THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Data Privacy for Big Automotive Data

BOEL NELSON

Divison of Networks and Systems
Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2017

Data Privacy for Big Automotive Data
Boel Nelson

Technical report 171L
ISSN 1652-876X
Department of Computer Science and Engineering
System Security Research Group

Divison of Networks and Systems
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
Phone: +46 (0)31 772 10 00

Author e-mail: boeln@chalmers.se

Printed by Chalmers Reproservice
Göteborg, Sweden 2017

Data Privacy for Big Automotive Data

Boel Nelson
Divison of Networks and Systems, Chalmers University of Technology

ABSTRACT

In an age where data is becoming increasingly more valuable as it allows for data analysis and machine learning, big data has become a hot topic. With big data processing, analyses can be carried out on huge amounts of user data. Although big data analysis has increased the ability to learn more about a population, it also carries a risk to individual users' privacy, as big data can contain or reveal unintended personal information. With the growing capacity to store and process such big data, the need to provide meaningful privacy guarantees to users thus becomes a pressing issue.

We believe that techniques for privacy-preserving data analysis enables big data analysis, by minimizing the privacy risk for individuals. In this work we have further explored how big data analysis can be enabled through privacy-preserving techniques, and what challenges arise when implementing such analyses in a real setting.

Our main focus is on differential privacy, a privacy model which protects individuals' privacy, while still allowing analysts to learn statistical information about a population. In order to have access to real world use cases, we have studied privacy-preserving big data analysis in the context of the automotive domain.

Keywords: big data, data privacy, differential privacy, privacy, vehicular data

Preface

This thesis is for the degree of Licentiate of Engineering, and includes the previously published papers listed next.

- **Boel Nelson**, Tomas Olovsson, "Security and Privacy for Big Data: A Systematic Literature Review", in *3rd International Workshop on Privacy and Security of Big Data (PSBD 2016)* in conjunction with *2016 IEEE International Conference on Big Data (Big Data)*, Washington DC, USA, December 7, 2016, pp. 3693-3702.
- **Boel Nelson**, Tomas Olovsson, "Introducing Differential Privacy to the Automotive Domain: Opportunities and Challenges", in *Proceedings of the 2nd International Workshop on Vehicular Security (V-SEC 2017)*, Toronto, Canada, September 24, 2017
- Mathias Johanson, Jonas Jalminger, Emmanuel Frécon, **Boel Nelson**, Tomas Olovsson, Mats Gjertz, "Joint Subjective and Objective Data Capture and Analytics for Automotive Applications", in *Proceedings of the 2nd IEEE International Workshop on Vehicular Information Services for the Internet of Things (VISIT 17)*, Toronto, Canada, September 24, 2017

The thesis also includes the following paper that is to be submitted as an extended version.

- **Boel Nelson**, Tomas Olovsson, “LDPModE: Modular Software for Differentially Private Data Collection”

Acknowledgments

In many ways, the acknowledgements section was the hardest thing to write in this thesis. First of all, I would like to thank my supervisor, Associate Professor Tomas Olovsson, for the advice and support he has provided during my PhD. Thank you for always seeing things from the bright side! I would also like to thank my co-supervisor, Professor David Sands, for managing to find a balance between challenging and coaching me.

Thank you to my friends, co-workers and collaborators at Chalmers, Alkit Communications, RISE SICS and Volvo Car Corporation. Out of fear of omitting someone important, I will not name anyone here. Instead, know that without the interaction with all of you, I would not have come this far today. Thank you!

Lastly, I would like to thank my friends and family, for bearing with me through this roller-coaster ride. To Mika and Karl for always being there to share my troubles and a pot of tea. To Hedvig, for always being interested in my research and never turning down a fika invite. And finally, to Andrej.

Boel Nelson

Göteborg, November 2017

Contents

List of Figures

List of Tables

Introduction

Data is becoming increasingly valuable as it can be used for many different kinds of data analyses. Already in 1985, Porter and Millar pointed out the competitive advantage of information [1]. For example, data can be used to improve sales by suggesting recommended products when shopping online [2] as well as provide personalized discounts [3], and it can also be used to improve traffic flow by suggesting the fastest route [4].

Along with increased data collection, attention is also being brought to what data is collected. For example, recent newspaper articles tell stories of data collection ranging from robotic vacuums [5], online restaurant reservations [6] to dating applications [7]. As data is getting closer to becoming hard currency, along with awareness of data collection increasing, this may also cause individuals to be more hesitant about giving away their data.

Then, is there a future for data analytics if access to data will be limited? Essentially there are two routes: either companies will have to stop collecting data, or, and more likely, companies will have to offer a certain degree of privacy to participants who contribute with their data. In fact, privacy enables data collection by working as an incentive for individuals to participate in data collection.

During this particular project [8], we have focused on the analysis of big data from vehicles. Especially, we have focused on how privacy-preserving tech-

nologies can be deployed in the automotive domain. Apart from privacy-preserving technologies enabling big data analysis by facilitating data collection, they can also eliminate answer bias [9] when data is gathered through surveys. Thus, it is our conviction that privacy should not be seen as an obstacle, but rather as an enabler when it comes to big data analysis.

1.1 Data Privacy

Perfect privacy can trivially be achieved by not releasing any data at all. However, for data to be useful in an analysis, the result from an analysis needs to achieve an adequate degree of accuracy. On the other hand, releasing too accurate data can infringe on the privacy of the user that produced the data. In other words, the challenge with privacy-preserving data analysis boils down to balancing a trade-off between privacy and accuracy.

1.1.1 On Data Privacy

Privacy does not have a single definition, which can cause a lot of confusion. In some contexts, privacy focuses on surveillance, but this is not the case in this thesis. Rather, privacy in this thesis refers to what can be learned from data that is released on purpose. Thus, our goal by providing data privacy is to guarantee that no other data than what was intended is leaked, as opposed to that no data is released as is typically achieved through cryptography.

Privacy can be thought of as having two extreme points, as we illustrate in Figure 1.1. One extreme is perfect privacy, where no data whatsoever is released. The other extreme is no privacy, which essentially means perfect accuracy, where all data is released. What we are trying to achieve in this thesis is to find a point in between the two.

Figure 1.1: The privacy spectrum, showing the two extremes of privacy. When performing privacy-preserving data analysis, the goal is to find the sweet spot in between the two extremes, not to achieve any of the extremes.

1.1.2 Privacy Expectations

One important part of providing privacy is identifying to whom privacy is granted. Do we protect all information about an individual, or only certain attributes? For example, do we hide everything about a certain individual, or maybe only what cereal they eat for breakfast? The range of what behaviour or what attributes we *can* protect is wide, but until we know which ones to hide we cannot implement and guarantee privacy. Thus, what information should be kept private and what information is considered public must first be known.

So, how do we identify what data to keep private? Usually, some data is considered sensitive, but how can we distinguish between public and sensitive data? As McSherry writes [10], there is a difference between *personal* data and *private* data. Individuals tend to want to protect their personal data, but this data is not necessarily also private. As an example, McSherry mentions that weight may be personal data that some individuals consider sensitive, but by just observing an individual one can make an educated guess about their approximate weight. Therefore, some personal data is not possible to keep private, as studies can be conducted with or without one individual's participation. Consequently, the main thing of importance before beginning data collection is to provide participants with information about what privacy they can expect, and also what they cannot expect.

For example, we may want to hide that a certain individual drives a diesel-fueled car in order to protect their privacy, but we may also want to release

statistics stating that diesel-fueled cars are harmful to the environment. Thus, if one already knows that their friend drives a diesel-fueled car, one can infer that he or she causes harm to the environment, but that information was not learned by the friend's participation in the survey. In conclusion, the sensitive part of information is not released because the friend chose to participate in the survey, since we would have learned about this correlation even without his or her participation. Therefore, in order to preserve privacy when we release statistical information about a population, the risk that we learn private data about an individual should not be governed by that person's participation or non-participation.

1.1.3 Myths of Anonymization

On a regular basis, individuals get asked if they would like to participate in different kinds of data analyses by providing their data. For example, one might be presented with a survey about how easy a website is to navigate when it is visited, or one might receive a request to rate a product one just purchased. In order to give an incentive to individuals to participate, data collectors often claim that the collected data will be "anonymized", implying that contributing data does not impose a privacy risk for the participant. However, as have been shown in several notable cases [11, 12, 13, 14, 15], side-information or auxiliary data is often available, and can be used to re-identified supposedly anonymized data.

In the well-known case with the governor of Massachusetts [11], medical records were anonymized by removing names as well as social security numbers from patients. Unfortunately, it turned out that a combination of ZIP code, gender as well as date of birth, which were all still available in the anonymized records, was enough to uniquely identify individuals when combined with a publicly available voter registration list. This mapping of the two sets is illustrated in Figure 1.2. In a similar manner, users from Netflix, an online video and tv streaming service, could be re-identified even though their data had supposedly

been anonymized, by cross-referencing the anonymized data with public data from the Internet movie database (IMDb) [13].

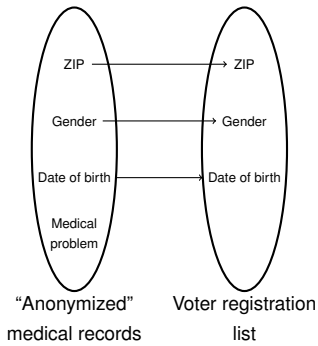

Figure 1.2: When background data is available, it is sometimes possible to combine two data sets in order to re-identify individuals. In this case it was possible to deduce the medical problem of individuals due to this mistake.

So why is it that anonymization often fails? Let us think of anonymization as a state, from which we cannot identify any one individual, no matter the amount of background data we have access to. Two problems immediately arise: first, we have no way of formally verifying that an anonymized state has been reached, and secondly, there are no standardized ways of achieving anonymization. In other words, there are no techniques for anonymization, but rather, anonymization is usually just arbitrary removal of attributes, as with the case of the governor of Massachusetts. What most surveys offer is therefore not true anonymization, but rather *conditional anonymization*, meaning that data is only anonymized while certain conditions (for example, there is no auxiliary data available) holds.

In other words, to be able to meet the privacy expectations of participants, we need robust privacy guarantees from the privacy model used for privacy-preservation, rather than arbitrary de-identification/anonymization techniques.

A privacy model that satisfies such rigorous requirements is *differential privacy*, which we introduce in the upcoming section.

1.2 Privacy Models

In this thesis we present two fundamentally different types of privacy models, those designed for privacy-preserving data publishing (PPDP) and differential privacy. The PPDP models focus on releasing whole sets of data, whereas differential privacy typically is used when answers to database queries are released. While our work focuses on differential privacy, the advantages of differential privacy become more apparent when it is compared to other privacy models, which is why we also introduce PPDP.

1.2.1 Privacy-Preserving Data Publishing (PPDP)

One branch of privacy models is the one that focuses on PPDP. Since PPDP focuses on publishing entire sets of data, models that are PPDP focus on how to modify a data set before it is released. This can be done by adding synthetic entries to the set, or by removing or changing existing entries. In this section we will introduce two of the existing PPDP privacy models, namely k-anonymity and l-diversity.

k-anonymity

k-anonymity [16] is the base privacy model that most PPDP models extend and build upon. In k-anonymity, the implementer picks a set of *quasi*-identifiers, attributes that together make up a unique identifier, and divides the data set into groups with k entries that share the same quasi-identifier in each group. Thus, each entry will be identified by its quasi-identifier, and all other attributes remain intact.

To further explain *k*-anonymity, we illustrate it with an example from the medical domain where privacy is of utmost importance. Imagine that a doctor has a set of raw patient data, as represented in Table 1.1.

Age	Gender	Medical Condition
50	Female	Cancer
55	Female	Cancer
60	Female	Bubonic plague
45	Male	Anthrax
60	Male	Anthrax
70	Male	Cancer

Table 1.1: Medical table with raw data

In order to apply *k*-anonymity, the value for k and the quasi-identifier must first be chosen. The doctor chooses $k = 3$ and the quasi-identifier *{age, gender}*, and creates Table 1.2. Notice how the age and gender now must have the same value in each group, and thus the age is generalized by a range. The table now satisfies 3-anonymity, and has two groups.

Age	Gender	Medical Condition
50-60	Female	Cancer
50-60	Female	Cancer
50-60	Female	Bubonic plague
45-70	Male	Anthrax
45-70	Male	Anthrax
45-70	Male	Cancer

Table 1.2: Medical table that satisfies 3-anonymity, containing two groups. The quasi-identifier is age and gender.

k-anonymity does not specify what attributes to pick for the quasi-identifier, or how the attribute values can be changed to fit into a group. This means

that the implementer must figure out on their own what attributes are sensitive, and how to best protect them. Therefore, *k*-anonymity does not give any formal privacy guarantees, as the implementer chooses what attributes will be protected.

l-diversity

l-diversity [17] was invented as an extension of *k*-anonymity. The main reason for this was that a weakness was discovered in *k*-anonymity, referred to as the similarity attack. We illustrate the weakness in Table 1.3. This table satisfies 3-anonymity, but since one of the groups share the same value, cancer, for their sensitive attribute, medical condition, the sensitive value is unintentionally leaked.

Age	Gender	Medical Condition
50-60	Female	Cancer
50-60	Female	Cancer
50-60	Female	Cancer
45-70	Male	Anthrax
45-70	Male	Bubonic plague
45-70	Male	Cancer

Table 1.3: Medical table that satisfies 3-anonymity. However, the first group share the same value for its sensitive attribute.

To avoid information being leaked through similarity attacks, the focus of *l*-diversity is therefore to have diversity among the value for the sensitive attributes. In comparison, table 1.4 not only satisfies 3-anonymity, but it also satisfies 3-diversity, as the sensitive attribute takes on three different values; anthrax, bubonic plague and cancer. In this case each value of the sensitive attribute is well-represented, meaning that each value occurs roughly the same amount of times within each group.

Age	Gender	Medical Condition
50-60	Female	Cancer
50-60	Female	Anthrax
50-60	Female	Bubonic plague
45-70	Male	Anthrax
45-70	Male	Bubonic plague
45-70	Male	Cancer

Table 1.4: Medical table that satisfies 3-diversity, and contains two groups. The quasi-identifier is age and gender. Notice that the values for medical condition are well-represented.

As *l*-diversity is an extension of *k*-anonymity, it also lacks formal privacy guarantees.

1.2.2 Differential Privacy

In this work, we have especially focused on *differential privacy* [18], which is the only privacy model that gives mathematical privacy guarantees. In other words, differential privacy is the only privacy model where privacy can be quantified and thus measured.

One of the main advantages of differential privacy, as opposed to the other privacy models presented in this thesis, is that it is resistant to background data. In fact, not only does differential privacy maintain its privacy guarantees in the presence of existing background data, but it also resists future releases of data. Furthermore, as expressed by Dwork et al. [19], the other main strengths of differential privacy is that queries can be composed and that any post-processing of data is possible without altering the privacy guarantees. That is, any results released by a differentially private algorithm can be combined and modified without violating privacy.

Since differential privacy is a mathematical definition [20], it can be achieved through several different implementations. In order to fulfill differential privacy, it is common to inject controlled random noise to a query answer, for example by using the Laplace mechanism, as was done in the original paper [18]. In the case where the answer to a query, $f(x)$, is a, a differentially private algorithm using the Laplace mechanism would respond with $f_{DP}(x) = a + noise_{Lap}$.

The Laplace mechanism is a way of achieving differential privacy when data is collected in a centralized manner, for example in a database. This is referred to as centralized differential privacy, and requires a trusted party to manage the database. In contrast, randomized response [9] provides local differential privacy. Randomized response lets each respondent introduce their own noise, which means no trusted party is required. Thus, differential privacy can be achieved either through a centralized or through a distributed algorithm.

A synonym used for differential privacy is ϵ-indistinguishability [21], since the variable ϵ is what determines how much alike elements appear, and thus ϵ controls the privacy offered. Essentially, ϵ quantifies the privacy loss a participant is subjected to when contributing to a query answer.

In the original definition of differential privacy, all participants share a global privacy budget. Every time a differentially private analysis is performed, even if the entire data set does not contribute to the answer, ϵ is deducted from the global privacy budget. However, there also exists a more fine-grained flavor of differential privacy called personalized differential privacy [22, 23], where each participant keeps track of their own privacy budget. When a participant contributes to multiple analyses, privacy degrades additively, meaning that the ϵ from each analysis is deducted from the participant's privacy budget.

To find a good balance between privacy and accuracy, ϵ needs to be set wisely, as it determines the trade-off between the two. If we recall the two extremes of privacy from Figure 1.1, the extreme where there is no privacy is depicted by an ϵ moving towards infinity, and the extreme with perfect privacy is represented by

an ϵ set to zero. Even though ϵ provides a quantifiable measure of privacy it is a relative measure, and must therefore be set on a case-by-case basis [24].

1.3 Big Data

Big data does not have one unanimous definition, but rather, several different ones. One of the most common definitions of big data is in terms of its characteristics, originally referred to as the three Vs: *volume*, *velocity* and *variety* [25]. Today, the amount of Vs have expanded to four or even five Vs, including *value* and *veracity* [26, 27, 28]. *Volume* refers to the size of the data set, while *velocity* addresses the speed the data set grows in. Moreover, *variety* is the different types of data collected, as well as the different formats this data is represented on. *Value* denotes the utility of data, indicating that an analysis of the data can create knowledge that is worth more than the data points would be individually. Lastly, *veracity* is the trustworthiness of data, concerning its accuracy and validity.

Sometimes, big data is defined as a data set that is in the range of exabytes (10^{18} bytes) large [27]. However, as the capacity of storage media is getting increasingly higher, defining big data as a specific volume is problematic since it requires the definition to be continuously updated to reflect what volume is considered big at a specific point in time. Therefore, another alternative definition of big data coined by Kaisler et al. that captures the issue of mentioning specific sizes, is "the amount of data just beyond technology's capability to store, manage and process efficiently" [27].

1.3.1 Vehicular Data

Big data is a relevant topic in the context of connected cars, as each vehicle produces data in real-time. As frequently expressed by Schneier [29], a modern car is no longer just a mechanical device, but rather, a distributed computer

network on four wheels. In fact, a modern car consists of up to 100 electronic control units (ECUs) which communicate over different in-vehicle networks using approximately 4600 different message types, which results in about 7700 unique signals sent between different ECUs [30].

The data sent over the networks can then be collected and uploaded to the cloud for further analysis, using an encrypted mobile channel, for example a 4G connection. In a previous research project [31, 32], a method for collecting data from such modern, connected, cars has been developed. The data is collected by physically separate hardware from the car, that is connected directly to the vehicular network, which means that big, automotive data can be collected.

Apart from real-time data from the vehicle, data about the vehicle owner and the vehicle itself is also stored centrally by the vehicle manufacturer. Thus, vehicular data is comprised of both dynamic, continuously growing, data as well as static data.

As an added privacy challenge, data gathered from a vehicle cannot be treated as independent variables, as they are produced by the same vehicle. Rather, vehicular data is high-dimensional. Treating vehicular data as independent variables has caused privacy issues in the past, for example when location data could be re-identified from speed data as shown by Gao et. al [14]. In this scenario, which was for privacy-preserving insurances, GPS location data was removed, and only speed data was gathered along with the home address, as the company had to be able to bill the driver. However, it turns out that by looking at a map of the area and comparing the speed limitations, it is possible to infer the location even though GPS data had been removed. In other words, even though some attribute has been explicitly removed, that does not ensure that it cannot be deduced from other data.

To complicate matters further, the fact that signals are collected over time means that the collected data will also be time-series data. In fact, the longer period of time data is gathered, the less sensors are required to uniquely identify different drivers [33]. That is, with big automotive data, identifying an individual driver

becomes easier than with smaller sets of data. Consequently, the sheer volume of data can cause privacy issues.

An additional challenge when it comes to collecting data from vehicles is where the privacy-preserving mechanism should be implemented. The ECUs have limited processing power, and therefore cannot handle resource intensive privacy-preserving techniques. Another option is to apply the privacy step once the data has been uploaded to the cloud, but then the cloud owner has to be trusted, since raw data will pass through the cloud. Thus choosing the particular flavor of differential privacy is highly dependent on the available resources.

1.4 Thesis Objective

The objective of this thesis is to investigate how privacy can enable big data analysis. We mainly focus on differential privacy, where we aim to identify both its strengths and its weaknesses when implemented in a real domain.

Since privacy can be achieved trivially by not releasing any data, we focus on the trade-off between privacy and accuracy, as we believe this is the condition that must be met before differential privacy will become widely adopted outside of the research community.

The research questions we set out to answer are as follows.

- What privacy model(s) are suitable for big data?
- How can differential privacy be achieved in the vehicular domain, and are there any additional challenges that apply in this domain compared to existing theoretical research?
- How can we improve accuracy of differentially private analyses in other ways than tweaking the privacy parameter ϵ?

1.5 Summary and Contribution of Included Papers

The papers included in this thesis began with a systematic literature review (Paper A) to create a snapshot of the current state of the art, which led to a closer investigation (Paper B) of the most popular security and privacy topic in big data right now: differential privacy. As a consequence of our findings in Paper A and Paper B, we continued by investigating real-world use cases, but this time with a focus on local differential privacy, in Paper C and Paper D.

1.5.1 Paper A

To investigate the the intersection between big data research and security and privacy research, we conducted a systematic literature review (Paper A) that created a snapshot of the current research field. We found that privacy is currently a popular topic to combine with big data research, and that differential privacy is particularly often used. Our conclusion is that differential privacy is especially well-suited for big data analysis, as it provides mathematically proven privacy guarantees that prevents overfitting of data that would lead to inference of information about individuals. Consequently, in this project [8] about big automotive data analysis, we have focused on differential privacy.

Our contribution in Paper A is a systematic categorization of recent research papers that span both research areas. We answer the following research questions.

- What recent security or privacy papers exists in the big data context?
- How many papers cover security or privacy for big data?
- Which security, privacy and big data topics are represented in the area?
- When a paper covers more than one category, which categories intertwine?

1.5.2 Paper B

We further connect differential privacy to the automotive domain in Paper B. Our main goal with this paper was to bridge the gap between theory and practice, by establishing the possible role of differential privacy within the context of the automotive domain, while at the same time identifying the challenges involved.

Paper B consists of a comprehensive introduction to differential privacy, and focuses especially on what challenges can arise when implementing differential privacy in a vehicular setting. Furthermore, we give advice to practitioners concerning where to start when implementing differential privacy in this domain. Lastly, we highlight the currently open research problems that apply to the entire differential privacy research community, and also discuss the specific problems encountered when dealing with vehicular data.

Thus, the contribution of this paper is as follows.

- a comprehensible introduction to differential privacy, including what type of differentially private analyses can be performed in the vehicular domain
- recommendations for how to proceed when implementing differentially private analyses in the vehicle domain
- a highlight of the challenges involved with the implementation

1.5.3 Paper C

In Paper C we showcase how both subjective and objective data can be collected from connected cars, by implementing a smartphone app that collects the subjective data. The idea is to capture how drivers experience certain scenarios right when it happens, rather than sending a poll in paper format months later. Consequently, the smartphone app collaborates with the in-vehicle network in order to send polls to driver's when interesting scenarios occur. We also discuss

what privacy implications our specific use case has for users, and propose a privacy architecture that relies on differential privacy to guarantee privacy. Our contribution is to provide answers to the following questions.

- How can we design the subjective data capture app in a way that makes it easy and safe to use in a vehicle, even while driving?
- How can we design a triggering mechanism to decide when a particular question or set of questions should be posed to a particular user? The triggering mechanism must be versatile and flexible to be usable for all relevant use cases.
- How can we cater for follow-up questions that depend on answers to previous questions?
- How can we protect the privacy of users while at the same time providing automotive engineers with as powerful data collection and data analytics tools as possible?

1.5.4 Paper D

In Paper D we implement a software solution, which we call Local Differential Privacy Modular Environment (LDPModE). LDPModE has two main purposes: to increase the usability of differential privacy, and to help investigate the accuracy-privacy trade-off of differential privacy. We develop open-source tools to help aid the process of creating a survey and to filter away noise from the collected answers. More specifically, our solution consists of the following three pieces of domain-independent software.

- a poll generator
- a simulation environment
- a filter for removing noise

LDPModE is modular in the sense that the simulation environment can be

exchanged seamlessly with a program for data collection to fit any target domain.

Furthermore, we conduct a case study that highlights how our software solution should be used, and develop a smartphone app to show how data collection can be implemented. Our case study covers the entire process from creating a poll, to evaluating the utility of the poll with the simulation environment. Thus, our contribution is as follows.

- A modular process, including software, for gathering data under local differential privacy
- A case study where the software is used
- An evaluation where privacy-accuracy trade-offs are explored using a domain specific poll

1.6 Conclusion and Future Work

In this thesis we have started paving the path towards utilizing differential privacy for big automotive data analysis. Differential privacy shows great promise when it comes to big data analysis in general, in fact, the noise introduced by many differentially private algorithms is independent of the data set's size, causing the noise to have less impact on accuracy. Thus, differential privacy is a particularly good choice for big data.

However, implementing differential privacy within the automotive domain has shown not to be trivial. Therefore we try to prepare practitioners by giving them advice and warning them of potential pitfalls in Paper B. Nonetheless, we have not been able to address every aspect of implementation in our work, and therefore interesting future work includes among others investigating how differential privacy works over time for cars. For example, finding out how long it takes before privacy budgets are depleted in a real-world setting.

Furthermore, our work has addressed different flavors of differential privacy. An interesting future research direction is investigating the challenge of composing results while maintaining meaningful privacy guarantees.

Our contribution so far have shown that there are several potential use cases for differential privacy in the automotive domain, and we will continue our work to scope out for which cases differential privacy is feasible, but also to find the limitations of differential privacy within our domain.

Bibliography

[1] M. E. Porter and V. E. Millar. *How information gives you competitive advantage*. Harvard Business Review, Reprint Service Watertown, Massachusetts, USA, 1985.

[2] J. Wang et al. "Utilizing Related Products for Post-purchase Recommendation in e-Commerce". In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys '11. Chicago, Illinois, USA: ACM, 2011, pp. 329–332.

[3] N. Werro et al. "Personalized Discount - A Fuzzy Logic Approach". en. In: *Challenges of Expanding Internet: E-Commerce, E-Business, and E-Government: 5th IFIP Conference e-Commerce, e-Business, and e-Government (I3E'2005)* (2005), pp. 375–387.

[4] T. H. Silva et al. "Traffic Condition Is More Than Colored Lines on a Map: Characterization of Waze Alerts". en. In: *Social Informatics: 5th International Conference, SocInfo 2013, Kyoto, Japan, November 25-27, 2013, Proceedings*. Springer International Publishing, Nov. 2013, pp. 309–318.

[5] M. Astor. "Your Roomba May Be Mapping Your Home, Collecting Data That Could Be Shared". en-US. In: *The New York Times* (July 2017). URL: https://www.nytimes.com/2017/07/25/technology/roomba-irobot-data-privacy.html (visited on 08/15/2017).

[6] P. David M. *The 'Internet of Restaurants' Is Coming for Your Info*. July 2017. URL: https://psmag.com/economics/your-favorite-restaurants-are-surveilling-you (visited on 08/15/2017).

[7] J. Duportail. "I asked Tinder for my data. It sent me 800 pages of my deepest, darkest secrets". en-GB. In: *The Guardian* (Sept. 2017). URL: http://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold (visited on 10/02/2017).

[8] VINNOVA. *BAuD II: Storskalig insamling och analys av data för kunskapsdriven produktutveckling | Vinnova*. Swedish. 2014. URL: https://www.vinnova.se/p/baud-ii-storskalig-insamling-och-analys-av-data-for-kunskapsdriven-produktutveckling/ (visited on 11/05/2017).

[9] S. L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". In: *Journal of the American Statistical Association* 60.309 (Mar. 1965), pp. 63–69.

[10] F. McSherry. *Differential privacy and correlated data*. Jan. 2017. URL: https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-29.md (visited on 06/22/2017).

[11] P. Samarati and L. Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Tech. rep. Technical report, SRI International, 1998.

[12] M. Barbaro and T. Zeller. "A Face Is Exposed for AOL Searcher No. 4417749". In: *The New York Times* (Aug. 2006). URL: http://query.nytimes.com/gst/abstract.html?res=9E0CE3DD1F3FF93AA3575BC0A9609C8B63 (visited on 07/16/2015).

[13] A. Narayanan and V. Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *IEEE Symposium on Security and Privacy, 2008. SP 2008*. May 2008, pp. 111–125.

[14] X. Gao et al. “Elastic Pathing: Your Speed is Enough to Track You”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’14. New York, NY, USA: ACM, 2014, pp. 975–986.

[15] J. Su et al. “De-anonymizing Web Browsing Data with Social Networks”. In: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 1261–1269.

[16] P. Samarati. “Protecting respondents identities in microdata release”. In: *IEEE transactions on Knowledge and Data Engineering* 13.6 (2001), pp. 1010–1027.

[17] A. Machanavajjhala et al. “L-diversity: Privacy beyond k -anonymity”. In: *ACM Transactions on Knowledge Discovery from Data* 1.1 (2007), 3–es.

[18] C. Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”. en. In: *Theory of Cryptography*. Ed. by S. Halevi and T. Rabin. Lecture Notes in Computer Science 3876. Springer Berlin Heidelberg, 2006, pp. 265–284.

[19] C. Dwork et al. “Exposed! A Survey of Attacks on Private Data”. In: *Annual Review of Statistics and Its Application* 4.1 (2017), pp. 61–84.

[20] C. Dwork. “Differential privacy”. In: *Automata, languages and programming*. Springer, 2006, pp. 1–12.

[21] C. Dwork. “Differential Privacy”. en. In: *Encyclopedia of Cryptography and Security*. Ed. by H. C. A. v. Tilborg and S. Jajodia. DOI: 10.1007/978-1-4419-5906-5_752. Boston, MA: Springer US, 2011, pp. 338–340.

[22] H. Ebadi et al. “Differential Privacy: Now it’s Getting Personal”. In: *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. POPL’15. ACM Press, 2015, pp. 69–81.

[23] Z. Jorgensen et al. “Conservative or liberal? Personalized differential privacy”. In: *2015 IEEE 31st International Conference on Data Engineering*. 2015, pp. 1023–1034.

[24] J. Lee and C. Clifton. “How Much Is Enough? Choosing ϵ for Differential Privacy”. en. In: *Information Security*. Ed. by X. Lai et al. Lecture Notes in Computer Science 7001. Springer Berlin Heidelberg, 2011, pp. 325–340.

[25] D. Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rep. META Group, Feb. 2001.

[26] M. Chen et al. “Big Data: A Survey”. In: *Mobile Networks and Applications* 19.2 (2014), pp. 171–209.

[27] S. Kaisler et al. “Big Data: Issues and Challenges Moving Forward”. In: *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. System Sciences (HICSS), 2013 46th Hawaii International Conference on. IEEE, Jan. 7, 2013, pp. 995–1004.

[28] R. Lu et al. “Toward efficient and privacy-preserving computing in big data era”. In: *Network, IEEE* 28.4 (Aug. 2014), pp. 46–50.

[29] Bruce Schneier. *Security and the Internet of Things - Schneier on Security*. Feb. 2017. URL: https://www.schneier.com/blog/archives/2017/02/security_and_th.html (visited on 06/22/2017).

[30] P. Kleberger et al. “Towards designing secure in-vehicle network architectures using community detection algorithms”. In: *2014 IEEE Vehicular Networking Conference (VNC)*. Dec. 2014, pp. 69–76.

[31] Alkit Communications AB. *BAuD - Big Automotive Data*. URL: http://www.alkit.se/baud/ (visited on 06/14/2017).

[32] VINNOVA. *BAuD: Storskalig insamling och analys av data för kunskapsdriven produktutveckling | Vinnova*. Swedish. Jan. 2013. URL: https://www.vinnova.se/p/baud-storskalig-insamling-och-analys-av-data-for-kunskapsdriven-produktutveckling/ (visited on 06/14/2017).

[33] M. Enev et al. “Automobile Driver Fingerprinting”. In: *Proceedings on Privacy Enhancing Technologies* 2016.1 (2015), pp. 34–50.