing them on cloud servers allow convenient data access and sharing. However, the large amount of data storage on cloud servers might cause serious data breach concerns. So practical access control and intrusion resilience are desired to prevent malicious parties from obtaining data they should not access.

5.2 Technical Approach

Our technical approach starts with characterizing multi-modal data and understanding their privacy implications, then two privacy preserving data uploading policies will be introduced.

5.2.1 Characterize Multi-modal Data

As we discussed above, collecting user data from multiple sources will benefit building a more comprehensive profile of each user. For example, real time driving data can be directly collected from vehicles through ODB-II or wireless channel in V2X applications. But when it is not easy to deploy data collection devices for vehicles, smartphone can be a alternative because it ensemble various sensors such as IMUs, GPS, bluetooth and network modules. Meanwhile, emerging wearable devices provide additional dimensions of data, such as heart rate and body temperature of the users, to understand the status of driving. However, these data from multiple sources not only fuel various applications/services, but also cause serious concerns. For example, some of these data can be non-identifiable (i.e., not PII), but combining them together might identify a specific user. The privacy concerns can vary from individual to individual, as we have mentioned different users have different privacy preference.

Based on these facts, we provide a privacy characterization for different kinds of user data based on to what extent they can impact user privacy. In order to ensure a privacy preserving data collection and and uploading, quantitative and qualitative analyses are given to discover privacy-impacting patterns. This characterization can be a reference for understanding privacy impacts and extracting usable information, and further used to define a variety of privacy-aware access control semantics. Techniques like access control or data encryption should be consider to secure data with high privacy impact.

Privacy Characterization

To characterize privacy impact, we not only consider data sources, but also consider properties of each data. A kind of data can be described using a set of properties, including:

- Identifiable or not-identifiable: certain kinds of data is possible to identify a user. For example, a user's phone number, social security number, VIN, are identifiable because they are associated with one user at a time uniquely. On the other side, some data are not identifiable, such as sensor readings from smartphone or vehicles.
- Context-aware or context-independent: Context-aware means the data may have different

implications under different contexts. Location is a context-aware data as some locations may be Top locations (e.g. home, work place) [MJ13] of a user. But some other locations locate at public area that will not disclose the user's privacy. So the GPS signal that directly reflect drivers' locations is context-aware. Another example is the timestamp. It indicates when a user is driving, which indirectly reflect daily activities of the user. So it is also context-aware.

- Semantic or not semantic: Some data, such as vehicle model, make, phone type, contain semantic information themselves. Some other data such as GPS readings indicate semantic information implicitly. GPS indicate locations of users, because the GPS coordinates can be transformed to semantic addresses using a map. So GPS data are semantic as well. On the other hand, sensor data such as IMUs, yaw rate and wheel speed, do not contain semantic meanings directly.
- Static or dynamic: Some data are rather linked to a static value, such as VIN, car/smartphone/smartwatch model, while some data such as sensors readings are dynamic because they are refreshed frequently.

Based on the discussion for privacy characterization, we summarize some typical data and their privacy characterization in Table 5.1.

Privacy Model on Large Data

Based on the characterization above, we can define a privacy vector to describe each data item's properties, i.e., $Z =$ [identifiableornot-identifiable,context$-$awareorcontext-independent,semanticornot-semantic, $static$ or $dynamic$,]. For example, if a data d is identifiable, context-independent, semantic, and dynamic, the privacy vector of d can be defined as $Z_d = [1, 0, 0, 0]$; if a data d' is not identifiable, context-aware, not semantic, and static, the privacy vector of d' can be defined as $Z_{d'} = [0, 1, 1, 1]$.

Table 5.1: Privacy characterization of data in connected vehicle applications

	Phone				Watch			Vehicle		
	Device Model	Carrier Info	Internet Addr.	Sensors	Heart Rate	Tempe. Value	Light	GPS	Speed	Brake Pedal
Identifiable(0) or not-identifiable(1)	1	1	0	0	0	0	0	0	0	0
Context-aware(0) or context-independent(1)	1	1	0	1	1	1	1	0	1	1
Semantic(0) or not semantic(1)	0	0	0	1	1	1	1	1	1	1
Static(0) or Dynamic(1)	0	0	0/1	1	1	1	1	1	1	1

The privacy vector can be used to describe the privacy characteristics in a general way. However, it is worth to note that the value of characterization in a specific data's privacy vector might change under different situations. And indirect privacy risks of sharing some dynamic data can also happen, which further depends on the time duration of data collection and frequency of data access. For example, if a driver's diving data are only accessed infrequently, an adversary might not be able to infer the driver's driving behaviors. However, if the adversary has the permission to frequently query driving data like GPS coordinates, e.g., read real time GPS every 30 minutes, the adversary is likely to infer the driver's location profile, and more personal information [LZD+16].

So based on the discussion, we can conclude that all kinds of data might potentially and implicitly cause concerns to the user privacy.

In order to aggregate multiple data sources and examine potential risks for the data aggregation in a descend way, a novel probabilistic mixture model learning framework that can adaptively evaluate privacy risks for multi-model data is further proposed. Here, we denote $\mathcal{D}$ as driving data produced by an individual driver, and it may contain multiple kinds of data, such as data from vehicle, smartphone, smartwatch. Then a specific type of data $d \in D$ is represented by a vector of characterizations v as we mentioned above. And a regression model for evaluating the risk score

$R(D)$ of the user's driving data can be presented as:

$$R(A) = \sum_{d} P(Z_d = v|d)P(R_i(D)|v, a) \tag{5.1}$$

where $R(A)$ is a risk score that indicate severity of privacy risk, such as [0, 0.25) indicates very low risk, [0.25,0.5) indicates low risk, [0.5, 0.75) indicates medium risk, and [0.75,1] indicate high privacy risk.

In practice, this model will need to be trained by some data associated with their risk levels that are pre-defined based on expertise and experience. Then new data can be input for evaluating privacy risks. Equation 5.1 models risk scores of driving data by adaptively combining risk factors from different sources. In particular, $P(Z_d = v|d)$ associates a kind of data to a specific combination of characterizations. For example, accelerometer reading is a kind of data that is not identifiable, context-independent, not semantic, and dynamic in most cases, so we have $v = [1, 1, 1, 1]$. But according to our discussion above, accelerometer reading might be used to infer users' identity given a large number of data [ETK+16], so it can be identifiable in this case, i.e., $v = [1, 1, 0, 1]$. The probability of associating a specific kind of data with its privacy vector is determined by previous knowledge (i.e., determined in the training phase), or validated by combining existing techniques [ETK+16; MSV+15; CCS17; NVB+16] to reduce the need of manual settings. Furthermore, the expression $P(R_i(D)|m, d)$ determines the risk score for a specific kind of data associated with a specific privacy vector. Finally, a sigmoid function (i.e., $\sigma(x) = \frac{1}{1+e^{-x}}$) can be used here to adjust the value score into the range of [0, 1].

In summary, the model computes a uniform privacy score that quantifies privacy risks of data from multiple sources. Given a set of labeled data, it can be trained and be used to evaluate the risks of the future data. In practice, the weight of each kind of data in Equation 5.1 should be adjusted to fit the real world requirement.

5.2.2 Privacy Preserving Data Uploading

According to the discussion above, privacy implications of a kind of data might vary under different scenarios. To cope with the complexity of measuring data privacy risks, fine-grained and context-aware privacy preserving data collection and uploading that optimize privacy requirements and resource consumption (e.g., processing, network bandwidth) is proposed in this section. Specifically, we present two data collection uploading policies: the fine-grained policy and context-aware policy.

Fine-Grained Uploading

Sensitive data are encrypted in the most cases during data transmission to achieve confidentiality and avoid internal or external leakage on cloud and beyond. However, encryption techniques might reduce the utilization, for example, keyword search is easy to implement over plain-text, but hard over cipher-text. Meanwhile, encryption also cause extra computational cost. In order to ensure both the privacy requirement and the utility of data, we propose to provide fine-grained uploading policies that allow users to hide specific kinds of data.

Different individuals might hold different viewpoints about what is privacy for themselves, so different people want to protect different sets of data. To meet each individual's requirement, firstly we provide a self-defined uploading policy, which gives users the right to select data that they would not be willing to share. The selected data will be encrypted or removed before uploading.

Meanwhile, people have uncertainty about the consequences of privacy related behaviors or even their own preferences over those consequences [ABL15]. So as supplementary, a set of pre-defined fine-grained policies is provided, which is based on characterizations of data in Table 5.1. For example, users are able to choose to encrypt all identifiable data or all semantic data so that their identity would not be leaked. In real world scenarios, risk scores of data in Equation 5.1 (after training) can be provided to users as references for deciding which kinds of data should be encrypted.

Context-Aware Uploading

The term “context-aware” (or “context-dependent”) refers that a user will have different privacy preferences and take different actions under different contexts. As we have discussed, location is a kind of context dependent data. For examples, frequent appearances at hospitals is likely to be more sensitive than appearances near a garden because the former may raise the health concerns while the latter is positioned as a social spot; top locations [ZB11] (or most frequently visited places, e.g., home or work place) are closely related to users’ identities, so they are much more sensitive than other locations. Since GPS signals and other context-aware data are often collected and shared in connected vehicle applications, context-aware uploading policies are designed to prevent critical privacy from being leaking.

In this thesis, we study the location as a context-aware data, and in our design, two options for context-aware uploading policies are provided. The first policy allows users to manually select some locations that needs to be protected. And users can also set a range (a radius of a circle whose center is a location needs to be preserved) for considering the protection within the area. With this policy, when a driver enters the area around a location that needs to be protected, the data collection will be paused in order to ensure data around these important location will not be leaked and leveraged by adversary.

Since the first policy requires manual settings for the protected locations by users, it may cause certain operation cost. To reduce operations of manually settings, we provide a method to automatically execute context-aware data protection by identifying top locations of a user. By recording a user’s historical location traces and storing the mobility history in the local database on smartphone or remote server (in our framework, we assume the data storage is secure, as they can be encrypted), the policy is able to compute the top locations (which are sensitive locations in the context) for this user. So similarly, the data collection will be paused when the user drives near her

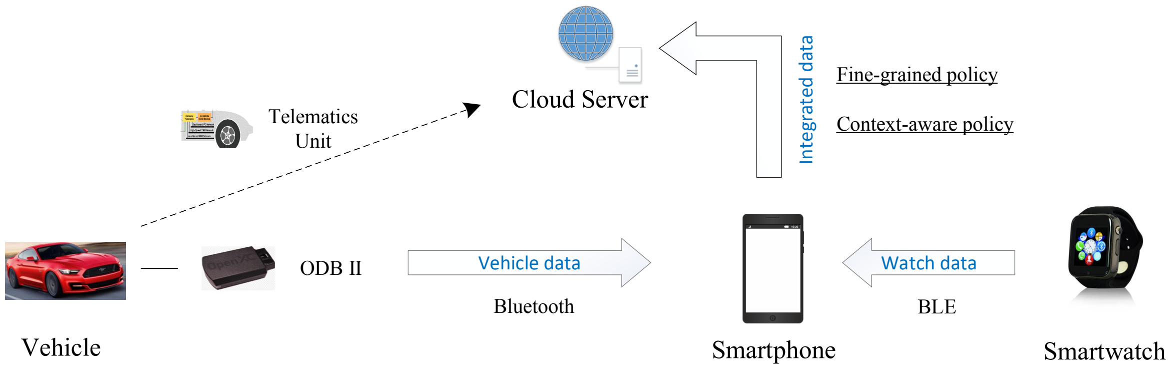


Figure 5.1: Architecture overview

top locations so that her top locations have no chance to be leaked.

5.3 Prototype Implementation

To implement the the secure and privacy preserving mechanisms, we develop a framework based on a vehicle data collection module, smartphone, and smartwatch as a prototype.

5.3.1 Implementation of Data collection

The multi-model data collection framework is presented in Figure 5.1, which consists of smartphone, smartwatch, and the vehicular device. Since smartphone has ubiquitous network connections and good computation capability, it is used to integrate and synchronize data from the vehicle and the smartwatch before uploading. And the privacy preserving policies are also implemented on the smartphone. In our prototype, Android phone and watch are used.

Collecting Data on Vehicle

Data collected from vehicles mainly come from in-vehicle networks (e.g., CAN bus networks) or vehicular embedded systems. The data mainly contains sensors readings and device parameters that directly reflect the driving process, vehicle condition, driver behaviors. In our implementation, vehicle data from a Ford car are read from OpenXC[1] and transmit to through Bluetooth socket. The data contain 19 kinds of signals, such as accelerator pedal position, engine speed, fuel level, etc. in total.

To set up the experiment phase, the OpenXC device should be plugged to ODB-II socket on the vehicle, and the smartphone broadcasts messages and searches for the OpenXC device for building Bluetooth connection . Upon the connection is set up, the smartphone is able to periodically read vehicle data through the Bluetooth connection.

Collecting Data on Smartphone

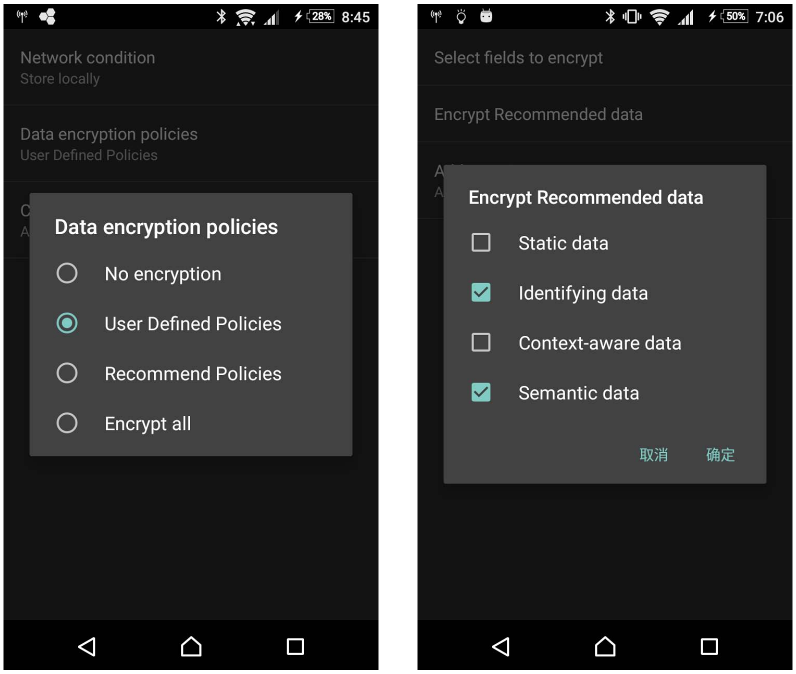
In our prototype implementation, smartphone not only integrates data from the vehicle and the smartwatch, but also provide abundant data such as IMU sensor data (accelerometer, gyroscope, magnetic), location data (Internet location and GPS location), network information (cellular network and Wi-Fi network). To collect these data on an Android device, Service Managers need to be registered and requested from $getSystemService()$ function respectively for different kinds of data. With a Service Manager, the data can be obtained by reading corresponding member function.

Collecting Data on Smartwatch

Compared to smartphone, smartwatch can provide more data about a driver's status, such as heartbeats and core body temperatures. In our implementation, an Android smartwatch is used. The approach for collection sensor data on a smartwatch is similar as the approach on a smartphone. In order to transmit smartwatch data to the smartphone, the smartwatch and the smartphone need to be connected through the Android Wear application[2] . Then a $DataApi.MessageListener$ on the smartphone will listen data messages from the corresponding smartwatch, and the data collected

[1] http://openxcplatform.com/

[2] https://developer.android.com/wear/index.html



Network condition
Store locally

Data encryption policies
User Defined Policies

Data encryption policies

○ No encryption
● User Defined Policies
○ Recommend Policies
○ Encrypt all

(a) Fine-grained uploading options

Select fields to encrypt

Encrypt Recommended data

Encrypt Recommended data

☐ Static data
☑ Identifying data
☐ Context-aware data
☑ Semantic data

(b) Recommended settings

Figure 5.2: Fine-grained uploading policies

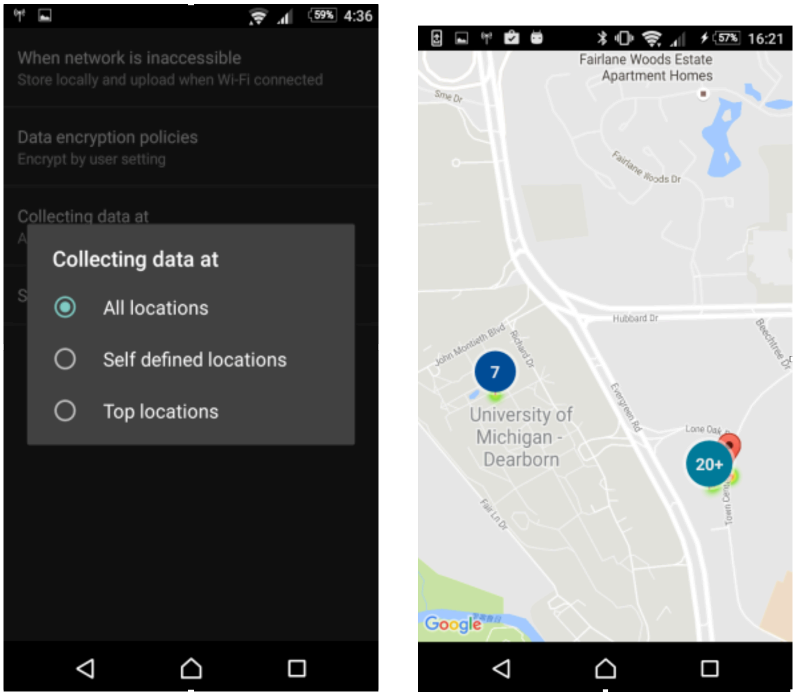
on the watch is send by a $GoogleApiClient$.

Integration from Multiple Sources

To integrate data from different sources and upload them to remote cloud, we set up a timer in the application on smartphone. Each time the timer times out, a event based function will be invoked to upload the latest data. The frequency for the timer can be determined by the user or according to service requirement.

5.3.2 Implementation of Data Uploading

To implement the two privacy preserving data uploading policies mentioned above, we set up a remote server on an Amazon Elastic Compute Cloud. The server will receive data from smartphone clients and store data in database.



When network is inaccessible
Store locally and upload when Wi-Fi connected

Data encryption policies
Encrypt by user setting

Collecting data at

Collecting data at

● All locations

○ Self defined locations

○ Top locations

Fairlane Woods Estate Apartment Homes
Fairlane Woods Dr
Hubbard Dr
John Montieth Blvd
Evergreen Rd
University of Michigan - Dearborn
7
20+
Google

(a) Context-aware uploading policies

(b) Top location identification

Figure 5.3: Context-aware uploading policies

Fine-Grained Uploading

To implement the fine-grained data uploading policy, a set of options are provided to users so that they are able to select which (or which group of) data needed to be encrypted. Here we just assume that users encrypt data for preventing the data from being eavesdropped or stolen by adversary. Without loss of generality, the operation can also be removing some specific data or not to collect some specific data.

As shown in Figure 5.2(a), four options are provided in the user interface: no encryption, users-defined encryption, recommended encryption, and all encryption. For the user-defined encryption option, users can choose which kinds of data they need to encrypt. But as discussed in a previous study [ABL15], users could be uncertain about the consequence about their data sharing and be unaware of whether sharing some data will lead to privacy leakage. So we also provide recommended encryption settings to users so that they can choose to encrypt sets of data according to the privacy characterization, as shown in Figure 5.2(b). The encryption techniques used to encrypt data is RSA encryption in our implementation, but in practice, other techniques can be used according to the design and requirement.

Context-Aware Uploading

The context-aware uploading policy is also based on the location protection, since locations are highly context dependent, as we discussed above. So a set of options that decide which places/areas need to be protected are provided in our implementation, as shown in Figure 5.3(a). The initial option is to allow data collection at all places. According to users' preference, the self-defined locations option allows users to set locations where they don't want to share data. So users can add locations by geo-coordinates and set a label for each locations. Meanwhile, as proposed in [Lan15], users may tend to ignore privacy protection measures due to the operation cost of settings. So a policy based on the top locations of a is also provided. The user's historical traces are stored in SQLite Database in Android system, and top locations of the user are computed on the historical traces using clustering algorithms (locations are clustered into clusters and the top k clusters are recognized as top k locations). After that, when the user drive into the area near the top locations, the data uploading will be paused or the data will be encrypted before uploading, so that the information around the user's top location can be preserved from remote server. An illustration of the top location option's interface is shown in Figure 5.3(b).

Besides, considering the resource efficiency and network usage, we also provide the option to store data locally when the Wi-Fi is inaccessible. Users can choose real-time data uploading under different network conditions (Wi-Fi/Cellular Network). When the required network is not inaccessible, our application will store the collected data locally in SD card of the smartphone, and upload the data at a time when the network is connected.

CHAPTER 6

Privacy Preserving Data Sharing in Connected Vehicle Services

Upon the given access to the data from the above application scenarios, a potential attacker largely has two strategies depending on how they acquire the information to breach user privacy. We define them as *homogeneous information aggregation* and *heterogeneous information aggregation*.

Homogeneous information means different data items from the same category of data. Attacks based on homogeneous information aggregation try to infer additional information by accessing the same data over a sufficiently long period. Unless there is a limitation in the amount of results for the database query, an attacker can collect enough amount of data to infer extra information about a user. For example, in a UBI scenario if an insurance company can query the GPS location of a vehicle, the company will be able to reconstruct the location history of a driver and infer meaningful information about a driver. By aggregating heartbeat rate, blood pressure, and glucose level of a user for a long time in drowsiness detection, an attacker will be able to even infer certain potential health issues of an individual. In a fleet management scenario with the driver information anonymized for privacy protection, if an attacker is able to access some vehicle sensor data over certain amount of time, the attacker will be able to fingerprint an individual, thus de-anonymizing the fleet driver's identity.

Heterogeneous information means information from different categories of data. A data consumer typically should not be allowed to query data which she is not authorized to access. Unauthorized access can be easily prevented by enforcing finely-granular data access control. However, if multiple data consumers collude with a malicious intent, heterogeneous information will be aggregated, thus breaching user privacy. In a smart mobility scenario, for example, if a credit card company provides user information to a drive-in business, the critical user privacy can be leaked together with the vehicle identity. In a fleet management scenario, it is also likely to infer the identity of users when multiple sources of anonymized information are aggregated.

In practice, an attacker uses both strategies for information inference. The longer time (the more data) the attacker has access to the data, the more types of data the attacker can access, the higher winning chance for it.

More specifically, we look at two scenarios, personalized mobility service and group mobility service.

In personalized mobility service applications, identity of user is usually not a concern as user identity is required for a user to enjoy personalized services. However, a user may worry about extra information leakage due to multi-sources of data aggregation over a sufficient long period. As illustrated in Fig. 6.1, a user's data are shared to some data consumers (e.g., data d_1, d_2, d_3 in Fig. 6.1) but the data consumers can infer additional private data (e.g., data d_x) using the shared data. This type of privacy breach can happen in scenarios including usage-based insurance, dangerous driving detection, and emerging smart mobility services.

In group mobility service applications, data consumer is allowed to access multiple user data in order to derive useful group information. However, the service provider should not be allowed to associate data with an individual user. As shown in Fig. 6.2, driver id should be anonymized before sharing data, but the data consumer is able to link data to a specific driver (e.g., the driver with id i_x in Fig. 6.2), thus breaching the privacy. This type of privacy leakage can happen in fleet management or real-time traffic monitoring, where drivers' id should be anonymized.

In the following sections, we present our case studies on these two types of privacy leakage respectively.

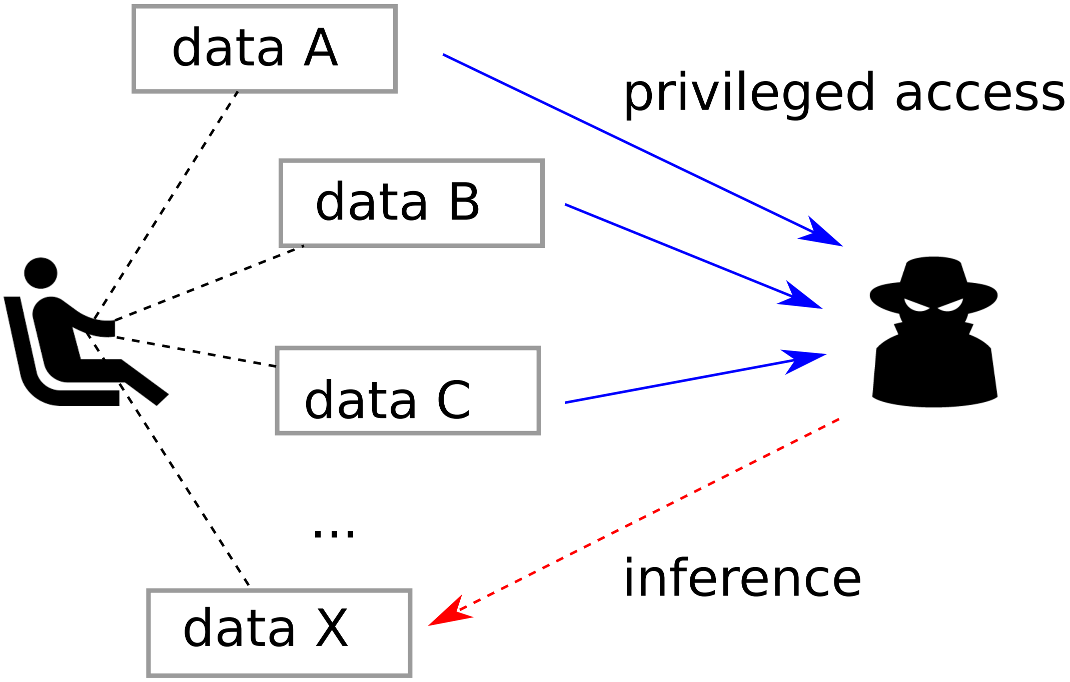


Figure 6.1: Privacy risk by information inference

6.1 Attacker Model

In this chapter, we will show two potential privacy leakage with notable impact on real-world connected vehicle services due to data sharing. In our model, we consider two business entities in the new mobility ecosystem: data producer and data consumer. Data producers collect data and store data on cloud server, and data consumers can query necessary data so that they can provide customized or group services accordingly. In our model, we assume the data producer is trusted and data is securely collected and stored in the cloud server [LMM+17; KSL+14; KSL+17] (the design and privacy issues of the data producer in Figure 3.1 is out of the scope of the paper). Some traditional OEMs (such as Toyota, Ford, etc.) which are aggressively building mobility platform and transforming themselves into emerging mobility business are examples of data producers in the future transportation ecosystem.

Service providers (data consumers) are semi-trusted. Although they receive data from the data producers and use the data to provide new mobility services, they are not trusted in not extracting

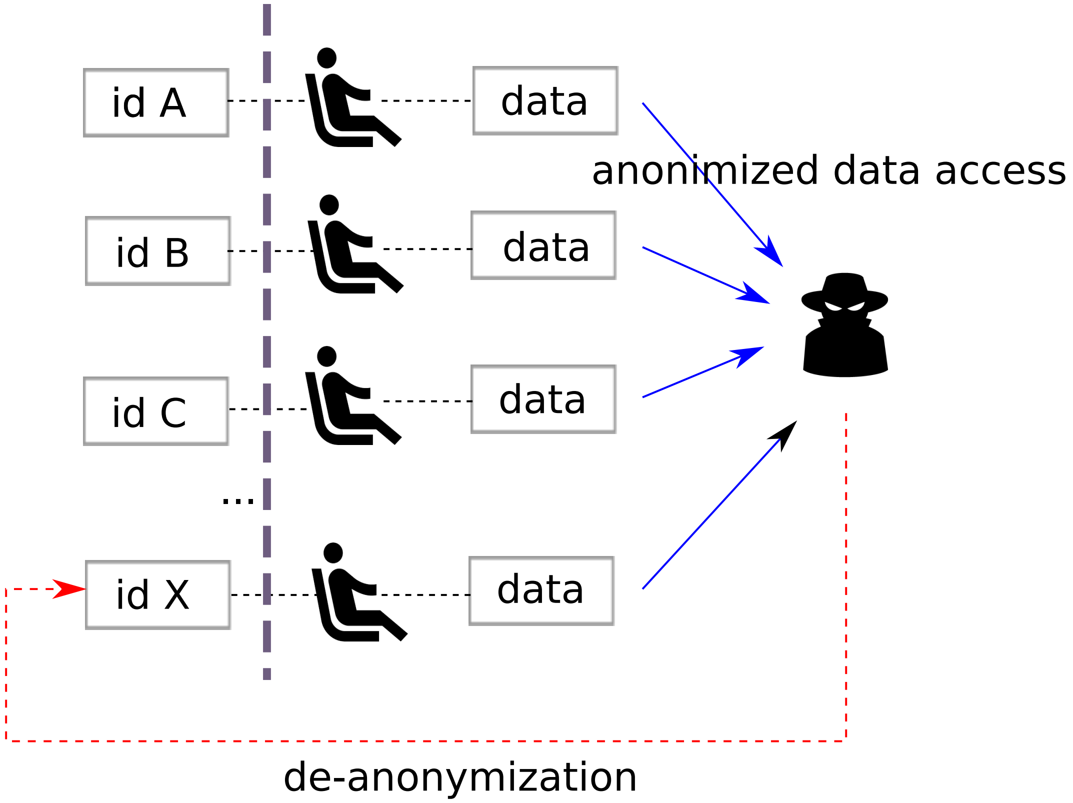


Figure 6.2: Privacy risk by de-anonymization

extra information from the data they receive. This means either they may try to leverage the data with malicious purposes besides normal usage for providing services (e.g., tracking individuals' activities and inferring individual's personal identifiable information), or the data can be leaked from data consumers to malicious parties who are aiming at discovering users' privacy without proper data management. Insurance companies and car sharing companies are examples of data consumers. In the remaining part of this paper, we also denote the malicious data consumers as attackers for convenience.

6.2 Privacy Risk of Top Locations Inference

Inference attack aims at illegitimately inferring or learning extra sensitive information about an individual with a high confidence through data analysis and data mining techniques [Kru07; LXZ+16;

LZM17; CMC+12; YGZ+17]. An attacker already knows the identity of victim through the result of data query. The goal is to infer extra information out of the given data query result, as shown in Fig. 6.1. Inference attack is usually neglected and hard to defend against because sensitive information can be inferred by seemingly non-sensitive data needed by applications.

In the context of connected vehicles, mobility location of an individual is a type of important information that should be paid attention to, because there are many applications, in which drivers' GPS locations [TDK+11] or vehicle/mobile device data that can be used to recover driving path [GFS+14; NVB+16; ZCL+17] are required to be shared. For example, some usage based insurances evaluate the premium based on how long a driver drive every day; connected vehicle applications require vehicles' real-time locations for communications.

Location traces are possible to expose sensitive information of individuals [Kru07; ZB11; LZD+16]. For example, after collecting comprehensive mobility traces of a driver, **top locations** are very likely to be inferred, and top locations are most correlated to users' identities [Kru07; ZB11]. Moreover, top locations can be linkable to her demographic information, hobby, and social relationship [LZD+16]. The privacy issue due to location sharing is called location privacy. The challenge here is how to protect individuals' *critical location privacy* from being inferred through data sharing while keeping usability of connected vehicle applications.

In this section, we present an attack to show the potential top location leakage during the location traces sharing. Here the location traces can be direct GPS locations sharing or indirect trace inference through driving sensor data [GFS+14; ZCL+17].

6.2.1 Problem Modeling

A user's trace refers to her mobile movements along the spatial and temporal domain. In our problem, a trace consists of a set of discrete GPS coordinates collected by vehicular sensors in time sequence. We model a trace of a specific user u as a function mapping timestamps in $\mathcal{T} =$

$\{t_1, t_2, ..., t_m\}$ to the user's GPS coordinates $\mathcal{M} = \{..., < t_i, lat_i, lng_i >, ...\}$, where $t_1 < t_2 < ... < t_i < ...t_m$ are in order.

From traces $\mathcal{M}$ of a user u, data owners can obtain an aggregate view on the user's mobility pattern by building the location profile, which includes the user's visited location set $\mathcal{L} = \{l_1, \ldots, l_n\}$ (by aggregating GPS coordinates nearby) and its discrete probability distribution as $\theta_i = P(l_i)$. Here $P(l_i)$ denotes percentage of visits of l_i among $\mathcal{M}$. So now our goal is to infer the top locations based on the location profile. One insight is that the most frequently visited locations are usually the top locations of an individual [Kru07].

6.2.2 Inference Algorithm

In order to achieve this goal, we estimate top locations for an individual's location traces based on k-means clustering algorithm. The k-means clustering aims to partition all the locations into k sets so as to minimize the within-cluster sum of squares of distance between each pair of points. So naturally, centroids of clusters are shifted toward to centers of dense data points, which are probably the most frequently visited places in our problem, so that within-cluster distances are minimized. Besides, as a previous work [ZB11] pointed out, "top N" locations can be interpreted as semantic information at different granularity levels (e.g., "top 2" locations are likely to be home and work locations, "top 3" locations probably correspond to home, work, and shopping/school/commute path locations). The k-clustering algorithm allows adjusting the parameter k to meet the requirements on inference granularities (i.e., how many top locations the attacker want to learn). Based on these intuitions, the centroids of clusters are estimated as closed to the top locations. Formally, given a set of GPS traces $\mathcal{M}$, the k centroid points $\mu_1, ..., \mu_k$ from the k clusters $\mathcal{C} = \{C_1, ..., C_k\}$ are computed so that:

$$\arg\min_{\mathcal{C}} \sum_{i=1}^{k} \sum_{l \in \mathcal{C}_i} Dist(l, \mu_i) \tag{6.1}$$

where $Dist(\cdot)$ compute distances on the earth sphere according the the latitude and longitude of two GPS coordinates.

6.3 Privacy risk by de-anonymization

In order to preserve individuals' privacy, an attacker sometimes has only anonymized DB query results. However, sufficient amount of driving data (speed, acceleration, brake) will allow an attacker to fingerprint each individual, and then an attacker will be able to deanonymize an individual later by a means such as machine learning classification algorithms [ETK+16; CCS17], as illustrated in Fig. 6.2. For example, in car fleeting services, driving data can be used to examine whether it is the same driver driving the car. In the mix zone technique [BS04], users' identifiers are hided, but if we can identify a specific user using driving data, the mix zone technique can be bypassed. So the fingerprinting attack brings privacy leakage risk. Previous work have considered to fingerprint drivers using sensor data [ETK+16]. However, the results are only evaluated in experimental environment with complex parameter optimization. In their experiments, drivers drove on different days respectively, which can discriminate driving due to road condition, traffic, etc., though the data collection during the same time of day for each driver. The more important is data were collected from each driver in a single day, but driving behaviors of a driver might change on different days because of different moods, emotions, or health conditions. So without collecting a driver's data from different days, it is hard to justify potential false positives. As a result, these facts could limit the attack in real-world scenarios.

In this paper, we consider real-world scenarios like the mix zone. Before driving into a mix zone, the IDs of drivers can be observed, and the driving data are being collected. After driving into the mix zone, new randomized IDs are assigned to drivers in this area to anonymize them. But we can still use the driving data to fingerprint the anonymous drivers, thus breach the mix zone. So in our problem, the fingerprinting is considered in each area (i.e., a piece of road). We select all drivers who have passed an area and denotes these drivers as a set $\mathcal{U}$. Given a series of driving data D collected from $\mathcal{U}$, we try to identify which driver in $\mathcal{U}$ is the owner of D. So this problem is formulated as a classification problem: each driver u in $\mathcal{U}$ has an ID, and has some historical driving data which has been collected. So we can extract features from the driving data, and train a machine learning model to correlate the driving data to the driver's ID. Now, when an anonymous driver driving through this road once again, we input the newly collected data into the model, and identify the ID of the anonymous driver based on the historical data. The insight is that each driver has its own unique driving behavior, which can be reflected on the driving data.

6.4 Countermeasures Methodology

In this section, we provide methods for thwarting privacy risks mentioned above.

6.4.1 Defending Inference Attacks in Location Sharing

To minimize leakage of location information, location obfuscation methods were proposed targeting at different scenarios, such as location-based services (LBS), mobile social networks, mobile device applications [ACV+11; STT+12; CDM15; STT17]. However, existing solutions may not be able to resist inference attacks (e.g., top location inference) while maintaining data utility. This is because obfuscating each location will not significantly change the overall profile (e.g., distribution, pattern) of location traces. To tackle this problem, shifting location coordinates towards relatively non-sensitive area (e.g., public area) was proposed in [LZD+16]. This can be a promising approach for LBS and mobile social networks check-ins. However, driving traces are continuously monitored in some scenarios, shifting locations towards several assigned areas will cause inconsistency for the driving data, and further affect corresponding services. Given the fact that drivers' locations are collected in so many scenarios, there is a dilemma (or a trade-off) between defending against location inference attacks and maintaining data utility.

In this section, we propose two different location sharing policies to balance the trade-off. Our goal is to increase the bar for top location inference by sacrificing a small portion of data.

Two-End Data Removing Policy

Our first policy proposes to remove a certain portion of data from both ends of a trip, i.e., starting point and destination of each trip. Formally, a trip here is a set of consecutive GPS data during a driving, i.e., $Trip = \{< t_1, lat_1, lng_1 >, < t_2, lat_2, lng_2 >, ..., < t_n, lat_n, lng_n >\}$, where $t_1 < t_2 < ... < t_n$ are in order. And the policy will discard data at the beginning, i.e., $Trip_b = \{< t_1, lat_1, lng_1 >, ..., < t_{i-1}, lat_{i-1}, lng_{i-1} >, < t_i, lat_i, lng_i >\}$, and the end, i.e., $Trip_e = \{< t_j, lat_j, lng_j >, < t_{j+1}, lat_{j+1}, lng_{j+1} >, ..., < t_n, lat_n, lng_n >\}$, of the trip. Here, indices i and j control potions of data that are going to be removed, and $Trip - (Trip_b \cup Trip_e)$ are the remaining data after processing by this policy.

The motivation of this policy is based on the observation that the start points and end points of a user's driving trips are PoIs that the user stays for a relatively long time. So removing a certain data points from the beginning and end of each trip will hide the exact locations of these PoIs, while retaining majority of data during the trip for data consumer. This policy is especially suitable for the scenarios that the driver has regular driving patterns between important locations. For example, a driver who works regularly drives from home to work place in the morning, and drives back home in the afternoon. The driving pattern is likely to repeat in weekdays though the routes between home and work place and may change due to factors such as traffic condition and intermediate visits. So this policy is able to preserve locations at both ends but retain intermediate driving data for high-quality services.

Context-Aware Data Sharing

In this section, we improve the end data removing strategy and show how to intelligently remove the portion of end data based on data usage pattern of the consumer (or attacker).

Context-aware (or context-dependent) means that individuals' data can, depending on different situations, have different privacy preferences and be treated with different actions. For example, given a set of GPS coordinates, some of them may indicate sensitive locations, while some of them are not sensitive, so GPS data are context-aware. It is important to understand whether a specific set of data is sensitive to individual's privacy before publishing the data, and a potential solution is to remove the sensitive part for preserving privacy while releasing the remaining non-sensitive part for fulfilling usability. Based on the motivation, we propose to consider top locations (e.g., most frequently visited locations) as sensitive locations to an individual, and identify whether a set of data is likely to expose top locations of this individual. There are different approaches to obtain top locations of a single user, such as allowing users to choose several important locations they want to preserve, analyzing users' mobility patterns to find out daily visited positions, etc. Note that in our attacker model, the data storage on cloud servers is secure, so we assume the storage or computation for ground truth top locations on server is secure.

Given top locations of a user stored on cloud server, we are able to evaluate whether a set of location data queried from a specific query is likely to reflect the user's important locations. In order to prevent the exact top locations from being inferred, we propose not to share data generated in a range around top locations. Formally, Given a top location's coordinate of the user is $< lat_{top}, lng_{top} >$, the data within a certain range R, i.e., $Dist(< lat_{top}, lng_{top} >, < lat_i, lng_i >) < R$ will be removed and not be shared to data consumers. In this way, it is expected to hide top locations' information by sacrificing data nearby. The evaluation of the two solutions will be presented in Section 6.5.2.

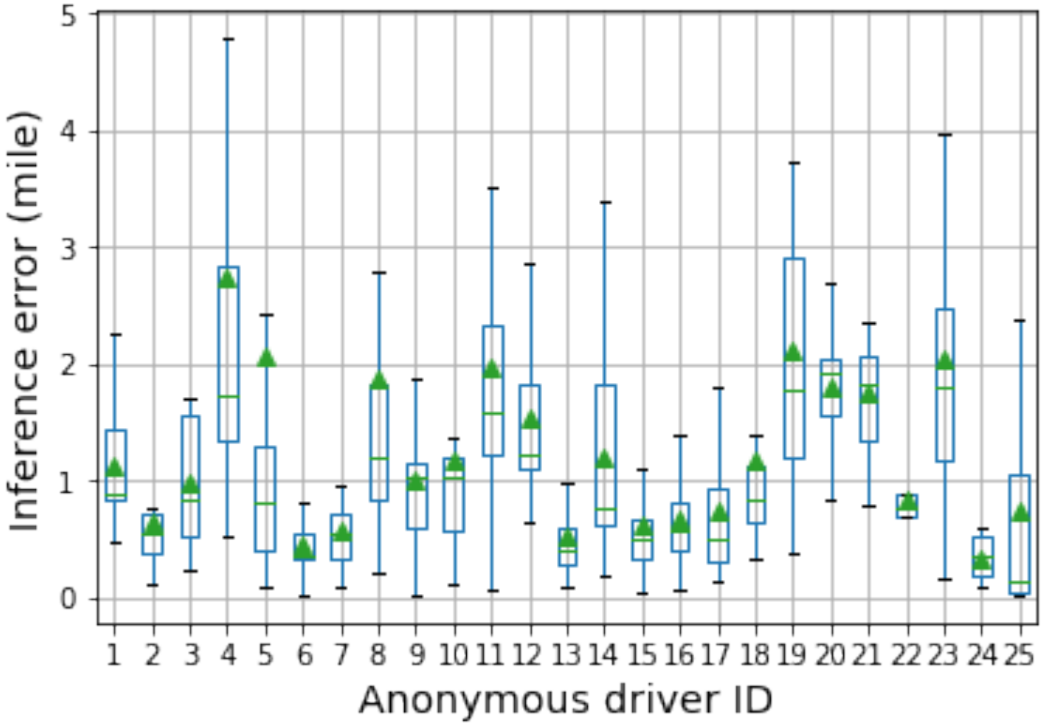


Figure 6.3: Top locations inference using daily data

6.4.2 Defending Fingerprinting Attacks

Adding Controllable Noise

Since unique driving behaviors extracted from driving data can be used to fingerprint and discriminate different drivers, adding extra noise to hide the discrimination is a potential solution. However, adding noise can reduce the utility of data for applications. So a solution that can control both privacy and utility is desired.

To resolve the dilemma, we adopt an approach, which is inspired by an approach [DBC16] based on differential privacy, that randomly selects noise from a Laplace distribution. According to the definition of differential privacy [Dwo11], a randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and all $S \subset Range(\mathcal{K})$:

$$Pr[\mathcal{K}(D_1) \in S] \leq e^{\epsilon} \times Pr[\mathcal{K}(D_2) \in S] \tag{6.2}$$

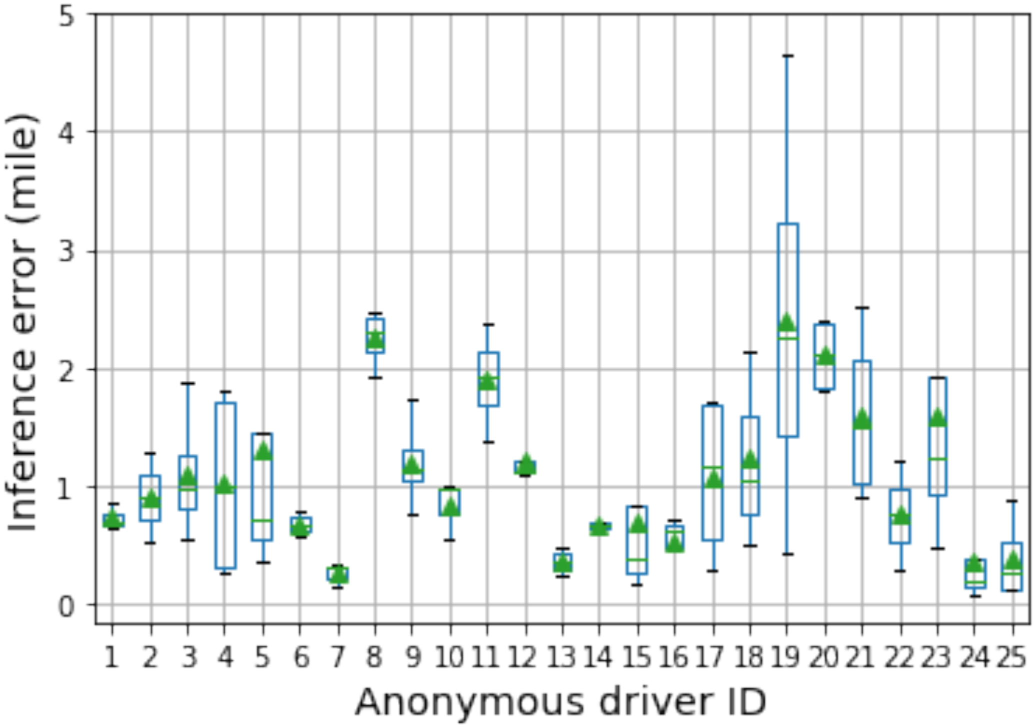


Figure 6.4: Top locations inference using weekly data

Specifically, if we denote $f(\cdot)$ is the original function for data query, X is the distribution of noise, we can rewrite Equ. 6.2 as:

$$\forall t, \frac{Pr[f(D_1) + X_1 = t]}{Pr[f(D_2) + X_2 = t]} \leq e^{\epsilon} \tag{6.3}$$

which means after adding noise, the query on D_1 and D_2 should return same output distribution. Let $d = f(D_1) - f(D_2)$, the distributions of noise are:

$$\forall x, \frac{Pr[X_1 = x]}{Pr[X_2 = x + d]} \leq e^{\epsilon} \tag{6.4}$$

where d is less or equal than the sensitivity Δf, i.e., $d \leq \Delta f = |f(D_1) - f(D_2)|$. So now, we know the distribution of noise is exactly Laplace noise $Lap(\frac{\Delta f}{\epsilon})$, i.e., $Pr[Lap(\beta) = x] = \frac{1}{2\beta}e^{-|x|/\beta}$, where $\beta = \frac{\Delta f}{\epsilon}$.

The similar idea can be adopted into our problem. Before statistics of driving data are published or queried, the noise is randomly selected from the Laplace distribution, $Lap(\mu;\beta)$ where μ = $mean$(statistics of data) and β = (max(statistics of data)-min(statistics of data))/ϵ, for each kind of statistics. So now, $\mathcal{K}$ is the process of selecting random noise, S is the outcome of applying random noise to driving data statistics, D_1 and D_2 are data sets before and after adding noise, respectively. Since the parameter ϵ determines how much difference exists between D_1 and D_2, if ϵ is smaller, the noise added onto data is heavier, and the data utility would reduce more, and vise versa. So by changing ϵ, we can control to what extent two statistics distributions are alike, and further control the balance between privacy and data utility.

Frequently-Changing Pseudonyms

The second defense is to use frequently-changing pseudonyms for “shorter" trip segments in order to increase the bar for an attacker to correlate these trip segments belonging to the same vehicle. Pseudonyms are commonly accepted as a mechanism to enhance anonymity and protect location privacy. In our problem, frequently-changing pseudonyms can prevent an attacker who uses homogeneous aggregation strategy over vehicular data sets. Since we assume that cloud server is secure in our attack model, it will be much more difficult for an attacker to build a proper fingerprint for a driver when a trip is segmented by multiple pseudonyms. Even an attacker can obtain some data to build driving profile of a user, data segments with less data points may not provide enough information for an attacker to accurately identify anonymous users.

6.5 Experimental Results

In this section, we mainly introduce our evaluations for countermeasures of the two privacy risks through data sharing using real-world vehicle datasets.

6.5.1 Dataset

Our experiments are based on BSM data collected in the Safety Pilot Model Deployment (SPMD) project [BS14]. In the SPMD project, driving data was collected from connected vehicle to find out how well safety technologies and systems work in a real-life environment with real drivers and vehicles. Basic driving data include timestamps, GPS location and heading, speed, brake pedal status, acceleration, yaw-rate, etc. These data are collected from drivers recruited by University of Michigan Transportation Research Institute (UMTRI). In this paper, we use a subset of data from Safety Pilot dataset. For location privacy inference, individual data will be studied independently. We mainly use GPS tracking coordinates to study how privacy-preserving policies affect the top location inference. The data we used involve 25 drivers who drove more than one week within one month (Apr. 2016) and their driving GPS data. Totally, 14,558,429 GPS records are included during 407.8 hours' driving. For fingerprinting, multiple user data are studied simultaneously. The dataset contains 43 drivers and 2,648,543 records of driving data. Driving data (listed in Table. 6.1) are used to perform fingerprinting attacks and corresponding countermeasures, and anonymized ids of drivers are used as ground truth to evaluate performance.

6.5.2 Top Location Inference and Preservation

Attack Evaluation

We evaluate the top location inference attacks based on a dataset containing 51 drivers' driving data within one month (Apr. 2016). In order to avoid bias due to small sample size, we only select drivers who drove more than one week (seven days) during April 2016. Then, due to the difficulty of obtaining top locations of each driver, we use residence/home to represent a kind of top locations in our experiment. The ground truth top locations are determined by manually examining each driving trip which is visualized on Google map. If most of trips of a driver start or end at a place near residential area, we recognized this place as this driver's residence, i.e., top locations

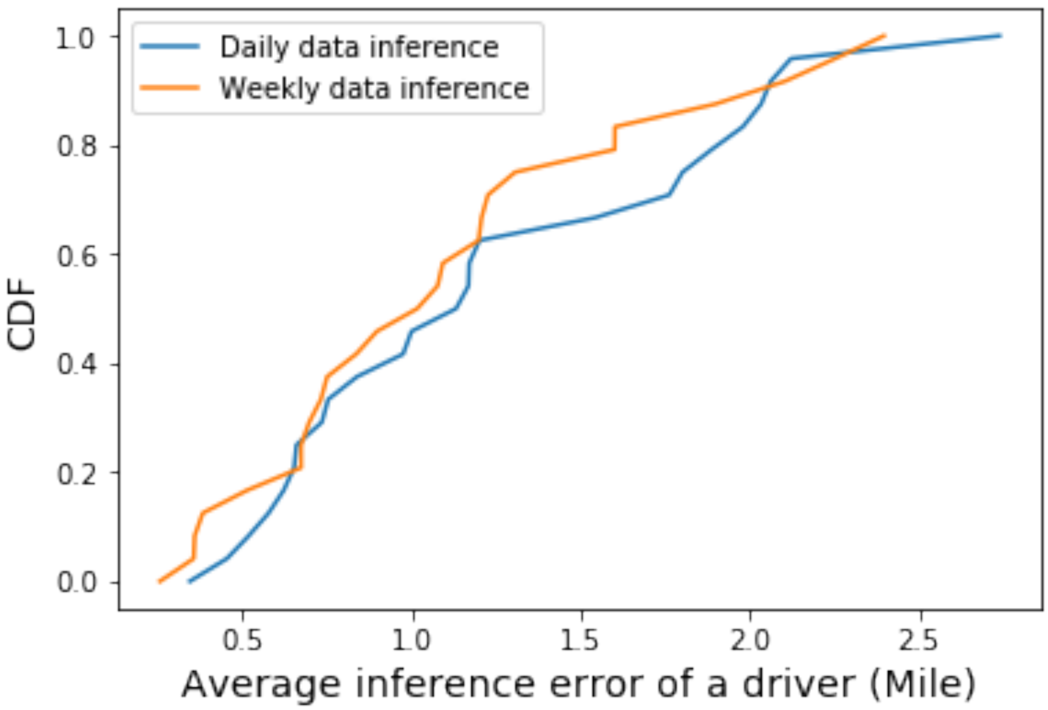


Figure 6.5: Cumulative distribution function (CDF) of average inference error

in the experiment. If a driver's top location can not be identified with confidence, this driver will not be considered into experiment. Finally, 25 out of the 51 drivers meet the requirement and are selected in our experiments. Then we perform the algorithm mentioned in Section 6.2 on each drivers' driving routes. The number of centroids is set as 3, i.e., $k = 3$ in Equ. 6.1. This is because humans' mobility patterns were considered to contain "top 3" locations: home, work, and shopping/school/commute path locations [ZB11].

We first consider to use daily driving data to perform inference on each driver respectively. This is to assume that an attacker can only obtain small amount of data, i.e., data generated in a day. Fig. 6.3 shows the mean (green triangle), minimum, first quartile (bottom edge of the box), median, third quartile (top edge of the box), and the maximum of inference error for trips of each driver respectively. Here the error is defined as the distance between the ground truth top location and its nearest inferred top location. Within the 25 drivers, 12 drivers have an average inference

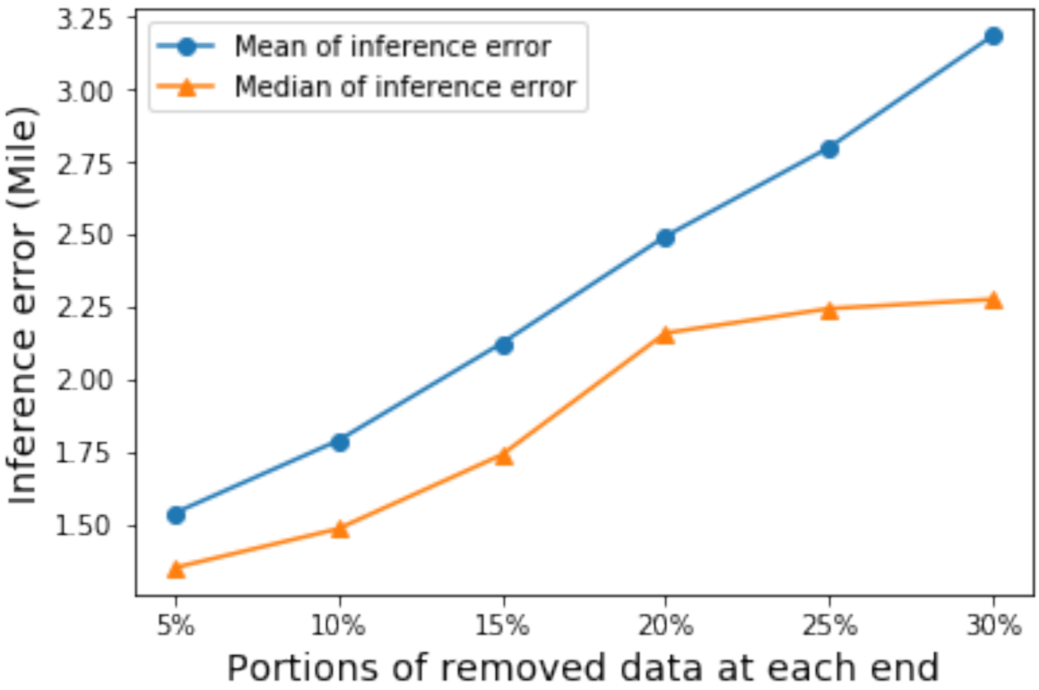


Figure 6.6: Experiments for two-end data removing policy

error less than one mile, which means their top locations can be guessed within a relatively small area. Overall, the average error for the 25 drivers top location inference is 1.23 mile and Fig. 6.5 shows cumulative distribution function of average inference error of each driver.

As discussed in Section 3.3, data consumers can aggregate more homogeneous information according to their capability. Here, the GPS trace is the homogeneous information that some data consumers (e.g., UBI, mobility service providers) request. So we also study whether more data will increase the privacy leakage. Assuming the party is able to collect data for a week, we separate data in one month into four weeks and perform inference experiments on location data points within each week. Fig. 6.4 shows the results. Compared with the results in Fig. 6.3, 14 drivers have lower average inference errors, and the total average error of all 25 drivers is 1.08 mile, which is also lower. Cumulative distribution function in Fig. 6.5 further supports the conclusion that given weekly data, drivers' top locations can be more accurately estimated. This is because more data will expose more about a driver's mobility patterns and location profile, thus frequently visited locations, such as residence, can be better estimated by algorithms.

Defense Evaluation

Firstly, we evaluate the two-end data removing policy. The trade-off here is how much data needed to be removed and how far the inferred top locations will be shifted after removing some data at the two ends of a trip. Fig. 6.6 shows the results of inferring top locations using daily driving data when different portions of data are removed from each trip's beginning and end. It can be seen that the average inference error increases proportionally when increasing the amount of data removal. When we remove 15% data from the beginning and 15% data from the end of each trip (30% of data are removed in total), the average inference error is 2.124 miles, which is obviously larger than original results.

Then we evaluate the context aware data sharing policy. In Fig. 6.7, given different ranges of data removing around top locations (from 500 meters to 3000 meters), the blue line shows the average inference error and the red line shows the average percentage of data which are removed within the range. We can see that when the range is set as 2500 meters (around top locations), the inference error is around 2.55 miles and the data loss is only 20%. The result shows that, compared with two-end data removing, the context aware data sharing can be more specifically targeted to resist top location inference by sacrificing less data. By scarifying a small amount of data, our goal is to avoid important location leakage to secure individuals' privacy. And further balance the right of privacy protection for customers and the usage of data for data consumers.

6.5.3 Fingerprint Attacks and Defense

Attack Evaluation

We first introduce experiments and results of our fingerprinting attacks.

Experiment settings: To evaluate the potential risks, we performed four (independent) experi-

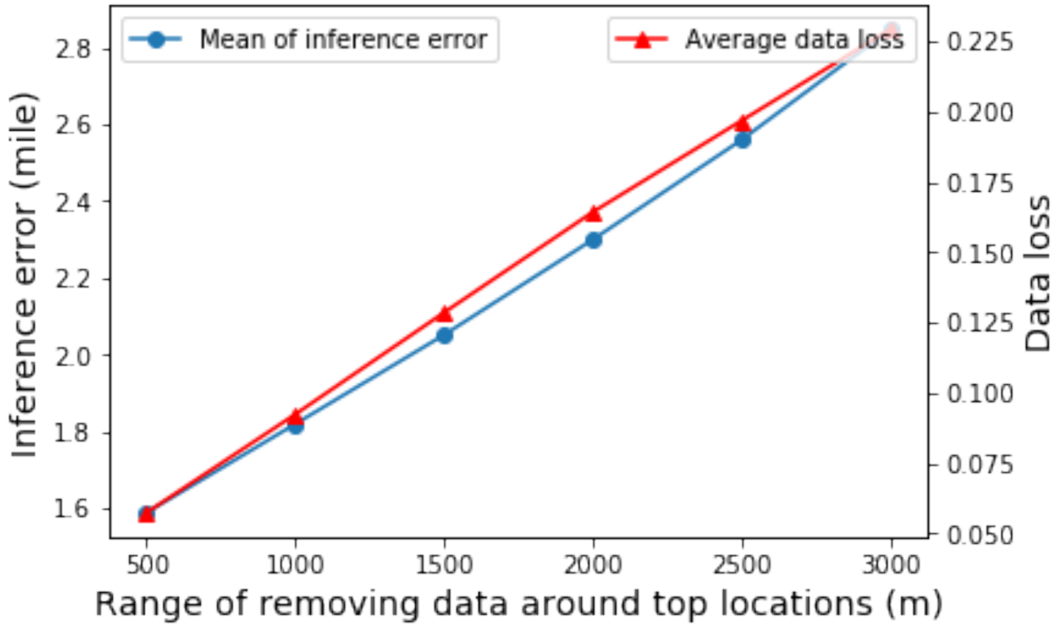


Figure 6.7: Experiments for context-aware data sharing policy

ments by selecting four roads around Ann Arbor, Michigan, US (Road 1-4). The average distance of driving trips through each road varies from 0.37 miles (short road) to 2.6 miles (long road), as we want to find out whether small scale of driving data can be used to fingerprint drivers (Fingerprinting attacks for long distance driving using ODB-II data have been validated in previous work [ETK+16]. However, it is suspected less data can intuitively reduce the success rate of the attack proposed in it). The number of drivers who have been taken into consideration is shown as "Num of Drivers" in Table 6.1, which is the number of drivers who passed this road for more than ten times in the testing set to minimize bias. Baseline is selected by randomly guessing an id for the testing data. This is to assume an external attacker does not know any information so it can only randomly guess who generates these data.

Feature and model selection: The driving data in our data set include: accelerator pedal, longitudinal acceleration from Conti IMU, lateral acceleration from Conti IMU, brake light active, vehicle speed from transmission, turn signal reading, yawrate from Conti IMU, left boundary signal, right boundary signal. But not all of them contribute to the best fingerprinting performance, so

Table 6.1: Evaluations of fingerprinting attacks

Experiments	Road 1	Road 2	Road 3	Road 4
Num. of Drivers	10	6	15	12
Random Guess (Baseline)	0.10	0.17	0.07	0.08
Fingerprint Attacks	0.83	0.80	0.81	0.86
Using single sensor to do fingerprinting				
Accel. Pedal	0.47	0.6	0.3	0.36
Longitudinal Accel.	0.57	0.43	0.24	0.26
Lateral Accel.	0.64	0.72	0.34	0.26
Boundary Left	0.37	0.29	0.28	0.28
Boundary Right	0.41	0.41	0.25	026
Brake	0.26	0.29	0.18	0.20
Speed	0.47	0.39	0.3	0.34
Turn Signal	0.31	0.38	0.20	0.21
Yaw Rate	0.65	0.63	0.33	0.28

we use Filter method [GE03] to filter out redundant sensor data and find an optimal combination of sensor data. Finally, six sensor data types (accelerator pedal reading, lateral acceleration, speed, turning signal reading, yawrate) are selected in our experiments.

To extract features for each driving trip from a sequence of sensor data, we calculate typical statistical features including minimum, maximum, average, standard deviation, kurtosis, skewness of sensor data of each driving across the roads. Since the fingerprinting attack is a classification problem as discussed in Section 6.3, we use 75% data (i.e., historical data with known id) as training set and use the remaining 25% data (i.e., anonymized data that we recover its id) as testing set.

Among various classification models, random forest classification model is chose as the classifier because its outstanding performance under most of scenarios. The choice of parameters of the random forest model is based on cross validation. The number of learning tree within the random forest model is set as 200, and entropy is used as criterion to split nodes in learning trees.

Results: Table 6.1 shows results of random guessing and our fingerprinting attacks for each of four experiments on four different roads respectively. The accuracy of all four sets of experiments exceed 80%, which is significantly higher than the baseline. The results reflect that the risks of de-anonymization using seemingly non-identifiable sensor data do exist. We also present the fingerprinting accuracy for each single kind of sensor data in Table 6.1. We can see if only using one kind of data to perform fingerprinting, inference accuracy is significantly lower compared with using multiple sensors data together. This shows the effects of heterogeneous information aggregation, as introduced in Section 3.3.

We also consider the data aggregation in time domain, which means different attackers can collect different amount of data to perform fingerprinting, according to their capabilities. In Fig 6.8, we select different percentage of data from driving traces as training sets, which corresponds different amount of data that can be obtained by attackers, and the remaining data as testing sets to evaluate the performance. The results illustrate that given more data, the attacker's inference accuracy increases (e.g., increased by 19% at most in Road 4). The results indicate that collecting more data (or performing long time tracking for anonymized drivers) will intensify the privacy leakage. However, even only given 10% of data traces, the inference accuracy is considerable, i.e., 73% averagely, which shows that small amount of data are enough to learn each driver's behaviors and distinguish different drivers. So countermeasures are required.

Defense Evaluation

We first validate the approach by adding controllable noise to experiments above. Table 6.2 shows the fingerprinting results after adding Laplace noise (ϵ is chose as $\epsilon = 4$) for the four experiments as an example. We can observe a significant decrease of fingerprinting attack accuracy. Fig. 6.9 shows the attack performance for varying ϵ. It illustrates that as we decrease the ϵ (i.e., as we increase the scale parameter of the Laplace distribution $Lap(\cdot)$), accuracy decreases. The figure clearly shows the trade-off between privacy protection through adding noise and the data utility

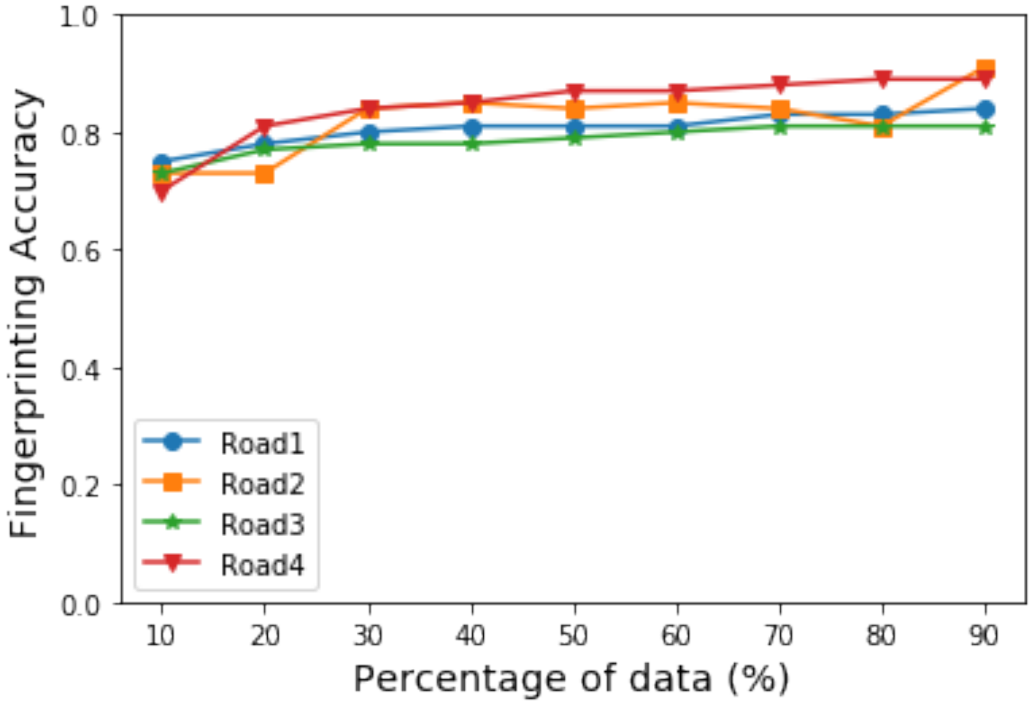


Figure 6.8: Results under different amount of data

Table 6.2: Evaluations of countermeasures for fingerprinting attacks

Experiments	Road 1	Road 2	Road 3	Road 4
Num. of Drivers	10	6	15	12
Random Guess (Baseline)	0.10	0.17	0.07	0.08
Fingerprint Attacks	0.83	0.80	0.81	0.86
Adding Noise ($\epsilon = 4$)	0.28	0.47	0.32	0.46
Frequently-Changing Pseudonyms	0.27	0.70	0.28	0.41

due to the noise. Adding more noise will reduce the possibility of identity inference, but it will change the data and affect usage in the mean time. The trade-off can be controlled by the parameter ϵ. So in a real world application, it is important to find a proper ϵ according to the scenarios and requirement of services.

Then, we evaluate effects of frequent-changing pseudonyms. We mimic an implementation

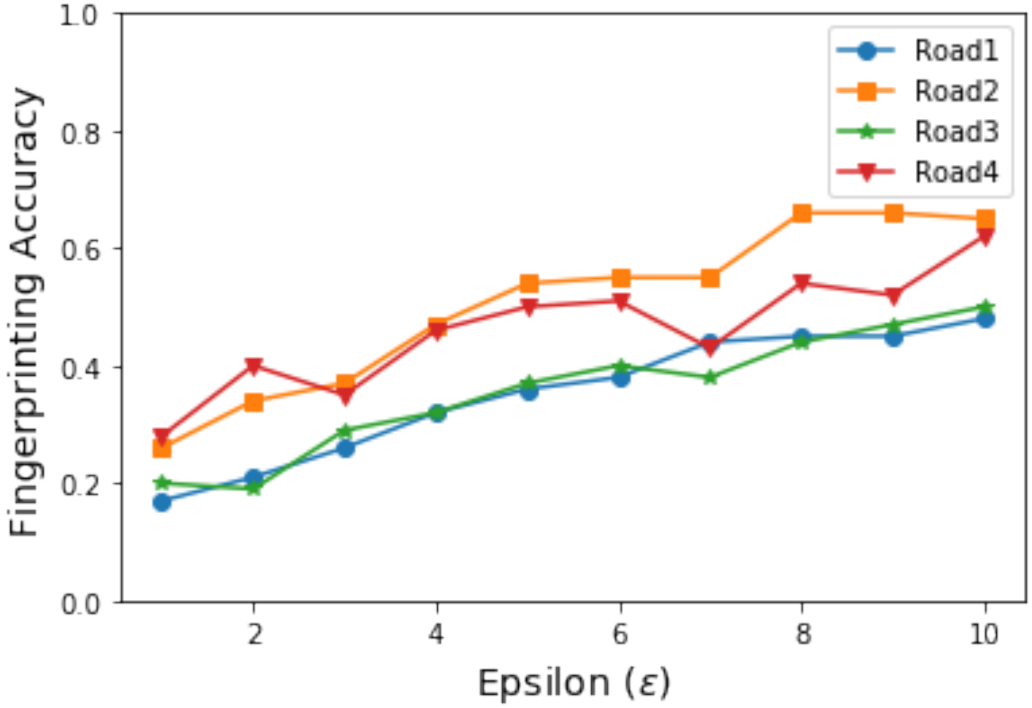


Figure 6.9: Results after adding Laplace noise

which assigns 10 pseudonyms to equal-length segments of data during each driving trip. And each segment is recognized as a single trip in the testing set, since the attacker can only observe data segments under pseudo ids. The results in Table. 6.1 indicate that the inference performances are reduced in varying degrees, especially in the cases that more than 10 drivers are considered (e.g., Road 1, Road 3, Road 4), the accuracy decreases to 0.28 and 0.41 respectively. This shows that frequent-changing pseudonyms make it difficult to re-identify a specific driver from a set of candidates.

CHAPTER 7

Conclusion

Data driven applications benefit user experience with connected and automated vehicle (CAV) technologies but also bring privacy concerns given huge amount of data being collected, utilized and potentially shared to third parties. In this thesis, we firstly provide an overview of the landscape of emerging connected vehicle and analyze potential data privacy under several scenarios. Our analyses and experiments show that direct and indirect privacy attacks are feasible in corresponding scenarios. To thwart privacy breaches while maintain data usability, we not only consider different policies to control at the data collection phase, but also propose different countermeasures to defend against attacks during data sharing. Our implementations and evaluations based on real world data validate our defense strategies and demonstrate the trade-off between user privacy and data utility.

Bibliography

[ABL15] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. “Privacy and human behavior in the age of information”. In: *Science* 347.6221 (2015), pp. 509–514.

[ACV+11] Claudio A Ardagna, Marco Cremonini, Sabrina De Capitani di Vimercati, and Pierangela Samarati. “An obfuscation-based approach for protecting location privacy”. In: *IEEE Transactions on Dependable and Secure Computing* 8.1 (2011), pp. 13–27.

[AHM+15] Moreno Ambrosin, Hossein Hosseini, Kalikinkar Mandal, Mauro Conti, and Radha Poovendran. “Verifiable and privacy-preserving fine-grained data-collection for smart metering”. In: *Communications and Network Security (CNS), 2015 IEEE Conference on*. IEEE. 2015, pp. 655–658.

[AR11] Mohammad Amin Assari and Mohammad Rahmati. “Driver drowsiness detection using face expression recognition”. In: *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*. IEEE. 2011, pp. 337–341.

[AS] S Abhishek Anand and Nitesh Saxena. “Speechless: Analyzing the Threat to Speech Privacy from Smartphone Motion Sensors”. In: *2018 IEEE Symposium on Security and Privacy (SP)*. Vol. 00, pp. 116–133.

[ASM+14] Elli Androulaki, Claudio Soriente, Luka Malisa, and Srdjan Capkun. “Enforcing location and time-based access control on cloud-stored data”. In: *Distributed Computing Systems (ICDCS), 2014 IEEE 34th International Conference on*. IEEE. 2014, pp. 637–648.

[BBO07] Mihir Bellare, Alexandra Boldyreva, and Adam O′Neill. “Deterministic and efficiently searchable encryption”. In: *Annual International Cryptology Conference*. Springer. 2007, pp. 535–552.

[BDO+04] Dan Boneh, Giovanni Di Crescenzo, Rafail Ostrovsky, and Giuseppe Persiano. “Public key encryption with keyword search”. In: *International conference on the theory and applications of cryptographic techniques*. Springer. 2004, pp. 506–522.

[BRS+17] Jasmine Bowers, Bradley Reaves, Imani N Sherman, Patrick Traynor, and Kevin Butler. “Regulators, Mount Up! Analysis of Privacy Policies for Mobile Money Services”. In: *Proc. USENIX Symp. Usable Privacy and Security*. 2017.

[BS04] Alastair R Beresford and Frank Stajano. "Mix zones: User privacy in location-aware services". In: *Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second IEEE Annual Conference on*. IEEE. 2004, pp. 127–131.

[BS14] D Bezzina and J Sayer. "Safety pilot model deployment: Test conductor team report". In: *Report No. DOT HS* 812 (2014), p. 171.

[BS16] Vincent Bindschaedler and Reza Shokri. "Synthesizing plausible privacy-preserving location traces". In: *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE. 2016, pp. 546–563.

[BTP+kh] Tara Siegel Bernard, Hsu. Tiffany, Nicole Perlroth, and Ron Lieber. *Equifax Says Cyberattack May Have Affected 143 Million in the U.S.* Sept. 2017, URL: https://www.nytimes.com/2017/09/07/business/equifax-cyberattack.html.

[CCS17] Dongyao Chen, Kyong-Tak Cho, and Kang G Shin. "Mobile IMUs Reveal Driver's Identity From Vehicle Turns". In: *arXiv preprint arXiv:1710.04578* (2017).

[CDM15] Mauro Conti, Roberto Di Pietro, and Luciana Marconi. "Privacy for LBSs: On Using a Footprint Model to Face the Enemy". In: *Advanced Research in Data Privacy*. Springer, 2015, pp. 169–195.

[CDP+17] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. "Social Fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling". In: *IEEE Transactions on Dependable and Secure Computing* (2017).

[CMC+12] Ningning Cheng, Prasant Mohapatra, Mathieu Cunche, Mohamed Ali Kaafar, Roksana Boreli, and Srikanth Krishnamurthy. "Inferring user relationship from hidden information in wlans". In: *MILITARY COMMUNICATIONS CONFERENCE, 2012-MILCOM 2012*. IEEE. 2012, pp. 1–6.

[Comom] Progressive Casualty Insurance Company. *Progressive*. Mar. 2018. http://www.progressive.com.

[DBC16] Anupam Das, Nikita Borisov, and Matthew Caesar. "Tracking Mobile Web Users Through Motion Sensors: Attacks and Defenses." In: *NDSS*. 2016.

[DDM+17] Andrea De Salve, Roberto Di Pietro, Paolo Mori, and Laura Ricci. "A Logical Key Hierarchy Based approach to preserve content privacy in Decentralized Online Social Networks". In: *IEEE Transactions on Dependable and Secure Computing* (2017).

[DPS+12] Ivan Damgård, Valerio Pastro, Nigel Smart, and Sarah Zakarias. "Multiparty computation from somewhat homomorphic encryption". In: *Advances in Cryptology–CRYPTO 2012*. Springer, 2012, pp. 643–662.

[DRK16] Sebastian Derikx, Mark de Reuver, and Maarten Kroesen. "Can privacy concerns for insurance of connected cars be compensated?" In: *Electronic Markets* 26.1 (2016), pp. 73–81.

[Dwo11] Cynthia Dwork. "Differential privacy". In: *Encyclopedia of Cryptography and Security*. Springer, 2011, pp. 338–340.

[DWS+15] Bruce DeBruhl, Sean Weerakkody, Bruno Sinopoli, and Patrick Tague. "Is your commute driving you crazy?: a study of misbehavior in vehicular platoons". In: *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM. 2015, p. 22.

[ETK+16] Miro Enev, Alex Takakuwa, Karl Koscher, and Tadayoshi Kohno. "Automobile driver fingerprinting". In: *Proceedings on Privacy Enhancing Technologies* 2016.1 (2016), pp. 34–50.

[GE03] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[GFS+14] Xianyi Gao, Bernhard Firner, Shridatt Sugrim, Victor Kaiser-Pendergrast, Yulong Yang, and Janne Lindqvist. "Elastic pathing: Your speed is enough to track you". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2014, pp. 975–986.

[HOO+14] Peter Händel, Jens Ohlsson, Martin Ohlsson, Isaac Skog, and Elin Nygren. "Smartphone-based measurement systems for road vehicle traffic monitoring and usage-based insurance". In: *IEEE Systems Journal* 8.4 (2014), pp. 1238–1248.

[HPF+15] Siniša Husnjak, Dragan Peraković, Ivan Forenbacher, and Marijan Mumdziev. "Telematics system in usage based motor insurance". In: *Procedia Engineering* 100 (2015), pp. 816–825.

[Jer14] J Jerome. "The Connected Car And Privacy Navigating New Data Issues". In: *The Future of Privacy Forum*. 2014.

[KFS+15] Thivya Kandappu, Arik Friedman, Vijay Sivaraman, and Roksana Boreli. "Privacy in Crowdsourced Platforms". In: *Privacy in a Digital, Networked World*. Springer, 2015, pp. 57–84.

[KMS+07] Kashif Kifayat, Madjid Merabti, Qi Shi, and David Llewellyn-Jones. "Applying secure data aggregation techniques for a structure and density independent group based key management protocol". In: *Information Assurance and Security, 2007. IAS 2007. Third International Symposium on*. IEEE. 2007, pp. 44–49.

[Kru07] John Krumm. "Inference attacks on location tracks". In: *Pervasive computing* (2007), pp. 127–143.

[KSL+14] Ghassan O Karame, Claudio Soriente, Krzysztof Lichota, and Srdjan Capkun. “Securing Cloud Data in the New Attacker Model.” In: *IACR Cryptology ePrint Archive* 2014 (2014), p. 556.

[KSL+17] Ghassan O Karame, Claudio Soriente, Krzysztof Lichota, and Srdjan Capkun. “Securing Cloud Data under Key Exposure”. In: *IEEE Transactions on Cloud Computing* (2017).

[Lan15] Susan Landau. “Control use of data to protect privacy”. In: *Science* 347.6221 (2015), pp. 504–506.

[LCZ+17a] Huaxin Li, Qingrong Chen, Haojin Zhu, and Di Ma. “Hybrid de-anonymization across real-world heterogeneous social networks”. In: *Proceedings of the ACM Turing 50th Celebration Conference-China*. ACM. 2017, p. 33.

[LCZ+17b] Huaxin Li, Qingrong Chen, Haojin Zhu, Di Ma, Hong Wen, and Xuemin Sherman Shen. “Privacy Leakage via De-anonymization and Aggregation in Heterogeneous Social Networks”. In: *IEEE Transactions on Dependable and Secure Computing* (2017).

[LLC15] Gang Li, Boon-Leng Lee, and Wan-Young Chung. “Smartwatch-based wearable EEG system for driver drowsiness detection”. In: *IEEE Sensors Journal* 15.12 (2015), pp. 7169–7180.

[LMM+17] H. Li, D. Ma, B. Medjahed, Q. Wang, Y.S. Kim, and P. Mitra. “Secure and privacy-preserving data collection mechanisms for connected vehicles”. In: *WCX17: SAE World Congress Experience*. SAE. 2017.

[LXZ+16] Huaxin Li, Zheyu Xu, Haojin Zhu, Di Ma, Shuai Li, and Kai Xing. “Demographics inference through Wi-Fi network traffic analysis”. In: *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*. IEEE. 2016, pp. 1–9.

[LZD+16] Huaxin Li, Haojin Zhu, Suguo Du, Xiaohui Liang, and Xuemin Shen. “Privacy leakage of location sharing in mobile social networks: Attacks and defense”. In: *IEEE Transactions on Dependable and Secure Computing* (2016).

[LZM17] Huaxin Li, Haojin Zhu, and Di Ma. “Demographic Information Inference through Meta-data Analysis of Wi-Fi Traffic”. In: *IEEE Transactions on Mobile Computing* (2017).

[MCH00] Robert John McMillan, Alexander Dean Craig, and John Patrick Heinen. *Motor vehicle monitoring system for determining a cost of insurance*. US Patent 6,064,970. May 2000.

[MJ13] Amirreza Masoumzadeh and James Joshi. “Top Location Anonymization for Geosocial Network Datasets.” In: *Trans. Data Privacy* 6.1 (2013), pp. 107–126.

[MSV+15] Yan Michalevsky, Aaron Schulman, Gunaa Arumugam Veerapandian, Dan Boneh, and Gabi Nakibly. “PowerSpy: Location Tracking Using Mobile Device Power Analysis.” In: *USENIX Security Symposium*. 2015, pp. 785–800.

[NVB+16] Sashank Narain, Triet D Vo-Huu, Kenneth Block, and Guevara Noubir. “Inferring user routes and locations using zero-permission mobile sensors”. In: *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE. 2016, pp. 397–413.

[PGS17] Mert D Pesé, Arun Ganesan, and Kang G Shin. “CarLab: Framework for Vehicular Data Collection and Processing”. In: *Proceedings of the 2nd ACM International Workshop on Smart, Autonomous, and Connected Vehicular Systems and Services*. ACM. 2017, pp. 43–48.

[PS14] Sai Teja Peddinti and Nitesh Saxena. “Web search query privacy: Evaluating query obfuscation and anonymizing networks”. In: *Journal of Computer Security* 22.1 (2014), pp. 155–199.

[RKS+16] Hubert Ritzdorf, Ghassan Karame, Claudio Soriente, and Srdjan Čapkun. “On Information Leakage in Deduplicated Storage Systems”. In: *Proceedings of the 2016 ACM on Cloud Computing Security Workshop*. ACM. 2016, pp. 61–72.

[SMS+17] Matthew Smith, Daniel Moser, Martin Strohmeier, Vincent Lenders, and Ivan Martinovic. “Analyzing privacy breaches in the aircraft communications addressing and reporting system (acars)”. In: *arXiv preprint arXiv:1705.07065* (2017).

[SSG+15] Nitesh Saxena, John J Sloan, Manasvee Godbole, Jun Yu Jacinta Cai, Michael Georgescu, Oliver Nick Harper, and David C Schwebel. “Consumer Perceptions of Mobile and Traditional Point-of-Sale Credit/Debit Card Systems in the United States: A Survey”. In: *International Journal of Cyber Criminology* 9.2 (2015), p. 162.

[SSM12] Arun Sahayadhas, Kenneth Sundaraj, and Murugappan Murugappan. “Detecting driver drowsiness based on sensors: a review”. In: *Sensors* 12.12 (2012), pp. 16937–16953.

[STT+12] 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 2012 ACM conference on Computer and communications security*. ACM. 2012, pp. 617–627.

[STT17] Reza Shokri, George Theodorakopoulos, and Carmela Troncoso. “Privacy games along location traces: A game-theoretic framework for optimizing location privacy”. In: *ACM Transactions on Privacy and Security (TOPS)* 19.4 (2017), p. 11.

[TDK+11] Carmela Troncoso, George Danezis, Eleni Kosta, Josep Balasch, and Bart Preneel. “Pripayd: Privacy-friendly pay-as-you-drive insurance”. In: *IEEE Transactions on Dependable and Secure Computing* 8.5 (2011), pp. 742–755.

[TSC+16] Vincent F Taylor, Riccardo Spolaor, Mauro Conti, and Ivan Martinovic. “Appscanner: Automatic fingerprinting of smartphone apps from encrypted network traffic”. In: *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. IEEE. 2016, pp. 439–454.

[VJ13] Christoffel Venter and Johan Joubert. “Use of Multisource Global Positioning System Data to Characterize Multiday Driving Patterns and Fuel Usage in a Large Urban Region”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2338 (2013), pp. 1–10.

[VVN16] Tien Dang Vo-Huu, Triet Dang Vo-Huu, and Guevara Noubir. “Fingerprinting Wi-Fi devices using software defined radios”. In: *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM. 2016, pp. 3–14.

[YGZ+17] Qing Yang, Paolo Gasti, Gang Zhou, Aydin Farajidavar, and Kiran S Balagani. “On inferring browsing activity on smartphones via USB power analysis side-channel”. In: *IEEE Transactions on Information Forensics and Security* 12.5 (2017), pp. 1056–1066.

[ZB11] Hui Zang and Jean Bolot. “Anonymization of location data does not work: A large-scale measurement study”. In: *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM. 2011, pp. 145–156.

[ZCL+17] Lu Zhou, Qingrong Chen, Zutian Luo, Haojin Zhu, and Cailian Chen. “Speed-Based Location Tracking in Usage-Based Automotive Insurance”. In: *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE. 2017, pp. 2252–2257.