University of Windsor
Scholarship at UWindsor

Electronic Theses and Dissertations | Theses, Dissertations, and Major Papers

2-15-2024

Cybersecurity and Privacy Issues in Intelligent Transportation Systems

Haesung Ahn
University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Part of the Transportation Commons

Recommended Citation

Ahn, Haesung, "Cybersecurity and Privacy Issues in Intelligent Transportation Systems" (2024). *Electronic Theses and Dissertations*. 9443.
https://scholar.uwindsor.ca/etd/9443

This online database contains the full-text of PhD dissertations and Masters’ theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email (scholarship@uwindsor.ca) or by telephone at 519-253-3000ext. 3208.

Cybersecurity and Privacy Issues in Intelligent Transportation Systems

By

Haesung Ahn

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Civil and Environmental Engineering
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Windsor

Windsor, Ontario, Canada

2024

© 2024 Haesung Ahn

Cybersecurity and Privacy Issues in Intelligent Transportation Systems

by

Haesung Ahn

APPROVED BY:

T. Kwon, External Examiner
University of Alberta

N. Biswas
Department of Environmental Engineering

H. Maoh
Department of Civil Engineering

C. Lee
Department of Civil Engineering

Y. Kim, Advisor
Department of Civil Engineering

January 25, 2024

DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION

I. Co-Authorship

I hereby declare that this thesis incorporates material that is result of joint research, as follows:

Chapter 2 of the thesis utilized the results from a published article in the journal *IEEE Transactions on Intelligent Transportation Systems*, which has the following co-authors: Dr. Juyeong Choi and Dr. Yong Hoon Kim. Dr. Yong Hoon Kim provided constructive supervisory and set the direction of the chapter. Dr. Juyeong Choi contributed feedback on refining ideas and editing the manuscript.

Chapter 3 incorporates unpublished material co-authored with Sungho Lim and Dr. Yong Hoon Kim, supervised by Dr. Yohee Han. In this chapter, Sungho conducted data analysis, while Drs. Kim and Han provided research direction. Chapter 4 also contains unpublished material co-authored with Dr. Umair Durrani, Sungho Lim and Dr. Yong Hoon Kim under the supervision of Dr. Yohee Han. For this chapter, Sungho contributed data and research ideas. Dr. Durrani assisted with data analysis and Drs. Kim and Han were involved in reviewing and refining the manuscript.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

II. Previous Publication

This thesis includes three original papers that have been previously published/submitted to journals for publication, as follows:

Thesis Chapter	Publication title/full citation	Publication status*
Chapter 2	H. Ahn, J. Choi, and Y. H. Kim, "A Mathematical Modeling of Stuxnet-style Autonomous Vehicle Malware," *IEEE Intell. Transp. Syst. Trans.*, 2022, DOI: 10.1109/TITS.2022.3213771	Published
Chapter 3	H. Ahn, S. Lim, Y. H. Kim, and Y. Han, "Quantifying Privacy Risks in Pandemic Contact Tracing Through Smart Card Data	Under review
Chapter 4	H. Ahn, U. Durrani, S. Lim, Y. H. Kim, and Y. Han, "Information Theory-based Quantifying Privacy Risks in Data	Under review

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

III. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

This dissertation addresses cybersecurity and privacy risks in ITS, focusing on the escalating cyberattacks on AVs and data breaches in areas like smart cards. We explore novel malware threats to AVs and hacker strategies impacting physical infrastructure, underscoring the need for enhanced security in the evolving AV sector. This dissertation also emphasizes the importance of researching privacy risks, especially with the rise in privacy breaches and extortion attempts involving sensitive personal information. Understanding the implications of publicly shared data is crucial in this context.

Chapter 1 raises research questions regarding the two main pillars, cybersecurity and privacy issues in ITS. Based on the questions, Chapters 2, 3 and 4 deal with these questions, presenting key insights this research found. Chapter 2 focuses on a case of vehicle hacking, specifically examining the implications of Stuxnet-style malware. Stuxnet attack methodology provides a critical context for understanding potential threats to AV systems. This chapter introduces a mathematical model to analyze how similar malware could spread both temporally and spatially in the context of AVs. Inspired by epidemiology and ecology, this approach conceptualizes malware as an infectious disease to study its propagation dynamics. This is the first attempt to apply such a model to the spread of Stuxnet-style malware in AV environments, paving the way for future research on the temporal and geographic spread of infectious malware in AV networks.

Chapter 3 delves into the privacy risks associated with the public sharing of COVID-19 patients' travel records during the pandemic. This measure, intended

for public health safety, inadvertently risked exposing sensitive personal travel details. The chapter examines how combining these records with other open-data sources might allow for the re-identification of individuals' private information. We quantify these re-identification risks, focusing on the volume and accuracy of the shared records, along with the variety of locations that the patients visited. This analysis is crucial for understanding the privacy implications of such data-sharing practices in public health contexts.

Chapter 4 introduces a method to quantify privacy risks using information theory, measuring information as entropy units. We use synthetic data to model individual travel patterns, combining these into a unit termed a *cube* that encapsulates both time and space elements. The study focuses on how adding these cubes affects privacy risk, particularly by assessing the novel information each cube contributes. This allows us to quantify the distinct information within various data sequences, using joint and conditional entropy to understand uncertainty fluctuations as more cubes are added.

Lastly, Chapter 5 concludes this dissertation’s insights that potential malware attacks can bring about magnificent physical destructions by manipulating infected AVs, privacy risks may emerge with the combination of external observations data, and information theory-based methodologies can quantify the risks in individual pieces of information. These insights can emphasize the necessity of our research.

DEDICATION

To my God for His guidance and love from birth, continuing now and forever.

To my mother, Geumsook Moon and late father, Yunseob Ahn, for their unwavering love.

To my wife Joy Kibbeum Lee, for her endless love and sacrifice, and to my fur babies Pepper and Maple.

To the composer of '*Tis so sweet to walk with Jesus'*, whose hymn provided the strength that enabled me to rise and face another day whenever my dissertation overwhelmed me.

"Step by step, step by step,
I would walk with Jesus,
All the day, all the way,
Keeping step with Jesus,"

ACKNOWLEDGEMENTS

I am deeply grateful to all my committee members: Drs. Tae Jung Kwon, Nihar Biswas, Hanna Maoh, Chris Lee and Yong Hoon Kim. Their insightful comments and suggestions greatly enhanced my dissertation. They were my role model to develop my teaching philosophies and research attitudes. Their example inspires me, and I aspire to emulate their excellence in my new role.

In particular, my advisor, Dr. Yong Hoon Kim, was an excellent and dependable mentor and researcher. I'd say becoming a student under his supervision was my life-changing moment.

Special thanks go to my co-authors, Drs. Yohee Han, Umair Durrani, and Juyeong Choi and Sungho Lim, for their invaluable mentorship and guidance. The collaborative experiences and numerous discussions we shared were like a compass throughout my PhD long journey.

I would like to express my special gratitude to my engineering mentors, Dr. Youngsang Park and Jonggu Heo. Their unwavering support and advice have been a guiding light during my PhD voyage.

My time at the University of Windsor has been unforgettable and incredibly rewarding. I am truly thankful to everyone I met at my home university.

TABLE OF CONTENTS

DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION.....iii
ABSTRACT.....vi
ACKNOWLEDGEMENTS.....ix
LIST OF TABLES.....xiv
LIST OF FIGURES.....xv
LIST OF APPENDICES.....xvi
CHAPTER 1 INTRODUCTION.....1
  1.1 Preface.....1
    1.1.1 Background.....1
    1.1.2 Preliminaries.....3
  1.2 Research Questions and Objectives.....5
    1.2.1 Malware Propagation in Autonomous Vehicles.....6
    1.2.2 Privacy Issues in Mobility Data.....7
    1.2.3 Quantification of Amounts of Information Regarding Privacy Issues.....8
  1.3 Dissertation Outline.....9
  Chapter 1 References.....12
CHAPTER 2.....14
A MATHEMATICAL MODELLING OF STUXNET-STYLE AUTONOMOUS VEHICLE MALWARE.....14
  2.1 Introduction.....14
    2.1.1 Background.....14
    2.1.2 Features of Stuxnet-style Attack.....15
    2.1.3 Reasons for a Growing Interest in Stuxnet-style Attacks in Targeting AVs.....16
    2.1.4 Contributions.....17
  2.2 Related Work.....19

2.2.1 Connected and Autonomous Vehicle Vulnerabilities 19
2.2.2 Risks of Stuxnet-style Malware 20
2.2.3 Malware Propagation Based on Epidemic...21
2.2.4 Lessons from Related Works 22
2.3 The Conceptual Framework for The Malware Propagation Model 24
2.3.1 Conceptual Similarities Between Malaria and Stuxnet 24
2.3.2 Model of Visiting Mechanic Shops 25
2.3.3 Malware Propagation Model Assumptions 26
2.4 Mathematical Modelling of the Spatiotemporal Propagation of Malware ..27
2.4.1 Notation...28
2.4.2 Transportation Operator Model...29
2.4.3 SI Epidemic Model..30
2.4.4 The Framework of Mathematical Equations 35
2.5 Experimental Design..37
2.5.1 Estimation of Transportation Operator Model (δij)..................................38
2.6 Results 40
2.6.1 The Speed of Infectious Malware Spreading...40
2.6.2 The Impact of Variable Probability of Malware Transmission (α).............41
2.6.3 The Impact of Different Connectivity 43
2.7 Conclusion 46
2.8 Chapter 2 References..50
CHAPTER 3 54
QUANTIFYING PRIVACY RISKS IN PANDEMIC CONTACT TRACING THROUGH SMART CARD DATA 54
3.1 Introduction..54
3.2 Related Work 58
3.3 Preliminary...61
3.3.1 Data Description..61
3.3.2 Resolutions in Travel Observations and Re-identification...........................62
3.4 Methods...64
3.4.1 Impact of the Number of Travel Observations 64
3.4.2 Impact of Spatial and Temporal Resolutions 67
3.4.3 Impact of Diverse Visited Locations..67

3.5 Results 68
3.5.1 Privacy Risks Based on the Number of Observations 68
3.5.2 Privacy Risks Based on Spatial and Temporal Resolutions..........................69
3.5.3 Privacy Risks Based on Diverse Visited Locations 74
3.6 Conclusion 77
3.7 Chapter 3 References..79

CHAPTER 4 INFORMATION THEORY-BASED QUANTIFYING PRIVACY RISKS IN DATA 81
4.1 Introduction..81
4.2 Preliminary: Information Theory 85
4.2.1 Introduction to Information Theory 85
4.2.2 Measuring the Uncertainty in Probability Distribution 87
4.2.3 Joint Entropy and Conditional Entropy...90
4.3 Related Work 94
4.3.1 Privacy Risk Measurement Research with External Observations............94
4.3.2 Information Theory-based Research Regarding Uncertainty 95
4.4 Methods of Analysis...97
4.4.1 Concepts of Rareness and Commonness in Datasets....................................98
4.4.2 Amount of Uncertainty and Privacy Risk..99
4.5 Experimental Design and Results...109
4.5.1 Data Description and Experimental Design 109
4.5.2 Amount of Total Uncertainty and Joint Entropy.......................................113
4.5.3 Anonymity Values and Entropy 117
4.5.4 Amount of Information Considering Order of Acquisition 123
4.7 Conclusion 126
4.8 Chapter 4 References..129

CHAPTER 5 CONCLUSIONS..131
5.1 Research findings and contributions..131
5.1.1 Understanding A Novel Type of Malware Attack......................................131
5.1.2 Exploring Privacy Issues Using Real-world Privacy Breach Case 132
5.1.3 Quantifying Privacy Risks Based on Information Theory.........................132
5.2 Research Limitations and Future Plans..133

5.2.1 For Cybersecurity Research with Diverse Methods of Spreading Malware 133
5.2.2 For language-based external observation privacy research.......................134
5.2.3 For Information Theory-based Quantifying Privacy Risks 134
5.2.4 For Inter-chapter Research Opportunities...135
5.3 Chapter 5 References..136
APPENDICES 137
Appendix A. Travel Records for Chapter 4 137
Appendix B. Box Plots of Entropy and Number of Information for Chapter 4....145
VITA AUCTORIS 148

LIST OF TABLES

Table 2-1 Summary of related works..23

Table 2-2. Notation (Conceptual Analogy Between AVs Epidemic and Vector-host Model) 28

Table 2-3. Lists of Maintenance by Mileage 39

Table 2-4. Result of Distribution by The Gravity Model 39

Table 2-5. Comparison of Infection Pace by Alpha 42

Table 2-6. Comparison of Infection Pace by Connectivity..46

Table 3-1 Notation..66

Table 4-1. Notation...90

Table 4-2. Sample data 99

Table 4-3. Identification results by varying information pieces and visit conditions 101

Table 4-4. Distribution of individuals with location information A and B 106

Table 4-5. Identification results by varying information pieces and visit conditions 107

Table 4-6. Resolved uncertainty rate by number of information 108

Table 4-7. Cubes by region and time..112

Table 4-8. Resolved uncertainty rate by number of information 115

Table 4-9. Example of the anonymity values under visit conditions with two pieces of information..118

Table 4-10. Combinations of visiting cubes based on information acquisition order 124

LIST OF FIGURES

Fig. 2-1. Similarity between malaria and Stuxnet AVs epidemic 24

Fig. 2-2. Contacts network between vehicles and mechanic shops 26

Fig. 2-3. Process of vehicle infection at mechanic shop j 32

Fig. 2-4. Process of OBD scanners infection at mechanic shop j 34

Fig. 2-5 Framework of the integrating SI epidemic model and transportation operator model 36

Fig. 2-6. Study area 37

Fig. 2-7. Infection rate of vehicles and scanners with α = 0.7 41

Fig. 2-8. Number of infected vehicles and scanners by alpha α = 0.7 43

Fig. 2-9. Number of infected vehicles and scanners by connectivity (ρ) 45

Fig. 3-1. Number of infected vehicles and scanners by connectivity (ρ) 64

Fig. 3-2 Re-identification risks by the number of observations 69

Fig. 3-3 Re-identification risks by spatial resolutions 71

Fig. 3-4 Re-identification risks by temporal resolution 72

Fig. 3-5. Missing rate by temporal resolution..74

Fig. 3-6. Privacy risks by combinations of visited locations 76

Fig. 4-1 Comparative illustration of two methods for the coin-finding game 86

Fig. 4-2 Classification of IDs over incremental location information.....................105

Fig. 4-3 Conceptual diagram of the study area 110

Fig. 4-4 Composition of recorded travel patterns by time slot and region 111

Fig. 4-5 Entropy fluctuation over the number of observations..............................117

Fig. 4-6 Composition percentage of the anonymity values category 121

Fig. 4-7 Box plots across the number of information pieces and anonymity values 122

Fig. 4-8 Joint and conditional entropy by information acquisition pattern 126

LIST OF APPENDICES

Appendix A. Travel Records for Chapter 4 137

Appendix B. Box Plots of Entropy and Number of Information for Chapter 4....145

CHAPTER 1

INTRODUCTION

1.1 Preface

1.1.1 Background

Since the first introduction of the concept of Intelligent Transportation Systems (ITS) emerged in the 20th century, the scope of ITS has significantly broadened. ITS is comprised of four key components: Smart Vehicles, Public Transportation, Internet of Things (IoT) Devices, and Controllers [1]. ITS leverages cutting-edge advancements in Information and Communication Technologies (ICT) to enhance efficiency, sustainability, safety and security in the transportation sector [2]. The advent of the IoT has expedited many ITS technologies, such as communicating with vehicles and infrastructures (V2X), analyzing big data, counting real-time traffic volumes and pedestrians, etc. The data collected from IoT contain numerous individuals' travel patterns, confidential vehicle information, and real-time road conditions.

While ITS has significantly enhanced vehicle safety and convenience, the increased computerization and connectivity pose serious cybersecurity concerns. ITS plays a crucial role in enhancing road safety by enabling the rapid sharing of traffic information. This is achieved by deploying advanced systems such as vehicular communication, navigation, and driver-assistance systems (ADAS). In addition to safety, ITS also contributes to convenience and entertainment in vehicles with features like Bluetooth connectivity [1]. This advancement has led to modern vehicles becoming more *computerized* and internally *connected*, primarily through Engine Control Units (ECUs) connected via the controller area network (CAN bus) communication protocol [3].

Access to a vehicle's internal network, including its CAN bus system, is often possible through the onboard diagnostic (OBD) port. This port supports various communication protocols, including the CAN protocol ISO 15765 [4], making the OBD port a crucial interface for vehicle diagnostics and system checks. However, computerized vehicles with outdated security systems can make vehicles susceptible to malware attacks. Numerous studies have highlighted the risks of malware attacks on CAVs and autonomous vehicles (AVs) through OBD ports [3], [5], [6]. The infected vehicles by malware can be exploited to destroy physical infrastructures and damage individuals' lives. Consequently, addressing this vulnerability to cyber-attacks has become a rapidly growing concern and is now considered a national priority.

On the other hand, cybersecurity and privacy risks in ITS are closely intertwined. Recently, we have seen cyberattacks targeting critical infrastructures, such as hospitals in southwest Ontario, Canada [7]. Adversaries can also target AVs and databases containing sensitive personal data. These massive attacks can lead to destruction and data breaches, with potential demands for ransom or selling of leaked data [8]. Such incidents highlight not only the vulnerabilities in our systems but also raise broader concerns about how personal information is collected, stored, and potentially exploited.

In today's era of abundant information, there is often a lack of awareness about how personal information is being collected and utilized. There is a saying that *Google/Apple knows everything about me* to reflect how our everyday devices, such as smartphones, continuously gather data about our activities. When people use their phones for activities like searching online or making purchases, information providers collect data on individual preferences and behaviours, often utilizing it for business purposes.

Beyond smartphones, a wide array of other information sources, such as smart cards and personal vehicles, including connected and autonomous vehicles (CAVs), contain substantial and sensitive personal data, like location-based service (LBS) data. Consequently, personal information within ITS is attracting considerable interest because these devices and technologies have the capacity to store sensitive information, creating potential risks of privacy breaches.

Considering the evolving challenges in cybersecurity and privacy, it is crucial to thoroughly examine the research that lays the groundwork for the discussions in this dissertation. Privacy issues in transportation data and cybersecurity attacks in AVs are not a topic for the distant future but close threats in ITS. This dissertation explores two critical issues in ITS: the malware propagation model among AVs and privacy risks in transportation databases.

1.1.2 Preliminaries

This section aims to lay the groundwork for understanding the dual focus of this study: cybersecurity risks, especially concerning malware attacks on AVs, and privacy concerns in transportation data systems.

1) Cybersecurity research

The threat of malware attacks on cybersecurity systems poses significant risks to user protection. Hackers can exploit personal information through malware, which nowadays not only aims to steal data but also to cause physical damage to systems and devices surrounding us. A notable case is the Stuxnet malware attack on Iran’s nuclear facilities in 2010, which caused physical destruction while concealing its actual activities

[9]. Reportedly, the malware falsified the monitoring systems in the nuclear facilities by showing the normal status messages of the uranium centrifuges on the monitoring system. However, the malware changed the electrical current that powers the centrifuge and the speed of the machine until it exploded, disguising its actual situation by showing regular messages on the monitor. This type of cyberattack represents a potential threat to the transportation engineering sector, especially AVs. These vehicles are particularly vulnerable because they are part of cyber-physical systems (CPS), which combine computer and physical functions. Therefore, research on malware targeting CPS components like AVs is vital to safeguard personal privacy and prevent physical damage.

2) Privacy research

Early research primarily addressed privacy threats in Location-Based Services (LBS). Privacy in LBS can be categorized into three distinct dimensions: identity, location, and query privacy [10]. Identity privacy concerns breaches of personal information. The service provider, such as the credit card operator, shall not know the identity of users who employ the card service. Location privacy involves the user’s spatial and temporal information issues. Query privacy relates to inferences about individuals based on data combination.

On the other hand, privacy protection methods fall into two major categories. One is to make data indistinguishable using generalization, substituting information attributes with a broader one, and suppression, removing sensitive data. Another method is to make data uninformative with differential privacy [11]. However, as we protect personal information, the value of data for research and business decreases. Hence, the researchers have tried to balance the use of personal information data and privacy protection.

Privacy research has also dealt with smart card data. Mayes' team defined several features of the smart card: it has a unique identifier, participates in an automated electronic transaction, can store data securely, and hosts/runs various security algorithms and functions [12]. Smart card data also contains large-scale individual LBS data, having a unique card ID and information on boarding/alighting stations. These days, many officials have tried to open the data for research purposes publicly. Even though techniques to protect privacy have been developed, the possibility of a breach of personal information from card data is still a high potential risk. If transit users have used the transit regularly, their routine travel itineraries will be stored in the smart card database. Adversaries could re-identify the users' identities from card data with additional information, such as the personal stories posted on social media in accordance with transit usage. When adversaries identify specific transit users, they can plan to commit a crime against the users. The aftermath of a breach of privacy in the real world would be detrimental to the user's life. Thus, further research on smart card data privacy issues should be undertaken.

1.2 Research Questions and Objectives

Cybersecurity and privacy have been considered indispensable in networked and computerized environments in the fields of ITS for many years. While existing researchers have studied the topics, there is little systematic study of the epidemic of autonomous vehicles and privacy breaches using smart card data. Thoroughly understanding the extent of the threat from malware and privacy breaches is crucial. This dissertation studies the severity of cybersecurity and privacy issues. A mathematical model was formulated to predict the spread of malware among AVs, and a simulation was conducted to identify an individual from the datasets. To this end, many research questions and objectives are outlined in the following subsections.

1.2.1 Malware Propagation in Autonomous Vehicles

We first need to understand the types of malware and propagation speed to cope with malware invasion. This dissertation posed a question regarding AVs' vulnerability moments, propagation method and speed. More specifically, the research questions can be specified as follows:

1. *What is the most vulnerable moment for AVs in terms of cybersecurity threats?*
2. *What types of malware can infiltrate AVs during these vulnerable moments?*
3. *Once inside an AV, how does the malware spread to other AVs within a city?*
4. *How quickly can this malware propagation occur among AVs?*

To address these enquiries, this research concentrated on a specific type of environment: CPS, which is based on integrated computational and physical capabilities to interact with humans [13]. While CPS offers numerous advantages, its blend of cyber systems and physical facilities also opens up vulnerabilities that hackers can exploit for malicious purposes. With the context of CPS, this dissertation analyzed the malware propagation among AVs based on an epidemic model to measure the threats to AVs. This approach is employed to assess and quantify the cybersecurity threats facing AVs.

1.2.2 Privacy Issues in Mobility Data

In recent years, open-data initiatives have enabled researchers to access and extensively utilize individual mobility data. However, this increased accessibility also poses significant privacy risks, as adversaries could potentially access this data and compromise individual privacy by re-identifying people from it. This dissertation specifically focuses on measuring the re-identification risks associated with matching quasi-identifiers (QIDs) to externally observed travel records, hereafter referred to as *external observation* data. The research investigates privacy concerns related to smart card data, particularly examining scenarios where individuals' travel logs, functioning as external observation records, are exposed or leaked. This research seeks to answer the following questions:

1. *How does the number of travel observations impact the risks?*
2. *How do variations in the resolution of travel observations impact the risks?*
3. *How do the diverse visited locations impact the risks?*

In order to answer the above questions, this dissertation adopts a method to measure the re-identification risks. The *k*-anonymity method was utilized [14]. It is a kind of anonymization in the scope of generalization by adding at least *k-1* other tuples to make it challenging to re-identify the target [11]. This study considers different QIDs' quantity and quality to measure the risks from real-world privacy breach cases: COVID-19 patients' travel logs data.

1.2.3 Quantification of Amounts of Information Regarding Privacy Issues

The last research subsection is related to quantifying the amounts of pieces of information regarding privacy risks. Other researchers, including our previous subsection, tend to concentrate on measuring privacy risks in accordance with given external information. This given information resulted from a specific individual's exposed external information. Accordingly, the observation data would not fit others who have different travel records compared to the specific individual. The researchers filtered out the subset of travel data that originated from others who had different travel logs compared to the external observation. While the filtering method can re-identify individuals effectively, this approach might be suitable only for identifying individuals with specific circumstances. With the filtered subset data, quantifying amounts of information would be challenging to gauge privacy risks in the entire dataset. With this regard, we shed light on the following research questions.

1. *What is the relationship between privacy risk and quantification of information amount?*
2. *What methods can be utilized to quantify privacy risks, considering the entire dataset?*
3. *How can we quantify the amount of information regarding newly added information pieces?*

In order to address the above questions, this dissertation utilized an information theory-based methodology to quantify a piece of information, which is external observation data. In particular, the concepts of joint entropy and conditional entropy were adapted to this research to quantify the amount of information in a piece.

1.3 Dissertation Outline

The remainder of this dissertation is organized as follows:

Chapter 2 introduces a mathematical malware propagation model among AVs. As discussed in the preliminary section, Stuxnet is a well-known malware that does not just snatch confidential information but destroys physical facilities [9], [15]. In autonomous vehicle cybersecurity research, many researchers have warned of hackers' attacks through a Controller Area Network (CAN bus), especially OBD scanners [3], [16]. If OBD scanners are infected by malware, the direct contact between OBD scanners and AVs will be the main entrance of the malware attacks. There is a conceptual similarity of virus/malware propagation between the malaria disease and Stuxnet-style malware. The infected mosquitoes and OBD scanners can take a role as a vector to transmit the virus/malware between hosts: humans and AVs. This is why this dissertation adopts an epidemic model, such as the susceptible-infected (SI) model. With the conceptual similarity, the research adopts the malaria model to formulate a mathematical malware propagation model. Also, the research formulates a transportation operator model, demonstrating AVs' visiting mechanic shops patterns. For the operator model, the average annual mileage records per vehicle and property tax information are considered to formulate the transportation gravity model.

Chapter 3 explores the re-identification risks from smart card data. This section analyzes real-world privacy leaked cases of COVID-19 patient travel logs in South Korea. The travel logs are examined as external observation data to measure the risks from the card data. According to the South Korean government, officials investigated travel logs from patient statements and shared them publicly to make people avoid visiting the places [17], [18]. The travel logs were collected mainly by interviews, and there were resolution differences in pinpointing the locations. Language-based travel observations can vary spatial and temporal resolutions regarding patients' visiting times and locations. This dissertation considers the differences in information resolutions to understand the impact of language-based travel observation data on re-identifying the patients from dataset. As a measurement to gauge the risks, the well-known method, k-anonymity, is adopted [14]. A series of re-identification analyses are undertaken to output the different risks to understand the impact of the language-based external observations. In particular, this research explores the significance of differences in the risks, considering the spatial and temporal resolutions of the observations.

Chapter 4 develops metrics for quantifying a piece of information in terms of privacy risk. Based on the aforementioned research questions, this section focuses on quantifying amounts of information in the dataset. To this end, we created synthetic data to show individual travel records and quantify the fluctuations of privacy risks, considering the entire dataset. We adapted Claude Shannon’s information theory methodologies. In Shannon’s research, entropy can estimate a dataset's uncertainty or unpredictability level [19]. When estimating entropy, the probabilities are input variables calculated from every distribution given conditions. Thus, the information theory-based method to measure privacy risks is not biased like filtered methods from existing studies. In particular, this dissertation focuses on joint entropy and conditional entropy to recognize the amount of a piece of information. Based on the basic concept of conditional and joint entropy, joint entropy can be estimated by accumulating conditional entropies as increasing the number of conditions. Chapter 4 delves into the impact of adding pieces of information, which can exacerbate privacy risks.

Lastly, Chapter 5 provides concluding comments and future plans.

Chapter 1 References

[1] D. Hahn, A. Munir, and V. Behzadan, “Security and Privacy Issues in Intelligent Transportation Systems: Classification and Challenges,” *IEEE Intell. Transp. Syst. Mag.*, vol. 13, no. 1, pp. 181–196, 2021, doi: 10.1109/MITS.2019.2898973.

[2] European Commission, “European Commission Mandate,” M/453 EN, 2009. [Online]. Available: https://ec.europa.eu/growth/tools-databases/mandates/index.cfm?fuseaction=search.detail&id=434#.

[3] S. Parkinson, P. Ward, K. Wilson, and J. Miller, “Cyber Threats Facing Autonomous and Connected Vehicles: Future Challenges,” *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 2898–2915, 2017, doi: 10.1109/TITS.2017.2665968.

[4] K. Khorsravinia, M. K. Hassan, R. Z. A. Rahman, and S. A. R. Al-Haddad, “Integrated OBD-II and mobile application for electric vehicle (EV) monitoring system,” in *2017 IEEE 2nd International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, Oct. 2017, vol. 2017-Decem, no. October, pp. 202–206, doi: 10.1109/I2CACIS.2017.8239058.

[5] T. Zhang, H. Antunes, and S. Aggarwal, “Defending Connected Vehicles Against Malware: Challenges and a Solution Framework,” *IEEE Internet Things J.*, vol. 1, no. 1, pp. 10–21, Feb. 2014, doi: 10.1109/JIOT.2014.2302386.

[6] K. Kim, J. S. Kim, S. Jeong, J.-H. Park, and H. K. Kim, “Cybersecurity for autonomous vehicles: Review of attacks and defense,” *Comput. Secur.*, vol. 103, p. 102150, Apr. 2021, doi: 10.1016/j.cose.2020.102150.

[7] J. La Grassa, “CEOs of Ontario hospitals hit by ransomware attack break down impact on operations , patients,” *CBC News*, Windsor, ON, Oct. 23, 2023.

[8] J. Ibarra, H. Jahankhani, and S. Kendzierskyj, “Cyber-Physical Attacks and the Value of Healthcare Data: Facing an Era of Cyber Extortion and Organised Crime,” 2019, pp. 115–137.

[9] J. P. Farwell and R. Rohozinski, “Stuxnet and the Future of Cyber War,” *Survival (Lond).*, vol. 53, no. 1, pp. 23–40, Feb. 2011, doi: 10.1080/00396338.2011.555586.

[10] P. A. Pérez-Martínez and A. Solanas, “W3-Privacy: the Three Dimensions of User Privacy in LBS,” in *12th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2011, pp. 1–2, [Online]. Available: http://crises2-deim.urv.cat/docs/publications/conferences/673.pdf.

[11] Y. Li, D. Yang, and X. Hu, “A differential privacy-based privacy-preserving data publishing algorithm for transit smart card data,” *Transp. Res. Part C Emerg. Technol.*, vol. 115, p. 102634, Jun. 2020, doi: 10.1016/j.trc.2020.102634.

[12] K. Mayes and K. Markantonakis, *Smart cards, tokens, security and applications: Second edition*. 2017.

[13] R. Baheti and H. Gill, “Cyber-Physical Systems,” *impact Control Technol.*, vol. 12, no. 1, pp. 161–166, Nov. 2011, doi: 10.1007/s10559-017-9984-9.

[14] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” *Int. J. Uncertainty,*

Fuzziness Knowledge-Based Syst., vol. 10, no. 05, pp. 557–570, Oct. 2002, doi: 10.1142/S0218488502001648.

[15] D. Kushner, “The real story of Stuxnet,” *IEEE Spectr.*, vol. 50, no. 3, pp. 48–53, Mar. 2013, doi: 10.1109/MSPEC.2013.6471059.

[16] J. Hong, “Cyber security issues in connected vehicle of intelligent transport system,” *Indian J. Sci. Technol.*, vol. 9, no. 24, pp. 1–7, 2016, doi: 10.17485/ijst/2016/v9i24/96027.

[17] The_Government_of_the_Republic_of_Korea, “Flattening the curve on COVID-19: How Korea responded to a pandemic using ICT,” 2020. [Online]. Available: https://www1.undp.org/content/seoul_policy_center/en/home/presscenter/articles/2019/flattening-the-curve-on-covid-19.html.

[18] G. Jung, H. Lee, A. Kim, and U. Lee, “Too Much Information: Assessing Privacy Risks of Contact Trace Data Disclosure on People With COVID-19 in South Korea,” *Front. Public Heal.*, vol. 8, no. June, Jun. 2020, doi: 10.3389/fpubh.2020.00305.

[19] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–656, Oct. 1948, doi: 10.1002/j.1538-7305.1948.tb00917.x.

CHAPTER 2

A MATHEMATICAL MODELLING OF STUXNET-STYLE AUTONOMOUS VEHICLE MALWARE

2.1 Introduction

2.1.1 Background

In mid-May 2017, powerful computer malware invaded networked computer systems around the globe. It triggered a wave of aftershocks, holding data for ransom on more than 200,000 computers in 150 countries. In September 2020, ransomware hit Universal Health Services, one of the largest healthcare providers in the United States. The attack forced a network shutdown throughout its facilities, impacting patient data and laboratory systems. It resulted in the cancellation of surgeries and attributed most of the unfavourable impact to facilities [1]. Such events become highly probable because a combination of widespread software homogeneity and unprecedented levels of connectivity creates an ideal climate for infectious malicious software (malware, for short) [2]. This evolving cyber-attack landscape leads to the sobering conclusion that malware and its self-propagating programs will be a new and growing threat to autonomous vehicle (AV) systems. Many cybersecurity experts [3-5] identified critical potential risks and challenges posed by epidemic-style AV malware, namely AV epidemic.

AVs are also susceptible to the same cybersecurity risks as networked computers. Vehicles are becoming computerized to control airbags, engines, brakes, ventilation, and advanced driver assistance systems (ADAS). Internally networked electronic systems with outdated security systems could offer hackers shortcuts to transmit malware through the infotainment system or diagnostic tool, and they can access the whole or subsection of controls related to the vehicle's security [6-7]. Further, vehicles will increasingly be connected by vehicle-to-everything (V2X) technologies for traffic efficiency, safety, and cooperative driving. Many researchers have identified vulnerabilities where the malware can control vehicles [8-9].

2.1.2 Features of Stuxnet-style Attack

Stuxnet is a well-known malware in the computer security community because it was first aimed at physical destruction. It is a convoluted program to invade and snatch the controlling system through semi-autonomous means [10]. Farwell and Rohozinski insist that over 60,000 computers were already infected in 2010. Stuxnet is believed to have been conceived to cripple Iran's uranium centrifuges, and it infected the industrial control systems through Windows computers. Its attack on Iran's centrifuges sped up or slowed down the centrifuges until they destroyed themselves, while the operators' computer screens showed everything was working as normal [11]. There was another attack with a modified Stuxnet. The U.S. National Security Agency tried to attack North Korean centrifuges using upgraded Stuxnet, but it failed [12]. Both cases demonstrate the threat of cyber-attacks to the physical system. Chen and Abu-Nimeh [13] highlighted the features and threats of Stuxnet. Stuxnet targets Windows PCs, which program the specific logic controllers that control automated physical processes in a standard industrial control system. They argued that Stuxnet is not restricted to computers but can affect critical physical infrastructure and threaten real lives. Hence, Stuxnet can attack cyber-physical systems more than other types of malware, which mainly focus on damaging and stealing intellectual properties and information.

Stuxnet targeted highly specialized industrial systems in critical high-security infrastructure in 2010. Karnouskos [14] argued that Stuxnet could be transmitted by plugging a USB flash drive into Windows computers. The infected computers and monitors fake industrial process control sensor signals; accordingly, there were no alarms or shutdowns due to abnormal working until the facilities were destroyed. Another study [7] pointed out the possibility of vehicle infection by first-infected diagnostic equipment. At the annual DerbyCon hacker convention in 2015, the security expert team [15] demonstrated an experiment attacking vehicles by simply infecting the mechanic's computer that runs the diagnostic software with Stuxnet-style malware. They turned off the vehicle's airbag system without the diagnostic software noticing the misdeed as proof of concept. Thus, hackers can easily infiltrate Stuxnet-style malware into AVs through infected diagnostic devices.

Therefore, it is noticeable that Stuxnet has lethal features that can devastatingly damage physical facilities, readily travel on USB sticks, spread through Windows computers stealthily and be a real threat to the security of AVs.

2.1.3 Reasons for a Growing Interest in Stuxnet-style Attacks in Targeting AVs

There are particular reasons that cyber security experts have focused on this specific malware. First, AVs are complex cyber-physical systems with many functions integrated by collaborative interactions between cyber (i.e., electronic control units, sensors) and physical systems (i.e., ADAS, steering, braking). Further, Stuxnet is malware targeting physical access to a vehicle, causing physical damage [15]. Therefore, the AV epidemic can cause severe physical consequences like vehicle accidents. Terrorists may exploit the infected vehicles to create a blast and massive collisions with all the vehicles filled up with fuel [16]. These are clear implications that the AV epidemic represents a significant threat to entire transport systems.

Second, a Stuxnet-style attack on AVs is plausible. Managing many computers at mechanics' shops with an up-to-date security upgrade would be challenging. Computers are used for various purposes, not only for diagnostics but also for the Internet, which could be a gateway to malware transmission [15]. Also, the on-board diagnostic (OBD-II) scanner randomly connects the computers with vehicles, and the hackers can spread the Stuxnet worm by connections. Stuxnet can be a scalable, similar way of spreading disease. Due to its sophistication, many experts [7], [15], [17] have warned that Stuxnet provides a blueprint for carrying out similar attacks on AVs.

Third, it can become more compatible with any vehicle brand and more scalable. Szijj et al. [15], who presented a Stuxnet attack on Audi brand vehicles, asserted that any brands using vulnerable diagnostic applications could easily be exposed to the malware. It can repeatedly replicate, hide, and cause an epidemic.

2.1.4 Contributions

Despite significant inroads into AV technology in recent years, a systematic study of the AV epidemic is lacking. Traditional analysis tools are insufficient to cope with the emerging security challenges of AVs or adequately predict their behaviours. This study aims to analyze the behaviour of the AV epidemic to bridge the methodological gaps between traditional tools and new challenges. Well-developed mathematical models in epidemiology [18-20] provide a conceptual platform that can be adapted to describe the AV epidemic. Therefore, this study seeks to understand and enhance the security of AV systems.

The primary contribution of this study is to develop the AV malware propagation model that provides a clear understanding of the replication methods. We believe there is a similarity regarding the propagation between malaria and Stuxnet as a vector-host model. Malaria cannot spread from person to person directly, but mosquitoes can spread malaria as a disease vector. The OBD-II scanners and computers at mechanic shops could take the role of malaria mosquitoes in spreading malware between AVs. Thus, adopting a vector model of epidemic propagation to express malware spreading could be a novel approach. Also, we incorporate a transportation operator model to present individual vehicles visiting mechanic shops. The model formulates the frequency and intensity of the mechanic shop visits. The novelty and pertinence of the model lie in its adaptation and modification from epidemiology to AV epidemic and its combination of the complementary strengths of network science to develop new methods. Therefore, the proposed model describes the characteristics of the AV epidemic built on the contact between mechanic shops and vehicles. Robust epidemic control techniques can be developed and integrated to eradicate the malware in a simulation environment. It also can be implemented in the real world with the proposed model in future studies. This study is the first attempt to recognize the significant threats of AV malware and defend against the AV epidemic.

The remainder of the chapter is organized as follows. Section 2 discusses related works, especially for AVs or connected and autonomous vehicles (CAVs) vulnerabilities and malware propagation models. Sections 3 and 4 present the conceptual framework of the proposed model and mathematical explanation. Section 5 consists of numerical experiments and the associated results. Section 6 summarizes the main findings and future research directions.

2.2 Related Work

2.2.1 Connected and Autonomous Vehicle Vulnerabilities

Modern vehicles consist of more than 100 electronic control units (ECUs); about 100 million lines of code are used to operate the vehicles [21]. They are tightly interconnected via internal networks based on the Controller Area Network (CAN) bus standard, which is an international standardized serial communications bus [22]. Corrigan described how information is conveyed between devices on a network, and the CAN transceiver connects physical subsections in the vehicle through the CAN communication protocols [22].

Many researchers have aroused attention regarding the vulnerability of the highly developed CAN network in AVs [23], [25], [26]. Amoozadeh et al. [28] analyzed AVs' security vulnerabilities in cooperative driving. They explained the physical types of attacks: falsification and eavesdropping of beacon information, inducing collision in congested areas with false distance information between vehicles. Also, Parkinson warned that access control within the OBD protocol could allow hackers to control the vehicles easily [24]. Hong [27] emphasized that security challenges could emerge due to automotive architecture design, insisting that OBD security should be developed to detect malware. Hackers can manipulate AVs physically by distributing malware.

Ethernet is attracting widespread interest with several advantages: higher throughput rates, smaller wiring than CAN, etc. [36]. Both CAN bus and Ethernet networking would be susceptible to infection during the direct connection to vehicles for diagnosis. Therefore, the interconnection between AVs and diagnostic devices via CAN bus and Ethernet can exacerbate the risk of invasion by hackers. However, there has been little discussion on malware propagation through physical contact.

On the other hand, according to Upstream Security's report, the shares of physical access attacks have decreased in the last decade [29]. They argued that remote contact-based attacks are more feasible. Even though the shares of physical contact-styled attacks have a decreasing tendency these years, we still need to study the physical malware propagation. Both remote and physical contact-based attacks related to hackers can bring devastating risks threatening people's lives.

2.2.2 Risks of Stuxnet-style Malware

Among several attacks on automotive control systems, Stuxnet-style malware may utilize vulnerabilities in equipment or computers at mechanic shops. Stuxnet mainly consists of .dll files that let hackers monitor and intercept data between the diagnostic application and the cable [13], [15], [17]. The detailed process of encroaching and manipulating vehicles is beyond the scope of this research, so we briefly explain the malware's propagation process. Stuxnet would be uploaded first to the mechanic shop's computer network, possibly using USB. Then, it infects the entire vehicle's computer system [11]. This is realistic because the mechanic's computer commonly accesses vehicles during routine maintenance for diagnosis. Thus, if an infected vehicle connects to the OBD scanners during the inspection, it could spread to any vehicle subsequently plugged into the same diagnostic device in the mechanic shop.

In addition, regional spreading of the malware could also be possible when the infected vehicles visit other shops. This is how hackers may dominate most vehicles without noticing any suspicious symptoms before their zero-day, which is the moment the hackers create massive and devastating attacks. As discussed, Stuxnet could spread via USB (OBD-II) in air-gapped facilities and deceive the self-monitoring system until it is detected [11], [13], [17]. The malware will encroach on the entire system or every registered vehicle confidentially but expeditiously. At last, the malware reaches devastating physical explosions or massive and severe collisions under hackers' deliberate terror plans.

2.2.3 Malware Propagation Based on Epidemic

Many studies have been published on the mathematical modelling of malware propagation based on the epidemic [32-33]. Meisel's team reviewed many computer networking studies inspired by biology. In particular, they found that the studies applied mathematical modelling of epidemics to spread computer viruses through computer networks [31]. Del Rey [30] reviewed several mathematical models that simulate propagation in network computers or mobile devices. In his review, he claims that the similarity between the behaviour of epidemic viruses and computer malware is the groundwork for creating mathematical models of malware propagation. Also, Peng et al. focused on similarities between mobile malware and biological viruses in their research [34]. They, especially, reviewed papers that leverage mathematical models and simulator-based approaches to understand mobile malware propagation.

Trullols-Cruces et al. conducted a simulation-based analysis to depict worm propagation in CAVs via V2V communication [35]. Their approach highlighted the vulnerability of CAVs with a numerical model, considering vehicle inter-arrival patterns and road network-wide in simulation circumstances. A simulation-based approach would be generally limited in obtaining universal results, and a well-developed mathematical method could be applied in various circumstances.

2.2.4 Lessons from Related Works

As discussed in the above studies, AVs could confront malware attacks through physical contacts, such as CAN bus and Ethernet. Understanding malware propagation is a fundamental prerequisite for protecting vehicles. The lessons from the prior studies indicated that similarities between malware and epidemic viruses could provide robust theoretical groundwork for the mathematical modelling of malware propagation. We considered a unique feature of Stuxnet propagation that transmits through OBD scanners similar to the transmission of malaria disease, a vector-host disease. We summarized the lessons from existing studies in Table 2-1. We manifested that AVs are susceptible to direct connection malware attacks, and well-organized epidemic mathematical modelling could prepare us for potential attacks on AVs.

Table 2-1 Summary of related works

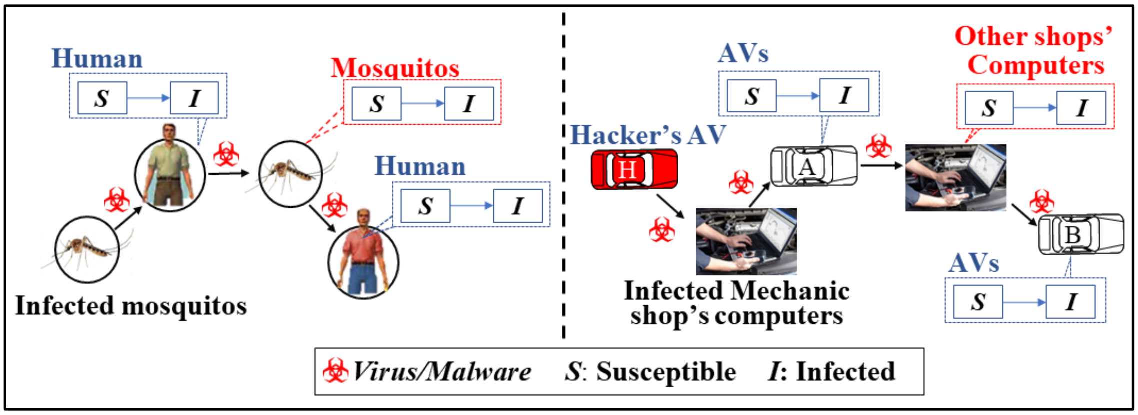
Topic category	Reference	Attack gateways	Approach and experiment
CAVs/AVs vulnerability & protection	[23]	OBD, OTA, Internet, Aftermarket equipment	Reviewing possible attacks on the CAVs and introducing a Cloud-based malware defense service
	[24-27]	OBD, Media (USB), Short-range communication (Bluetooth), Radio, etc.	Reviewing published papers to categorize vulnerabilities of AVs or CAVs from a variety of attack gateways
	[28]	Wireless communication	Analyzing security attacks in the car-following model using a simulator
	[29]	Remotely 80 %, physically 20 %	Analyzing the reported automotive attack incidents since 2010 and classified attack type
Risk of Stuxnet malware	[11], [13], [15], [17]	USB	Introducing an overview of the Stuxnet worm, especially for its ability to propagate by installing a malicious .dll file through a USB
Malware propagation model based on epidemic	[30-33]	Computers	Reviewing published papers to analyze proposed mathematical models in computers
	[34]	Smartphones	Reviewing published papers handling malware propagation in smartphones with mathematical models and simulators
Simulation-based worm propagation	[35]	V2V communication	Analyzing worm propagation in CAVs through V2V communication by using simulation

2.3 The Conceptual Framework for The Malware Propagation Model

2.3.1 Conceptual Similarities Between Malaria and Stuxnet

When a mosquito bites a malaria-infected human, the microscopic malaria parasites mix with the mosquito's saliva. Then, the virus can be delivered to other people in the mosquito's next bite [37]. Conceptual similarities exist since the spreads share similar mathematical underpinnings [38]. Biological diseases and AV malware spread rely on contacts between entities, and the Stuxnet-style attack is achieved by an indirect transmission involving an intermediate host. In this respect, the spread of Stuxnet (from vehicle to vehicle via infected mechanic's computers) is similar to the spread of malaria (from human to human via mosquitoes). Inspired by an indirect transmission mechanism through a mosquito carrying malaria, the model captures malware propagation in AVs (humans) and mechanics/dealerships (mosquitos).

The model adapted and modified the Susceptible-Infected (SI) epidemic model [19] to describe the malware infection process. Fig. 2-1 (a) describes how malaria spreads to humans through infected mosquito bites, and the similarity of transmission of Stuxnet was depicted in Fig. 2-1 (b).



(a) Malaria spreading via mosquitos (b) Malware spreading via OBD & computers

Fig. 2-1. Similarity between malaria and Stuxnet AVs epidemic

2.3.2 Model of Visiting Mechanic Shops

Unlike the anonymous and remote world of computer malware over the Internet, Stuxnet can spread only when they have physical contact with AVs. Thus, it is important to include a realistic visiting mechanic shop model.

Fig. 2-2 illustrates malware propagation from an initial hacker's attack to the overall propagation in a city over time. Hacker's house is on the bottom left of the figure. When the hacker visits a mechanic shop, the OBD scanners and computers in the shop will be infected by malware, and whoever visits the shop will continue to be infected. If the infected vehicles visit other susceptible mechanic shops in different areas, regional malware propagation will be implemented over time. The speed of malware propagation depends on the frequency and combination of visiting mechanic shops. Thus, another critical factor of this study is to model human behaviour, such as visiting mechanic shops, which will be discussed in Section 4.

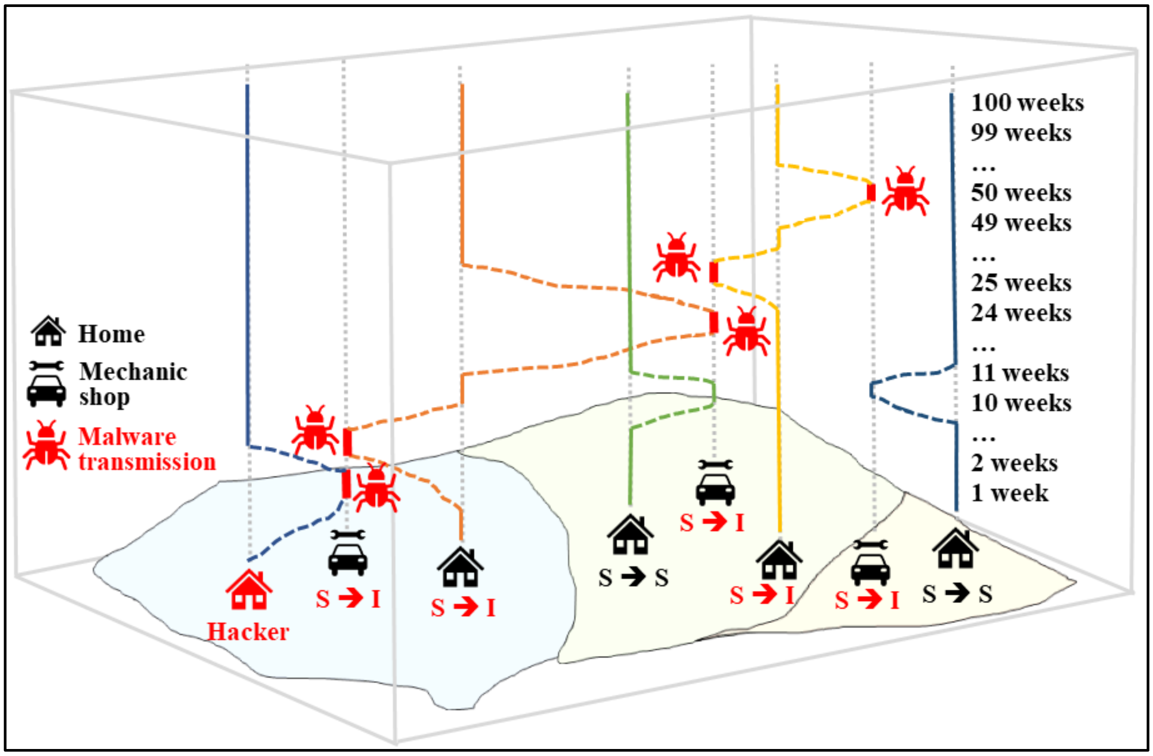


Fig. 2-2. Contacts network between vehicles and mechanic shops

2.3.3 Malware Propagation Model Assumptions

For our model, we used several assumptions. First, the OBD scanner would take the role of checking vehicles as a universal device in the future. Stuxnet can transport via OBD scanners and infect mechanic shops’ computers with the removable drive propagation method: a way of copying itself to inserted removable drives like USB and be executed by the auto-execution while scanning the vehicles [17].

Second, the security software updates would not detect Stuxnet before its zero-day. As we mentioned in the previous section, the main feature of Stuxnet is to lurk before zero-day, which could hinder detection even if the vehicles are updated [23]. Thus, the updates would not be able to detect Stuxnet until zero-day.

Third, we adopted the SI epidemic model instead of the SIR or SIRS epidemic model due to the features of Stuxnet-style malware that will not be recognized until the zero-day and will not heal spontaneously. The stealth ability of the malware can also preclude recovery from the infection, which is why we considered the SI epidemic model.

Fourth, hackers can spread Stuxnet through physical contact instead of wireless communication. Remote and physical contact attacks could bring devastating results equally.

Lastly, Stuxnet can spread to different car brands and models. Also, hackers would not create variants to enhance malware's ability to lurk.

2.4 Mathematical Modelling of the Spatiotemporal Propagation of Malware

Based on the review of related works and conceptual framework in the previous sections, we created a mathematical malware propagation model among AVs adopting the malaria propagation model. The primary reason for the adoption is the conceptual similarities between malaria and AV Stuxnet-style malware. Stuxnet-style malware can transmit malicious codes to susceptible objects through direct contact via OBD scanners like malaria mosquitoes. Also, the malaria model consists of a well-developed theoretical and mathematical model that brings about universal results compared to simulation-based approaches [39].

For the reasons mentioned above, our proposed model is a coupled system of differential equations in the discrete space (residential zones and mechanic shops) and discrete-time (weeks) domains. The model consists of a malaria model and a transportation operator model.

2.4.1 Notation

As we adapt and modify the malaria model (Vector-host), we compare the notation to the traditional malaria model.

Table 2-2. Notation
(Conceptual Analogy Between AVs Epidemic and Vector-host Model)

Notation	AVs Epidemic	Malaria (Vector-host) model
N_i	total number of AV population in zone i, $(X_i + Y_i)$	total number of the human population (host) in community i, $(X_i + Y_i)$
$X_i(t)$	number of susceptible vehicles at time t, in zone i	number of susceptible humans at time t, in community i
$Y_i(t)$	number of infected vehicles at time t, in zone i	number of infectious humans at time t, in community i
R_j	total number of OBD scanners population in mechanic shop j, $(P_j + Q_j)$	total number of mosquitoes (vectors population) in community j, $(P_j + Q_j)$
$P_j(t)$	number of susceptible OBD scanners at time t, in mechanic shop j	number of susceptible mosquitos at time t, in community j
$Q_j(t)$	number of infected OBD scanners at time t, in mechanic shop j	number of infectious mosquitos at time t, in community j
α	probability that an infectious mechanic shop's OBD scanner transmits the malware to susceptible AV or vice versa	probability that an infectious mosquito transmits the infection to susceptible humans or vice versa
δ_{ij}	number of visiting vehicles from zone i to mechanic shop j	(No similar variables)

A city has E resident zones and M mechanic shops (i = 1, 2, …, E, j = 1, 2, …, M). The population of AVs in zone i, N_i are partitioned into two explicit and disjoint infection groups at time t. The number of susceptible and infected vehicles at time t, in zone i is denoted $X_i(t)$ and $Y_i(t)$, respectively. It is noted that $X_i(t) + Y_i(t) = N_i$. The total number of OBD scanners in mechanic shops, denoted $R_j(t)$ is subdivided into mutually exclusive compartments of susceptible OBD scanners $P_j(t)$ and infected ones $Q_j(t)$ so that $P_j(t) + Q_j(t) = R_j$. Susceptible (if previously unexposed to the malware) passes into an infectious group. The population of vehicles (N_i) and OBD scanners (R_j) are constant values.

2.4.2 Transportation Operator Model

If we had applied real-world data: records of visiting mechanic shops from residential areas, we could have measured malware propagation speed directly and more concisely. However, it is challenging to get the data in most cases, so we modelled the transportation operator model to capture the patterns of visiting mechanic shops. The transportation operator model consists of two essential factors. One is estimating the demand for visiting mechanic shops per residential zone i, similar to trip production in the traditional four-step transportation modelling. The amount of production can be estimated using statistical data, and we elaborate on this in the next section.

Another factor relates to trip distribution to capture malware propagation in a city. The trip distribution can be estimated by a gravity model [40]. Longini utilized the transportation gravity model to deal with the issue of the absence of real-world data for estimating the propagation of Hong Kong influenza [41].

We defined δ_{ij} as a transportation operator model representing the number of weekly vehicle visits from zone *i* to mechanic shops *j*. The *E* × *M* matrix δ_{ij} is assumed to be symmetrical, i.e., $\delta_{ij} = \delta_{ji}$. In (1), δ_i accounts for the number of departing vehicles from zone *i* to *M* mechanic shops, and δ_j indicates the number of arriving vehicles from *E* resident zones to mechanic shop *j*. *D* is the distance between zone *i* and mechanic shop *j* that restrains the trip between origin and destination as an impedance factor. Lastly, the K_{ij} is a calibration parameter. This parameter is applied to calibrate the gap between the result of $\frac{\delta_i \times \delta_j}{D^2}$ and the summation of the number of vehicles visiting from *i* to *j*.

$$\delta_{ij} = K_{ij} \times \frac{\delta_i \times \delta_j}{D^2} \qquad (1)$$

2.4.3 SI Epidemic Model

The mathematical equations could present the infection progress from *X* to *Y* and *P* to *Q* groups [20]. Susceptible vehicles at time *t*, which is $X_i(t)$ in the notation, are infected by infected OBD scanners, $Q_{\mathrm{j}}(t)$. Also, the susceptible OBD scanners at time *t*, $P_j(t)$ are infected by the malware from the infected vehicles $Y_i(t)$. The initial condition of the malware transmission is that a mechanic shop *h* is exposed to the malware at time 0. The initial conditions are for (*i* = 1, 2, ..., *E*, *j* = 1, 2, …, *M*, *t* = 0, 1, 2, …, 100) are following:

$$X_i(0) = N_i,\ \forall i \qquad (2)$$

$$P_j(0) = R_j,\ \forall j \qquad (3)$$

$$Y_i(0) = 0,\ Q_j(0) = 0, \quad \forall j \text{ but } h \qquad (4)$$

$$Q_h(0) = 1 \qquad (5)$$

It is noted that the malware can be transmitted only through the connection between vehicles and OBD scanners. Equations (6) and (7) represent the changing number of susceptible and infected vehicles in zone *i* at time *t*, respectively. In the equations, the term A, $\delta_{ij}\cdot\frac{X_i(t)}{N_i}$, represents the number of susceptible vehicles from zone *i* to mechanic shop *j*. The term B, $\alpha\cdot\frac{Q_j(t)}{R_j(t)}$, is a probability of infection of a vehicle that visits mechanic shop *j* at time *t*. α is a probability that an infectious mechanic shop computer transmits the malware to a susceptible AV. The result of multiplying the number of susceptible vehicles visiting the mechanic shop and the infection probability is a newly infected number of vehicles at the mechanic shop *j* at time *t* (term C). When we estimate the number of susceptible vehicles in zone *i* at time t+1, term C is subtracted from $X_i(t)$ in (6). Term C also can be added to $Y_i(t)$ when we calculate the number of infected vehicles in (7). Fig. 2-3 details the process of vehicle infection at the mechanic shop *j* from the terms A to C.

$$X_i(t+1) = X_i(t) - \sum_{j=1}^{M} \underbrace{\underbrace{\delta_{ij}\cdot\frac{X_i(t)}{N_i}}_{\text{Term A}}\cdot\underbrace{\alpha\cdot\frac{Q_j(t)}{R_j(t)}}_{\text{Term B}}}_{\text{Term C}} \tag{6}$$

$$Y_i(t+1) = Y_i(t) + \underbrace{\sum_{j=1}^{M} \delta_{ij}\cdot\frac{X_i(t)}{N_i}\cdot\alpha\cdot\frac{Q_j(t)}{R_j(t)}}_{\text{Term C}} \tag{7}$$

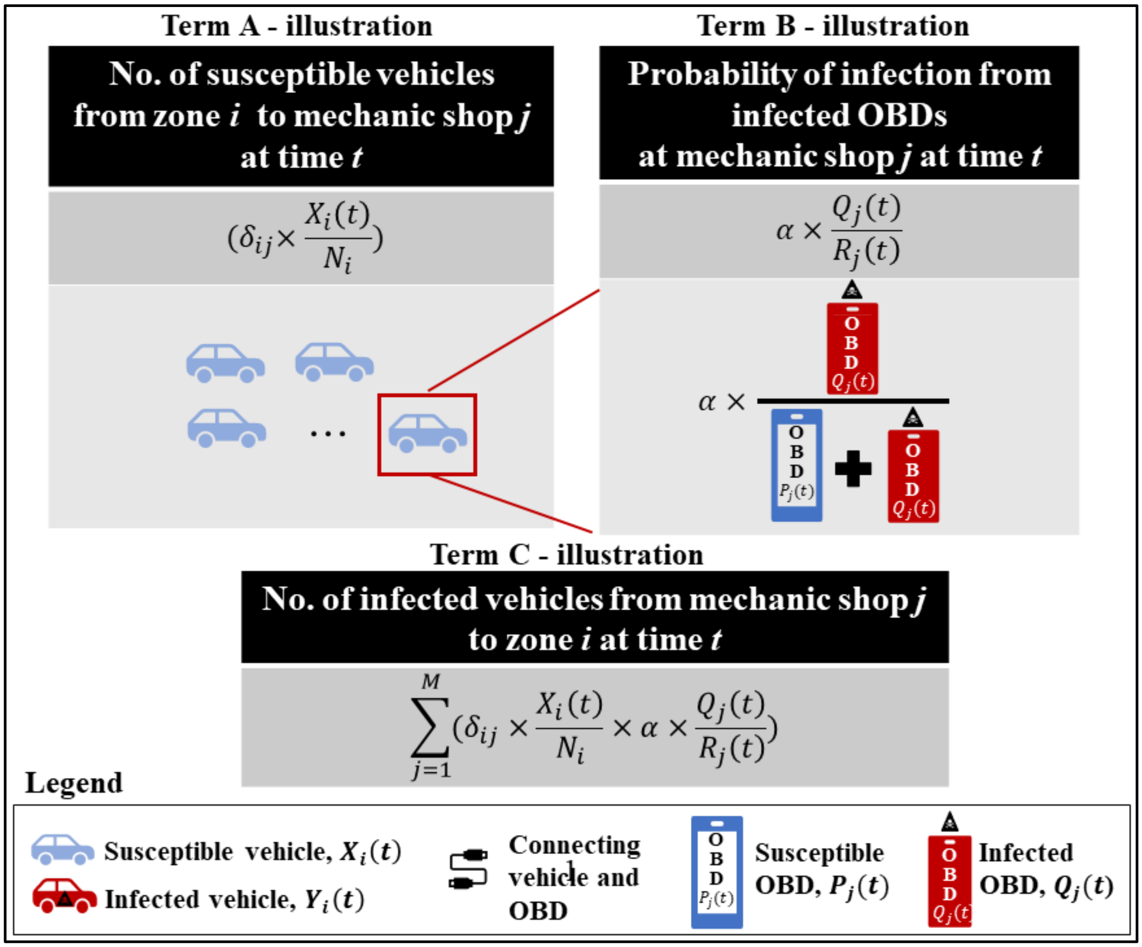

Term A - illustration
No. of susceptible vehicles from zone i to mechanic shop j at time t
$(\delta_{ij}\times\frac{X_i(t)}{N_i})$
Term B - illustration
Probability of infection from infected OBDs at mechanic shop j at time t
$\alpha\times\frac{Q_j(t)}{R_j(t)}$
$\alpha\times$
OBD $Q_j(t)$
OBD $P_j(t)$
OBD $Q_j(t)$
Term C - illustration
No. of infected vehicles from mechanic shop j to zone i at time t
$\sum_{j=1}^{M}(\delta_{ij}\times\frac{X_i(t)}{N_i}\times\alpha\times\frac{Q_j(t)}{R_j(t)})$
Legend
Susceptible vehicle, $X_i(t)$
Infected vehicle, $Y_i(t)$
Connecting vehicle and OBD
OBD $P_j(t)$
Susceptible OBD, $P_j(t)$
OBD $Q_j(t)$
Infected OBD, $Q_j(t)$

Fig. 2-3. Process of vehicle infection at mechanic shop *j*

Equations (8), (9) and Fig. 2-4 represent a process of OBD scanner infection at mechanic shops.

$$P_j(t+1) = P_j(t) - \underbrace{P_j(t)\cdot(\underbrace{1-\underbrace{(1-\alpha)^{\underbrace{\sum_{i=1}^{E}\delta_{ij}\cdot\underbrace{\frac{Y_i(t)}{N_i}}_{\text{Term D}}\cdot\frac{1}{R_j}}_{\text{Term E}}}}_{\text{Term F}}}_{\text{Term G}})}_{\text{Term H}} \qquad (8)$$

$$Q_j(t+1) = Q_j(t) + \underbrace{P_j(t) \cdot \left(1 - (1-\alpha)^{\sum_{i=1}^{E} \delta_{ij} \cdot \frac{Y_i(t)}{N_i} \cdot \frac{1}{R_j}}\right)}_{\text{Term H}} \quad (9)$$

The term D, $\delta_{ij} \cdot \frac{Y_i(t)}{N_i}$, represents the number of infected vehicles visiting from zone *i* to the mechanic shop *j*. Term *E* indicates how many infected vehicles are scanned by one susceptible OBD scanner at mechanic shop *j* at time *t* that could infer the infection risk of the susceptible scanner. Term F shows a case of the probability of no infection of the susceptible OBD scanner even with many contacts with infected vehicles. We subtract the α from the entire probability one, which is 100%, and the $1 - \alpha$ to the exponent of term E to calculate the probability of no infection. Lastly, we take term F from one to get the probability of infection of susceptible OBD scanner at mechanic ship *j* at time *t*. Therefore, term H represents the number of infected OBD scanners after connecting with the infected vehicles that visited the mechanic ship *j* from zone *i* at time *t*. By adding or subtracting the term H from the number of susceptible $P_j(t)$, or infected OBD scanners $Q_j(t)$, we can obtain the numbers of susceptible or infected OBD scanners at time t+1 in (8) and (9).

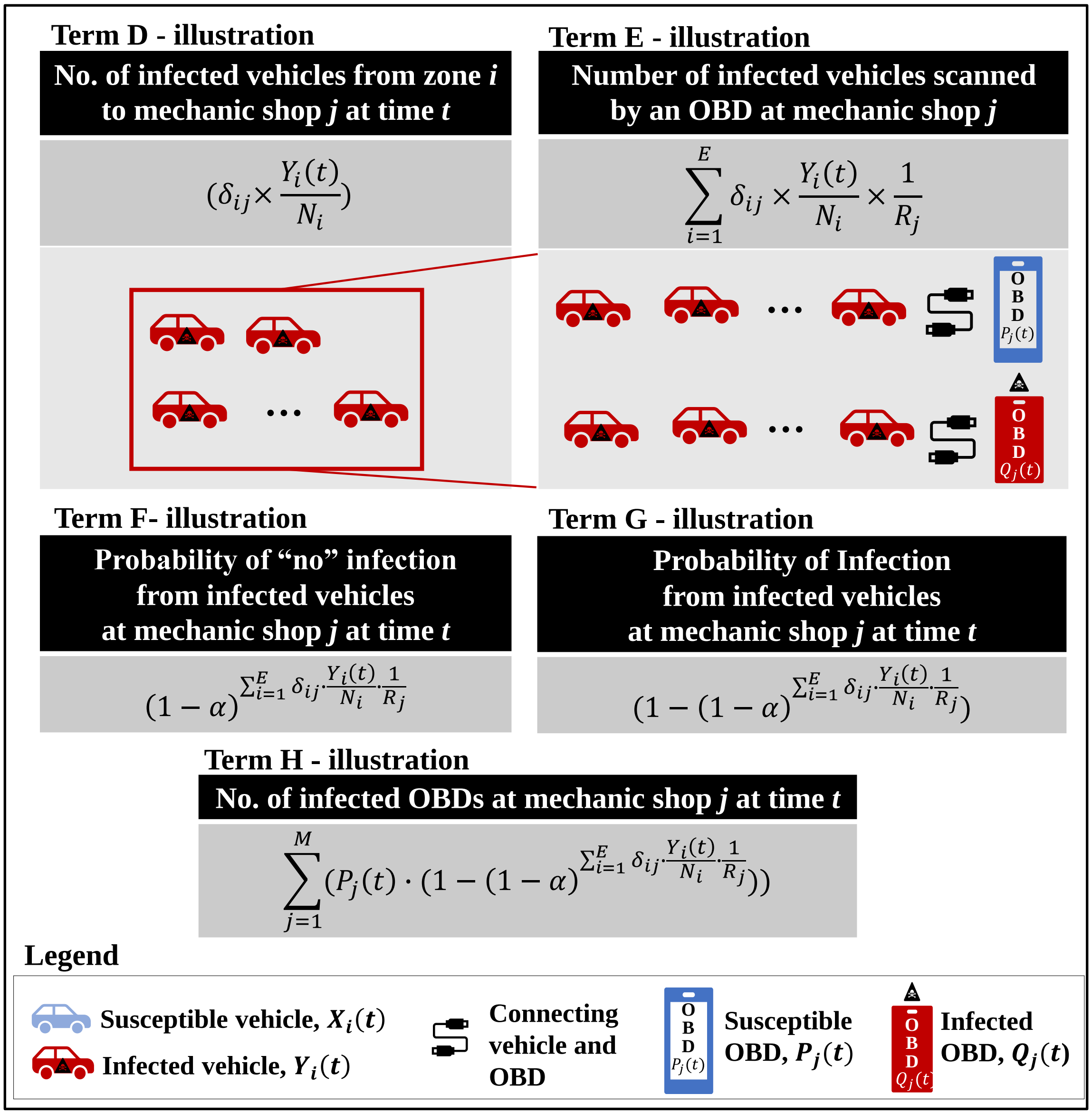
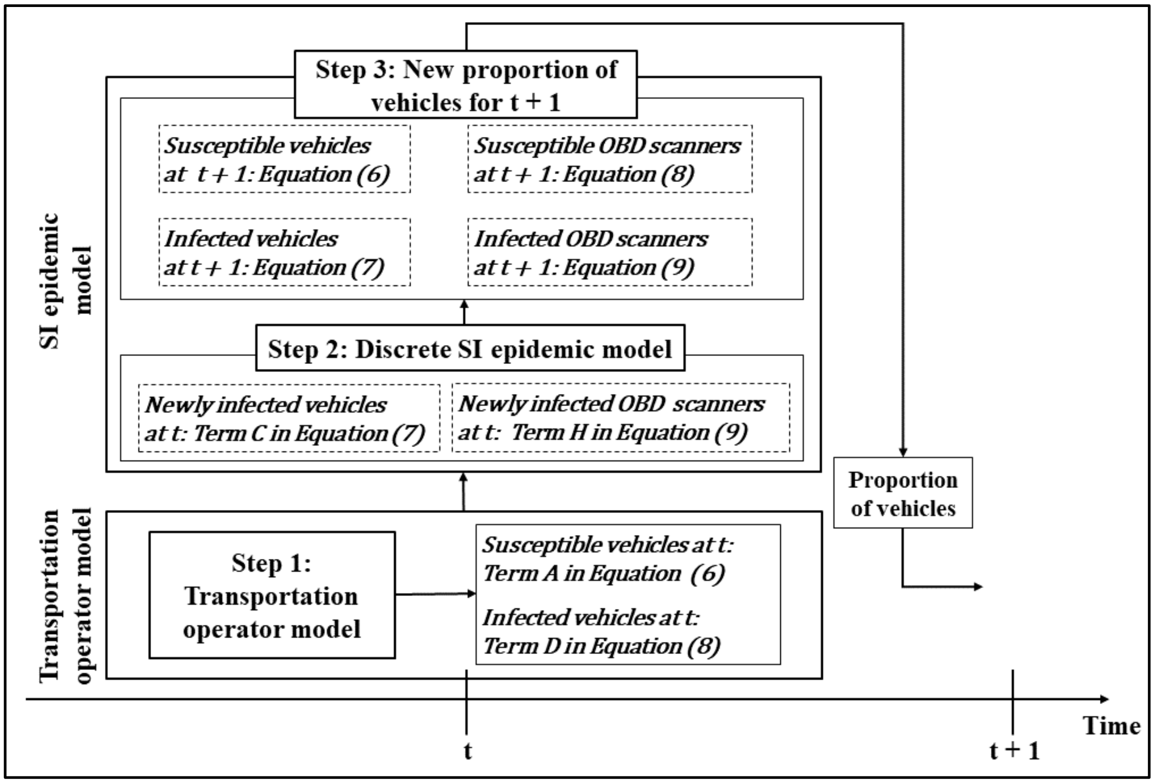

Term D - illustration
No. of infected vehicles from zone i to mechanic shop j at time t
$(\delta_{ij} \times \frac{Y_i(t)}{N_i})$
Term E - illustration
Number of infected vehicles scanned by an OBD at mechanic shop j
$\sum_{i=1}^{E} \delta_{ij} \times \frac{Y_i(t)}{N_i} \times \frac{1}{R_j}$
OBD $P_j(t)$
OBD $Q_j(t)$
Term F- illustration
Probability of "no" infection from infected vehicles at mechanic shop j at time t
$(1-\alpha)^{\sum_{i=1}^{E} \delta_{ij} \cdot \frac{Y_i(t)}{N_i} \cdot \frac{1}{R_j}}$
Term G - illustration
Probability of Infection from infected vehicles at mechanic shop j at time t
$(1-(1-\alpha)^{\sum_{i=1}^{E} \delta_{ij} \cdot \frac{Y_i(t)}{N_i} \cdot \frac{1}{R_j}})$
Term H - illustration
No. of infected OBDs at mechanic shop j at time t
$\sum_{j=1}^{M} (P_j(t) \cdot (1-(1-\alpha)^{\sum_{i=1}^{E} \delta_{ij} \cdot \frac{Y_i(t)}{N_i} \cdot \frac{1}{R_j}}))$
Legend
Susceptible vehicle, $X_i(t)$
Infected vehicle, $Y_i(t)$
Connecting vehicle and OBD
Susceptible OBD, $P_j(t)$
Infected OBD, $Q_j(t)$

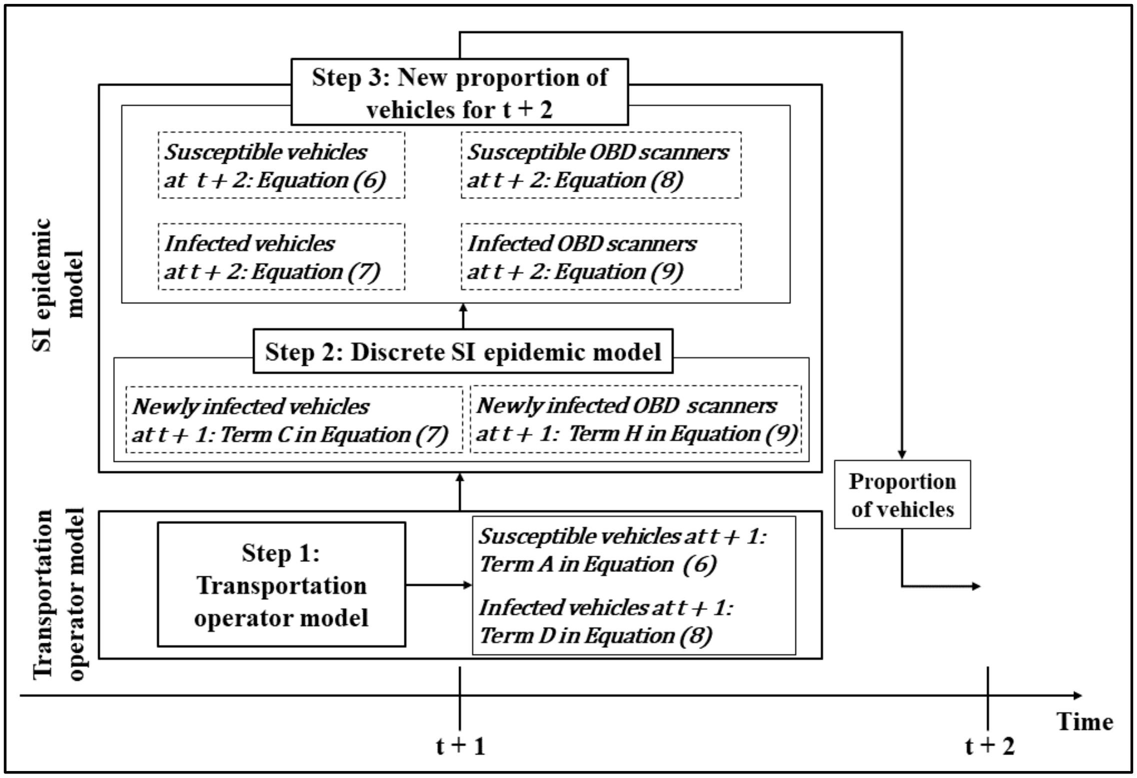
Fig. 2-4. Process of OBD scanners infection at mechanic shop *j*

2.4.4 The Framework of Mathematical Equations

The modelling framework in Fig. 2-5 summarizes the whole process of the two integrated models to depict malware propagation over time. After the initial attack by the hacker at time $t = 0$, the malware can propagate from one mechanic shop to the entire city by the two models. The transportation operator model can model visiting patterns from residential areas to mechanic shops, and vehicles regularly go to the mechanic shops. In Step 1, the transportation operator model selects the number of susceptible and infected AVs to visit mechanic shops. In the next step, the infection process between susceptible AVs and OBD scanners is modelled with the SI epidemic model. After visiting mechanic shops at time t, the newly infected AVs and OBD scanners will increase the proportion of infection. Then, the new proportion of susceptible and infected vehicles will be applied to choose the number of vehicles for the visits at $t + 1$. The new proportion of infected OBD scanners will be used in Step 2 at $t +1$.


Step 3: New proportion of vehicles for t + 1
Susceptible vehicles at t + 1: Equation (6)
Susceptible OBD scanners at t + 1: Equation (8)
Infected vehicles at t + 1: Equation (7)
Infected OBD scanners at t + 1: Equation (9)
SI epidemic model
Step 2: Discrete SI epidemic model
Newly infected vehicles at t: Term C in Equation (7)
Newly infected OBD scanners at t: Term H in Equation (9)
Proportion of vehicles
Transportation operator model
Step 1: Transportation operator model
Susceptible vehicles at t: Term A in Equation (6)
Infected vehicles at t: Term D in Equation (8)
Time
t
t + 1

(a) from time *t* to *t + 1*


Step 3: New proportion of vehicles for t + 2
Susceptible vehicles at t + 2: Equation (6)
Susceptible OBD scanners at t + 2: Equation (8)
Infected vehicles at t + 2: Equation (7)
Infected OBD scanners at t + 2: Equation (9)
SI epidemic model
Step 2: Discrete SI epidemic model
Newly infected vehicles at t + 1: Term C in Equation (7)
Newly infected OBD scanners at t + 1: Term H in Equation (9)
Proportion of vehicles
Transportation operator model
Step 1: Transportation operator model
Susceptible vehicles at t + 1: Term A in Equation (6)
Infected vehicles at t + 1: Term D in Equation (8)
Time
t + 1
t + 2

(b) from time *t + 1* to *t + 2*

Fig. 2-5 Framework of the integrating SI epidemic model and transportation operator model

2.5 Experimental Design

This section explains a numerical experiment based on the mathematical modelling of malware propagation. We utilized network and geostatistics data from Windsor, Ontario, Canada, as a study area. Windsor has a population of about 233,000 and 112,000 registered vehicles. We assumed that all 112,000 registered vehicles in our study area were AVs. This assumption is based on the premise that AVs are extensively computerized, posing a greater risk of hacker manipulation than traditional vehicles. Traditional vehicles largely depend on human control for critical functions like steering and braking. If a hacker successfully attacks AVs, we anticipate that the societal damages would be significantly more severe than those involving traditional vehicles. This potential for a more severe impact is why we modelled all vehicles in the experimental area as AVs. Then, we created 104 residential and 116 mechanic shop zones to simulate Stuxnet propagation. Fig. 2-6 shows the locations of mechanic shops and the residential zone centroids in Windsor.

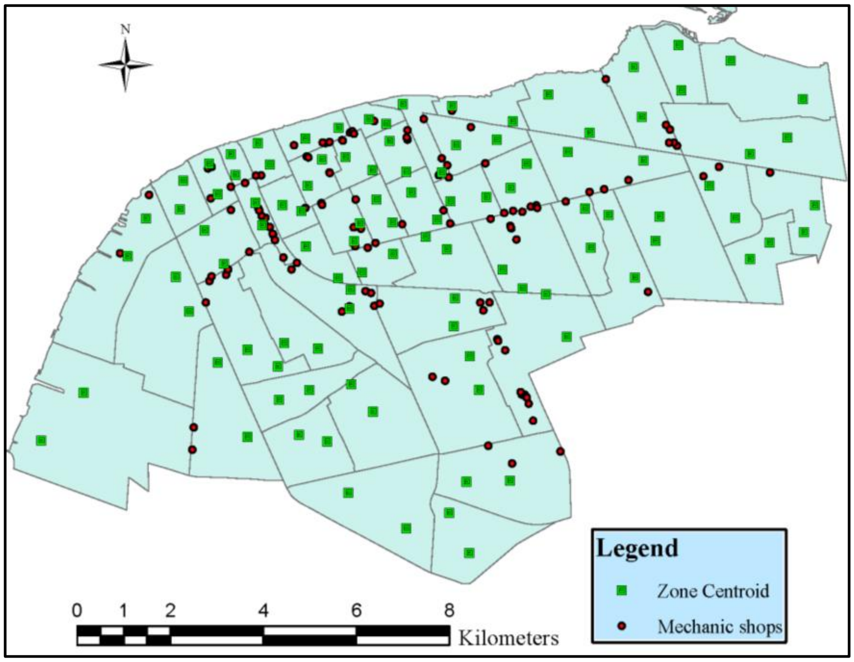


Fig. 2-6. Study area

2.5.1 Estimation of Transportation Operator Model (δ_{ij})

This study estimated the annual frequency of visiting mechanic shops for maintenance using the number of registered vehicles and the annual vehicle kilometres from Statistics Canada [42-43]. We estimated the average annual mileage per vehicle in Ontario for ten years ('00-'09) at 17,569 km/year.

Also, we investigated the maintenance items for vehicles per a specific mileage and considered the visiting mechanic shop cycles [44]. Table 2-3 shows the lists of maintenance by mileage. In our view, most vehicles would be diagnosed by the OBD scanners every 5,000 km mileage for a regular check-up with the maintenance items. Thus, the annual average number of visiting mechanic shops for maintenance using the OBD scanners was estimated at 3.5 visits/year.

As discussed in the previous section, it was challenging to investigate real-world data regarding records of visiting mechanic shops (δ_{ij}). To address the absence of real-world data, we adopted the transportation gravity model with alternative data: property tax records to consider the magnitude of the destination. The higher the property tax, the more operating income is needed to maintain the shops. Hence, we believe that property tax could represent the operating profits inferring the number of mechanic shops visiting. We investigated the tax records of the 116 mechanic shops from the City of Windsor data [45]. Table 2-4 indicates the tax and the number of OBD scanners in accordance with the size of the tax.

Based on the tax information, we estimated the transportation operator model δ_{ij} by using (1). The matrix δ_{ij} was estimated using the doubly constrained gravity model equation. As a result, the total weekly production was estimated at 7,580 vehicles/week.

The final δ_{ij} matrix is the symmetrical table consisting of zone centroids and mechanic shops.

Table 2-3. Lists of Maintenance by Mileage

Mileage (km)	Maintenance items
Every less than 5,000 km	- Engine oil and filter - Belts, Power steering fluid - Tire inflation and condition
Every 5,000 - 10,000 km	- Brake inspection - Tire rotation and wheel balancing - Automatic transmission fluid - Battery and cables
Every 10,000 - 20,000 km	- Brake, chassis lubrication - Coolant (Antifreeze), Wiper blades - Steering and suspension - Power steering fluid
Every greater than 20,000 km	- Replace fuel filter - Cooling system flush and refill - Cabin air filter, Fuel tank air filter

Table 2-4. Result of Distribution by The Gravity Model

Property Tax	Number of mechanic shops	Percentage	Number of OBDs
$30000 Over	11	9.0%	4
$15,000 - $30,000	27	23.0%	3
$10,000 - $15,000	38	33.0%	2
$0 - $10,000	40	34.0%	1
Total	116	100.0%	241

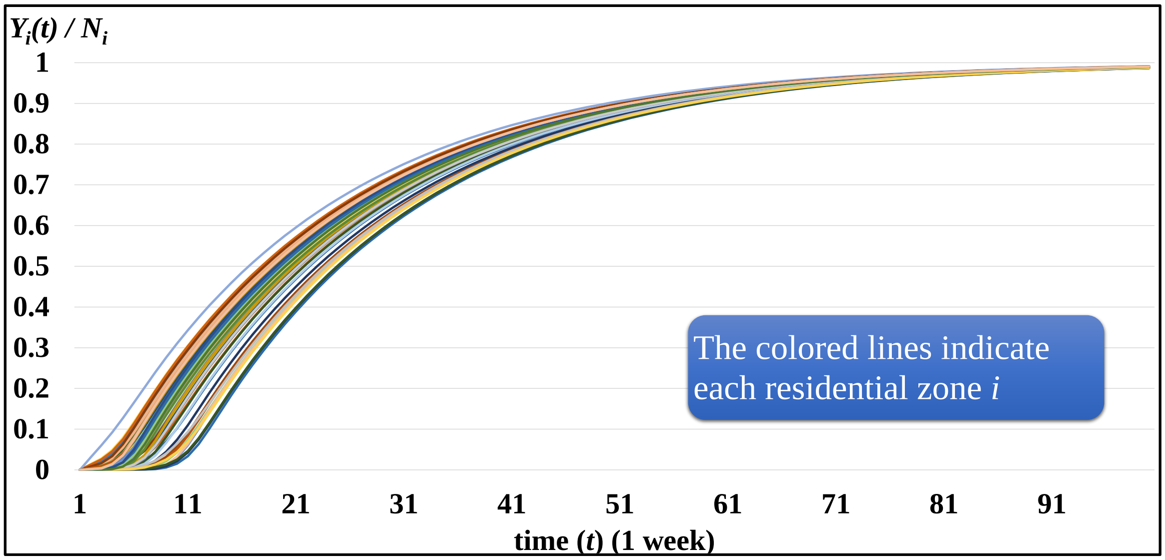
2.6 Results

2.6.1 The Speed of Infectious Malware Spreading

We analyzed the pattern and speed of the AV epidemic spread in detail with a combination of different variables. The first analysis is the base condition with a constant probability of transmitting the malware from infected OBD scanners (or AVs) to susceptible AVs (or OBD scanners) (α = 0.7 per week).

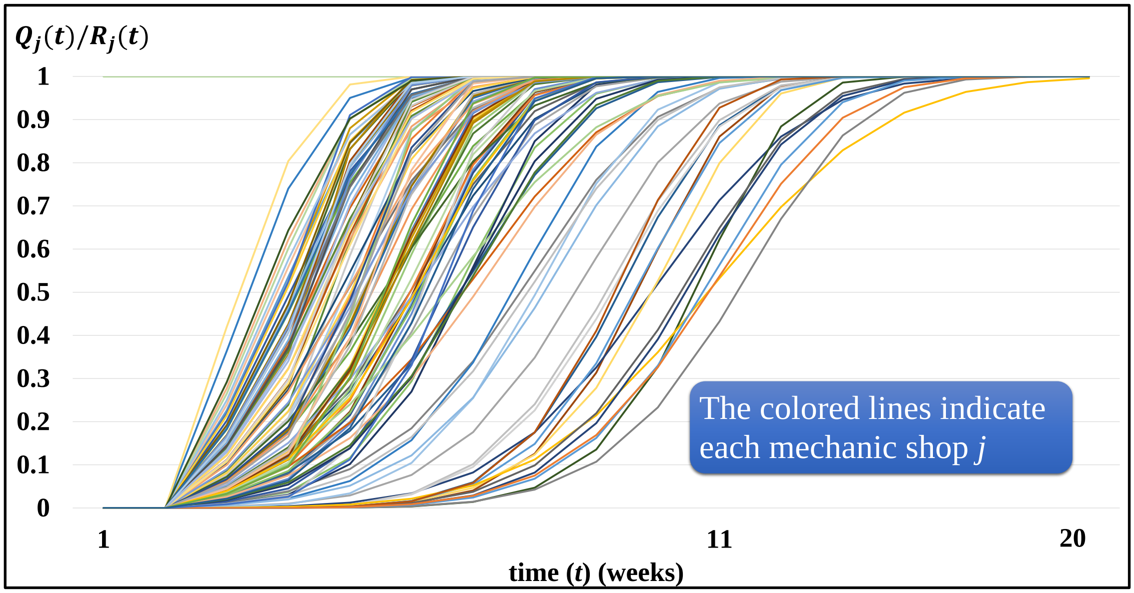
In this base condition, we can analyze the speed of infection with the proportion of infected vehicles ($\frac{Y_i(t)}{N_i}$) and OBD scanners ($\frac{Q_j(t)}{R_j}$) over time. The proportion of infected vehicles in zone i, $X_i(t)$, and OBD scanners in mechanic shop j, $Q_{\mathrm{j}}(t)$, over time (week) are shown in Fig. 2-7 (a) and (b), respectively. Each coloured line indicates residential zones or mechanic shops. The transmission probability of malware, α, has a constant value of 0.7, and the infected vehicles showed an S-shape curved increasing pattern. Considering that the infection started at a single mechanic shop, it only took 21 weeks to infect about half of the vehicles in Windsor.

The infection rate of OBD scanners is much faster than that of vehicles, and it only took 20 weeks to transmit the malware to all OBD scanners. This is because each OBD scanner connects to anonymous vehicles several times daily. This routine usage pattern can result in vulnerability to exposure to deliberate attacks that could spur the soaring infection speed. Also, fewer OBD scanners than the registered vehicles are another reason for the different infection speeds.



The colored lines indicate each residential zone *i*

(a) Proportion of infected vehicles by zone *i*



The colored lines indicate each mechanic shop *j*

(b) Proportion of infected OBDs by mechanic shop *j*

Fig. 2-7. Infection rate of vehicles and scanners with α = 0.7

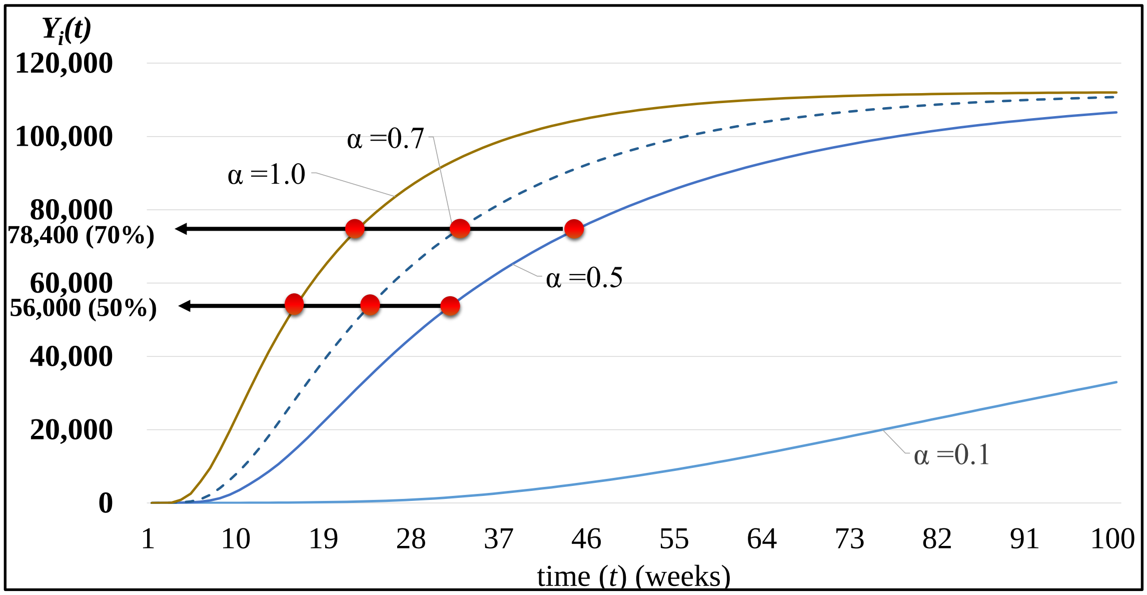
2.6.2 The Impact of Variable Probability of Malware Transmission (α)

We considered the variable probability of transmission (α) of the malware from 10 % to 100 %. The transmission speed could lag behind due to the regular security update. Thus, through the changing probability, we can understand the fluctuation of the speed at which malware spreads.

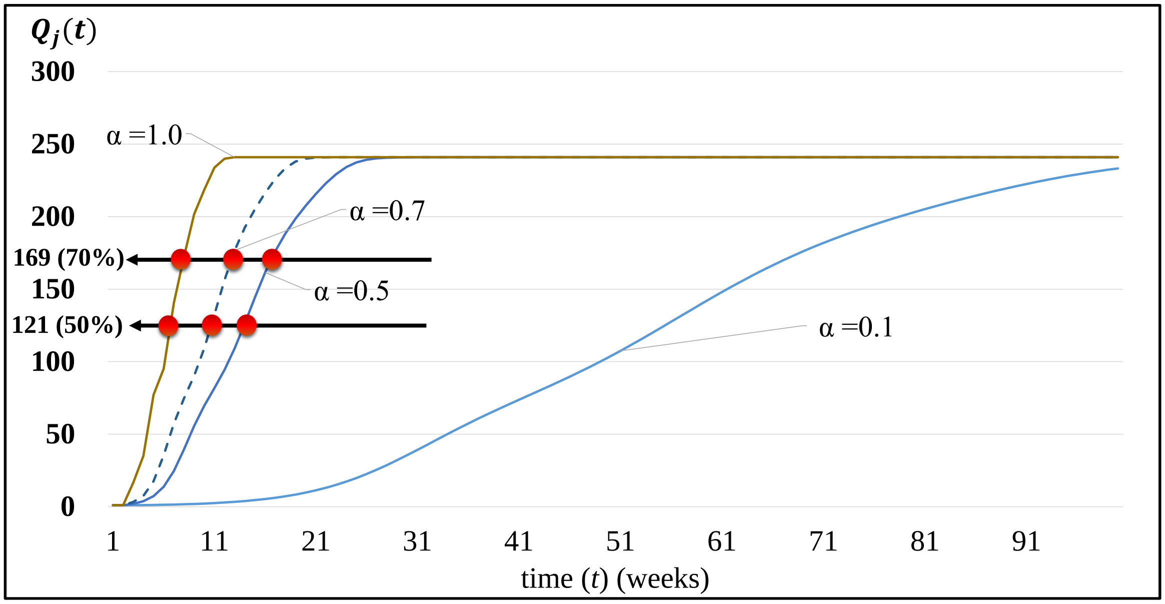
We compared the speed of the transmission with various probabilities of α (from 0.1 to 1). In particular, we focused on situations when the proportion of infected vehicles and OBD scanners reached 50 % and 70 % by α. The impact of different α is illustrated in Fig. 2-8 and Table 2-5. The infection speed with α = 1.0, which we shall call complete transmissibility, increases rapidly. It took only 17 weeks to infect half of the city's vehicles, which is about 56 thousand vehicles, and it also only needed 24 weeks to spread the malware to 70 % of vehicles. For both cases, the speed was about 8 weeks and 11 weeks faster than the speed of α = 0.7, respectively. The speed of spreading malware among OBD scanners is more noticeable than in AVs. With the smaller number of OBD scanners than vehicles, the speed of infection showed an abrupt increase. In the case of α = 1.0, only 7 or 8 weeks were needed to exceed 50 % or 70 % of infection, showing about 4 or 5 weeks faster than base case α = 0.7. Even though the transmissibility (α) was reduced to 0.5, less than 17 weeks were needed to reach the percentage. Hence, cyber security experts need to prepare for the tremendous speed of spreading malware.

Table 2-5. Comparison of Infection Pace by Alpha

Classification			α = 0.1	α = 0.5	**α = 0.7**	α = 1.0
AVs	$\frac{Y_i(t)}{N_i} > 50$ %	When	After 100 weeks	After 33 weeks	**After 25 weeks**	After 17 weeks
		Compare	Inestimable	8 weeks slower	**Criteria**	8 weeks faster
	$\frac{Y_i(t)}{N_i} > 70$ %	When	After 100 weeks	After 48 weeks	**After 35 weeks**	After 24 weeks
		Compare	Inestimable	13 weeks slower	**Criteria**	11 weeks faster
OBDs	$\frac{Q_j(t)}{R_j} > 50$ %	When	After 55 weeks	After 14 weeks	**After 11 weeks**	After 7 weeks
		Compare	43 weeks slower	3 weeks slower	**Criteria**	4 weeks faster
	$\frac{Q_j(t)}{R_j} > 70$ %	When	After 67 weeks	After 17 weeks	**After 13 weeks**	After 8 weeks
		Compare	54 weeks slower	4 weeks slower	**Criteria**	5 weeks faster



(a) Number of infected vehicles



(b) Number of infected OBD scanners

Fig. 2-8. Number of infected vehicles and scanners by alpha α = 0.7

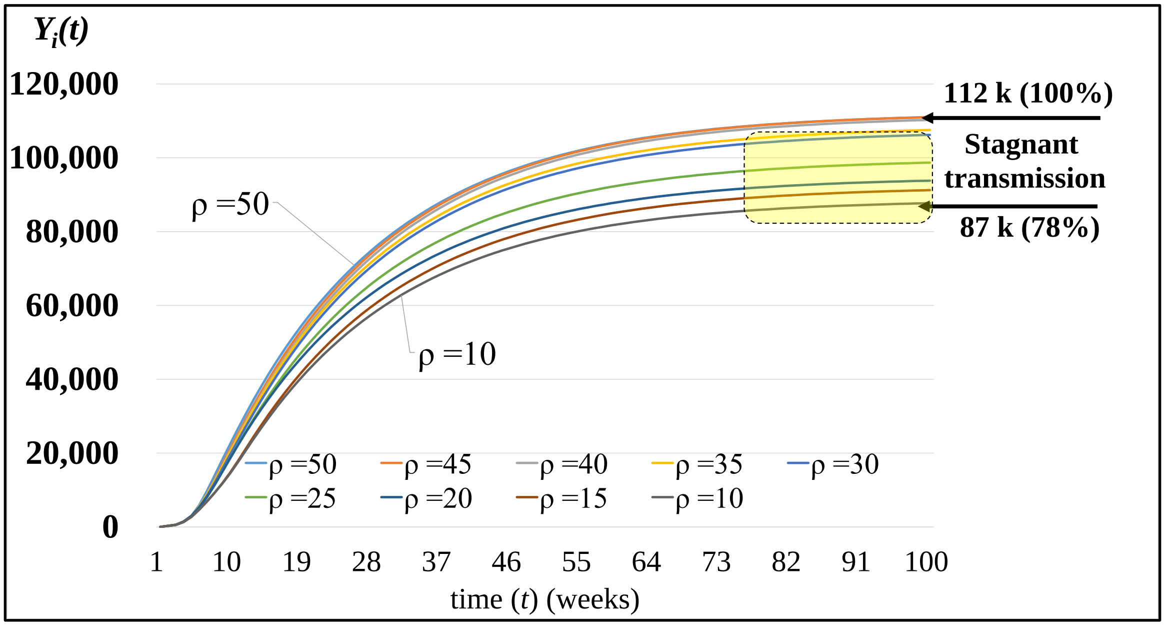
2.6.3 The Impact of Different Connectivity

This subsection analyzes the impact of rank-based visiting mechanic shops for each residential zone. The degree of connectivity (ρ) accounts for the rank-based number of mechanic shops to visit per residential zone.

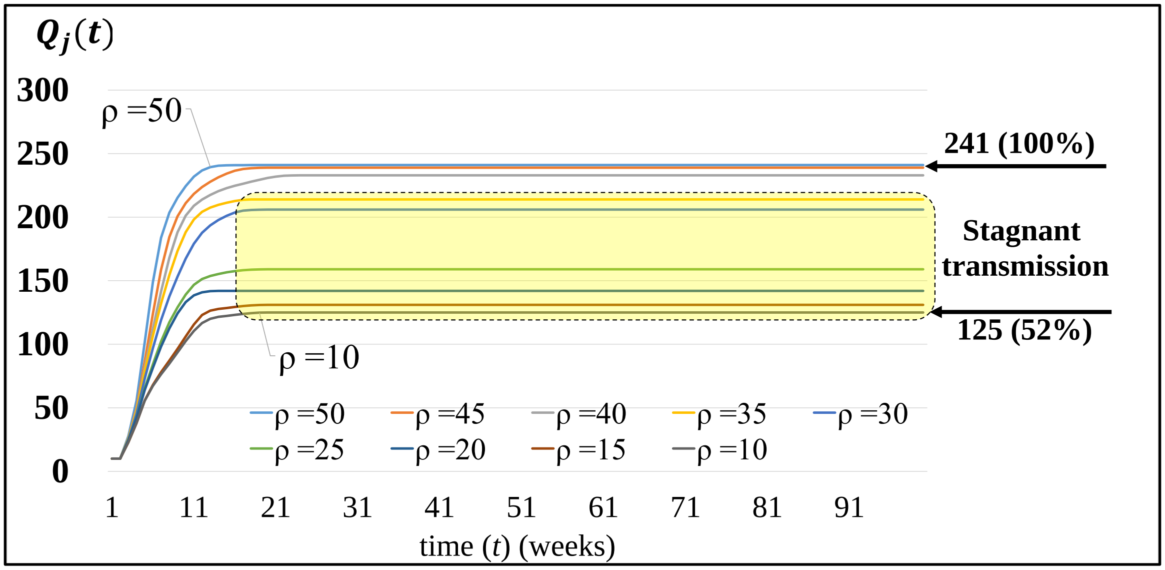
From subsections (A) and (B), we considered the propagation rate with the entire number of mechanic shops (116 shops) under the gravity model. In the general gravity model, the visiting patterns could be fully mixed in proportion to the mechanic shops' magnitude or reduced inversely proportional to the distance. This method could be appropriate to figure out the overall trends of malware transmission, assuming that susceptible and infected AVs and OBD scanners will be connected to each other in a fully mixed circumstance across the city. However, the fully mixed circumstance is not realistic because people are prone to visit their favourite shops. The people in the same zone would have a similar propensity to visit the shops because they could recommend specific shops considering the distance from the zone and service rating. Due to people's propensity, the malware would be hindered from propagating its malicious codes rapidly.

In subsection (C), we intended to capture the propensity by changing the number of visiting mechanic shops. The top rankings per residential zone were estimated using the gravity model from subsections (A) and (B). More specifically, if we consider only the top 10 mechanic shops for each zone, the AVs will visit the top 10 shops in accordance with the ranking of their residential zones. This new condition for choosing the number of shops could create a more realistic propagation model. In this dissertation, we shall call the selected number of top shops a degree of connectivity (ρ). We simulated the propagation speed with the different connectivity degrees (ρ) from 10 to 50 shops.

Fig. 2-9 and Table 2-6 show the malware transmission speed by the connectivity degree (ρ). As highlighted in Fig. 2-9 (a), some graphs having a low connectivity degree (ρ) of less than 35 showed that the proportion of infected vehicles did not reach 100 % even near 100 weeks. The increasing speed of infected vehicles shows stagnant transmission after 73 weeks; the stagnant transmission of OBD scanners starts at 21 weeks and has a more considerable duration than vehicles. As the connectivity degree decreases, the infection speed will be more stagnant.



(a) Number of infected vehicles



(b) Number of infected OBD scanners

Fig. 2-9. Number of infected vehicles and scanners by connectivity (*ρ*)

Table 2-6. Comparison of Infection Pace by Connectivity

Classification			**α = 0.7** **(Base case)**	ρ = 10	ρ = 50
AVs	$\frac{Y_i(t)}{N_i} > 50\ \%$	When	**After 25 weeks**	After 28 weeks	After 20 weeks
		Compare	**Criteria**	3 weeks slower	5 weeks faster
	$\frac{Y_i(t)}{N_i} > 70\ \%$	When	**After 35 weeks**	After 51 weeks	After 31 weeks
		Compare	**Criteria**	16 weeks slower	4 weeks faster
OBDs	$\frac{Q_j(t)}{R_j} > 50\ \%$	When	**After 11 weeks**	After 14 weeks	After 6 weeks
		Compare	**Criteria**	3 weeks slower	5 weeks faster
	$\frac{Q_j(t)}{R_j} > 70\ \%$	When	**After 13 weeks**	Stagnant	After 7 weeks
		Compare	**Criteria**	Inestimable	6 weeks faster

Table 2-6 compares the infection speed of the base condition (α = 0.7) with the various connectivity (ρ). As we discussed, the lower connectivity had a slower infection speed. With the low connectivity (ρ =10), it was not able to reach the 70 % of infection proportion ($\frac{Q_j(t)}{R_j}$) of OBD scanners. This finding indicated that some OBD scanners could avoid malware threats even for a long time without infection. Notably, the low connectivity degree could bring about less contact between susceptible and infected vehicles and OBD scanners.

2.7 Conclusion

Today's AVs are susceptible to the same cybersecurity risks as networked computers. In particular, when AVs are connected with infected OBD scanners for a regular check-up or repair in the mechanic shops, the risk of malware transmission could soar. Stuxnet is a vector-host malware, similar to the malaria model, which is an exemplary vector-host epidemic. This study adopted the methodologies of epidemiology to understand and address the AV epidemic. We developed the mathematical model to embody malware transmission in various conditions: analysis of a fixed probability of malware transmission, with variable probability, and consideration of vehicle owners' visiting mechanic shop patterns. A primary contribution of this study could enhance understanding regarding the novel malware propagation and help make alternative plans to protect AVs.

While this model describes the dynamics of AV epidemics, it is worth noting that the proposed model can also help design optimal control strategies. The use of the transportation gravity model in our study revealed trends in malware propagation both within and between regions. In practical terms, if cybersecurity experts can identify the initial sites of malware infections, they could then estimate the potential speed of its spread. This allows for inferring the initial areas of attack and implementing quarantine measures for vehicles travelling from infected areas to mechanic shops in other regions. Consequently, our proposed models can be effectively integrated with network science and control theory methodologies. This integration will enable cybersecurity experts to develop efficient strategies, significantly reducing the spread of malware infections among vehicles.

This dissertation offers the possibility of developing more sophisticated AV epidemic models by leveraging well-developed mathematics theories in epidemiology and ecology. Although this study is the first step in understanding the novel AV malware, it has several limitations. We focused on malware propagation through physical contact between OBD scanners and vehicles. We assumed that OBD scanners would be a universal device for checking vehicles in the future and that studying physical contact-styled attacks could be rewarding for AV security.

Also, we did not consider the effect of different AV brands, models and malware variants. The differences in brands, models and variants could impact the speed of malware propagation. We assumed that Stuxnet malware could transmit any brand and model. Also, hackers would not need to create variants to enhance the malware’s ability to lurk.

On top of that, we did not compare our results with other models. To the best of our knowledge, this dissertation is the first attempt at applying the malaria propagation model to malware propagation through OBD scanners as a vector. Even though some studies considered the malware propagation models in computers or smartphones based on the epidemic, the unique circumstances of AVs were not considered. Accordingly, it was not able to compare the speed with different epidemic-based models. Also, if the real-world mechanic shop visiting data had been included, the results would have been more persuasive. For instance, the frequency of visiting mechanic shops and the types of maintenance and repair would be impacted by the age of vehicles and vehicle owners’ decision-making process. Considering the life cycle of vehicles and individual owner’s preferences would be a significant topic in follow-up studies. It characterizes the topology of complex connectivity and develops models to mimic a network's growth and reproduce its structural properties. Consequently, we hope that further analysis considering malware transmission inter cities or provinces will confirm our findings: the risk of Stuxnet. Lastly, research regarding individual and specific automotive requirements for Stuxnet-style attacks can be a crucial and potential topic to prepare for malware attacks.

2.8 Chapter 2 References

[1] L. H. Newman, “A ransomware attack has struck a major US hospital chain,” *WIRED*, San Francisco, Sep. 2020. [Online]. Available: https://www.wired.com/story/universal-health-services-ransomware-attack/

[2] S. Singh, C. Estan, G. Varghese, and S. Savage, “Automated worm fingerprinting,” in *6th Symposium on Operating Systems Design and Implementation*, 2004, pp. 45–60. [Online]. Available: https://www.usenix.org/legacy/events/osdi04/tech/full_papers/singh/singh.pdf

[3] M. Koczerginski, “The Cybersecurity Implications of Driverless Cars,” *McMillan LLP*, 2016. https://mcmillan.ca/The-Cybersecurity-Implications-of-Driverless-Cars

[4] McAfee, “2017 Threats Predictions,” Santa Clara, 2016. [Online]. Available: https://www.mcafee.com/enterprise/en-us/assets/reports/rp-threats-predictions-2017.pdf

[5] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, “Connected Vehicles: Solutions and Challenges,” *IEEE Internet Things J.*, vol. 1, no. 4, pp. 289–299, Aug. 2014, doi: 10.1109/JIOT.2014.2327587.

[6] K. Koscher et al., “Experimental Security Analysis of a Modern Automobile,” in *2010 IEEE Symposium on Security and Privacy*, 2010, pp. 447–462. doi: 10.1109/SP.2010.34.

[7] S. Checkoway et al., “Comprehensive experimental analyses of automotive attack surfaces,” in *20th USENIX Security Symposium*, 2011, pp. 77–92. [Online]. Available: https://www.usenix.org/conference/usenix-security-11/comprehensive-experimental-analyses-automotive-attack-surfaces

[8] E. Fernandes, B. Crispo, and M. Conti, “FM 99.9, Radio Virus: Exploiting FM Radio Broadcasts for Malware Deployment,” *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 6, pp. 1027–1037, Jun. 2013, doi: 10.1109/TIFS.2013.2259818.

[9] J. Petit and S. E. Shladover, “Potential Cyberattacks on Automated Vehicles,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 1–11, 2014, doi: 10.1109/TITS.2014.2342271.

[10] J. P. Farwell and R. Rohozinski, “Stuxnet and the Future of Cyber War,” *Survival (Lond).*, vol. 53, no. 1, pp. 23–40, Feb. 2011, doi: 10.1080/00396338.2011.555586.

[11] D. Kushner, “The real story of Stuxnet,” *IEEE Spectr.*, vol. 50, no. 3, pp. 48–53, Mar. 2013, doi: 10.1109/MSPEC.2013.6471059.

[12] J. Hsu, “Stuxnet-Style Virus Failed to Infiltrate North Korea’s Nuclear Program,” *IEEE Spectrum*, New York, Jun. 2015. [Online]. Available: https://spectrum.ieee.org/nsa-stuxnetstyle-virus-failed-to-infiltrate-north-koreas-nuclear-program

[13] T. M. Chen and S. Abu-Nimeh, “Lessons from Stuxnet,” *Computer (Long. Beach. Calif).*, vol. 44, no. 4, pp. 91–93, Apr. 2011, doi: 10.1109/MC.2011.115.

[14] S. Karnouskos, “Stuxnet worm impact on industrial cyber-physical system security,” in *IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society*, 2011, pp. 4490–4494. doi: 10.1109/IECON.2011.6120048.

[15] A. Szijj, L. Buttyán, and Z. Szalay, “Hacking cars in the style of Stuxnet,” 2015. [Online]. Available: http://www.hit.bme.hu/~buttyan/publications/carhacking-Hacktivity-2015.pdf

[16] S. Malik and W. Sun, “Analysis and Simulation of Cyber Attacks Against Connected and Autonomous Vehicles,” in *2020 International Conference on Connected and Autonomous Driving (MetroCAD)*, Feb. 2020, pp. 62–70. doi: 10.1109/MetroCAD48866.2020.00018.

[17] N. Falliere, L. O. Murchu, and E. Chien, “W32. Stuxnet Dossier,” Cupertino, 2011. [Online]. Available: https://www.wired.com/images_blogs/threatlevel/2011/02/Symantec-Stuxnet-Update-Feb-2011.pdf

[18] D. Mollison, “Possible velocities for a simple epidemic,” *Adv. Appl. Probab.*, vol. 4, no. 2, pp. 233–257, Aug. 1972, doi: 10.2307/1425997.

[19] A. Hastings, “Models of spatial spread: A synthesis,” *Biol. Conserv.*, vol. 78, no. 1–2, pp. 143–148, 1996, doi: 10.1016/0006-3207(96)00023-7.

[20] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*, 1st ed. New Jersey: Princeton University Press, 2008.

[21] R. N. Charette, “This Car Runs on Code,” *IEEE Spectrum*, 2010. https://spectrum.ieee.org/this-car-runs-on-code

[22] S. Corrigan, “Introduction to the controller area network (CAN),” 2002. [Online]. Available: https://www.rpi.edu/dept/ecse/mps/sloa101.pdf

[23] T. Zhang, H. Antunes, and S. Aggarwal, “Defending Connected Vehicles Against Malware: Challenges and a Solution Framework,” *IEEE Internet Things J.*, vol. 1, no. 1, pp. 10–21, Feb. 2014, doi: 10.1109/JIOT.2014.2302386.

[24] S. Parkinson, P. Ward, K. Wilson, and J. Miller, “Cyber Threats Facing Autonomous and Connected Vehicles: Future Challenges,” *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 2898–2915, 2017, doi: 10.1109/TITS.2017.2665968.

[25] K. Kim, J. S. Kim, S. Jeong, J.-H. Park, and H. K. Kim, “Cybersecurity for autonomous vehicles: Review of attacks and defense,” *Comput. Secur.*, vol. 103, p. 102150, Apr. 2021, doi: 10.1016/j.cose.2020.102150.

[26] I. Studnia, V. Nicomette, E. Alata, Y. Deswarte, M. Kaaniche, and Y. Laarouchi, “Survey on security threats and protection mechanisms in embedded automotive networks,” in *2013 43rd Annual IEEE/IFIP Conference on Dependable Systems and Networks Workshop (DSN-W)*, Jun. 2013, pp. 1–12. doi: 10.1109/DSNW.2013.6615528.

[27] J. Hong, “Cyber security issues in connected vehicle of intelligent transport system,” *Indian J. Sci. Technol.*, vol. 9, no. 24, pp. 1–7, 2016, doi: 10.17485/ijst/2016/v9i24/96027.

[28] M. Amoozadeh et al., “Security vulnerabilities of connected vehicle streams and their impact on cooperative driving,” *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 126–132, 2015, doi: 10.1109/MCOM.2015.7120028.

[29] Upstream Security, “2021 Global Automotive Cybersecurity report,” 2021. [Online]. Available: https://upstream.auto/2021report/

[30] A. M. del Rey, “Mathematical modeling of the propagation of malware: a review,” *Secur. Commun. Networks*, vol. 8, no. 15, pp. 2561–2579, Oct. 2015, doi: 10.1002/sec.1186.

[31] M. Meisel, V. Pappas, and L. Zhang, “A taxonomy of biologically inspired research in computer networking,” *Comput. Networks*, vol. 54, no. 6, pp. 901–916, Apr. 2010, doi: 10.1016/j.comnet.2009.08.022.

[32] J. O. Kephart and S. R. White, “Directed-graph epidemiological models of computer viruses,” in *Proceedings. 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, 1991, pp. 343–359. doi: 10.1109/RISP.1991.130801.

[33] C. C. Zou, W. Gong, and D. Towsley, “Code red worm propagation modeling and analysis,” in *Proceedings of the 9th ACM conference on Computer and communications security - CCS ’02*, 2002, p. 138. doi: 10.1145/586110.586130.

[34] S. Peng, S. Yu, and A. Yang, “Smartphone Malware and Its Propagation Modeling: A Survey,” *IEEE Commun. Surv. Tutorials*, vol. 16, no. 2, pp. 925–941, 2014, doi: 10.1109/SURV.2013.070813.00214.

[35] O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas, “Worm Epidemics in Vehicular Networks,” *IEEE Trans. Mob. Comput.*, vol. 14, no. 10, pp. 2173–2187, Oct. 2015, doi: 10.1109/TMC.2014.2375822.

[36] D. Teshler, “Entering the Ethernet era: The difference between CAN and Ethernet,” *GUARDKNOX*, 2021. https://blog.guardknox.com/entering-the-ethernet-era-the-difference-between-can-and-ethernet.

[37] “Frequently asked questions about malaria,” *Centers for Disease Control and Prevention (CDC)*. https://www.cdc.gov/malaria/about/faqs.html#:~:text=Is+malaria+a+contagious+disease,to+someone+who+has+malaria.

[38] Y. H. Kim, S. Peeta, and X. He, “Modeling the information flow propagation wave under vehicle-to-vehicle communications,” *Transp. Res. Part C Emerg. Technol.*, vol. 85, pp. 377–395, 2017, doi: 10.1016/j.trc.2017.09.023.

[39] Y. H. Kim, S. Peeta, and X. He, “An Analytical Model to Characterize the Spatiotemporal Propagation of Information Under Vehicle-to-Vehicle Communications,” *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 3–12, Jan. 2018, doi: 10.1109/TITS.2015.2512523.

[40] A. Wilson, *Entropy in urban and regional modelling*. London: Pion, 2021. doi: 10.2307/142897.

[41] I. M. Longini, “A mathematical model for predicting the geographic spread of new infectious agents,” *Math. Comput. Model.*, vol. 12, no. 9, pp. 1179–1180, 1989, doi: 10.1016/0895-7177(89)90250-1.

[42] Statistics_Canada, “Table 23-10-0145-01 Canadian vehicle survey, number of vehicles on the registration lists, by type of vehicle, province and territory, annual,” *Statistics Canada*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2310014501&pickMembers%5B0%5D=1.7.

[43] Statistics_Canada, “Table 23-10-0148-01 Canadian vehicle survey, vehicle-kilometres, by type of vehicle, province and territory, annual (x 1,000,000),” *Statistics Canada*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2310014801&pickMembers%5B0%5D=1.7.

[44] “Car care guide,” *Car Care Council*. https://www.carcare.org/car-care-resources/car-care-guide/.

[45] “Public Property Inquiry,” *City of Windsor*. https://publicpropertyinquiry.citywindsor.ca/.

CHAPTER 3

QUANTIFYING PRIVACY RISKS IN PANDEMIC CONTACT TRACING THROUGH SMART CARD DATA

3.1 Introduction

Our daily lives are continually monitored and recorded through various means, including CCTV, credit cards, and mobility data from taxi and transit smart cards. Due to rapid advancements in data handling and storage technology, the collected data – which covers almost the entire population – can be stored in a database with numerous sensitive personal data attributes [1]. The collected data attributes can be categorized into *explicit identifiers* and *quasi-identifiers (QIDs)* [2]. The explicit identifiers, such as social security numbers, addresses, and names, can directly identify the users, as there may be only one unique person with identical information to these identifiers. Accordingly, the explicit identifiers have been anonymized thoroughly for privacy protection. Meanwhile, using QIDs, like zip code, birth date, gender, and mobility data, to identify an individual does not pose the same privacy risk levels as explicit identifiers. This is because a number of individuals may share similar QID information, limiting the potential for unique identification based solely on QIDs. However, unique and publicly known individuals, such as politicians or celebrities can be re-identified using their several QIDs and public information [3].

Sweeney demonstrated such re-identification risks with an example that involved identifying the medical diagnosis and medications of a governor of Massachusetts. The study linked QIDs – ZIP code, birth date, and gender – derived from voter registration lists and health insurance databases [4]. In that case, Sweeney already had access to the governor’s 5-digit ZIP code, gender, and birth date, which are publicly available information. Sweeney filtered out irrelevant IDs in the two databases by comparing the QIDs: ZIP code, birth date, and gender with the governor’s public information. Ultimately, Sweeney identified one person with identical information to the governor’s and revealed the governor’s sensitive medical information. In Sweeny's example, the publisher of the two databases might have deemed the databases safe to publish since they did not contain explicit identifiers. The researchers, however, identified a unique ID and found the governor's medical sensitive information. Therefore, matching some QIDs from different sources with public information can significantly increase re-identification risks.

After Sweeney’s research, many studies have demonstrated that re-identification attacks are possible using several QIDs in databases and public information because of the uniqueness of individual travel records [5-7]. Re-identification risks can also be found in smart card data. The QIDs in smart card data (hereafter referred to as card QIDs) consist of travel records, including boarding/alighting times and stations, fares, and approximate age information such as student, adult, and senior card classification. The card QIDs can be exploited to re-identify users’ card IDs when combined with external observations by adversaries.

Such public sharing of external observations consequently raised significant privacy concerns, especially during the COVID-19 pandemic. Several countries thoroughly traced COVID-19 patients’ travel records and released them publicly after excluding explicit identifiers, such as names, to encourage people to avoid the locations visited by the patients [8-10]. However, these travel records can be considered as external observation data, provoking privacy risks to patients. Adversaries could extract unique attributes such as a patient's smart card ID. Once adversaries identify the card ID, they can trace the user’s every itinerary in the card database, resulting in significant privacy risks. Therefore, understanding the re-identification risks of COVID-19 patients stemming from the linkage between the card QIDs and travel observations is pivotal for protecting individual privacy in potential future pandemics.

However, unexplored aspects exist regarding quantifying re-identification attacks using card QIDs and travel observations. Many researchers have investigated security and privacy issues stemming from government policies, particularly contact tracing [11-14]. These studies have reviewed potential privacy risks from the policies and employed ordinal scales or rubrics as categorized analysis approaches to measure the relative extent of privacy risks. Despite previous researchers’ efforts, quantitatively assessing the severity of these risks remains unaddressed.

Also, existing studies have not delved deeply into the accuracy of language-based external observations. COVID-19 travel observations, primarily derived from patients' verbal reports, may contain discrepancies, thereby affecting re-identification risks. The issue of data accuracy extends beyond COVID-19 travel observations to include social media posts about visits to popular locations. These posts are also subject to potential inaccuracies inherent in language-based observations. For example, the actual date of a visit might not always correspond to the date of the post; there can be a discrepancy of a few days. Furthermore, posts that do not specify an exact location name may only offer a rough estimate of the actual location. Consequently, when analyzing re-identification risks, it is essential to take into account these unique characteristics of language-based travel observations.

Lastly, other researchers have tended to focus mainly on the number of observations rather than the diversity [15-17]. Even though the number of observations may be identical, the combinations of visited locations can vary. For example, some people might visit only one place with four observations, while others may visit four different places. Re-identification risks can fluctuate significantly according to the combination of diverse visited locations. Therefore, exploring the impact of diverse visited locations on re-identification risks can enhance our understanding of real-world privacy breach case studies.

Given the above knowledge gaps, our study aims to address three primary objectives. First, we seek to quantify the privacy risks in smart card data associated with patient travel observations induced by contact-tracing policies. Second, we aim to elucidate the impact of language-based travel observations on the extent of privacy risk. Given their inherent variability in spatial and temporal accuracy, we contend that language-based travel observations can significantly affect the risk spectrum. Lastly, we explore the varied combinations of locations in patients' itineraries to quantify the re-identification risks stemming from these diversities.

This study provides three primary contributions. First, we uniquely quantify patients' re-identification risks within a smart card database, leveraging real-world privacy breach cases from language-based travel observations. Second, we elucidate the influence of spatial and temporal accuracy variations in language-based travel observations on re-identification risks. Finally, we explore the fluctuations in the risks associated with the diversity of visited locations.

The remainder of this chapter is structured as follows. Section 2 reviews the existing research regarding COVID-19 privacy breaches and re-identification risks. Section 3 describes smart card and COVID-19 patient travel record data and presents a preliminary analysis. The methods of quantifying re-identification risks under diverse circumstances are presented in Section 4. Sections 5 and 6 present the results and conclusion.

3.2 Related Work

This section provides a comprehensive overview of various studies on COVID-19 privacy and re-identification risks, aiming to bridge the knowledge gap. We particularly focused on existing methodologies aimed at quantifying the risks and studied how QIDs and external observations' characteristics affected the extent of the risks.

The COVID-19 contact tracing measures have provoked significant interest and concerns about privacy risks. Jung et al. assessed relative privacy risk levels using an ordinal scale in five categories: demographics, significant places (homes and workplaces), sensitive information (hobbies and religions), and routine behaviour, aiming to compare the risks across different cities [11]. They quantified the privacy risk levels on a scale from 0 to 3 based on contact tracing data. Similarly, Krehling and Essex applied consensus principles to compare the relative privacy risks of 55 digital contact tracing apps worldwide [14]. Lastly, Ahn et al. proposed the COVID index, which balances personal privacy and public health safety by adjusting policy magnitude. They proposed that governments should modulate the intensity of contact tracing based on the severity of virus outbreaks [12]. Many researchers have made substantial efforts to assess the relative risk levels using categorized measures. However, the extent of re-identification risks remains unresolved. To the best of our knowledge, research is lacking, considering quantifying the risks using real-world privacy breach data.

Additionally, we explored other privacy studies that quantified the risks using external observations and other open data sources: mobile phone usage data, credit card transactions, and smart card data. De Montjoye et al. quantified the re-identification risks by filtering out irrelevant data based on the observed points and mobile phone user datasets or credit card data [15-16]. They randomly chose the observed points from datasets and concluded that about 90% of individuals could be re-identified with four observations. They also highlighted the importance of spatial and temporal resolutions of observations for privacy risk levels. In general, trajectory data resolution signifies the degree of data precision in terms of spatial and temporal attributes. In our study, resolution can be associated with the accuracy levels of the spatial and temporal observation data. De Montjoye found that reducing data resolution decreased the likelihood of re-identification risk. Conversely, they asserted that increasing the number of exposed locations could increase the risk.

In transportation mobility data, Gao et al. quantified the risks with a license plate recognition (LRP) data set collected from 516 stationary detectors [17]. They randomly changed temporal resolutions of external observations and the traffic volume in the surveyed locations of the original LRP data set to quantify the risks. They concluded that five external observations can identify 90% of individuals. Additionally, Li et al.

reviewed the risks in transit smart card data. They asserted that if adversaries know two observations of a passenger, including departure and arrival station names and bus transfer information, about 30% of the transit users can be re-identified [18].

Existing studies quantified the levels of re-identification risks from open data sources based on external observations. These studies, however, manipulated the resolutions of observations within controlled experimental circumstances. Their findings could have been more applicable if they had considered real-world privacy breach cases. COVID-19 patient travel observation data has language-based varying data resolutions. Patients reported their travel records mainly based on their memories because the data was collected verbally and not recorded automatically during their visits. Human memory is unreliable; false memories can often be indistinguishable from true ones [19-20]. Thus, travel observations based on such memories could entail errors, and the levels of resolution in terms of visited locations and times could vary. With language-based travel observations, researchers do not need to manipulate the resolutions for analysis. This aspect is crucial for a more realistic quantification of privacy risks.

On the other hand, in the introduction, we emphasized the significance of diverse combinations of visited locations. While existing studies have provided invaluable insights into the context of privacy risks, to the best of our knowledge, they have not fully explored the role of diverse combinations of exposed locations. We believe the quantified risks could be overestimated or underestimated without considering the visited location diversity. Therefore, we trace travel observations consecutively, and this approach allowed us to consider location diversity and understand the associated risks more accurately.

3.3 Preliminary

3.3.1 Data Description

1) Smart Card Dataset

This study employed the Seoul smart card dataset, which consists of approximately 131.1 million public transit trip records per week from 16.6 million anonymized smart card IDs. Seoul has an extensive public transportation network encompassing 301 subway stations and 19,540 bus stops to cover an area of 605.2 km2. The smart card data attributes contain essential trip information such as modes and routes, origins and destinations, and boarding/alighting time. Additionally, the data provides transaction information, including anonymized card IDs, the number of transfers, and the fare [21]. To calculate fares proportional to distance and provide subsidies, the government encourages users to tag their cards while boarding and alighting [22]. Thus, smart card data contain individual users' detailed itineraries, which can exacerbate privacy risks when combined with other sensitive records like adversaries' observations or COVID-19 patient travel records.

2) COVID-19 Patient Travel-records

This study analyzed the travel records of COVID-19 patients in South Korea, specifically reviewing data released by the Seoul government from March to June 2020 [23]. We extracted public transit-related information from the language-based travel records, including travel mode, approximate or exact departure/arrival times, and visited places. These places are categorized into two groups: those with station information and those without. We referred to this extracted transit-related information as the external observations and linked it with smart card data to re-identify the patients' card IDs. As a result, this study identified 236 patients who utilized the subway, accounting for 927 travel observations.

From the travel records, we found that 72% of subway-using patients had a travel record of four or fewer trips before being diagnosed with COVID-19. Also, the trip purposes and the nearest stations to their homes and workplaces – considered sensitive personal information – were publicly disclosed without adequate protection. More specifically, regular trips, such as home-based work and school trips, constituted 61% of the records. Approximately 56% of the travel records revealed both the nearest subway stations to patients' homes and workplaces simultaneously. Robust protective measures should have been implemented to shield this highly sensitive data against privacy breaches before its public release.

3.3.2 Resolutions in Travel Observations and Re-identification

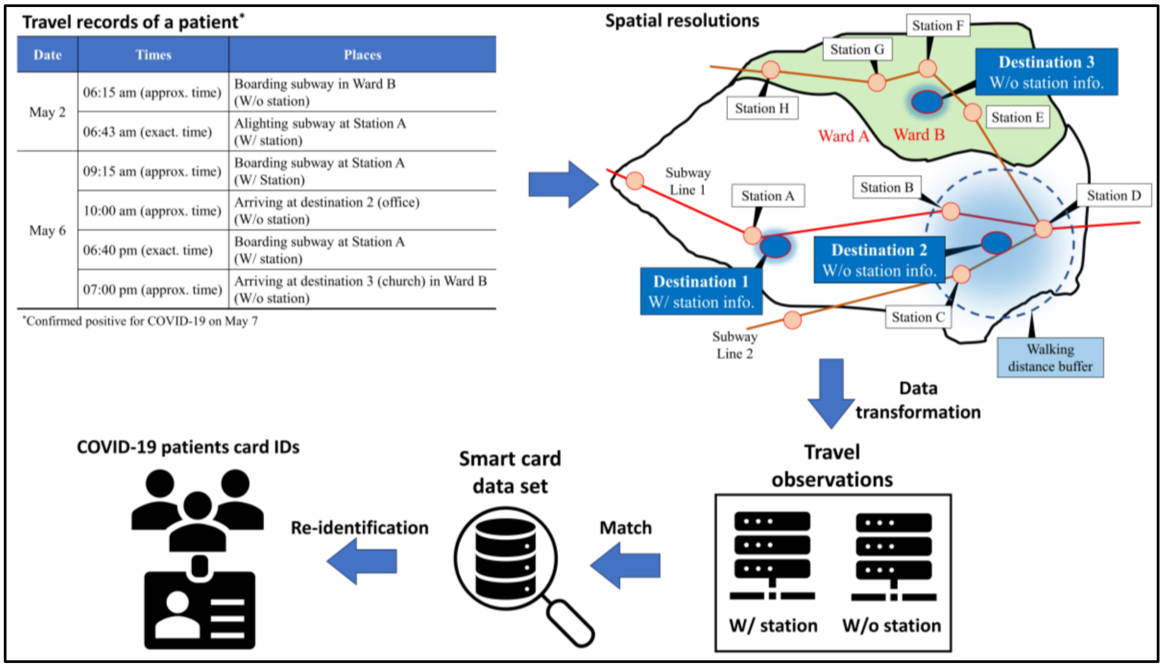
As discussed in the related work, language-based travel observations can vary spatial and temporal resolutions, significantly impacting the re-identification risks. In this regard, we have examined how the spatial and temporal resolutions were generated from the patient travel records.

As shown in Fig. 3-1, if a patient reports travelling to destination 1 via station A, an adversary could potentially re-identify the patient's card ID using only the smart card transaction records from station A. However, suppose the patient reports visiting destination 2 without providing station information. In that case, the adversary would need to consider all subway stations within the surrounding area of destination 2. This is because, without exact station information, adversaries would consider how far people walk from the final transit stations to reach their destinations. In our research, we assumed a walkable distance radius of about 500 meters from the final destinations, which are the circle's centers with the radius based on existing research [24]. Fig. 3-1 highlighted the walkable distance from destination 2 in shaded blue, containing three subway stations (B, C, and D).

Similarly, if a patient provided only administrative units like ward information, adversaries would need to broaden their search to all stations within that administrative boundary. As illustrated in Fig. 3-1, a patient who visited destination 3 via station G, might report being in ward B without specifying the station. In such cases, due to the lack of precise station details, adversaries would examine the transaction records of the stations from E to H within Ward B rather than focusing on G.

On top of that, people generally remember the time in an approximate manner, not pinpointing the exact moment. For instance, a patient who visited a place at 4:21 pm may report the time as 4:21 pm exactly, approximately 4:30 pm, or even roughly 4 pm. Given the potential errors resulting from people’s imperfect memory for the moment, temporal resolutions also can impact quantifying the re-identification risks.

In conclusion, the discrepancies in spatial and temporal resolutions of language-based travel observations can increase the difficulty of re-identifying COVID-19 patients’ smart card IDs. Thus, it is crucial to research re-identification risks across varying resolutions of travel observations to enhance our understanding of the related privacy concerns.



Travel records of a patient[*]

Date	Times	Places
May 2	06:15 am (approx. time)	Boarding subway in Ward B (W/o station)
	06:43 am (exact. time)	Alighting subway at Station A (W/ station)
May 6	09:15 am (approx. time)	Boarding subway at Station A (W/ Station)
	10:00 am (approx. time)	Arriving at destination 2 (office) (W/o station)
	06:40 pm (exact. time)	Boarding subway at Station A (W/ station)
	07:00 pm (approx. time)	Arriving at destination 3 (church) in Ward B (W/o station)

[*]Confirmed positive for COVID-19 on May 7

Fig. 3-1. Conceptual diagram of re-identification from travel records with spatial resolutions

3.4 Methods

Our approach aimed to quantify re-identification risks using real-world privacy breach cases. We focused on addressing the following research questions: 1) How does the number of travel observations impact the risks? 2) How do variations in the resolution of travel observations impact the risks? 3) How do the diverse visited locations impact the risks?

3.4.1 Impact of the Number of Travel Observations

To address the first research question, we quantified re-identification risks across varying numbers of exposed travel records, employing an approach similar to previous research [15-16]. We introduced an *anonymity value*, which represents the number of IDs that possess travel records identical to a patient's travel observations.

The notations for the anonymity value are as follows. U stands for the population of individual smart card ID databases, and E denotes the entire set of smart card transaction records. We focused on COVID-19 patients who took the subway. These patients are represented by the variable i, $(i = 1, 2, \dots, P)$, and their card IDs are included in the card databases, denoted by $(i \in U)$. The travel observations of the patient i are described with O^i, and these observations are included in the transaction records, denoted by $\left(O^i \in E\right)$. We traced the trajectories of the patients using travel observations, considering the number of observations. We denoted observation numbers by j (where $j = 1, 2, \dots, N$), with 1 indicating the most recent travel record. These observation numbers are represented in the subscript of the symbol (O_j^i).

Our approach then identified other transit users who have identical travel records to the patients' observations (O_j^i). We grouped these users into a subset $S\left(O_j^i\right)$; and counted the cardinality of the subset $S\left(O_j^i\right)$ to indicate the *anonymity value*, denoted by $\left|S\left(O_j^i\right)\right|$. Thus, the anonymity value can be an indicator of re-identification risks. The magnitude of these risks changes with the number of travel observations. Theoretically, $\left|S\left(O_j^i\right)\right|$ can be from 1, the patient him/herself, to U, the total IDs.

Lastly, we estimated the *uniqueness value*, denoted by ε, a concept introduced by de Montjoye in 2013 for quantifying the re-identified individuals given j exposed observations [15]. The uniqueness value aggregates the individual anonymity value and considers the number of travel observations, j; we express our uniqueness value as ε_j. The uniqueness value represents a proportion of re-identified patients compared to the total cardinality of COVID-19 patient subsets, P, according to j. The uniqueness value is calculated using the following equation and notations are summarized in Table 1.

$$\varepsilon_j = \sum_{i=1}^{P} \frac{\left|S(O_j^i)\right|}{P}, \quad (j = 1, 2, \ldots, N) \tag{1}$$

Table 3-1 Notation

Notation	Description
U	Population of smart card IDs
i	COVID-19 patient, a subset of U. $(i \in U)$
P	Total cardinality of COVID-19 patient subsets
E	Entire smart card transactions
O^i	Travel observations of patient i
j	Number of travel observations
O_j^i	Travel observation of patient i considering j
$S(O_j^i)$	Subset of individual card IDs from U having identical travel records to O_j^i
$\left\|S(O_j^i)\right\|$	Anonymity value of $S(O_j^i)$. (i.e., the cardinality of $S(O_j^i)$)
ε_j	Uniqueness value, according to the number of observations

3.4.2 Impact of Spatial and Temporal Resolutions

To address the 2nd research question, we divided the locations visited by patients into two categories based on the spatial resolution of observations. Locations with precise station names were classified as high-resolution. In contrast, the locations categorized as low-resolution included information about destinations and administrative units instead of exact station names. Subsequently, we attempted to re-identify smart card IDs whose travel records according to the high and low spatial resolution levels. This categorization allowed us to estimate anonymity and uniqueness values in accordance with the respective resolution levels.

As previously discussed, the imperfect memory and recognition of time among people create variations in temporal resolutions. To explore the impact of temporal resolutions on privacy risks, we classified the recorded visiting times into several categories based on the granularity of the data: exact time, within 30 minutes, within an hour, within three hours (morning, afternoon, evening), and a day. We then applied corresponding time buffers to these categories – 15 minutes, 30 minutes, 1 hour, 3 hours, and a day. Similarly to spatial resolution, we estimated the anonymity and uniqueness values according to the temporal resolutions.

3.4.3 Impact of Diverse Visited Locations

Lastly, to address the final question, this study examined the impact of the variety of visited locations on re-identification risks while keeping the number of observations constant. We categorized the combinations of travel observations into three types: 1) observations where a single location was visited multiple times, 2) observations where two distinct locations were visited, and 3) observations where three different locations were visited. We used box plots to visualize the distribution of re-identification risks and understand the variability of risks according to these combinations. From this, we deduced that the variety of visited locations within a given number of observations could significantly impact the anonymity values.

3.5 Results

3.5.1 Privacy Risks Based on the Number of Observations

We focused on three specific exposed travel observation numbers, 2, 3, and 4, to analyze the anonymity value, $\left|S\left(O_j^i\right)\right|$. With only one travel observation, it was challenging to re-identify the IDs. Numerous IDs display the same travel records, often limited to a single travel observation. For example, if 1000 individuals use the subway from Station A to Station B simultaneously on a specific day, it results in 1000 identical card IDs. In this example, the risk of re-identifying individual transit users might be negligible. Hence, we excluded the case of one observation and considered only the three specific numbers of observations. Considering that people's memory can fade over time, we analyzed patients' most recent four travel records to ensure high data accuracy. This approach resulted in a sample of 168 patients. We also assumed anonymity values less than 10 would pose a significant risk for re-identification attacks. Thus, our analysis was centred on these cases with anonymity values of less than 10.

Fig. 3-2 presents the detailed results and findings from the analysis. The x-axis represents the grouped number of exposed observations, while the y-axis indicates the cumulative uniqueness value. The legend shows the anonymity value, which is classified into three categories: 1, less than or equal to 2, and less than or equal to 10. We traced smart card IDs using 2, 3, and 4 exposed numbers and found that the risks increased correspondingly with the number of exposed travel observations. Considering 4 exposed travel observations, the uniqueness value with an anonymity value of 1 was 0.18. The anonymity value was lower than the corresponding value of 0.3 reported in a previous smart card study [18]. We attribute this difference to the imperfect accuracy of language-based travel observations compared to other studies.

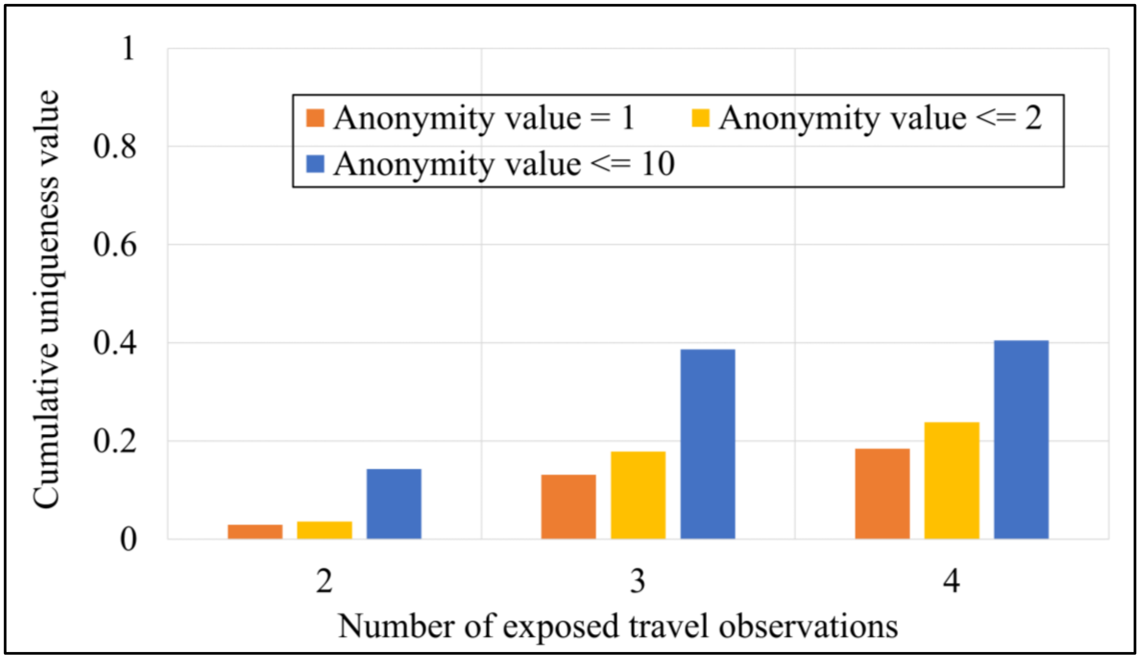


Fig. 3-2 Re-identification risks by the number of observations

3.5.2 Privacy Risks Based on Spatial and Temporal Resolutions

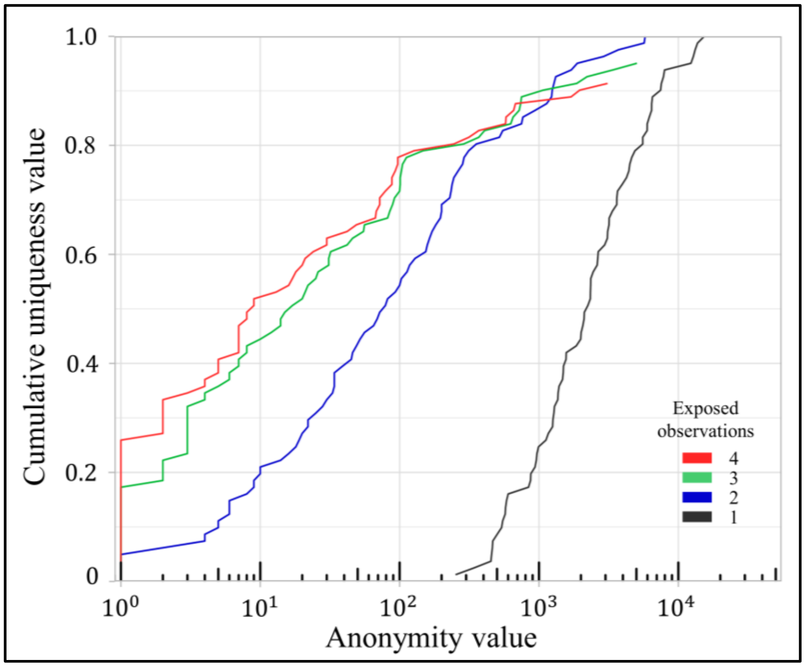
This subsection explores the impact of spatial and temporal resolutions derived from language-based observations on re-identification risks. The 168 patients were divided into 81 patients for high spatial resolution and 87 patients for low spatial resolution.

1) *Spatial Resolutions*

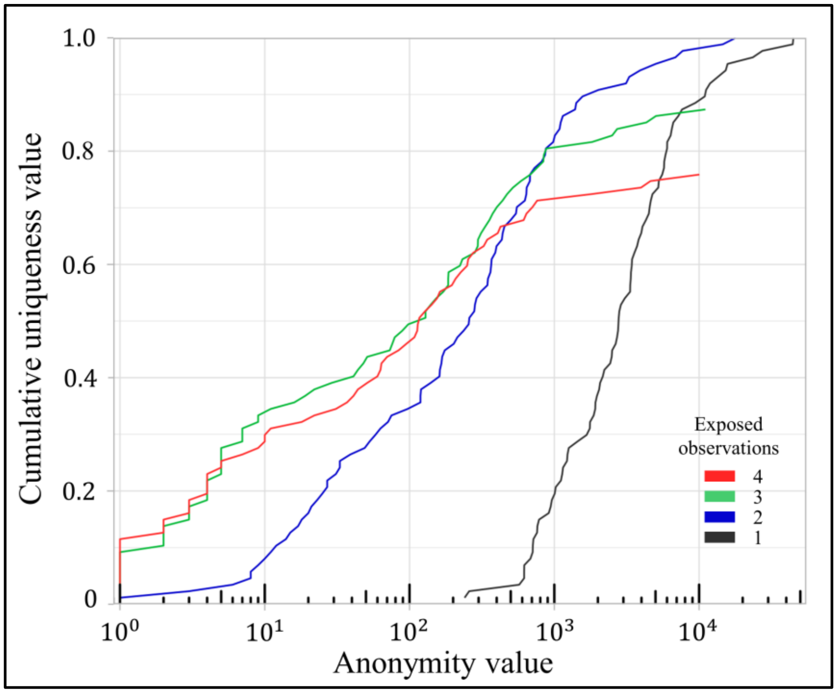
Fig. 3-3 illustrates how the re-identification risks fluctuated based on the number and spatial resolutions of exposed travel observations. Fig. 3-3 (a) and (b) indicate the high and low spatial resolutions, respectively. In both figures, the x-axis stands for the anonymity value in the logarithmic unit from 10^0 to 10^4. For instance, the anonymity of 10^0 indicates that adversaries can re-identify a unique individual from card data. As introduced in subsection 5.1, the y-axis indicates the cumulative uniqueness value, and the legend shows the number of exposed observations from 1(black line) to 4 (red line).

More specifically, Fig. 3-3 (a) showed a sharp rise in the uniqueness value to 0.26 at an anonymity value of 10^0 for 4 exposed travel observations (red line). This is significantly larger than 2 travel observations (blue line) with just 0.03. Similarly, Fig. 3-3 (b) highlighted the risk fluctuation of the low spatial resolutions. At an anonymity value of 10^0, the cumulative uniqueness value of the 4 travel observations was about 0.12, nearly half what is shown at high resolutions. The cumulative uniqueness values of other numbers of observations at 10^0 were significantly smaller than high-resolution observations. Our results, consistent with previous findings [15-16], indicated that the uniqueness values increased with higher spatial resolution observation data.

On the other hand, we observed fluctuations within the same spatial resolutions across the four lines (from black to red). For example, when considering 4 travel observations, 0.26 of patients could be re-identified at an anonymity level of 1. However, in the same red line, the uniqueness value increased by around 0.1 (from 0.4 to 0.5) at an anonymity level of 10. This led us to question what characteristics of the observations could account for such a discrepancy given the same number of exposed travel observations and spatial resolutions. We assumed that this variation stemmed from the diverse visited locations of individuals. While some people have repetitive travel records with the same stations and times, others may have heterogeneous patterns. We analyzed this in the following subsection, 5.3.



(a) By station



(b) by ward/town or destination

Fig. 3-3 Re-identification risks by spatial resolutions

2) *Temporal Resolutions*

We analyzed the variation in re-identification risks based on different temporal resolutions using the 4 exposed observations of the 81 patients in the high-spatial-resolution group. Fig. 3-4 is similar to Fig. 3-2, indicating re-identification risks associated with various temporal resolutions. It might be intuitive to expect higher temporal resolutions to result in a greater chance of re-identification. However, our experiments found a distinct characteristic concerning temporal resolutions. Surprisingly, the 1-hour temporal resolution exhibited the highest re-identification risk, with a value of 0.26 when the anonymity was 1 and 0.52 when the anonymity was 10. These values even surpassed those of the highest resolution, which was 15 minutes.

This discrepancy between our results and prior studies [17] could be attributed to the unique characteristics of language-based travel observations. People tend to recognize time in hourly units, which can unexpectedly influence re-identification risks. This is because many patients reported their travel times approximately; therefore, attempts to track them using higher temporal resolutions may result in lower uniqueness values.

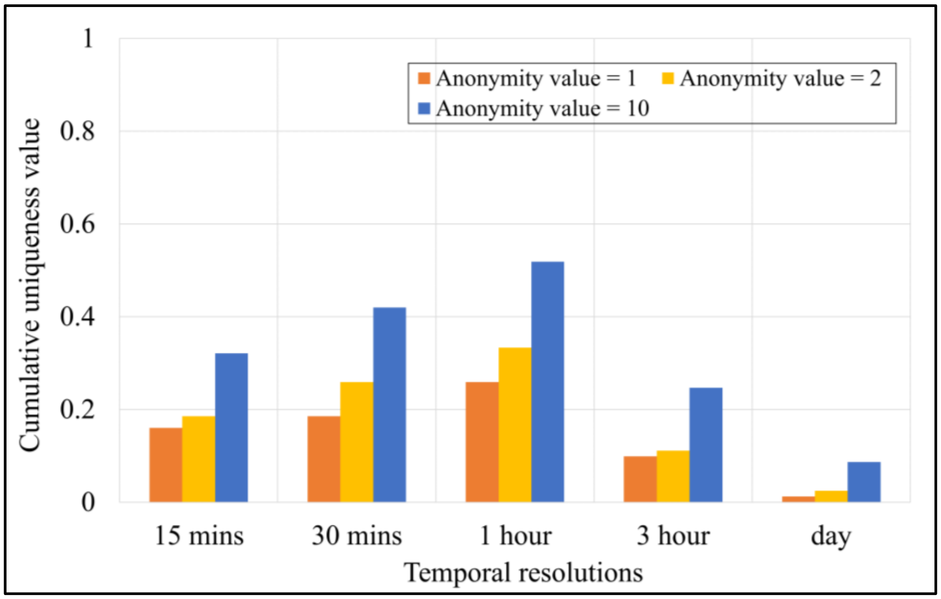


Fig. 3-4 Re-identification risks by temporal resolution

We further explored the impact of imperfect memory on time recognition, particularly focusing on the rate of missing data from tracing IDs. Fig. 3-5 shows the cumulative uniqueness value corresponding to temporal resolutions, represented by two legends: an *anonymity value of 1* and an *anonymity value of at least 1*. The dark-shaded bar graphs represent instances where the anonymity value equals 1, signifying that only one person's card ID matched the patients' travel itineraries after the re-identification attack. The light-shaded bar graphs indicate instances where at least 1 ID was re-identified, displaying itineraries identical to the patients. Thus, the anonymity values can range from 1 ID to the total number of patients, P. Given that the sum of the cumulative uniqueness value is 1.0, the missing rate can be estimated by subtracting the uniqueness value of at least 1 anonymity from the total. At the top of Fig. 3-5, the missing data rate is represented with arrows.

Intuitively, a day range of trace has zero missing rates since there should be at least one person who has identical travel records to the patients. In contrast, a temporal resolution of 15 minutes showed a high missing rate of about 55%. This high missing rate is attributed to the disparity between a high temporal resolution of data and people's imperfect memory. We believe these missing data rates resulted from the language-based travel observations.

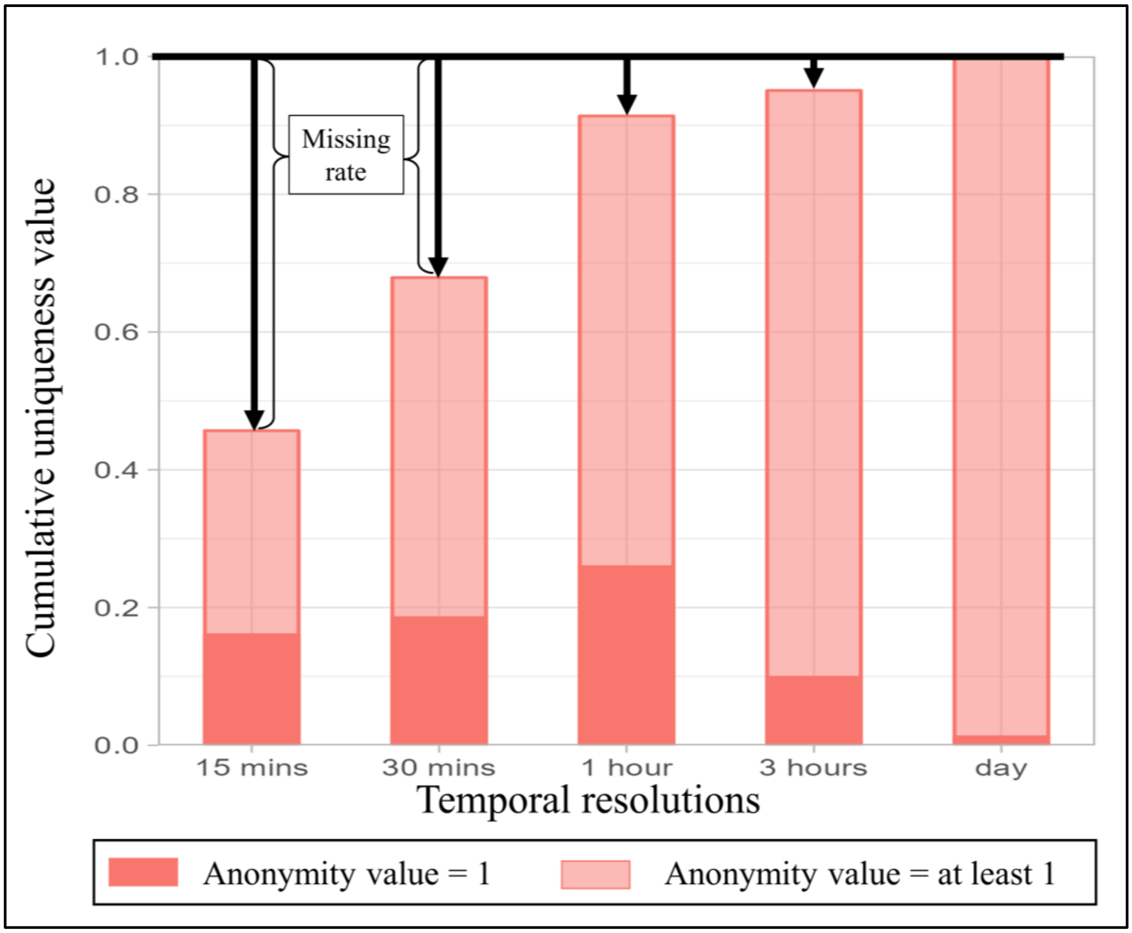


Fig. 3-5. Missing rate by temporal resolution

3.5.3 Privacy Risks Based on Diverse Visited Locations

Fig. 3-3 (a) showed that the cumulative uniqueness value fluctuated, even with a consistent number of exposed travel records. To explore the cause of these fluctuations, we examined how the re-identification risks would vary when the number of observations remained constant, but the combinations of visited stations changed. As discussed in the methods section, the diversity of visited locations can significantly impact the re-identification risks.

In Fig. 3-6, the grey box plots indicate the anonymity value with 3 observations, which did not consider diverse combinations of exposed locations. Like existing studies, the grey box plot considered the risks solely associated with the number of observations. Consequently, researchers using this approach could estimate the risks, but not exactly because they overlooked the potential impact of different visited locations.

In contrast to the previous researchers' methods, we classified the exposed travel observations in a grey box plot into three location categories based on the variety of visited locations: single location (visited repeatedly), 2 different locations, and 3 different locations. We illustrated these categories with three box plots in green, brown, and red, each representing various visited locations. Subsequently, we estimated the risks associated with the varying diversity of visited locations.

More specifically, Fig. 3-6 illustrates the quantified risks using the median value in the box plots. The existing method, a grey box plot, identified 7 IDs with identical travel records within a 1-hour data range. However, we discovered that the anonymity value varied depending on the number of visited locations: 80 IDs for 1 location, 6 IDs for 2 locations, and 1 ID for 3 locations. This indicated a decrease in anonymity value as location diversity increased and privacy protection was lowered.

Also, we evaluated the dispersion of anonymity values by comparing the interquartile ranges represented by the length of box plots. For the grey box, the upper and lower quartiles were 70 and 1, respectively, resulting in a quartile range of 69. The interquartile range of green, yellow, and red box plots were approximately 490, 19, and 0, respectively. Our analysis showed that as location diversity increased, the range of anonymity value became denser, implying a greater potential for re-identification.

This analysis demonstrated that a diverse combination of visited locations could amplify re-identification risks. We believe the differences in the risks between traditional methods and our approach can offer valuable insight into the impact of diverse visited locations on re-identification risks. Consequently, when the disclosure of patients' travel records is required in future pandemics, governments should be mindful of these real-world data features to ensure patient privacy.

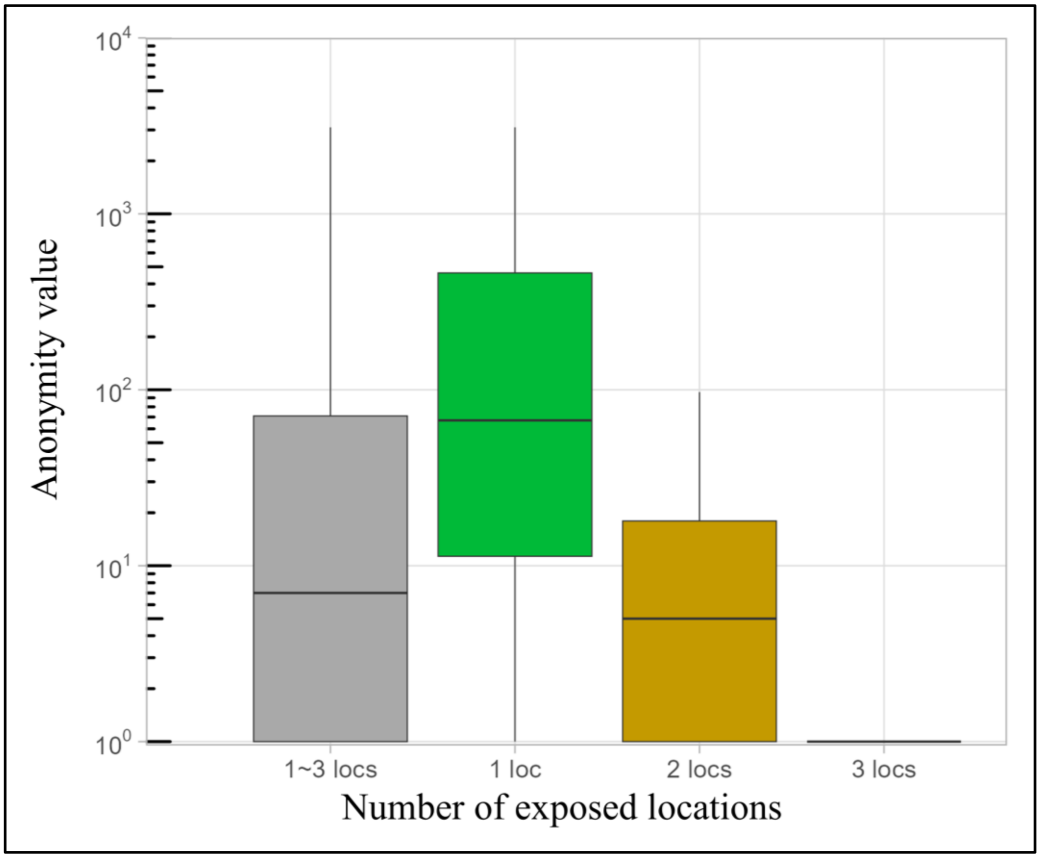


Fig. 3-6. Privacy risks by combinations of visited locations

3.6 Conclusion

During the COVID-19 pandemic, substantial amounts of sensitive personal information of the patients were shared publicly without sufficient precautions. The disclosed patient travel records contained distinct characteristics such as language-based varying data accuracy and diversity of the visited locations. Given the absence of research quantifying re-identification risks stemming from language-based travel records, this study brought a new perspective on the implication of these data characteristics.

This study thoroughly explored privacy risks associated with patient travel observations. First, we quantified the re-identification risks of smart card data based on travel records, a byproduct of government-led personal information breaches. Our findings indicated that the re-identification risks associated with data derived from language-based travel observations were not as severe as those suggested by existing studies. Secondly, we found that higher spatial data resolution can intensify the re-identification risks. Interestingly, a temporal data resolution of 1 hour showed a higher risk than a 15-minute resolution due to people's preferences for approximate time recognition. Lastly, the impact of various combinations of diverse visited locations can deteriorate the risks. Our findings highlighted the varying likelihood of re-identification risks based on the number, resolutions, and combinations of patient travel observations.

The vast amount of data derived from language-based sources, including Google Maps reviews and social networking stories, can weaken privacy protection. Our analysis of these language-based travel observations has yielded important insights into the associated re-identification risks. The inherent heterogeneity of observation numbers, spatial and temporal resolutions, and diverse combinations of visited locations provide a more comprehensive understanding of the real-world scenario. This knowledge will be a fundamental basis for devising future privacy protection measures.

Our research, though significant, does come with certain limitations. The travel record data samples were primarily insufficient to capture comprehensive and diverse individual travel patterns. South Korea's effective contact tracing policy made it challenging to obtain a large dataset of patient samples during the early stages of COVID-19 [25]. Additionally, the government has omitted detailed content from patient travel records from the pandemic’s mid-stages, further complicating data collection during the peak of virus propagation. Broadening the data collection scope to include other countries could enhance our understanding of the risks of publishing individual patients' travel records.

Second, this study did not explore protection methods for the publication of travel records. Our work represents a pioneering effort to consider real-world exposed data, including individual travel records, in quantifying re-identification risks. Our primary focus was understanding the aspects of travel observations that can amplify risks when dealing with other open data sources. Therefore, comparing protection methods could be an essential area of research for future studies.

Future work could aim to quantify a more comprehensive range of re-identification risks with larger data samples. This would facilitate a deeper understanding of privacy concerns associated with language-based individual information breaches.

3.7 Chapter 3 References

[1] L. Sun, K. W. Axhausen, D. Lee, and X. Huang, “Understanding metropolitan patterns of daily encounters,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 34, pp. 13774–13779, 2013, doi: 10.1073/pnas.1306440110.

[2] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,” in *2007 IEEE 23rd International Conference on Data Engineering*, Apr. 2007, pp. 106–115. doi: 10.1109/ICDE.2007.367856.

[3] T. Dalenius, “Finding a needle in a haystack or identifying anonymous census records,” *J. Off. Stat.*, vol. 2, no. 3, pp. 329–336, 1986, [Online]. Available: https://www.proquest.com/scholarly-journals/finding-needle-haystack-identifying-anonymous/docview/1266806751/se-2

[4] L. SWEENEY, “k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 05, pp. 557–570, Oct. 2002, doi: 10.1142/S0218488502001648.

[5] M. Terrovitis and N. Mamoulis, “Privacy Preservation in the Publication of Trajectories,” in *The Ninth International Conference on Mobile Data Management (mdm 2008)*, Apr. 2008, pp. 65–72. doi: 10.1109/MDM.2008.29.

[6] H. Ke, A. Fu, S. Yu, and S. Chen, “AQ-DP: A New Differential Privacy Scheme Based on Quasi-Identifier Classifying in Big Data,” in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2018, pp. 1–6. doi: 10.1109/GLOCOM.2018.8647941.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “L-diversity: privacy beyond k-anonymity,” in *22nd International Conference on Data Engineering (ICDE’06)*, 2006, pp. 24–24. doi: 10.1109/ICDE.2006.1.

[8] The_Government_of_the_Republic_of_Korea, “Flattening the curve on COVID-19: How Korea responded to a pandemic using ICT,” 2020. [Online]. Available: https://www1.undp.org/content/seoul_policy_center/en/home/presscenter/articles/2019/flattening-the-curve-on-covid-19.html

[9] S. Park, G. J. Choi, and H. Ko, “Privacy in the Time of COVID-19: Divergent Paths for Contact Tracing and Route-Disclosure Mechanisms in South Korea,” *IEEE Secur. Priv.*, vol. 19, no. 3, pp. 51–56, May 2021, doi: 10.1109/MSEC.2021.3066024.

[10] M. Borrelli, “WECHU warns of possible COVID-19 exposure risk at Leamington grocery store,” *CTV News*, Windsor, ON, Mar. 12, 2021. Accessed: Jun. 05, 2021. [Online]. Available: https://windsor.ctvnews.ca/wechu-warns-of-possible-covid-19-exposure-risk-at-leamington-grocery-store-1.5344750

[11] G. Jung, H. Lee, A. Kim, and U. Lee, “Too Much Information: Assessing Privacy Risks of Contact Trace Data Disclosure on People With COVID-19 in South Korea,” *Front. Public Heal.*, vol. 8, no. June, Jun. 2020, doi: 10.3389/fpubh.2020.00305.

[12] N. Y. Ahn, J. E. Park, D. H. Lee, and P. C. Hong, “Balancing Personal Privacy and Public Safety During COVID-19: The Case of South Korea,” *IEEE Access*, vol. 8, pp. 171325–171333, 2020, doi: 10.1109/ACCESS.2020.3025971.

[13] A. Majeed and S. O. Hwang, “A Comprehensive Analysis of Privacy Protection Techniques Developed for COVID-19 Pandemic,” *IEEE Access*, vol. 9, pp. 164159–164187, 2021, doi: 10.1109/ACCESS.2021.3130610.

[14] L. Krehling and A. Essex, “A Security and Privacy Scoring System for Contact Tracing Apps,” *J. Cybersecurity Priv.*, vol. 1, no. 4, pp. 597–614, Oct. 2021, doi: 10.3390/jcp1040030.

[15] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the Crowd: The privacy bounds of human mobility,” *Sci. Rep.*, vol. 3, no. 1, p. 1376, Mar. 2013, doi: 10.1038/srep01376.

[16] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. “Sandy” Pentland, “Unique in the shopping mall: On the reidentifiability of credit card metadata,” *Science (80-.).*, vol. 347, no. 6221, pp. 536–539, Jan. 2015, doi: 10.1126/science.1256297.

[17] J. Gao, L. Sun, and M. Cai, “Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data,” *Transp. Res. Part C Emerg. Technol.*, vol. 104, pp. 78–94, Jul. 2019, doi: 10.1016/j.trc.2019.04.022.

[18] Y. Li, D. Yang, and X. Hu, “A differential privacy-based privacy-preserving data publishing algorithm for transit smart card data,” *Transp. Res. Part C Emerg. Technol.*, vol. 115, p. 102634, Jun. 2020, doi: 10.1016/j.trc.2020.102634.

[19] G. Duggan, “Partial recall: Why we can’t trust our own memories | Nature of Things,” *CBC*, 2018. https://www.cbc.ca/natureofthings/features/partial-recall-why-we-cant-trust-our-own-memories (accessed Oct. 02, 2022).

[20] J. Shaw and S. Porter, “Constructing Rich False Memories of Committing Crime,” *Psychol. Sci.*, vol. 26, no. 3, pp. 291–301, Mar. 2015, doi: 10.1177/0956797614562862.

[21] Y. Han, J. Ku, Y. Kim, and J. Hwang, “Analyzing the accessibility of subway stations for transport-vulnerable population segments in Seoul: Case of bus-to-subway transfer,” *Case Stud. Transp. Policy*, vol. 10, no. 1, pp. 166–174, Mar. 2022, doi: 10.1016/j.cstp.2021.11.014.

[22] S. Lee and Y. G. Hur, “One Card Fits All : Integrated Public Transport Fare System,” Seoul, 2017. Accessed: Feb. 24, 2022. [Online]. Available: https://www.seoulsolution.kr/en/content/one-card-fits-all-integrated-public-transport-fare-system

[23] Y. Lu, J. Zhao, X. Wu, and S. M. Lo, “Escaping to nature during a pandemic: A natural experiment in Asian cities during the COVID-19 pandemic with big social media data,” *Sci. Total Environ.*, vol. 777, 2021, doi: 10.1016/j.scitotenv.2021.146092.

[24] A. El-Geneidy, M. Grimsrud, R. Wasfi, P. Tétreault, and J. Surprenant-Legault, “New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas,” *Transportation (Amst).*, vol. 41, no. 1, pp. 193–210, Jan. 2014, doi: 10.1007/s11116-013-9508-z.

[25] D. Scott and M. J. Park, “South Korea’s Covid-19 success story started with failure,” *Vox*, 2021. https://www.vox.com/22380161/south-korea-covid-19-coronavirus-pandemic-contact-tracing-testing (accessed Sep. 30, 2022).

CHAPTER 4

INFORMATION THEORY-BASED QUANTIFYING PRIVACY RISKS IN DATA

4.1 Introduction

We live in an era of excess information with the development of high-speed Internet, information and communications technology, and massive data handling and storage technologies [1-2]. In such environments, where information overload is common, people have questioned which information has more value and importance. Information value, however, can vary depending on the situation or individual perspectives. For instance, access to confidential news about mergers and acquisitions (M&A) or company profit data can significantly influence the return on investment (ROI) in the stock market. This type of news might be highly valuable to individuals aiming for substantial ROI. On the contrary, the same information might not appeal to someone who has never engaged in stock market investments. Therefore, the value of information can vary widely, influenced by people's individual interests and unique circumstances.

Similarly, there is a diverse range of sensitivities regarding privacy breaches in the realm of information privacy. Some individuals readily share personal details on social media, such as airline tickets, containing temporal and spatial information about their travel destinations. This information could potentially be used by adversaries to infer the individual's travel itinerary [3]. Conversely, others are much more cautious about revealing their whereabouts. For example, the installation of CCTVs by municipalities for security enhancement has often been met with complaints from residents, necessitating officials to engage in persuasive efforts. These examples illustrate the varying degrees of privacy risk awareness among individuals. It is challenging to determine which is more hazardous: personal travel itineraries exposed through social media or those captured by CCTV. This difficulty arises from the challenge of quantifying the risk level associated with individual exposures. Developing comprehensive mechanisms to quantify the amount of information and its associated risks can provide a solid foundation for protecting individual data.

Numerous researchers have focused on measuring privacy risks associated with open data, including medical records, license plate recognition (LPR), mobile phone data, and credit card transactions [4-6]. Various methodologies have been developed to assess the privacy risk in publicly opened data. Their methods typically assume that adversaries can access additional external information, like an individual’s daily commuting times, which could be observed through surveillance. By integrating this external information with open data, there is a potential for adversaries to reidentify individuals within the datasets. While this approach is useful for indicating privacy risks under certain conditions where external information aligns with the dataset, it may not fully capture the comprehensive risks present in the dataset. More specifically, existing methods tend to focus on measuring the privacy risks of the remaining datasets by filtering out data irrelevant to external information. In our view, their findings may offer a limited perspective on the overall privacy risks due to the lack of filtered-out data. The findings could be suitable for only the specific external observation circumstances. Also, to measure the amount of information in the dataset, considering the whole dataset with all possible circumstances is crucial. Therefore, adopting an alternative approach that accounts for every possible situation within the dataset could improve our understanding of associated privacy risks.

Entropy can be the alternative approach in privacy risk research. Claude Shannon laid the foundation for understanding how information is processed and transmitted [7]. Shannon introduced a concept regarding entropy in the context of information theory, measuring a data set's unpredictability or randomness. Based on Shannon's research, entropy can also be used to assess a dataset's uncertainty or unpredictability level. For example, let's consider the scenario of rolling a die. If the frequency of rolling a 4 is extremely low, occurring only once in 100 attempts, the event of rolling a 4 becomes a significant surprise. This low probability of rolling a 4 implies that seeing it chosen would be unexpected, and this element of surprise can be regarded as information due to the rareness of occurring [8]. We can say that as the probability of an event occurring decreases, the uncertainty about the event increases. Measuring privacy risks and the amount of information also needs the probability of data. Suppose we are playing with a friend to guess the number of dice in the friend's hands. There will be no surprise or uncertainty when the friend says it is a 3. However, if the friend says the number is less than 4, we still need to guess the answer from 1, 2 and 3. There is still uncertainty to guess, and if we need to guess it with only one attempt, we should consider the probability of the remaining numbers occurring compared to the entire dice rolling [8]. Entropy can be a measurement to guess the correct answer. If we adopt it to identify people in datasets, we could measure the privacy risk level by reducing the uncertainties of guessing.

This dissertation introduces a methodology for quantifying the amount of information and the level of privacy risk in a dataset. For this purpose, the research constructs a synthetic dataset that captures every individual visit, encompassing all possible spatial and temporal combinations of visited information. The research identified combinations where every person is uniquely distributed or classified into the smallest number of people, which cannot be identified individually. We regard these unique distribution combinations as *resolved uncertainty situations*. The conditional and joint entropy have been utilized to represent specific combinations, considering the fluctuation of uncertainty and privacy risk by adding a piece of visit information. This chapter's primary contribution is exploring information theory methods to quantify the resolved uncertainty and establish a connection between this resolved uncertainty and privacy risks in the dataset. A sensitivity analysis was conducted to identify the characteristics of travel information combinations. This is crucial as certain combinations of travel records may lead to earlier identification of individuals than other combinations. The results could provide significant insights into enhancing data privacy when it is published.

The remainder of the chapter is organized as follows. Section 2 explores the significance and concepts of information theory. Section 3 discusses related works, especially for the existing privacy research and information theory-based uncertainty studies. Section 4 introduces the methodologies to quantify the dataset's information amount and privacy risks; Section 5 consists of results and findings. Section 6 summarizes the main findings and future research directions.

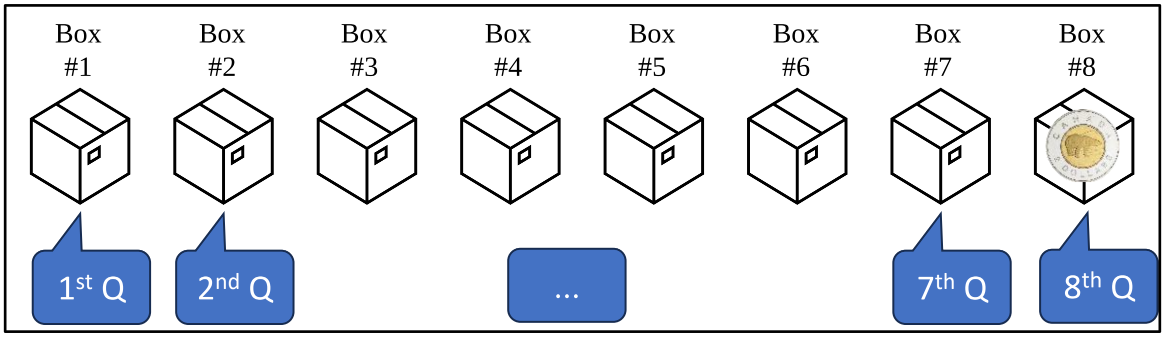
4.2 Preliminary: Information Theory

This section explores the fundamental concepts of information theory, which form the cornerstone of this chapter.

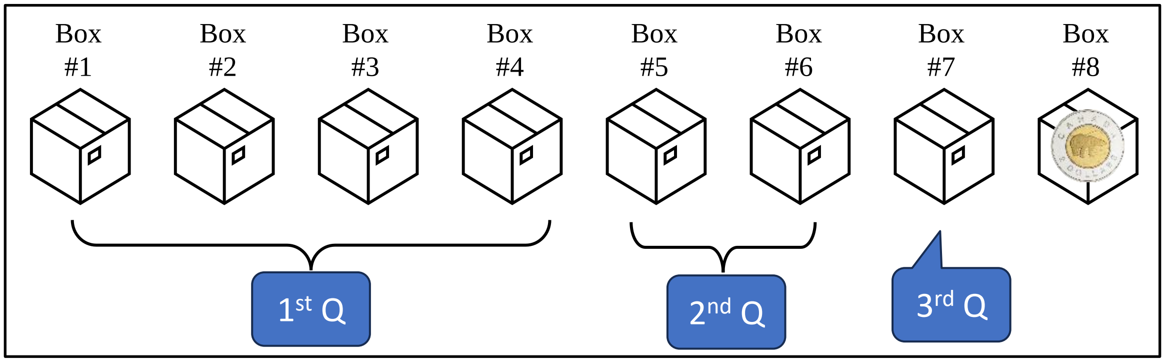
4.2.1 Introduction to Information Theory

In information theory, *Shannon Entropy* mainly represents the uncertainty of a single random variable and indicates the amount of information. This concept has been introduced by many researchers using familiar examples, such as a game depicted in Fig. 4-1 [8-10]. Consider a game where the objective is to find a coin hidden in one of the eight boxes. In this game, the information we need is “where the coin is.” To find the coin, we are allowed to ask only binary questions. One approach is to ask about each box individual: “Is it in box 1?”, “Is it in box 2?”, and so on, up to “Is it in box 8?” as shown in Fig. 4-1 (a). If the coin were in the 1st box, we could find it with just one question, though the probability of this happening is very low. Conversely, if the coin were in the 8th box, we would need to ask eight questions. On average, this method would require 4.5 questions to find the coin. This average is calculated by summing the number of questions needed to find the coin in each scenario (from 1 to 8), and then dividing by 8.

Alternatively, we can optimize the process by asking more strategic binary questions. For instance, we can begin by asking, “Is the coin in the right half of the eight boxes?” as depicted in Fig. 4-1 (b). Depending on the answer, we then focus on either the right or left half, successively halving the search area with each question. This method allows us to locate the coin using just three questions: first by reducing the eight boxes to four, then to two, and finally identifying the specific box. While this method invariably requires three questions, it may initially seem less efficient than the one or two questions needed in the individual box inquiry method. However, given the low probability of finding the coin with fewer questions in the individual method, the binary halving inquiry method offers a more reliable and often more efficient solution.



(a) Individual box inquiry method



(b) Binary halving inquiry method

Fig. 4-1 Comparative illustration of two methods for the coin-finding game

We can estimate the probability of finding the coin in any box by diving the total number of boxes (n), as $p = \frac{1}{n}$. The equation to determine the number of questions (n_Q) required in this game can be derived as follows [8]:

$$n_Q = \log_2 n = \log_2(\frac{1}{p}) = -\log_2 p \equiv h \qquad (1)$$

Now, we can calculate the number of questions needed to *resolve our uncertainty* about the outcome of the game. In this context, h represents the amount of uncertainty, which is also known as *Shannon Entropy* or simply *Entropy* [8]. This approach remains efficient even when the number of boxes is doubled. For instance, if the number of boxes increases to 16, the required number of questions – or the amount of uncertainty – can be calculated as $\log_2(1/16) = 4$; this is just one additional question compared to the eight-box scenario. In contrast, the individual box inquiry method would require 16 questions, doubling the number of questions with the doubled number of boxes. This illustrates why information theory is so powerful: it provides a method to quantify uncertainty in a scalable and efficient manner, regardless of the size of the database.

In the coin-finding game example, we explored how entropy serves as a measure of uncertainty in a simplified scenario, illustrating a fundamental concept in information theory. The principle of *uncertainty reduction* is pivotal for understanding the basics of information theory and plays a significant role in our research. Measuring uncertainty through this lens is vital for analyzing the amount of information that can be derived from individual travel records, such as assessing the likelihood of someone visiting a specific place. However, the coin-finding game is just an introduction to the broader applications of information theory. We will delve into more complex aspects, with a particular focus on probability distributions. These distributions offer a more detailed perspective on uncertainty and information theory, which is essential for our research.

4.2.2 Measuring the Uncertainty in Probability Distribution

In contrast to the straightforward scenario of the coin-finding game, which revolves around a single event, probability distributions represent a sophisticated and multifaceted application of information theory. This section delves into how information theory quantifies the inherent uncertainty in these distributions.

Consider X as a discrete random variable representing outcomes ranging from 1 to n. The probability distribution of X, denoted as P_i, can be expressed as $P_1, P_2, \dots, P_n$. Using the coin-finding game as an analogy, let us say X represents which box contains the coin. With each of the eight boxes having an equal probability of 1/8 of containing the coin, we can apply the uncertainty equation (1) for a single outcome of X. In a scenario where each of the eight boxes contains the coin exactly once across multiple trials, each probability would be identical, such as $P_1 = P_2 = \cdots = P_8 = 1/8$. To calculate the average uncertainty associated with the probability of finding the coin across all eight outcomes (from P_1 to P_8), we use the following equation (2) for the random variable X:

$$H(X) = E[h(X)] = -\sum_{i=1}^{n} P_i \log_2 P_i \qquad (2)$$

Equation (2) represents the sum of the expected value of the uncertainty for all outcomes, as defined by Shannon Entropy [7]. For each outcome, the uncertainty is quantified by $\log_2 P_i$, and when this is multiplied by the probability, P_i, it gives the expected value of uncertainty for that particular outcome (P_i). Summing these values for all possible outcomes of X, we obtain a measure of the *average uncertainty (or entropy)* associated with the probability distribution of finding the coin in one of the eight boxes. This sum is essentially an average because it combines the uncertainties of all outcomes, taking into account how likely each outcome is. In other words, the more likely an outcome is, the more it contributes to the overall uncertainty calculation. This method of combining uncertainties, where the likelihood of each outcome influences its contribution to the total, exemplifies what Shannon Entropy measures. For instance, in the coin-finding game, where each of the eight boxes has an equal probability (1/8) of containing the coin, the Shannon Entropy calculation considers the probability of the coin being in each box. The uncertainty for each box is calculated as $-\log_2(\frac{1}{8})$. Since each outcome (box) has the same probability, the entropy is the sum of these individual uncertainties multiplied by their probability, as follows,

$$H(X) = -8 \times \frac{1}{8} \times \log_2\left(\frac{1}{8}\right) = 3\ bits$$

When the base is 2 of the logarithms, we deemed the entropies to be measured in bits [10]. If the probability of containing the coin is not equal to 1/8, such as 1/16, 1/16, 1/16, 1/16, 1/8, 1/8, 1/4, 1/4, the calculation of the uncertainties will be as follows,

$$\begin{aligned} H(X) = &-\frac{1}{16} \times \log_2\left(\frac{1}{16}\right) - \frac{1}{16} \times \log_2\left(\frac{1}{16}\right) - \frac{1}{16} \times \log_2\left(\frac{1}{16}\right) - \frac{1}{16} \times \log_2\left(\frac{1}{16}\right) \\ &- \frac{1}{8} \times \log_2\left(\frac{1}{8}\right) - \frac{1}{8} \times \log_2\left(\frac{1}{8}\right) - \frac{1}{4} \times \log_2\left(\frac{1}{4}\right) - \frac{1}{4} \times \log_2\left(\frac{1}{4}\right) \\ &= 2.75\ bits \end{aligned}$$

This reflects the average amount of information or surprise typically associated with learning which specific box contains the coin. Hence, it is often described as the *average uncertainty* of the probability distribution. This calculation quantifies the *expected level of uncertainty across all possible outcomes*. Table 4-1 shows the notation of equation (2).

Table 4-1. Notation

Notation	Description
X	Discrete random variable representing possible outcomes (e.g., which box contains the coin)
$H(X)$	Average uncertainty or a measure of the expected uncertainty across all possible outcomes of X
$h(X)$	Uncertainty of a single outcome of X, quantified using equation (1)
P_i	Probability of the i^{th} outcome of X, where i ranges from 1 to n

4.2.3 Joint Entropy and Conditional Entropy

1) Understanding Joint and Conditional Entropy

Joint entropy and conditional entropy are important metrics for quantifying the amount of information for multiple variables. This chapter adopts conditional and joint entropy equations as established in previous studies [8], [10]. Suppose we have two discrete random variables X and Y, with n and m possible outcomes, respectively. Their joint probability distribution is denoted as $P(X = i, Y = j)$, and the individual distribution of variables X and Y are $P_i \ and\ P_j$, respectively. The joint entropy, which measures the total uncertainty in both X and Y together, can be expressed as:

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(X = i, Y = j) \cdot \log_2(P(X = i, Y = j)) \qquad (3)$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{i,j} \cdot \log_2(P_{i,j}) \qquad \text{simplified (3)}$$

This equation sums the product of the joint probability and the logarithm of the joint probability over all possible states of the variables X and Y to measure the total uncertainty.

Following joint entropy, we also consider conditional entropy. Conditional entropy quantifies the average amount of information or uncertainty of one variable, X, given the knowledge of another variable, Y. More specifically, when the value of variable Y is known, conditional entropy measures the remaining uncertainty or information in X, considering the conditional probability distributions. This remaining information represents the *unique* information content of X that is not shared with Y. If the variables X and Y have joint probability denoted as $P_{i,j}$, where $P_i \ and \ P_j$ are their respective individual probabilities, the conditional entropy of X given Y can be defined by the following equations:

$$H(X|Y) = -\sum_{j=1}^{m} P(Y = j) \cdot H(X|Y = j) \qquad (4)$$

$$= \ -\sum_{j=1}^{m} P_j \cdot H(X|Y) \qquad \text{simplified (4)}$$

$$= -\sum_{j=1}^{m} P(Y = j) \cdot \sum_{i=1}^{n} P(X = i|Y = j) \cdot \log_2(P(X = i|Y = j)) \quad (5)$$

$$= -\sum_{j=1}^{m} P_j \cdot \sum_{i=1}^{n} \left(P_i \middle| P_j\right) \cdot \log_2(P_i|P_j) \qquad \text{simplified (5)}$$

$$= -\sum_{i=1}^{n} \sum_{j=1}^{m} P(X = i, Y = j) \cdot \log_2 \frac{P(X = i, Y = j)}{P(Y = j)} \qquad (6)$$

$$= -\sum_{i=1}^{n} \sum_{j=1}^{m} P_{i,j} \cdot \log_2 \frac{P_{i,j}}{P_j} \qquad \text{simplified (6)}$$

More specifically, equation (4) calculates the conditional entropy of X given Y. $P(Y = j)$ is the probability of Y being in state j. $H(X|Y = j)$ is the entropy of X when Y

is known to be in state j. Thus, multiplying $P(Y = j)$ by $H(X|Y = j)$ is a process to calculate the average or expected uncertainty in X when the value of Y is known. Equation (5) is an expanded form of equation (4). It gives us the breakdown of the conditional entropy equation $H(X|Y = j)$ by dividing it by multiplying the conditional probability of X given Y, $P(X = i|Y = j)$, by the surprise or information content associated with X in Y, $\log_2(P(X = i|Y = j))$. Lastly, equation (6) is another expanded form of the conditional entropy. Equation (6) provides a direct calculation method using the joint probability distribution $P(X = i, Y = j)$ and marginal probability distribution $P(Y = j)$. The logarithmic term $\log_2(\frac{P(X=i,Y=j)}{P(Y=j)})$ calculates the surprise or information associated with the occurrence of X given Y. This means that the logarithm process quantifies the additional information provided by X when Y is known.

Therefore, these equations from (4) to (6) effectively capture the average uncertainty in X that remains after observing Y, providing insights into the interdependence of these variables.

2) Applying the Chain Rule

This subsection delves into the relationship between joint entropy and conditional entropy within the framework of information theory. The *chain rule for entropy* is a key concept that illustrates how joint entropy is fundamentally the cumulative sum of conditional entropies [10]. This principle is particularly insightful for measuring the uncertainty in travel information by aggregating various information pieces.

As previously discussed, conditional entropy captures the unique information of a variable in the context of a given variable. The following equations (7) and (8) demonstrate the relationship between joint entropy and conditional entropy for two and three variables [10].

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1) \quad (7)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1) \quad (8\text{-}1)$$

$$= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \quad (8\text{-}2)$$

In equation (7), $H(X_1)$ represents the total information of the variable X_1, and $H(X_2|X_1)$ signifies the unique information of the variable X_2 given X_1. The sum of the two parts of the equation (7) can be equivalent to the joint entropy, $H(X_1, X_2)$, of the two variables. For three variables, $X_1, X_2, and\ X_3$, equation (8) can be divided into two parts, as shown in equations (8-1) and (8-2). Equation (8-1) consists of the total information of X_1 and the conditional entropy of X_2 and X_3 given X_1. The term $H(X_2, X_3|X_1)$ in equation (8-1) represents joint entropy of $X_2\ and\ X_3$, excluding the information shared with variable X_1. Thus, the joint entropy of three variables is the sum of the complete information of X_1 and remaining joint entropy of $X_2\ and\ X_3$, excluding X_1.

Equation (8-2) further dissects $H(X_2, X_3|X_1)$, into two parts: $H(X_2|X_1)$ and $H(X_3|X_2, X_1)$. This breakdown implies the new information obtained from variables X_2 and X_3 is incremental, excluding the previously accounted information. More specifically, $H(X_1)$ encompasses the total information of X_1, being the initial information source. Adding $H(X_2|X_1)$ introduces new information from X_2, independent of X_1. Subsequently, $H(X_3|X_2, X_1)$ derives information from X_3 that is not contained in $X_2\ and\ X_1$.

This approach is vital for our study as it allows us to quantify the total uncertainty in travel records by sequentially adding information from specific times or locations. The following equation (9) is the sum of the conditional entropies using the chain rule for multivariates from 1 to n [10].

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) \dots + H(X_n|X_{n-1}, \dots, X_1) \qquad (9)$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}, \dots, X_1)$$

4.3 Related Work

Over the past few decades, privacy risk research has received significant attention. This section delves into existing studies on privacy risk, with a particular focus on risk measurement methodologies, including re-identification risk research and approaches for uncertainty utilizing information theory.

4.3.1 Privacy Risk Measurement Research with External Observations

Numerous researchers have focused on privacy risks associated with anonymized and publicly shared data, especially when analyzed through external observation. Anonymized datasets contain attributes known as *quasi-identifiers (QIDs),* which may include details like ZIP code, birth date, gender, prices of purchased items, visited locations and times [11]. Dalenius pointed out that QIDs in anonymized datasets can be utilized to reidentify individuals, especially those with unique attributes or who are publicly known [12]. Sweeney exemplified a notable case of reidentifying a public figure using QIDs and external observation. Sweeney successfully reidentified the governor of Massachusetts's records using anonymized data from voter registration and health insurance records [4]. This was achieved by matching the governor’s publicly known information, in other words, external observations, such as ZIP code, birth date and gender. Consequently, Sweeney found the governor’s sensitive medical records. Therefore, using QIDs and external observations has posed significant privacy risks and attracted the interest of many researchers.

Researchers have utilized QIDs to demonstrate the vulnerability of shared data. By reidentifying individuals and narrowing down the number of candidates with identical travel records, they have effectively demonstrated the inherent privacy risks. This process highlights the vulnerability of shared data and underscores the need for robust privacy-preserving techniques in data anonymization.

However, existing research has focused on cases where QIDs directly match external observation data, leading to a narrowed perspective. By concentrating on matching data, researchers might overlook non-matching data. This selective approach can result in an analysis that emphasizes specific subsets of data, potentially missing out on the complete picture. Consequently, such methods may not fully capture the overall privacy risks as they do not consider the entire dataset. An alternative approach, which encompasses a comprehensive analysis of all data, including both matching and non-matching subsets, is necessary for a more accurate quantification of privacy risks.

4.3.2 Information Theory-based Research Regarding Uncertainty

As introduced in the preliminary section, information theory offers a framework to measure the amount of uncertainty in information. A key aspect of this approach is its ability to aggregate all probabilities, including those from both exactly matched and non-matching data. This inclusive nature of information theory allows for the consideration of every possible case, irrespective of whether there is a match or not.

In particular, the use of conditional entropy and joint entropy in information theory enables the quantification of the incremental amounts of information added up to the total uncertainty level achieved by aggregating various information pieces of information. Some studies have been based on information theory to understand the amount of uncertainty. For instance, Wellmann and Regenauer-Lieb applied information theory to visualize uncertainties in a structural geological model [13]. Geological models are used to understand and predict the distribution of rocks and fluids in the subsurface of the Earth [14]. Given the significant uncertainties below ground, structural geological models serve as vital tools for providing essential information about underground circumstances, thereby helping minimize investigation costs. Furthermore, Wellmann also examined how additional information could reduce uncertainty regarding potential drilling locations [15]. The research team explored the overall reduction of uncertainty by incorporating information from multiple drilling sites, employing conditional entropy and joint entropy methods and treating the multiple drilling locations as multivariate input data.

This method can be effectively adapted to the field of transportation engineering, where travel data are collected from multiple locations, and the value of information derived is often uncertain. In particular, research that measures the fluctuation of uncertainty related to privacy risks in transportation data is essential for understanding privacy implications. The subsequent section delves into methods based on information theory for quantifying the amount of uncertainty derived from multiple information sources.

4.4 Methods of Analysis

Our study focused on quantifying the amount of information inherent in data. To achieve this main goal, we hypothesized that privacy risks can be quantified by measuring the amount of resolved uncertainty in units of entropy. In addition, we explored how the resolved uncertainty regarding the re-identification of individuals changes with the addition of more information pieces. On top of that, this chapter considered how different amounts of resolved uncertainty emerge based on various combinations of information acquisition. With this regard, the following three fundamental questions guided our research:

1) What is the relationship between privacy risk and resolved uncertainty, and what methods can be utilized to quantify this uncertainty?
2) How does the amount of resolved uncertainty increase with the addition of more information pieces?
3) How does the uncertainty vary with different combinations of information acquisition?

The subsequent subsections explore the methods and provide examples of approaches to the above questions.

4.4.1 Concepts of Rareness and Commonness in Datasets

Suppose a dataset containing the travel records of 8 individuals across 6 different locations. These records specify whether each person visited a particular location, resulting in a binary outcome: ‘visited’ or ‘not visited’. These binary outcomes can be numerically represented for analytical convenience with ‘visited’ coded as 1 and ‘not visited’ as 0. Table 4-2 illustrates these travel records, organized by person and location.

In the context of privacy, information about location D is particularly sensitive, as it uniquely identifies a person with ID 5 as the only person who visited the location. As discussed in the preliminary section, this information contains a high level of surprise due to its rareness, with the probability of occurrence being $\frac{1}{8}$. Conversely, for the other seven individuals who did not visit this location, the probability is $\frac{7}{8}$. Applying the entropy equation, the average entropy for this information is calculated as $-\frac{1}{8} \cdot \log_2(\frac{1}{8}) - \frac{7}{8} \cdot \log_2(\frac{7}{8}) = 0.544\ bits$. Also, we can calculate the entropy of each location with the equation. As indicated by the entropy values in Table 4-2, 0.544 bits is the lowest entropy value compared to other locations. Thus, while the information about visiting location D uniquely identifies ID 5, *it does not resolve uncertainty sufficiently* to identify all eight individuals in the dataset. It would be challenging to understand privacy risks accurately, focusing on specific uniquely identified cases only. This illustrates that the significance of information about visiting a location should be evaluated comprehensively, considering both its rareness and commonness.

Table 4-2. Sample data

ID	Location A	Loc. B	Loc. C	Loc. D	Loc. E	Loc. F
1	1	0	0	0	0	0
2	0	1	0	0	0	1
3	0	0	1	0	0	0
4	1	1	0	0	0	0
5	0	0	0	1	1	0
6	0	0	0	0	1	0
7	0	1	1	0	0	1
8	0	0	1	0	1	0
# of visited	2	3	3	1	3	2
# of not visited	6	5	5	7	5	6
Entropy	0.811	0.954	0.954	0.544	0.954	0.811

4.4.2 Amount of Uncertainty and Privacy Risk

1) Identification Process

In the preliminary section, we discussed how Shannon entropy can estimate the amount of uncertainty to identify a coin among eight boxes. Applying this concept to privacy risks, we can quantify the amount of the resolved uncertainty when all eight individuals from Table 4-2 are identified, which is the most critical privacy risk situation. When all eight individuals in a dataset are uniquely identifiable, each individual can be distinctly separated based on the available information. In such a case, the resolved uncertainty of identifying these eight individuals can be calculated as $H(X) =$

$-8 \times \frac{1}{8} \times \log_2\left(\frac{1}{8}\right) = 3\ bits$. This value represents the *total resolved uncertainty* involved in uniquely identifying each individual. However, as indicated by the data in Table 4-2, uniquely identifying all eight individuals is not feasible based solely on the information from one location. Therefore, it is necessary to accumulate information from multiple locations to assess whether the individuals can be uniquely identified.

Table 4-3 demonstrates the process of identifying individuals as we accumulate and analyze more location information, applying binary visit conditions: visited (denoted as ‘1’) or not visited (denoted as ‘0’). In this table, ‘A(1)’ signifies a visit to location A, while ‘A(0)’ indicates that location A was not visited. Each condition has its own condition number, which consists of the number of information pieces (*i*) and the order number (*j*) of combinations in the same pieces (*i*). Initially, we cannot uniquely identify any individual with information from location A alone. However, incorporating data from additional locations enables distinguishing individuals. For instance, with the first set of information about location A, we categorize IDs 1 and 4 as having visited (coded as A(1)) and the other individuals as not having visited (coded as A(0)). At this stage, no individual has been uniquely identified.

The identification process becomes more refined by adding information from the second location B. Now, we have two pieces of information: one from location A and another from location B. This dual information allows for a more detailed division of individuals based on their visit conditions at both locations. More specifically, ID 1, who visited location A but not B, is represented as B(0) | A(1), and ID 4, who visited both locations, is shown as B(1) | A(1). As we continue to integrate information from additional locations C, D, and E, our ability to distinguish between individuals becomes even more precise. With five pieces of information – encompassing locations A, B, C, D, and E – we reach a point where each of the eight individuals can be uniquely identified based on their specific combinations of visit conditions. This scenario represents the most significant privacy risk, where individuals are identifiable solely through their visit records.

Table 4-3. Identification results by varying information pieces and visit conditions

# of info.	Con.	Visit conditions	IDs
1 piece	C1-1	A(0)	ID 2, 3, 5, 6, 7, 8
	C1-2	A(1)	ID 1, 4
2 pieces	C2-1	B(0) | A(0)	ID 3, 5, 6, 8
	C2-2	B(1) | A(0)	ID 2, 7
	C2-3	B(0) | A(1)	ID 1
	C2-4	B(1) | A(1)	ID 4
3 pieces	C3-1	C(0) | B(0), A(0)	ID 5, 6
	C3-2	C(1) | B(0), A(0)	ID 3, 8
	C3-3	C(0) | B(1), A(0)	ID 2
	C3-4	C(1) | B(1), A(0)	ID 7
	C3-5	C(0) | B(0), A(1)	ID 1
	C3-6	C(1) | B(0), A(1)	-
	C3-7	C(0) | B(1), A(1)	ID 4
	C3-8	C(1) | B(1), A(1)	-
4 pieces	C4-1	D(0) | C(0), B(0), A(0)	ID 6
4 pieces	C4-2	D(1) | C(0), B(0), A(0)	ID 5
	C4-3	D(0) | C(1), B(0), A(0)	ID 3, 8
	C4-4	D(1) | C(1), B(0), A(0)	-
	C4-5	D(0) | C(0), B(1), A(0)	ID 2
	C4-6	D(1) | C(0), B(1), A(0)	-
	C4-7	D(0) | C(1), B(1), A(0)	ID 7
	C4-8	D(1) | C(1), B(1), A(0)	-
	C4-9	D(0) | C(0), B(0), A(1)	ID 1
	C4-10	D(1) | C(0), B(0), A(1)	-
	C4-11	D(0) | C(1), B(0), A(1)	-
	C4-12	D(1) | C(1), B(0), A(1)	-
	C4-13	D(0) | C(0), B(1), A(1)	ID 4
	C4-14	D(1) | C(0), B(1), A(1)	-
	C4-15	D(0) | C(1), B(1), A(1)	-
	C4-16	D(1) | C(1), B(1), A(1)	-

# of info.	Con.	Visit conditions	IDs	# of info.	Con.	Visit conditions	IDs
5 pieces	C5-1	E(0) | D(0), C(0), B(0), A(0)	-	5 pieces	C5-17	E(0) | D(0), C(0), B(0), A(1)	ID 1
	C5-2	E(1) | D(0), C(0), B(0), A(0)	ID 6		C5-18	E(1) | D(0), C(0), B(0), A(1)	-
	C5-3	E(0) | D(1), C(0), B(0), A(0)	-		C5-19	E(0) | D(1), C(0), B(0), A(1)	-
	C5-4	E(1) | D(1), C(0), B(0), A(0)	ID 5		C5-20	E(1) | D(1), C(0), B(0), A(1)	-
	C5-5	E(0) | D(0), C(1), B(0), A(0)	ID 3		C5-21	E(0) | D(0), C(1), B(0), A(1)	-
	C5-6	E(1) | D(0), C(1), B(0), A(0)	ID 8		C5-22	E(1) | D(0), C(1), B(0), A(1)	-
	C5-7	E(0) | D(1), C(1), B(0), A(0)	-		C5-23	E(0) | D(1), C(1), B(0), A(1)	-
	C5-8	E(1) | D(1), C(1), B(0), A(0)	-		C5-24	E(1) | D(1), C(1), B(0), A(1)	-
	C5-9	E(0) | D(0), C(0), B(1), A(0)	ID 2		C5-25	E(0) | D(0), C(0), B(1), A(1)	ID 4
	C5-10	E(1) | D(0), C(0), B(1), A(0)	-		C5-26	E(1) | D(0), C(0), B(1), A(1)	-
	C5-11	E(0) | D(1), C(0), B(1), A(0)	-		C5-27	E(0) | D(1), C(0), B(1), A(1)	-
	C5-12	E(1) | D(1), C(0), B(1), A(0)	-		C5-28	E(1) | D(1), C(0), B(1), A(1)	-
	C5-13	E(0) | D(0), C(1), B(1), A(0)	ID 7		C5-29	E(0) | D(0), C(1), B(1), A(1)	-
	C5-14	E(1) | D(0), C(1), B(1), A(0)	-		C5-30	E(1) | D(0), C(1), B(1), A(1)	-
	C5-15	E(0) | D(1), C(1), B(1), A(0)	-		C5-31	E(0) | D(1), C(1), B(1), A(1)	-
	C5-16	E(1) | D(1), C(1), B(1), A(0)	-		C5-32	E(1) | D(1), C(1), B(1), A(1)	-

2) Quantifying Uncertainty by Process

This subsection builds upon the methods outlined in the previous subsection, where we identified individuals by accumulating location information. In this section, our focus shifts to quantifying the amount of uncertainty based on information theory, particularly on joint entropy and conditional entropy [10]. This requires estimating the probability under each visit combination in terms of joint entropy and conditional entropy.

As calculated, the total uncertainty of identifying every eight individuals was 3 bits, representing the highest level of uncertainty. Fig. 4-2 illustrates the identification process, along with the corresponding probabilities used to quantify the amount of uncertainty at each process. Initially, individuals are categorized into two groups based on their visit information for location A: those who did not visit (A(0)) and those who did (A(1)), referred to as condition C1-1 (six people) and C1-2 for (two people), respectively. The probabilities for these categories are thus $\frac{6}{8}$ and $\frac{2}{8}$. Using these probabilities, we calculated the amount of uncertainty for this information, $H(A)$, using the equation $H(A) = -\frac{6}{8} \cdot \log_2(\frac{6}{8}) - \frac{2}{8} \cdot \log_2(\frac{2}{8})$, which equals 0.811 bits. With only one location's information at this stage, there is no need to differentiate conditional entropy and joint entropy separately.

In the second process, we analyzed data from both locations A and B. This addition allowed us to calculate the joint entropy of locations A and B, as well as the conditional entropy of B, given A. As depicted in Fig. 4-2, the six individuals from the C1-1 visit condition were further classified into C2-1 (four individuals) and C2-2 (two individuals), resulting in probabilities of $\frac{4}{6}$ and $\frac{2}{6}$, respectively. Conversely, the two individuals in the C1-2 condition were divided into C2-3 and C2-4 conditions, each with one individual, leading to a probability of $\frac{1}{2}$ for each division. This division indicates unique identification for these two individuals.

With the probabilities: $\frac{4}{6}$, $\frac{2}{6}$, $\frac{1}{2}$, and $\frac{1}{2}$, we applied predefined equations for joint entropy and conditional entropy. Specifically, the conditional entropy of B given A is quantified using the simplified equation (5) as follows,

$$H(X|Y) = -\sum_{j=1}^{m} P_j \cdot \sum_{i=1}^{n} \left(P_i \middle| P_j\right) \cdot \log_2(P_i|P_j) \qquad \text{simplified (5)}$$

We changed the variables X and Y to B and A to adopt the above equation to our example. Accordingly, we converted P_j into $P_{A(j)}$ and P_i into $P_{B(i)}$, where $i \; and \; j$ are from 0 to 1, respectively. From Fig. 4-2, we had the following probabilities for locations A and B:

$$P_{A(0)} = \frac{6}{8}, P_{A(1)} = \frac{2}{8},$$

$$P_{(B(0)|A(0))} = \frac{4}{6}, P_{(B(1)|A(0))} = \frac{2}{6}, P_{(B(0)|A(1))} = \frac{1}{2}, and \; , P_{(B(1)|A(1))} = \frac{1}{2}$$

The simplified equation (5) can be converted into the following equation (10), and the conditional entropy of B given A can be calculated with the above probabilities.

$$H(B|A) = -\sum_{j=0}^{1} P_{A(j)} \cdot \sum_{i=0}^{1} \left(P_{B(i)} \middle| P_{A(j)}\right) \cdot \log_2 (P_{B(i)}|P_{A(j)}) \qquad (10)$$

$$= \frac{6}{8} \, H\left(\frac{4}{6}, \frac{2}{6}\right) + \frac{2}{8} \, H\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$= -\frac{6}{8} \cdot \left(\frac{4}{6} \cdot \log_2 \frac{4}{6} + \frac{2}{6} \cdot \log_2 \frac{2}{6}\right) - \frac{2}{8} \cdot \left(\frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{2} \cdot \log_2 \frac{1}{2}\right)$$

$$= 0.939 \; bits$$

This value indicated that we gained 0.939 bits of uncertainty by adding information for location B. As discussed in the previous subsection, our goal is to reach the total uncertainty of 3 bits, the value when all eight individuals are uniquely identified. Therefore, the 0.939 bits represented the *unique* amount of new information contributed by location B, excluding the influence of location A. This method enables us to quantify the unique amount of new information. By integrating the information from location A with that from location B, we can calculate the overall uncertainty of both locations combined. The subtotal of uncertainty from locations A and B is calculated as follows:

$$H(A,B) = H(A) + H(B|A) = 0.811 + 0.939 = 1.75\ bits$$

This equation sums the uncertainty from each location, yielding a comprehensive measure of the total uncertainty for the identification process based on the combined information from both locations.

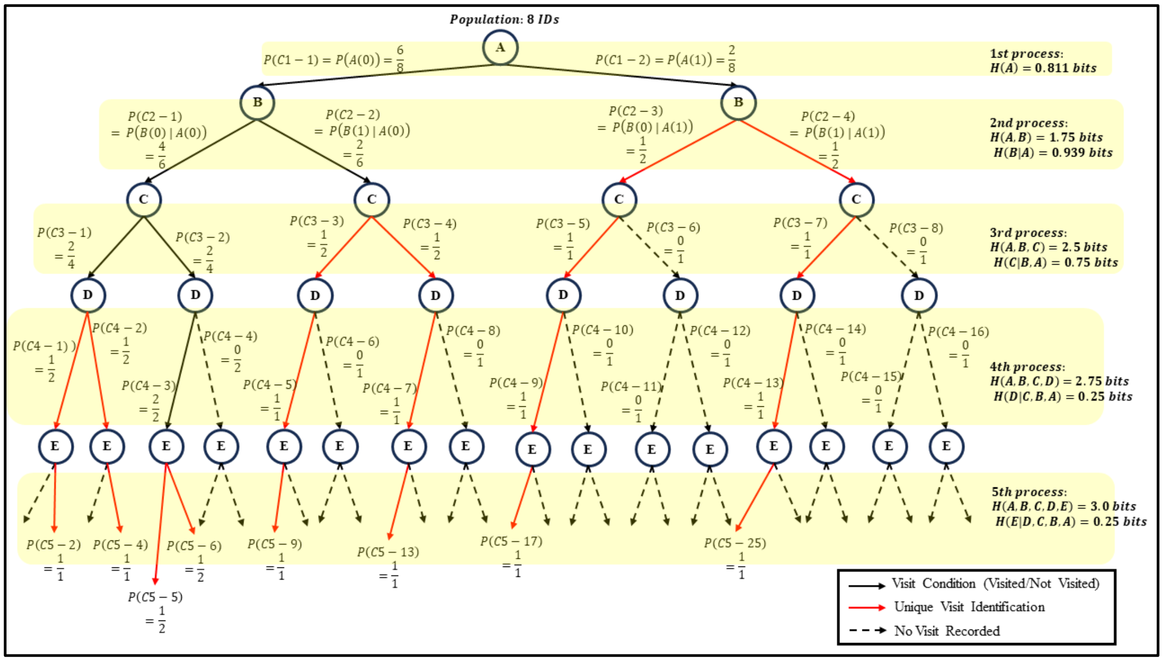


Fig. 4-2 Classification of IDs over incremental location information

3) More Efficient Way to Quantify Uncertainty

We can simplify the calculation of both joint entropy and conditional entropy by using probabilities based on the total populations of eight individuals categorized by visit conditions. In equation (10), we applied the probabilities of location A, ($\frac{6}{8}$ and $\frac{2}{8}$), the marginal distribution of $P_{A(j)}$, as multipliers for the uncertainty values of location B, represented by $H\left(\frac{4}{6},\frac{2}{6}\right)$ and $H\left(\frac{1}{2},\frac{1}{2}\right)$, in the subsequent process.

However, there is a more direct method for this calculation. This method involves directly multiplying the probabilities from the previous process with those of the current process. For example, multiplying $\frac{6}{8}$ (the probability of A(0)) with $\frac{4}{6}$ (the probability of B(0)) results in $\frac{4}{8}$. Applying this method to the other visit conditions allows us to use the total population of 8 as the common denominator. Table 4-4 shows the outcomes of probability calculations for the various visit conditions of locations A and B. In this table, the values $\frac{4}{8}, \frac{2}{8}, \frac{1}{8}$ and $\frac{1}{8}$ represent the outcomes obtained by multiplying the marginal distribution of the previous process (location A) with those of the current process (location B).

Table 4-4. Distribution of individuals with location information A and B

B A	B(0)	B(1)	Marginal distribution of $P_{A(j)}$
A(0)	$\frac{6}{8} \times \frac{4}{6} = \frac{4}{8}$	$\frac{6}{8} \times \frac{2}{6} = \frac{2}{8}$	$\frac{6}{8}$
A(1)	$\frac{2}{8} \times \frac{1}{2} = \frac{1}{8}$	$\frac{2}{8} \times \frac{1}{2} = \frac{1}{8}$	$\frac{2}{8}$
Marginal distribution of $P_{B(j)}$	$\frac{5}{8}$	$\frac{3}{8}$	$\frac{8}{8}$

We can estimate the joint entropy $H(A,B)$ and conditional entropy of B given A, $H(B|A)$, with the values as follows,

$$H(A,B) = -\left(\frac{4}{8} \cdot \log_2 \frac{4}{8} + \frac{2}{8} \cdot \log_2 \frac{2}{8} + \frac{1}{8} \cdot \log_2 \frac{1}{8} + \frac{1}{8} \cdot \log_2 \frac{1}{8}\right) = 1.75 \; bits$$

$$H(B|A) = H(A,B) - H(A) = 1.75 - 0.811 = 0.939 \; bits$$

With this method, we can quantify the uncertainty in a more efficient manner. By simplifying Table 4-3, we focused only on visit conditions where at least one ID has been identified, along with their corresponding probabilities, as shown in Table 4-5. Utilizing Table 4-5, we can estimate the joint entropy and conditional entropy for conditions that incorporate subsequent pieces of information about locations A, B, C, D and E. The calculations proceeded as follows,

Table 4-5. Identification results by varying information pieces and visit conditions

# of info.	Con.	Visit conditions	Probability
3 pieces (Locations A, B, and C)	C3-1	ID 5, 6	2/8
	C3-2	ID 3, 8	2/8
	C3-3	ID 2	1/8
	C3-4	ID 7	1/8
	C3-5	ID 1	1/8
	C3-7	ID 4	1/8
4 pieces (Locations A, B, C and D)	C4-1	ID 6	1/8
	C4-2	ID 5	1/8
	C4-3	ID 3, 8	2/8
	C4-5	ID 2	1/8
	C4-7	ID 7	1/8
4 pieces	C4-9	ID 1	1/8
	C4-13	ID 4	1/8
5 pieces (Locations A, B, C, D and E)	C5-2	ID 6	1/8
	C5-4	ID 5	1/8
	C5-5	ID 3	1/8
	C5-6	ID 8	1/8
	C5-9	ID 2	1/8
	C5-13	ID 7	1/8
	C5-17	ID 1	1/8
	C5-25	ID 4	1/8

- For three pieces of information

$$H(A,B,C) = -\left(2 \cdot \frac{2}{8} \cdot \log_2 \frac{2}{8} + 4 \cdot \frac{1}{8} \cdot \log_2 \frac{1}{8}\right) = 2.5\ bits$$

$$H(C|B,A) = H(A,B,C) - H(A,B) = 2.5 - 1.75 = 0.75\ bits$$

- For four pieces of information

$$H(A,B,C,D) = -\left(\frac{2}{8} \cdot \log_2 \frac{2}{8} + 6 \cdot \frac{1}{8} \cdot \log_2 \frac{1}{8}\right) = 2.75\ bits$$

$$H(D|C,B,A) = H(A,B,C,D) - H(A,B,C) = 2.75 - 2.5 = 0.25\ bits$$

- For five pieces of information

$$H(A,B,C,D,E) = -\left(8 \cdot \frac{1}{8} \cdot \log_2 \frac{1}{8}\right) = 3.0\ bits$$

$$H(E|D,C,B,A) = H(A,B,C,D,E) - H(A,B,C.D) = 3.0 - 2.75 = 0.25\ bits$$

4) Resolved Uncertainty Rate

The previous subsections discussed the methodologies for identifying individuals based on various pieces of information and quantifying the amounts of uncertainty. In this subsection, we focus on measuring the level of privacy risks, termed the *resolved uncertainty rate*, which is based on the concept of normalized entropy. Normalized entropy serves as a measure of privacy risk level. We have established that the total uncertainty in our context is 3 bits, representing the highest level of uncertainty. Normalized entropy is utilized to gauze the magnitude of entropy in a specific condition relative to this highest possible entropy, as indicated in references [16] and [17]. In our example, the resolved uncertainty rate for each process is calculated by dividing the entropy values by the highest of 3 bits. Table 4-6 shows the resolved uncertainty rate corresponding to the number of pieces of information gathered.

Table 4-6. Resolved uncertainty rate by number of information

Number of information	Joint entropy (bits)	Resolved uncertainty rate
1	0.811	27%
2	1.75	58%
3	2.5	83%
4	2.75	92%
5	3.0	100%

By accumulating visit locations in the sequence of A-B-C-D-E, we achieved a 100% resolved uncertainty rate. Assuming that a 90% resolved uncertainty rate poses significant privacy risks, it is evident that only four pieces of information are required to reach this threshold. However, it should be noted that trends in increasing resolved uncertainty rate will vary according to the specific data and its components. The following section explores diverse data scenarios and their impact on the resolved uncertainty rate.

4.5 Experimental Design and Results

4.5.1 Data Description and Experimental Design

1) Data

Our research employs synthetic data, which is structured around four subway stations in a city, each representing a distinct region. This synthetic data was created by considering real-world travel patterns of transit users derived from Smart card data. As a result, the synthetic data accurately reflects realistic travel behaviours and associated privacy risks. These regions are characterized by their unique land uses, which influence the travel patterns of visitors. Region A is an industrial complex with manufacturing plants, warehouses and other industrial facilities. Regions B and C are primarily residential areas, while region D is notable for its university and numerous companies, attracting a diverse mix of students, professionals, and visitors. A single camera is installed at ticket booths at each station to record individual visits, and the recorded video clips are stored. These recordings are made during two distinct time slots over three days: from 5 am to 2 pm (termed the *first time slot*) and from 2 pm to 11 pm (termed the *second* *time slot*), over three days. Consequently, we have twenty-four distinct databases corresponding to the two-time slots across three days at four stations, as illustrated in Fig. 4-3 and Table 4-7.

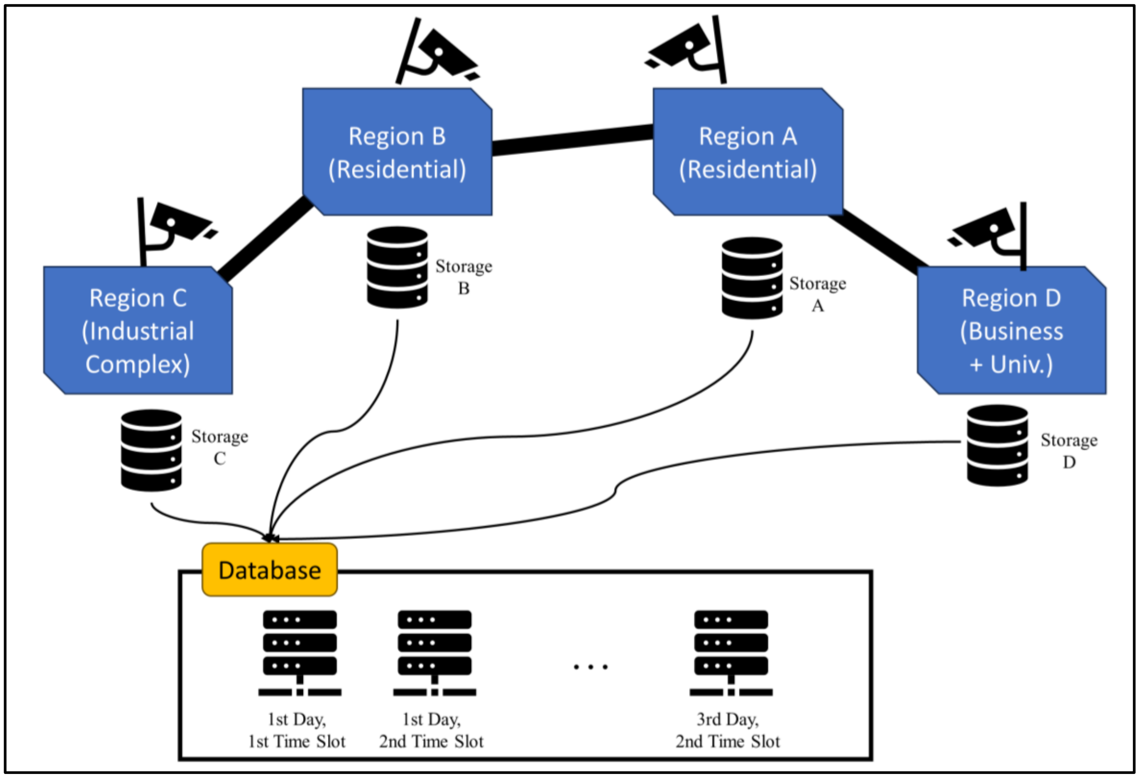


Fig. 4-3 Conceptual diagram of the study area

The travel patterns observed in the four regions vary according to the land uses. Fig. 4-4 illustrates these travel patterns, segmented by time slot and region. For instance, regions A and D exhibit typical business or industrial area travel patterns, with two peak demand periods. In region D, the majority of visitors arrive during the first time slot (before 2 pm) for work (48%), and depart in the second time slot (after 2 pm) to return to their homes (40%). Region A also displays similar travel patterns, with two peak periods of demand accounting for 37% and 30%, respectively. Conversely, regions B and C, which are residential areas, show a different pattern. Most visitors use the subway during the first time slot to leave for their day's activities and return to these regions in the evening.

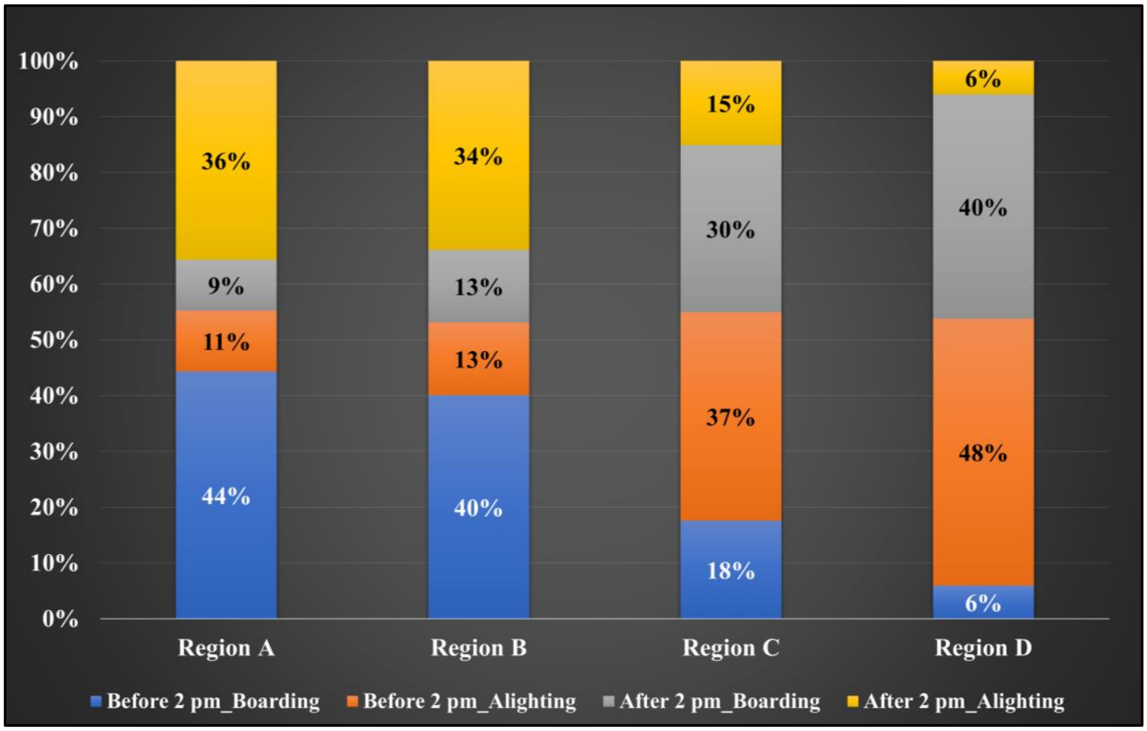


Fig. 4-4 Composition of recorded travel patterns by time slot and region

2) Experimental Design

Our study involves four different regions where video clips were recorded. We considered a hypothetical situation where an individual is reported missing at the end of the third day. In response, police officers were assigned to investigate this case, utilizing the recorded data to track the missing person. Due to the absence of facial recognition technology, the officers were required to review all the recordings manually. This task involves analyzing data from 24 separate databases to identify the individual. After an extensive investigation, the officers successfully recognized 212 transit users and their travel records across the 24 distinct databases. For the purpose of our research, we refer to these 24 databases as *cubes*, as described in Table 4-7.

Table 4-7. Cubes by region and time

Region	1st day		2nd day		3rd day	
	Before 2 pm	After 2 pm	Before 2 pm	After 2 pm	Before 2 pm	After 2 pm
A	Cube 1	Cube 2	Cube 9	Cube 10	Cube 17	Cube 18
B	Cube 3	Cube 4	Cube 11	Cube 12	Cube 19	Cube 20
C	Cube 5	Cube 6	Cube 13	Cube 14	Cube 21	Cube 22
D	Cube 7	Cube 8	Cube 15	Cube 16	Cube 23	Cube 24

In our database, the travel records for all 212 transit users across the 24 cubes are recorded based on their visits. For each individual, a '1' is recorded if they visited a cube and a '0' if they did not. This binary method efficiently tracks each individual's visits to the cubes, providing a complete dataset for our analysis. As outlined in the methods of analysis section, this study conducted multiple analyses using the travel records of 212 individuals across 24 cubes. The analyses consist of estimating total uncertainty, identifying individuals based on their visit conditions, calculating joint entropy and conditional entropy from the identification results, and estimating resolved uncertainty rates.

The 24 cubes led to a vast number of combinations for providing information. For instance, if police officers need to review just one video clip, they can select from any of the 24 cubes. This selection process would be repeated 24 times, once for each cube to cover all cubes. With the two pieces of information, the number of potential combinations for cube data review increases. This is calculated as 24 choose 2 ($_{24}C_2$), resulting in 276 combinations, examples being Cube1-2, Cube 1-3, …, and Cube 23-24.

Considering 12 pieces of information out of 24 pieces, which was half of the entire piece, increases the number of combinations extremely. For example, combinations like Cube 1-2-3-4-5-6-7-8-9-10-11-12 or Cube 12-13-14-15-16-17-18-19-20-21-22-23 are possible. Using the combinatorial formula for choosing 12 items from a set of 24 (denoted as 24 choose 12 or ${}_{24}C_{12}$), we find that the total number of these combinations reaches 2,704,156. Therefore, when we sum up all possible combinations from 1 to 24 pieces of information – ${}_{24}C_{1} + {}_{24}C_{2} + {}_{24}C_{3} + \ldots + {}_{24}C_{23} + {}_{24}C_{24}$ – the aggregate figure reaches approximately 16.7 million combinations.

Given the extensive number of combinations, it is challenging to demonstrate the identification process with tables and figures like Table 4-3, Table 4-5 and Fig. 4-2. Accordingly, this section primarily focused on showing the results using the same methods from the previous section. These included the identification process, uncertainty quantification, joint entropy and conditional entropy through the more efficient approach, and the calculation of the resolved uncertainty rate.

4.5.2 Amount of Total Uncertainty and Joint Entropy

1) Total Uncertainty

The first step in our analysis involved calculating the total uncertainty based on the number of individuals involved. We uniquely identified each of the 212 individuals by utilizing the travel records. In our view, uniquely identifying all 212 individuals is significant, as it presents the most hazardous privacy risks. The probability of uniquely identifying an individual is inversely proportional to the number of individuals. Therefore, in our case with 212 individuals, the most hazardous probability is calculated as $\frac{1}{212}$. Using this probability, we determined the total entropy with the formula $-212 \times \frac{1}{212} \times \log_2 \frac{1}{212}$, resulting in approximately 7.728 bits

2) Joint Entropy and Resolved Uncertainty Rate by Information Piece

Then, this study estimated the joint entropy to measure the fluctuation of privacy risks based on the number of information pieces from 1 to 24 cubes. Table 4-8 depicts these fluctuations of joint entropy corresponding to the number of information sources. As more information about visit locations was added, we observed that the joint entropy values approached 7.728 bits, equating to the total uncertainty. We found that even though we had the same number of information, the amount of uncertainty varied significantly. From the methods section, we assumed that 90% of the resolved uncertainty rate poses significant privacy risks. When analyzing the resolved uncertainty rate with the 10 pieces of information, we found that certain combinations of cubes reached approximately 90%, while others showed only about 65%. Our findings indicate that, in terms of the minimum resolved uncertainty rate, at least 19 pieces of information were needed to reach the 90% threshold.

Table 4-8. Resolved uncertainty rate by number of information

No of info	Minimum uncertainty (entropy)	Maximum uncertainty (entropy)	Minimum resolved uncertainty rate	Maximum resolved uncertainty rate
1	0.804	0.986	10.4%	12.8%
2	1.422	1.951	18.4%	25.3%
3	1.873	2.862	24.2%	37.0%
4	2.264	3.712	29.3%	48.0%
5	2.605	4.506	33.7%	58.3%
6	2.995	5.216	38.7%	67.5%
7	3.772	5.780	48.8%	74.8%
8	4.291	6.264	55.5%	81.1%
9	4.594	6.609	59.4%	85.5%
10	5.021	6.912	**65.0%**	**89.4%**
11	5.299	7.133	68.6%	92.3%
12	5.392	7.286	69.8%	94.3%
13	5.674	7.390	73.4%	95.6%
14	5.935	7.479	76.8%	96.8%
15	6.206	7.545	80.3%	97.6%
16	6.425	7.611	83.1%	98.5%
17	6.645	7.662	86.0%	99.1%
18	6.849	7.709	88.6%	99.8%
19	7.108	7.718	**92.0%**	99.9%
20	7.172	7.728	92.8%	**100.0%**
21	7.327	7.728	94.8%	100.0%
22	7.519	7.728	97.3%	100.0%
23	7.690	7.728	99.5%	100.0%
24	7.728	7.728	**100.0%**	100.0%

Additionally, Fig. 4-5 provides a more detailed view of the fluctuations in uncertainty across different numbers of information. In this figure, we employed a box plot to illustrate the variability in uncertainty amount for the same number of information pieces and to show the trend of approaching the total uncertainty. We want to clarify that while outliers in a box plot generally indicate extreme cases or potential data errors, in our study, these outliers represent unique user travel patterns rather than errors or extreme values.

When analyzing scenarios with fewer pieces of information, we observed that the interquartile range of the box pots was relatively narrow, and outliers were limited in their spread. This can be linked to the limited diversity in the information available with a smaller number of information pieces. As the number of information pieces increases, reaching a peak at 12 – where the diversity is maximum – both the interquartile range and the outliers in the box plots expand noticeably. However, as we approach 24 pieces of information, the diversity begins to decrease. The enlargement of the interquartile range and outliers can be attributed to the increased diversity of travel patterns. As the number of information pieces increases, the entropy value eventually reaches the total uncertainty of 7.778 bits. This level of uncertainty is achieved regardless of the diversity in the data, indicating the point of highest privacy risk. The most hazardous privacy risk occurs when the entropy value equals the total uncertainty. Further analysis can expand on this observation by illustrating the trends between entropy and the number of reidentified individuals.

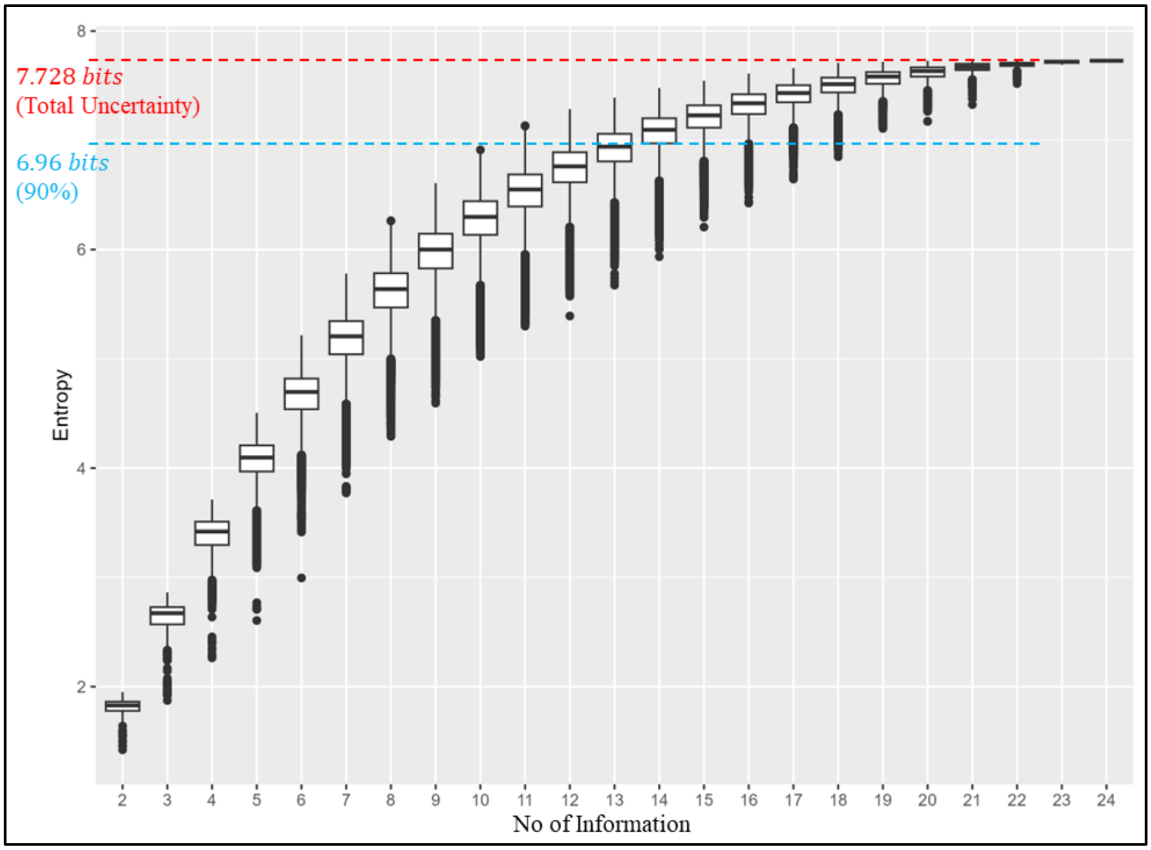


Fig. 4-5 Entropy fluctuation over the number of observations

4.5.3 Anonymity Values and Entropy

1) Anonymity Values

This section explored the relationship between entropy and the *anonymity value*, a concept introduced in Chapter 3 regarding the privacy risks associated with COVID-19 patient data. The anonymity value refers to the count of individuals sharing identical travel records corresponding to the visit conditions across the pieces of information. Table 4-9 indicates a part of the data, focusing on two pieces of information and visit conditions, which are defined as *visited (yes)* or *not visited (no)*, along with the corresponding anonymity value, denoted as (n).

We note that even with an equal number of information pieces, there are noticeable variations in entropy and anonymity values across the visit conditions. For analytical purposes, this study grouped these varying anonymity values into six categories: n = 1, 1<n≤5, 5<n≤10, 10<n≤20, 20<n≤50, 50<n≤100, and 100<n≤212. Particularly, an anonymity value of n=1 indicates the re-identification of a single unique individual from the combinations of information. When only one individual's data is distinguishable, it is considered the most hazardous regarding privacy. This situation completely contrasts with the principle of *k*-anonymity, where *k* represents the threshold number of individuals required in a dataset to obscure individual identities effectively. Under *k*-anonymity, each dataset should be indistinguishable from at least *k-1* others, ensuring anonymity. Hence, when this threshold is markedly reduced to 1, as with n=1, the risk to privacy significantly increases, marking it as the most vulnerable scenario for individual re-identification. In this research, we assumed that the anonymity value of fewer than 10 individuals is dangerous because the small number of remaining individuals can be easily reidentified.

Table 4-9. Example of the anonymity values under visit conditions with two pieces of information

No of info	Information combination	Entropy	Visit conditions	Anonymity values	Category
2	Cube1 – Cube 10	1.677	No (0) – Yes (1)	29	20<n≤50
2	Cube1 – Cube 10	1.677	Yes (1) – Yes (1)	41	20<n≤50
2	Cube1 – Cube 10	1.677	No (0) – No (0)	118	100<n≤212
2	Cube1 – Cube 10	1.677	Yes (1) – No (0)	24	20<n≤50
2	Cube1 – Cube 11	1.823	No (0) – No (0)	78	50<n≤100

2	Cube1 – Cube 11	1.823	Yes (1) – No (0)	49	20<n≤50
2	Cube1 – Cube 11	1.823	No (0) – Yes (1)	69	50<n≤100
2	Cube1 – Cube 11	1.823	Yes (1) – Yes (1)	16	10<n≤20
…					
2	Cube 22 – Cube 24	1.822	Yes (1) – No (0)	41	20<n≤50
2	Cube 22 – Cube 24	1.822	No (0) – No (0)	90	50<n≤100
2	Cube 22 – Cube 24	1.822	Yes (1) – Yes (1)	20	10<n≤20
2	Cube 22 – Cube 24	1.822	No (0) – Yes (1)	61	50<n≤100
2	Cube 23 – Cube 24	1.878	No (0) – No (0)	91	50<n≤100
2	Cube 23 – Cube 24	1.878	Yes (1) – Yes (1)	49	20<n≤50
2	Cube 23 – Cube 24	1.878	Yes (1) – No (0)	40	20<n≤50
2	Cube 23 – Cube 24	1.878	No (0) – Yes (1)	32	20<n≤50
2	**Total**			**1104**	

The previous analyses in subsection 4.5.2 demonstrated fluctuations in entropy and anonymity values across different numbers of information pieces. We discussed that using 10 pieces of information can reach 90% resolved uncertainty. For further analysis, this study analyzed the composition percentages by the anonymity values category across various numbers of information pieces.

To estimate these percentages, we divided the count of each anonymity values category by the total number of anonymity values for each set of information pieces. More specifically, for the two pieces of information, the total number of anonymity values was 1104, categorized as follows: 5<n≤10 with 1 individual, 10<n≤20 with 173

individuals, 20<n≤50 with 347 individuals, 50<n≤100 with 534 individuals, and 100<n≤212 with 49 individuals. The corresponding percentages were 0.1%, 15.7%, 31.4%, 48.4% and 4.4% respectively.

Fig. 4-6 describes the composition percentages of the anonymity value categories for 2 pieces of information to 24 pieces. As the number of information pieces increased, the percentage of the lower-risk anonymity values, such as 50<n≤100 and 100<n≤212, decreased and eventually vanished. On the other hand, the percentages for n≤10 categories increased. Previously, we discussed that with 10 pieces of information, 90% of resolved uncertainty was first reached. Fig. 4-6 details this with the following composition: n=1 with 54%, 1<n≤5 with 41%, 5<n≤10 with 4%, and other categories contributing less than 1%. Accordingly, we found that approximately 99% of individuals faced significant privacy risks as they were reidentified in groups of less than 10 individuals when considering 10 pieces of information.

This approach allowed us to illustrate how different numbers of information pieces contribute to privacy risks using a united measurement of percentages. We believe that using composition percentages provides a more effective way to understand the impacts of varying numbers of information pieces on privacy risks compared to relying solely on varying anonymity values.

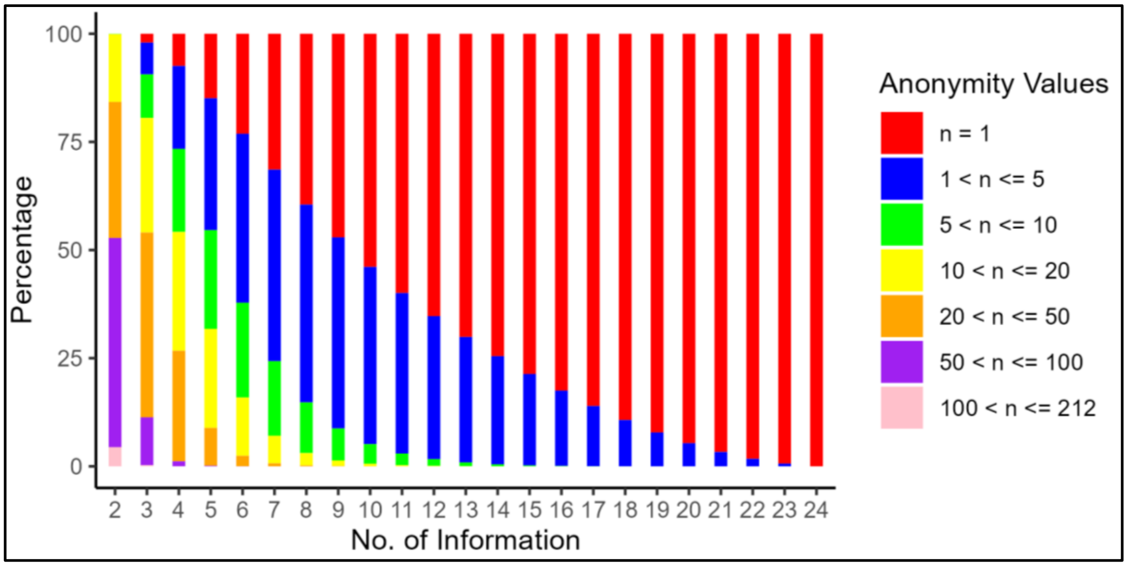


Fig. 4-6 Composition percentage of the anonymity values category

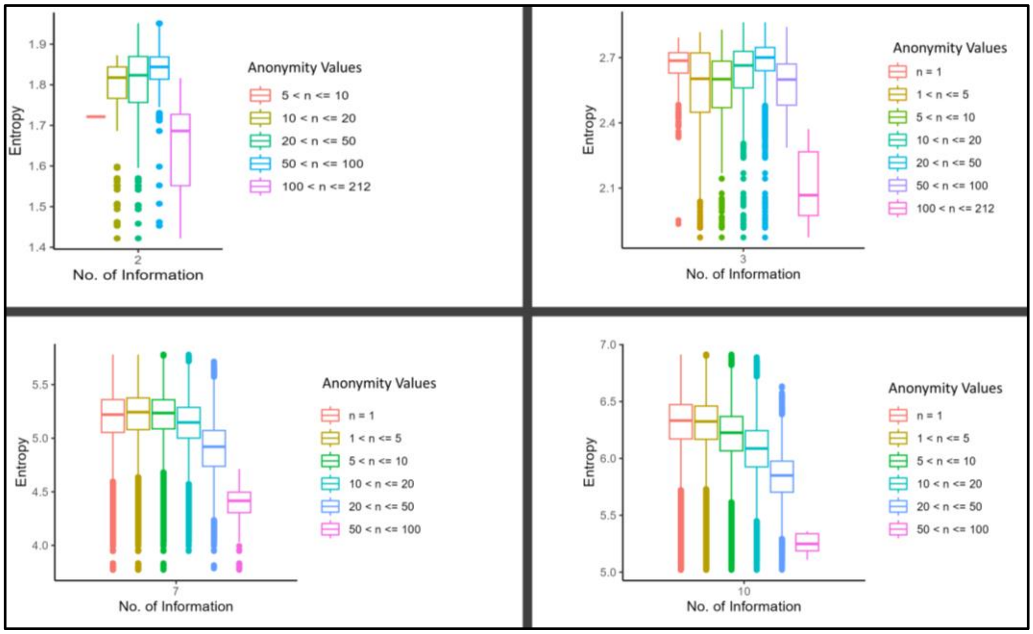
2) Relationship Between Anonymity Values and Entropy

Lastly, this study explored the relationship between the categories of anonymity values and entropy across varying numbers of information pieces. Fig. 4-7 displayed box plots that visualize this relationship, and the plot corresponding to two pieces of information is located in the upper left corner of Fig. 4-7 (a). Considering the median values in these box plots of the two information pieces, categories such as 50<n≤100, 20<n≤50, and 10<n≤20 exhibited higher entropy values compared to other categories: 5<n≤10 and 100<n≤212.

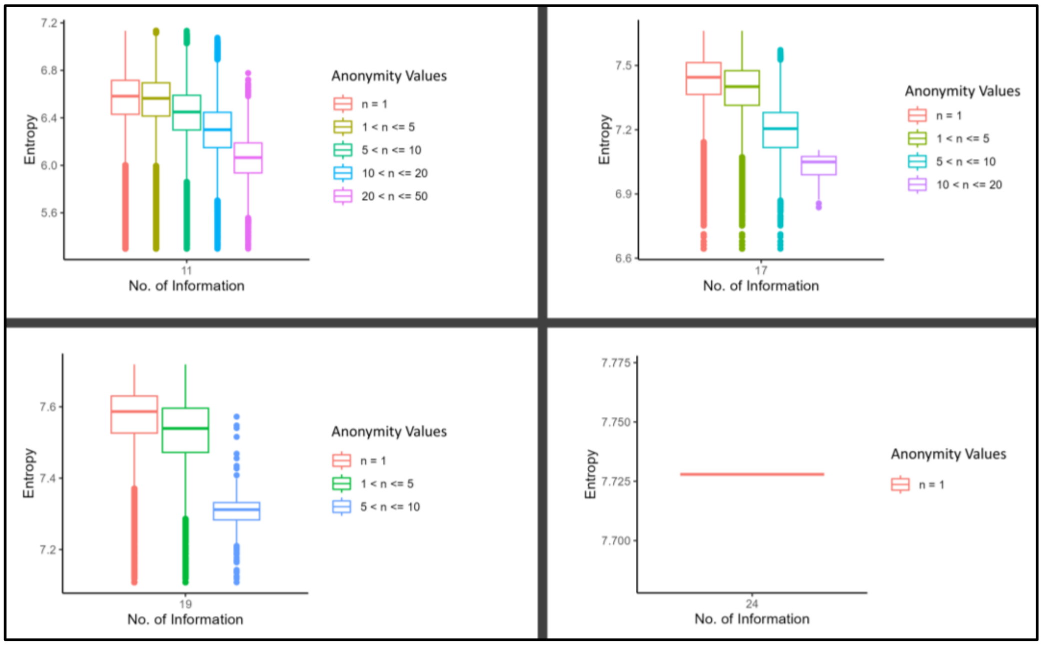
When focusing on three information pieces, as shown in the upper right corner of Fig. 4-7 (a), the median entropy values for all categories tended to converge at around 2.7 bits. Notably, the n=1 category, indicating the highest privacy risk, exhibited the highest entropy values. In contrast, the category with the least privacy risk, 100<n≤212, showed the lowest entropy values. A similar pattern was observed with seven and ten information pieces: the highest risk category, n=1, consistently showed the highest entropy values, while lower risk categories, such as 50<n≤100 or 100<n≤212, showed the lowest entropy

values. This pattern was consistent across information pieces from 3 to 24, indicating that higher privacy risks were associated with increased entropy values.

Detailed box plots for each individual number of information pieces are included in Appendix B.



(a) For numbers of information: 2, 3, 7, and 10 pieces



(b) For numbers of information: 11, 17, 19, and 24 pieces

Fig. 4-7 Box plots across the number of information pieces and anonymity values

4.5.4 Amount of Information Considering Order of Acquisition

This subsection examines information quantity with a particular focus on the order of acquisition and the value of newly acquired information concerning the existing data set. As previously mentioned in Section 4.1, the value of information can significantly vary based on the circumstances and the sequence of its collection. This study aims to quantify the value of new information, considering how it integrates with and complements previously collected data. For instance, if information about Cube 11 is gathered fifth in a sequence, its value is influenced by the nature of the data collected in the preceding steps. If Cube 11's information is distinct from the earlier data, it is considered more valuable due to its novelty. Conversely, if it largely overlaps with the existing data, the incremental value of Cube 11's information could be minimal.

One of the key objectives of this study was to quantify the variation in the amount of newly added information. For this purpose, we analyzed patterns by calculating the fluctuation of conditional entropy, utilizing a more efficient method described in the methods of analysis section. By employing the conditional entropy method, we specifically focused on estimating the *novel* value of each new cube's information, considering the extent of its overlap with previously obtained data. This approach enabled us to effectively quantify the amount of novel information for different sequences of information pieces.

To achieve this, we focused on scenarios that involve acquiring 10 distinct pieces of cube information. As shown in Table 4-8, when considering 10 cube information pieces, the resolved uncertainty rate reached approximately 90% for the first time. Based on this, we selected 12 specific combinations of cube information acquisition from these scenarios. Table 4-10 presents these 12 combinations. For instance, the first combination, termed Com1, consists of visiting cubes in the following sequence: Cube 1-2-4-7-11-12-14-18-23-24.

Table 4-10. Combinations of visiting cubes based on information acquisition order

Acquisition order Combination	Cube information acquisition order									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Com1	Cube 1	2	4	7	11	12	14	18	23	24
Com2	1	2	4	7	11	12	16	18	23	24
Com3	1	2	4	7	11	12	16	21	23	24
Com4	1	2	4	7	11	12	16	22	23	24
Com5	1	2	4	7	11	14	16	18	23	24
Com6	1	3	4	8	11	12	14	18	23	24
Com7	1	7	8	11	12	16	18	19	20	23
Com8	1	7	8	11	12	16	18	20	21	23
Com9	2	3	4	7	11	12	16	18	23	24
Com10	2	4	5	7	11	12	16	18	23	24
Com11	2	4	7	11	12	16	18	21	23	24
Com12	3	4	7	8	9	11	12	18	23	Cube 24

Colour highlights indicate the same information

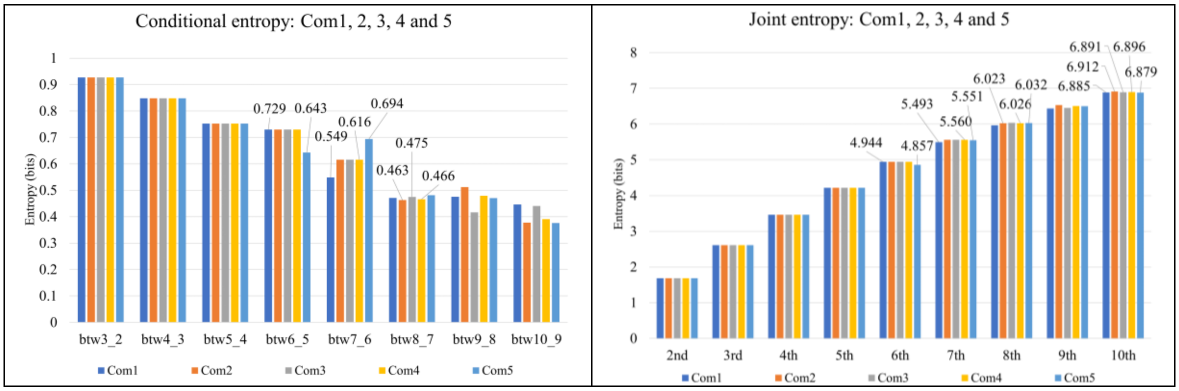
In our analysis, we employed colour highlights to group visiting cube information, facilitating the comparison of different combinations and their respective information acquisition patterns. Combinations 1, 2, 3, 4 and 5 (Com1 to Com5) displayed almost identical sequences. However, starting from the 6th piece of information, these combinations began to differ. For instance, in the 6th piece, Combination 5 (Com5) included information from Cube 14, whereas the other combinations used information from Cube 12. This difference resulted in different conditional entropy values between the 5th and 6th pieces. In Fig. 4-8 (a), Com5 had

0.643 bits, and the others had 0.729 bits, respectively. In this case, the value of the 6th piece of information, which involved Cube 12, was greater than that for Cube 14, given the previous sequence of Cubes 1-2-4-7-11.

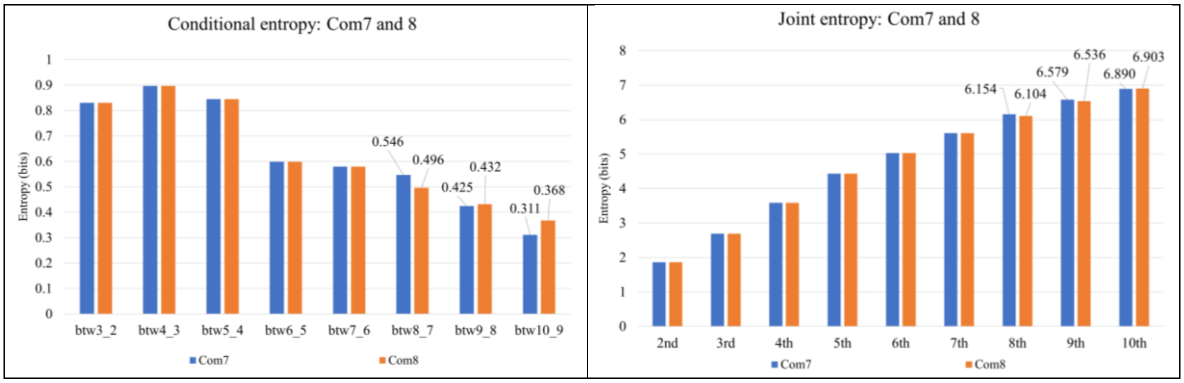
Moving to the 7th piece of information, Combination 1 (Com1) contained Cube 14, while the other combinations chose Cube 16. This choice led to conditional entropy values of 0.549, 0.616 and 0.694 bits. Notably, Com5 recorded the highest conditional entropy for the 7th piece, in contrast to having the lowest value for the 6th piece. This observation suggests that the value of information is significantly influenced by the sequence in which it is acquired. The fluctuation in the amount of information continued through to the 10 pieces of information, with the joint entropy estimation eventually converging to approximately 6.8 bits.

Similarly, in Fig. 4-8 (b), Combinations 7 and 8 exhibited almost identical patterns, with the only difference being in the 8th and 9th pieces of information. When analyzing the conditional entropy between the 7th and 8th pieces, which involved Cube 19 and 20, respectively, we found it to be 0.546 bits for Cube 19 and 0.496 bits for Cube 20. This indicated that the information amount from Cube 19 was slightly greater than that from Cube 20. Despite having the same 10th piece of information, Cube 23, the differences in the 8th and 9th pieces resulted in different conditional entropies for the 10th piece, with both combinations converging to 6.9 bits.

This analysis enabled us to quantify the amount of each piece of information by considering the different sequences in which they were acquired. Employing this method allows for understanding the novel amount of information each new cube adds. It highlights how the sequence of information acquisition can significantly impact the overall value of the information.



(a) For Combinations 1, 2, 3, 4, and 5



(b) For Combinations 7 and 8

Fig. 4-8 Joint and conditional entropy by information acquisition pattern

4.7 Conclusion

The value of information is inherently variable, influenced by the recipient's interests, circumstances, and timing. Quantifying this value, particularly in terms of newly added pieces of information, presents notable challenges. Travel data contains many sensitive details. When these details are reidentified, they pose a significant risk to privacy. Therefore, accurately measuring the levels of privacy risk in travel data is crucial to protect this valuable information.

In this study, we investigated methods to quantify the amounts of a piece of information from travel data in relation to privacy risk. Initially, we conducted an identification process by adding pieces of information and uniquely identifying 212 individuals based on various combinations of obtained information.

We then applied joint entropy from information theory to quantify the amounts of information concerning the identification process. A key metric we introduced was the resolved uncertainty rate, which served as an indicator of privacy risk levels associated with the added information. Our findings revealed that with just 10 pieces of information, the resolved uncertainty rate reached a 90% level, indicating a significant privacy risk. These 10 pieces of information were less than half of the total 24 pieces. It is critical to note that certain combinations of these pieces can pose greater privacy risks, even when they make up less than 50% of the total information piece. While these results are specific to our dataset, privacy experts should be cautious in assuming that using only a small proportion of the total information pieces is safe.

Furthermore, we introduced the concept of anonymity value to denote the count of individuals who have identical visit records in each travel information combination, highlighting the re-identification risk. It was observed that as more pieces of information were added, both the entropy values and privacy risks proportionally increased.

Additionally, we employed the conditional entropy method to quantify the amount of each newly added piece of information. By analyzing how conditional entropy varied across different combinations of travel information, especially those with almost identical records, we were able to discern the novel value each new piece of information brought to the dataset.

Although this study introduced an information-theoretic approach to quantifying amounts of information from travel records, further developments are necessary. We utilized synthetic data that reflected real-world travel patterns across 24 travel combination cases to demonstrate our methodology. A broader and larger transportation-related dataset, such as smart card and location-based service data, should be analyzed to expand our approach. Additionally, examining how the amount of information fluctuates across different user clusters can provide valuable insights into the relationship between individual travel patterns and associated privacy risks.

Lastly, this dissertation did not consider certain people’s preferences regarding using non-digital or cash transactions to hide their travel records. The consideration of non-digital or cash transactions as a factor in privacy risk quantification certainly presents a significant area for future research. Many individuals prefer to use cash to maintain privacy in their movements, making this a critical variable in quantifying the risk. However, the primary goal of this dissertation was to establish foundational methodologies for quantifying privacy risks based on information theory. Therefore, exploring non-digital payment methods as a variable in privacy risk quantification is a promising direction for future research.

4.8 Chapter 4 References

[1] L. Sun, K. W. Axhausen, D. Lee, and X. Huang, “Understanding metropolitan patterns of daily encounters,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 34, pp. 13774–13779, Aug. 2013, doi: 10.1073/pnas.1306440110.

[2] E. W. T. Ngai and A. Gunasekaran, “A review for mobile commerce research and applications,” *Decis. Support Syst.*, vol. 43, no. 1, pp. 3–15, Feb. 2007, doi: 10.1016/j.dss.2005.05.003.

[3] S. R. Kelleher, “Why you should never share your boarding pass on social media,” *Forbes*, Aug. 03, 2023. [Online]. Available: https://www.forbes.com/sites/suzannerowankelleher/2023/08/03/never-share-boarding-pass-social-media/?sh=3ca003772c38

[4] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 05, pp. 557–570, Oct. 2002, doi: 10.1142/S0218488502001648.

[5] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. “Sandy” Pentland, “Unique in the shopping mall: On the reidentifiability of credit card metadata,” *Science (80-.).*, vol. 347, no. 6221, pp. 536–539, Jan. 2015, doi: 10.1126/science.1256297.

[6] J. Gao, L. Sun, and M. Cai, “Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data,” *Transp. Res. Part C Emerg. Technol.*, vol. 104, pp. 78–94, Jul. 2019, doi: 10.1016/j.trc.2019.04.022.

[7] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–656, Oct. 1948, doi: 10.1002/j.1538-7305.1948.tb00917.x.

[8] A. Lawrence, “Information, Uncertainty, and Surprise,” in *Probability in Physics*, Springer, Cham, 2019, pp. 209–231. doi: 10.1007/978-3-030-04544-9_11.

[9] A. Ben-Naim, *A Farewell to Entropy*, vol. 6, no. August. World Scientific, 2008. doi: 10.1142/6469.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2005. doi: 10.1002/047174882X.

[11] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,” in *2007 IEEE 23rd International Conference on Data Engineering*, Apr. 2007, pp. 106–115. doi: 10.1109/ICDE.2007.367856.

[12] T. Dalenius, “Finding a needle in a haystack or identifying anonymous census records,” *J. Off. Stat.*, vol. 2, no. 3, pp. 329–336, 1986, [Online]. Available: https://www.proquest.com/scholarly-journals/finding-needle-haystack-identifying-anonymous/docview/1266806751/se-2

[13] J. F. Wellmann and K. Regenauer-Lieb, “Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models,” *Tectonophysics*, vol. 526–529, pp. 207–216, Mar. 2012, doi: 10.1016/j.tecto.2011.05.001.

[14] X. Zhan, C. Lu, and G. Hu, “A Formal Representation of the Semantics of Structural Geological Models,” *Sci. Program.*, vol. 2022, pp. 1–18, Jan. 2022, doi: 10.1155/2022/5553774.

[15] J. Wellmann, “Information Theory for Correlation Analysis and Estimation of Uncertainty Reduction in Maps and Models,” *Entropy*, vol. 15, no. 12, pp. 1464–1485, Apr. 2013, doi: 10.3390/e15041464.

[16] I. Brito, “The normalized expected utility – entropy and variance model for decisions under risk,”

Int. J. Approx. Reason., vol. 148, pp. 174–201, Sep. 2022, doi: 10.1016/j.ijar.2022.06.005.

[17] J. Vang, *Data Science Topics*. One-Off Coder, 2019. [Online]. Available: https://datascience.oneoffcoder.com/normalized-entropy-mi.html

CHAPTER 5

CONCLUSIONS

5.1 Research findings and contributions

This dissertation has explored two principal areas: cybersecurity threats targeting AVs and privacy breach risks associated with data sets in relation to ITS. This section summarized the findings and conclusions in light of the research questions and objectives discussed in Chapter 1.

5.1.1 Understanding A Novel Type of Malware Attack

Our focus was on AVs' most vulnerable moments - during scanning for repairs and maintenance in mechanic shops. We modified the well-developed, mathematically-based Susceptible-Infected (SI) epidemic model, drawing parallels with the malaria spreading model. By adopting this vector-host model, we introduced a novel approach to understanding malware propagation in the context of AV cybersecurity.

Furthermore, we utilized the transportation gravity model to estimate the speed of the spread of malware. Due to the lack of real-world data on visits to mechanic shops, we employed the gravity model as an alternative approach. Our study enriched the gravity model by incorporating various factors, such as the probability of malware spread and vehicle owners’ preferences for certain mechanic shops.

This comprehensive analysis contributes to a deeper understanding of novel malware propagation among AVs and builds a groundwork to prepare for future potential attacks through CAN-bus communication.

5.1.2 Exploring Privacy Issues Using Real-world Privacy Breach Case

This dissertation examined privacy risks in mobility data, particularly focusing on the effects of language-based external observations from the COVID-19 patient travel logs. Our findings indicated a relationship between the number of travel observations, varying spatial/temporal resolutions, and the risk of re-identification. Another key finding of our research was that various visited locations can increase the likelihood of re-identification.

This research contributed to illuminating the significant impact of language-based external observations on privacy risks. Google Map reviews and social networking stories are prime examples of language-based observations. These observations include details about where and when events happen but also contain spatial and temporal resolutions, which can affect privacy risk levels. Research into these types of observations lays essential groundwork for developing strategies to protect data privacy, especially when data publication is necessary.

5.1.3 Quantifying Privacy Risks Based on Information Theory

In this part of the dissertation, we delve into our core research question: how can we quantify the amount of information related to privacy risks? We utilized methodologies based on information theory. We illustrated how resolved uncertainty values can mirror privacy risks and applied concepts of joint entropy and conditional entropy to quantify the amount of information from newly added pieces of information.

The main contribution of this research is to reduce the research gap by developing an information theory-based methodology that quantifies the amount of information in transportation datasets. This approach highlights the importance of considering privacy risks from the perspective of the entire dataset rather than solely from selected or filtered subsets, thereby offering a more comprehensive understanding of privacy research in the transportation data sector.

5.2 Research Limitations and Future Plans

Future research endeavours can expand upon our current insights and methodologies, addressing various aspects of cybersecurity and privacy risks in ITS.

5.2.1 For Cybersecurity Research with Diverse Methods of Spreading Malware

Our study primarily focused on malware propagation through physical contact between AVs and OBD-II scanners during maintenance and repair. However, it is crucial to consider both remote and direct attack methods, as they could lead to substantial damage regardless of their current prevalence. We hypothesized that Stuxnet malware could spread across all vehicle brands and models via universal communication through the CAN bus. Future research could explore different malware propagation rates, factoring in the varied usage of OBD-II scanners by vehicle model.

Given the innovative nature of our research, we could not compare our findings with other models of malware propagation in AVs. The cybersecurity field for AVs is uncertain due to its unique environment. Future studies could compare propagation speeds and success rates of different attacks against various security measures in AVs. Additionally, incorporating real-world data from mechanic shop visits would provide a stronger foundation for this research.

5.2.2 For language-based external observation privacy research

One limitation of our study stemmed from government policies. For instance, the Korean government, acknowledging potential privacy breaches, concealed sensitive personal data. This policy shift meant that our research was based on a comparatively smaller dataset, limiting our ability to review risks thoroughly over extended periods and across diverse travel patterns.

Expanding this study to include privacy breach cases of COVID-19 patients from other countries could have provided a more comprehensive analysis. As the world transitions into the endemic phase of COVID-19, broadening our research scope to incorporate global data could yield valuable insights for academic research and practical data privacy protection.

5.2.3 For Information Theory-based Quantifying Privacy Risks

This dissertation utilized information theory-based methodologies to quantify information in transportation-related data, but there were some limitations. Our analysis used synthetic data modelled on real-world travel patterns. This approach was crucial in helping us understand how to measure the uncertainty resolution. However, employing larger and more diverse transportation-related datasets, such as extensive smart card big data, could provide deeper insights.

In our study, we primarily focused on information about locations and times, which we represented as cubes. Future research could improve by clustering individuals based on their travel patterns. This method would provide a clearer understanding of how different travel behaviours relate to privacy risks.

5.2.4 For Inter-chapter Research Opportunities

This dissertation explores two intertwined pillars within the realm of ITS: cybersecurity and privacy. These closely intertwined areas demonstrate that a cyber attack can lead to both substantial physical damages and privacy breaches, as recently witnessed in hospitals in Southern Ontario, Canada [1].

Chapter 2 investigates malware propagation through infected OBD-II scanners and computers at mechanic shops. Chapter 3 presents real-world cases of privacy breaches and their impacts on the privacy risks associated with transportation smart card databases. Chapter 4 develops a methodology to quantify privacy risks from databases. The interconnected nature of these chapters underscores the potential for future research.

Mechanic shops often hold sensitive customer information, including addresses, telephone numbers, and license plate numbers. In addition, vehicles equipped with embedded navigation systems and dash cameras add another layer of risk. These factors present a significant security concern. Hackers could potentially access and misuse customers' travel records and personal conversations, linking them to criminal activities. Future research, therefore, can simultaneously address the intertwined concerns of cybersecurity and privacy in ITS, thus enhancing our understanding of these critical areas.

5.3 Chapter 5 References

[1] J. La Grassa, “CEOs of Ontario hospitals hit by ransomware attack break down impact on operations , patients,” *CBC News*, Windsor, ON, Oct. 23, 2023. [Online]. Available: https://www.cbc.ca/news/canada/windsor/hospitals-southwestern-ontario-ceo-ransomware-attack-1.7031544

APPENDICES

Appendix A. Travel Records for Chapter 4

ID	Cube1	Cube2	Cube3	Cube4	Cube5	Cube6	Cube7	Cube8	Cube9	Cube10	Cube11	Cube12
ID1	0	0	0	0	0	0	0	0	1	1	0	0
ID2	1	0	1	0	0	0	0	0	0	1	0	1
ID3	0	0	0	0	1	0	1	0	0	0	0	0
ID4	0	0	0	0	1	1	1	1	0	0	0	0
ID5	0	0	1	1	0	0	1	1	0	0	1	0
ID6	1	1	1	1	0	0	0	0	1	1	1	1
ID7	0	0	0	0	1	1	1	1	0	0	0	0
ID8	0	0	1	0	0	0	1	0	0	0	1	1
ID9	1	1	0	0	0	0	1	1	1	1	0	0
ID10	0	0	1	0	1	0	0	0	0	0	1	1
ID11	0	1	0	0	0	0	0	1	1	0	0	0
ID12	0	0	1	0	0	0	1	0	0	0	1	1
ID13	0	0	1	1	1	1	0	0	0	0	1	0
ID14	1	1	0	0	1	1	0	0	1	1	0	0
ID15	0	0	0	0	0	0	0	0	0	0	1	1
ID16	0	1	0	0	0	0	0	1	0	1	0	0
ID17	0	0	1	0	1	0	0	0	0	0	1	1
ID18	0	0	1	1	1	1	0	0	0	0	1	1
ID19	1	0	0	0	0	0	1	0	1	0	0	0
ID20	0	0	1	1	0	0	1	1	0	0	1	1
ID21	0	1	0	1	0	0	0	0	1	1	1	1
ID22	0	0	1	1	0	0	1	1	0	0	0	0
ID23	0	0	1	1	1	1	0	0	0	0	1	1
ID24	1	1	1	1	0	0	0	0	1	1	1	1
ID25	0	0	1	1	1	1	0	0	0	0	1	0
ID26	1	1	0	0	0	0	1	1	1	1	0	0
ID27	0	0	1	1	0	0	1	1	0	0	0	0
ID28	0	1	0	0	0	0	0	1	0	0	0	0
ID29	1	0	0	0	0	0	1	0	1	1	0	0
ID30	1	0	0	0	0	0	1	0	0	1	0	0
ID31	0	1	0	0	0	1	0	0	1	1	0	0
ID32	0	0	1	1	0	0	1	1	0	0	1	1
ID33	0	0	1	1	0	0	1	0	0	0	1	1
ID34	1	1	0	0	0	0	1	1	1	0	0	0
ID35	0	0	1	0	1	0	0	0	0	0	1	0
ID36	0	0	1	1	0	0	1	1	0	0	1	1
ID37	0	0	0	0	1	1	1	1	0	0	0	0
ID38	0	1	0	0	0	0	0	1	1	1	0	0
ID39	0	0	1	0	1	1	0	1	0	0	0	0
ID40	0	0	0	1	0	1	0	0	0	0	0	1
ID41	0	0	1	1	1	1	0	0	0	0	1	1
ID42	0	0	1	0	0	0	1	0	0	0	1	0
ID43	0	0	1	0	0	0	1	0	0	0	1	1
ID44	0	0	1	1	0	0	1	1	0	0	1	0
ID45	0	1	0	0	0	0	0	1	1	0	0	0
ID46	0	0	1	1	0	0	1	1	0	0	0	0
ID47	1	1	0	0	0	0	1	1	1	0	0	0
ID48	1	0	1	0	0	0	0	0	1	1	1	1
ID49	1	1	0	0	1	1	0	0	1	1	0	0
ID50	1	0	0	0	0	0	1	0	1	1	0	0
ID51	1	0	0	0	0	0	1	0	1	0	0	1
ID52	0	0	0	0	0	1	0	1	0	0	0	0
ID53	0	0	0	1	0	0	0	1	0	0	1	1

ID	Cube13	Cube14	Cube15	Cube16	Cube17	Cube18	Cube19	Cube20	Cube21	Cube22	Cube23	Cube24
ID1	1	0	0	0	1	1	0	0	1	1	0	0
ID2	0	0	0	0	1	1	1	1	0	0	0	0
ID3	1	0	1	0	0	0	0	0	1	1	1	1
ID4	1	1	1	1	0	0	0	0	1	1	1	1
ID5	0	0	1	0	0	0	1	0	0	0	1	0
ID6	0	0	0	0	1	0	1	0	0	0	0	0
ID7	1	1	1	1	0	0	0	0	1	0	1	0
ID8	0	0	1	1	0	0	1	0	0	0	1	0
ID9	0	0	1	1	1	1	0	0	0	0	1	1
ID10	1	1	0	0	0	0	1	0	1	0	0	0
ID11	0	0	1	0	1	1	0	0	0	0	1	1
ID12	0	0	1	1	0	0	0	0	0	0	0	0
ID13	1	0	0	0	0	0	1	1	1	1	0	0
ID14	1	1	0	0	1	1	0	0	1	1	0	0
ID15	0	0	1	1	0	0	1	1	0	0	1	1
ID16	0	0	0	1	0	1	0	0	0	0	0	1
ID17	1	1	0	1	0	0	1	1	1	1	0	0
ID18	1	1	0	0	0	0	1	1	1	1	0	0
ID19	0	0	1	0	1	1	0	0	0	0	1	1
ID20	0	0	1	1	0	0	1	0	0	0	1	0
ID21	0	0	0	0	1	0	1	0	0	0	0	0
ID22	0	0	0	0	0	0	1	1	0	0	1	1
ID23	1	1	0	0	0	0	0	0	0	0	0	0
ID24	0	0	0	0	1	1	1	1	0	0	0	0
ID25	1	0	0	0	0	0	1	0	1	0	0	0
ID26	0	0	1	1	1	0	0	0	0	0	1	0
ID27	0	0	0	0	0	0	1	0	0	0	1	0
ID28	0	0	0	0	0	1	0	0	0	0	0	1
ID29	0	0	1	1	0	1	0	0	0	0	0	1
ID30	0	0	0	1	1	1	0	0	0	0	1	1
ID31	1	1	0	0	0	1	0	0	0	1	0	0
ID32	0	0	0	1	0	0	1	1	0	0	1	1
ID33	0	0	1	1	0	0	1	1	0	0	1	1
ID34	0	0	1	0	1	1	0	0	0	0	1	1
ID35	1	0	0	0	0	0	1	1	1	1	0	0
ID36	0	0	1	1	0	0	1	1	0	0	1	1
ID37	1	0	1	0	0	0	0	0	1	0	1	0
ID38	0	0	1	1	1	1	0	0	0	0	1	1
ID39	1	0	1	0	0	0	1	0	1	1	0	1
ID40	0	1	0	0	0	0	1	1	1	1	0	0
ID41	1	1	0	0	0	0	1	0	1	0	0	0
ID42	0	0	1	0	0	1	1	0	0	0	1	1
ID43	0	0	1	1	0	0	1	1	0	0	1	1
ID44	0	0	1	0	0	0	0	0	0	0	0	0
ID45	0	0	1	0	1	0	0	0	0	0	1	0
ID46	0	0	0	0	0	0	0	1	0	0	0	1
ID47	0	0	1	0	0	1	0	0	0	0	0	1
ID48	0	0	0	0	1	0	1	0	0	0	0	0
ID49	1	1	0	0	1	0	0	0	1	0	0	0
ID50	0	0	1	1	1	0	0	0	0	0	1	0
ID51	0	0	1	1	1	0	0	0	0	0	1	0
ID52	1	1	1	1	0	0	0	0	0	1	0	1
ID53	0	0	1	1	0	0	1	1	0	0	1	1

ID	Cube1	Cube2	Cube3	Cube4	Cube5	Cube6	Cube7	Cube8	Cube9	Cube10	Cube11	Cube12
ID54	0	0	1	0	1	0	0	0	0	0	1	1
ID55	0	0	0	1	0	1	0	0	0	0	1	1
ID56	1	0	1	0	0	0	0	0	1	1	1	1
ID57	0	1	0	0	0	1	0	0	1	1	0	0
ID58	0	0	0	0	0	0	0	0	1	1	0	0
ID59	0	0	0	0	1	1	1	1	0	0	0	0
ID60	0	0	0	0	0	0	0	0	1	1	0	0
ID61	0	0	1	1	0	0	1	1	0	0	1	1
ID62	0	0	0	0	0	0	0	0	1	1	0	0
ID63	1	0	1	0	0	0	0	0	1	0	1	0
ID64	0	1	1	0	1	1	0	0	0	0	1	0
ID65	0	0	0	1	0	0	0	1	0	0	1	0
ID66	1	1	0	0	1	1	0	0	1	1	0	1
ID67	0	0	0	0	1	1	1	1	0	0	0	0
ID68	1	1	0	0	1	1	0	0	1	0	0	0
ID69	1	1	0	0	0	0	1	1	1	1	0	0
ID70	0	1	0	0	0	0	0	1	1	1	0	0
ID71	0	0	1	1	0	0	1	1	0	0	1	0
ID72	0	0	1	0	1	0	0	0	0	0	1	0
ID73	1	1	0	0	0	0	1	1	1	0	0	0
ID74	0	0	0	0	0	0	0	0	0	0	1	1
ID75	1	0	0	0	1	0	0	0	1	1	0	0
ID76	0	0	0	1	0	0	1	1	0	0	0	1
ID77	0	0	1	0	0	0	1	0	0	0	1	0
ID78	1	1	1	1	0	0	0	0	1	1	1	0
ID79	1	1	1	1	0	0	0	0	1	1	1	1
ID80	0	0	1	1	1	1	0	0	0	0	0	0
ID81	0	0	0	0	0	0	0	0	0	1	0	0
ID82	0	0	0	1	0	0	0	1	0	0	0	1
ID83	0	0	1	0	0	0	1	0	0	0	1	0
ID84	0	0	0	0	0	0	0	0	0	0	0	0
ID85	0	0	1	1	1	1	0	0	0	0	1	0
ID86	0	0	0	0	1	1	1	1	0	0	0	0
ID87	0	0	0	0	0	1	0	1	0	0	0	0
ID88	0	0	0	0	1	0	1	0	0	0	0	0
ID89	0	0	0	1	0	0	0	1	0	0	1	1
ID90	0	0	0	0	0	0	0	0	1	0	1	0
ID91	0	0	0	0	0	0	0	0	0	0	1	1
ID92	1	0	0	1	0	0	1	1	0	0	1	1
ID93	0	0	1	0	0	0	1	0	0	0	1	1
ID94	1	1	0	0	0	0	0	1	1	1	0	1
ID95	0	0	0	0	1	1	1	1	0	0	0	0
ID96	0	0	1	0	0	0	1	0	0	0	1	0
ID97	0	0	1	0	0	0	1	0	0	0	1	0
ID98	0	0	0	1	0	0	0	1	0	0	1	0
ID99	1	1	0	0	0	0	1	1	0	1	0	0
ID100	0	0	1	1	0	0	1	1	0	0	0	1
ID101	1	1	0	0	0	0	1	1	0	0	0	0
ID102	0	0	0	0	1	0	1	0	0	0	0	0
ID103	0	0	0	1	1	0	1	1	0	0	0	1
ID104	0	0	1	1	1	1	0	0	0	0	1	1
ID105	1	1	0	0	1	1	0	0	0	1	0	0
ID106	0	0	1	1	0	0	1	1	0	0	1	1

ID	Cube13	Cube14	Cube15	Cube16	Cube17	Cube18	Cube19	Cube20	Cube21	Cube22	Cube23	Cube24
ID54	1	1	0	0	0	0	1	1	1	1	0	0
ID55	1	1	0	0	0	0	1	1	1	1	0	0
ID56	0	0	0	0	1	1	1	1	0	0	0	0
ID57	1	1	0	0	1	1	0	0	1	1	0	0
ID58	0	0	1	1	1	1	0	0	0	0	0	1
ID59	1	0	1	0	0	0	0	0	1	1	1	1
ID60	0	0	1	1	1	1	0	0	0	0	1	1
ID61	0	0	1	1	0	0	0	0	0	0	0	0
ID62	1	1	0	0	1	1	0	0	1	1	0	0
ID63	0	0	0	0	1	1	1	1	0	0	0	0
ID64	1	0	0	0	0	1	0	0	0	1	0	0
ID65	0	0	1	0	0	0	1	0	0	0	1	0
ID66	1	1	0	0	0	1	1	0	1	1	0	0
ID67	1	0	1	0	0	1	0	0	1	1	1	0
ID68	1	0	0	0	1	0	0	0	1	0	0	0
ID69	0	0	1	1	0	1	0	0	0	0	0	1
ID70	0	0	1	1	1	0	0	0	0	0	1	0
ID71	0	0	1	0	0	0	1	1	0	0	1	1
ID72	1	0	0	0	0	0	1	1	1	0	0	0
ID73	0	0	1	0	1	0	0	0	0	0	1	0
ID74	1	1	0	0	0	0	1	1	1	1	0	0
ID75	1	1	0	0	1	1	0	0	1	1	0	0
ID76	0	0	1	1	0	0	0	1	0	0	1	1
ID77	0	0	1	0	0	0	1	0	0	0	1	0
ID78	0	1	0	0	1	0	1	0	0	0	0	0
ID79	0	0	0	0	1	1	1	0	0	0	0	1
ID80	0	0	0	0	0	0	1	1	1	1	0	0
ID81	0	0	0	1	0	1	0	0	0	0	0	1
ID82	0	0	0	1	0	0	1	1	0	0	1	1
ID83	0	0	1	0	0	0	1	1	0	0	1	1
ID84	1	1	1	1	0	0	0	0	1	1	1	1
ID85	1	0	0	0	1	0	1	0	1	0	0	0
ID86	1	0	1	0	0	0	0	0	0	1	0	1
ID87	1	1	1	1	0	0	0	0	1	1	1	1
ID88	1	1	1	1	0	0	0	0	0	1	0	1
ID89	0	0	1	1	0	0	1	0	0	0	1	0
ID90	0	0	0	0	1	1	1	1	0	0	0	0
ID91	0	0	1	1	0	0	1	1	0	0	1	0
ID92	0	0	1	1	0	0	1	1	0	0	1	1
ID93	0	0	1	1	0	0	0	1	0	0	0	1
ID94	0	0	1	1	1	1	0	0	0	0	0	1
ID95	1	1	1	1	0	0	0	0	0	0	0	0
ID96	0	0	1	0	0	1	1	1	0	0	1	1
ID97	0	1	1	1	0	0	1	1	0	0	1	1
ID98	0	0	1	0	0	0	1	1	0	0	1	1
ID99	0	0	0	1	0	1	0	0	0	0	0	1
ID100	0	0	0	1	0	0	1	1	0	0	1	1
ID101	0	0	0	0	1	1	0	0	0	0	1	1
ID102	1	1	1	1	0	0	0	0	1	1	1	1
ID103	1	0	1	1	0	0	1	0	1	0	1	0
ID104	1	1	0	0	0	0	0	1	0	1	0	0
ID105	0	1	0	0	0	1	0	0	0	1	0	0
ID106	0	0	1	1	0	0	0	1	0	0	0	1

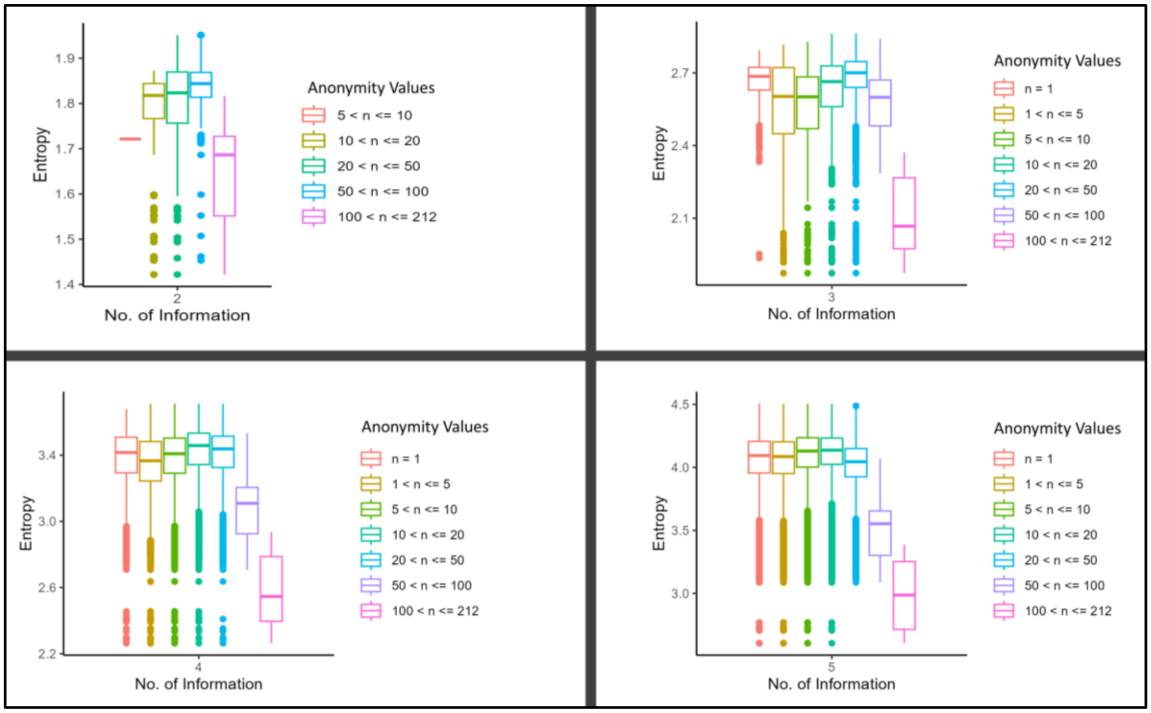
ID	Cube1	Cube2	Cube3	Cube4	Cube5	Cube6	Cube7	Cube8	Cube9	Cube10	Cube11	Cube12
ID107	1	1	0	0	1	1	0	0	1	1	0	0
ID108	1	1	0	0	0	0	1	1	0	1	0	0
ID109	0	0	1	1	0	0	1	1	0	0	1	0
ID110	1	1	0	0	1	1	0	0	1	1	0	0
ID111	0	0	1	1	1	1	0	0	0	0	1	1
ID112	1	1	0	0	0	0	0	1	1	1	0	0
ID113	0	1	0	1	0	0	0	0	0	1	0	1
ID114	1	1	1	1	0	0	0	0	1	1	1	1
ID115	1	1	1	1	0	0	0	0	1	0	1	0
ID116	0	0	0	0	1	0	1	0	0	0	0	1
ID117	1	0	0	0	0	0	1	0	1	1	0	0
ID118	1	0	0	0	0	0	1	0	1	0	0	0
ID119	1	0	0	0	1	0	0	0	1	0	0	0
ID120	0	0	1	0	0	0	1	0	0	0	0	1
ID121	1	0	0	0	1	0	0	0	0	1	0	0
ID122	0	0	0	0	0	0	0	0	1	1	1	1
ID123	0	0	0	0	0	0	0	0	0	0	1	1
ID124	0	0	0	1	0	0	0	1	0	0	1	1
ID125	1	1	0	0	0	0	1	1	1	1	0	0
ID126	0	1	0	1	0	0	0	0	0	1	0	1
ID127	0	0	1	1	1	1	0	0	0	0	0	1
ID128	1	0	1	0	0	0	0	0	1	1	0	0
ID129	0	1	1	0	0	0	1	1	1	1	0	0
ID130	0	0	1	0	0	0	0	1	0	0	1	1
ID131	0	1	0	0	0	0	0	1	0	1	0	0
ID132	0	0	0	0	0	0	0	0	1	1	1	0
ID133	1	1	1	1	0	0	0	0	0	1	0	1
ID134	0	0	0	0	0	0	0	0	0	0	0	0
ID135	0	0	1	1	0	0	1	1	0	0	0	1
ID136	1	0	0	0	1	0	0	0	1	1	0	0
ID137	0	0	1	0	0	0	1	0	0	0	1	1
ID138	0	0	0	1	1	0	1	1	0	0	0	1
ID139	0	0	1	1	0	0	1	1	0	0	1	1
ID140	1	0	0	0	0	0	1	0	1	1	0	0
ID141	0	0	1	1	0	0	1	1	0	0	0	0
ID142	0	0	1	0	0	0	1	0	0	0	0	1
ID143	1	1	1	1	0	0	0	0	0	0	0	0
ID144	0	0	1	0	0	0	1	0	0	0	0	0
ID145	1	1	0	1	1	0	0	0	1	1	0	1
ID146	0	0	0	0	0	0	0	0	0	0	1	0
ID147	1	0	0	0	0	0	1	0	1	0	0	0
ID148	1	1	1	1	0	0	0	0	0	1	0	1
ID149	0	1	0	0	0	1	0	0	1	0	0	0
ID150	0	0	0	0	1	1	1	1	0	0	0	0
ID151	0	0	1	1	0	0	1	1	0	0	1	0
ID152	1	1	1	1	0	0	0	0	1	0	1	0
ID153	1	1	0	1	1	1	0	0	1	1	0	0
ID154	1	1	1	0	0	0	0	1	1	0	1	0
ID155	1	1	0	0	1	1	0	0	1	1	0	1
ID156	1	0	0	0	0	0	1	0	1	0	0	0
ID157	0	0	1	1	1	1	0	0	0	0	0	1
ID158	0	0	1	1	1	1	0	0	0	0	1	0
ID159	0	1	0	1	0	0	0	0	1	1	1	1

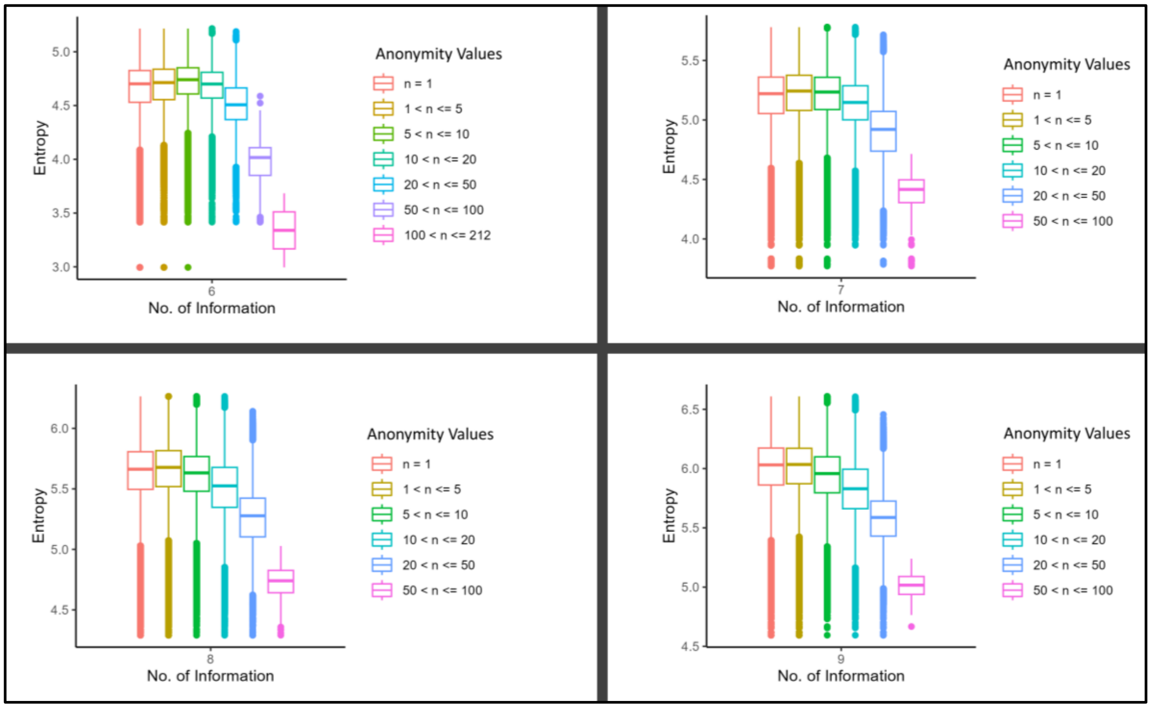
ID	Cube13	Cube14	Cube15	Cube16	Cube17	Cube18	Cube19	Cube20	Cube21	Cube22	Cube23	Cube24
ID107	1	1	0	0	0	0	0	0	0	0	0	0
ID108	0	0	0	1	1	1	0	0	0	0	1	1
ID109	0	0	1	0	0	0	0	1	0	0	0	1
ID110	1	0	0	0	1	1	0	0	1	1	0	0
ID111	1	1	0	0	0	0	1	1	1	1	1	0
ID112	0	0	1	1	0	1	0	0	0	0	0	1
ID113	0	0	0	0	0	1	0	1	0	0	0	0
ID114	0	0	0	0	0	0	0	0	0	0	0	0
ID115	0	0	0	0	1	0	1	0	0	0	0	0
ID116	1	1	1	0	0	0	0	0	1	1	1	1
ID117	0	0	1	1	1	0	0	1	0	0	1	1
ID118	0	0	1	0	1	0	0	0	0	0	1	0
ID119	1	0	0	0	1	0	0	0	1	0	0	0
ID120	1	0	1	1	0	0	1	0	0	0	1	0
ID121	0	1	0	0	1	1	0	0	1	1	0	0
ID122	0	0	0	0	1	1	1	1	0	0	0	0
ID123	0	0	1	1	0	0	1	0	0	0	1	0
ID124	0	0	1	1	0	0	0	1	0	0	0	1
ID125	0	0	1	1	0	0	0	0	0	0	0	0
ID126	0	0	0	0	0	0	0	0	0	0	0	0
ID127	0	1	0	1	0	0	1	1	1	1	0	0
ID128	0	0	1	1	1	0	1	0	0	0	0	0
ID129	0	0	1	1	0	0	1	0	0	0	1	0
ID130	0	0	1	1	0	0	1	0	0	0	1	0
ID131	0	0	0	1	1	0	0	0	0	0	1	0
ID132	0	0	1	1	1	0	0	0	0	0	1	0
ID133	0	0	0	0	1	0	1	0	0	0	0	0
ID134	1	0	1	0	0	0	0	0	1	0	1	0
ID135	0	0	0	1	0	0	1	0	0	0	1	0
ID136	1	1	0	0	1	0	0	0	1	0	0	0
ID137	0	0	0	1	0	0	1	0	0	0	1	0
ID138	0	0	0	1	0	0	0	0	0	0	0	0
ID139	0	0	1	1	0	0	1	0	0	0	0	1
ID140	0	0	1	1	1	1	0	0	0	0	1	1
ID141	0	0	0	0	0	0	0	1	0	0	1	1
ID142	0	0	0	1	0	0	1	1	0	0	1	1
ID143	0	0	0	0	1	1	1	1	0	0	0	0
ID144	0	0	0	0	0	0	1	0	0	0	1	0
ID145	1	0	0	0	1	1	0	1	1	0	0	0
ID146	0	0	1	0	0	0	1	0	0	0	1	0
ID147	0	0	1	0	0	0	0	0	0	0	0	0
ID148	0	0	0	0	1	1	1	1	0	0	0	0
ID149	1	0	0	0	1	1	0	0	1	1	0	0
ID150	0	0	0	0	0	0	0	0	1	1	1	1
ID151	0	1	1	1	0	0	1	0	0	1	1	1
ID152	0	0	0	0	0	1	0	1	0	0	0	0
ID153	1	1	0	0	1	1	0	0	1	1	0	0
ID154	0	0	0	0	0	1	0	1	0	0	0	0
ID155	1	0	0	0	1	0	0	0	1	0	0	0
ID156	0	0	1	0	1	1	0	0	0	0	1	0
ID157	0	1	0	0	0	0	0	1	0	1	0	0
ID158	1	0	0	0	0	0	0	0	0	0	0	0
ID159	0	0	0	0	1	1	1	1	0	0	0	0

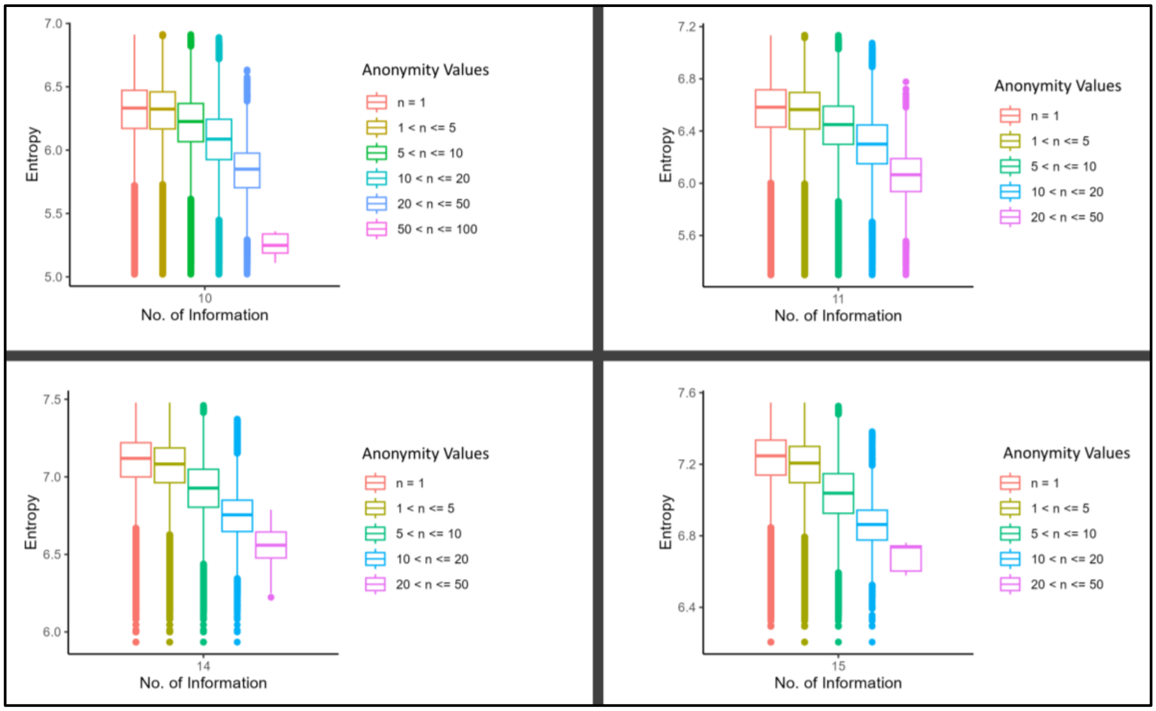
ID	Cube1	Cube2	Cube3	Cube4	Cube5	Cube6	Cube7	Cube8	Cube9	Cube10	Cube11	Cube12
ID160	0	0	1	0	1	0	0	0	0	0	0	1
ID161	1	0	0	0	1	1	0	1	0	0	0	0
ID162	0	0	0	1	1	0	1	1	0	0	1	0
ID163	0	0	1	1	1	1	0	0	0	0	1	0
ID164	0	1	0	1	0	0	0	0	0	0	1	1
ID165	0	1	0	1	0	0	0	0	1	1	1	1
ID166	1	0	0	1	0	0	1	1	0	0	1	0
ID167	0	0	0	0	0	1	0	1	0	0	1	0
ID168	0	0	1	1	0	0	1	1	0	0	0	1
ID169	0	0	0	0	1	0	1	0	0	0	0	0
ID170	0	1	0	0	0	1	0	0	1	1	0	0
ID171	0	0	1	1	0	0	1	1	0	0	0	1
ID172	0	0	0	1	0	1	0	0	0	0	1	1
ID173	0	0	0	0	1	1	1	1	0	0	0	0
ID174	1	1	0	0	1	1	0	0	0	0	0	0
ID175	1	1	1	1	0	0	0	0	1	0	1	0
ID176	1	1	0	0	1	1	0	0	1	1	0	0
ID177	0	1	0	1	0	1	0	0	0	0	0	0
ID178	0	0	0	1	1	1	1	0	0	0	0	1
ID179	0	0	0	0	0	0	0	0	1	1	1	0
ID180	0	0	1	1	1	1	0	1	0	0	1	1
ID181	0	1	1	0	1	1	0	0	1	1	0	0
ID182	0	0	0	1	0	0	0	1	0	0	0	0
ID183	0	0	1	0	0	0	1	0	0	0	0	1
ID184	0	0	0	0	1	0	1	0	0	0	0	0
ID185	0	0	0	0	0	0	0	0	0	0	0	0
ID186	0	0	0	1	0	0	0	1	0	0	0	1
ID187	1	1	1	1	0	0	0	0	1	1	1	1
ID188	0	1	0	1	0	0	0	0	0	0	0	0
ID189	0	0	0	0	0	1	0	1	0	0	0	0
ID190	0	0	0	0	0	0	0	0	0	1	0	0
ID191	0	1	0	0	0	0	0	1	0	1	0	0
ID192	0	0	0	0	0	0	0	0	1	1	0	0
ID193	1	1	0	0	1	1	0	0	0	1	0	0
ID194	0	0	0	0	1	1	1	1	0	0	0	0
ID195	0	0	1	1	0	0	0	1	0	0	1	1
ID196	1	1	0	0	1	1	0	0	1	0	0	0
ID197	0	1	0	0	1	1	1	0	0	1	0	0
ID198	1	0	0	1	1	1	0	0	1	0	0	0
ID199	0	0	0	1	0	0	0	1	0	0	0	1
ID200	0	1	1	0	0	0	1	1	0	1	1	0
ID201	0	1	0	0	0	1	0	0	0	1	0	0
ID202	0	1	0	1	0	0	0	0	1	1	0	1
ID203	1	0	0	0	0	0	1	0	0	0	0	0
ID204	1	1	0	0	1	1	0	0	1	1	0	1
ID205	0	0	1	1	0	0	1	1	0	0	1	1
ID206	0	1	1	1	1	0	0	0	0	0	1	0
ID207	0	1	1	0	0	0	1	1	0	0	1	1
ID208	0	0	0	0	0	0	0	0	0	0	0	1
ID209	0	0	0	0	1	1	1	1	0	0	0	1
ID210	1	1	1	1	0	0	0	0	1	1	1	1
ID211	0	0	0	0	1	0	1	0	0	0	1	0
ID212	0	1	0	1	0	0	0	0	0	0	1	0

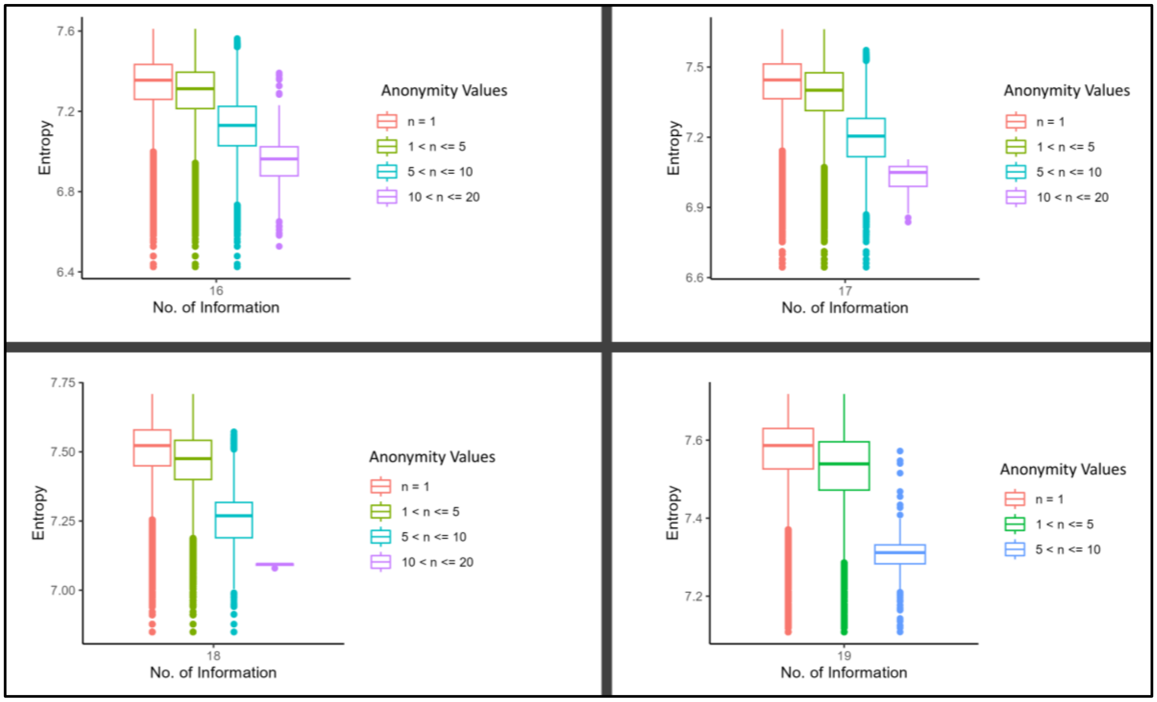
ID	Cube13	Cube14	Cube15	Cube16	Cube17	Cube18	Cube19	Cube20	Cube21	Cube22	Cube23	Cube24
ID160	0	1	0	0	1	0	1	0	1	0	0	0
ID161	1	0	1	0	0	1	0	0	1	1	1	0
ID162	0	0	1	0	0	0	0	1	1	0	1	1
ID163	1	0	0	0	0	0	0	1	0	1	0	0
ID164	0	0	1	1	0	0	0	1	0	0	0	1
ID165	0	0	0	0	0	1	0	1	0	0	0	0
ID166	0	0	1	0	0	0	1	1	0	0	1	1
ID167	1	1	0	1	0	0	1	0	1	1	0	1
ID168	0	0	0	1	0	0	0	1	0	0	0	1
ID169	1	1	1	1	0	0	0	0	1	0	1	0
ID170	1	1	0	0	1	0	0	0	1	0	0	0
ID171	0	0	0	1	0	0	0	0	0	0	0	0
ID172	1	1	0	0	0	0	0	1	0	1	0	0
ID173	0	1	0	1	0	0	0	0	1	1	1	1
ID174	0	0	0	0	1	1	0	0	1	1	0	0
ID175	0	0	0	0	1	1	1	1	0	0	0	0
ID176	1	1	0	0	0	1	0	0	0	1	0	0
ID177	0	0	0	0	0	1	0	0	0	1	0	0
ID178	0	1	0	0	0	0	0	1	0	1	0	0
ID179	0	0	0	1	0	1	0	0	0	0	0	1
ID180	1	1	0	1	0	0	1	0	1	1	0	1
ID181	1	1	0	0	1	1	0	0	1	1	0	0
ID182	0	0	0	0	0	0	0	1	0	0	0	1
ID183	0	0	0	1	0	0	0	1	0	0	0	1
ID184	0	1	0	1	0	0	0	0	1	1	1	1
ID185	0	0	0	0	0	0	1	1	0	0	1	1
ID186	0	0	0	1	0	0	0	1	0	0	0	1
ID187	0	0	0	0	1	1	1	1	0	0	0	1
ID188	0	0	0	0	1	1	1	0	0	0	0	1
ID189	1	0	1	0	0	0	0	0	1	1	1	1
ID190	0	0	0	1	1	1	0	0	0	0	1	1
ID191	0	0	0	1	0	0	0	0	0	0	0	0
ID192	0	0	1	1	1	0	0	0	0	0	1	0
ID193	0	1	0	0	1	1	0	0	1	1	0	0
ID194	0	1	0	1	0	0	0	0	1	0	1	0
ID195	0	0	1	1	0	0	0	0	0	0	0	0
ID196	1	0	0	0	1	1	0	0	1	1	0	0
ID197	1	1	1	0	0	1	0	0	1	1	1	0
ID198	1	0	0	0	1	1	0	0	1	1	0	0
ID199	0	0	0	1	0	0	0	0	0	0	0	0
ID200	0	0	1	1	0	0	0	0	0	0	0	0
ID201	0	1	0	0	1	1	0	0	1	1	0	0
ID202	0	0	0	0	0	1	0	1	0	0	0	0
ID203	0	0	0	0	1	0	0	0	0	0	1	0
ID204	1	0	0	0	1	1	0	0	1	1	0	0
ID205	0	0	1	1	0	0	1	1	0	0	1	0
ID206	1	0	0	0	0	0	1	0	1	0	0	0
ID207	0	0	1	1	0	1	1	0	0	0	1	1
ID208	0	0	0	1	0	0	0	1	0	0	0	1
ID209	0	0	0	1	0	0	0	0	1	1	1	1
ID210	0	0	0	0	0	1	0	1	0	0	0	0
ID211	0	0	1	0	0	0	0	0	1	1	1	1
ID212	0	0	1	0	0	0	1	0	0	0	1	0

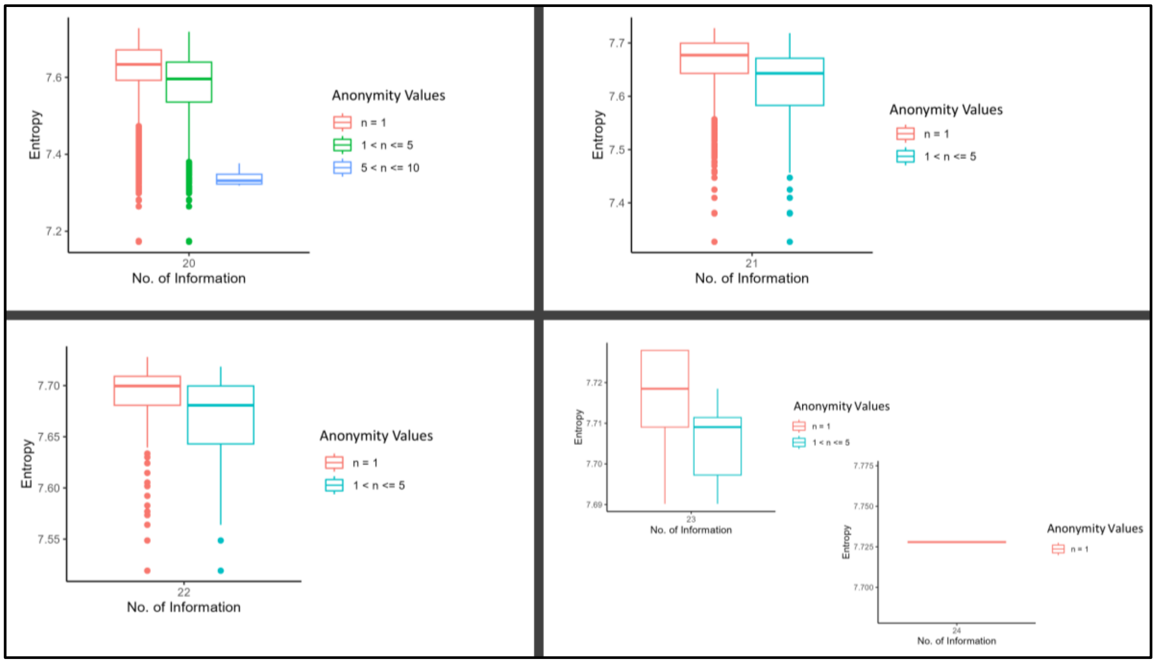
Appendix B. Box Plots of Entropy and Number of Information for Chapter 4






Entropy
7.0
6.5
6.0
5.5
5.0
10
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10
10 < n <= 20
20 < n <= 50
50 < n <= 100
Entropy
7.2
6.8
6.4
6.0
5.6
11
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10
10 < n <= 20
20 < n <= 50
Entropy
7.5
7.0
6.5
6.0
14
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10
10 < n <= 20
20 < n <= 50
Entropy
7.6
7.2
6.8
6.4
15
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10
10 < n <= 20
20 < n <= 50


Entropy
7.6
7.2
6.8
6.4
16
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10
10 < n <= 20
Entropy
7.5
7.2
6.9
6.6
17
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10
10 < n <= 20
Entropy
7.75
7.50
7.25
7.00
18
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10
10 < n <= 20
Entropy
7.6
7.4
7.2
19
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10


Entropy
7.6
7.4
7.2
20
No. of Information
Anonymity Values
n = 1
1 < n <= 5
5 < n <= 10
Entropy
7.7
7.6
7.5
7.4
21
No. of Information
Anonymity Values
n = 1
1 < n <= 5
Entropy
7.70
7.65
7.60
7.55
22
No. of Information
Anonymity Values
n = 1
1 < n <= 5
Entropy
7.72
7.71
7.70
7.69
23
No. of Information
Anonymity Values
n = 1
1 < n <= 5
Entropy
7.775
7.750
7.725
7.700
24
No. of Information
Anonymity Values
n = 1

VITA AUCTORIS

NAME:	Haesung Ahn
PLACE OF BIRTH:	Gwangju, Jeollanam-do, South Korea
YEAR OF BIRTH:	1981
EDUCATION:	Dongsung High School, Seoul, South Korea, 2000
	University of Seoul, B.Sc.- Transportation Engineering, Seoul, South Korea, 2006
	University of Seoul, M.A.Sc. - Transportation Engineering, Seoul, South Korea, 2014