UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Ekler Paulino de Mattos

Smart Privacy: An Anonymization-based Framework for Smart Mobility Open Data

Belo Horizonte
2024

Ekler Paulino de Mattos

Smart Privacy: An Anonymization-based Framework for Smart Mobility Open Data

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Antonio Alfredo Ferreira Loureiro
Co-Advisor: Heitor Soares Ramos Filho

Belo Horizonte
2024

Mattos, Ekler Paulino de.

C444s Smart privacy: an anonymization-based framework for smart mobility open data / Ekler Paulino de Mattos. – 2024.

1 recurso online (266 f. il.) : pdf.

Orientador: Antonio Alfredo Ferreira Loureiro.
Coorientador: Heitor Soares Ramos Filho.

Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.

Referências: f. 238-264.

1. Computação - Teses. 2. Redes de computadores - Teses. I. Loureiro, Antonio Alfredo Ferreira. II. Ramos Filho, Heitor Soares. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título.

CDU 519.6*22(043)

Ficha catalográfica elaborada por Célio Resende Diniz, bibliotecário CRB 6/2403 - Universidade Federal de Minas Gerais – ICEx.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇAO

FOLHA DE APROVAÇÃO

Smart Privacy: An Anonymization-based Framework for Smart Mobility Open Data

EKLER PAULINO DE MATTOS

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ANTONIO ALFREDO FERREIRA LOUREIRO - Orientador
Departamento de Ciência da Computação - UFMG

Documento assinado digitalmente
ANTONIO ALFREDO FERREIRA LOUREIRO
Data: 20/12/2024 10:22:24-0300
Verifique em https://validar.iti.gov.br

PROF. HEITOR SOARES RAMOS FILHO - Coorientador
Departamento de Ciência da Computação - UFMG

Heitor Soares Ramos Filho:78751241404
Assinado de forma digital por Heitor Soares Ramos Filho:78751241404
Dados: 2024.12.20 10:05:20 -03'00'

PROF. EDUARDO COELHO CERQUEIRA
Instituto de Tecnologia - UFPA

Documento assinado digitalmente
Eduardo Coelho Cerqueira
Data: 18/12/2024 13:06:46-0300
Verifique em https://validar.iti.gov.br

PROF. MÁRIO SÉRGIO FERREIRA ALVIM JÚNIOR
Departamento de Ciência da Computação - UFMG

Documento assinado digitalmente
MARIO SERGIO FERREIRA ALVIM JUNIOR
Data: 19/12/2024 09:37:54-0300
Verifique em https://validar.iti.gov.br

PROFª. MICHELE NOGUEIRA LIMA
Departamento de Ciência da Computação - UFMG

Documento assinado digitalmente
MICHELE NOGUEIRA LIMA
Data: 18/12/2024 15:35:00-0300
Verifique em https://validar.iti.gov.br

PROF. LEONARDO BARBOSA E OLIVEIRA
Departamento de Ciência da Computação - UFMG

Documento assinado digitalmente
LEONARDO BARBOSA E OLIVEIRA
Data: 18/12/2024 21:54:10-0300
Verifique em https://validar.iti.gov.br

PROF. FABRÍCIO AGUIAR SILVA
Departamento de Informática - UFV

Documento assinado digitalmente
FABRICIO AGUIAR SILVA
Data: 18/12/2024 13:19:46-0300
Verifique em https://validar.iti.gov.br

Belo Horizonte, 16 de dezembro de 2024.

Aos meus pais, irmãs, esposa e filhos.

Acknowledgments

I want to thank God, my fountain of faith, light and hope in all the days of my life.

I appreciate my advisor, Prof. Antonio A. F. Loureiro, for their encouragement, guidance, and support throughout my graduate studies. Who gave me valuable advice about research and life. Professor, thank you for believing in my potential and encouraging me to go further.

I am also very thankful to my co-advisor, Prof. Heitor S. R. Filho, for accepting and helping me with the challenging research topic of privacy in smart mobility. Thank you for the technical support and valuable guidance, and be attentively present to my progress in each step of this work.

I am also grateful to Prof. Fabrício (UFV). Your recommendations were also fundamental to the success of this work.

I would definitely like to express my deepest gratitude to my dear wife, Desirée Pires Diniz. The cornerstone is that without her unconditional help, patience, and love, this work would not be possible in good times and bad. I would not succeed without you by my side. I will be eternally grateful for your presence in my life. To my children Cecília and Rael, my source of inspiration and strength. Loves of my life, thank you for making me a better person!

To my life teachers: my parents, Mattos and Elisa, and my sisters, Thaiz and Rosane. Thank you for teaching me everything about life, love, and people. On countless occasions, you were present with words of encouragement and affection.

To my friends and professors at UFMS, Angelo, Deiviston, Gedson, Glasielly, Juliana, Kleber, Priscila, Edson Cáceres, and Nalvo who helped me through this doctoral process, thank you for your support, conversations, and recommendations during this journey.

To my long-time friends Bruno Zarpelão and Rodrigo Miani, for their support and encouragement in this process.

I would also like to thank the members of the WISEMAP lab, in particular Augusto, for their technical support and valuable contributions to this work; thank you very much for your sincere friendship. Thanks to UFMG friends Keiller and Balbino for the conviviality, necklaces, and exchanges of knowledge that provided great insights into my research. I certainly made lots of friends here.

CAPES, CNPq, FAPESP, and FAPEMIG partially supported this research.

"*A career is born in public - talent in privacy*"
(Marilyn Monroe)

Resumo

Nas cidades inteligentes, pessoas e veículos são entidades móveis que produzem dados massivos de geolocalização por meio de vários sensores, chamados dados de mobilidade. Esses dados interessam a vários setores e podem ser protegidos, publicados parcialmente ou totalmente e usados para diversos propósitos, chamados de *dados abertos de mobilidade inteligente* (SMOD). No entanto, surgem preocupações sobre a privacidade, qualidade e utilidade desses dados, especialmente ao considerar a mobilidade em ambientes dinâmicos como as cidades inteligentes. Para mitigar os riscos à privacidade, como a reidentificação de indivíduos e a exposição de informações sensíveis, foram desenvolvidos Mecanismos de Proteção de Privacidade de Localização (LPPMs). No entanto, muitos LPPMs são projetados para operar em modo estático ou são mal calibrados e não consideram a mobilidade das cidades inteligentes. Outra questão essencial é compreender o comportamento dos LPPMs, que refletem na qualidade de proteção destes dados. Além disso, surge uma questão sobre a utilidade dos SMOD, que está associada ao seu uso generalizado para diversas finalidades, tornando o processo de proteção destes dados mais complexo em termos práticos. Nesse sentido, há um desafio em publicar os SMOD para fins específicos, como os domínios, aplicações e serviços das cidades inteligentes. Considerando esses desafios, o objetivo desta tese é estudar como a privacidade de localização baseada em anonimização pode ser aplicada aos SMOD para atingir os requisitos de privacidade, utilidade e qualidade de anonimização em cidades inteligentes. Orientando este estudo, introduzimos um framework baseado em anonimização para SMOD, que considera os requisitos de privacidade, utilidade e qualidade de anonimização. Assim, avançamos o estado da arte em quatro frentes: (i) propusemos um framework para caracterizar e encontrar similaridades nas distribuições estatísticas extraídas de métricas de mobilidade que evidenciam o impacto da mobilidade na privacidade; (ii) projetamos uma solução capaz de identificar domínios, aplicações e serviços que melhor aproveitam os dados de mobilidade anonimizados; (iii) projetamos uma solução que mensura a qualidade dos dados anonimizados e do funcionamento de mix-zones, um tipo de LPPM baseado em anonimização; e (iv) projetamos um ataque de reidentificação de trajetória eficiente e um esquema de mix-zone dinâmica que ajusta o nível de privacidade ao longo do tempo frente às flutuações de tráfego. Nossas contribuições avançam no projeto de LPPMs, considerando a privacidade, utilidade e qualidade de anonimização de SMOD, aspectos essenciais para o desenvolvimento de cidades inteligentes.

Palavras-chave: Privacidade de Localização; Anonimização; Dados Abertos; Cidades Inteligentes

Abstract

In smart cities, people and vehicles are mobile entities that produce massive geolocation data through various sensors, called mobility data. This data interests various sectors and can be protected, partially or fully published, and used for various purposes, called Smart Mobility Open Data (SMOD). However, concerns arise about this data's privacy, quality, and utility, especially when considering mobility in dynamic environments such as smart cities. To mitigate privacy risks, such as the individuals' re-identification and exposure of latent information, Location Privacy Protection Mechanisms (LPPMs) have been developed. However, many LPPMs are designed to operate in static mode or are poorly calibrated and do not consider the mobility of smart cities. Another issue is understanding the behavior of LPPMs, which reflects on the quality of protection of this data. In addition, a question arises about the utility of SMOD, which is associated with their widespread use for various purposes, making the process of protecting this data more complex in practical terms. In this sense, there is a challenge in publishing SMOD for specific purposes, such as smart city domains, applications, and services. Considering these challenges, this thesis aims to investigate how anonymization-based location privacy can be applied to SMOD to achieve the privacy, utility, and anonymization quality requirements in smart cities, considering the particular characteristics of this environment. We investigate privacy attacks and LPPM schemes in mobile networks, noting their limited application in SMOD and the need for privacy mechanisms adapted to this context. Guiding this study, we present an anonymization-based framework for SMOD, which considers privacy requirements, utility, and anonymization quality. Thus, we advance state of the art in four fronts: (i) we propose a framework to characterize and find similarities in the statistical distributions extracted from mobility metrics that evidence the impact of mobility on privacy; (ii) we design a solution capable of identifying domains, applications, and services that best leverage anonymized mobility data; (iii) we design a solution that measures the quality of anonymized data and the functioning of mix-zones, a type of anonymization-based LPPM; and (iv) we design an efficient trajectory re-identification attack and a mix-zone dynamics scheme that adjusts the privacy level over time in response to traffic fluctuations. Our contributions advance the design of LPPMs by considering the privacy, utility, and anonymization quality of SMOD, which are essential for developing smart cities.

Keywords: Smart Cities; Open Data; Location Privacy; Anonymization

List of Figures

List of Tables

List of Abbreviations

AC Area Coverage
ACC$_{pred}$ Accuracy in k's Predictions
AHP Analytic Hierarchy Process
AIC Akaike Information Criterion
ANP Analytic Network Process
APA Aggregated Presence Attack
AQ Anonymization Quality
AQM Anonymization Quality Framework for Mix-zones
AR Anonymization Rate
ASS Anonymity Set Size
ATM Activation Time of the Mix-zone
BWM Best Worst Method
CA Certification Authority
ChA Cheating Attack
CODU Contact Duration between a Pair
CONEN Contact Entropy
CS charging station
CV coefficient of variation
DBSP DBSCAN Stay Points
DEA De-Anonymization
DGA dictionary guessing attacks
DM decision-maker
DPS Data Privacy Specialist
DSRC Dedicated Short Range Communication
DTW Dynamic Time Warping algorithm
EV Electric Vehicles
FFA FIFO Attack
FMP Future Mobility Prediction
FPMT Frequency of Points from the Middle of the Trajectory
GDPR Data Protection Regulation
GEO-I Geo-indistinguishability
GPA Global Passive Adversary
HA Higher Anonymization
IDM Interval of Departure Time between Cars on Mix-zones
INCO Inter-contact Time
IoD Internet of Drones
IoHT Internet of Health Things
IoT Internet of Things
IoV Internet of Vehicles
ITM Interval of Arrival Time between Cars on Mix-zones
k-DynMix k-Dynamic Mix-zone
LAA Local Active Adversary
LBS Location Based Service
LBSN Location-based Social Network
LPA Location Privacy Attack
LPAd Local Passive Adversary
LPPM Location Privacy Protection Mechanism
MA Mix-zone Activation
MAXCON Maximum of Connections between a User Pairs
MCDM Multi-Criteria Decision-Making
MD Mix-zone Deactivation
MDA Meeting Disclosure Attack
ME Mix-zone Efficacy
MLE Message-locked encryption
MLTA Maximum Likelihood Tracking Attack
MMBA Maximum Movement Boundary Attack
MPA Mobility Prediction Accuracy

MQA Multiple-query Attack
MSN Mobile Social Network
MT-MT multimodal traces with multimodal trajectories
MTT Multi-target tracking
MT-UT multimodal traces with unimodal trajectories
NAR Non-Anonymization Rate
NCM Number of Cars in Mix-zone
NNPDA Nearest-Neighbor Probabilistic Data Association
NTC Number of Trips Completed within the Mix-zone
NUMVIS Number of Visits per User
OA Optimal Anonymization
OSN Online Social Network
PA Position Attack
PAC POIs' amount by Cell
PADA Presence and Absence Disclosure Attack
PIA Personal Identification Attack
PMA POI Matching Accuracy
POI Point of Interest
POIA Points of Interest Attack
PPT Probabilistic Polynomial Time
PSEU Pseudonyms per User
RGYR Radius of Gyration
RIA Region Intersection Attack
ROI Region of Interest
RRET Re-identification Risk in entire Trip
RSU Road Side Unit
SC spatial crowdsourcing
ScA Scrambler Attack
SeLA Semantic Linking Attack
SMA Simple Moving Average
SMOD smart mobility open data
SP Stay Point
SpA Speed Attack
SPAD Spatial Distortion
SPAP Spatial Distortion of POIs
SPC Stay Point Count
SPD Stay Point Duration
SPF Smart Privacy Framework
SRA Shrink Region Attack
SSE sum of squared estimate of errors
SyA Sybil Attack
SyLA Syntactic Linking Attack
TA Tracking Attack
TAC Timeout of Arriving Cars in Mix-zones
TAS Trips' Average Speed
TC Trajectory Classification
TDi Trips' Distance
TDSL Travel's Distance Straight Line
TDu Trips' Duration
TiA Timing Attack
TMA Trajectory Matching Accuracy
TrA Transition Attack
TRIPTIME Trip Time
U Uniqueness
UAFAT Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains
UT unimodal traces
V2I vehicle to infrastructure
V2V vehicle to vehicle
V2X Vehicle-to-Anything
VANET vehicular ad-hoc network
VL Visits per Location
VLBS Vehicle Location-Based Services
VSN Vehicular Social Network
WEMA Weighted Exponential Moving Average
WM Wasserstein metric
WPM Weighted Product Method
WS Wasserstein metric

List of Publications

Refereed Conference Papers

1. **Ekler Paulino Mattos**, Augusto CSA Domingues, and Antonio AF Loureiro. Reidentificação de trajetórias de veículos baseada na caracterização das preferências de caminho. In **Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**, pages 820–833. SBC, 2019.

2. **Ekler P de Mattos**, Augusto CSA Domingues, and Antonio AF Loureiro. Give me two points and i'll tell you who you are. In **2019 IEEE Intelligent Vehicles Symposium (IV)**, pages 1081–1087. IEEE, 2019.

3. Augusto CSA Domingues, **Ekler Paulino de Mattos**, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Social Mix-zones: Anonymizing Personal Information on Contact Tracing Data. In Proceedings of the **18th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks**, pages 81–88, 2021.

4. **Ekler Paulino de Mattos**, Augusto CSA Domingues, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Behind the Mix-Zones Scenes: On the Evaluation of the Anonymization Quality. In Proceedings of the **19th ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks**, pages 133–140, 2022.

5. **Ekler Paulino de Mattos**, Augusto CSA Domingues, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Protect your Data and I'll Show Its Utility: A Practical View about Mix-zones Impacts on Mobility Data for Smart City Applications. In Proceedings of the Int'l **ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks**, pages 45–52, 2023.

6. **Ekler Paulino Mattos**, Augusto CSA Domingues, Fabrício A. Silva, Heitor S Ramos, and Antonio AF Loureiro. k-DynMix: Um Mecanismo de Proteçã Dinâmica de Privacidade em Mix-Zones. In **Anais do XXIV Simpósio Brasileiro de**

Segurança da Informação e de Sistemas Computacionais, pages 709–724. SBC, 2024.

Refereed Journal Papers

1. **Ekler P de Mattos**, Augusto CSA Domingues, Bruno P Santos, Heitor S Ramos, and Antonio AF Loureiro. The impact of mobility on location privacy: A perspective on smart mobility. **IEEE Systems Journal**, 16(4):5509–5520, 2022.
2. **Ekler Paulino de Mattos**, Augusto CSA Domingues, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Slicing who slices: Anonymization quality evaluation on deployment, privacy, and utility in mix-zones. **Computer Networks**, 236:110007, 2023.
3. **Ekler Paulino de Mattos**, Augusto CSA Domingues, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Protect your data and I'll rank its utility: A framework for utility analysis of anonymized mobility data for smart city applications. **Ad Hoc Networks**, page 103567, 2024.

Future Work under Review by Advisor

1. Augusto CSA Domingues, **Ekler Paulino de Mattos**, Fabrício A Silva, and Antonio AF Loureiro. Trajectory Definition, Assessment, and Recommendation: A Survey. IEEE Transactions on Intelligent Transportation Systems, (2025). **Under revision by advisor - periodic**.
2. **Ekler Paulino de Mattos**, Augusto CSA Domingues, Bruno P. Santos, Alisson Renan Svaigen, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. SELECT Location Privacy Attacks FROM Mobile Networks ORDER BY Usability: A Survey. IEEE Communications Surveys and Tutorials, 2025. **In production - periodic**.
3. **Ekler Paulino de Mattos**, Augusto CSA Domingues, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. k-DynMix: A Dynamic Privacy Protection in Mix-zones, 2025. **In production - periodic**.

Contents

Chapter 1

Introduction

In the digital age, modern cities worldwide seek to integrate information and communication technologies into all aspects of city life to promote urban centers, considering different purposes where the citizen plays a central role. The ultimate goal is to have the so-called *smart cities* [1, 4] that provide their citizens with a proper quality of life. Many smart city applications are expected to be designed and organized to achieve this urban revolution using nine building blocks, as depicted in Figure 1.1.

A *smart utility* is a building block aimed at reducing the consumption of resources such as energy, gas, and water. In this way, it can contribute to economic growth, sustainability, and efficiency. *Smart buildings* enable the use of sensors and smart grids to implement smart infrastructures, such as water systems, energy, streets, and buildings. *Smart environment* aims to improve the sustainability of cities and the quality and safety of citizens' lives, such as pollution control. *Smart governance* intends to increase transparency, improve local government efficiency, and tailor government services to its citizens. *Smart public services* enable the deployment of public resources efficiently and effectively and include applications such as adaptive waste management. *Smart health* includes the efficient and effective provision of health care. *Smart economy* aims to create economic growth, employment, and urban growth. *Smart citizens* focus on forming more conscious citizens through smart educational systems to increase their employability and digital inclusion, among other goals.

Smart mobility is a key building block most associated with smart cities [1]. In particular, it involves a smart transportation system that improves traffic safety and efficiency, reduces the time citizens spend in transit, and enhances the quality of life [5]. For this, it uses transportation networks with improved, embedded real-time monitoring and control systems [6, 7]. Users of smart mobility can combine different transportation modes to reduce travel times, traffic, and air pollution. Moreover, it also includes intelligent vehicles with driver assistance systems, smart adaptive traffic lights to improve traffic flow and reduce travel times [8], bus route optimization [9], shared-bike solutions [10], and scooter programs to decrease car traffic and air pollution [11, 12].

These building blocks – people, vehicles, and devices – are ubiquitous entities that produce a large amount of data using different sensors and services, such as location-based

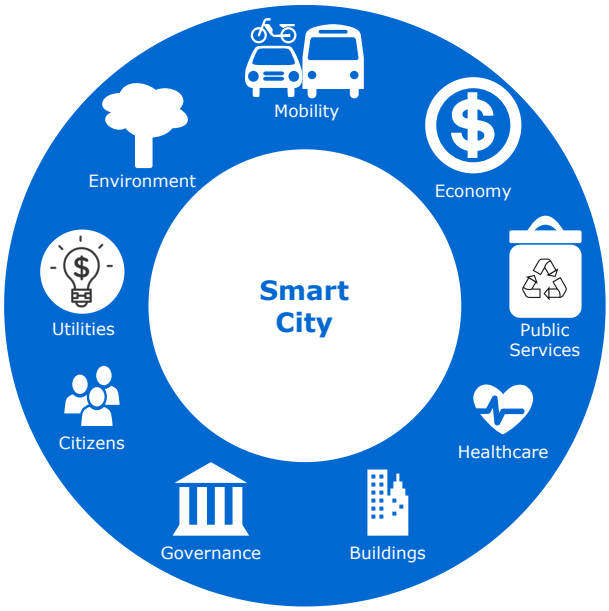


Figure 1.1: Smart city building blocks [1].

services, pollution management, climate monitoring, smart highway infrastructure, and health care [13]. Naturally, several of these entities are mobile in a cyber-physical environment, and produce data containing timestamped geolocation records, called mobility data.

Mobility data is valuable because it provides information about people's location over time, i.e., their trajectories containing spatial and temporal information. These mobility data can be fully or partially published in a public repository that can be collected, processed, and analyzed for many purposes in developing smart cities, named *smart mobility open data (SMOD)* [7, 14]. For instance, the statistical analysis of mobility data can contribute to traffic monitoring applications, route plans, multimodal transport integration, and so on [15, 16, 17]. We can use mobility data for urban planning, like Point of Interests (POIs) mining and strategic places identification, such as cities for deploying services and electric vehicles (EV) charging; utilized the design of data dissemination protocols for Vehicle to infrastructure (V2I) and Vehicle to vehicle (V2V) communication in Vehicular ad-hoc networks (VANETs) [17, 18, 19, 20]. The mobility data can also be used in health and private companies, such as the spread of diseases and targeted marketing, and among several other possibilities [16, 21, 22, 23].

Despite the importance of mobility data for smart city development, they present concerns about the privacy, utility, and quality of these data, whether they are consumed in offline or online mode [16, 17, 24, 25]. When we think of utility in the abstract sense, we are referring to the widespread use of mobility data, which makes the processing task, such as protecting this data to obtain utility, more complex in practical terms. Another way of thinking about utility is to process this data for specific purposes, such as application domains and services [25]. In this sense, there is a challenge in making data available for

specific application domains.

Another important issue is measuring the quality of mobility data when it is protected for publication. Quality requirements guide utility. Data that has been protected by protection algorithms with poorly calibrated parameters or with little data flow to produce reasonable levels of protection will affect the quality of the data and, consequently, its usefulness for certain purposes. Therefore, it is necessary to investigate the quality of the execution of the protection algorithms, which will affect the data quality.

Privacy is the principal requirement of open mobility data because it directly impacts utility and quality requirements [16, 17, 25]. The mobility data present significant privacy risks when published to the public [26, 27]. Several studies have shown that it is possible to use the open mobility data to re-identify people/vehicles and their latent information such as identity, residence, place of work, and even religious preferences with Location Privacy Attacks (LPAs) [28, 29, 30]. In this way, Location Privacy Protection Mechanisms (LPPMs) have been proposed to address privacy leakage, which are divided into anonymization and obfuscation techniques [31, 32, 33]. Anonymization techniques are identity perturbations by concealing the relationship between users and their mobility data by pseudonyms changing. In VANETs, vehicles receive a pseudonym to protect the real identity of the users. These pseudonyms are changed for certain events, like random periods, to avoid statistical analysis and inference attacks. Obfuscation techniques consist of adding noise to mobility data to send intentionally inaccurate location information to applications that use these locations to provide some service to the user called Location Based Service (LBS) applications.

1.1 Problem Statement

A critical issue about LPPM design is understanding how mobility impacts users' privacy. This investigation is mandatory because many privacy mechanisms are set up with statics mode, usually applied to homogeneous datasets, with only one modal type that does not consider mobility aspects in dynamic environments, such as smart mobility. Also, the mobility data protected by a LPPM misconfigured can degrade the operation of online applications, such as the quality of service of LBS, and in offline applications, using these data can biasing results, such as distorting the strategic places identification in the urban planning context. This investigation can contribute to designing efficient LPAs in which the adversary knowledge is some aspect of mobility that can yield a high re-identification rate. Also, efficient attacks can drive the development of more resilient LPPMs that also consider the mobility aspects–for example, producing LPPMs capable of

getting privacy levels following vehicle traffic fluctuations over time for different transport modes.

Another issue is that the LPPM improves users' location privacy but at the cost of data utility [34]. In the anonymization-based LPPMs, the change of pseudonyms must be frequent to obtain adequate privacy and reduce the length of the trajectories in the mobility data that could be explored by statistical analysis [15]. Also, it diminishes the utility of this data, hiding the temporal relationship between the user's locations, which can be critical in the utility of some systems. In the obfuscation-based LPPMs, adding noise to the mobility data can also degrade the performance of systems sensitive to absolute values of location information, such as a ride-sharing LBS system. Further, a comprehensive discussion about the trade-off between privacy and utility happens once the data is protected. However, issues still arise about the application of anonymized data to smart city development: What are the smart city domains, applications, and services that can best leverage mobility data protection by mix-zones: an anonymization-based LPPM?

The Anonymization Quality (AQ) is another issue that must be addressed. Although well-known anonymization-based LPPMs exist to degrade the anonymization rate and enable linking and inference attacks, a few have investigated their behavior and analysis of the data quality protected by LPPMs. The AQ is a concept related to LPPM quality of internal functioning that impacts the quality of anonymized mobility data.

In this way, the research problem we tackle in this thesis can be summarized with the following problem statements:

- **RQ1: What are the impacts of mobility on location privacy?**
- **RQ2: How to build efficient LPAs based on mobility characteristics and resilient LPPMs based on anonymity, which considers privacy, utility, coverage, and anonymization quality for SMOD?**
- **RQ3: How do you measure the quality over time of a LPPM and the data it anonymizes?**
- **RQ4: What are the smart city domains, applications, and services that can best leverage mobility data anonymized by mix-zones?**

These questions represent open challenges in the Vehicular Networks research field, specifically in the privacy aspects of open data in Smart Mobility. Obtaining the answers to these questions can bring significant advancements regarding the design of LPPMs in

real-world smart city environments, considering the privacy, utility, and anonymization quality of SMOD, and contribute to the planning and development of smart cities.

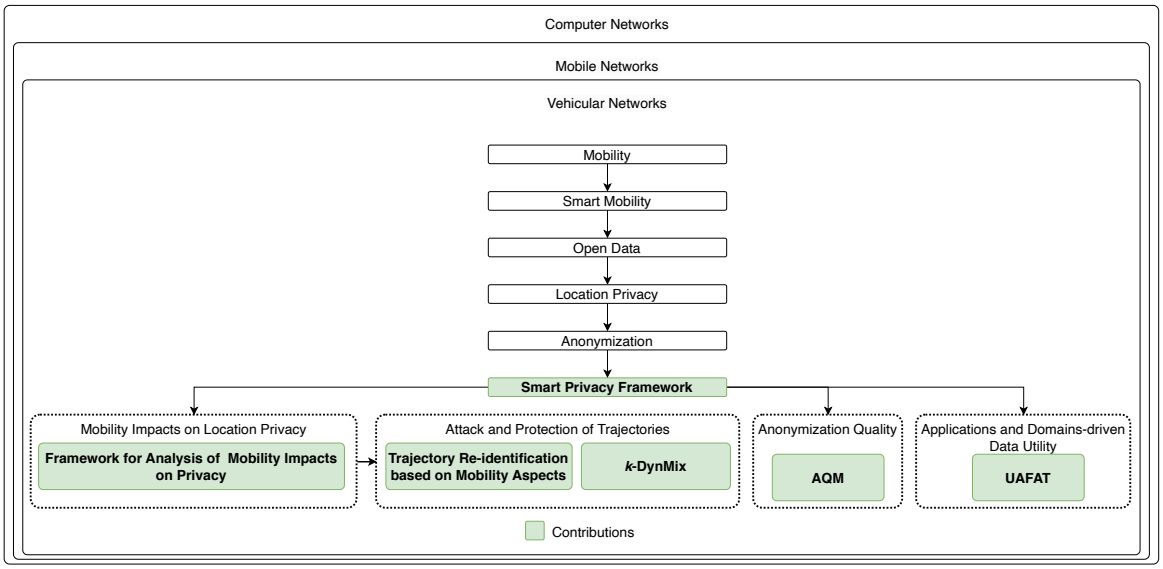


Figure 1.2: Related research field and thesis's contributions.

1.2 Objective

The main goal of this thesis is to investigate how location privacy based on anonymization can be applied to SMOD to achieve the privacy, utility, and anonymization quality requirements in smart cities, considering the particular characteristics of this environment.

The following steps were followed to achieve this objective. First, we survey the current location privacy in mobile networks, categorizing major types of LPAs and corresponding LPPM schemes, including privacy and utility metrics. This study reveals that these mechanisms are underexplored in the context of open mobility data for smart cities. Specifically, there is a need to develop attack and defense mechanisms that account for the mobility characteristics of the smart city environment. Next, guided by this study, we have introduced the Smart Privacy Framework (SPF), a novel anonymization-based framework for smart mobility open data. This framework considers the privacy, utility, and anonymization quality requirements. This framework is organized on four fronts: Mobility Impacts on Location Privacy, Attack and Protection of Trajectories, Anonymization Quality, and Applications and Domains-driven Data Utility.

The first front – Mobility Impacts on Location Privacy – concerns the impacts of mobility on privacy. We conduct a comprehensive study of mobility characteristics found

in mono- and multimodal datasets that could affect privacy in the smart mobility context.

The second front – Attack and Protection of Trajectories – concerns the design of a LPA and a LPPM algorithms that consider the mobility aspects, the first to re-identify and the second to protect the users' trajectories. The results obtained on the first front brought perspectives for the design of the second front.

The third front – Anonymization Quality – refers to the study of the quality of anonymity of LPPMs. Understanding the AQ enables performing analysis, such as electing better-deployed privacy locations, analyzing the trade-off between privacy, utility, and quality, and measuring the LPPM performance per period, resulting in the quality of anonymized mobility data.

In the fourth front – Applications and Domains-driven Data Utility – we study the utility of anonymized mobility data in the sense of identifying smart cities domains, applications, and services that can better leverage anonymized data by anonymization-based LPPMs. Also, it enables the identification of applications from various smart cities applications that can make the best use of anonymized mobility data.

1.3 Contributions

This study advances the state-of-the-art in location privacy applied to SMOD in four different directions. Figure 1.2 summarizes the main contributions of this thesis. We describe them as follows.

► **Mobility Impacts on Location Privacy**

C1) Framework for Analyzing the Impacts of Mobility on Privacy. Answer to RQ1: We proposed a framework to characterize and find similarities in the statistical distributions extracted from two Stay Points (SPs) metrics. SP is an aspect of mobility data analysis, defined as a region where an entity stays for a minimum time interval. SPs operate as substrates to build location privacy protection mechanisms, and studying their behavior with different types of mobility datasets can reveal valuable insights about privacy.

► **Attack and Protection of Trajectories**

On this front, we have two contributions in answer to RQ2. A trajectory re-identification algorithm and a dynamic anonymization-based LPPM mitigate this attack.

- **C2) Trajectory Re-identification Attack Algorithm.** We presented a simple and efficient re-identification attack technique that uses only two geo-referenced points as input data. The attack considers the mobility characteristic where adversary knowledge is information about drivers' behavior in a city, such as their road preferences, can be used to re-identify their trajectories anonymized by mix-zones. The results showed that the attack approach re-identified up to 95% of anonymized cabs trajectories.

- **C3) k-DynMix - A Dynamic Privacy Protection in Mix-zones.** Mix-zones are anonymization-based LPPM that depend on factors that affect their performance, e.g., defining fair privacy levels (k) over time. To tackle these issues, we proposed the k-DynMix, a dynamic mix-zone that adjusts the level of privacy over time in online mode and linear complexity, according to the flow of vehicles, to achieve higher anonymization. We validate our approach with two prediction engines using accuracy to predict privacy. Also, we have compared the k-DynMix with classic mix-zones using mix-zone coverage, Anonymization Quality (AQ) metrics, and the AQM framework. Finally, we explored the potentialities of k-DynMix in anonymizing trajectories against our trajectory re-identification attack proposed above. The k-DynMix simultaneously maximized privacy over time and achieved performance similar to the best result of classic mix-zones in coverage metrics and AQ. Also, the k-DynMix outperformed the classic mix-zones against the trajectory re-identification attack.

► **Anonymization Quality**

C4) Anonymization Quality Framework for Mix-Zones (AQM). Answer to RQ3: The Anonymization Quality (AQ) is a concept related to the protection, efficacy, and internal functioning of an LPPM, enabling an understanding of how privacy occurs and analyzing its performance over time. Anonymization quality framework for mix-zones (AQM) enables characterizing and evaluating the impacts of anonymization over time and space in mobility data. For instance, elect mix-zones that do not consider the traffic but its operating requirements, such as quality, coverage, privacy, and utility metrics. Also, with AQM, it is possible to elect a positioning algorithm from two approaches.

► **Applications and Domains-driven Data Utility**

C5) Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT). Answer to RQ4: The trade-off between privacy and utility is a critical issue in data protection. Although the literature provides a comprehensive discussion about this trade-off, issues still need to be solved about the application of anonymized data to smart city development. For instance, to identify smart cities

domains, applications, and services that can best leverage mobility data anonymized by mix-zones. To address this issue, we propose the Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT), which aims to identify domains, applications, and services where the anonymized data will provide more or less utility in various aspects. Moreover, it measures the utility through twelve metrics related to privacy, mobility, and social, including mix-zones performance metrics from anonymized trajectories produced by mix-zones.

1.4 Thesis Outline

The remainder of this thesis is organized into nine chapters, described as follows:

- Chapter 2 focuses on the comprehensive study about location privacy on mobile networks. First, we present the fundamentals of location privacy attacks concerning mobility. Next, we present the LPPMs based on obfuscation and anonymization, including they privacy and utility metrics. Finally, we focused on the state-of-art mix-zones, a type of anonymization-based LPPM widely used in Vehicular Networks.

- Chapter 3 presents the Smart Privacy Framework (SPF): an Anonymization-based Framework for Smart Mobility Open Data and their privacy, quality, and utility indicators assists the publishing data analysis.

- Chapter 4 brings a general discussion of location privacy issues in smart mobility. Then, we present a characterization framework that evidences the relationship between location privacy and mobility, which can be a shining light to address location privacy in smart mobility.

- Chapter 5 presents the privacy contribution regarding the privacy. Particularly a trajectory re-identification attack that exploits mobility characteristics to identify anonymized trajectories by mix-zones.

- The contribution of the anonymization quality is on Chapter 6, which presents an Anonymization quality framework for mix-zones (AQM). This framework is capable of analyzing mix-zones in functioning terms, enabling measuring the anonymization quality of the data from them.

- In Chapter 7, we present the utility contribution, in which we present a new perspective on the utility with a Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT). This framework identifies and ranks smart cities

domains, applications, and services that can best leverage mobility data anonymized by mix-zones.

- Chapter 8 presents another privacy contribution regarding anonymization-based LPPMs. We present the proposed dynamic mix-zone scheme called k-DynMix, which adjusts the privacy level over time, considering traffic fluctuations. We validate this approach regarding coverage, privacy, and anonymization quality.

- Chapter 9 summarizes all the contributions addressed in this thesis. Additionally, this chapter lists new challenges to tackle, shedding light on new research directions regarding the location privacy applied to SMOD in smart cities.

Chapter 2

Location Privacy: An Overview

This Chapter presents a comprehensive study of location privacy on mobile networks. Specifically, we present the fundamentals of Location Privacy Attacks (LPAs) and Location Privacy Protection Mechanisms (LPPMs), detailing some approaches and metrics for evaluating them. Also, we highlight related work in this area. Finally, we present issues and new directions about location privacy and open questions when applied in a smart mobility open data context.

The Chapter is organized as follows. Section 2.1 introduces location privacy and the importance of mobility data protection in smart cities. Section 2.2 presents terminology and definitions found in location privacy. In Sections 2.3 and 2.4, details about LPAs, and we propose a taxonomy for it. Section 2.5 emphasizes the LPPMs; particularly, we focused on anonymization-based LPPMs. In Section 2.6, we bring the metrics used for measuring both location attacks and defense. In Section 2.7, we bring the discussion and future trends about anonymization-based LPPMs and LPAs, including in the smart mobility open data for smart cities context. Finally, we present our Chapter remarks in Section 2.8.

2.1 Introduction

The substantial adoption of data protection regulations, such as the General Data Protection Regulation (GDPR), is a driving factor that makes privacy a fundamental building block in smart cities. People generally demand privacy while searching for the comfort of the connected cyber world. In this sense, with the popularity and flourishing of VANETs and LBSs, for example, maps (Waze, WeGo, and Uber) and delivery services (Uber Eats, Zomato, and Rappi). The popularity of Location-based Social Networks (LBSNs), for example, Instagram, Foursquare, and Twitter, causes millions of users to share location-related data daily. They facilitate the lives of their users by providing context-aware services, promoting greater integration of users with the environment around them,

and increasing the quality of experience and service. Also, the location data are valuable in both the private market and public sectors.

In private market, LBSs commonly share this location data with third parties who collect user data for business, then analyze it for market insights, and sell to companies to make, e.g., targeted advertising campaigns, identify market trends, and so on. In public sector, they use this data to help better plan for the future and allocate resources. Specifically, transport and transit planning, purposes of traffic safety, identifying demographic trends, improving infrastructure and tourism [23], understanding the spread of disease and can be used to develop more effective strategies against natural disaster [35, 36, 37, 38].

However, there are several risks in sharing location data [32]. When sharing location data with service providers and repositories for public use, users no longer have control over managing and using sensitive information, which can be passed on to third parties or used for spurious purposes. For example, it reveals sensitive information about users like their home, workplace, social relationships and level, health status, political options, lifestyle, sexual orientation, religion, and much more [39].

The scientific community has broadly explored Location Privacy Protection Mechanism (LPPM) and Location Privacy Attack (LPA) to mitigate these privacy issues. The LPAs is very useful for understanding the behavior of LPPMs, while the goal of the LPPM is protecting the location and identity of users while still allowing them to use geo-located services [32, 33, 40]. Also, LPPM guarantees privacy levels when the mobility datasets are published for use for the public, as open data, and private organizations [31, 41, 42].

In front of this universe of proposals for attack and defense mechanisms, there is no consensus in the literature regarding terminology, metrics, types of attacks, and defense mechanisms for location privacy. In this Chapter, we propose categorizing the types of LPAs and their relationship regarding the sensitive data they attack. In addition, we present an LPPM category focusing on anonymized-based LPPMs, particularly in mix-zones. This study aims to identify the research frontiers of these works and possible research opportunities on location privacy within the context of smart mobility open data. In the remaining sections of this Chapter, we discuss some aspects of LPAs and LPPMs that are related to the contributions made in this thesis.

2.2 Terminology and Definitions about Location Privacy

In this thesis, we consider the following definitions as defined in [31] [43] [44]:

- *Anonymity*: the state of not being identifiable within a set of subjects, the anonymity set.
- *Unlinkability*: it means that within the system, from the attacker's perspective, the items of interest are no more and not less related to their observation than they are related to their *a priori* knowledge.
- *Identifiability*: the state of being identifiable within a set of subjects, the *identifiability* set.
- *Pseudonymous*: it is the action of using a pseudonym as an ID. Petit et al. [44] define the following requirements that pseudonyms should follow to ensure privacy requirements in networks: time-limited, uniqueness, pseudonym change block, and link to another identifier.
- *Obfuscation*: refers to reducing the knowledge precision - be that about location, time, or other feature - of the attacker regarding a mobile entity, intending to increase the entity's unlinkability.
- *Inference attack*: according to Krumm [45], this consists of analyzing data to gain knowledge about a subject illegally.
- *Adversary*: the unauthorized entity that can acquire some data exchanged as part of the LBS, which may pose a privacy threat. Many privacy protection techniques require accurate adversary modeling to guarantee their effectiveness [46].

2.2.1 Adversary Model

An important aspect of the success of a location privacy attack is defining the adversary model. Designing a well-defined adversary model allows for a better analysis of the outcome of privacy metrics, enabling the identification of vulnerabilities in privacy approaches and sensitive information on the dataset. For these facts, when we design a location privacy attack, we must consider the many factors about the adversary model, such as resources and knowledge available and goals to be achieved. Diaz et al. [47, 48] cited some important characteristics concerning adversary models:

- Attacker External: an adversary is not a part of the system but observes and extracts victim knowledge to generate an attack. For example, an adversary observes sensitive places where he has passed and uses this information to re-identify the victim's trajectories from anonymized datasets.

- Attacker Composition: an internal adversary is a part of the system, e.g., servers providing location-based services, energy providers in smart metering, or third parties controlling nodes in the system, such as third parties that apply privacy policies.
- Attacker Coverage: a local adversary acts on a specific part of the system, for example, a geographical location, a subset of nodes, or traces. A global adversary has access to the entire system.
- Types of Action: the active adversary can interfere with the system by adding, removing, or modifying information or communication. Otherwise, the passive adversary can only read, observe, and infer from his observations.
- Adaptation Capacity: a static adversary does not change his strategy and prior resources, independent of how the attack progresses. An adaptive adversary can adapt their strategy while the attack is ongoing, e.g., by observing system parameters.
- Prior Knowledge: adversaries can learn more about the system and potentiate the privacy attack. Prior Knowledge can be general domain-specific, about the world, or scenario-specific knowledge, e.g., a prior probability distribution or contextual information. Specifically, contextual information is any information that could be used to help the attack. We can cite as contextual information the users' home and work location, the number of users in an area at a given time, the relationships between different users, the linkage between a user's identity and their location, the location restrictions of an area, such as road networks and POIs. Prior knowledge could also be information obtained from the process before the attack, e.g., a Neural Network's training phase or a Markov chain's transition matrix. Usually, prior knowledge for these cases has a high computational cost.
- Resources: resources are intimately related to adversary prior knowledge. Refers to resources available that can be used in an attack, either a computational resource, structural, or contextual information. The efficient adversaries are restricted to Probabilistic Polynomial Time (PPT) algorithms in computational resources. Conversely, unbounded adversaries are not restricted to any computational models. Also, some types of resources use bandwidth or the number of malicious nodes available to the adversary.

In general, an adversary can be classified according to their resources and prior knowledge. If the adversary has too much prior knowledge and resources to perform an attack, he is called a strong adversary. On the other hand, the lack of prior knowledge, a poor amount of resources, or both define a weak adversary [49]. However, when we define prior knowledge when designing a privacy attack, it is essential to analyze the difficulty

level practically to get prior knowledge. In practice, many privacy attack proposals in the literature do not consider this aspect, resulting in useless attacks.

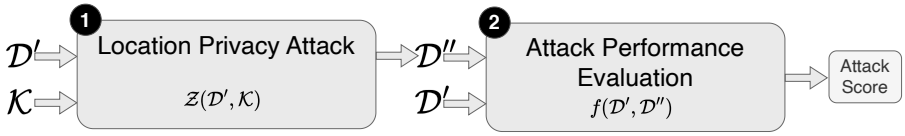


Figure 2.1: Location Privacy Attack Model.

2.2.2 A Generic Model of the Location Privacy Attack

Before introducing the types of attacks and their respective formal definitions, we present a general privacy attack model to clarify the objectives and resources related to an attacker. Let's consider the following sets:

- $\mathcal{D}$ is a released dataset that contains information about users and/or vehicles with respective unique identifier id, where ID represents the set of all unique $id \in \mathcal{D}$;
- $\mathcal{F}$ represents an LPPM and its privacy methods;
- $\mathcal{D}'$ is a resulting dataset of privacy mechanisms from $\mathcal{F}$ applied to $\mathcal{D}$, being $\mathcal{D}' \leftarrow \mathcal{F}(\mathcal{D})$;
- K is the attacker's background knowledge, which can be obtained through some auxiliary information, for instance, a geographic map. It is important to emphasize that K is not derived from $\mathcal{D}$;
- $\mathcal{Z}$ represents a privacy attack function;
- $\mathcal{D}''$ is information re-identified of a user with $\mathcal{Z}$.

The generic model of privacy attack $\mathcal{Z}$ is defined in Figure 2.1. As input, an attacker needs the protected dataset $\mathcal{D}'$ and the prior knowledge $\mathcal{K}$. As output, an attacker has a set $\mathcal{D}''$, composed of the information resulting from $\mathcal{Z}$. The content of $\mathcal{D}'$ varies according to the privacy attack goal. For instance, in de-anonymization attacks, $\mathcal{D}''$ has a set of user identifiers, whereas $\mathcal{D}'$ has a set of social relationships in meeting disclosure attacks.

From that output $\mathcal{D}''$, the attacker applies an attack performance evaluation, step 2 in Figure 2.1, which is a function f that indicates the amount of correct information obtained in $\mathcal{D}''$ by processing $\mathcal{D}'$, i.e., $f(\mathcal{D}', \mathcal{D}'')$. As a result of this evaluation, he obtains

an attack score which expresses how much information in $\mathcal{D}''$ is similar to the information in $\mathcal{D}$ (i.e., $\mathcal{D}'' \approx \mathcal{D}$). Many performance evaluation functions are used in the literature, with the accuracy rate being the most commonly used (see Section 2.6.1).

2.3 Location Privacy Attacks

In this section, we present a general discussion about Location Privacy Attacks (LPAs) applied in VANETs and Internet of Vehicless (IoVs) that also can be used in the context of smart mobility open data to identify latent information about users, such as their home, work, routines, and even identity. We provide a general definition for each attack category, basic functioning, mathematical formalism, and prominent papers.

2.3.1 De-Anonymization (DEA)

De-Anonymization (DEA), also known as *re-identification*, originates from database research and extends to various contexts, such as location privacy. It is a general term to define the identity of individuals [50, 51, 52, 53], vehicles [54, 55], and entities trajectory [28, 56] whose information is recorded as records within a de-identified database through data linkage techniques [50, 57, 58].

Formally, the de-anonymization of $\mathcal{D}'$ consists in connecting its containing sensitive information K to obtain the set of users' id [59], denoted by $\mathcal{D}''_{id}$. Equation 2.1 presents a general definition of de-anonymization.

$$\mathcal{D}''_{id} \leftarrow \mathcal{Z}(K, \mathcal{D}') \tag{2.1}$$

In the location privacy area, there are several approaches related to privacy attacks that use de-anonymization to define identity attacks. However, in some cases, this terminology is not adequate. There are privacy attacks that do not necessarily bring up just the identity of entities but the identification of trajectories, social relationships, social status, religion, locations, mobility prediction, etc. Thus, in the following, we present the related work of these privacy attacks, which use de-anonymization terminology in their papers, but in some situations, these attacks can be considered as a subclass of de-anonymization.

▶ Related Studies about De-Anonymization (DEA)

.

In location privacy, there are many privacy attacks considered a kind of DEA, but it does not explicitly re-identify the user's identity. Another way is to re-identify sensitive places, trajectories, transportation modes of the victims, and so on. Examples of Re-identification attack derivations include Points of Interest Attack, Uniqueness (U), Trajectory Classification, and Tracking Attack.

Trajectory Classification (TC) (or *transport mode detection*) is a research area that classifies the transport type from in a mobility dataset, being this anonymized or not. In general, transport mode techniques can be used as a preliminary step for generates re-identification attacks. For example, by executing a transport mode detection attack on a protected mobility dataset, the victim's trajectory set can be narrowed down to a single vehicle category, enabling the execution of an inference attack. Some studies classify vehicle traces, such as cars, bicycles, and trucks [28, 30], while others go further by also classifying pedestrians [60].

Zan et al. [28] developed a de-anonymization attack which used mobility trace classification to identify vehicles protected by mix-zones. In this attack, the anonymized traces were classified by vehicle type: car, motorcycle or truck. They claimed that vehicles of different types have distinct acceleration/deceleration profiles and produce different mobility traces profiles that can be used as fingerprints, making it possible to classify them and reconstruct the traces. They used machine learning classification, and used a simulator to produce the experiment and freeways instead of an urban environment, which is a less challenging environment to perform the re-identification.

Another proposal applying acceleration and deceleration approaches to classifying vehicles was proposed by Sun et al. [30]. They have presented a machine learning approach with GPS data to identify multi-classes of vehicles: passenger cars, single trucks and multi-trailer trucks. The overall result of the rating for all three vehicle classes is about 75%. The challenge in question is to differentiate the two categories of trucks since they have similar mobility models.

Das and Winter [60] proposed a characterization model based on a multi-input and multiple-output Fuzzy system capable of identifying types of mobility, such as walking, bus, tram, and train, from smartphone trajectories. This technique has proved efficient about other models based on machine learning, which need training and are limited in interpreting wrong trajectories when submitted to unreliable training.

De-anonymization can be realized not just with mobility data from datasets but with another kind of source, e.g., datasets of the apps installed in smartphones [61], even obtained from several types of non-sensory and sensory on the smartphones [62].

Sekara et al. [61] showed that it is possible to capture users' behavior in data from smartphone applications, taking into account the time. Usage behavior was used as a fingerprint to re-identify users. The data was collected from the Google Play Store, which

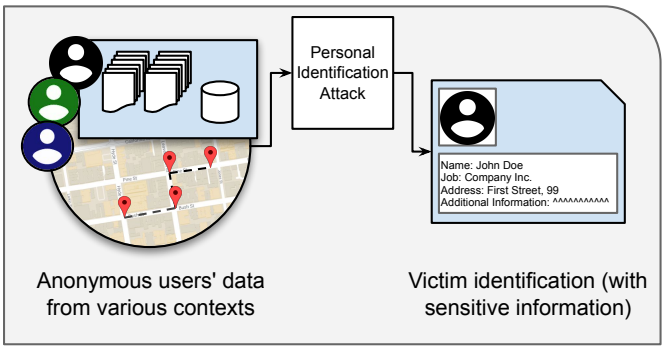


Figure 2.2: Personal Identification Attack.

had about 12 months of data from 3.5 million users. The technique re-identified 91.2% of users, using the strategy inspired by [63].

Mosenia et al. [62] proposed a De-anonymization approach called PinMe, which combines non-sensory/sensory data stored on the smartphone (e.g., the environment's air pressure and device's timezone), and public information (e.g., elevation maps) for user's location even when all location services, e.g., GPS, are turned off. The accuracy of PinMe was estimated by the user's location during four activities (walking, traveling on a train, driving, and traveling on a plane). They have also proposed countermeasures to mitigate these vulnerabilities.

2.3.2 Personal Identification Attack (PIA)

Personal Identification Attack (PIA) is also called as *Single Identity Attack*, consists of identifying a user (or a set of users) based on their home address or determining a person's gender, education level, and religion through an anonymous trace [33, 45, 64, 65]. Figure 2.2 shows this attack.

For example, be $\mathcal{Z}$, the function that groups the users' trace according to the location and time-frequency. If the attacker has background knowledge $K_{i,t}$ about religious temple location i and the schedules of each section t, it is possible to identify the religion of the victim (or several victims), denoted by $\mathcal{D}''_{id}$, applying the following function [45, 66].

$$\mathcal{D}''_{id} \leftarrow \mathcal{Z}(K_{i,t}, \mathcal{D}') \tag{2.2}$$

▶ Related Studies about Personal Identification Attack (PIA)

PIA is one of the first attacks explored in human mobility. It can cause several damages since it is possible to match a single user with his own mobility. For instance,

Tockar [67] presented a study that shows how to track celebrities in New York City, using public "paparazzi" pictures and data from a public dataset of New York cabs. PIA is explored through social network user's information, trace mobility, and home location discovery. A variety of studies is presented as follows.

Krumm [45] presented a study that shows what an attacker can do with a large volume of location data from several individuals. The author introduced four algorithms, based on probabilistic and cluster heuristics, that identify personal home location and, consequently, user's identity through GPS location data gathered from volunteers. Mahmud et al. [68] inferred the user's home location by extracting information from the user's tweets on the Twitter social network. They proposed an algorithm that uses statistic and heuristics classifiers, combining several tweets from the same user.

PIA also has been combined with other attacks such as Points of Interest Attack (POIA) to identify personal information about users. Gambs et al. [58] created a probabilistic mobility model called Mobility Markov Chain (MMC), considering that a person transits between different POI following a probabilistic function. Li et al. [65] used data provided by Mobile Social Networks (MSNs) to extract user's POIs and, consequently, to infer demographic information about the users. Gu et al. [69] inferred home location identification using LBSNs data. They developed a trust-based unified probabilistic model that models edge in LBSNs based on POIs, social relationship, and user-centric data.

2.3.3 Aggregated Presence Attack (APA)

Aggregated Presence Attack (APA) is also called as *Multiple Identity Attack*. This strategy consists of identifying an expected set of users, denoted by $R_{i,t}$, being $R_{i,t} \subset \mathcal{D}'$, which were in the same region i at time t [33, 70, 71, 72]. Figure 2.3 presents this attack.

Exemplifying, let's consider that an attacker, called $Trudy$, has a background knowledge K about a user, called *Alice*. He knows that *Alice* usually goes shopping i with her friends *Bianca* and *Catherine*. If $Trudy$ tracks *Alice*'s position in a mall and observes that two other users are close to *Alice* during a considerable time t, $Trudy$ could infer that these users are *Bianca* and *Catherine*. Equation 2.3 brings a general definition of APA.

$$\mathcal{D}''_{id} \leftarrow \mathcal{Z}(K, R_{i,t}). \tag{2.3}$$

▶ Related Studies about Aggregated Presence Attack (APA)

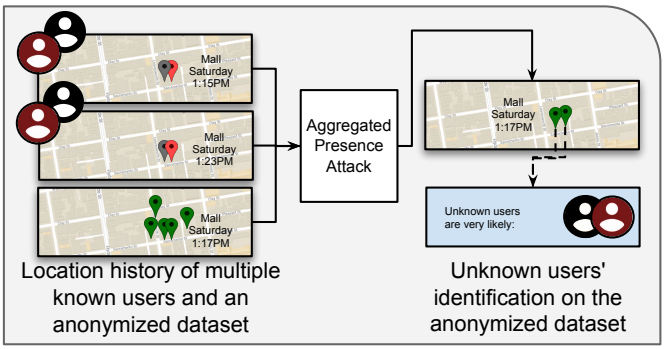


Figure 2.3: Aggregated Presence Attack.

Aggregated Presence Attack (APA) can be treated as extension of PIA. In this attack, large datasets contribute to better de-anonymization results. Moreover, just as in PIA, Points of Interest Attack (POIA) can be used as a facilitator to APA. For instance, Shokri et al. [71] have proposed metrics to evaluate the privacy of users based on location data using APA. Specifically, they modeled the users' mobility as a Markov Chain and applied an APA to infer the number of users at a specific place and time. To discover the specific places, POIA could be used.

Pyrgelis et al. [72] formalized the concept of membership inference and developed an APA through machine learning algorithms. In this approach, the attacker's prior knowledge is considered as a training set used to infer membership groups.

Social network data also are used as prior information to APA. Wondracek et al. [73] exploited group membership information from social networks to identify a people's group or, in the best scenario, uniquely identify a person. They developed a history stealing-based attack that allows an attacker to probe the browser history of a user, revealing personal social network information.

2.3.4 Meeting Disclosure Attack (MDA)

Meeting Disclosure Attack (MDA) is also called as *Link Prediction Attack* and *Social Connections Inference Attack*, it is an attack that infers the relationship among two or more entities [33, 71, 74, 75, 76]. To obtain this information, an attacker observes different aspects of users, such as discovering social relationships by meeting the frequency of the entities. That is, given a region R_i and a set of users $\mathcal{D}$, how many times the users in $\mathcal{D}$ met in the region R_i and how close the users in $\mathcal{D}$ are from each other. Figure 2.4 shows this attack.

Combining this different information, it is possible to infer social relationships SR,

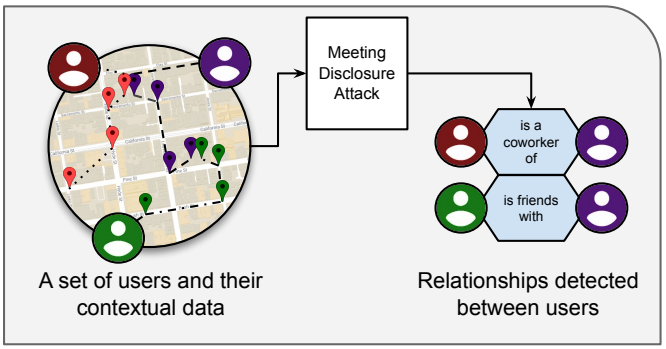


Figure 2.4: Meeting Disclosure Attack.

for instance, friends, relatives, and coworkers. Equation 2.4 presents a general definition of MDA, where SR is a set of social relationships inferred, $\mathcal{D}$ is a set of users considered and R_i is a region observed.

$$SR \leftarrow \mathcal{Z}(D, R_i) \tag{2.4}$$

▶ Related Studies about Meeting Disclosure Attack (MDA)

MDA can be used for many different applications than just privacy attacks. These include predicting a friendship in a social network, relations between entities in a knowledge graph, affinities between users and items in a recommender system, and potential biological interactions between drugs and diseases [77].

Shokri et al. [71] have defined a formal model of location privacy framework for mobile users, named Location-Privacy Meter. Additionally, it introduced a generic attack-based Markov Chain that can be used to answer information disclosure attack questions. Specifically, with the generic attack, it was possible to derive attacks as MDA, APA, and Position Attack (PA).

Backes et al. [78] proposed a social relationship inference attack without prior knowledge about existing relationships. For this, advanced learning features were used to automatically summarize the users' mobility features. The attack was compared with existing approaches and could predict the social relationship of any two individuals with 13% to 15% improvement over the state-of-the-art. They also proposed three defense mechanisms: hiding, substitution, and generalization. The results showed that the mechanisms of hiding and substitution outperform the generalization. In addition, hiding and substitution achieved a trade-off between utility and privacy. The first preserved utility while the second provided better privacy.

Yu et al. [79] considered both contextual information and relationship propagation for measuring an effective relationship between pairs of users that meet with low-frequency in a MDA scenario. Specifically, they proposed a graph embedding method to account for the relationship propagation, and a frequency-based method considers the pairwise

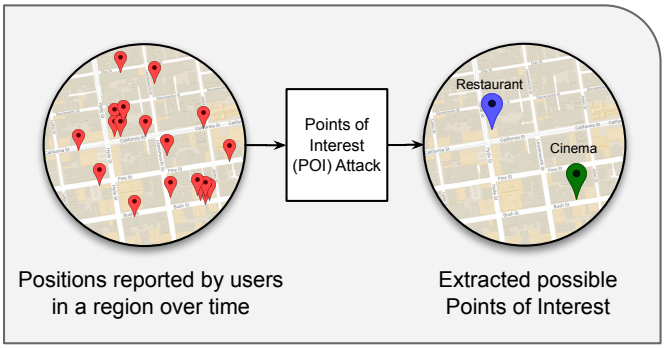


Figure 2.5: Points of Interest (POI) Attack.

relationships independently. Second-order random walks were used to select the neighborhood for each user to reduce noise and better capture the user interaction. Also, it was proposed that the external venue information of meeting venues be encoded in the user graph. The method was evaluated with Gowalla and Brightkite datasets [80], and the recall and precision results outperformed the state-of-the-art methods, including approaches in [78, 81].

2.3.5 Points of Interest Attack (POIA)

Points of Interest Attack (POIA) uses location points that people commonly stay to characterize users' profiles [82] and re-identify them. POIs are geographic regions that users visit frequently due to their interests, such as stores, gas stations, restaurants, and so on. We can find two types of POIs in the literature. POI at the user level are locations where only one user u has stayed frequently, e.g., home, workplace, etc. POI at group level, also called *Region of Interest (ROI)*, are regions where more than one user (n users) commonly have stayed, like a shopping mall, gas station, football stadium [83, 84].

POIA is widely used as a preliminary step to generate other types of LPAs, for instance, Tracking Attack [54, 55, 85], Personal Identification Attack [58, 65, 69], and Aggregated Presence Attack [71, 72]. For example, in many Tracking Attack approaches, POIA is used to identify POIs, and based on a temporal sequence, it is possible to link them, reproducing the trajectory of a user. In these scenarios, there are a considerable number of researches related to human mobility [46, 68, 69, 82, 86, 87, 88, 89, 90], however, vehicular mobility is few explored [66, 91]. Figure 2.5 presents the intuition of this attack.

Consider a given set of regions R, a period t, and $\mathcal{D}$ users. An attacker can be interested in discovering what the most visited places P by $\mathcal{D}$ at t, expressed in Equation 2.5.

$$P \leftarrow \mathcal{Z}(R, \mathcal{D}, t) \tag{2.5}$$

POIA differs from the APA and MDA. The focus of the POIA is to infer regions where the users concentrate, not necessarily at the same time, and this information can be used for to identify users. As presented, the APA focuses in to identify population frequency in a specific region, i.e. not focuses on regions, but on people that stayed in a region at same time. On the other hand, MDA is used to identify social relationship between people, vehicles, and so on.

▶ Related Studies about Points of Interest Attack (POIA)

Maouche et al. [82] proposed a POIA called AP-Attack, based on which aggregates user mobility into a probability distribution acting as a fingerprint of user mobility. The attack was validated with state-of-art POI attacks on four real mobility datasets, protected with obfuscation LPPMs based on perturbation (GEO-I [2] and Promesse [92]), and hiding (W4M [93]). AP-Attack outperformed these attacks by more than 27%. They concluded that there is no LPPM fit-to-all users protection, causing the protection level of LPPMs intrinsically dependent on the mobility data. New directions for POIs protection is Multi-LPPM user-centric approaches.

Kaplan et al. [94] have presented an attack for discovering if an unknown private trajectory passes (or does not pass) through Region of Interest (ROI). The ROI could be areas used for a privacy attack, e.g., the adversary could learn if the victim visited a hospital. The attack uses a set of known trajectories and their pairwise distances to the unknown trajectory. In detail, the proposal generates a set of candidate trajectories that resemble the private trajectory. After this explores properties of the candidate trajectories like ROI that could be in the private trajectory. However, the adversary needs many public trajectories as input to identify some similarities with private trajectories.

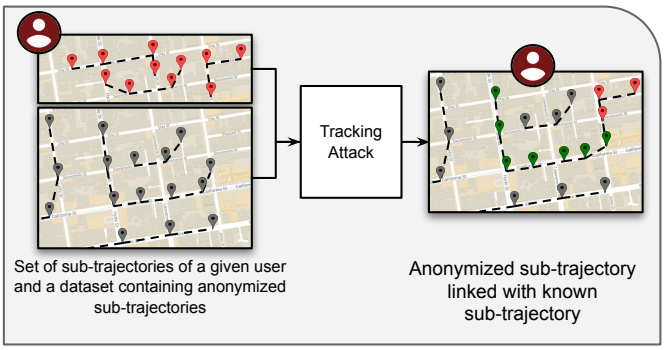


Figure 2.6: Tracking Attack.

2.3.6 Tracking Attack (TA) and Uniqueness (U)

Tracking Attack (TA) is also known in the literature as *Reconstruction Attack* [71], *Trajectory Re-identification* [95], *Sequential Tracking Attack* [96], *Linkage Attack* [97], and *Location Tracking* [98]. This attack constructs a partial or entire sequence of the events to develop a user trace [71]. This attack can also be the "gateway" to others that target a specific entity [56]. This attack can be used in human mobility [53, 63] in vehicular mobility [56], or both [58]. Figure 2.6 shows the intuition of this attack. For example, consider P points or sub-traces the attacker collected from the victim. The Tracking Attack consists in to reconstruct the original traces $\mathcal{D}''_{traces}$ that were anonymized into $\mathcal{D}'$, i.e.:

$$\mathcal{D}''_{traces} \leftarrow \mathcal{Z}(P, \mathcal{D}') \tag{2.6}$$

In VANETs, there are many strategies of Tracking Attacks (TAs) [99]. It is worth noting that each TAs depends strictly on adversary knowledge, such as access beacon messages, positioning of victims, LPPM locations, and so on. The definition of the prominent types of TAs found in literature is defined as follows [99, 100]:

- Global Passive Adversary (GPA): This type of adversary listens to all communications in the network with a radio transceiver and tracks vehicle locations without actively intervening. GPA collects data over a wide area and analyzes location and beacon messages to uncover sensitive vehicle information [101, 102].

- Local Passive Adversary (LPAd): A local passive adversary refers to an entity in a specific area and observes network traffic or communications without actively interfering or manipulating them. This adversary aims to analyze patterns and infer users' locations or movements based solely on the information passively collected [103].

- Syntactic Linking Attack (SyLA): In this attack, the attacker connects an old pseudonym of a vehicle with its new one. If only one vehicle changes its pseudonym at a given time, such as Vehicle V switching from V_1 to V_1 at time t, the attacker can link V_1 to V_1, exposing the vehicle's identity [103, 104, 105].

- Semantic Linking Attack (SeLA): This attack involves extracting helpful information from a vehicle's safety message, such as predicting its future location. Using this data, the attacker can link a vehicle's pseudonyms if multiple vehicles change pseudonyms at once [103, 105].

- Scrambler Attack (ScA): A scrambler attack occurs at the link layer, where the attacker uses scrambling values to correlate messages regardless of pseudonyms,

undermining privacy. This attack becomes effective when the vehicle operates with static beacon frequencies [106].

- Cheating Attack (ChA): In this attack, a compromised vehicle broadcasts beacon messages with a flag set to 1, forcing nearby vehicles to change pseudonyms. The attacker is among the k neighbors of the network passing as a valid vehicle and can track a target by exploiting the pseudonym change [107].

- Sybil Attack (SyA): The attacker creates numerous false identities to disseminate misinformation in the network. Next, the attacker creates virtual vehicles with fake details and false identities. Then, the attacker disrupts regular traffic and collects location data of targeted vehicles [108, 109].

- Timing Attack (TiA): Also known as Time Inference Here, the attacker does not modify the messages' content but introduces delivery delays. By studying the timing information of the vehicle's movements, the attacker can link pseudonyms to real vehicle identities [110, 111].

- Transition Attack (TrA): The attacker tracks vehicle movements at intersections and estimates the likelihood of each turn based on past observations. The attacker can predict the vehicle's route by monitoring vehicles at entry and exit points [112, 113].

- FIFO Attack (FFA): In a First-In-First-Out (FIFO) attack (or Location Inference), vehicles passing through a mix-zone in a fixed sequence allow the attacker to link their new pseudonyms to old ones. The predictable timing makes the mapping process easier for the attacker [114, 115].

- Local Active Adversary (LAA): This adversary operates in a specific region and actively intercepts vehicle communications. The attacker can track vehicle movements within the localized area by exploiting parameters such as transmission range [116, 117].

- Speed Attack (SpA): The attacker obtains the entering and exiting speed of the vehicle in the anonymous set and calculates the mapping probability of the vehicles in the anonymous set according to the average speed. In this way, the new and old pseudonyms with the same average speed have higher mapping probabilities [100].

- Maximum Likelihood Tracking Attack (MLTA): This attack is a derivation of a tracking attack whose objective is to find the most likely traces for all users, given the observed traces [71].

The *Uniqueness (U)* is a metric that measures the level of the singularity of mobility register in location datasets. This metric verifies how many location points are needed to

identify a trace within a dataset. Uniqueness is a metric used to verify the identification level in traces. Only uniqueness does not guarantee re-identification of users, pseudo-anonymized mobility traces themselves do not disclose the identity of a user [51]. Thus, the uniqueness can be used as external knowledge and combined with other approaches to inferring the identity of users [118, 119]. For instance, we can use the knowledge obtained with *Uniqueness (U)* and apply a DEA to link each individual inside a training set of mobility traces to its anonymous counterpart inside a testing set. In this thesis, we consider the uniqueness as a part of TA, because it is possible to recover traces efficiently, even if the data is obfuscated [51, 52, 63].

▶ **Related Studies about Tracking Attack (TA)** and **Uniqueness (U)**

Tracking Attack (TA) explores many strategies and other location privacy attacks as a preliminary step for reconstructing traces of vehicles and users. LPAs commonly used are Points of Interest Attack [54, 55, 85], and Uniqueness [51, 63, 120, 121].

Sui et al. [85] proposed the parking spots attack, in which the adversary considers the parking habits of taxi drivers extracted from the taxi mobility traces to re-identify victims. As a countermeasure, they presented a protection scheme that exchanged sub-trajectories of the most relevant parking-point taxis. One of the restrictions of this attack is that the attacker needs prior knowledge of the opponents and habits of the taxi drivers to infer the parking spots.

An attack on individual privacy that uses independent data sets is called a composition attack. Cecaj et al. [122] presented a compositional tracking attack based on georeferenced social network data. They proposed a data fusion-based re-identification that used georeferenced social networks to re-identify mobile users from an anonymized Call Description Records dataset.

Chang et al. [54] claimed that trajectories have user profile indicators, such as preferences and usual behaviors, which are unique and with little change over time, aiding in re-identifying vehicles. The indicators found the stops of interest (e.g., malls and gas stations), as well as road segment preferences. They re-identified the victim's trajectories by comparing the indicators found in trajectory segments collected by the victim's observation and those of anonymized mobility histories. In this study, they used the dataset of taxicabs from Shanghai and Shenzhen. Results showed that it was possible to re-identify the anonymous trajectories with sub-paths 8 to 9 days in size with an accuracy of 96.64% and 77.03% for Shanghai and Shenzhen, respectively.

Murakami has been concerned with producing tracking attacks in which the attacker holds a small amount of location information, called by the sparsity problem in de-anonymization attacks [55, 123, 124]. Murakami et al. [123] proposed two approaches for mitigating data sparsity problems in de-anonymization attacks: tensor factorization and group sparsity regularization. Specifically, they presented a new training model, used

in the Markov Chain, based on sparsity tensor factorization to train the personalized transition matrices. This approach captures the spatial group structure from a small amount of training data. The results showed that the training method using tensor factorization outperforms the re-identification based on the Maximum Likelihood estimation method. In [55], Murakami proposed a LPA that combined Regions of Interest (ROIs) attack that uses a small amount of training data, with up to only one location for re-identifying users. The method is based on the Jensen-Shannon divergence between two probability distributions about the traces trained and anonymized. The proposed method outperformed the state-of-the-art, tensor factorization-based Markov Chain and the random approach; both use a large amount of data for training.

Murakami and Watanabe [124] have used tensor factorization (or matrix factorization) to accurately estimate personalized transition matrices and a population transition matrix from a small amount of training data. The technique was compared with the Maximum Likelihood (ML) Estimation method in both the personalized matrix mode and the population matrix mode. The experiments used four datasets: Geolife [125], Gowalla [80], epfl/mobility dataset [126], and Rome/taxi dataset [127]. The approach proposed proved that the method significantly outperforms the ML estimation method in all of the four datasets.

In location privacy literature, few studies have been concerned with identifying gaps in de-anonymization attacks. One of them was proposed for Wang et al. [53], that identified gaps in state-of-art de-anonymization algorithms. They submitted 7 de-anonymization algorithms for re-identifying users in cellular and social network datasets. The insights from the analysis proposed a framework with four new algorithms to evaluate practical factors, for example, tolerate spatial and temporal mismatches, location contexts, and user-level errors. The algorithms achieved 17% in terms of the hit-precision.

An efficient tracking attack strategy is to explore mobility characteristics for rebuilding trajectories anonymized for mix-zones [95]. Mattos et al. [56] proposed a trajectory re-identification based on the characterization of the road preferences that occur in urban environments, in particular in taxi cabs. The re-identification technique has time complexity in the order of $O(E + V\log V)$, where E represents the number of edges (intersections) in the graph of the roads of the city, and V the number of nodes (streets). The technique has been validated with a taxicabs dataset of the San Francisco city-USA, protected by mix-zones, containing about 231,230 trips, which re-identified up to 100% of the anonymized trajectories.

Li and Li [128] proposed a trajectory-tracking attack using the Matrix Completion method, which leverages incomplete location data from RSUs to reconstruct users' movement paths using matrix completion techniques. This approach is the first of its kind. By creating a sampling matrix of vehicle positions, adversaries can accurately recover trajectories, enhanced by hierarchical clustering, to minimize the number of RSUs needed.

Simulations show that ATM effectively tracks precision, adaptability to variations, and performance under different sampling rates.

Regarding place selection for generating attacks, Saini et al. [101] proposed an attacker placement strategy that can be used to select locations for eavesdropping and allow attackers to collect helpful information for linking pseudonyms. The strategy is to select locations where it can listen to beacons from many vehicles, and pseudonym changes are likely to occur.

Uniqueness research has been introduced by De Montjoye et al. [63], which analyzed the quantification of the uniqueness in human mobility traces from call logs. Based on the change in the granularity of the dataset's spatiotemporal data, the authors presented a formula for calculating the uniqueness of human mobility. From this, proposed an inference model in which four spatiotemporal points were sufficient to identify up to 95% of individuals from a call logs dataset containing human mobility data of 1.5 M users with 15 months of data.

Boutet et al. [51] also presented a uniqueness analysis on human mobility in three types of user data sources (GPS, WiFi, and GSM antennas) focused on Lausanne (Switzerland) and Lyon (France) cities and compared two uniqueness approaches: probabilistic [63] and deterministic based [51]. The results showed that only four spatio-temporal points from the WiFi, GSM, and GPS traces are necessary to identify 94% of the individuals on both datasets uniquely. In addition, a POI-based re-identification approach was proposed, applying different LPPMs, including spatial filtering, temporal cloaking, adding spatial noise to mobility data, and using a generalization. Next, they analyzed the impact of these mechanisms, both the uniqueness of users' mobility traces and the outcome of the de-anonymization. They concluded that spatially obfuscating mobility data is not enough to protect users and that classical LPPMs cannot protect users.

Further to human mobility, some studies have investigated uniqueness in vehicular mobility [120, 121]. Rossi et al. [120] presented a user re-identification technique that exploits the uniqueness of GPS data, even if not in the mobility dataset. Specifically, the technique calculates the minimum distance between a set of geo-localized points of the victim and the points of the anonymized traces. The victim's trace is the one with minimal distance between the points. Three real-world datasets were applied: epfl/mobility [126], CenceMe [129], and GeoLife [125]. They concluded that two spatial points were needed to identify nearly 100% of users. However, they did not provide details on which LPPM was applied.

Tan et al. [121] also investigated the uniqueness of GPS data. Specifically, they identified differences in privacy issues between LBSs and their derivation for vehicles, called Vehicle Location-Based Services (VLBS). They claimed that vehicles are restricted to roads, so their trajectories are unique and possibly re-identified. The proposed heuristic model was able to re-identify trajectories of anonymized vehicles with up to 95% accuracy

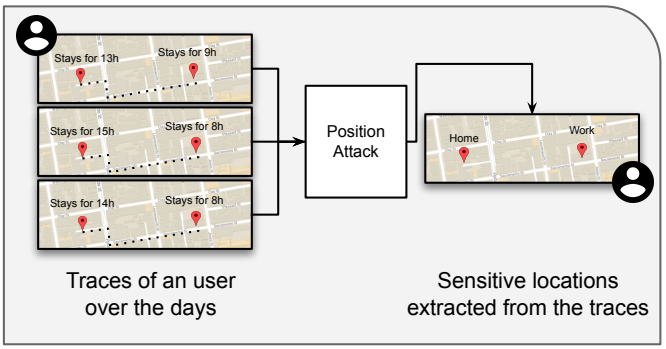


Figure 2.7: Position Attack.

from four spatiotemporal points of these vehicles and urban maps. However, the authors considered in this attack that the attacker would have full access to the VLBS data server and that the attacks would occur by collecting the spatiotemporal points of the victim's car.

Some studies concerning uniqueness have used multiple data sources, for instance, application installation data of smartphones [51, 61]. Sekara et al. [61] used the application installation data to capture users' behavior in data from applications collected from smartphones, taking into account the time and using this information as uniqueness.

2.3.7 Position Attack (PA)

Position Attack (PA) is also called *Sensitive Place Attack*, is used to discover and/or to predict locations related to a specific user or a set of users, for instance, home and work [46]. This type of attack needs special attention to be neutralized because sensitive places can induce the discovery of personal information about a specific user.

Let's consider that an attacker got infiltrated in a LBS and observed the trace data of a set of users, but these data are obfuscated. Over the days, the attacker notes that a user u spends most of the day regularly in a specific obfuscated area. Using a geographic map, the attacker perceives that this area is surrounded by mountains, where there is a deluxe hotel. This way, the attacker could infer that the user u probably works at the hotel. Figure 2.7 presents the intuition of this attack. Equation 2.7 presents a general definition of PA, where $K_{l,t}$ is the attacker's background knowledge related to a set of locations (or regions) l and a time t, and P is a single or a set of positions (or sensitive places) related to users in $\mathcal{D}'$.

$$P \leftarrow \mathcal{Z}(K_{l,t}, \mathcal{D}') \tag{2.7}$$

▶ Related Studies about Position Attack (PA)

Krumm [45] introduced four heuristics of Position Attack aiming to identify individuals' homes, which then would be used to infer their personal identifications. The heuristics are Last Destination, Weighted Median, Largest Cluster, and Best Time. To prove their point, they collected a GPS points dataset from 172 users, which also contained ground truth information about the users' homes. Additionally, they introduce cloaking, noising, and rounding as obfuscation methods to the position attack and explore their effectiveness. The most successful heuristic was Last Destination, being able to re-identify 12.8% of the individuals' homes. They concluded by stating that even simple heuristics, such as the ones presented, can be used to obtain private information and increase the attacker's knowledge about its victims.

Gu et al. [69] introduced a trust-based probabilistic model to predict users' home locations using social and check-in information. Considering the social spectrum, they assumed that individuals tend to form social relationships with others closer to their home location. Similarly, considering the check-in information, they also assumed that users tend to visit venues near their home locations. To validate, they analyzed a dataset extracted from Foursquare, a LBSN, containing information about user check-ins and users' social connections. They evaluated using only data from users with check-ins and all users in the dataset. For the first one, considering a radius of 100 meters or less, they obtained an accuracy of 92.1%. For the second one, considering the same radius, they obtained an accuracy of 63.1%, outperforming the state-of-the-art algorithms in the literature.

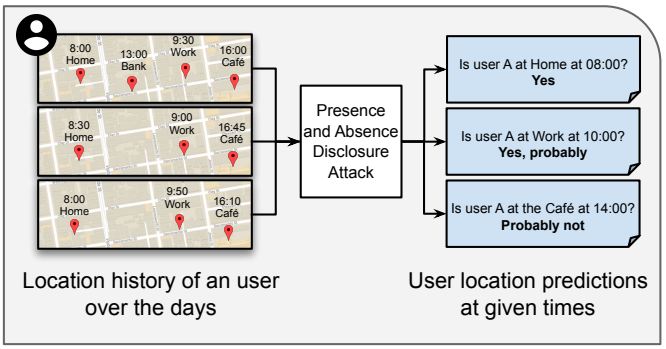


Figure 2.8: Presence and Absence Disclosure Attack.

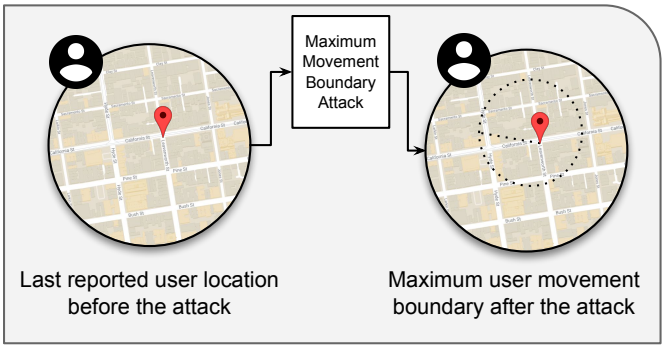


Figure 2.9: Maximum Movement Boundary Attack.

2.3.8 Presence and Absence Disclosure Attack (PADA)

Presence and Absence Disclosure Attack (PADA) is also known as *Position and Discrete Time Attack*, this approach tries to determine whether or not a user u is present at a place l at a specific time t [70, 130, 131]. For instance, empty homes are good targets for burglary; even further, physically attacking a person requires you to know exactly the location where the person is. Figure 2.8 shows this attack.

For example, let's assume that an user u visits a bank b weekly at very close periods of time. If an attacker wants to obtain some sensitive information about u to harm him, the attacker could apply a sensitive place attack and predict the probability of u being at the bank in a given period of time t. The Equation 2.8 defines the PADA, where Pr_u is the probability of a user u being in a set of places P at a time t.

$$Pr_u \leftarrow \mathcal{Z}(P, \mathcal{D}', t) \tag{2.8}$$

▶ Related Studies about Presence and Absence Disclosure Attack (PADA)

Similar to the Position Attack, the Presence and Absence Disclosure Attack (PADA) aims at affirming, with a certain level of accuracy, if a victim is (or is not) at a specific location, e.g., their home or work location. Although very close in their definitions, the Presence and Absence Disclosure Attack (Section 2.3.7) aims at tracking a user through his/her movements, knowing every location they are as they move. On the other hand, the Presence and Absence Disclosure Attack (PADA) focuses on a single location, with no intention to track the victim as they move. Attackers have many motivations to perform these types of attack, such as ensuring that a victim is not at home (or even workplace) so thieves can come in [130], or the contrary, in case the victim is the target.

Most strategies for this attack rely on publicizing personal data on location-based social networks like Facebook, Twitter, etc. The attacker lures the victims' latest updates (e.g., posts, pictures, and tweets) to infer their whereabouts. To obtain information about

the victim's current whereabouts, the attack strategy must have as recent data as possible about the victim. Moreover, it must be able to produce its output as quickly as possible, considering that the victim may move during the processing, invalidating the result.

2.3.9 Multiple-query Attack (MQA)

In Multiple-query Attack (MQA), the adversary tries to compromise the actual location of the query sender with the help of a series of two or more spatial queries involving different cloaking regions. Cloaking region is a kind of LPPM that obfuscates the location data of the user, adding noising or distorting his data, based on the region where this user is supposed stay [31]. The multi-query attack can be performed in three different forms:

- *Maximum Movement Boundary Attack (MMBA)*: The attacker calculates the maximum boundary area in which the user/victim could have moved between two succeeding position updates or queries, making it possible to infer its possible position [97, 31]. This attack works for Cloaking Region (CR). For example, consider CRs A and B, which are reported by user u at timestamps t_A and t_B, respectively. Without loss of generality, let $t_A < t_B$. Let be v the maximum user velocity, and let $\delta t = |t_B - t_A|$. Determine if there is any location $x \in A$ from which the user cannot reach some location $y \in B$, even by traveling at maximum speed v. Formally, an attack is successful if:

$$\exists x \in A \; s.t. \; \forall y \in B, d(x, y) > v\delta t \tag{2.9}$$

  The $d(x, y)$ results from a distance measure function between location x and y. Figure 2.9 shows the intuition of this attack.

- *Region Intersection Attack (RIA)*: tries to reduce position inaccuracy generated for obfuscation method, e.g., cloaking region (CR). The attacker uses several imprecise position updates or queries from a user to calculate their intersection. From the intersections, the attacker can infer privacy-sensitive regions where the user is located [132]. For example, consider n CRs denoted by R regions which are reported by users U over some time $\delta t = |t_b - t_a|$, where t_a, t_b are timestamps of beginning and end of the analysis, respectively. The region intersection attack works by inferring the location where most users are. The intersection of R_n regions in the δt,

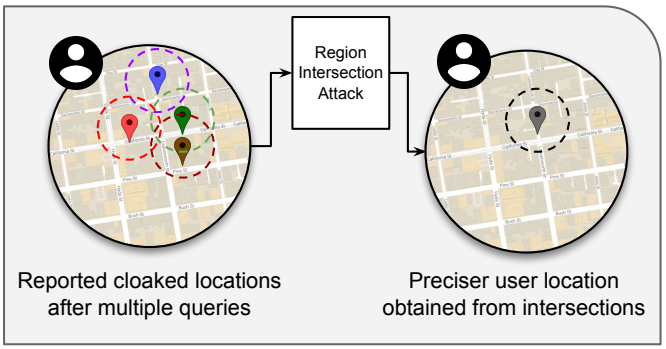


Figure 2.10: Region Intersection Attack.

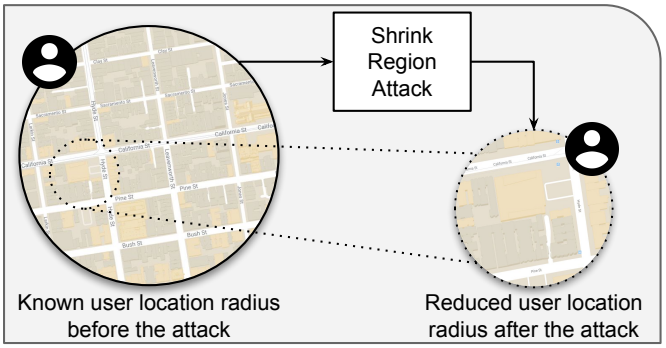


Figure 2.11: Shrink Region Attack.

generated by the function Z at δt time. That is,

$$Z^t = \bigcap_{i \in S}^{t} R_i^t \mid S = \{1..n\},\ t = \{1..\delta t\} \tag{2.10}$$

In this attack, the actual location of the user may be revealed through an intersection operation of anonymity sets from different queries that generate overlapping CRs. Figure 2.10 presents this attack.

- *Shrink Region Attack (SRA)*: occur when two or more different subsequent queries cause shrinkage of an obfuscated region (OR), which can reveal the user's identity and position [132]. The idea is that an attacker monitors consecutive queries and the corresponding members of the k-anonymity set. The attacker can infer which user sent an initial update or query if the member of the set changes, revealing his location and identity. For example, consider that users A, B, and C are in different positions in the city. User A is moving to different parts of the town and sends two queries to the LBS server X. Consider that a k-anonymity approach is used by A that generates the k-anonymity set (A, B) for the first query and the anonymity set (A, C) for the second query. So, X can infer that A initially issued the query, inferring his position and identity. Figure 2.11 illustrates the intuition of this attack.

▶ Related Studies about Multiple-query Attack (MQA)

We discuss the three types of attacks that occur through a series of user location queries: Maximum Movement Boundary Attack (MMBA), Region Intersection Attack (RIA), and Shrink Region Attack (SRA). Although there are differences in their formulations, all the attacks have as main objective increasing the precision regarding the current user location.

Ghinita et al. [97] presenting two MMBA approaches. One approach is the attacker has background knowledge about the sensitive locations on the map. They also have presented privacy-preserving algorithms to oppose the application of the attacks, showing a complete system architecture of how personal location data can be collected and used without compromising the user's privacy.

Talukder and Ahamed [132] discussed both RIA and SRA, and surveyed cloaking solutions in the literature regarding their efficiency and their complexity. However, they have not explored the attack strategies for both types further.

2.3.10 Future Mobility Prediction (FMP)

Future Mobility Prediction (FMP) is a kind of attack whom the attacker can perform a location prediction to obtain with probability p where the user will be at a specific time t in the future [88, 89, 132, 133, 134, 135]. Some proposals predict the next location of vehicles [136, 137, 138, 139], and people [140]. In general, approaches are based on likelihood functions and historical data to predict the next venue, road, and so on. For example, Krumm [139] proposed a Markov approach for predicting a vehicle's near-term future route based on its past route. Figure 2.12 presents the intuition of this attack.

Given a sequence of traversed road segments as $X(t)$, where t is a discrete-time variable and $X(.)$ is a road segment (e.g., an integer unique among all the road segments). Therefore, the configuration of the road segments is $\{...X(-2), X(-1), X(0), X(1), X(2)...\}$, where $t < 0$ represents the past, $t = 0$ is the present and $t > 0$ the future, which we intend to predict. Considering the Markov model, it gives a probabilistic prediction over future road segments based on past road segments, but the events are independent. For example, if we want prevent the next road $P[X(1)]$, it is necessary consider previous events, e.g., $X(0), X(-1), ...$ That is, $P[X(1)] = P[X(1)|X(0), X(-1), ...] = P[X(1)|X(0)]$. Even the Markov model can predict beyond just the next road segment. That is, we can build $P[X(2)|X(0)]$. In general, we can build an n^{th} order Markov model $n \geq 1$ to predict the m^{th} next encountered segment ($m \geq 1$). Equation 2.11 brings its definition.

$$P_n[X(m)] = P[X(m)|X(n-1), X(-n+1), X(-n+2), ..., X(0)] \tag{2.11}$$

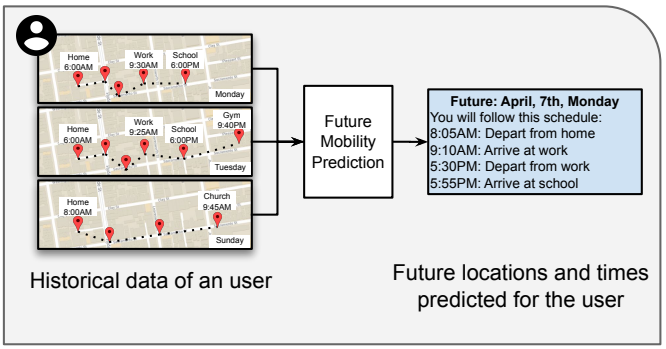


Figure 2.12: Future Mobility Prediction

▶ Related Studies about Future Mobility Prediction (FMP)

From the viewpoint of FMP on mobile networks, Zhang and Dai [141] have presented a survey of mobility prediction and characteristics regarding movement predictability, prediction outputs, and performance metrics. Specifically, they proposed a future mobility prediction architecture composed of six components: data source, required information, prediction algorithms, prediction outputs, performance metrics, and categories of applications. Relating to state-of-the-art approaches, techniques were identified, including Bayesian Networks, Artificial Neural Networks, Markov Chain, Hidden Markov Model, and data mining based on different kinds of knowledge. Also, they presented open research challenges concerning the fifth-generation mobile system and some potential trends.

Boukerche et al. [142] presented a perspective for future localization systems for VANETs. They argued that it is possible that future localization systems use some Data Fusion technique to improve the position information for vehicles, and it is accurate and robust enough to be applied in VANETs critical applications. Also, data fusion techniques can be used to compute positions based on inaccurate estimations. However, both proposals above focus little on mobility prediction regarding threats and location privacy.

Studies about FMP that consider location privacy can be grouped based on individual historic data [89, 143], and crowding movement historic data [144, 145]. The first class consists of FMP analysis of individual users rather than all users, in which predicting his future movement is possible, revealing knowledge hidden under mobility history. However, the information known by a user does not work to predict other users. The second class predicts movement based on the similar mobility pattern extracted from crowding people instead of worrying about individual movements, i.e., random movements from the entire body of mobile users' profiles. This approach is efficient in predicting the movement of new individuals that have the same pattern. The drawback is that it is possible to degrade the clustering process if the users' movement is very random.

One of the first studies on mobility prediction focused on location privacy was proposed for Sadilek and Krumm [134]. They presented a prediction model of future

locations within a time window of one hour. The model, named Far out, used combined Fourier analysis to find periodicities in human mobility and Principal Component Analysis (PCA) based on Singular Value Decomposition (SVD) to extract strong, meaningful patterns from location data, which are subsequently leveraged for prediction. The results showed accuracy in predicting the location of up to 93%.

Kuruvatti et al. [143] have combined Markov Chain and context information extracted from diurnal mobility to improve the future user location prediction accuracy (e.g., cells, routes). From this, it's possible to anticipate running into a coverage hole and initiate context-aware resource allocation in a network. Specifically, from the trajectories, specific landmarks of users were identified, such as the origin and destiny of users, and this information was used to improve the accuracy of mobility prediction.

Augir et al. [135] combined the semantic information and Bayesian networks to improve the accuracy of the next check-in of users. They evaluate geographical and semantic location privacy with data of 1065 users collected from tweets generated from Foursquare. The results showed that the median accuracy of predicting the next check-in is between 100 and 150 meters in six major cities. They concluded that semantic information improves prediction success, even if obfuscation techniques protect the user data.

2.4 A Taxonomy of Location Privacy Attacks

This section looks at LPAs with an analytical study about type mobility, victim data types, and data sources used on attack proposals based on related work. Figure 2.13 summarizes the LPAs and categorizes them by sensitive data that is a threat to obtain when using these attacks. Each attack category identifies specific types of latent information about entities like identity, location, or social relationships. However, there are classes of attacks aiming to identify more than one piece of information about the entities. For instance, De-Anonymization (DEA) is a generic term for many proposals of attacks, such as Points of Interest Attack (POIA), Tracking Attack (TA), and Position Attack (PA). De-anonymization attack category has attacks to recover identity [58], tracking paths of entities [146, 53], and also identify the social relationships [78]. Also, Multiple-query Attack (MQA) is a generic term for Shrink Region Attack (SRA), Region Intersection Attack (RIA), and Maximum Movement Boundary Attack (MMBA), that aim to infer the location of the entities[1].

[1]For simplicity, we have omitted sub-classes of some attacks, but details about them are defined in Section 2.3.

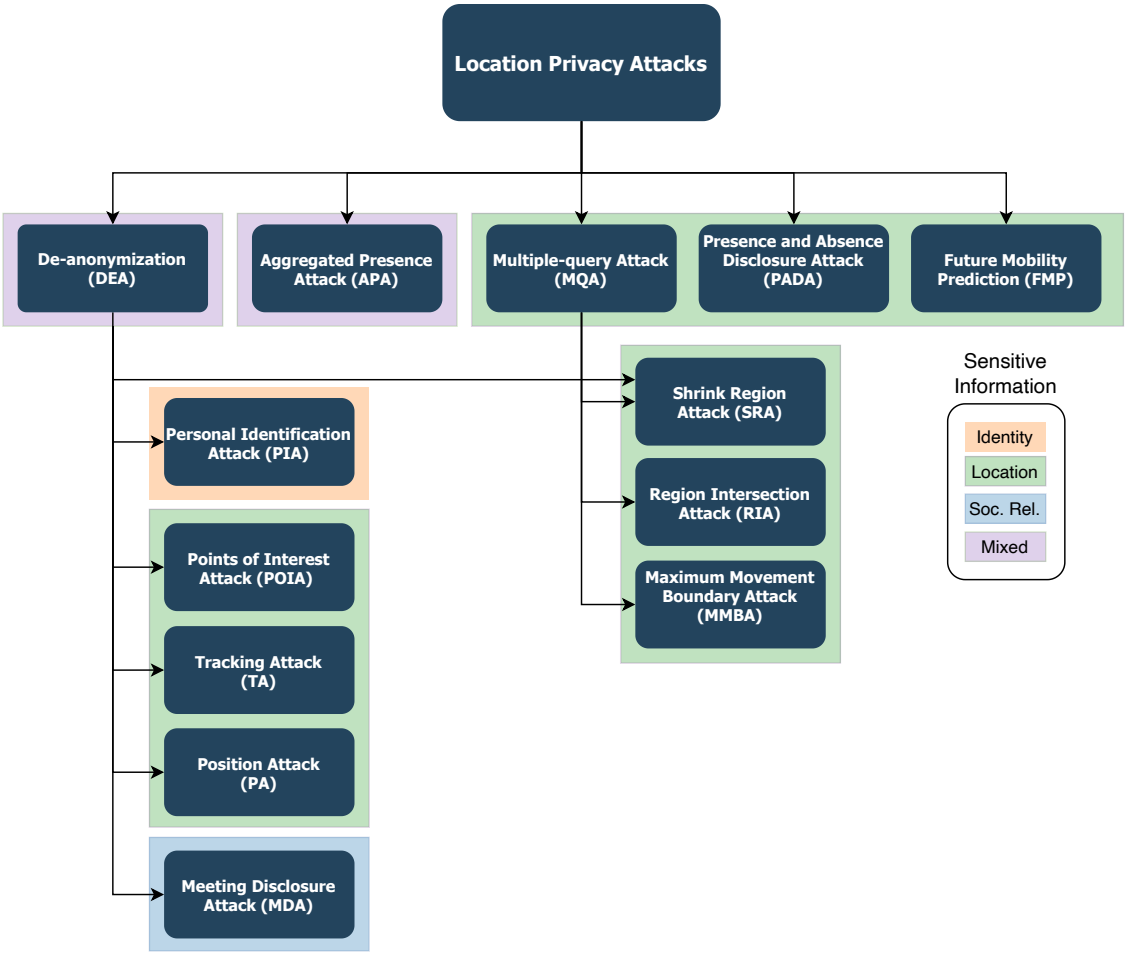


Figure 2.13: Categorization of LPAs classified by user latent information.

Some attacks may be genuinely aimed at recovering location, such as PA, which seeks to identify the location of victims. However, some attacks can obtain more than two latent pieces of information, such as the location and identity of victims, such as TA, which can track a user (location), identify a user's trajectories, and consequently infer their identity.

Another key aspect is that most research focuses more on humans than vehicular mobility. Or in studies that use both human and vehicular mobility. We see the absence of LPAs dedicated to vehicular mobility, which may also give insights for future research.

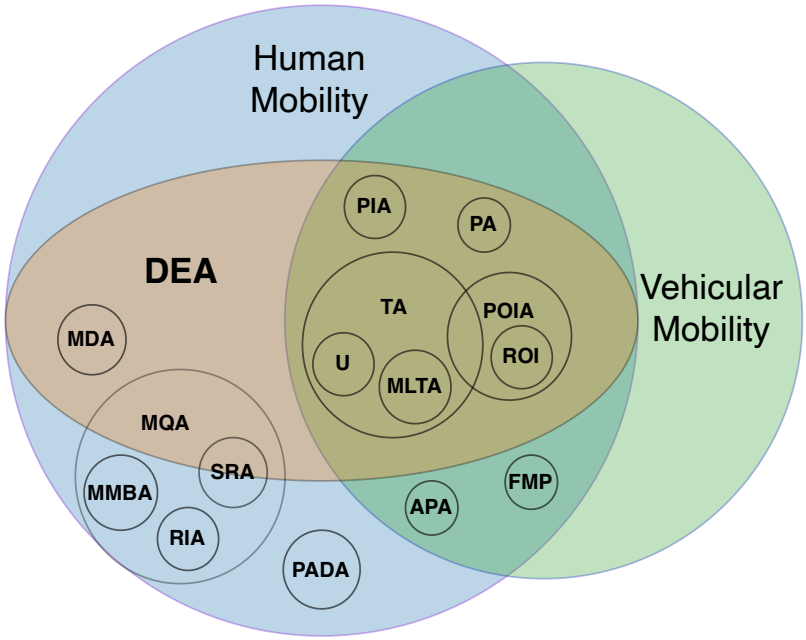


Figure 2.14: A taxonomy of the location privacy attacks in the mobility data.

2.4.1 Location Privacy Attack vs. Mobility Types

The diagram in Figure 2.14 represents the disposition of LPAs proposals defined in Section 2.3 according to the type of mobility used in these proposals. The big circles on the left and right in the figure represent the type of mobility (human or vehicular) in which the attack was applied. For example, for MDA papers, the authors have used only datasets of human mobility. On the other way, PIA there are human and vehicular mobility studies. Notably, we have the prevalence of proposals in LPAs that use human mobility about vehicular mobility, using the human mobility dataset. PIA, PA, TA, Uniqueness (U), POIA contain studies found in both human and vehicular mobility. We can also observe that none of the attack category studies deal exclusively with vehicular mobility.

Although de-anonymization is a generic term of attack in location privacy, in practice, not all attacks are considered de-anonymization because there are attacks that are not directly associated with revealing the identity or location. For instance, the aim of FMP is to infer possible entities' paths or locations. Thus, attacks that are a derivation of de-anonymization are: MDA, PIA, PA, SRA, TA, Uniqueness, and POIA. The only DEA category that features exclusively human mobility papers is MDA.

In MQA, specifically in Shrink Region Attack (SRA) there are some proposals considered de-anonymization, and others are not. Because some works in SRA deal with attacking the victim's position rather than his identity. In this sense, Presence and Ab-

sence Disclosure Attack (PADA), Maximum Movement Boundary Attack (MMBA), and RIA are attacks that use human mobility and do not pertain to a de-anonymization derivation. Although APA and FMP do not belong to de-anonymization, they contain works that use human and vehicular mobility.

Another point to consider is that there are several proposals in the literature where an attack is a composition of other attacks. For example, many papers use POIA as an intermediate step to TA. Or a prior analysis of the dataset uses the uniqueness metric. Additionally, we can find in the literature derivations of TA and POIA, called Maximum Likelihood Tracking Attack (MLTA) and ROI respectively.

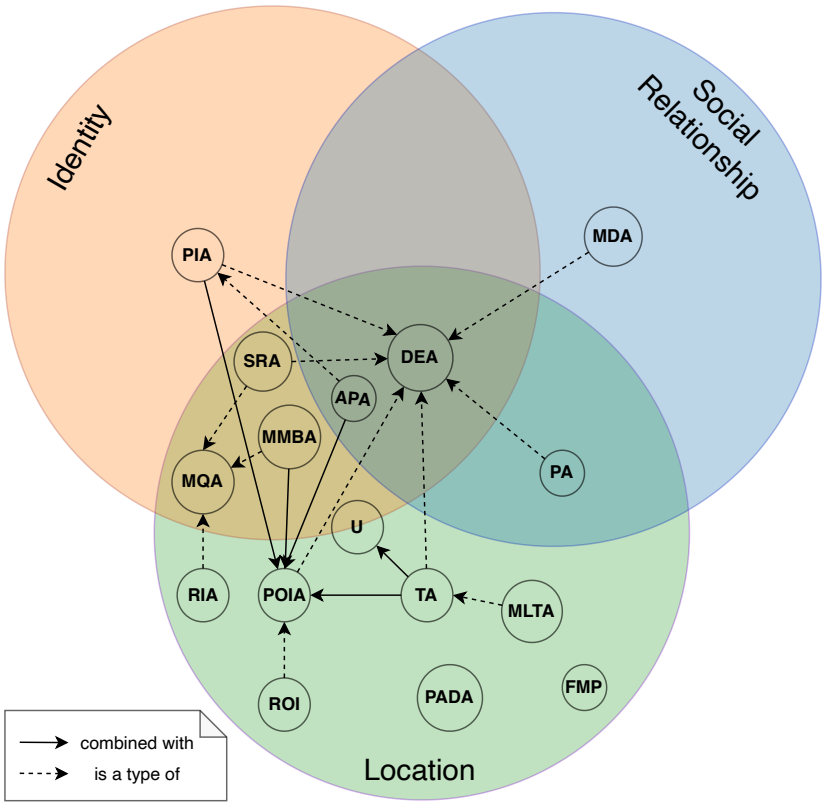


Figure 2.15: Relation between location privacy attacks and data type information

2.4.2 Location Privacy Attacks vs. Victim Data Types

The diagram denoted in Figure 2.15 provides an insight into the relationship between LPAs and information latent of entities revealed by attacks. We can find attack proposals in which DEA and APA focused on revealing the identity, location, and social

relationships. Attack proposals aiming to recover the identity and location are a derivation of MQA. Just Position Attack contains studies to heal the social relationships and location of victims. There is little research on LPAs that recovers identity and social relationships.

Another critical point is the evolution of the amplitude of attacks on the victim's sensitive data types. We can identify the evolution of the attacks in terms of information to be obtained from the victim. For example, de-anonymization was originally an attack on associating user identity with protected data, but several proposals in the literature have expanded this concept to location and social relationships. This fact also can be observed with Uniqueness, whose initial purpose was to use it as a metric to measure the level of privacy of protected data [63]. However, it can be considered an attack against the identity and also the location of users [118, 119, 120, 121].

Many LPAs use attacks as a preliminary step to perform other attacks. This is the case with POIA, which is considered an intermediate process to generate TA, PIA, MMBA, and APA. These attacks can use POIA for producing robust attacks that consider semantic information obtained with POIA [147, 148]. In this sense, the scientific community has focused on developing efficient POIA algorithms [149, 150].

The LPAs studies have used datasets with only one transport mode. In this way, there is a gap of LPAs strategies that use multi-modal transportation datasets to have trajectories of different vehicle types, or even a single trajectory can have different vehicles taken by a user. Another opportunity is the design of composition attacks, in which an adversary uses independent anonymized releases to breach privacy. Multi attacks- types composed of two or more location attacks- must be explored to understand and develop more robust LPPMs.

2.5 Location Privacy Protection Mechanisms

Location privacy is a fundamental concern in mobile networks. Mobile entities in the urban environment massively yield location data, valuable information that can be used for various purposes. The location data can be used to obtain accurate information from LBS, stored in the data market, monetized to the target markets, and used as open data, where these data are published in open repositories and used publicly by many areas for urban development. However, the location data brings privacy concerns. Threats using LPAs it is possible to identify sensitive data of users, routines, home, workplaces, their identity, and spatial and temporal location. To address the LPAs, LPPMs has been developed to protect the location and identity of the users.

The LPPMs are well-studied approaches in mobile networks, such as VANETs. However, for an urban environment dynamics like Smart Cities, in which the smart mobility building block yields mobility data with different characteristics about VANETs, it is mandatory to study the potentialities of classic LPPMs in terms of protection. Although there are few studies about LPPMs designed for Smart Cities, we discuss two well-studied categories related to this type of mechanism to protect the location and privacy of location data. The LPPMs can be classified in techniques anonymization and obfuscation-based.

2.5.1 Obfuscation-based LPPMs

Obfuscation-based LPPMs are to protect the location, rather than identity like anonymization-based LPPMs. Obfuscation methods reduce the accuracy and precision of the spatiotemporal information of location data [71]. Some obfuscation techniques are based on *k-anonymity*. Obfuscation-based LPPMs can be divided into four techniques: perturbation, dummy, reducing precision, and location hiding.

- **Perturbation-based approaches (adding noise)**

Perturbation-based approaches are LPPMs that cause mobility data to be effectively changed in space and time before being sent to an LBS. Thus, the LBS will not easily know the actual location of the mobile entity. At this point, the trade-off between privacy and utility should be considered. Data needs to be distorted enough to be protected, but if the data is too distorted, LBS cannot be used. Most LPPMs work by randomly adding noise to the raw mobility data (adding noise), changing the latitude/longitude or trace segments with another one (confusion approaches) [151]. *Location Hiding* and *Location Replacement* are derivation of Perturbation-based approaches.

Location Hiding is a derivation of the perturbation approach. In the location hiding mechanism, every event is independently eliminated (i.e., its location is replaced by $\emptyset$) with probability λ_h where h is the location hiding level [71, 78]. Many studies have used this approach to hide sensitive locations, e.g., home, work, shopping malls, etc. *Location Replacement* is a mechanism that replaces a certain proportion of locations with others to cheat the adversary [78]. The data types to be replaced can be check-in locations, pieces of traces, and GPS points [152].

- **Differential privacy-based LPPMs**

Introduced by Dwork [153], differential privacy is a model that defines a formal and provable privacy guarantee. The basic idea assumes that an answer to the query on the dataset should be almost the same whether or not a single element is present inside the dataset. In other words, differential privacy ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis. The advantage over *k-anonymity* is that differential privacy is resilient to the external knowledge an attacker may have.

The relationship between differential privacy and other fields, such as quantitative information flow, has been explored. Alvim et al. [154] have investigated the relationship between information theory and differential privacy. They showed how to model the query system based on an information-theoretic channel and compared the differential privacy with min-entropy leakage. They concluded that differential privacy implies a bound on the min-entropy leakage but not vice-versa. Also, they investigated the utility of the randomization mechanism and the proximity between randomized and real answers. They concluded that differential privacy implies a bound on utility. From these results, they proposed an optimal randomization mechanism that provides the best utility while guaranteeing ϵ-differential privacy.

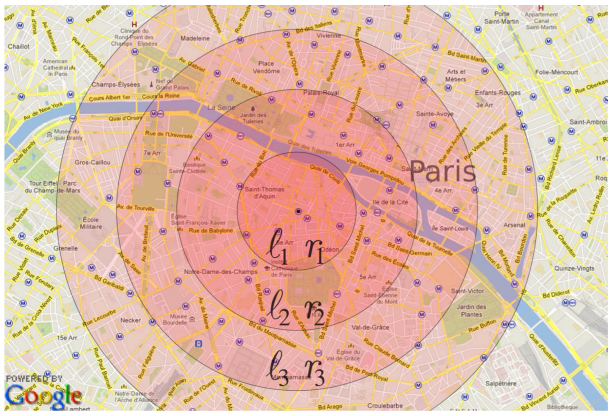


Figure 2.16: Variation of privacy level in relation to r. Source: [2]

.

Geo-indistinguishability (GEO-I) approach proposed by Andrés et al. [2], is a specialization of differential privacy for location privacy based in [155, 153]. Specifically, GEO-I is a formal notion of location privacy that guarantees privacy levels limiting the likelihood of two points to be reported locations of the same actual location within a given radius. In other words, the privacy level l is inversely proportional to the radius. A user has *l-privacy* with radius r if any two l locations at a distance at most r produce observations with "similar" distributions, where the "level of similarity" depends on l. In this way, an LPPM reaches GEO-I if the probability of reporting an obfuscated location z is similar for two close locations l and l' and the more different the further if l and l' are distant from each other. Informally, a simple way to specify the privacy requirements to the user is by variables l and r, and the ϵ-geo-indistinguishability is $\epsilon = l/r$. The ϵ

ensures a level of privacy for l within r and a proportionally selected level for all other radii, where the privacy levels decrease with the radii, see Figure 2.16.

▶ Dummy-based LPPMs

Dummy-based LPPMs similes fake nodes and generates privacy, adding the mobility data and false information with true information to the LBS server, which can not distinguish the real node. Dummies information can be users, GPS positions, locations, and trajectories [32, 156, 157]. They are strictly related to environmental factors, mainly the communication model, and both the node's trajectory and track flow.

▶ Generalization-based LPPMs

Generalization-based approaches preserve privacy by reducing the precision of the position/location information in which an entity mobile sends a merged region to the LBS server, making identifying the actual entity location challenging. In location privacy, the generalization approaches synergize with *k-anonymity*. For instance, it is possible to obfuscate the location sent by entity mobile at the same time that other k entities are in the obfuscated region, called *cloaking area*, then the adversary will have uncertainty in matching between the entities and their actual locations accurately [26, 98, 158].

One of the prominent papers on generalization was proposed by Ardagna et al. [159, 160]. They presented a generalization technique that sends a circular area instead of the precise user position to the LBS server. Also, it presented the relevance metric associated with the degree of privacy introduced into a location measurement. The purpose of the relevance metric is to measure the trade-off between the required accuracy of the LBS server location and the needs of users while minimizing the disclosure of personal location information.

▶ Protocol-based LPPMs

Unlike anonymization or obfuscation techniques that hide or alter location data, protocol-based LPPMs focus on transmitting, sharing, and managing data in real-time communications [32, 161]. These protocol-based approaches are typically more tailored to specific tasks (e.g., finding nearby vehicles) but can offer superior privacy protection. These approaches depend substantially on encryption techniques to provide robust privacy assurances for particular use cases.

2.5.2 Anonymization-based LPPMs

Identity privacy refers to preserving a user's real identity when this user is in VANETs, where vehicles constantly exchange information about their location, speed, and real-time traffic conditions. Often, this information is linked to the identity of the drivers. The problem is that if the user's identity of these drivers is not protected, it will be possible to apply DEAs, particularly TAs, and track the individual, compromising their privacy. One way to mitigate a TA is through constant pseudonyms changing.

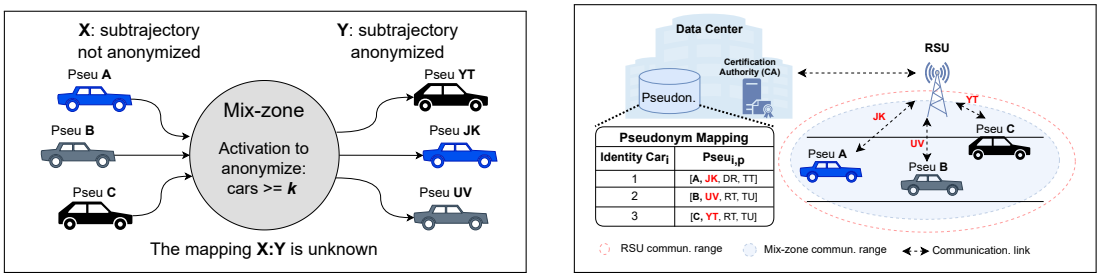
The anonymization-based LPPMs are techniques to break the links between identity and location information without creating noises or distortions by pseudonym-changing mechanisms. Pseudonyms are fictitious or altered names users use to hide their real identities. The pseudonyms changing refer to techniques used to protect the privacy of vehicles and their drivers by changing or exchanging the pseudonyms that vehicles use for communication within the VANET. A vehicle uses a pseudonym for a time (to ensure stable communication), then the pseudonym is changed according to the adopted privacy scheme [113]. The European standard ETSI TS 102 867 [162] recommends changing a pseudonym every five minutes, while the American SAE J2735 [163] standard advises changing it every 120 seconds or after 1 kilometer, whichever occurs later.

2.5.3 Mix-zones: A Privacy Protection Scheme

The Mix-Zone is one of the main anonymization-based LPPMs that uses the pseudonym changing strategies to anonymize mobility data in various mobile network types and mobility contexts, such as human mobility [164], vehicular ad-hoc network (VANET) [165], Internet of Vehicles (IoV) [166], tracing contact data [167], Location Based Service (LBS) [168], and in the Internet of Drones (IoD) [169]. In this thesis, we explore the mix-zones applied to vehicular networks.

▶ Mix-zones Schema

Mix-zones are anonymization areas defined by a radius r where entities change their pseudonyms according to a trigger function (e.g., when reaching a minimum of k entities simultaneously inside it) [167, 170]. Consider the vehicle set $\mathcal{V} = \{V_1, V_2, ..., V_i, ..., V_m\}$ where $1 \leq i \leq m$. Each V_i that makes a trip T_a composed of GPS points and $T_a \in \mathcal{T} = \{T_1, T_2, ..., T_a, ..., T_n\}$ where $1 \leq a \leq n$. Mix-zone M_s is a geographical area of k-anonymity



(a) Mix-zone workflow. (b) Pseudonym system model.

Figure 2.17: Mix-zone schemes: 2.17a Mix-zone toy-example with $k = 3$, where three cars A, B, and C enter it and meet the minimal k. When the cars exit the mix-zone, they receive new pseudonyms (JK, UV, and YT, respectively) without any association with previous ones (i.e., an external observer does not know the mapping function). 2.17b when a vehicle $i = 1$ with pseud. A is within the mix-zone M_s with at least k vehicles inside it, i can opt to change its pseudonym. So, each vehicle receives a symmetric session key from the RSU, which initiates the pseudonym changing and the symmetric key updates. After the pseudonym changing of i (pseud. $A \rightarrow JK$), the RSU communicates to Certification Authority (CA) about the change and updates the set on the mapping database from $Pseu_{1,1} = A$ to $Pseu_{1,2} = JK$.

that vehicles go through, causing their pseudonyms to be modified [171]. When a moving vehicle goes inside a mix-zone with radius r, its trajectory will be sliced into two sub-trajectories delimited by different pseudonyms – one corresponding to the part before the mix-zone and the other corresponding to the part after the mix-zone (see Figure 2.17a). Vehicles change their pseudonyms inside the mix-zone if there is a set of vehicles A, denoted as the anonymity set, present in it simultaneously and obey the condition of $|A| \geq k$, where k is anonymity mix-zone parameter [164, 172]. A vehicle population $P_M \subset \mathcal{V}$ can be through the M and can or not be anonymized depending on the condition cited. A trip T_a can travel through multiple mix-zones $\mathcal{M} = \{M_1, M_2, ..., M_s, ..., M_n\}$ in its path and, consequently, has its pseudonym changed multiple times, resulting in several sub-paths delimited by different pseudonyms.

Figure 2.17b denotes the system model of classical pseudonym changing [170]. The RSUs connect vehicles to the infrastructure networks giving them access to the internet and various remote services. The data center is responsible for generating a legal pseudonym for each vehicle by a Certification Authority (CA) and storing the mapping of pseudonyms and the driver's real identity in a dataset. In the mix-zones scheme, a vehicle V_i equipped with a GPS device has to register with CA and preload a pseudonym set $Pseu_{i,p}$ with pseudonym p in the pseudonym set $\mathcal{P} = \{1, ..., m\}$. Details about the pseudonym change process and thread-based mix-zones algorithm can be found in [173].

An essential aspect of the mix-zone is the definition of the privacy level k. Because it is closely related to the anonymity set's size ($|A|$) – a measure of the level of location privacy available in the mix-zone M_s at that time –, it is responsible for the activation of the M_s, that is a step for pseudonym changing of vehicles. Also, k specifies the privacy levels at any given time, which enables users to be protected while using a LBS. For

example, a user may determine that having an anonymity set of at least 15 individuals is enough to ensure their pseudonyms can not be linked between different application areas. Thus, the k parameter of the mix-zone M_s can be defined as 15. Users might opt to withhold sharing their location data with an application until a mix-zone reaches a certain anonymity threshold, i.e., $k = 15$ [164].

The critical point about k-anonymity approaches overall, including the mix-zones, is defining the k value [164, 174]. Previously defined the k, it's possible to bring benefits to users' privacy. For instance, consider a mix-zone M_s that protects against a LBSs in which the k was previously set with an average of $|A|$. In this way, when a user signs up for a new location-based service, the middleware can get the k value from the nearby mix-zone, in the case M_s, thus estimating the level of privacy that can be expected. This information can be shared with users before they decide whether to use a new location-based service [164, 175]. Generally, the classic mix-zones schemes the k is defined previously [164, 171, 175].

Three key questions related to the design of mix-zones-based mechanisms are: (1) What are the optimal areas for deploying mix-zones? and (2) How many mix-zones are required to ensure adequate location privacy? (3) How can mix-zones perform well in terms of privacy and coverage in relation to anonymized data? The questions (1) and (2) define the challenge known as the mix-zones placement [175, 176], and the question (3) concerns the mix-zones performance [177, 178]. Below, we classify the mix-zones and their pseudonyms changing schemes, and next, we present some relevant works on mix-zones and their positioning.

▶ Flavors of Mix-zones and Pseudonyms Changing

The literature has comprehensive discussions about mix-zones and pseudonym-changing schemes where research communities have tried to improve the mix-zones' performance, resulting in many proposals based on different features [99, 104, 179, 180, 181]. Although some proposals consider certain regions for pseudonym change, they do not use k-anonymity concepts from classical mix-zones. However, this study shows it is a substantial part of forming dynamic mix-zones. Thus, a consensus about terminology, evaluation, or standardization needs to be reached. We categorize mix-zones within the following perspectives that can be combined in a proposal: pseudonyms changing type, infrastructure, map, pseudonym change policies, application domain, execution mode, cryptography protocol usage, modality cooperation, and hybrid LPPMs.

An important aspect that categorizes mix-zones is the pseudonym-changing mechanism subjacent. There are several pseudonyms changing policy found in the literature, and some approaches are based on group-signature, trigger, reputation, silent period, and encryption techniques. The pseudonym changing by group-signature means vehicle

Figure 2.18: Mix-zones and Pseudonym Changing Taxonomy.

groups are formed on the fly in regions and use generated keys instead of authority-certified keys, in which vehicles can exchange their pseudonyms successively [182, 183]. In group-signature techniques, a broadcast message is signed with a group key and the sender key, anonymizing the message's sender in the group.

Trigger-based schemes regroup solutions where the pseudonym update occurs when certain conditions (context) are satisfied or at a specific time and region, such as the number of a vehicle's neighbors, position, velocity, acceleration, heading direction, the age of pseudonym, and so on [184, 185, 186, 187]. Trigger-based schemes refer to solutions where the pseudonym update occurs only when specific conditions are met, such as under certain contexts, times, or within designated zones. The trigger can be an external trigger (e.g., when k-neighbors surround a vehicle) or requires inter-vehicle coordination (e.g., a simultaneous pseudonym change) [113, 182]. In contrast, trigger-free solutions involve pseudonym updates that are not restricted to specific conditions or strategies, allowing pseudonym changes to occur more flexibly without predefined triggers.

Reputation is a feature in which users earn reputation "credit" by implementing a change of pseudonym [188, 189]. Then, users with higher reputations can generate

actions, such as creating anonymization regions and proposing pseudonym exchange.

Silent-based mix-zones is a cooperation model schema in which vehicles remain silent by ceasing the transmission of beacons or other identifying messages and changing their pseudonyms in a synchronization way [105, 113, 190, 191]. This break in communication prevents attackers from linking a vehicle's identity to its location once it leaves the mix-zone. Also, in the cooperation mode, vehicles can use encryption in their communication messages, named encryption-based mix-zones [170, 192, 193].

Infrastructure-dependent refers to the mix-zones family where the RSU infrastructure or particular device is necessary to participate in the pseudonym change process. This involvement may include tasks like setting up mix-zones or broadcasting road-related messages [187, 189, 194]. Map-dependent mix-zones refer to a pseudonym change scheme limited to specific locations, such as intersections or high-traffic areas like shopping centers, restaurants, and gas stations [115, 185, 195]. Application domain refers to the application domain purpose of the mix-zone. We can have mix-zones intended to preserve privacy when users use LBS or send safety messages in V2X communication [168, 177, 196, 197]. The execution mode of the mix-zone can be online if the mix-zone proposal is executed in real-time or offline when the mix-zone is used in batch to protect data for publishing [171].

Pseudonym-changing strategies also can be cooperative and non-cooperative. The cooperative means that vehicles agree to change their pseudonyms synchronously. Vehicles can securely facilitate this cooperation through encryption or by coordinating periods of silence among vehicles. Within this group, vehicles can either change or swap their pseudonyms [184, 198]. When a vehicle changes its pseudonym, it replaces it with a new one that was previously issued to it by an authoritative entity.

On the other hand, non-cooperated strategies are vehicle-centered, which means that selfish vehicles can change their pseudonyms independently of their surrounding vehicles or infrastructure [166, 199]. The problem is that each vehicle can identify and track other vehicles even after changing their pseudonyms. Some schemes mitigate correlation tracking by combining obfuscation techniques, which add noise to location data and alter speed information, making it difficult to link pseudonyms [199]. Cooperation may be map-dependent, meaning that it takes place in specific locations, such as hotspots or mix-zones [178, 191, 195, 196]. These techniques use the concept of "hiding in the crowd" to protect vehicles from precise tracking by attackers. The idea is that a vehicle can achieve anonymity if it remains active within a crowd of indistinguishable vehicles, making it harder to single out any individual vehicle. This strategy assumes that blending in with a large group reduces the likelihood of being tracked or identified accurately.

In cooperation mode, we have two pseudonym change policies [104, 186]: pseudonym change and pseudonym exchange. In the pseudonyms change, the vehicles simply change their pseudonym collects in their pseudonym set. In the pseudonym exchange (or pseudonym swap), vehicles exchange pseudonyms with each other [185, 187, 191]. The process of ex-

changing pseudonyms uses a swapping protocol, where the RSU randomly selects two vehicles to swap their pseudonyms. After each exchange, the RSU notifies the CA to maintain accountability. To ensure security, all messages involved in the pseudonym exchange process are encrypted [104].

Some studies have identified Hybrid LPPMs as a future trend for more robust solutions [34, 99, 100, 171]. Hybrid LPPMs combine multiple schemes, such as cooperation, obfuscation and silence, obfuscation and encryption, and real and fake trajectories [100, 171]. Hybrid strategies for changing pseudonyms can enhance location privacy protection and address shortcomings in single-type schemes. Additionally, if properly implemented, they can ensure both quality of service and privacy. Figure 2.18 presents a categorizes and Table 2.1 summarizes this discussion. Next, we will detail relevant studies about mix-zones, pseudonyms-changing mechanisms, and mix-zone placement strategies.

▶ Mix-zones Related Studies

In the literature, the mix-zone for vehicular networks (or vehicular mix-zone) is distinct from the classical concepts of mix-zones because of the mobility type involved. Vehicular mobility is constrained by many spatial and temporal factors, such as physical roads, directions, speed limits, traffic conditions, and road conditions [200]. So, in the vehicular network context, Freudiger et al. [170] proposed a protocol to provide local privacy in vehicular networks. Specifically, they introduced a protocol called CMIX, based on mix-zone encryption for road intersections that ensures changing pseudonyms. Palanisamy and Liu [165, 190] considered the geographic aspects, constraints on movement patterns, and statistical behavior of the users for generating mix-zones. Specifically, it proposed a mix-zone model for vehicular networks that construct non-rectangular, adaptive mix-zones that start from the center of a road segment intersection on its outgoing road segments. The mix-zone length is determined based on the average speed of the road segment, the time window, and the minimum pairwise entropy threshold.

Some mix-zones derivations combine obfuscation approaches. For example, Chen et al. [171] presented a mix-zone technique that makes extra perturbations outside mix-zones and recursively submits the sub-traces recursively to the technique to make sure that every sub-trajectory is of an appropriate length. The proposed technique was evaluated with two re-identification algorithms, Feature STC [201] and DBHMM [202]. The results of de-anonymization accuracy were up to 20% and 90%, respectively. Then, an analysis of the privacy and utility of the protected dataset is made with two trajectory analyzers. The results showed that the protected dataset, in general, provides a utility level that is very close to that of the original dataset.

Li et al. [187] proposed a pseudonym swap mechanism, PAPU, designed to enhance privacy VANETs. The PAPU is based on differential privacy to ensure provable

unlinkability between users' pseudonyms, making it difficult for adversaries to track vehicles over time. They measured privacy protection through metrics like unlinkability and privacy leakage alongside performance metrics such as computational overhead and communication cost. The results demonstrated that PAPU effectively improves privacy while maintaining acceptable performance in terms of resource usage.

Boualouache and Moussaoui [191] proposed a pseudonym strategy based on the silence that swaps pseudonyms between two vehicles at mix-zones deployed at a traffic light when it is red. They evaluated the approach using the anonymity set entropy and attacker success rate as privacy metrics.

Zuberi and Ahmad [110] proposed dynamic mix-zones that explore spatial and temporal features to reduce time attacks called transient mix-zones. The transient mix-zones are positioned in traffic lights and are activated for all the green signals at all the traffic junctions. Also, they mathematically modeled the vehicular traffic flow in mix-zones as Poisson distribution. They emphasized the importance of the number of mix-zones that users must cross to achieve effectiveness.

Memon et al. [189] proposed a pseudonym-changing schema based on the multi-mix-zones generation and vehicle reputation that allows users to change their dynamic pseudonyms inside and outside a mix-zone. The pseudonym changes outside a mix-zone occur with road infrastructure, such as Reported Servers that intermediate the communication between vehicles and RSUs. They evaluated the proposal using the SUMO simulator and compared the results with several existing pseudonym-changing techniques.

Mengjia et al. [198] proposed mix-context-based pseudonym-changing privacy-preserving authentication (MPCPA) through a mutual authentication mechanism to prevent attack vehicles from sneaking into a VANET system. The approach preserves the integrity of transmitted messages with an anonymous authentication mechanism. In addition, MPCPA adopts a mix-context-based pseudonym-changing strategy to prevent vehicle tracking. Performance analysis demonstrates that MPCPA incurs low computational costs and offers a privacy-preserving scheme that is more secure than existing authentication schemes.

Zhou and Zhang [100] proposed the mix-zone scheme based on pseudonym change and dummy location to protect against attacks caused by time, location, and speed information entering and leaving the confusion zone. Particularly, the approach controls the time of the vehicle's silence state and adds noise to the location information, adjusts the speed value of the vehicle, and improves the defense ability against the inferred attack. Also, consider the quality of LBS in the mix-zone is guaranteed due to the gradual noise attenuation.

Kalaiarasy et al. [203] proposed minimizing the number of pseudonyms changing notifications in the network based on creating mix-zones that enhance the location privacy level to a remarkable level. Their approach is based on ring signatures to guarantee only

the reliability of vehicles involved in the process of pseudonym-changing notifications, such that the number of pseudonym changes is comparatively reduced to an acceptable level. The results showed that the approach had a higher success rate than pseudonym techniques quantified under different mix-zones, vehicular distances, and vehicular nodes in the network.

Li et al. [196] presented a dynamic silent-based mix-zone in which the pseudonym-changing scheme is based on the traffic condition at the traffic light region. They propose that the mix zone length is dynamically configured according to the traffic flow predicted in the green traffic light cycle and the pseudonym-changing scheme at the red light. To estimate the mix-zone length, they explored four prediction algorithms. Also, they compared the proposal with three typical silent, anonymous schemes; comprehensive experimental results show that TLAS can achieve better performance in both the anonymous effect and driving security.

Kalaiarasy and Sreenath [177] proposed a self-generated mix-zones by facilitating the pseudonym-changing anonymity of vehicles based on the generated incentive degree. They proposed a mechanism that uses a one-way hash function and an improved pseudonym scheme for estimating vehicular incentives to facilitate privacy protection. The results showed that their approach outperformed the others ' changing proposals regarding location privacy under different distances, the number of vehicles, and the number of mix-zones.

Deng et al. [178] proposed a pseudonym change scheme for location privacy preservation in VANETs, in which vehicles can adopt different pseudonym change strategies based on various network and traffic scenarios to resist tracking by global passive adversaries. Particularly, they proposed a pseudonym-changing region created by the vehicle without RSUs. Also, the registration protocol, authentication protocol, pseudonym issuance protocol, and pseudonym revocation protocol are introduced for the pseudonym management mechanism.

▶ Mix-zone Placement Related Studies

An emergent research area in mix-zones is to choose good places to position in urban environments. The fair distribution of mix-zones throughout the city must consider many factors that may affect coverage for anonymization, e.g., traffic flows and the seasonality and noise caused by buildings [176, 204]. For this, mix-zone placement is an NP-hard problem [175, 197, 205]. The seminal study to address the mix-zone placement problem was proposed by Freudiger et al. [205]. They treated the mix-zone placement problem as an optimization problem and used mobility profiles to compute the effectiveness of a mix-zone and identify the best places. Notably, they considered the optimization problem of the distance-to-confusion and the cost induced by mix zones on mobile nodes.

Liu et al. [197] designed two heuristic algorithms, based on the problem of independent sets, that are effective for positioning multiple mix-zones and thereby reducing the privacy risks of mobile users' trajectories. This approach considered the effect of traffic density in entropy terms for each road segment and intersection. The efficiency of the algorithms was evaluated with taxicabs of the city of San Francisco dataset [126].

The work proposed by Palanisamy and Liu [190] is an extension of MobiMix [165]. Additionally, they proposed a heuristic for mix-zones placement considering the road network topology, user mobility patterns, and road characteristics. In detail, mix-zones are placed at intersections with high traffic density and low skewness in the transition probability distribution. First, top-n mix-zones are selected by a cost associated with their size (radius) and the average estimated anonymity levels. Posteriorly, the top-n mix-zones are placed at intersections so that users can go through sufficient mix-zones along their path. For this, a road network was divided into grid cells using a quadtree index partition, and the average distance between any pair of mix-zones within each quadrant (grid cell) was maximized.

Sun et al. [206] proposed a statistics-based metric to evaluate a mix-zone's effectiveness and select candidates regarding privacy requirements. They proposed a mix-zone placement scheme in which vehicles from anywhere pass through a mix-zone at a certain driving time, and the extra overhead of adjusting routes is small. They modeled the positioning problem as an instance of the set-covering problem. They proposed solving it with greed set cover and selecting mix-zones from eligible intersections, stating that any travel exceeding a certain distance must pass through at least one mix-zone, with the additional delay caused by route adjustment limited to a predetermined amount.

Xu et al. [168] treated the problem of optimal multiple mix-zones as a transportation problem and proposed a greedy-based heuristic algorithm for this. Specifically, given a privacy threshold, the mix-zones candidates, and demanded points for each candidate, a cost is calculated, i.e., a likelihood distribution that is the total time of all users traveling from the demand points. Initially, all candidates are selected and then are greedily removed one by one. The proposed method was validated by protecting two datasets against inference attacks, which proved to be efficient in reducing the risk.

Memon and Arain [207] have presented an approach to solving mix-zone placement for urban environments in more than one direction (within 1D, 2D, and multiple directions). They proposed two heuristic algorithms to minimize the costs of the optimal location while the average mix-zone capacity can be maximized, increasing privacy in a high-traffic vehicular environment.

Ravi et al. [208] presented a heuristic to place mix-zones considering the trade-off between privacy and cost. The heuristic is called the Anonymity Enhancing Mix Protocol (AEMP), which alters the exit order of vehicles from the mix-zone to enhance privacy. Then, it uses AEMP in a placement mix-zone strategy composed of simulated annealing

and genetic algorithms. In the first steps of the algorithm are necessary starting points for placement, which is selected by a ranking of intersections defined by mixability, a metric that determines the usefulness of initially placing a mix-zone. AEMP was submitted against a tracking attack based on machine learning - Random Forest, which has proven to be resilient for this attack and outperformed the mix-zone baseline.

Svaigen et al.[175] argued a mix-zones positioning approach based on the premise that mix-zone needs to change according to the flow behavior of mobile entities. They proposed two constraint metrics with a positioning algorithm based on a k-means algorithm. The first metric considers only the graph topology; the second one aims to place them considering $n - 1$ road flows between two mix-zones in a window size.

Ravi et al. [176] proposed a dynamic mix-zone placement for VANETs considering traffic flow over time. The proposal has two components: an offline component and an online component. The online component calculates aggregate traffic flow associated with time slots defined by the engineer's empirical knowledge. These time slots can vary from an hour to an entire day. Next, the component calculates mix-zone placement for each traffic pattern over time and stores this information in a library. The algorithm updates the library periodically so that the optimization process is not on the critical real-time path of the placement algorithm. The online component takes the current prevailing traffic pattern as input and uses the Kalman-Takens Estimator to predict traffic for the next time slot. Next, the system then finds the closest match to the patterns stored in the library. Once the algorithm has found a match, it recovers the appropriate mix-zone placement from the library.

2.6 Evaluating Location Privacy

The literature on privacy metrics is wide [32, 48, 49, 71, 209, 210, 211]. However, there is no standard or which metric would be best suited to apply to LPPM types [32]. Some metrics are dedicated to quantifying privacy applying for vehicular networks [211], or human mobility [212]. To address this issue, Wagner and Eckhoff [49] did notable work discussing over eighty privacy metrics using examples from six different privacy domains. An overview of privacy metrics can be found in [32, 33]. Detailed view about the metrics, e.g., k-anonymity, uncertainty, and differential privacy, can be found in [48, 153, 209, 213, 210]. Location privacy metrics can be divided into two viewpoints: attack and protection metrics.

2.6.1 Location Privacy Attacks Metrics

The location privacy attack metrics evaluate the effectiveness of attacks aimed at compromising location privacy. They measure how successful an adversary is in identifying users or their locations despite privacy-preserving measures [32, 33, 49]. In general, the metrics that measure the effectiveness of the adversary are based on correctness. Attack correctness is a kind of attack metric that depends strongly on the adversary model and is widely used in many privacy domains, e.g., communication, database, and location privacy. Concerning location privacy, the attack metrics measure the information leakage in the LPPM obtained by the adversary attack to gain knowledge about users. Attack correctness measures the adversary's success, where the adversary successfully identifies the correct individual or the true positive rate. However, it also may consider the false positives and false negatives rates, i.e., cases where the adversary identifies an individual incorrectly from the dataset [49].

In location privacy, an LPPM is expected to protect a dataset from mitigating the outcome of an attack. Thus, if LPPM is efficient, the value of this metric should be small. This metric directly assesses the attack on the protected dataset and uses the actual dataset as the basis for verifying the effectiveness of the attack.

Shokri et al. [71] have presented extensive work about attack correctness metrics based on probability distribution. Specifically, given $\mathcal{X}$ a victims' traces set, the information obtained is with posterior distribution $Pr(x|o)$, which $x \in \mathcal{X}$, and observed traces o of a set of observed traces by an adversary, $o \in \mathcal{O}$. However, the attacker does not have infinite resources. For this reason, the attack is only an estimate $\hat{Pr}(x|o)$. They proposed three attack metrics for evaluating the effectiveness of an attack: certainty, accuracy, and correctness.

Accuracy assumes that the outcome of your attack results in an approximate value, assuming that the attacker's knowledge is limited. For example, in a Tracking Attack scenario, the attacker has only a sample of the victim's geo-location points. Accuracy is calculated with each element of distribution $\hat{Pr}(x|o)$ trying to converge to $Pr(x|o)$.

Certainty is a metric based on the uncertainty of the outcome of the attack. The attack generates uncertainty greater than zero if it presents a result that concerns more than one possible user and null uncertainty when the attack identifies the victim, independently if this is the correct victim or not. It's quantifying with an entropy of the distribution $\hat{Pr}(x|o)$. The higher the entropy is, the lower the certainty of the adversary (Equation 2.12).

$$\hat{H}(x) = \sum_{x} \hat{Pr}(x|o) \log \frac{1}{\hat{Pr}(x|o)} \tag{2.12}$$

Correctness measures the expected distance between the true outcome of the attack $x_c \in \mathcal{X}$ and the estimate based on the $\hat{Pr}(x|o)$. The distance $\|.\|$ between victims' traces set $\mathcal{X}$ can be computed as sum, which is the *adversary's expected estimation error* (Equation 2.13).

$$\sum_{x} \hat{Pr}(x|o)\|x - x_c\| \tag{2.13}$$

where $\hat{Pr}(x|o)$ is posterior distribution of $Pr(x|o)$ that is the probability of trace x given observed trace o. The distance is equal to 0 in the case of $x = x_c$ and equal to 1 otherwise.

Trajectory Matching Accuracy (TMA) is a type of correctness to measure the effectiveness of re-identification attacks [54, 56], which is defined as:

$$\text{TMA} = \frac{N_{\text{reid}}}{|T_p|}, \tag{2.14}$$

where $N_{\text{reid}} \in [0, |T_p|]$ and $|T_p|$ represent the total of re-identified trajectories and the total of anonymized trajectories, respectively. This formula can also be generalized to measure the correctness of LPPMs that protect different aspects of location privacy, such as measuring the correctness in LPPMs that protect POIs [26, 214, 215]. For example, how much POIs could be retrieved in a Points of Interest Attack (POIA). The formula would be the ratio between the amount of POIs re-identified and total POIs of anonymized, called by *POI Matching Accuracy (PMA)*, see Equation 2.15. Given two sets of POIs, the function $Matched$ matches non-obfuscated with obfuscated POIs of user i, $poi(T_i)$ and $poi(T'_i)$ respectively. Two of the POIs are matched if they are sufficiently close to the other at maximal distance threshold d_{max}. PMA metric also can be extended for measure Region of Interest (ROI), called Coverage of Sensitive Region [49].

$$PMA = \frac{|Matched(poi(T_i), poi(T'_i))|}{|poi(T'_i)|} \tag{2.15}$$

2.6.2 Location Privacy Protection Metrics

The location privacy protection metrics assess the effectiveness of mechanisms designed to protect an individual's location privacy. They measure how well a privacy-preserving method prevents adversaries from correctly inferring sensitive information from location data. Key protection metrics include Anonymity Set Size, Entropy; k-anonymity; l-diversity; Location Obfuscation; and Mix-zones Effectiveness;

Anonymity Set Size (ASS). The anonymity set (AS) refers to the group of vehicles that cannot be distinguished from the target vehicle, including the target itself [216]. The size of this set, denoted as $|AS|$, corresponds to the total number of vehicles within it. The larger the anonymity set, the higher the level of location privacy protection, which in this context should be greater than 1.

Another class of formal metrics is **uncertainty metrics**, responsible for measuring the uncertainty level of the adversary against revealing protected data by an LPPM [49]. This metric assumes that an adversary uncertain of his estimate cannot breach privacy as effectively as one who is certain; this notion tends to increase when applying LPPMs to mobility data. Usually, many uncertainty metrics build on entropy. Concerning location privacy, entropy has been used in tracking attacks, i.e., the adversary tracks victims by linking many places that have been visited for a time period. In this case, the entropy is calculated for each location in time, and the probabilities are updated after each period using Bayesian belief tables [217]. Another application of entropy is to measure the position of a user. However, if more than one location is close to each other, locations may be high entropy, i.e., high uncertainty about revealing these places [218].

Further to ***k-anonymity*** being a privacy model in which LPPMs are implemented, it is also considered a metric for measuring the level of privacy applied to the dataset. *K-anonymity* states that during a given time window and inside a given area, there should be at least k entities. Many LPPMs use this concept to generate privacy, where k entities simultaneously change their pseudonyms or cloak geolocated data. The privacy levels of k-anonymity are defined by tuning the parameters: the area size, time window, and k value. In the literature, there are extensions of k-anonymity to improve privacy levels, considering the contextual knowledge an adversary could have. **Mix-zones Effectiveness**. The *Mix-zones Effectiveness* evaluates how well mix-zones (regions where users can change pseudonyms to break linkability) protect users' movements from being tracked [197].

L-diversity is an extension of k-anonymity by ensuring that within an anonymity set, there are at least L diverse values for sensitive attributes, which prevents linking users to specific sensitive information.

The **Location Obfuscation** measures the degree to which the proper location is hidden through techniques such as adding noise or reducing the precision of the reported location. The ***ϵ-differential privacy*** is a kind of Local Obfuscation metric that measures the privacy of obfuscation-based LPPMs in different forms. In some cases, LPPMs guarantee privacy to protect the absence and presence of entities. In others, the guarantee of protecting locations and sensitive places of users, like POI. In this case, the goal is not to protect the entities themselves but the places visited for it.

2.6.3 Utility Metrics

In addition to evaluating privacy, another issue of LPPM design is valuing the utility of data protection. For this, it is essential to define efficient utility metrics for measuring the quality of protected mobility data. One of the approaches to measuring the utility is Information loss, which refers to data quality degradation during processes like anonymization and obfuscation, particularly in location privacy scenarios. It quantifies how much useful information is lost to protect user privacy, using metrics including Shannon Entropy, which assesses the unpredictability of data; Kullback-Leibler Divergence and Jensen-Shannon Divergence, which compare the differences between original and anonymized data distributions; and Mutual Information, defined as a function *WS*, which evaluates the distortion in the shared information between original and transformed data [15, 26, 33, 214].

The data distortion approach compares the mobility data properties before and after applying an LPPM [15]. The optimal value of this metric is expected that the data distortion will not be severe enough so that it does not become useless. In the context of anonymization approaches for mobility data, such as anonymized vehicular trajectories datasets, we can measure its utility by checking the distortion level *ZD*, with distortion function *WS* between the original dataset $\mathcal{D}$ and the anonymized dataset $\mathcal{D}'$, in which the trajectories are sliced to anonymize the user's identity, for example, $ZD = WS(\mathcal{D}, \mathcal{D}')$.

We can apply the same principle to location data protected by obfuscation techniques, in which the distortion metrics include evaluating the spatial and temporal imprecision and comparing the covered area. We can cite metrics distortion-based as Area Coverage, Spatial Distortion [26].

Area Coverage (AC) computes how much the distortion on data affected the regions visited by a user, e.g., POIs [92]. That is, removing location data in regions less important for user mobility, e.g., removing POIs in these regions and preserving data in regions considered important. On the other hand, adding fake POIs and changing them to new regions can degrade the analysis mobility data. For example, an analyst uses distorted data in the transport planning scenario. He may conclude that in certain areas, public transport public is needed, but this is not true. Another example is the public health department's case, which surveys noise in urban areas by running a crowd-sensing campaign to measure noise levels in the city. In this case, the locations on the dataset do not need high precision, but it is important to cover the correct regions of the city. The Area Coverage (AC) is calculated as follows. Consider a map equally divided into square regions. The function $C(T)$ returns a number of regions whose user trajectory T goes through Equation 2.16, where cell c_i from region set $\mathcal{C}$, i.e., cell $c_i \subseteq \mathcal{C}$, $1 \leq i \leq |\mathcal{C}|$, and $p \odot c_i$ is geo-location point p insides c_i.

$$C(T) = \{c_i \in \mathcal{C} | \exists p \in T : p \odot c_i\} \tag{2.16}$$

The AC of obfuscated trace T' from T is calculated with the F-score, which needs to calculate the recall and precision. Recall that Equation 2.17 measures the proportion of cells of the non-obfuscated trace still found in the obfuscated trace. The precision, Equation 2.18, evaluates the proportion of cells that users cross in the obfuscated trace, which is present in the n the non-obfuscated trace. From the recall and precision, we can calculate the F-score, denoted by Equation 2.19.

$$AC_{Recall}(T, T') = \frac{|C(T) \cap C(T')|}{|C(T)|} \tag{2.17}$$

$$AC_{Precision}(T, T') = \frac{|C(T) \cap C(T')|}{|C(T')|} \tag{2.18}$$

$$AC_{F-score}(T, T') = \frac{2AC_{precision}(T, T')AC_{Recall}(T, T')}{AC_{precision}(T, T') + AC_{Recall}(T, T')} \tag{2.19}$$

An obfuscation-based LPPM can distort a mobility dataset to the point that it does not extract POIs located in specific regions of the urban map before the anonymization was possible. *POIs' amount by Cell (PAC)* does this work. PAC metric evaluates if a certain POI amount is found in a specific region after applying obfuscation. We count the number of POIs for each c_i of the urban map, both actual dataset D and obfuscated datasets D', with the function $POI_{Count}(.)$. Then we calculate the recall and precision Equations 2.20 and 2.21 respectively. The distortion of the amount of POIs is the average of all the F-scores of the cells, denoted by Equation 2.22.

$$PAC_{Recall}(D, D', c_i) = \frac{|POI_{Count}(D, c_i) \cap POI_{Count}(D', c_i)|}{|POI_{Count}(D, c_i)|} \tag{2.20}$$

$$PAC_{Precision}(D, D', c_i) = \frac{|POI_{Count}(D, c_i) \cap POI_{Count}(D', c_i)|}{|POI_{Count}(D', c_i)|} \tag{2.21}$$

$$PAC_{F-score}(D, D', \mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \frac{2PAC_{Precision}(D, D', c_i)PAC_{Recall}(D, D', c_i)}{PAC_{Precision}(D, D', c_i) + PAC_{Recall}(D, D', c_i)} \tag{2.22}$$

The PAC and AC are utility metrics defined in the range of $[0, 1]$. A higher value represents better utility, meaning better spatial accuracy of the LBS results.

Spatial Distortion (SPAD) is an average distance metric that measures the spatial error of geo-located points between trajectory $T = \{p_1, p_2, p_3, ..., p_n\}$ and its obfuscation version $T' = \{p'_1, p'_2, p'_3, ..., p'_n\}$ [214, 26]. For each point, p' in T' is searched for minimal

projection on T, function *distance* measures the minimal distance between T and T' points (Equation 2.23).

$$SPAD(T, T') = \frac{\sum_{p' \in T'} \min_{0<p<|T|} distance(p, p')}{|T'|} \tag{2.23}$$

Spatial Distortion of POIs (SPAP) is a derivation of Spatial Distortion, representing an average distance metric that measures the spatial error of POIs as denoted by Equation 2.24.

$$SPAP(T, T') = \frac{\sum_{l' \in T'} \min_{0<l<|T|} distance(l, l')}{|T'|} \tag{2.24}$$

where l and l' represent the location of the users' original trajectory T and protected trajectory T', respectively. This metric can be used to measure data's utility level in urban planning for public safety. For instance, it can be used to analyze mobility data and identify location points for security camera installations in locations with high population density.

2.7 Discussion and Future Trends about Anonymization-based LPPMs and LPAs

This section discusses new directions for the Anonymization-based Location Privacy. Particularly, we identify issues about the GlsplLPA, anonymization-based LPPMs, and utility focused on SMOD. First, we identify which aspects of mix-zones and pseudonym changes have been explored, detailed in Section 2.5.3, and new research opportunities. Next, we present new trends in the general context about LPAs, LPPMs, and the utility of anonymized mobility data.

Table 2.1 presents the proposed pseudonym-changing mechanisms and mix-zones for vehicular networks discussed in Section 2.5.3. We can observe a prevalence of approaches based on the cooperative modality for pseudonym-changing. In contrast, few works have addressed the non-cooperative modality and pseudonym-changing between vehicles.

Most approaches present solutions to protect mobility data for VANETs, where the mix-zone aims to protect safety messages exchanged in Vehicle-to-Anything (V2X). However, the mix-zones protecting data sent to LBSs and Vehicular Social Networks (VSNs) is a little-explored issue.

Another issue is that most mix-zone proposals depend on the infrastructure of vehicular networks such as RSUs and specific servers for pseudonym-changing. They also

depend on specific locations on the urban map to be positioned, such as traffic lights, intersections, and certain points on the road. This fact suggests that there is a prevalence of pseudonym-changing and mix-zones approaches in which they are positioned in these locations statically, immutable over time. Thus, there is a need to investigate dynamic mechanisms independent of VANET infrastructure and road locations to generate anonymization, as is the case of ad hoc mix-zones formed by groups of vehicles.

Regarding validation, most of the proposals found in Table 2.1 are evaluated through GPA that use location privacy protection metrics, such as Entropy and Anonymity Set Size (ASS). There is also a significant number of proposals that use synthetic datasets and simulated environments, using mobility simulators, networks, and adversarial networks, such as SUMO [219], Veins [220], and MobiSim [221] that are mobility simulators the most frequently used. SUMO exports mobility models to well-known network simulation tools such as OMNET++ [222], NS2, and NS3 [223]. In the investigated proposals, some attacks, such as Nearest-Neighbor Probabilistic Data Association (NNPDA), are implemented in Veins.

On the other hand, when analyzing Table 2.1, we can identify several little-explored issues that require investigation regarding the validation of the protection algorithms. For example, it is necessary to validate the privacy approaches from more robust adversary models, considering mobility characteristics and even metrics of attacks on location privacy.

From the point of view of the intrinsic characteristics of the mix-zones, such as privacy, geometry, and location, most of the proposals present static solutions for these parameters. Thus, there needs to be more dynamic approaches in which the mix-zones' privacy, geometry, and positioning are configured over time according to mobility aspects.

Another research opportunity for anonymization-based LPPMs is to explore the trade-off between privacy, utility, and quality of protected data in online and offline execution modalities. Another issue is exploring the design of hybrid LPPMs, which consider both anonymization and obfuscation to protect mobility data. We can mitigate user identity and location attacks by considering these two protection mechanisms.

The positioning of mix-zones is also an issue that should be investigated further. In particular, the positioning of mix-zones over time, considering aspects such as mobility, fluctuations in traffic volume, and points of interest over time, to obtain greater coverage of the data to be protected. Also, the selection of mix-zones already positioned, based on their operation and the quality of anonymized data over time, is an open question.

Regarding smart mobility open data, in which privacy, quality, and utility requirements must be met, these works need to consider these requirements simultaneously, as well as the trade-off between privacy, utility, and quality of the protected data. In particular, the quality of the internal functioning of mix-zones and their anonymized data.

Finally, the proposal in the literature has explored utility as a distortion relation

based on an information loss function. However, there appears to be progress on utility in relation to the use of trajectories anonymized by mix-zones in certain application and service domains in the context of smart cities, which is an important issue when dealing with data to be published.

Table 2.1: Mix-Zones and Pseudonym Changing Proposals

Proposal		**LPPM Characteristics**					**LPPM Parameters**	**Validation**			
Year	Ref	Approach	Type/ Pseud.[1]	Crowd Type/ Infra Dep.	Coop.	Network/ Data Prot.	Priv/Geo/Pos	Trade-offs/ Utility	Adversary/ Attack[2]	Metric	Environ./ Traffic
2003	[164]	Mix-Zone	A/C	Traffic flow/ Infra. and Map	Yes	Human mobility/ LBS	Stat./Stat./Stat.	No	GPA	ASS	Simulated/ Realistic
2007	[170]	Mix-Zone, Cryptography	A/C	Traffic flow/ Infra. and Map	Yes	VANET/ safety	Stat./Stat./Stat.	No	GPA	Entropy, TMA	Realistic/ Realistic
	[194]	Multiple Mix-Zones	A/C	Traffic flow/ Infra. and Map	No	VANET/ safety	Stat./Stat./Stat.	No	GPA / Mix-zone aware	Attacker success	Simulated/ Realistic
2009	[205]	Placement (Stat.)	A/C	Traffic flow/ Infra. less	No	VANET/ safety, LBS	Stat./Stat./Dyn.	No	TA / Timing attack	Entropy, TMA, Mixing effectiveness	Simulated/ Synthetic
2010	[192]	Ad Hoc Mix-Zones, Placement (Dyn.)	A/C	Traffic flow/ Infra. less	Yes	VANET/ safety	Dyn./Stat./Dyn.	No	GPA	ASS	Synthetic
2011	[165]	Mix-Zone	A/C	Traffic flow/ Infra. dep	Yes	VANET/ LBS	Stat./Dyn./Stat.	No	GPA, TiA, TrA	Entropy	Simulated/ Synthetic
	[195]	Placement (Stat.)	A/C	Global and social spots/ Infra. and Map	Yes	VANET/ safety	Dyn./Stat./Stat.	No	GPA	ASS	Synthetic
	[193]	Mix-Zone, Cryptography	A/C	Traffic flow / Infra. and Map	Yes	VANET/ safety	Stat./Stat./Stat.	No	Internal, Passive	Successful track	Simulated/ Synthetic
	[182]	Ad Hoc Mix-Zones, Placement (Dyn.)	A/C	Traffic flow, Group identifiers/ Infra.: RSU	Yes	VANET/ safety	Dyn./Stat./Dyn.	No	GPA	ASS TMA	Simulated
2012	[197]	Placement (Stat.)	A/C	Traffic flow and POIs / Infra. and Map: Intersections	No	LBS/ LBS	Stat./Stat./Dyn.	No	GPA TA: POIs	Entropy	Realistic/ Realistic
	[224]	Mix-Zone	A/C	Traffic flow/ Infra. dep	Yes	VANET/ safety	Dyn./Stat./Stat.	No	–	Pseu. changes Protection rate	Simulated

Table 2.1 continued from previous page

Proposal		LPPM Characteristics					LPPM Parameters	Validation			
Year	Ref	Approach	Type/ Pseud.[1]	Crowd Type/ Infra Dep.	Coop.	Network/ Data Prot.	Priv/Geo/Pos	Trade-offs/ Utility	Adversary/ Attack[2]	Metric	Environ./ Traffic
2013	[225]	Placement (Stat.)	A/C	Traffic flow/ Infra. and Map: Intersections	No	VANET/ safety	Stat./Stat./Dyn.	Privacy and cost	–	Number of vertices, Solution quality, Execution time	Realistic/ Realistic
	[226]	Ad Hoc Mix-Zones, Placement (Dyn.)	A/C	Traffic flow, Predicted location, Privacy level/ Infra.: Control Server	Yes	VANET/ safety	Stat./Stat./Dyn.	No	GPA	Successful track Entropy	Synthetic
	[184]	Ad Hoc Mix-Zones, Placement (Dyn.)	A/C	Neighboring cars/ Infra. less	Yes	VANET/ safety	Stat./Stat./Dyn.	No	GPA	ASS	Realistic/ Synthetic
2014	[190]	Mix-Zones, Placement (Dyn.)	A/C	Traffic flow/ Infra.	Yes	VANET/ LBS	Stat./Dyn./Dyn.	No	TiA, TrA/ Inference	ASS, Entropy, Success rate	Simulated/ Synthetic
	[191]	Silence-Based Mix-Zones	A/S	Traffic flow/ Infra. and Map	Yes	VANET/ safety	Dyn./Dyn./Stat.	No	GPA	ASS Entropy, Attacker success	Simulated
2015	[206]	Placement (Stat.)	A/C	Traffic flow/ Infra.	No	VANET/ Open data	Stat./Stat./Dyn.	No	IA	Entropy, Mix deployed	Simulated
	[188]	Mix-Zones, Placement (Dyn.)	A/C	Traffic flow / Infra.: Control Server	Yes	VANET/ safety	Dyn./Stat./Dyn.	No	GPA	Location Privacy Strength	Synthetic
	[227]	Ad Hoc Mix-Zones, Placement (Dyn.)	A/C	Traffic flow/ Infra. less	Yes	VANET/ safety	Stat./Stat./Dyn.	No	GPA	ASS, Success rate	Synthetic
	[183]	Ad Hoc Mix-Zones, Placement (Dyn.), Silent Period	A/C	Traffic flow indep./ Infra. and Map	Yes	VANET, VSN/ safety	Dyn./Dyn./Dyn.	No	GPA, LPAd	Entropy	Simulated
	[168]	Placement (Stat.)	A/C	Traffic flow/ Infra.: intersections	No	VANET/ LBS	Stat./Stat./Dyn.	No	GPA, TA	Success rate	Realistic/ Realistic

Table 2.1 continued from previous page

Proposal		LPPM Characteristics					LPPM Parameters	Validation			
Year	Ref	Approach	Type/ Pseud.[1]	Crowd Type/ Infra Dep.	Coop.	Network/ Data Prot.	Priv/Geo/Pos	Trade-offs/ Utility	Adversary/ Attack[2]	Metric	Environ./ Traffic
2016	[116]	Silent Period	A/C	Traffic flow/ Infra. less	Yes	VANET/ safety	Stat./Stat./Stat.	Traceability and QoS	GPA, LAA/ MTT	Silence Time, Traceability	Simulated/ Realistic
	[110]	Multiple Mix-Zones	A/C	Traffic flow/ Infra.	No	VANET/ safety	Dyn./Stat./Stat.	No	TiA	TMA	Simulated/ Realistic
	[228]	Placement (Stat.), Silent Period, Mix-Zones	A/C	Traffic flow/ Infra. and Map	Yes	VANET/ safety	Stat./Dyn./Dyn.	No	GPA, LPAd SyLA, SeLA	ASS	Simulated
	[166]	Mix-Zone, Silent Period	A/C	Traffic flow indep./ Infra.	No	IoV/ safety, LBS	Stat./Stat./Stat.	No	GPA/ Obs-Map, Link-Map	Entropy	Simulated
2017	[207]	Placement (Stat.)	A/C	Traffic flow/ Infra.: RSU	No	Road Networks/ Open data	Stat./Stat./Dyn.	No	GPA/ TA	Entropy, Mix deployed cost, TMA	Realistic/ Synthetic
	[113]	Silent Period	A/C	Traffic flow/ Infra. less	Yes	VANET/ safety	Stat./Stat./Stat.	Traceability and QoS / Dataset Distortion	GPA/ MTT	Traceability, Trace Distortion, Entropy, ASS	Realistic/ Realistic
	[114]	Placement (Stat.)	A/C	Traffic flow and POIs / Infra. less	No	VANET/ LBS	Stat./Stat./Dyn.	Privacy and usability	–	Endpoint Deviation, Comp. cost, Entropy	Simulated/ Realistic
	[189]	Multiple Mix-Zones	A/C	Traffic flow/ Infra.	Yes	VANET/ safety, LBS	Dyn./Stat./Stat.	No	GPA	Successful rate	Simulated
	[105]	Ad Hoc Mix-Zones, Cryptography, Placement (Dyn.), Silent Period	A/C	Traffic flow/ Infra. less	Yes	VANET/ safety	Stat./Stat./Dyn.	No	GPA, LPAd, SyLA, SeLA	ASS Entropy, Attacker success	Simulated/ Synthetic
	[171]	Mix-Zone, Hiding	A-O/C	Traffic flow/ Infra.: Intersections	Yes	VANET/ Open data	Stat./Stat./Stat.	Privacy and utility	Feature STC, IA: DBHMM	Attacker success	Realistic/ Realistic

Table 2.1 continued from previous page

Proposal		LPPM Characteristics					LPPM Parameters	Validation			
Year	Ref	Approach	Type/ Pseud.[1]	Crowd Type/ Infra Dep.	Coop.	Network/ Data Prot.	Priv/Geo/Pos	Trade-offs/ Utility	Adversary/ Attack[2]	Metric	Environ./ Traffic
2018	[115]	Placement (Dyn.)	A/C	Traffic flow, POIs/ Infra. less	No	VANET/ LBS	Stat./Stat./Dyn.	Privacy and usability	–	Endpoint Deviation, Comp. cost, Entropy	Simulated/ Synthetic
	[185]	Mix-Zone	A/S	Traffic flow/ Infra. and Map: Traffic light, Road regions	Yes	VANET/ safety	Dyn./Stat./Stat.	No	GPA/ NNPDA	TMA, Max entropy, ASS	Synthetic
	[186]	Multiple Mix-Zones	A/S	Traffic flow/ Infra.	Yes	VANET/ safety	Dyn./Stat./Stat.	No	GPA / Linkability, Cheating attack	ASS	Simulated
	[199]	Mix-Zone	A/C	Traffic flow indep./ Infra. and Map less	Both	VANET/ safety	Dyn./Stat./Dyn.	No	GPA	ASS, Performance, Location privacy strength	Synthetic
2019	[208]	Placement (Stat.)	A/C	Traffic flow/ Infra.	No	VANET/ safety, LBS	Stat./Stat./Dyn.	Privacy and cost	TA	Mixability, TMA, Perc. protection	Simulated/ Synthetic
	[175]	Placement (Dyn.)	A/C	Traffic flow/ Infra. less	Yes	Road Networks/ send to LBS	Stat./Stat./Dyn.	No	–	Coverage rate, Pseud. rate	Realistic/ Realistic
	[100]	Mix-Zone, Dummy Location	A-O/C	Traffic flow/ Infra.	Yes	VANET/ safety	Stat./Stat./Stat.	No	LPAd/ Loc., Time, and Speed inference	TMA	Simulated/ Realistic
	[198]	Mix-Zone	A/C	Traffic flow/ Infra. less	Yes	VANET/ safety	Stat./Stat./Stat.	No	Pseu. linking	TMA,	–
2020	[187]	Mix-Zone	A/S	Traffic flow/ Infra.: RSU	Yes	VANET/ safety	Stat./Stat./Stat.	Dataset Utility loss	GPA	ASS	Realistic/ Realistic
	[203]	Mix-Zone	A/C	Traffic flow/ Infra.	Yes	VANET/ safety	Stat./Stat./Stat.	No	–	Success rate, Location privacy	Simulated

Table 2.1 continued from previous page

Proposal		LPPM Characteristics					LPPM Parameters	Validation			
Year	Ref	Approach	Type/ Pseud.[1]	Crowd Type/ Infra Dep.	Coop.	Network/ Data Prot.	Priv/Geo/Pos	Trade-offs/ Utility	Adversary/ Attack[2]	Metric	Environ./ Traffic
2021	[196]	Mix-Zone, Silent Period	A/C	Traffic flow/ Infra.	Yes	VANET/ safety	Dyn./Dyn./Stat.	No	GPA/ MTT, NNPDA	Traceability	Simulated/ Realistic
2022	[178]	Mix-Zone	A/C	Traffic flow/ Infra. and Map less	Yes	VANET/ safety, LBS	Dyn./Dyn./Dyn.	No	GPA	ASS	Simulated/ Realistic
	[177]	Mix-Zone, Cryptography, Placement (Dyn.)	A-O/C	Incentive / Infra. and Map less	Yes	VANET/ safety	Dyn./Dyn./Dyn.	Privacy loss and utility	GPA	Success rate vs Num. mix-zones, Location Privacy vs Distance	Simulated
2023	[176]	Placement (Dyn.)	A/C	Traffic flow/ Infra. and Map: Intersections, RSU	Yes	VANET/ safety, LBS	Stat./Stat./Dyn.	No	GPA	Clusters number ASS Avg Mix-zones/day Avg vehicles/day	Simulated/ Synthetic

[1] LPPM Type: Anonymization (A), Obfuscation(O); Pseudonym Changing Type: Pseudonym Change (C), Pseudonym Swap (S).

[2] Inference Attack (IA), Global Passive Adversary (GPA), Tracking Attack (TA), Timing Attack (TiA), Transition Attack (TrA), Local Passive Adversary (LPAd), Local Active Adversary (LAA), Multi-target tracking (MTT), Syntactic Linking Attack (SyLA), Semantic Linking Attack (SeLA), Nearest-Neighbor Probabilistic Data Association (NNPDA).

2.8 Concluding Remarks

This chapter discussed the key location privacy mechanisms, as well as both attack and defense approaches. Initially, we present the Location Privacy Attacks, describing a taxonomy, formalization, and the state-of-art. Next, we presented fundamentals about the LPPMs based on obfuscation and anonymization. We presented the state-of-the-art anonymization-based LPPMs and identified several issues open to research regarding the mix-zones schemes, including their application on smart mobility open data. This study enabled the construction of Smart Privacy Framework (SPF): an anonymization-based framework for Smart Mobility Open Data, in which we can answer the research questions of this thesis. In Chapter 3 details the SPF.

Chapter 3

Smart Privacy: An Anonymization-based Framework for Smart Mobility Open Data

This chapter presents the Smart Privacy Framework (SPF), an Anonymization-based Framework for Smart Mobility Open Data. Initially, we present an overview of the framework in Section 3.1. Then, we detail each step that composes the SPF, representing the next sections defined as follows. Section 3.2 details the Mobility Impacts on Location Privacy step. Section 3.3 presents the second and third step, Privacy Design and Protection. The fourth step is the Privacy, Quality, and Utility Indicator Extraction is detailed in Section 3.4. The fifth step, Data Publishing Analysis, is defined in Section 3.5. The concluding remarks are in Section 3.6.

3.1 Anonymization-based Framework for Smart Mobility Open Data

The studies carried out on LPA and LPPM in Chapter 2 allowed us to identify important research questions within the context of SMOD in privacy, quality, and utility terms. These research questions were defined in Section 1.1 that justify the development of this framework. Following the steps within SPF we can answer these questions. Next, we will detail SPF and present each research question within the context of this framework.

The Smart Privacy Framework (SPF) proposed in this work comprises five steps: Mobility Impacts on Location Privacy (1), Privacy Design (2), Protection (3), Privacy, Quality, and Utility Indicator Extraction (4), and Data Publishing Analysis (5), as depicted in Figure 3.1.

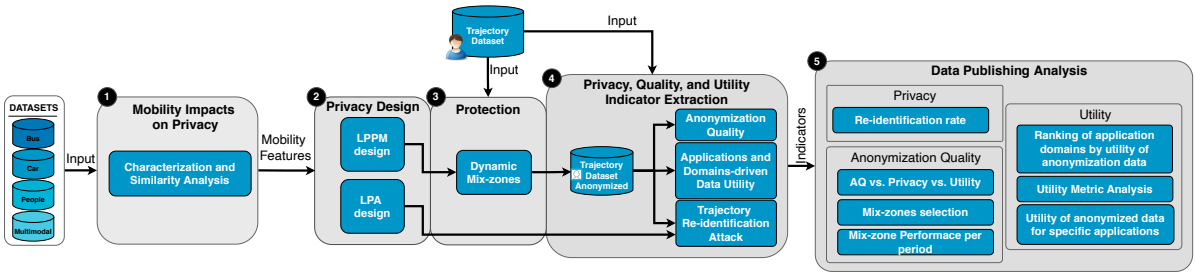


Figure 3.1: Smart Privacy Framework (SPF): An Anonymization-based Framework for Smart Mobility Open Data.

3.2 Mobility Impacts on Location Privacy

RQ1: What are the impacts of mobility on location privacy?

The first step, Mobility Impacts on Location Privacy, represents a framework to characterize and find similarities in the statistical distributions extracted from two Stay Point (SP) metrics. SP is an aspect of mobility data analysis, defined as a region where an entity stays for a minimum time interval. SPs operate as substrates to build privacy mechanisms; both LPAs and LPPMs. Regarding LPAs, a user's SPs can identify mobility patterns and extract user profiles and behaviors. Also, two or more SPs can track these users on a map. Clustering SPs enables the discovery of hot spots in people's concentration over time, making it possible to identify them. On the other hand, SPs can contribute substantially to the design of LPPMs. For anonymization-base LPPMs, SPs mineration can be used for mix-zones deployment, identify hot spots, and anonymize trajectories of greatest vehicle circulation. For obfuscation-based LPPMs, it is possible to create cloaked areas, dummy trajectories, and users where a hot spot exists.

Studying SP behavior with different types of mobility datasets can reveal valuable insights about privacy, such as: Are the SPs distribution metrics from mobility datasets of different transportation models similar? If the answer is not affirmative, it means that different mobility patterns and privacy will tend to have different behaviors for different transportation modes. In this case, to produce attacks, it is necessary to understand mobility and extract specific characteristics of each case to produce the adversary model and, consequently, the privacy attack. Regarding protection, the LPPM to be designed should be dynamic over space and time to meet dynamic environments containing different types of vehicles and traffic fluctuations over time.

The required framework input is real different datasets from different transportation modes, uni-and multimodal, to represent the scenarios found in Smart Mobility. The framework output are three analyses.

- WMA1: Similarities of distributions between the two transport generic types: Vehicular and human.
- WMA2: Similarities between distributions of the same category of vehicles. For example, if a taxi distribution is more similar to the distribution of another taxi than are distribution of buses or people.
- WMA3: Similarities between distributions of the same generic type but extracted from monomodal and multimodal transport datasets. For instance, if the taxicab dataset matches a vehicle category from a multimodal dataset, such as car, cabs, or bus.

From this analysis, we can get an answer about the impact of mobility on privacy. Also, we have mobility features to help with the LPA and LPPM design. Chapter 4 provides details about this structure. We used seven datasets with different transport modes. We analyzed datasets of human and vehicular mobility and multimodal transport. Additionally, we selected at least two datasets for each type of vehicle to check for possible similarities.

3.3 Privacy Design and Protection

RQ2: How to build efficient LPAs based on mobility characteristics and resilient LPPMs based on anonymity, which considers privacy, utility, coverage, and anonymization quality for SMOD?

The second step, Privacy Design, involves developing LPA and LPPM. The output obtained in the first step makes it possible to design efficient LPAs and LPPMs considering mobility characteristics. In the case of LPAs design, it is possible to exploit certain mobility characteristics to design adversarial models that produce more specific attacks but with high re-identification rates. For example, in the context of trajectory re-identification attacks, by analyzing the characteristics of taxi datasets, we can hypothesize that drivers' preferences in trips, such as choosing a shorter path to complete their routes, can produce a trajectory tracking attack for anonymized mobility data. Only two geolocated points are needed to reconstruct their trajectory through an algorithm based on the shortest path problem. In Chapter 5, we detail a case study of LPA about the design of this attack.

The output of first step also enables the development of LPPMs based on mobility characteristics. In a dynamic mobility model of smart cities, such as smart mobility, where there are different types of vehicles and constant traffic fluctuations, dynamic anonymization-based LPPMs are required. In particular, mix-zones depend on factors such as location, geometry, mobility patterns, vehicle density, and arrival rates, which can influence their performance. Based on this, inappropriate configurations of their parameters can reduce the anonymization rate or protect the data with a low level of privacy, facilitating tracking and inference attacks. These issues require the production of a dynamic mix-zone that adjusts the privacy level k over time in an online mode according to events such as fluctuations in vehicle traffic to achieve the highest anonymization notation. In Chapter 8, we proposed a dynamic mix-zone that adjusts the level of privacy over time in online mode and linear complexity, according to the flow of vehicles, to achieve higher anonymization. In our scenario, the output of step 2 corresponds to an attack of a dynamic mix-zone scheme to anonymize trajectories at different privacy levels over time. Also, a trajectory re-identification attack will be used to evaluate the effectiveness of privacy.

The third step, Protection, consists of anonymizing a dataset of trajectories to be published. The input consists of the dataset of trajectories and the LPPM developed, in this case the dynamic mix-zone scheme. Instances of this scheme are eventually distributed throughout the city by a mix-zone placement algorithm to obtain better data coverage. The output is the dataset of trajectories anonymized by the mix-zones.

3.4 Privacy, Quality, and Utility Indicator Extraction

RQ3: How do you measure the quality over time of a LPPM and the data it anonymizes?

The fourth step, Privacy, Quality, and Utility Indicator Extraction, represents the analysis and extraction of indicators from the data anonymized by the mix-zones in the previous step in terms of privacy, quality, and utility. In this step, the input represents a dataset of trajectories anonymized in the previous step, and the output is an indicator of privacy, quality, and utility of the data. The privacy effectiveness is assessed by submitting the anonymized trajectories to one or more LPAs, making it possible to combine several

attacks on location privacy. In our scenario, we apply the trajectory re-identification attack designed in step 2. We use TMA to measure the effectiveness of the attack in the mix-zones distributed throughout a city.

The quality of anonymization is another indicator measured in this step. The anonymization quality is achieved by Anonymization Quality Framework for Mix-zones (AQM). This framework, which has as input the anonymized trajectory dataset, is composed of quality metrics that measure the internal functioning of the mix-zones, such as NCM, Activation Time of the Mix-zone (ATM), IDM, and ITM. The quality metrics make up an objective function, AQ, which measures the quality of a mix-zone's functioning and consequently reflects on the quality of its anonymized data. In addition to analyzing the quality of the mix-zones and their anonymized data, AQ makes it possible to perform several analyses, for example, identifying and selecting mix-zones with the highest AQ over time. The output of AQ are indicators corresponding to the metrics and the AQ overall and per mix-zone. The implementation of Anonymization Quality Framework for Mix-zones (AQM) is found in Chapter 6.

RQ4: What are the smart cities domains, applications, and services that can best leverage mobility data anonymized by mix-zones?

Finally, the utility, concerns using anonymized data from mix-zones by applications and domains of smart cities. The utility questions are implemented by Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT), which takes the dataset of anonymized trajectories as input. This framework has two algorithms, the one based on Multi-Criteria Decision-Making (MCDM) and the other is an Utility for Specific Applications that measures the utility through twelve metrics related to privacy, mobility, and social, including mix-zones performance metrics from anonymized trajectories produced by mix-zones. Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT) aims to identify smart cities' domains, applications, and services where the anonymized data will provide more or less utility in various aspects. Also, it identifies which among several smart city applications a given dataset has the greatest utility. The output is various utility indicators to help in decision-making on the five steps.

3.5 Data Publishing Analysis

The fifth step, Data Publishing Analysis, corresponds to the analysis process for publishing the anonymized trajectory dataset, whether by the public or private sector, such as the data market. Using quality, privacy, and utility indicators, professionals, such as Data Privacy Specialist (DPS) and Privacy Engineer, can evaluate these data, including verifying where these data can be published in compliance with privacy regulations, such as Data Protection Regulation (GDPR).

The privacy indicator has a result in the TMA, a kind of re-identification rate, being for each region of the city protected by mix-zone and also an overall average TMA value, allowing for an overview of the level of privacy. This way, the DPS can identify which mix-zone has more and less privacy. For the regions with less privacy, the DPS can combine other privacy approaches, such as adding dummy trajectories to improve the privacy level for that region.

Regarding the quality, the indicators produced by AQM are quality and coverage metrics, including AQ of all mix-zones and AQ average. With these metrics, it is possible to have a more precise diagnosis of the functioning of mix-zones and particular anonymization concerns that coverage metrics cannot identify. Identifying precisely which mix-zones are malfunctioning and why, such as a lack of activations or low vehicle flow over time. In this way, the DPS can make decisions more quickly and efficiently about these mix-zones. Additionally, it is possible to get the performance of a mix-zone per period, such as anonymization and efficacy rates. Also, with quality metrics, the DPS enables the selection of mix-zones that yield data anonymization considering the quality, privacy, and utility analysis. Also, the DPS may select the best mix-zones deployment algorithm from an algorithm group in quality, utility, privacy, and coverage terms.

One of the utility indicators represents the utility of twelve privacy, mobility, and social metrics, including mix-zone performance metrics from anonymized trajectories. These utility and performance metrics can be used by DPS analysis, which one between social, privacy, and mobility aspects from the anonymized dataset had more and less utility in various aspects. Another utility indicator is the comparison of the utility of the mobility data based on privacy level for applications and domains, in which the DPS can decide which privacy level is adequate for those mobility data. Another utility indicator is the ranking of smart city application domains that best leverage mobility data anonymized by mix-zones. The ranking can help the DPS decide which area the anonymized dataset can be published from the following application domains: Statistical Analysis Urban Planning, Driver Behavior, Social Networks, Opportunistic Networks, Targeted Market, and Privacy & Safety. Another utility indicator is the utility of instances of datasets protected with different privacy levels. This indicator assists the DPS decides which anonymized mobility

dataset and privacy level is most useful for an application. The last utility indicator is the ranking of applications of smart cities ranked by application that can best leverage an anonymized mobility dataset protected by a privacy level.

In summary, the LPA can analyze the privacy, quality, and utility indicators, for instance, to verify the trade-off between AQ, privacy, and utility to decide the publishing domains for anonymized mobility data. We believe that SPF can substantially contribute to data analysis, both in the open data space and in smart mobility open data (SMOD).

3.6 Concluding Remarks

In this Chapter, we presented the Smart Privacy Framework (SPF), an Anonymization-based Framework for Smart Mobility Open Data, as well as the functioning of its components, is organized by steps, which, when followed within the framework, answer the research questions of this thesis.

We also present the privacy, quality, and utility indicators produced as Smart Privacy Framework (SPF) output. These indicators can assist Data Privacy Specialist (DPS) in analyzing anonymized data and in its decision-making for publishing smart mobility open data. The following chapters of this thesis detail the SPF implementation of each step. Specifically, the next Chapter, Chapter 4, presents the Mobility Impacts on Location Privacy step.

Chapter 4

The Impact of Mobility on Location Privacy in Smart Mobility

Location privacy is an issue addressed in many mobility contexts where privacy is a concern. Some proposals to tackle this problem exist, and some questions naturally arise: Are these proposals suitable for a dynamic environment, such as smart mobility? How are privacy and mobility related?

In this chapter, we answer these questions to explore location privacy in smart mobility, considering open and online data. We propose a characterization framework that evidences the hypothesis that mobility can affect privacy approaches of anonymization and obfuscation in the context of smart mobility. In Section 4.2, we presented relevant proposals for mobility analysis. In Section 4.3, we identified gaps in the privacy approaches, anonymization (mix-zones), and obfuscation (GEO-I) in the context of smart mobility. Next, in Section 4.4, we present the framework for analyzing location privacy with stay points. We evaluated the framework applied to 7 real datasets (mono or multi-modal mobility) in Section 4.5. Section 4.6 presents the new directions about location privacy in Smart Mobility. Finally, in Section 4.7, we present the final remarks of this chapter.

4.1 Introduction

Human mobility refers to the movement of human beings in space and time [229]. The study of human mobility has a fundamental role in developing smart cities, such as urban planning, estimating migratory flows, and developing traffic forecasting applications to help people, vehicles, and things move safely and efficiently [229, 230]. In this context, Smart Mobility emerged as an essential feature associated with smart cities [1]. For example, Smart Mobility can foster a smart transportation system that improves traffic safety and efficiency, reduces citizens' time commuting, and enhances the quality of life [5]. In such a smart transportation scenario, users can combine different trans-

portation modes (e.g., bike, bus, car, and walking) to reduce travel times, traffic, and air pollution. Moreover, people, vehicles, and things act as sensors and produce geo-tagged mobility data, which is valuable to help the management of assets and efficiently interact with resources and services in smart mobility scenarios. As a result, we typically find three dataset classes according to the transport mode prevalence and its granularity level:

1. Unimodal traces (UT): all trajectories contained in a dataset have a single transportation mode, i.e., there is only one vehicle type in the dataset (Figure 4.1a I)).

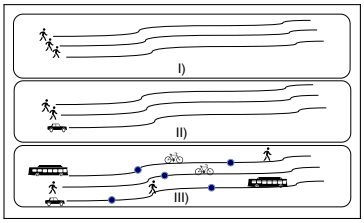
2. Multimodal traces with unimodal trajectories (MT-UT): in a single dataset, there may be at least more than one transport mode, but the trajectories are associated with a single transport mode (Figure 4.1a II)).

3. Multimodal traces with multimodal trajectories (MT-MT): in a single trajectory, there may be more than one mobility type. In this scenario, a user can take a taxi, then walk, and, finally, get a bus to reach his/her destination (Figure 4.1a III)).

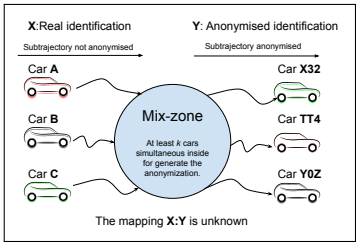
The benefits of smart mobility are clear, but there are also many privacy concerns. For example, mobility datasets contain not only a set of positions on a map or sensitive places such as home and workplace, among other Points of Interest (POIs). The contextual information attached to a trace tells much about the individuals' habits, interests, activities, and relationships [71]. Thus, malicious entities can be mining latent information on these datasets to identify and track users without their consent, which aggravates the privacy threats related to sharing multimodal mobility data, voluntarily or not [32]. Location privacy is a longitudinal issue in mobility. It is a particular type of information privacy for preventing other entities from learning one's current or past location [164]. Notably, it gained attention when the Internet of Things (IoT) and the Internet of Vehicles (IoV) contributed to smart mobility connecting objects sharing location information, but unrestricted [231].

Some traditional strategies for anonymization and obfuscation, such as Mix-zones [56] and GEO-I [232], respectively, are being applied to provide location privacy. However, when this happens, we have to ask whether these traditional location privacy approaches suit such a new smart mobility environment, as Scenarios 2 and 3 described above. Location privacy solutions based on obfuscation or anonymization are generally static regarding the setup of their parameters. They are not tuned for different types of datasets and their scenarios and, thus, are not resilient against heterogeneous mobility data containing other mobility behaviors. Furthermore, what is the degree of impact of mobility on location privacy?

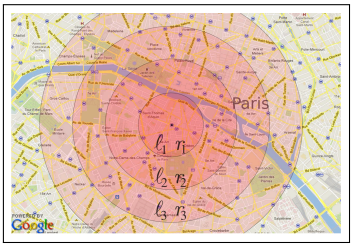
In this chapter, we explore the influence of mobility on location privacy. For this, we propose a framework that allows us to characterize and analyze the similarity between types of transport modes through metrics extracted from Stay Point (SP) – a region in



(a) Mobility in smart mobility



(b) Mix-zone with $k = 3$.



(c) GEO-I: privacy levels about r. Source: [232]

Figure 4.1: 4.1a Mobility scenarios present in smart mobility datasets. Item I) UT; II) MT-UT; III) MT-MT. 4.1b Mix-zone where three cars with pseudonyms A, B, and C enter a mix-zone and attend the minimal $k = 3$ and at the exit, receive new pseudonyms (TT4, Y0Z and X32, respectively) without any association with previous ones, cloaking their identities. 4.1c GEO-I scenario where the privacy level is proportional to the radius.

which an entity stays for a minimum time interval [233]. Particularly, we analyze two SPs metrics: Stay Point Count (SPC), which represents the total of different locations visited by users; Stay Point Duration (SPD), which refers to the time a user spends at a location. Unlike other mobility metrics, SPs can provide directions for parameter settings of LPPMs techniques, such as the coverage radius size and noise level. Those SPs can be used as indicators of POIs and traffic intensity to choose the best place to apply LPPMs [233, 234, 235]. Thus, when studying the distributions of SPs metrics from different transport modes, we found that each transport mode can influence the LPPMs parameters tuning. In the literature, several approaches have used equal LPPMs parameters for different transport modes [41, 82, 146]. However, this work aims to provide empirical evidence on the impact of mobility on location privacy for different transport modals, without delving into the theoretical analysis, that location data from different modes of transport must be protected by LPPMs calibrated independently.

We conducted a comprehensive evaluation applied to seven real datasets for these analyses (in which six of them refer to the case UT, and one refers to the cases MT-UT and MT-MT). The results showed that the SPC metric reached 100% and 83.3% accuracy for coarse-grained (person and vehicle) and fine-grained (bus, taxi, and person) data, respectively. Additionally, we show a similarity between distributions for the same vehicle type for mono and multimodal datasets. Results suggest that mobility has a high impact on privacy in different granularity levels, enabling us to build a resilient mobility-aware privacy solution. To our knowledge, this is the first study that analyzes stay points to observe the impacts of mobility on location privacy. Since stay points are an essential step in location privacy research, they serve as a powerful mechanism for positioning and setting privacy approaches.

4.2 Mobility Analysis and Fallacies on Location Privacy

This section presents some relevant proposals in the literature about mobility analysis and fallacies found in studies of location privacy.

Table 4.1: Related work about statistical analysis of mobility.

Ref	Metric	Acc. Model[1]	#Datasets	Modal	Sim. Analysis	Sim. Method[1]	Loc. Privacy
[236]	Banknotes	AIC	1	people	No	-	No
[237]	Trip Displacement	AIC	2	walk/run,bike, train/subway, car/taxi/bus	No	-	No
[230]	Trip Displacement	AIC	3	bus, taxi, subway	partially	-	No
[238]	Trip Displacement, Trip Duration, Trip Interval	MLE, BIC	2	taxi, subway	No	-	No
[239]	Degree Distribution, Friendship Node, Hashtag	AIC, BIC, SSE Approach proposed	2	OSN	No	-	No
[240]	Yellow Intervals, Driver's Age, Gender, Phone Status, Maximum Decel./Accel., Vehicle's approaching Speed, Distance between Vehicle's Position, Stopping line when yellow light go up.	AIC, BIC	1	car	No	-	No
[241]	Distr. of Displacements, Waiting Time	AIC	3	people	No		No
[242]	Distance and Transfer Time between POIs	AIC	4	bus, taxi, subway	partially	Pearson Corr.	No
[243]	Srrive-stay-leave, Trajectory Entropy, Number of Trips, Average Velocity, Trip Length, Total Driving Days, and Average Mileage per Day	MLR, DCNN, Approach proposed	3	private car, taxi	yes	MSE, RMSE, MAE, KL, R^2	No
Our work	Stay Point Count (SPC), Stay Point Duration (SPD)	AIC, SSE	7	people, bus, taxi, private car, multimodal	Yes	Wasserstein	Yes

[1] AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; MLE: Maximum Likelihood Estimation; SSE: Sum of Squared Estimate of Errors; MLR: Multiple Linear Regression; DCNN: Deep Convolutional Neural Network; MSE: Mean-square Error; RMSE: Root-mean-square Error; MAE: Mean Absolute Error, KL: Kullback–Leibler divergence; R^2: degree-of-fit test and the closer the R^2.

4.2.1 Statistical Analysis of Mobility

The analysis of statistical properties of human mobility can reveal valuable insights for developing various services, including opportunistic networks, traffic monitors, and recommender systems [237, 238, 244, 245]. Thus, there is a broad study on the characterization of distributions extracted from metrics of different mobility contexts, such as banknotes in human mobility, trip displacement in mono/multimodal transport, degree distribution in Online Social Networks (OSNs), yellow intervals on intersections from monomodal transport, and distance and transfer time between POIs in-place semantics context. However, there is little research to understand location privacy from the statistical analysis of mobility aspects, such as stay point metrics. Following, we highlight some relevant literature proposals that pursue distribution characterization of the modal datasets focusing on stopping metrics, such as stopping time in semaphores' yellow light, waiting times between displacements (or trip interval), and traffic accident duration.

Brockmann et al. [236] explored traveling statistics of human mobility over a million

individual displacements by analyzing banknotes' circulation in the United States. They concluded that the distribution of the traveling distances decays as a power law, indicating that trajectories of banknotes are similar to Lévy flights [246]. Also, they showed that the probability of pause time distribution (staying in a restricted region) is characterized by a long tail leading to a sub-diffusive process.

Zhao et al. [237] explored the Lévy walk behavior of human mobility [246]. They decomposed mobility patterns of multimodal datasets into different classes according to transport modes such as Walk/Run, Bike, Train/Subway, and Car/Taxi/Bus [237]. They concluded that human mobility could be modeled as a mixture of different transport modes. Moreover, single movement patterns can be approximated by a log-normal distribution rather than a power-law distribution.

Xia et al. [238] also analyzed human mobility in both subway and taxi with three metrics: trip displacement (TDis), trip duration (TDu), and trip interval (TIn). The results showed that TDis patterns by subway and taxi are similar and follow log-normal distribution rather than an exponential model. Additionally, TDu on weekends differs from weekdays, no matter the modal. The TDu metric is fitted to Weibull distribution for subway and log-normal distribution for taxis. For TIn, they concluded that the Weibull distribution could fit the probability curve by taxi rather than the log-normal distribution. For the subway, the TIn obeys the distribution composed of Weibull and log-normal distributions.

Li et al. [240] explored the stopping behavior during yellow intervals on the semaphores. Notably, they evidenced that the survival curves extracted from stopping time confirm the existence of group-specific effects on drivers based on two metrics: stopping time and drivers' age. The results showed that the log-logistic-based frailty model with age as a grouping variable presents the best goodness of fit and prediction accuracy.

Zhang et al. [244] investigated the prediction curves of traffic accident duration, which provide an important basis for traffic mitigation measures after accidents. They applied AIC and BIC to fit the probability distribution of the accident duration, and the results showed that the log-normal distribution fitted best.

Alessandretti et al. [241] verified the relationship between spatial and temporal properties of human mobility using trajectories of 850 individuals of Copenhagen Network Study composed of GPS and Wi-Fi data. They showed that a log-normal distribution best describes the distribution of displacements and waiting times between them. They also noticed correlations between displacement length and the waiting time at the destination.

There are also studies about understanding place semantics in mobility [242] and hot zones evolution [243]. Papandrea et al. [242] noticed that POIs have some statistically similar properties among individuals. They classified POIs based on their relevance on a per-user basis. Furthermore, they applied travel metrics (spatial and temporal distances) to four datasets: trajectory, continuous mobility datasets, and two CDR datasets. The

trajectory dataset is defined as a unique trip with origin and destiny. In contrast, after starting in the continuous mobility dataset, the user sampling never stops unless the sample collector gets switched off, yielding many trips in a trajectory. The results showed a correlation between these metrics in trajectory datasets.

Xiao et al. [243] investigated the spatiotemporal evolution of urban hot zones (a kind of POIs) from stay points behavior on private cars dataset. They noticed that the hot zones' formation is intricately related to the spatiotemporal coupling correlation of stay points, and its spatiotemporal variation shows certain predictability. Furthermore, they analyzed mobility patterns between taxis and private cars with trajectory entropy, number of trips, average velocity, trip length, total driving days, and average mileage per day. They concluded that taxi trips, unlike private cars, are different, irregular, and have a high degree of randomness.

Despite the vast literature on the statistical analysis of mobility, few proposals analyze privacy from the perspective of mobility metrics. Further, there are few studies about the similarity between different data sources and transport modes. Alessandretti et al. [241] initially analyzed Pearson's correlation between the two datasets. Although the proposal of Papandrea et al. [242] and Xiao et al [243] presented relevant contributions for human mobility analysis, they did not compare similarity levels between datasets. Further, to the best of our knowledge, no previous proposal analyzed these metrics when characterizing location privacy.

Unlike previous studies, we advance the state of the art w.r.t. SPs. We explore the stay points metrics for characterizing and evaluating the impacts of mobility data on location privacy. The stay point metrics stand out in location privacy over the mobility metrics discussed above. Once mining the SPs, it is possible to get valuable information about the users' mobility profile (e.g., whereabouts and diary routines) and then define the best placement and configuration of LPPM's instances, such as radius, noise level (ϵ), and k[1]. For this, we propose an analytical framework to analyze two SP metrics extracted from different transport modal datasets (see Section 4.4). Table 4.1 summarizes the related work discussion.

4.2.2 Mobility Effects on Security and Location Privacy

In a smart mobility scenario, people with smartphones, vehicles, and things can be seen as mobile devices, allowing users to access a rich set of mobile services. Nevertheless, these resource-constrained devices need fast and easy access to mobile services without

[1]Minimum of entities into a region for anonymizing.

multiple credentials of the users. In this way, password-based single-sign-on authentication has been widely applied in mobile environments. An authentication token is generated on an identity server, and one can request mobile services from related service providers without multiple registrations [247]. However, this model introduces privacy and security threats. For instance, if an adversary accesses the identity server, one can retrieve users' passwords by performing Dictionary guessing attacks (DGA) and overissue authentication tokens to break the security [247, 248].

Other security and privacy issues occur when cloud services provide data deduplication to users equipped with mobile devices to save storage space. For instance, the user's data must be cyphered by a symmetric encryption method like Message-locked encryption (MLE) to avoid leaking private information. However, MLE is vulnerable to brute-force DGA. Additionally, MLE schemes are subject to key management problems, mainly if users access different devices. Thus, to mitigate DGA and key management problems, some proposals have focused on applying secure distributed secret sharing protocols [248, 249].

In different location privacy studies [41, 82, 146], we can see evidence of the impact of mobility on privacy by analyzing the performance discrepancy of an LPPM applied to different transport modes. For instance, in vehicular mobility, some proposals showed significant differences in the re-identification rate between the datasets of buses and cars of the same city [41]. Other studies also found divergences in the re-identification rate in datasets of different cities, such as datasets of cabs in Rome and buses in Shanghai [146]. In both studies, datasets were submitted to LPPMs and configured with the same parameters.

Some investigations have identified privacy divergences of user registers [82] in datasets, which are not all equal in the face of re-identification attacks. This means that some users' profiles can never be re-identified even in the absence of LPPMs, while others can be easily re-identified. The authors argued that this difference in users' protection level is that no generic LPPM provides the same protection level for different users' profiles. Moreover, the resilience of an LPPM against re-identification attacks depends on the underlying data. For instance, the LPPM settings to protect the location data of a user walking and using his/her smartphone may differ from those of a user driving a private car due to mobility characteristics such as speed, direction, and frequency of visits to places.

One of the reasons for these accuracy differences of re-identification attacks is that possibly these datasets were protected by inappropriate or misconfigured LPPMs or with the static setting, which did not consider the dynamic scenarios with different transport modes. As a result, we have datasets with low protection and utility. Next, we emphasize some essential privacy issues when using classical LPPMs to smart mobility.

4.3 Location Privacy Issues in Smart Mobility

This section shows the classical LPPMs and collection of issues concerning privacy and mobility aspects addressed in smart mobility, organized in three scenarios: general, anonymization, and obfuscation. Nevertheless, we must first understand privacy threats through an adversary model.

4.3.1 Adversary Model

Defining a consistent adversary model is important to outline a location privacy attack's limits. This model allows a panoramic view of privacy threats and defines more appropriate mitigation actions. Therefore, we present an adversary model capable of carrying out both anonymized and obfuscated data attacks.

The adversary model can be defined as follows. Let $\mathcal{F}$ and $\mathcal{G}$ be two functions that represent the LPPMs anonymization and obfuscation, respectively. Also, let $\mathcal{D}'$ be an open dataset $\mathcal{D}$, but protected by $\mathcal{F}$ or $\mathcal{G}$, being $\mathcal{D}' \leftarrow \mathcal{F}(\mathcal{D})$ or $\mathcal{D}' \leftarrow \mathcal{G}(\mathcal{D})$. The adversary may also have access to some training traces (possibly noisy or incomplete) of users and other public contextual information, represented by a profile B_u for each user u. The above information applied to $\mathcal{D}$ represents the adversary's background knowledge about the users $\mathcal{B} = (B_1, \ldots, B_b, \ldots, B_m)$, where $[1 \leq b \leq m]$ and b represents the number of elements in $\mathcal{B}$ known by the adversary, enabling the adversary to execute a Tracking Attack $\mathcal{Z}$ or Points of Interest Attack $\mathcal{W}$, for example.

In a Tracking Attack (TA), the adversary's objective is to determine the whole sequence (or a partial sub-sequence) of events in a user's trace. Given an anonymized dataset $\mathcal{D}'$ composed of users and background B_u, a tracking attack is defined as $T_u \leftarrow \mathcal{Z}(\mathcal{D}', B_u)$, where T_u represents the re-constructed trajectory of user u.

A Points of Interest Attack (POIA) uses location points or regions where people commonly stay at a given instant, such as home or workplace, to characterize users' profiles. Given a set of regions on map $R \in \mathcal{B}$, a period of time t, and an obfuscated dataset $\mathcal{D}'$ composed of users. An adversary can be interested in discovering the most visited locations L_s of users s in $\mathcal{D}'$ at time t. That is, $L_s \leftarrow \mathcal{W}(\mathcal{D}', R, t)$. The exact location is not needed, but the region on the map.

We can observe privacy threats through an adversary model, even if an LPPM protects data $\mathcal{D}$ from adversary's background knowledge $\mathcal{B}$, which is the type of LPPM (anonymization $\mathcal{F}$, or obfuscation $\mathcal{G}$) applied in $\mathcal{D}$. For example, if the adversary has $\mathcal{B}$

in which the $\mathcal{D}$ was protected by $\mathcal{G}$. The adversary knows that users' sensitive locations have been obfuscated and may have low success with a POI attack $\mathcal{W}$, but the users' identity was not protected. Thus, the adversary can be highly accurate in identifying the identity of users with tracking attack $\mathcal{Z}$. Likewise, if the adversary has $\mathcal{B}$, which $\mathcal{D}$ was protected by $\mathcal{F}$, a POI attack $\mathcal{W}$ will have a high accuracy in identifying the location, as the POIs was not protected. A way to protect open data is to apply a hybrid LPPM based on anonymity and obfuscation. However, we encountered issues between privacy and utility, as defined below.

4.3.2 General Scenario

In smart mobility open data, there are issues related to the decision about the type of privacy to achieve, the order of applying these approaches, and how to set up the LPPMs to get an optimal trade-off between obfuscation and anonymization. However, applying these techniques does not mean achieving full location privacy; instead, it can be effective for datasets used for a specific purpose that requires one type of protection over another. Depending on the systems that use the protected dataset (called consumers), these issues can affect/change privacy and utility goals.

Suppose a congestion reduction scenario where it is necessary to test a system that monitors traffic flow. In the test dataset, fine-grained geolocation points are needed for measuring the traffic efficiently, and a high distortion of this data may lose its utility. In this case, the test dataset should consider obfuscation since it is more sensitive than anonymization. Thus, this dataset should have a high level of anonymization and a low level of obfuscation. These questions are described in Figure 4.2.

Questions **Q1** and **Q2**, hybrid mechanisms (obfuscation and anonymization) deal with a new trend for the LPPM design but, at the same time, pose new challenges. For instance, hybrid mechanisms are intricately dependent on the context in which each customer's family will use the protected dataset.

Question **Q3** addresses the issue that there is no one-size-fits-all LPPM for many privacy scenarios without losing privacy and utility, such as in datasets MT-UT and MT-MT. We detail this issue and directions for a possible solution in the following.

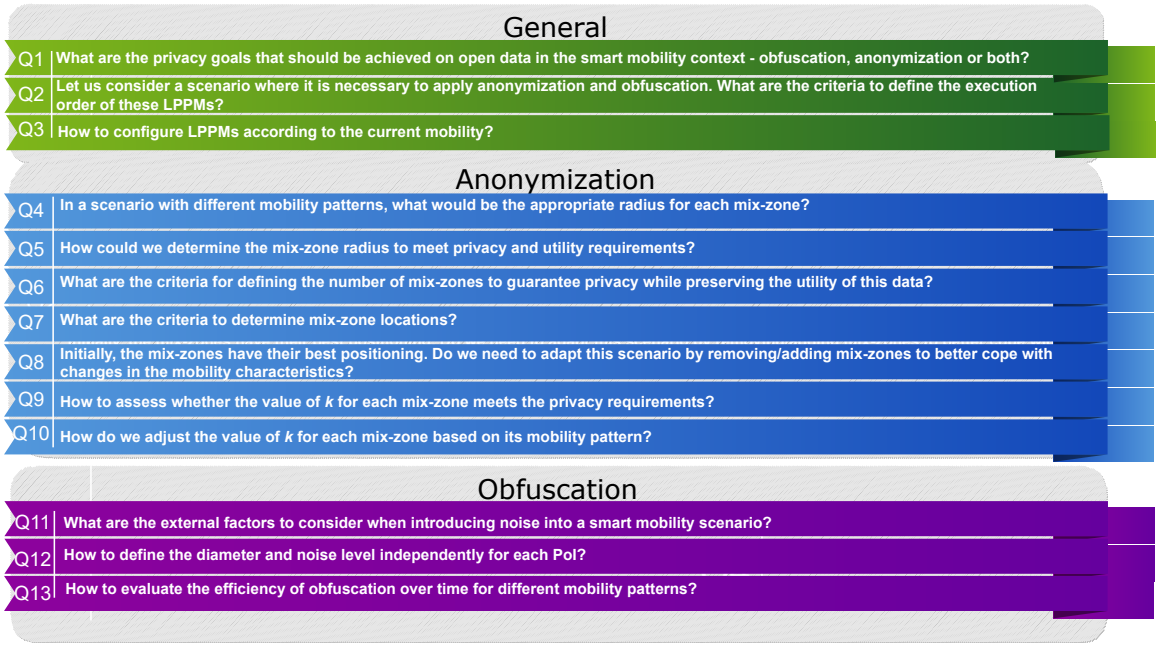


Figure 4.2: Location privacy issues in smart mobility.

4.3.3 Anonymization Scenario

Mix-zones is a technique widely used in VANETs to anonymize mobility data. It selects urban regions where the simultaneous anonymization of vehicles (or people) occurs by changing their current pseudonyms [56]. To anonymize them, we must have at least k entities within the mix-zone (see Figure 4.1b). The mix-zone parameters are the radius (r), the minimum number of entities in the mix-zone to change the pseudonym (k), and the geo-position. Although there are studies on mix-zones to optimize anonymization effectiveness, whether in silence period strategies [186, 250], cryptography [170], positioning [190], or modeling of its geometric region [165], few efforts have been conducted to investigate mix-zones in the context of smart mobility open data. Specifically, the privacy and data utility side effects when the LPPM parameters are not calibrated in an environment with different modals. Here, we have identified some issues related to these concerns.

From a spatial point of view, we need to calibrate the mix-zone radius (r) for different entities, considering the trade-off between privacy and utility. We can gain anonymity coverage with a larger radius since more entities are likely to be inside the zone simultaneously. However, within a mix-zone, there is a period of silence in which location records are discarded. Therefore, the larger the mix-zones radius, the greater the data gaps, which might compromise the dataset's utility. Additionally, people's mix-zone radius may be smaller, as people tend to have a lower speed than vehicles. But what are the radii for two different types of entities (like MT-UT and MT-MT scenarios) within mix-zones?

Other spatial issues are the number and location of mix-zones on a map, depending on that area's mobility characteristics. If high data privacy is desirable or needed, the mix-zones must be positioned in higher traffic regions [175]. The positioning of mix-zones over time is a temporal issue that we need to consider. Once the best positioning for mix-zones is defined, these points may lose their effectiveness to anonymizing over time. With the flow variation of entities, some mix-zones may not make sense anymore, whereas other regions may need them. Thus, the problem is how to define the lifetime of a mix-zone.

Tuning the parameter k of a mix-zone is also affected by a given area's mobility characteristics. In high vehicle traffic or crowd, it is desirable to have a high value of k. In contrast, a low-traffic region requires a small value of k, which must be defined to cover data anonymization and a significantly higher privacy level. Also, k is decisive for pseudonyms changing. A k low value may yield excessive pseudonym changes that imply a high anonymity level. However, this harms open data once it produces a significant number of sliced trajectories, losing its utility. Further, in an online context, the excess on pseudonyms changing can negatively affect communication protocols (e.g., routing task) and applications that need long-term communication relationships (e.g., file transfer or interactive chat sessions) [250, 186]. Figure 4.2 presents a summary of these anonymization issues.

4.3.4 Obfuscation Scenario

An obfuscation scenario also presents many issues concerning privacy and utility. For example, GEO-I [232] is an obfuscation technique based on data perturbation. It protects the user's location by adding spatial noise extracted from a Laplace distribution to the actual user's location in the mobility trace [232]. GEO-I considers the privacy level l to be proportional to the radius r, and defines an ϵ-geo-indistinguishability as $\epsilon = l/r$. The value of ϵ represents a level of privacy for l within r and proportionally selects a privacy level for all other radii, observing that the lower the ϵ, the higher the noise (see Figure 4.1c). The GEO-I approach is necessary to consider the noise level applied to data. If it is high, one may risk distorting the data by losing its utility. Nevertheless, if the noise level is low, the data may not be properly protected to ensure privacy. Therefore, the noise level must be adjusted to handle this trade-off.

Another parameter to consider is the GEO-I radius to identify the user's POIs and introduce noise. The greater the radius, the greater the chance of identifying the POIs of a mobile entity, but also the greater the noise to be applied to protect the region. Thus, the radii may have different values for datasets with multimodal trajectories. The radius

to identify the vehicle's POIs may differ from that of people's. People's POIs usually have smaller perimeters, such as homes, workplaces, tourist points, and campus buildings. In contrast, vehicles' POIs have a larger perimeter, such as the airport area for POI of taxis or a parking yard for car rental. However, what would be the ideal radius for MT-UT and MT-MT scenarios? Must these POIs radius/noise levels be varied with time? Figure 4.2 presents a summary of these obfuscation issues.

Based on these facts, we can state the following hypotheses:

> ***Hypothesis 1****: Mobility reflects directly on location privacy.*
>
> If we consider Hypothesis 1 valid, which is quite reasonable, then we have:
>
> ***Hypothesis 2****: Different types of mobility need different profiles of privacy and utility, independent of the applied privacy model.*

These hypotheses are not trivial to validate. However, a possible direction is to understand how privacy behaves in different mobility patterns to address Hypotheses 1 and 2. In this work, we propose analyzing the impacts of mobility on location privacy by analyzing SPs, as a substrate used in many privacy algorithms, to identify specific behaviors in different transport modes. Specifically, we intend to characterize and analyze distributions extracted from SPs of different datasets and compare them. The goal is to verify the similarity between distributions of the same and different transport modes. The similarity level is expected to be low between different transport modes and high for the same modes of transportation. In this way, we highlight the hypothesis. More details are presented below.

Table 4.2: Datasets details.

Id	Name	Location	Transp. type	#users	#reg.	#Staypoints[1]
brig	Brightkite [251]	LBSN	human	51,406	4,747,287	2,679,758
gow	Gowalla [251]	LBSN	human	107,092	6,442,892	3,733,344
cabs	Cabspotting [252]	San Francisco, USA	taxi	532	6,837,027	28,589
t-dri	T-Drive [253]	Beijing, China	taxi	10,320	17,652,648	187,183
dubl	Dublin Bus [254]	Dublin, Ireland	bus	911	43,851,182	52,961
rio	Rio Bus [252]	Rio de Janeiro, Brazil	bus	13,954	51,845,217	570,894
geo	Geolife [125]	Beijing, and 36 cities in China, USA, South Korea, and Japan.	multimodal	182	24,876,978	27,862
geo-car	Geolife cars [125]	—	cars	36	512,807	770
geo-cabs	Geolife cabs [125]	—	taxi	29	242,018	449
geo-bus	Geolife bus [125]	—	bus	43	1,276,632	1679
geo-human	Geolife human [125]	—	human	61	2,535,433	3320

[1] Stay points setup $SP_p\langle R = 250, T = 30\rangle$ for brig and gow datasets, $SP_v\langle R = 500, T = 30\rangle$ for cabs, t-dri, dubl, rio, geo, and geo-* datasets.

[2]This algorithm is also used for the *stay point duration* metric. To do this, replace Line 7 with the function that calculates the stay point duration metric.

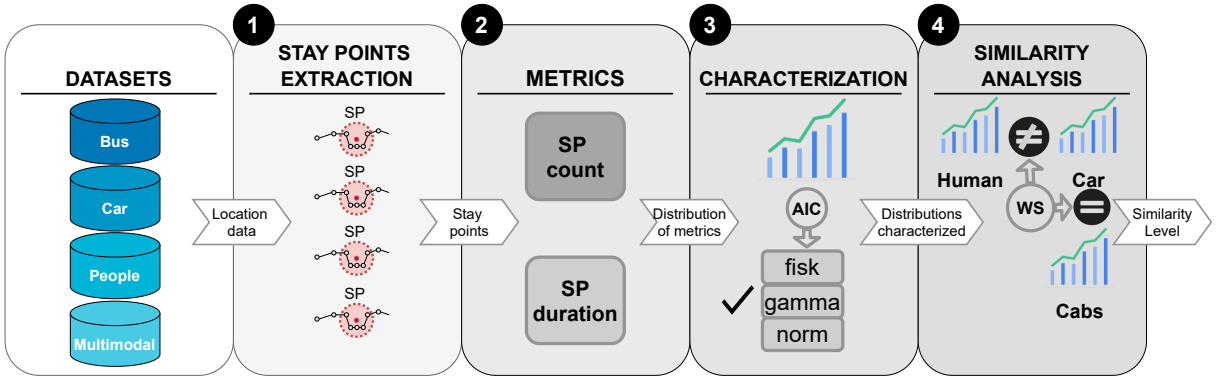


Figure 4.3: The framework for analyzing location privacy with stay points.

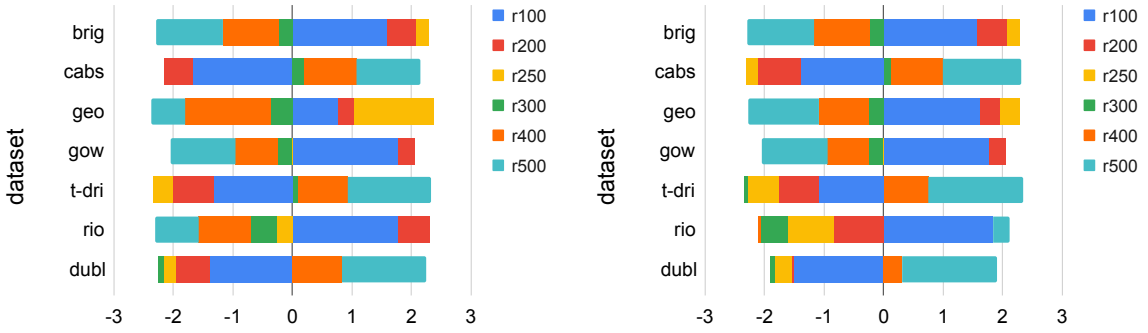
Algorithm 1: Characterization and Similarity Analysis of Stay Points Count metric[2].

Data: $\mathfrak{D} = \{\mathcal{D}_1, \mathcal{D}_1, ..., \mathcal{D}_n\}$ set of distinct datasets.
Result: Υ: best fit distribution for each $D_i \in \mathfrak{D}$; S: similarity matrix; Φ: Accuracy matching of mobility group;

for $\mathcal{D}_i \in \mathfrak{D}$ **do**
  $P_i \leftarrow$ Extract_SP_Param($\mathcal{D}_i$)
  $SP_i \leftarrow$ Extract_SP($\mathcal{D}_i, P_i$)
  $\mathcal{SP} \leftarrow \mathcal{SP} \cup SP_i$
end
for $SP_i \in \mathcal{SP}$ **do**
  $SPU_i \leftarrow$ Extract_SP_by_User(SP_i)
  $\mathcal{SPU} \leftarrow \mathcal{SPU} \cup SPU_i$
end
for $SPU_i \in \mathcal{SPU}$ **do**
  $distr_i \leftarrow$ Best_Fit_Distr(SPU_i)
  $\Upsilon \leftarrow \Upsilon \cup distr_i$
end
for $SPU_i \in \mathcal{SPU}$ **do**
  for $SPU_j \in \mathcal{SPU}$ **do**
    $S[i,j] \leftarrow$ WM(SPU_i, SPU_j)
  end
end
$\Phi \leftarrow$ ACC(S)
return Υ, S, Φ

Table 4.3: Stay points values from analysis radius and time to stay.

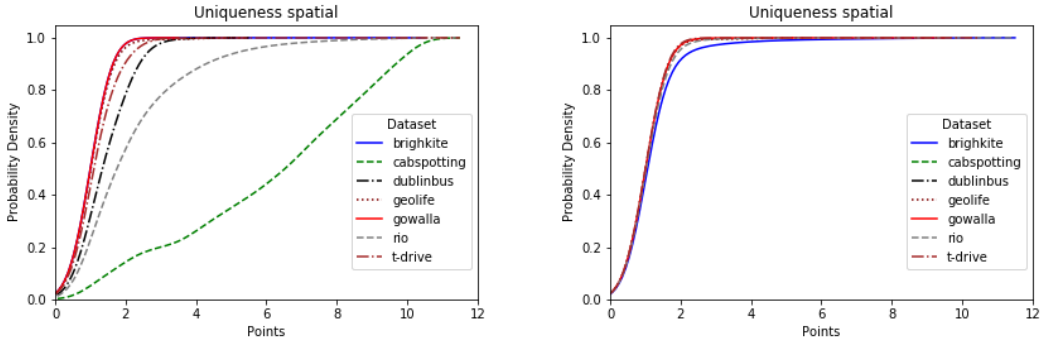
	t15						t30					
Dataset	**r100**	**r200**	**r250**	**r300**	**r400**	**r500**	**r100**	**r200**	**r250**	**r300**	**r400**	**r500**
brig	**3050695**	2874132	2826407	2753173	2642387	2611167	**2882499**	2722982	2679758	2612758	2509555	2481030
cabs	31437	31467	31479	31484	31501	**31506**	28453	28487	28513	28530	28566	**28589**
geo	37349	37295	**37409**	37229	37117	37206	**28433**	27864	27862	27603	27360	27198
gow	**4340465**	3967355	3894363	3838986	3718058	3620534	**4146983**	3799599	3733344	3683107	3571545	3480181
t-dri	319472	356156	380332	405721	450533	**486246**	162470	166066	167868	171871	179502	**187183**
rio	**893318**	839561	803565	797655	777507	783632	**582907**	562466	563038	565332	568527	570894
dubl	93308	96572	98176	98668	102506	**104893**	51879	52390	52308	52371	52511	**52961**



(a) Radius [100-500] and time to stay at least 15 mins.

(b) Radius [100-500] and time to stay at least 30 mins.

Figure 4.4: Stay points extraction with many radius and time to stay time threshold.



(a) Uniqueness of trajectories.

(b) Uniqueness of stay points.

Figure 4.5: Uniqueness spatial for datasets defined on Table 4.2.

4.4 A Framework for Location Privacy Analysis with Stay Points

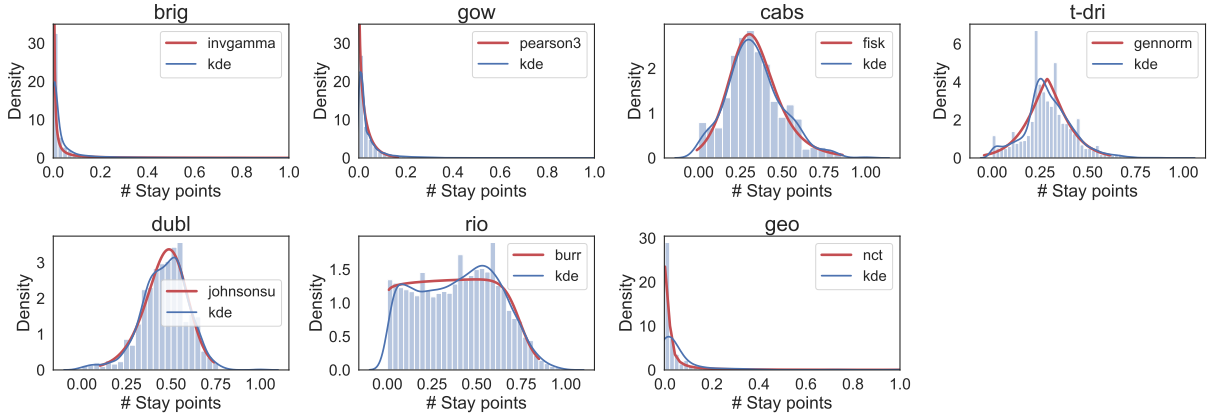
This section presents an analytical framework for analyzing location privacy with SPs. Firstly, we define the SPs and their relationship to the LPPMs. Next, we detail the steps for extraction, characterization, and similarity analysis of stay points metrics.

4.4.1 Analyzing Location Privacy with Stay Points

Stay Point (SP) is a region where an entity stays for a minimum time interval [233]. The parameters of an SP are the radius r in meters of the region and the minimum time to stay there t in minutes. These points are relevant for detecting many mobility characteristics, such as traffic lights and even traffic jams. Stay points are commonly used

Best Distr.	brig	gow	cabs	t-dri	dubl	rio	geo
invgamma	**−207,956**	−392,287	−3,218	−54,411	−4,966	−111,489	−430
pearson3	−78,572	**−452,154**	−3,189	−54,387	−5,406	−112,594	72
fisk	−136,175	−416,304	**−3,377**	−55,407	−5,332	−110,197	122
gennorm	−73,731	−321,543	−3,171	**−55,626**	−5,276	−123,696	−180
johnsonsu	−83,889	−421,870	−3,214	−55,570	**−5,496**	−112,781	−207
burr	−114,167	−437,034	−3,167	−55,216	−5,431	**−134,550**	223
nct	−205,646	−387,561	−3,228	−55,435	−5,112	−112,661	**−451**

(a) Best distribution and AICc(SSE) value for the SPC metric.



(b) Distribution of each dataset and best-fit distribution (in red color: probability density function (pdf)), calculated with AICc(SSE).

Figure 4.6: Best fit distribution for SPC metric.

as a substrate for many privacy mechanisms in the context of location privacy. In LPPM design, stay points are typically used to detect POIs and apply obfuscation methods. Additionally, stay points can be used for mix-zones placement [32]. In location attacks, stay point mining enables identifying and characterizing behaviors in the users' trajectory, revealing sensitive information, such as social preferences, since victims regularly go to those places [32, 197]. In this way, stay points bring valuable information w.r.t. location privacy.

Figure 4.3 details the steps for analyzing location privacy with stay points. Step 1, we extract SP from seven different real mono and multimodal mobility datasets (in which six of them refer to the case UT and one referring to the cases MT-UT and MT-MT). Step 2, from each SP set, we extract the distributions of two SP metrics: SPC and SPD. Step 3, we characterize the distributions according to the best-fit statistic model. Step 4, for each SP metric, we compare the similarity between datasets' distributions. Specifically, we verify if there is a divergence between the distributions for vehicles and people in terms of SPs. Next, we refine the vehicle category for cars, cabs, and buses. We use an accuracy metric to quantify the number of distributions that match each other. The steps are defined in Algorithm 1 and detailed in the following sections.

4.4.2 Extraction of Stay Points

The number of collected SPs on stay points extraction is related to their radius and time to stay parameters. Many empirical studies in the literature set up the SP parameters [233, 255]. For instance, Zheng et al. [233] argued that radius and time to stay parameters enable finding significant places, such as restaurants and shopping malls. At once, it is possible to ignore geo-regions without semantic meaning, like places where people wait for traffic lights. They extracted 10,354 stay points from the dataset with 107 users using mobile devices in Beijing, including 36 cities in China and a few cities in the USA, South Korea, and Japan. Chen et al. [255] used spatial density clustering and temporal Gaussian Kernel Density to extract spatiotemporal features from sparse and non-stationary stay behavior data. Concerning SP set-up parameters, questions naturally arise: What is the setup of these parameters to obtain optimal stay point extraction? Is there a setup pattern for datasets of different natures, such as transport modals?

Here, we analyzed the parameter tuning for the SPs extraction, considering various transport modals to answer these questions (Line 2 of Algorithm 1). The goal is to identify how tuning the radius and time parameters affects the result set of SPs in different transport modals collected from seven datasets (see Section 4.5). We defined a testing scenario of ranging the radius threshold r from 100 meters to 500 meters ($S_R = \{100, 200, 250, 300, 400, 500\}$), and the time threshold S_T from 15 minutes to 30 minutes ($S_T = \{15, 30\}$). For each combination of radius $r \in S_R$ and time value $t \in S_T$, we extracted the set of SPs for all users in each mobility trace (see Table 4.3). We can see that the settings to extract stay points for people and vehicles are different.

For the people datasets, the best configuration was the smallest radius ($r100$), whereas for the vehicle datasets was the largest radius ($r500$). This fact is best seen in Figures 4.4a and 4.4b that represent the z-score standardization of the stay points count in the datasets for all the radius and time values $t15$ and $t30$. For both time configurations, the vehicle datasets – cabs, t-dri, rio and dubl – got a positive score above value 2, i.e., collected more SPs with $r500$. The only exception was the rio dataset, which for $t15$ prevailed $r200$. The people datasets gow, brig and geo prevailed radius $r250$, $r200$ and $r100$, respectively. Geolife is a multimodal dataset, but most of its mobility records refer to people, which leads to bias, with lightning configurations below $r250$. The SP extraction results suggest that the transport modal significantly influences the parameters' choice to collect a more significant amount of stay points. Therefore, we adopted two parameters set, $SP_p\langle R = 250, T = 30\rangle$, $SP_v\langle R = 500, T = 30\rangle$, applied to human and vehicular datasets, respectively. These stay points setups are already used in [233] for people datasets and in [255] for vehicle datasets, as they have contexts close to ours.

Optimal tuning of SPs parameters is context-dependent and has several open issues [233, 255].

4.4.3 Uniqueness and Stay Points

In a preliminary analysis, we evaluate whether SPs can be used to understand users' privacy of different trajectory datasets. Also, verify if the privacy of SPs is similar between the same and different transport models. For this, we explore the uniqueness of the SPs. *Uniqueness* is a privacy metric that estimates the number of points p needed to identify the mobility trace of an individual uniquely. The fewer points required, the more unique the trajectories are and the easier to re-identify a user using outside information [63]. In the same way as trajectory, uniqueness can be used to measure privacy with SPss. People can be re-identified by where they often spend their time, such as at home or work, which can be a SP. We analyzed the uniqueness with two aspects – trajectories and SPs – to verify the similarities and divergences regarding privacy.

Uniqueness is calculated as follows. Let a dataset D contain mobility trace T_u for each user u. The uniqueness of a trace given a set I_p of p randomly chosen spatiotemporal points. A trace is said to be compatible with I_p if $I_p \subseteq T_u$. Let $S(I)$ be the set of users in which mobility traces T is compatible with I_p. All mobility traces in the dataset T_u are successively tested for matching with I_p. A trace is characterized if the set of traces compatible with the points contains at most x users, that is, $|S(I_p)| \leq x$. A trace is uniquely characterized if the set contains precisely one trace, i.e., $|S(I_p)| = 1$. The definition of uniqueness can be adapted to stay points. In this case, we use stay points (I_{SP}) instead of spatiotemporal points I_p.

4.4.4 Stay Points Metrics

After the SPs extraction step, we use the metrics related to the stay points aspects: *Stay Point Count (SPC) per user* and *Stay Point Duration (SPD)* (Lines 6–9 of Algorithm 1).

Stay Point Count (SPC) per user refers to the total of different locations one visits. This metric can be used to understand the different mobility characteristics of users. Its distribution contains the locations visited by users in which some may have

only a few places, while others may have a large collection. That can be used for POI extraction, routing algorithms, and contagion models.

Stay Point Duration (SPD), also called the *duration of stay at a stay point*, refers to the time a user spends at a location. The lower bound is defined by the stay time algorithm's parameter, and the upper bound has no limit. Understanding the time users (or population) spend on average at a location can be a good indicator of its capacity regarding data offloading, helping to design handoff solutions.

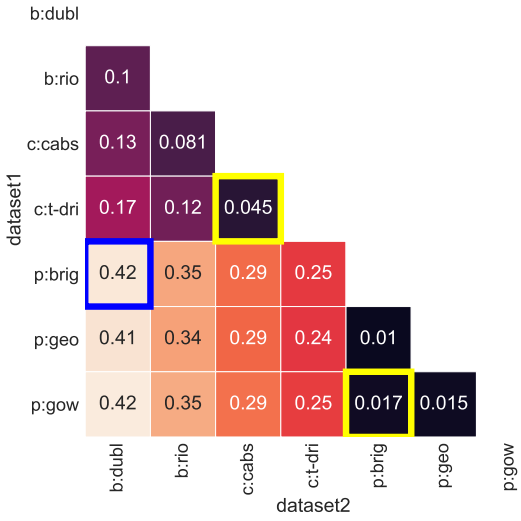
4.4.5 Distributions characterization

The fundamental step for identifying divergences between transport mode datasets is to characterize the distributions (*dist*) obtained from SP metrics (Lines 11–12 of Algorithm 1). That is, identify a representative model of *dist* from a set of candidate models. For this, we calculated the Akaike Information Criterion (AIC) from the Sum of squared estimate of errors (SSE) of each distribution candidate (AIC (SSE))[3]. AIC is an estimator of out-of-sample prediction error, widely used in statistical analysis, that evaluates a collection of models for the data and estimates the quality of each model w.r.t. the other models [237, 230, 238, 244]. The lowest value of AIC will be the best-fit distribution for *dist*. Thus, AIC provides a means for model selection. Specifically, we use an AIC correction (AICc) to address potential overfitting for small sample sizes. SSE is a measure of the discrepancy between the data and an estimation model. It is used as an optimal criterion in parameter selection and model selection. A small SSE indicates a tight fit of the model to the data. AIC works for samples of different sizes, including non-normal distributions.

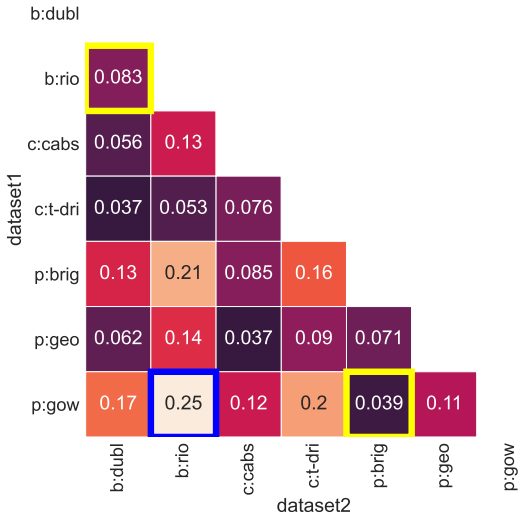
4.4.6 Analysis of Similarities between Distributions

Statistical distance is the approach we use to identify the distance between two probability distributions. We applied the Wasserstein metric (WM), which measures the difference between two distributions by the optimal cost of rearranging one distribution

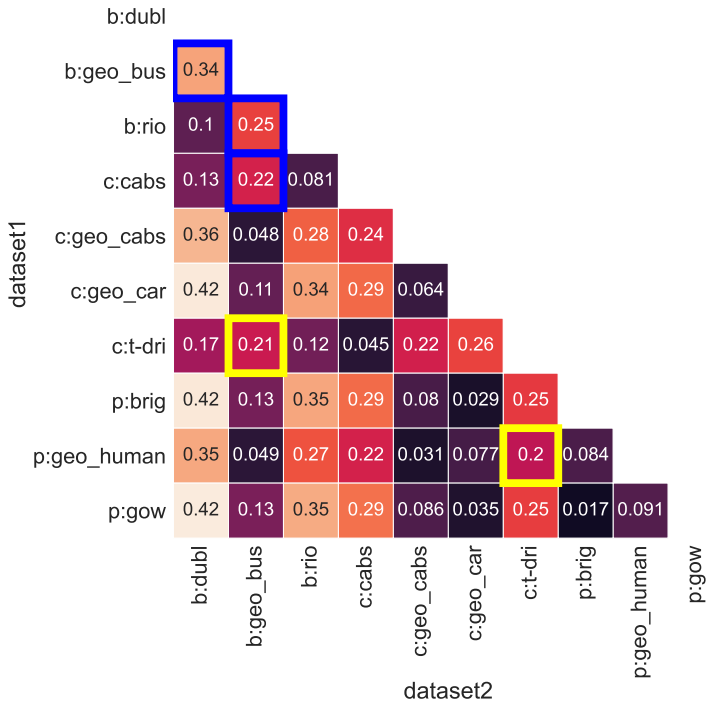
[3]For more details, please refer to K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.



(a) SP count.



(b) SP duration.



(c) SP count for the categories inside Geolife: human, bus, car, and cabs.

Figure 4.7: Wasserstein distance for SPC and SPD metrics. Analysis with people=r250 and vehicles=r500, both t=30 minutes. The metrics are grouped by vehicle category: people (p), car (c), and bus (b).

into the other[4] (Lines 14–18 of Algorithm 1). The smaller the WM value is, the less effort is needed to transform one distribution into another, and consequently, the two distributions show high similarity (Line 19 of Algorithm 1). The Wasserstein distance is asymmetric, (weakly) continuous, and ideal for analyzing corrupted data, in contrast to common distributions divergence approaches, such as Kullback-Leibler or Jensen-Shannon [256]. For the WM, we carry out three types of analysis to identify:

WMA1 similarities of distributions between the two transport generic types: vehicular and human.

WMA2 similarities between distributions of the same category of vehicles. For example, if a taxi distribution is more similar to the distribution of another taxi than a distribution of buses or people.

WMA3 similarities between distributions of the same generic type but extracted from monomodal and multimodal transport datasets. For instance, if the taxicab dataset matches a vehicle category from a multimodal dataset, such as car, cabs, or bus.

4.5 Experimental Results

We perform three analyses of SPs extracted from different transport mode datasets – one analysis of distribution characterization and two analyses concerning the similarity between the transport modes. For the analysis, we used seven datasets with different transport modes (see Table 4.2). We analyzed datasets of human and vehicular mobility and multimodal transport.

In the first similarity analysis, we analyze the trajectories and stay points extracted from different transport datasets and verify their uniqueness. In the second similarity analysis, we used distributions extracted from stay points in datasets of different transport modes. We use SP metrics to extract the distributions, as previously defined. We selected at least two datasets for each type of vehicle to check for possible similarities between them. Also, for Geolife, we analyzed the trajectories of humans, cars, buses, and cabs categories[5] separately. We evaluated the performance of the WM in the task of associating the distributions generated by the stay points metrics using the accuracy metric, i.e., $Accuracy = \frac{distributions\ match\ correctly}{total\ distributions} \times 100$. For instance, the WMA2 analysis with two datasets' distributions for each type of vehicle is distinct: cabs, people, and buses. It

[4]For more details, please refer to C. Villani, "Topics in Optimal Transportation". American Mathematical Soc., 2003, no. 58.

[5]We used all datasets, except the Rio de Janeiro bus dataset, in which we used a sampling of 10 days corresponding to 33% of all dataset

totals six distributions. During the WM matching, there was a more significant similarity between distributions of the same type, such as taxis-taxis and buses-buses, matching a total of four datasets and providing an accuracy of 67%.

4.5.1 Trajectory and Stay Points Uniqueness Analysis

The spatial uniqueness of trajectories shows that human and vehicular datasets have different behavior (see Figure 4.5a). However, human dataset distributions tend to have a slight similarity according to each category—the same concerns vehicular datasets. For instance, the Brightkite and Gowalla datasets, two social networks, have almost the same curve, meaning they have the same number of points needed to re-identify users. Both require two points to re-identify about 97% of users. Geolife also presents similarities with Brightkite and Gowalla datasets in terms of uniqueness. This fact occurs because, although Geolife is a multi-modal dataset, there is a prevalence of its location registers being of human mobility (see Table 4.2).

The uniqueness of vehicle datasets tends to have the most significant discrepancy. However, it is more distant than the human dataset. We can observe that vehicle datasets are more challenging to re-identify than human datasets. In which two points re-identify 90%, 80%, 65%, and 15% for T-Drive, Dublin, Rio de Janeiro, and Cabspotting, respectively. The Rio de Janeiro and Cabspotting datasets are harder to re-identify because vehicle routes are not evenly distributed across a geographic region. Rio de Janeiro's location registers were collected during the Olympic Games, on the route where the games occur. Cabspotting produced location registers predominately on the expressway connecting downtown to San Francisco airport. In this way, vehicles can travel along the same route, with a greater probability that vehicles will generate equal-value location records.

Figure 4.5b represents the Uniqueness of the stay points for each dataset. Users' stay points were extracted for each dataset. Then, we randomly select ten stay points from each user. Finally, we apply over the set of stay points of each dataset to re-identify users. We can see that all the datasets showed a high degree of similarity, including vehicles and people. In other words, in all datasets, users were re-identified from the same number of stay points; that is, only two stay points are enough to re-identify more than 90% of users for all datasets, even for datasets with a high uniqueness of the trajectories, such as Rio de Janeiro and Cabspotting. These facts suggest that stay points can be used as the user's fingerprint and, consequently, a powerful approach to privacy leakage.

4.5.2 Stay Points Distribution Characterization Analysis

For the characterization of the distributions produced by the metric *SPC* for each dataset, we used 89 types of distributions available in the library in Scipy[6]. Table 4.6a shows the results of AIC(SSE), in which the best-fit distribution to the SPC metric for each dataset[7]. For example, for the stay points count of the Cabspotting and Brightkite datasets, the distributions that obtained the best fit were the fisk and invgamma distributions, respectively, among the set of distributions.

Figure 4.6b shows the distribution of the number of stay points by users for each dataset. We can see the similarity between datasets of the same category. For instance, human datasets like Gowalla and Brightkite tend to be similar. Both distributions indicate that many users have only one stay point. The vehicular category, like cabs datasets Cabspotting and T-Drive, also have similar distributions. Still, they tend to a normal distribution, in which we observed that many vehicles show more than one stay point. However, datasets of different categories, such as human and vehicular, tend to present divergences, such as Brightkite and Rio de Janeiro Bus. Differences between stay points distributions can affect LPPMs regarding the number of instances, parameters (*radius*, ϵ, and k), lifetime, and placement. For example, when applying user-level obfuscation with GEO-I to people and using SP as a preliminary step, we will have one GEO-I instance per user, due to the nature of the SP distribution for people, with few places to obfuscate. In contrast, for vehicles with many SPs per user, we will have more than one GEO-I instance per user.

4.5.3 Stay Points Similarity Analysis

Similarity analysis in Figure 4.7 shows the Wasserstein distance of SPC and SPD metrics for the datasets. The WM values in squares are similarity levels between the pairs of datasets analyzed. Low WM near zero (dark-colored squares) represents more similarity between the pairs of datasets distributions than high WM near 0.43 (light-colored squares). The WMA1 analysis for the SPC metric (see Figure 4.7a) reaches 100% match accuracy between distributions of the same mobility group. Distributions of datasets of the same transport mode tend to be similar, while distributions of different transport modes tend to be distant. For example, the dataset pairs (Gowalla, Brightkite)

[6]For more details, visit `https://docs.scipy.org/doc/scipy/reference/stats.html`.
[7]We omitted the *SPD* AICc(SSE) results due to space limitations in this work.

and (Cabspotting, T-Drive) have WM values of 0.017 and 0.045 (see squares highlighted in yellow border in Figure 4.7a). In the meantime, the WM value for (Brightkite, Dublin) is 0.42 (see blue square in the figure). This behavior is also similar to the SPD metric, where it shows 100% accuracy in matching for distributions of people and vehicles (Fig 4.7b). For instance, the WM value of bus distributions (Dublin, Rio de Janeiro) and people distributions (Brightkite, Gowalla) is 0.083 and 0.039 (highlighted with yellow squares). In contrast, the WM value for (Rio de Janeiro, Gowalla) is 0.25 (blue square).

In the WMA2 analysis, for the SPC metric, the distributions of the same transport modes tend to be similar, while the distributions of different transport modes tend to distance themselves. The matching accuracy for distributions of the same transport mode is 83.3%. Considering taxicabs (Cabspotting, T-Drive) distribution, the WM reaches 0.045, while the people (Brightkite, Gowalla) distribution reaches 0.017 (Figure 4.7a). Although Geolife is a multimodal transport dataset, it is close to people's transport mode, with a WM value 0.01 for the pairs (Geolife, Gowalla). However, the SPD metric achieves a matching accuracy of 34% for the same transport mode. Although there is no direct matching for this metric, the WM values for distributions of the same transport mode category are very close, such as (T-Drive, Dublin), (T-Drive, Rio de Janeiro) pairs, in which both belong to vehicular mobility (Figure 4.7b). There are two reasons why Geolife, a multimodal dataset, resembles people's mobility. First, the data was sensed by users carrying cell phones, different from vehicles with fixed sensors. Second, 55% of the dataset records labeled are assigned as people.

Further, we analyzed the association between monomodal datasets and categories of the multimodal dataset. To do so, we extracted four datasets from Geolife that represent the labeled data of people (p:geo human) and vehicles as private cars, taxis, and buses (c:geo car, c:geo cabs, and b:geo bus, respectively), totalizing ten datasets. We extracted their SPs, then the SPC metric to analyze their similarities with the WS distance and verify the accuracy as depicted in Figure 4.7c. In this analysis, for WMA1, the accuracy reaches 70% in classifying the distributions according to the transport modes (vehicular and human). For the WMA2, the accuracy is 50%. The Geolife categories had a more significant similarity, except c:geo car, which had a considerable similarity with the p:brig and p:gow datasets.

Concerning WMA3 is 50% accuracy of association between monomodal and Geolife categories. Additionally, among the four distributions of monomodal transport (Cabspotting, T-Drive, Rio de Janeiro, and Dublin datasets), three had a match with the Geolife vehicle categories, highlighted by blue squares in the figure. The bus distributions, such as Rio de Janeiro and Dublin, were associated with the Geo-Bus category. Although the T-Drive distribution is associated with Geo-Human, we can see that it is close to the vehicle distribution Geo-Bus (highlighted with yellow squares in the figure), with a difference of 0.01 between WM values.

Based on the analysis of different types of datasets with the SPC and SPD metrics, we have the following insight:

> Independently of granularity level, whether in coarse (person and vehicle) or fine granularity, such as type of vehicle (taxi, bus, and person), distributions of the same mobility type tend to converge to each other. However, distributions of different types of mobility tend to diverge from each other.

Indeed, the SPs metrics applied to different transport mode datasets can be considered a fingerprint for each type of mobility, containing their inherent characteristics. Therefore, these fingerprints reflect the context of location privacy approaches, evidencing Hypotheses 1 and 2. This is especially true in smart mobility, in which multimodal trajectories can be joined in a unique trace with an identity from a set of signatures.

4.6 A light at the end of the tunnel

The previous results show strong evidence of the hypothesis that mobility affects location privacy. The classic LPPMs approaches, such as mix-zone and GEO-I, reveal promising perspectives for the future of location privacy in smart mobility. Remarkably, we can model the privacy issues about mix-zone and GEO-I as optimization problems, given smart mobility's dynamic nature. Hence, it encourages designing adaptive LPPMs aware of mobility to preserve its subtleties of privacy and utility. Each LPPM instance must be independent, including a variation of its parameters. For instance, for mix-zones, the radius can vary according to the type of vehicle in it, as well as the LPPM instances' lifetime and positioning, which can also be modeled as optimization problems.

The results suggest strong evidence for the hypotheses that mobility affects location privacy. The classic LPPMs approaches, such as mix-zone and GEO-I, reveal promising perspectives for the future of location privacy in smart mobility. Remarkably, we can model the privacy issues about mix-zone and GEO-I as optimization problems, given smart mobility's dynamic nature. Hence, it encourages designing adaptive LPPMs aware of mobility to preserve its subtleties of privacy and utility. Each LPPM instance must be independent, including a dynamic variation of its parameters. For example, for mix-zones, the radius can vary according to the type of vehicle in it and the LPPM instances' lifetime and positioning. This can also be modeled as optimization problems in which optimization approaches, such as Bayesian optimization, can eventually be applied [257]. Further, the mix-zones privacy level k can be tuned over time, considering various mobility aspects,

such as vehicular traffic fluctuations.

4.7 Concluding Remarks

In this Chapter, we have explored location privacy in smart mobility. We identified gaps in the privacy approaches, anonymization (mix-zones), and obfuscation (GEO-I) in the context of smart mobility. We have evidenced the hypothesis that mobility can affect privacy. For this purpose, we carried out experiments to find similarities in the distributions extracted from two SP metrics, applied to seven datasets of mono and multimodal mobility. Specifically, an accuracy metric was used to quantify the datasets' distributions that matched each other. The results showed that the SPC metric reached 100% and 83.3% accuracies for coarser-grained (person and vehicle) and finer-grained (bus, taxi, and person), respectively. Additionally, we showed a similarity between distributions of the same vehicle type for mono and multimodal datasets. We also showed that vehicles and people have specific patterns about stay points extraction. The results suggest that privacy has a high dependence on mobility at different levels of granularity. In the context of smart mobility, these facts reveal the viability of building location attacks and LPPMs-aware of mobility type, which may bring potential contributions about privacy and utility in both open data and online environments.

Based on the principles presented in this chapter, we demonstrate in the next chapter that mobility behavior can be used to design efficient location privacy approaches, like location privacy attacks. Specifically, we propose a simple but efficient tracking attack approach that explores the mobility characteristics, such as driver's behaviors, to generate attacks against anonymized data protected by classical mix-zones and achieve a high re-identification rate.

Chapter 5

Exploring Mobility Characteristics for a Vehicular Tracking Attack

This chapter analyzes the potentialities of exploring mobility characteristics for generating re-identification attacks, e.g., tracking attacks. Specifically, we show how information about drivers' behavior in a city, such as their road preferences, can be used to re-identify their trajectories, even with their trajectories anonymized by mix-zones. We present a simple and efficient re-identification technique that uses only two geo-referenced points as input data. We validate our technique with a real dataset of taxicabs, being able to re-identify up to 100% of the anonymized trajectories.

The organization of this Chapter is as follows: Section 5.2 presents a literature review of re-identification attacks for vehicular networks. Section 5.3 defines the terminology, notations, and problem description used in this work. Section 5.4 details the privacy attack that uses the minimal path to re-identify trajectories. The experiments and discussion about re-identification attack is defined in Section 5.5. Finally, in Section 5.6, we present the final remarks of this chapter.

5.1 Introduction

In the current ubiquitous computing age, a huge amount of mobility data produced by mobile entities such as smartphones instigates the interest of companies and researchers. This data can be used to understand human behavior and to develop various services in the area of traffic engineering, such as vehicle congestion monitoring, flow control, infrastructure planning, and others [121, 54].

However, compiling these trajectory data brings serious risks to users' privacy. Malicious agents can exploit information found in trajectories, even when submitted to Location Privacy Protection Mechanisms (LPPMs), to generate attacks on people and vehicles in various ways [258]. For example, to identify points of interest such as residences

and workplaces, predict the absence or presence of a user in a specific location [33], track and locate mobile entities [121], and associate an identity for each anonymized trajectory (trajectory re-identification attack) [40, 31].

Re-identification is a kind of privacy attack that identifies anonymized trajectories and, consequently, their source's identity from limited information about the victim [32]. This is an attack that can also be the "gateway" to others that target a specific entity. The information includes geo-localized points, trajectory segments obtained from historical records in LBSs servers and intrinsic characteristics of anonymous trajectories. This information can represent fingerprints and be used to infer a vehicle's path and, hence, the users' location.

In the literature, several re-identification approaches are based on the characterization of mobility traces using techniques such as machine learning [28, 85], and statistical inference [121]. However, most require high computational costs and training datasets, which may not always be available.

In this work, we propose a simple yet efficient re-identification approach in the order of $O(E + V\log V)$, where E represents the number of edges (intersections) in the graph of the roads of the city, and V the number of nodes (streets). We present the re-identification of trajectories based on characterizing the road preferences that occur in urban environments, particularly in taxi cabs. Assuming the premise that vehicles, especially taxis, tend to follow the shortest path between two points, the idea is to construct the taxi minimum path from two geo-localized points (beginning and end of the trajectory) and compare it with anonymized trajectories as a way of re-identifying vehicles. The approach eliminates the need for historical data collection or training sets for the re-identification calculation.

We verified the effectiveness of our attack model against a privacy scheme called mix-zones. Mix-zones are urban regions that promote the simultaneous anonymization of vehicles by changing their pseudonyms. In our experiment, we modified the level of anonymization of the mix-zones to validate the ability of our technique to re-identify vehicle trajectories. We validate our re-identification technique in traces of taxis in the city of San Francisco, USA, containing about 231,230 trips in which we re-identify up to 100% of the anonymized trajectories.

5.2 Related Work

In the literature, there are several studies about the re-identification of vehicles and users in which they use features extracted from trajectories.

Sui et al. [85] proposed the parking spots attack, in which the adversary considers the parking habits of taxi drivers extracted from the taxi mobility traces to re-identify victims. As a countermeasure, they presented a protection scheme that exchanged sub-trajectories of the most relevant parking-point taxis. One of the restrictions of this attack is that the attacker needs prior knowledge of the opponents and habits of the taxi drivers to infer the parking spots.

Zan et al. [28] developed a re-identification attack that used mobility trace classification to identify vehicles anonymized by mix-zones. In this attack, the anonymized traces were classified by vehicle type: car, motorcycle, or truck. The authors claimed that vehicles of different types have distinct acceleration/deceleration profiles and produce different mobility trace profiles, which work like fingerprints, making it possible to classify them and rebuild the traces. They used machine learning classification, and used a simulator to produce the experiment and freeways instead of an urban environment, which is a less challenging environment to perform the re-identification.

De Montjoye et al. [63] proposed one of the first studies on quantification of the uniqueness in human mobility traces. Based on the change in the granularity of the dataset's spatiotemporal data, the authors presented a formula for calculating the uniqueness of human mobility. They also proposed an inference model in which four spatiotemporal points are sufficient to identify up to 95% of individuals from a cellphone dataset containing human mobility data of 1.5M users with 15 months of data.

Rossi et al. [120] presented a user re-identification technique that exploits the uniqueness of GPS data, even if not present in the mobility dataset. Specifically, the technique calculates the minimum distance between a set of geo-localized points of the victim and the points of the anonymized traces. The victim's trace is the one with minimal distance between the points. Three real-world datasets were applied: Cabspotting [126], CenceMe [129], and Geolife [125]. The authors concluded that two spatial points were needed to identify nearly 100% of users. However, they did not provide details on which LPPM was applied.

Tan et al. [121] also investigated the uniqueness of GPS data. The heuristic proposed can re-identify anonymized trajectories with up to 95% accuracy from four spatiotemporal. However, the attacker needed full access to the data server.

Sekara et al. [61] showed that it is possible to capture users' behavior in data from smartphone applications, taking into account the time. Usage behavior was used as a fingerprint to re-identify users. The data was collected from the Google Play Store, which had about 12 months of data from 3.5 million users. The technique re-identified 91.2% of users, using the strategy inspired by [63].

Chang et al. [54] claimed that trajectories have user profile indicators, such as preferences and usual behaviors, which are unique and with little change over time, aiding in re-identifying vehicles. The indicators found the stops of interest (e.g., malls and gas

stations), and road segment preferences. They re-identified the victim's trajectories by comparing the indicators found in trajectory segments collected by the victim's observation and those of anonymized mobility histories. In that study, they used the dataset of taxicabs from the cities of Shanghai and Shenzhen. The results showed that it was possible to re-identify the anonymous trajectories with subpaths with 8 to 9 days in size with 96.64% and 77.03% of hit for Shanghai and Shenzhen, respectively.

Kaplan et al. [94] have presented an attack for discovering if an unknown private trajectory passes (or does not pass) through regions of interest. The regions of interest could be areas used for a privacy attack, e.g., the adversary could learn if the victim visited a hospital. The attack uses a set of known trajectories and their pairwise distances to the unknown trajectory. It generates a set of candidate trajectories that resemble the private trajectory. After this, it explores properties of the candidate trajectories, like regions of interest that could be in the private trajectory. However, the adversary needs many public trajectories as input to identify some similarities with private trajectories.

The studies above extract characteristics of the victims' mobility traces to carry out re-identification attacks. Some approaches use machine learning, which requires training data and high computational costs. Other approaches require additional information about the victim and their context to succeed in the attack. The proposals about singularity inference, although presenting expressive results such as the studies of [120, 61] are sensitive to spatiotemporal resolution that directly compromise the re-identification rate. This fact is proven in [63], whose re-identification rate degrades up to 50% with the spatiotemporal resolution change. In addition, certain approaches have used synthetic data produced by simulators and highway datasets, which poorly reflect the real traces of vehicles in urban environments.

Different from the previous approaches, our proposal re-identifies vehicles with the minimum of known information, being necessary basically two geo-located points and with a computational cost of $O(E + V\ logV)$. This work differs from [120], which calculates the minimum distance between a set of points of the victim, owned by the attacker, and anonymized mobility traces to generate the attack. Our proposal uses two geo-localized points to construct a minimum path and obtain a correlation with some of the anonymized trajectories. The proposal of Kaplan et al. [94] is slightly different from our work because it aims to identify regions of interest that the private trajectory has passed. However, the accuracy could be degraded if the private trajectory is submitted by mix-zone privacy schema. More details on the approach will be discussed in Section 5.5.

5.3 Background and Problem Description

This section defines the terminology and notations used in this work. We describe the general structure of the mix-zone privacy scheme and the trajectory re-identification attack. The dataset of Cabspotting [126] is used, in which the trajectories were anonymized, and the taxis' identity was replaced by a pseudonym: a unique and random identifier. Additionally, the trajectories were anonymized through the mix-zones technique.

5.3.1 Mix-zones: A Privacy Protection Scheme

A mix-zone is a geographical area of k-anonymity that vehicles go through, causing their aliases to be modified [171]. When a vehicle is at a mix-zone, its trajectory will have at least two segments (or sub-trajectories) delimited by different pseudonyms. Vehicles change their nicknames in the mix-zone if and only if at least k vehicles are present at the same time [164, 172]. For more details on mix-zones, see Section 2.5.3.

Next, we formalize the definition of mix-zones and the pseudonyms change process:

- W is an urban geographic map which has an associated global time $W.t$. All objects in W observe $W.t$.
- A set of vehicles $S = (v_1, v_2, v_3, ..., v_i, ..., v_n)$, $1 \leq i \leq n$ and $|S| = n$. Also, $S \subset W$.
- A vehicle v_i has the following attributes: a pseudonym $v_i.alias$, which is distinct from the other $n-1$ vehicles. All v_i contains a path T.
- A trajectory T of the vehicle v_i is formed by a temporal sequence of spatiotemporal points $T = (p_1, p_2, p_3, ..., p_i, ..., p_m)$, $1 \leq i \leq m$. A geo-location point $p_i = (x, y, t)$ where (x, y) represents latitude and longitude at time t. A set of trajectories of different vehicles is denoted by $T_a = (v_1.T, v_2.T, \ldots, v_i.T, \ldots, v_n.T)$, $1 \leq i \leq n$ where $T_a \subset S$. We consider that each vehicle v_i has a unique trajectory T.
- An anonymized trajectory T' is a path T in which it was processed by some anonymization function $T' = anon[T]$ by passing through a mix-zone $M_j \in Mx$ and suffered the change of pseudonym in some $W.t$. A set of anonymized trajectories is denoted by $T_p = (v_1.T', v_2.T', \ldots, v_j.T', \ldots, v_n.T')$, $1 \leq j \leq n$ where $T_p \subset S$.

- A mix-zone M_i is a geographic area that has dimensions $M_i.r$ and requires the minimum level of anonymity K where $K > 1$. The time duration of v_i in M_i can be any. A set of mix-zones $Mx = (M_1, M_2, \ldots, M_p)$, where $Mx \subset W$.
- The change of the pseudonym of T_a vehicles in a mix-zone M_i is a function $M_i(T_a)$, which occurs at a current time $W.t$ if and only if $\forall v_i$, $\exists$ some $T.p_j \in M_i.r$ and $|T_a| \geq K$.
- Consider an open dataset of anonymized trajectory $\mathcal{D}'$ composed by anonymized trajectories T_p eventually published in a public repository.

5.3.2 Adversary Model

Our tracking attack design is based on Local Passive Adversary (LPAd) that applies the Semantic Linking Attack (SeLA), particularly the Multi-target tracking (MTT) for re-identifying users' trajectory, defined as follows.

The Local Passive Adversary (LPAd) strategically deploys low-cost receivers over local regions of the road network near some mix-zones, for instance, a mix-zone M_i, and eavesdrops on exchanged messages, collecting sub-trajectories composed by GPS points of each vehicle v_j before and after in each M_i. It also assumes that the Local Passive Adversary (LPAd) collets $\mathcal{D}'$ in a repository. The GPS points and $\mathcal{D}'$ information represent the adversary's background knowledge about the users $\mathcal{B} = (B_1, \ldots, B_b, \ldots, B_m)$, where $[1 \leq b \leq m]$ and b represents the number of elements in $\mathcal{B}$ known by the adversary, enabling the adversary to execute a Tracking Attack $\mathcal{Z}$, and consequently perform the Multi-target tracking (MTT) for multi-users. In a Tracking Attack (TA), the adversary aims to determine the whole sequence (or a partial sub-sequence) of events in a user's trace. Given an anonymized dataset $\mathcal{D}'$ composed of users and background B_u, a tracking attack is defined as $T_u \leftarrow \mathcal{Z}(\mathcal{D}', B_u)$, where T_u represents the re-constructed trajectory of user u. Next, we detail our attack proposal.

5.3.3 The Privacy Attack Model

In urban mobility, time is a valuable factor reflected in mobility dynamics. When moving around in urban environments, people typically intend to avoid long roads, look

for shorter roads, and consider the traffic conditions to avoid congested routes, roads, or accidents. Domingues et al. [259] presented a mobility characterization of the taxis' trace of San Francisco, showing that about 70% of all trips tend to follow the minimum path (i.e., the shortest path) between two points. The minimum path is defined as the shortest distance between two points, considering the road's infrastructure as the base. This definition will remain the same throughout this work. In addition, trips that do not follow the minimum path tend to make short detours in the same way, with about a 5% increase to the minimum distance.

Based on this principle, we propose the following hypotheses to construct the adversary model:

Hypothesis 1: *Most vehicles, especially taxis, choose minimal paths to complete their routes.*

If we consider the Hypothesis 1 true, which is quite reasonable, then we have:

Hypothesis 2: *It is possible to re-identify anonymous trajectories.*

The adversary's goal is to re-identify as many anonymous trajectories as possible. Thus, we consider that the adversary, defined in Section 5.3.2, has access to the following information (B_u):

- trajectories anonymized by the mix-zone, which were published on a public access server;
- two geo-located points *start* and $finish$, which correspond to points near the beginning and end of a trajectory, respectively ($start \approx p_1$ and $finish \approx p_m \in T$), of the victim v_i.

The general idea of the re-identification algorithm is to construct the minimum path of v_i ($\min[v_i.T]$) generated from the points *start* and $finish$ provided from each path, one before and one after the mix-zone. The path is calculated by applying a minimum path algorithm in graphs (e.g., Dijkstra algorithm) in a graph composed of city roads represented by vertices and their intersections represented by the edges. The lengths of the roads are represented by the weights of the edges. After constructing the minimum path, the next step is to compare it to the trajectory of the individual in an attempt to find a correlation between them. The opponent will succeed if he/she finds some path of greater correspondence than the minimum path. Further algorithm details are presented in Section 5.4.

5.4 Vehicles Re-identification from the Minimum Path

This section details the privacy attack that uses the minimal path to re-identify trajectories.

5.4.1 Re-identification Algorithm

Algorithm 2 presents the steps to re-identify trajectories through the minimum path. Its inputs are the geo-localized points of the trajectories reported before the passage of the vehicles by a mix-zone M_i (T_a), and the geo-localized points of the trajectories reported after the passage of the vehicles by the same mix-zone (T_p). Additionally, it also receives the graph (G) containing the roads and intersections of the city. As a result, the algorithm returns a bijective mapping Φ of the trajectories in T_a on the trajectories in T_p.

Algorithm 2: Re-identification of trajectories

Data: Subtrajectories T_a before the mix-zone,
Subtrajectories T_p after the mix-zone,
Graph G of roads and intersections
Result: Mapping $\Phi : T_a \rightarrow T_p$

```
1  costs ← Matrix (T_a rows, T_p columns);
2  for Trajectory i ∈ T_a do
3      for Trajectory j ∈ T_p do
4          start ← arg min_p f(i.p) | f(i.p) = i.p.t;
5          finish ← arg max_p f(j.p) | f(j.p) = j.p.t;
6          minimal-path ← Dijkstra(G, start, finish);
7          trajectory-choosen ← ⟨i, j⟩;
8          error ← DTW(trajectory-chosen, minimal-path);
9          costs[i, j] ← error;
10     end
11 end
12 Φ ← minimize(costs);
13 return Φ;
```

The algorithm works as follows: Line 1 defines a cost matrix, which stores the results of each possible match of the trajectories from T_a to T_p. For each trajectory in T_a (Line 2), it is iterated on each of the possible trajectories in T_p (Line 3), extracting the start and end points of the trajectory represented by *start* and $finish$ (Lines 4 and 5). Then, the minimum path is calculated from these points (Line 6). Line 7 defines the

candidate trajectory as the junction of the before trajectory i to the mix-zone and the trajectory j after the mix-zone. That is, it is assumed that the victim of the trajectory i is also the same as the one from the anonymous trajectory j.

Finally, the correlation between the candidate trajectory and the minimum path is calculated through the Dynamic Time Warping algorithm (DTW) [260], which calculates the optimal non-linear correlation of two-time series. In our implementation, the DTW returns the correlation level represented by an error. If the error is small, it means that there is a high correlation between the two trajectories, otherwise there is not. The error is stored in the cost matrix (Lines 7 through 9). The mapping of the trajectories T_a and T_p is calculated by solving the problem of minimization of costs from the cost matrix:

$$\Phi : \min\{costs[i, j] : i \rightarrow j, i \in T_a, j \in T_p\}. \tag{5.1}$$

Assuming that vehicles tend to follow the minimum path between two points, we expect that the smaller the distance between the candidate trajectory and the minimum path, the more likely that the path is a minimum path and, consequently, the more likely that the driver followed it. In other words, candidate trajectories with a large error do not represent a minimum path between their points of origin and destination and, therefore, have a small probability of being the real trajectory chosen.

5.4.2 Complexity Analysis of the Re-identification Algorithm

The computational complexity of Algorithm 2 is

$$\mathcal{O}(C(E + VlogV) + D^3) \approx \mathcal{O}(E + VlogV) \tag{5.2}$$

where $C = |T_a||T_p|MN$, $|\cdot|$ represents the cardinality of the set of trajectories that are in a mix-zone, M and N represents the size of the largest trajectory in T_a and T_p, respectively, D represents the dimension of the cost matrix, $|T_a| = |T_p| = D$.

This complexity is derived from the Dijkstra minimum path algorithm (Line 6), which the proposed algorithm uses to construct the route to be compared to the anonymized trajectories. The DTW algorithm (Line 8) has complexity $\mathcal{O}(MN)$, and the minimize method (Line 12) can be solved by attribution linear, with complexity $\mathcal{O}(D^3)$. However, since $|T_a|, |T_p|, D, M, N \ll E \approx V$, the complexity of Algorithm 2 can be resolved by approximation in Equation 5.2.

Table 5.1: Mix-zones Setup

Mix-zone	Latitude	Longitude	Radius (m)
mixzone0	37,614350	−122,395635	500
mixzone1	37,633000	−122,419134	300
mixzone2	37,628569	−122,432339	300
mixzone3	37,768201	−122,406079	500
mixzone4	37,769199	−122,453495	500
mixzone6	37,635672	−122,403605	500
mixzone7	37,735237	−122,406974	500
mixzone8	37,768914	−122,406881	500

5.5 Experiments

The main goal of our experiments is to evaluate the efficiency of the re-identification algorithm, considering different privacy levels (K).

5.5.1 Experiment Setup

In this study, we used a dataset containing mobility traces of taxicabs of the city of San Francisco, USA, with data from 500 unique taxis collected over a period of 30 days [126]. However, we considered the trajectories of the taxis in which they actually went through some previously defined mix-zone. Thus, the duration of the experiments was about 25 days (May 17 – June 10, 2008) and had a granularity of 10 secs.

As mentioned above, the privacy mechanism used in this work is the mix-zones approach. Eight mix-zones were strategically positioned in the intersections with a great traffic flow of taxis. The goal was to position the mix-zones to reach as many taxicabs as possible during the 25-day experiment. For this, we conducted a previous study on the flow of taxicabs in the city. Table 5.1 represents the information on the mix-zones. Some mix-zones have distinct radii to avoid possible overlaps.

The experiment is split into two phases. The first is the anonymization of trajectories. The second phase is the re-identification attack of those by the proposed algorithm.

In the anonymization phase, there were 231,230 trajectories that crossed the analyzed mix-zones: 104,449 trips were anonymized, and 126,781 were not (due to the privacy level, see Table 5.2). In each experiment, the mix-zones' privacy level (K-*anonimity*) was changed. At each level K, trajectories that went through one or more mix-zones and met their requirements were anonymized.

Table 5.2: Results of the anonymization process through the mix-zones technique

K	Non-Anonymized	Anonymized	Total	Rate
2	2981	30364	33345	0.91
6	11048	22139	33187	0.67
10	15696	17404	33100	0.53
14	19622	13391	33013	0.41
18	22868	10068	32936	0.31
22	25971	6893	32864	0.21
26	28595	4190	32785	0.13

We can observe that the coverage of anonymized vehicles is inversely proportional to the value of K. This is because the anonymity of n vehicles occurs only if there is simultaneously $n \geq K$ inside a mix-zone. Thus, if there is a small K, more vehicles will change their pseudonym, but the chance of re-identifying a vehicle will be greater. We also considered that the trajectories of the vehicles that entered but did not leave the mix-zone and did not reach the minimum value of K were not anonymized. Table 5.2 presents the anonymization rate for the different levels of privacy K, in which $K = 2, 6, 10, 14, 18, 22,$ and 26 were obtained, for all mix-zones 91%, 67%, 53%, 41%, 31%, 21% and 13% of anonymized trajectories, respectively.

In the re-identification phase of the trajectories, we compared the proposed strategy (Algorithm 2) with the random re-identification approach, which served as a baseline in experiments [28]. In this algorithm, the candidate trajectory is formed by joining a sub-trajectory of T_a and T_p that are randomly selected.

5.5.2 Efficiency Validation

We use a metric, called Trajectory Matching Accuracy (TMA) to measure the effectiveness of re-identification attacks [54], which is defined as:

$$\text{TMA} = \frac{N_{\text{reid}}}{|T_p|}, \tag{5.3}$$

where $N_{\text{reid}} \in [0, |T_p|]$ and $|T_p|$ represent the total of re-identified trajectories and the total of anonymized trajectories that have passed in any of the mix-zones, respectively.

5.5.3 Results and Discussion

The results of the re-identification of trajectories for each mix-zone and each level of K are represented in Figures 5.1(a)–(g). In a panoramic view, the highest TMA reached, with 100% of hit, was to the mix-zone settings where the K had significantly higher values. That is, for values of $K > 2$. The discrepancy with the results of the random re-identification algorithm (represented by dashed lines in the graphs) is evident.

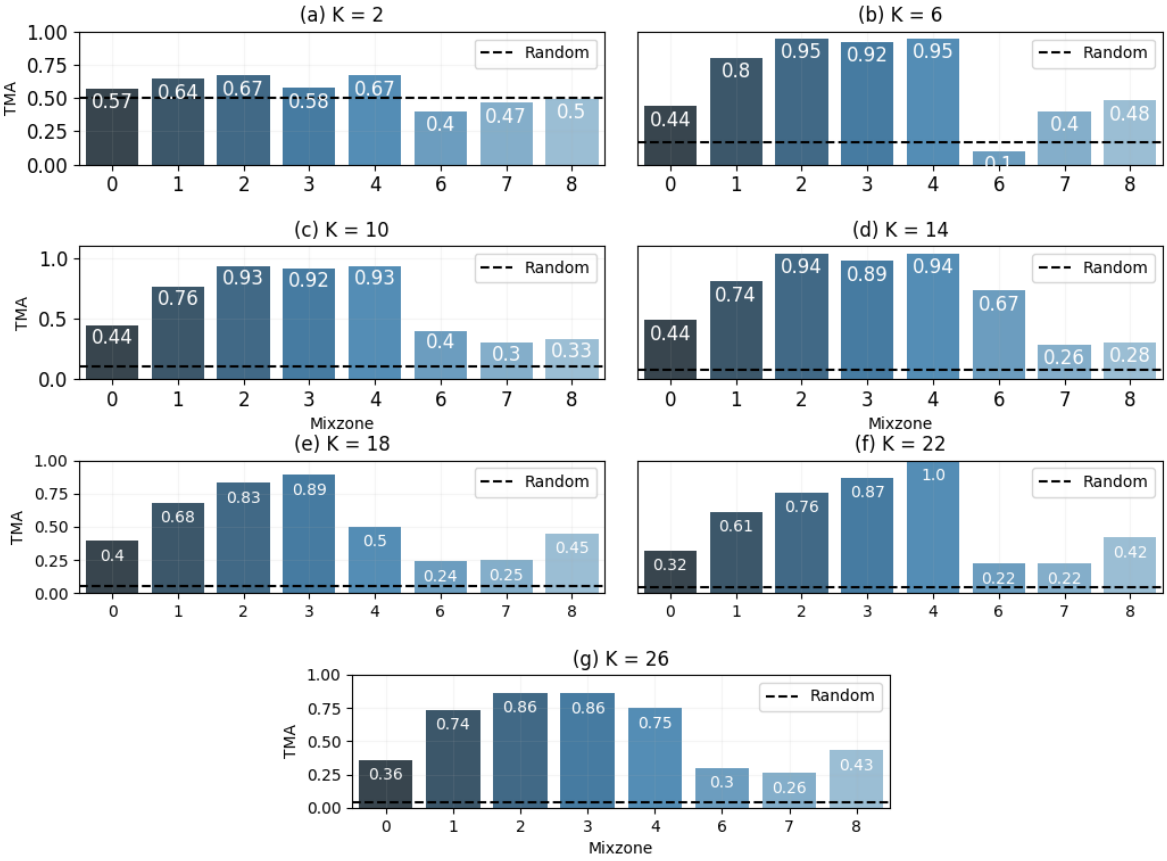


Figure 5.1: The distribution of the TMA for the different levels of anonymization for each of the defined mix-zones

This reality is different for mix-zones 0, 6, 7, and 8, located closer to the start and end points of the trajectories and produce the anonymization of the initial and final data. Although they have a low re-identification value, these mix-zones are not good solutions for a privacy mechanism since the anonymization that occurs very close to the beginning and end places of the trajectory allows the adversary to infer that they are the true starting and ending points of the trajectory [171, 115].

In Figure 5.1(a), only two of the eight mix-zones obtained lower re-identification rates than the random algorithm. Mix-zones 2 and 4 obtained the highest values for the others, with 67% re-identification each. Compared to the other K levels of anonymization,

the $K = 2$ level has the lowest rates due to the minimum number of vehicles present in the mix-zone, reducing the algorithm's search space. For $K = 6$ (Figure 5.1 (b)), it is observed that the increase in the number of vehicles in the mix-zone provides better results by the algorithm. In this case, only one of eight mix-zones presented results below those produced by the random algorithm. The algorithm produced good results for the rest, especially the mix-zones 2, 3, and 4, with re-identification rates above 90% of the trips. This same behavior can be observed for $K = 10$ and $K = 14$ (Figure 5.1(c) and 5.1(d), respectively), where the algorithm presented rates above 90% for the mix-zones located at central points of the trajectories, and even for those located at the beginning or end of the trajectories, it was able to re-identify with considerable precision when compared to the random algorithm. For $k = 18$ (Figure 5.1(e)), the same patterns observed before can be seen, with mix-zones 1 to 4 reaching the maximum values. This trend goes on in both cases of $K = 22$ and $K = 26$ (Figures 5.1(f) and 5.1(g), respectively), which show how the positioning of the mix-zones definitely affects the final results. We can also highlight the results on mix-zone 4 for $K = 22$, in which the algorithm obtained 100% of correct results.

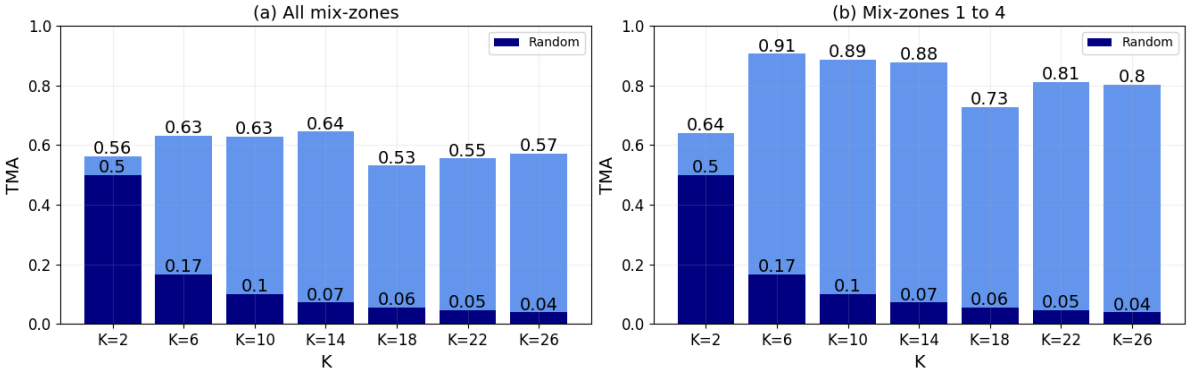


Figure 5.2: Aggregate TMA for each level of privacy K

In Figure 5.2, the aggregate TMA is represented, considering the simple average among all mix-zones (Figure 5.2(a)) and between mix-zones 1, 2, 3, and 4 (Figure 5.2(b)), for each privacy levels K defined. In addition, the average TMA obtained for the random algorithm is also shown for comparison. It is possible to observe how the increase in privacy drastically reduces the re-identification of the random algorithm, making its use unfeasible for any value of K greater than 2. Considering all the mix-zones (Figure 5.2(a)), the algorithm presents similar values for the different levels of anonymization.

Although it is considerably more accurate than the random algorithm, the low precision, when re-identifying anonymized trajectories by mix-zones located in points at the beginning or end of the trajectories, causes a decrease in the aggregate TMA. Because of this, Figure 5.2(b) presents the aggregate TMA considering only mix-zones 1, 2, 3 and 4, whose location tends to be at the center of the trajectories. To prove this fact, we performed a second experiment.

Table 5.3: Top 5 mix-zones Setup

Mix-zone	Latitude	Longitude	Radius (m)
mixzone top 1	37,7156331	−122,3987989	500
mixzone top 2	37,7247622	−122,4012375	500
mixzone top 3	37,7331701	−122,404774	500
mixzone top 4	37,6754004	−122,3891533	500
mixzone top 5	37,7098094	−122,3943708	200

The idea of the second experiment was to deploy the mix-zones in regions that generate the anonymization of the traces in the middle of the trajectories. A sample (S) was generated with 20% of the data from the dataset in [126], selected randomly. The mean segments (S_{avg}) from each trace of S were collected. The middle segment is the set of geo-located points in each trajectory's middle. From S_{avg}, we calculated the traffic flow for all intersections in San Francisco with the frequency of vehicles that traveled through them. The intersections were sorted by frequency in descending order. The top 5 most frequent intersections of the ranking were selected as mix-zones (Table 5.3). The radius was selected empirically, aiming to maximize the number of trips anonymized in each mix-zone. Similarly, the privacy levels K were selected with the aim of maximizing the number of trips to be anonymized in each mix-zone. The lower privacy levels, compared to the previous analysis, can be explained by the positioning of such mix-zones, which are over a high-speed highway, making it harder to detect a higher number of vehicles simultaneously.

Figure 5.3 shows the aggregated TMA for the top 5 mix-zones, for privacy levels ranging from $K = 2$ to $K = 5$. Additionally, for comparison reasons, we also present the TMA for the random algorithm. Again, it is possible to see how the random solution is unpractical for privacy levels greater than 2. Different from what was seen in Figure 5.2, for the top-5 mix-zones, the aggregated TMA for $K = 2$ is lower than the one obtained through the random algorithm. As stated before, the reduced search space makes it harder for the proposed algorithm to find the correct solution. However, when we look at the other privacy levels $K = 3, 4, 5$, it is evident that the algorithm can re-identify the drivers with high precision, reaching up to 87% for $K = 5$.

It is possible to observe that the algorithm is efficient in re-identifying trajectories, with values above 90% at one of the anonymization levels. These results confirm the Hypotheses I and II raised in Subsection 5.3.3.

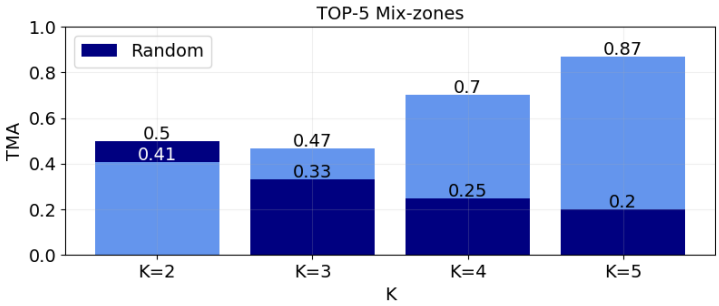


Figure 5.3: Aggregated TMA for privacy levels K2 to K5 considering the Top-5 mix-zones

5.6 Concluding Remarks

In this chapter, we proposed and evaluated a trajectory re-identification approach based on the characterization of vehicle path preferences. To do so, we assumed that most vehicles, especially taxis, choose minimum paths to complete their routes, making it possible to re-identify up to 100% of the anonymized trajectories through our algorithm. To validate this approach, we applied it to a dataset of mobility traces of San Francisco, USA, and compared the results to a baseline solution. We were able to advance the state of the art in the following directions: a new re-identification approach that uses the characterization of the minimum path's preferences, with the computational cost of the order of $O(E + V\ logV)$; Additionally, we use only two location points of the victim as input to generate re-identification and validate our proposed approach through real and wide-scale mobility.

Mix-zones present many factors that maximize the success of tracking attacks, such as the one we propose in this chapter. Factors such as high dependency on positioning, geometry, mobility patterns, vehicle density, and arrival rates can degrade the anonymization rate. Therefore, it is necessary to explore the internal functioning of mix-zones to understand how anonymization happens, including measuring the quality of the anonymized data. In the next chapter, we propose the AQM, a framework for the analysis of anonymization quality in mix-zones over space and time for mobility data capable of electing mix-zones, periods of good anonymization, and mix-zones deployment positioning algorithms.

Chapter 6

Anonymization Quality in Mix-zones

This chapter presents the Anonymization Quality Framework for Mix-zones (AQM). A framework that enables characterizing and evaluating the impacts of anonymization over time and space in mobility data. We conducted experiments with a cab mobility dataset and two positioning algorithms to explore one of the potentialities of the anonymization quality: elect mix-zones that do not consider the traffic but its operating requirements too. The anonymization quality enables the selection of mix-zones that yield data anonymization considering the quality, privacy, and utility analysis.

This chapter is organized as follows. Section 6.1 brings an introduction and motivation discussing factors that affect the mix-zones performance and presents the anonymization quality definition. Section 6.2 discusses the related work. Section 6.3 presents essential concepts and the problem of this work. Section 6.4 introduces the anonymization quality framework and details its metrics, quality function, privacy attack, and utility analysis. Section 6.5 shows and analyzes the results and examines the lessons learned from this work. Finally, Section 6.6 presents chapter remarks.

6.1 Introduction

In the era of ubiquitous computing, technologies like the Internet of Things (IoT) and IoV have contributed to connecting objects and sharing location services in broad environments such as smart cities, bringing many benefits to citizens [231]. However, these services yield massive and unrestricted location data of citizens, which poses privacy concerns. By mining location data, it is possible to identify latent information about places of interest, individual and collective habits, and even the identity of citizens [261]. To address the privacy of users' identity, LPPMs based on anonymization have been proposed, such as mix-zones. In this mechanism, mix-zones are anonymization areas defined by a radius r where entities change their pseudonyms according to a trigger function (e.g., when reaching a minimum of k entities simultaneously inside it) [170, 167].

Although mix-zones are LPPMs widely used in VANETs, they present drawbacks. For instance, mix-zones depend highly on positioning, geometry, mobility patterns, vehicle density, and arrival rates [261, 262], which can degrade the anonymization rate and enable linking and inference attacks. Some proposals address these issues, but only a few have investigated the anonymization behavior in the spatiotemporal context, particularly on anonymization quality.

Anonymization Quality (AQ) is a concept related to the protection, efficacy, and internal functioning of an LPPM, enabling an understanding of how privacy occurs and analyzing its performance over time. Mix-zones with high anonymization quality are expected to present high vehicle flow, adequate privacy, and high efficacy at anonymizing mobility data. These factors make it possible to identify specific behaviors of mix-zones that cannot be identified by analyzing conventional traffic metrics, allowing the creation of sophisticated applications for smart cities' urban planning, such as management, selection, and election of mix-zones. In this context, questions naturally arise, such as:

- **Q1** - Are positioning and setting up mix-zones based on the average vehicle traffic flow sufficient to achieve anonymization quality, privacy, and utility?
- **Q2** - Do mix-zones deployed in the same region have the same performance?
- **Q3** - How to measure mix-zone performance over time and/or space regarding anonymization quality?
- **Q4** - What factors can lead to anonymization quality?
- **Q5** - What is the best positioning algorithm that deploys mix-zones to obtain the best trade-off between AQ, privacy, and utility?
- **Q6** - What are the anonymization quality potentialities in application terms?

This chapter proposes the Anonymization Quality Framework for Mix-zones (AQM). The AQM is capable of getting valuable information to understand the anonymization behavior, not just in terms of traffic flow but also regarding quality metrics that compose it, such as mix-zone activation, the interval of arrival and departure of vehicles, vehicles that finish their trips inside the mix-zones, and the number of vehicles inside mix-zones along the day. Thus, the AQM can provide valuable insights into the design of robust anonymization proposals. For example, it can help elect more effective mix-zones from a set given a privacy budget, identify day periods, and select the best set of mix-zones according to an anonymization rate variation. It is also helpful to decide how to position mix-zones, considering other parameters besides traffic flow and tuning their parameters. We conducted a comprehensive framework evaluation applied to mobility datasets of real taxicabs and two positioning algorithms to deploy mix-zones. The results showed that the anonymization quality framework enables the selection of mix-zones that yield data anonymization in terms of quality, privacy, and utility. To the best of our knowledge,

this is the first study that analyzes mix-zone coverage and quality metrics to observe the anonymization quality.

6.2 Related Studies

The analysis of anonymization properties can reveal various insights for developing more robust LPPMs against linkage attacks. Thus, a broad study about anonymization approaches – including mix-zones – exists. In this context, there are studies to improve the anonymization rate by addressing issues on which mix-zones are highly dependent, like positioning, geometry, mobility patterns, vehicle density, and arrival rates [261, 167, 262]. Following, we highlight some relevant literature proposals that pursue these issues.

Vehicular mix-zones differ from classical mix-zones because they are constrained by many spatial and temporal factors, such as trajectories strictly on physical roads, heading directions, traffic regulations (e.g., speed limits), traffic conditions, and road conditions [167, 200]. The first studies considering spatiotemporal factors were introduced by [170], who present a protocol for vehicular networks called CMIX based on mix-zone encryption for road intersections that ensures changing pseudonyms.

Table 6.1: Mix-zones' issues and contributions of each proposal.

Author	Operation	Mix-zones Issue	Contribution	Evaluation	Anon. Aspects
[170]	online	Comm. security	Encryption protocol	simulated env.	Privacy
[190]	online	Anonymization rate	Non-rectangular mix-zones based on speed and entropy; mix-zones placement based on road network topology, user mobility patterns.	simulated env.	Privacy
[263]	online	Low traffic in mix-zones	Dummy vehicles scheme when mix-zone traffic is low.	simulated env.	Privacy
[262]	online	Low traffic in mix-zones	Dummy vehicles emulated by real vehicles when mix-zone traffic is low.	simulated env.	Privacy
[100]	online	Anonymization rate	Double trajectory protection with swap pseudonyms and add noise.	simulated env.	Privacy
[208]	offline	Anonymization rate	Placement heuristic using simulated annealing and genetic algorithms.	simulated env.	Privacy/Cost
[175]	offline	Anonymization rate	Placement metric-driven based on graph topology and road flows.	real dataset	Privacy
[261]	online/offline	Impacts of mobility on privacy	Methodology to verify similarities between different modals in the smart mobility.	real dataset	Privacy/Utility
our work	online/offline	Anonymization behavior	Framework enabling understanding the anonymization behavior.	real dataset	Privacy/Utility/Quality

Palanisamy and Liu [190] considered the constraints on the users' movement patterns and statistical behavior for generating vehicular mix-zones. They proposed non-rectangular-shaped, adaptive mix-zones in which the mix-zone length is determined based on the average speed of the road segment, the time window, and the minimum pairwise entropy threshold.

Concerning low traffic density, Vaas et al.[263] proposed a mix-zones scheme that generates fictive chaff vehicles and broadcasts their traces when the traffic is low at the mix-zone. Their proposal smoothed the mix-zones' traffic flow and enhanced the protection up to 76%. Khodaei and Papadimitratos [262] proposed mix-zones where a subset of vehicles is selected to emulate non-existing cars, reducing the probability of linking pseudonyms from 68% to 18%.

Zhoun and Zang [100] addressed speed limitations on intersections and traffic lights where mix-zones are positioned against inference attacks. They proposed an LPPM with

two protection layers – anonymization and perturbation – which swap pseudonyms and add noise to trajectories that pass through mix-zones.

Mix-zone placement in urban environments is another concerning issue, which considers many factors that may affect anonymization coverage, e.g., the seasonality of traffic flow and the noise caused by buildings. It is an NP-hard problem [208, 175] that asks for sub-optimal solutions. Ravi et al. [208] proposed a mix-zone placement heuristic using simulated annealing and genetic algorithms considering a trade-off between privacy and cost. In it, mix-zones are selected by a ranking of intersections defined by a mixability metric that determines the usefulness of initially placing a mix-zone. Svaigen et al. [175] proposed a mix-zone placement approach based on the premise that mix-zones need to change according to the flow behavior of mobile entities. They also proposed metrics that consider the graph topology and aim to place mix-zones considering road flows between two mix-zones in a given window size.

In Chapter 4, we evidenced that mobility can impact location privacy approaches, like mix-zones, in the context of smart mobility. We explored the distributions extracted from two stay points metrics for designing LPPMs and found high similarities between distributions for the same vehicle type for monomodal and multimodal datasets. Also, we showed the side effects regarding privacy and data utility when mix-zone parameters are misconfigured in a multimodal environment.

Despite the vast literature addressing mix-zone issues, few investigations have been conducted about anonymization behavior, including mix-zones' quality and other anonymization-based LPPMs (see Table 6.1). Most proposals operate only in online mode and simulated environments. In addition, they only consider privacy, leaving aside questions about utility and quality. Regarding anonymization behavior, in Chapter 4, we proposed identifying mobility impacts on location privacy through the behavior of stay points metrics. However, we did not detail the anonymization behavior. Further, to the best of our knowledge, no previous proposal analyzed quality metrics when characterizing anonymization. Unlike previous studies, we advance the state-of-the-art w.r.t. anonymization quality. We explore the mix-zones metrics for characterizing and evaluating the impacts of anonymization quality in mobility data. Once the mix-zone metrics are extracted, it is possible to get information to understand the anonymization in terms of quality, not just traffic flow, which can help with new anonymization proposals.

In this chapter, we extended the work [173], having as the main contribution the AQM framework that enables many applications concerning mix-zones performance, such as elects more effective mix-zones from a set given a privacy budget. The framework uses the anonymization quality function built with the previous study metrics. Also, we propose two new quality metrics (the interval of departure of vehicles and vehicles that finish their trips inside the mix-zones) that compose the framework. We expanded the evaluation of AQM of the capability of elect mix-zones with good anonymization quality

from mix-zones positioned with two positioning algorithms. Further to the privacy and anonymization analysis, we present a data utility analysis with a metric for designing data dissemination protocols.

6.3 Mix-zones Problem Description

This section describes the mix-zones problem and questions about the anonymization quality in mix-zones. Mix-zones are a k-anonymity-based approach widely used in VANETs to pseudonym changing that yields anonymization of users. However, they present many drawbacks: mix-zones depend highly on positioning, arrival rates, mix-zone geometry, mobility patterns, and vehicle density [261, 262], which can degrade the anonymization rate and enable linking and inference attacks with a high success rate. Regions with low traffic are expected to have low anonymization, so deploying mix-zones in regions with high to moderate traffic is crucial. But even when positioning mix-zones in areas with high traffic, there may be different periods of the day when the minimum number of vehicles does not suffice to ensure anonymization. However, many proposals for addressing the issues mentioned above are based on static solutions that use traffic flow snapshots that do not consider events like traffic fluctuations. Moreover, they do not consider events in a real environment, like the effects of mix-zones activation time in terms of saving energy, vehicle retention, and reasonable traffic flow over time. To evaluate an LPPM, we need a deep understanding of its working and operational requirements. Specifically, mix-zones require reasonable traffic flow, in and out throughput, lower vehicle retention, and anonymizing when activated. These are fundamental pillars of anonymization quality that reflect the coverage, protection, and utility of the data protected by mix-zones.

Concerning anonymization issues, questions as mentioned in Section 6.1 arise. We must understand the anonymization behavior and its quality to answer these questions. Thus, a deeper study of mix-zones characterization is needed. We present a mix-zone characterization framework that measures the quality of anonymization produced by mix-zones.

6.4 Anonymization Quality Framework

This section introduces the anonymization quality framework and details its metrics, quality function, privacy attack, and utility analysis.

6.4.1 Framework

The Anonymization Quality Framework proposed in this work comprises six steps: Deployment (1), Protection (2), Anonymization Quality (3), Attack (4), Utility (5), and Analysis and Applications (6), as depicted in Figure 6.1.

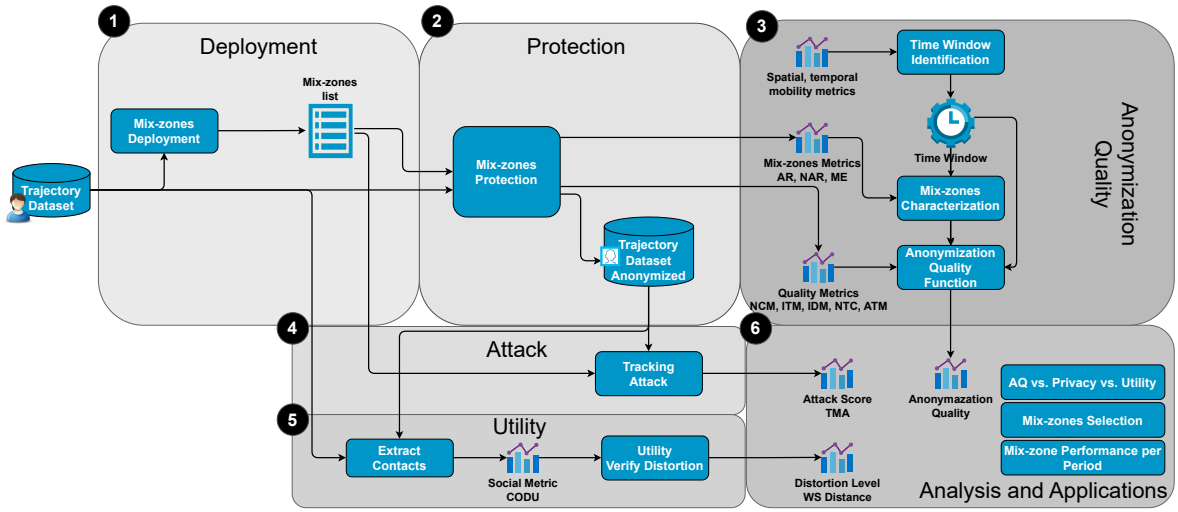


Figure 6.1: Anonymization Quality Framework for Mix-zones (AQM).

The first step, Deployment, involves positioning the mix-zones with a deployment algorithm. A sample of the trajectory dataset D is the algorithm's input, and the output is a list of mix-zones. In the second step, Protection, the dataset is protected using the list of previously deployed mix-zones, set up with radius r and k values. At the end of anonymization, we have the anonymized dataset D' and mix-zones' metrics, such as (no) anonymization rate and efficacy of each mix-zone. The quality metrics are another output of the Protection step. Quality metrics are information about the anonymization in quality terms, with variables about the mix-zones operation, such as activation time period, arrival and departure rates, etc. The mix-zones and quality metrics are output to the Anonymization Quality step.

The third step, Anonymization Quality, is the process of extracting information from the mix-zones and quality metrics to calculate the anonymization quality metric. This metric enables a deep understanding of how and why anonymization occurs over

time, allowing the identification of potentialities and limitations about certain mix-zones. For this, a well-defined time window is needed. The time window identification occurs with the analysis of spatial, temporal, and mobility metrics defining ideal periods of the day. With the time window W, it is possible to characterize the mix-zones with mix-zone metrics and calculate the anonymization quality function with quality metrics. In the fourth step, Attack, we verify the privacy of the anonymized dataset with a tracking attack. The Attack step has as input a list of mix-zones and an anonymized dataset, resulting in an attack score that defines the vulnerability level. The fifth step, Utility, identifies the anonymized dataset's utility level. It consists of computing the distortion level between the original trajectory dataset D and the anonymized trajectory dataset D'_k, which is protected with a privacy level k. As a utility, we define the application of anonymized data in the context of VANETs communication infrastructure, specifically in context-aware data dissemination protocols. We use a social metric, CODU, to extract the datasets and verify their distortion $WS(D, D'_k)$. To calculate CODU, it is necessary to get the contact data from the trajectory datasets. Finally, in the sixth step, Analysis and Application, we analyze the utility, privacy, and anonymization quality metrics, making it possible to make several applications, such as: verify the trade-off between utility vs. privacy vs. anonymization quality; elect and rank the best mix-zones in quality terms from the given mix-zones budget; elect positioning algorithms based on utility vs. privacy vs. anonymization quality; identify mix-zones performance per period; identify the advantages and limitations of mix-zone deployment algorithms, and so on. Each step of the framework is detailed as follows.

6.4.2 Coverage, Quality, and Mobility Analysis in Mix-zones

Let D be a trajectory dataset that needs to be anonymized by mix-zones. Consider a privacy budget represented by a mix-zones set $\mathcal{M}$, deployed with a positioning algorithm (Step 1 of Figure 6.1). These mix-zones have anonymization quality, can extract the quality metrics, and have anonymized D (Step 2 of Figure 6.1). Now, the current step is identifying the time window of the mobility data, so this knowledge can be used to characterize mix-zones and their anonymization quality over time (Step 3 of Figure 6.1).

6.4.2.1 Spatial and Temporal Mobility Analysis

The goal of mobility analysis is to show how different contexts, such as weekdays and day periods, have different mobility characteristics, which can be explained by traffic volume, driving conditions, as well as the existing motivations for the trips (e.g., work, leisure, and daily routines). For this, we divide the trips according to the weekday into four time windows: dawn, comprising the period from 0:00 to 5:59; morning, including the period from 6:00 to 11:59; afternoon, consisting of the period from 12:00 to 17:59; and night, including the period from 18:00 to 23:59.

After defining the time window, we apply four mobility metrics to these periods to understand the effects of traffic over the periods. Additionally, it validates if the selected time window is well-defined according to traffic fluctuation. Following, we detail these mobility metrics:

- **Trips' Average Speed (TAS)**: The average trip speed is obtained from the speed measured at each location reported at the trip's intermediary points;
- **Trips' Duration (TDu)**: Trips' duration measured in minutes;
- **Trips' Distance (TDi)**: Trips' total distance measured in km;
- **Visits per Location (VL)**: The total number of visits by all the individuals in every location during the period of observation [264].

6.4.2.2 Coverage and Quality Mix-zone Analysis

This analysis aims to identify inherent characteristics of mix-zones that lead us to understand their anonymization quality. Mix-zones with anonymization quality are mix-zones that yield high efficacy and considerable privacy levels, anonymizing the mobility data at the moment they are activated, which can thus lead to energy savings. To this end, we use the following mix-zones' coverage and quality metrics extracted from mix-zones designed with anonymization quality. Considering a mix-zone M, the mix-zone coverage metrics are:

- **Anonymization Rate (AR)**: The AR_M is the number of trajectories that passed and were anonymized by$M$;

- **Non-Anonymization Rate (NAR)**: The NAR_M is the number of trajectories that passed and were not anonymized by M;
- **Mix-zone Efficacy (ME)**: The Refers to the ratio between a number of users anonymized by M, denoted by AR_M, and the population P_M that crossed it, i.e., $ME_M = |AR_M|/|P_M|$.

The mix-zones quality metrics are:

- **Number of Cars in Mix-zone (NCM)**: Number of cars crossing the mix-zone over time. NCM allows us to understand how the parameters of the mix-zones should behave in the face of traffic fluctuations over time, such as the behavior of the parameter k. If k follows NCM, it is expected that k must be set close to the lower and upper bounds during periods with low and high traffic. This results in best-effort anonymization with different levels of privacy over time and the best possible coverage;
- **Interval of Arrival Time between Cars on Mix-zones (ITM)**: Measures the time interval in seconds between vehicles entering the mix-zone, enabling the measuring of traffic volume inside the mix-zone. If ITM is low, it means that the cars are close from each other in time, and there is a high probability of anonymization. But a high ITM means that there is a temporal distance between the vehicles and a high tendency to not reach k to anonymize;
- **Interval of Departure Time between Cars on Mix-zones (IDM)**: Measures the time interval, in seconds, between vehicles going out of the mix-zone, enabling the measurement of traffic volume inside the mix-zone. If the IDM is low, the cars are temporally close, and there is a high probability of anonymization. However, a high IDM means that there is a temporal distance between the vehicles and a high tendency not to reach k to anonymize;
- **Number of Trips Completed within the Mix-zone (NTC)**: Measures the number of vehicles that terminate their trips within the mix-zones. Depending on the mix-zone design, these vehicles may not be anonymized. Mix-zones with high NTC means low AR and ME.
- **Activation Time of the Mix-zone (ATM)**: Time period in which a mix-zone is activated for anonymization when the number of cars inside it is greater or equal to the k parameter. This metric can reflect the mix-zone quality in anonymization terms. High ATM denotes mix-zones with anonymization for a long period of time.

6.4.3 Anonymization Quality Function

The AQ considers aspects of privacy and behavior of the mix-zone over time. These aspects make it possible not only to identify the performance of anonymization but also to understand the functioning of the mix-zones over time. From there, it becomes possible to select mix-zones or operating periods of mix-zones that produce better anonymization. Next, we describe these aspects that make up the quality of anonymization. Finally, we define the formal notion of the AQ metric, shown in Step 3 of Figure 6.1.

For the AQ calculate, we consider $W = \{d, m, a, n\}$ to be a time window divided into four equal periods where d, m, a, and n represent the dawn, morning, afternoon, and night periods, respectively. Consider the s as one of the time window of the day period, e.g., s can be equal to d or m.

Mix-zones are measured as unlinkability, the capacity to link old and new pseudonyms. For a user with pseudonym $u \in A$, entering and exiting a mix-zone M with a new pseudonym u', there is a probability of mapping u to u', $p_{u' \to u}$. That is, each user u of k users exiting M with the pseudonym u' has $p_{u' \to u} = \frac{1}{|A|}$ probability of being any of the k users in the anonymity set. This way, the unlinkability can be measured with the uncertain level of each outgoing pseudonym u'. Entropy is a metric used to measure the uncertainty level provided by the anonymity technique in the system. The Entropy of pseudonym u' in a s period is denoted by $H_s(u') = -\sum_{u=1}^{|A|} p_{u' \to u} \log_2 p_{u' \to u}$.

We use entropy as privacy weight in the AQ. The entropy is calculated from a previously defined k parameter. The greater the number of vehicles within the mix-zones that have $|A| \geq k$, the higher the $H_s(u')$, that is, the level of re-identification uncertainty over time. The total number of ATM (ϕ_s) in a period s is another aspect of AQ. It is denoted by $\phi_s = \mathrm{N}(ATM)\mathrm{AVG}(ATM)$, where $\mathrm{N}(ATM)$ is the number of ATM and $\mathrm{AVG}(ATM)$ is the ATM average in the day period s. Long ATMs tends to produce greater anonymity than short ATMs. Additionally, many ATMs, even with a short duration in one period, can also bring good performance in anonymization.

Another important aspect of anonymization quality is the throughput of the mix-zone over time in the period s, denoted by $\mathcal{I}_s$. The $\mathcal{I}_s$ can be calculated by the ratio between the average of ITM (τ_y) and the average of IDM (τ_x) in a period s, denoted by $\mathcal{I}_s = \tau_x/\tau_y$, where $\mathcal{I}_s \in [0, 1]$ (after normalized). $\mathcal{I}_s$ represents the quality of vehicle flow in M. The use of ITM and IDM to determine $\mathcal{I}_s$ is important to understand the behavior of vehicle flow in M, both about the entry and exit of vehicles in M, which directly reflects the anonymization performance.

In the case where $IDM \leq ITM$, it means that more vehicles are leaving than entering M. That is, $\mathcal{I}_s \approx 0$ means higher throughput, where more vehicles leave the mix-zone than enter in a period. A mix-zone with higher throughput tends to have

better anonymization over time, as long as the policy for pseudonym changing is for vehicles to leave the mix-zone and the anonymity set complies with $|A| \geq k$. In terms of anonymization quality, it is desirable that mix-zones present $\mathcal{I}_s \approx 0$. On the other hand, if $IDM > ITM$ means that more vehicles are arriving than leaving M. In this case, $\mathcal{I}_s \approx 1$ means that more vehicles are concentrated within the mix-zone than outgoing, and the tendency is to have a smaller amount of anonymized vehicles. Also, it can indicate a vehicle jam scenario in M.

Finally, μ is the NTC that measures the number of vehicles that terminate their trips within the mix-zones, where $\mu \in [0, 1]$ and 1 is the maximum value for NTC. The AQ function can be formulated with these metrics.

The AQ of a day period s is denoted by Boltzmann distribution, represented by $AQ_s = H_s(u')\phi_s\epsilon^{-(\mathcal{I}_s+\mu)/\lambda}$. The $H_s(u')$, ϕ_s, $\mathcal{I}_s$, and μ represent privacy weight, the total number of ATMs, throughput, and the number of vehicles that terminate their trips within the mix-zone M, in the day period s, respectively. The λ is the Boltzmann constant. The $\epsilon^{-(\mathcal{I}_s+\mu)/\lambda}$ is used to penalize the result when many vehicles enter the mix-zones and do not leave it. Consequently, these vehicles are not anonymized. Specifically for the relation $\mathcal{I}_s + \mu > 1$ then $\epsilon^{-(\mathcal{I}_s+\mu)/\lambda}$ tends to zero. Otherwise, if $\mathcal{I}_s + \mu \leq 1$ then $\epsilon^{-(\mathcal{I}_s+\mu)/\lambda}$ tends to one. $AQ \in [0, 1]$ where 1 is the maximum value for the anonymization quality.

6.4.4 Tracking Attack

To explore the potentialities of the quality metrics, we investigate their behavior in re-identification attacks against mobility data protected with different privacy levels (Step 4 of Figure 6.1). The goal is to verify its capacity to identify the best mix-zones against re-identification attacks varying the k privacy levels. We used the trajectories re-identification attack proposed in [56] to validate the framework in privacy terms. This attack uses only two location points of trajectories as knowledge. The attack hypothesis is that most vehicles, especially taxis, choose minimal paths to complete their routes. Thus, it is possible to re-identify anonymous trajectories by comparing candidates' trajectories with a minimal path built between two points.

The attack efficacy denoted by Trajectory Matching Accuracy (TMA) is a privacy metric to measure the effectiveness of re-identification attacks [54], which is defined as $TMA = N_{\text{reid}}/|AR_M|$, where $N_{\text{reid}} \in [0, |AR_M|]$ and $|AR_M|$ represent the total of re-identified trajectories and the total of anonymized trajectories that have passed in the mix-zone M, respectively. TMA $\in [0, 1]$ where TMA equals 1 is the max attack success. In contrast, TMA equals 0, representing no success, and the trajectory is protected against

the attack.

6.4.5 Utility Analysis of Anonymized Data with Social Metrics

One of the characteristics that impact human mobility is our social behavior. Social relationships generate mobility, and thus by looking into mobility, we can obtain insights into how we interact, such as the frequency and duration of contacts between user pairs. We can use this information to boost technologies like context-aware data dissemination protocols. Although the metrics used to measure contacts are a subgroup of the location metrics, they can have applications in various areas, such as identifying social relationships and assisting in designing data dissemination protocols [265]. For instance, pairs of users with a longer contact duration are likelier to have some social relationship than pairs with a short contact duration. Consequently, they allow you to disseminate data more efficiently [266]. In this work, we use a metric called Contact Duration between a Pair (CODU), defined as the average time two users spend inside each other's transmission range without interruptions [265]. CODU represents an opportunity to transmit a message in an opportunistic network in seconds. The higher the CODU value, the more data can be delivered in each encounter.

Data utility is an essential aspect of Anonymization Quality. Anonymized data with enough utility can be used for the public good or monetized in the data marketplace and consumed by smart city applications [18]. Here, we discuss the utility's impact of applying mix-zones to mobility data. For this, we investigate the behavior of utility metrics on anonymized data that can be used in particular scenarios of smart city applications.

Mobility characteristics metrics can be used as utility metrics. Assuming that when applying an LPPM, data can be lost, we can explore mobility aspects to identify which mobility data types were most affected. For instance, the distance between trips, time at specific points of interest, and social relationships with others. The idea is to identify data distortions with a distance between the mobility model metrics measured before and after applying an LPPM, such as mix-zones, in the dataset $\mathcal{D}$. So, the distortion of the statistical distance of mobility metrics has been used as a utility metric in the mobility context.

Statistical distance is the approach we use to identify the distance between two probability distributions. We applied the Wasserstein metric (WS), which measures the difference between two distributions by the optimal cost of rearranging one distribution into the other[1].

[1]For more details, please refer to C. Villani, "Topics in Optimal Transportation". American Mathe-

Table 6.2: Mix-zones positioned with FPMT and their locations.

Name	Latitude	Longitude	Coef. Var. (%)	Location	Places Covered by Mix-zone
mix0	37.714801	-122.397982	71.20	101 express way, Visitacion Valley.	cafes, subways station, restaurants.
mix1	37.724830	-122.400157	74.42	101 express way, Bay View.	supermarkets, restaurants, clinics.
mix2	37.735133	-122.404532	79.27	101 express way, Bernal Heights.	road interchange, gas station, supermarket.
mix3	37.676005	-122.391491	63.93	101 express way, Firth Park.	tourist area, hotels, docks.
mix4	37.615315	-122.393566	54.08	entrance road to the airport.	cabs park, bus stop, restaurant, garages
mix5	37.774378	-122.401540	59.21	downtown, near 101 express way.	exit to Oakland city, shopping, church.
mix6	37.768990	-122.419450	**90.68**	downtown, Mission District.	tourist area, museums, subways station.

Definition: The distortion associated with a random variable X is the result over the original dataset $D \in \mathcal{D}$ as $WS(D, D'_k) = Wasserstein[X(D), X(D'_k)]$. The smaller the WS value, the less effort is needed to transform one distribution into another; consequently, the two distributions show high similarity. The Wasserstein distance is asymmetric, (weakly) continuous, and ideal for analyzing corrupted data, unlike common distribution divergence approaches, such as Kullback-Leibler or Jensen-Shannon [256]. In this paper, we verify the utility of datasets in the context of dissemination protocols in VANETs, denoted in Step 5 of Figure 6.1. Specifically, we analyze the WS of CODU metric of the dataset D before and after being submitted to mix-zones, previously configured, i.e., $WS[CODU(D), CODU(D'_k)]$.

6.5 Results and Discussion

This section presents the results and discusses the Anonymization Quality Framework for Mix-zones (AQM). Firstly, we present the experiment setup, including the dataset details and deployment algorithms. Next, we show the window definition and the relationship between mix-zones metrics for characterization. We also state the results of mix-zones metrics and quality metrics. Next, we discuss the utility of anonymized data. In the sequence, we present the AQ function and the relation between each positioning algorithm's mix-zones metrics and quality metrics. Finally, we present the analysis of AQ vs. privacy. vs. utility as a form to elect better mix-zones.

6.5.1 Experiments Setup

In this study, we used the *Cabspotting* dataset, containing mobility traces of taxicabs of San Francisco, USA, with data from 500 unique taxis collected over 25 days

matical Soc., 2003, no.58

Table 6.3: Mix-zones positioned with DBSP and their locations.

Name	Latitude	Longitude	Coef. Var. (%)	Location	Places Covered by Mix-zone
mix0	37.615837	-122.387579	69.20	inside airport, board/onboarding region.	airport, cab stop point, bus stop.
mix1	37.705587	-122.393410	73.37	101 express way, near Visitacion Valley.	bay coast, tourist area.
mix2	37.679245	-122.388355	**93.61**	101 express way, Firth Park.	tourist area, hotels, docks.
mix3	37.789005	-122.402847	93.48	downtown, Belden Place.	tourist area, museums, subways station.
mix4	37.778224	-122.392002	79.31	downtown, Oracle Park.	tourist area, docks, baseball stadium.
mix5	37.782250	-122.410558	84.30	downtown, Union Square.	bus stop, subways station, civic center.
mix6	37.746171	-122.394069	73.45	280 express way, Islais Creek.	near coast, supermarket, ship center.

[126]. Collected in 2008, this trace contains data about the taxicabs' location, sampled periodically through an embedded GPS sensor. Additionally, each record contains a flag indicating if a passenger is being transported, i.e., if a trip is happening (if not, the driver is moving around looking for customers). Overall, there are approximately 440,000 trips, with an average of 17,600 daily trips. With this mobility behavior, more than 70% of the road segments within the city limits are visited at least once, generating an average of 400,000 contacts between vehicles per day, affirming the potential of this dataset for our analysis.

To validate the framework, we used two mix-zones deployment algorithms – FPMT and DBSP – which yield the mix-zones candidates' list, each one respectively sorted by vehicle frequency on intersections and centroids in descending order. From each candidate list, only seven were empirically selected based on two conditions. First, the mix-zones must anonymize trajectories in at least two of the three k setups. Second, the mix-zones could not overlap areas between the mix-zones already selected. Given the anonymization condition above and the restricted size of the San Francisco region, it resulted in a privacy budget of seven mix-zones.

Tables 6.2 and 6.3 show the mix-zone locations positioned with FPMT and DBSP algorithms, respectively. They also show the relevant places covered by them. We extract a sample D of high traffic from the *Cabspotting* dataset corresponding to the 19th day of the collection period with 417,781 registers, 454 users, and 2036 trips. Then we anonymized it, varying the parameter k with values 2, 4, and 6, but with a radius r of 500 meters equal to all mix-zones for the cab dataset [261]. The radius r of 500 meters was defined by a coverage empirical analysis ranging the radius threshold r from 100 to 600 m ($r = \{100, 200, 250, 300, 400, 500, 600\}$). From $r = 600$ m, we had mix-zones overlapping, significantly reducing the privacy budget.

6.5.1.1 Mix-zones Deployment Algorithms

We choose two distinct approaches of mix-zone deployment algorithms to evaluate the AQ framework to elect mix-zones that yield anonymization quality. The first one,

FPMT, uses the frequency of points from the segments of the middle of trajectories for positioning the mix-zones in the road intersections. The second one, the DBSP, is based on the clustering of hot points of the users, such as SPs, for positioning the mix-zones. In the following, we detail each algorithm.

The basic principle of the deployment algorithm based on the Frequency of Points from the Middle of the Trajectory (FPMT) (Algorithm 3) is to deploy the mix-zones on road intersections with the highest occurrence of central segment points of the trips that crossed the intersections [56]. The algorithm's inputs are a trajectory dataset D and a city map. The first step of the algorithm is to extract the sample of intersections I from the map (line 1). Next, the trajectories T (line 2) will be randomly sampled, and the middle portion of intersections $S[T]$ (line 3) will be selected. With the middle segment of the trajectory and the intersections, the next step is to calculate the frequency of cars visiting each intersection. It returns the intersections list I_{freq} sorted by frequency in descending order (line 4). The next step is to re-calibrate the I_{freq} position according to the intersections on the road generating the mix-zones candidates MZ (line 5). Finally, it must remove mix-zones that overlap in terms of space resulting in a mix-zones set $\mathcal{M}$ (line 6).

Algorithm 3: FPMT algorithm.

Data: Trajectory Dataset D; City Map Map
Result: Mix-zones set $\mathcal{M}$

1 $I \leftarrow loadIntersections(Map)$;
2 $T \leftarrow getTrips(D)$;
3 $S \leftarrow getMiddleSegments(T)$;
4 $I_{freq} \leftarrow freqCarsOnIntersections(S, I)$;
5 $MZ \leftarrow matchPoint(I_{freq}, Map)$;
6 $\mathcal{M} \leftarrow chooseMZDistanceLimit(MZ)$;

Algorithm 4: DBSP algorithm.

Data: Trajectory Dataset D; City Map Map
Result: Mix-zones set $\mathcal{M}$

1 $T \leftarrow getTrips(D)$;
2 $SP \leftarrow getStayPoints(T, r, t)$;
3 $SP_{Clist} \leftarrow getClusteringDBSCAN(SP)$;
4 $C_{centroid} \leftarrow getCentroids(SP_{Clist})$;
5 $MZ \leftarrow matchPoint(C_{centroid}, Map)$;
6 $\mathcal{M} \leftarrow chooseMZDistanceLimit(MZ)$;

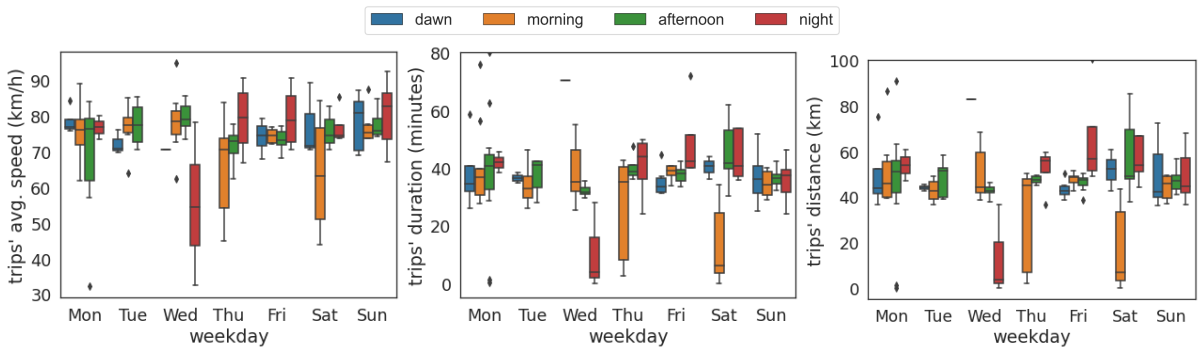
The DBSCAN Stay Points (DBSP) is based on stay points clustering extracted from trajectories to deploy mix-zones. Stay Point (SP) is a region where an entity stays for a minimum time interval [233]. The parameters of a SP are the radius r (in meters) of the region and the minimum length of stay t (in minutes). These points are relevant for detecting many mobility characteristics, such as traffic lights and traffic jams. Furthermore, stay points are commonly used as a substrate for many privacy mechanisms in the context of location privacy. In LPPM design, stay points are typically used to detect Points of Interest (PoIs) and apply obfuscation methods. Additionally, we can use stay points for mix-zones placement [32]. This way, stay points bring valuable information w.r.t. location privacy [261].

The DBSP algorithm (Algorithm 4) has as inputs a trajectory dataset D and a city map Map. The first step of the algorithm is randomly sampling T from D (line 1). Next, it collects the stay points set SP from trajectories (line 2). The stay points have as parameters the radius r and a minimum length of stay t. Here we use r as 500

meters (as suggested in [261]) and one minute for collecting the maximum of stay points, including traffic lights. The next step is to obtain the cluster list SP_{Clist} of stay points, for which we use the DBSCAN clustering algorithm[2] (line 3). The SP_{Clist} is sorted by descending order of cluster size. We configured the distance between the two samples as 100 meters for the clustering process. The next step is to get the centroid $C_{centroid}$ from SP_{Clist} (line 4). Then, the algorithm re-calibrates the $C_{centroid}$ position according to the point on the road generating the mix-zones set candidates MZ (line 5). Finally, it must remove mix-zones that overlap in terms of space resulting in a mix-zones set $\mathcal{M}$ (line 6).

Further to flow-based and clustering algorithms discussed above, we can use alternative positioning algorithms, such as those based on genetic algorithms, simulated annealing, or bio-inspired algorithms like ant-colony optimization [208, 267]. Because the positioning corresponds to an independent module and the AQ framework works like a complementary process for selecting already positioned mix-zones.

6.5.2 Definition of Time Window



(a) Trips' Average Speed (TAS) (b) Trips' Duration (TDu) (c) Trips' Distance (TDi)

Figure 6.2: Mobility metrics distribution per period, grouped by weekday.

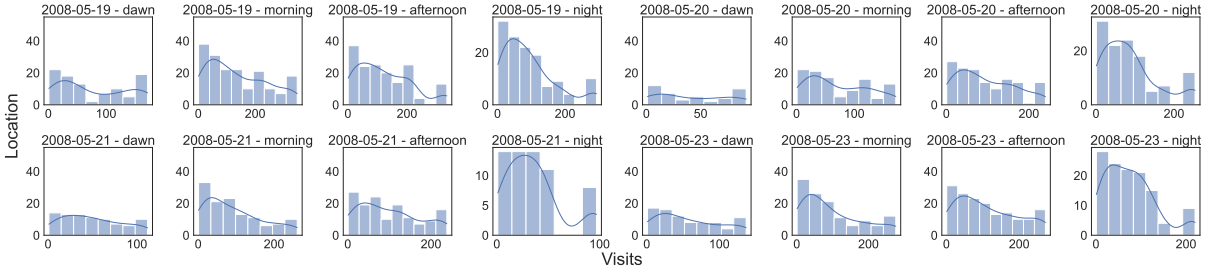


Figure 6.3: Visits per Location (VL) per day periods.

We consider the concept of the time window to identify differences in mobility behavior due to temporal aspects, assuming that the population inside each window will

[2]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

behave differently according to the traffic conditions. The division into four periods enables us to aggregate the trips in intervals that well represent the variation of the traffic intensity in the city in periods such as peak times or not [268]. For this, we evaluated four metrics: TAS, TDu, TDi, and VL and the results can be seen respectively in Figures 6.2a, 6.2b, 6.2c, and 6.3. In addition to the daily time window, we split our analysis into weekdays, allowing us to compare with the distributions in the other periods. We used a sample of seven days, from Monday to Sunday – this sampling corresponds to 30% of the dataset with 1,982,509 GPS records generated by all cab users. It also starts on the first business day of the week.

For TAS (Figure 6.2a), the increase in traffic during working hours will lead to slower average speeds, which can be noticed by lower medians for morning and afternoon periods in comparison with dawn and night. Similarly, TDu (Figure 6.2b) varies according to the time window, which can be evidence for different trip purposes (work or leisure, for example) as well as evidence of traffic increasing the overall duration of trips. Finally, TDi also varies during the time window for the same reasons, that is, different trip purposes and drivers taking detours and other routes in an attempt to avoid traffic (Figure 6.2c).

Regarding the VL metric (see Figure 6.3), there is a similarity between the VL distributions of the same day period on four different days. For example, at night, the VL distribution curve presents many locations with few visits (when cab drivers are at home or in the taxi parking lot). In contrast, during daytime periods, the VL tends to be uniform, where cabs are distributed throughout the city, e.g., cab points, parking lots, and cab trips. At dawn, the VL tends to be more uniform due to the reduction of vehicles in circulation. The results show that each time window presents particular behaviors due to temporal effects, such as traffic volume and trip purposes. Thus, to evaluate mix-zone performance in every different scenario, we apply, in our analysis, the defined time window W of four periods: dawn with a period between 00:00:00 and 05:59:59, morning with a period of 06:00:00 to 11:59:59, afternoon from 12:00:00 to 17:59:59, and night from 18:00:00 to 23:59:59.

6.5.3 Mix-zone Characterization

After we positioned the mix-zones with the positioning algorithms, we anonymized the dataset D with them. Figure 6.4 depicts the mix-zones metrics of mix-zones positioned with FPMT. We note that k is inversely proportional to the AR. For instance, in this scenario, for $k = 6$, we have the highest privacy level, and the re-identification probability is $1/6$ in the worst case. Still, the AR is lowest than $k = 2$ and $k = 4$ setups. The

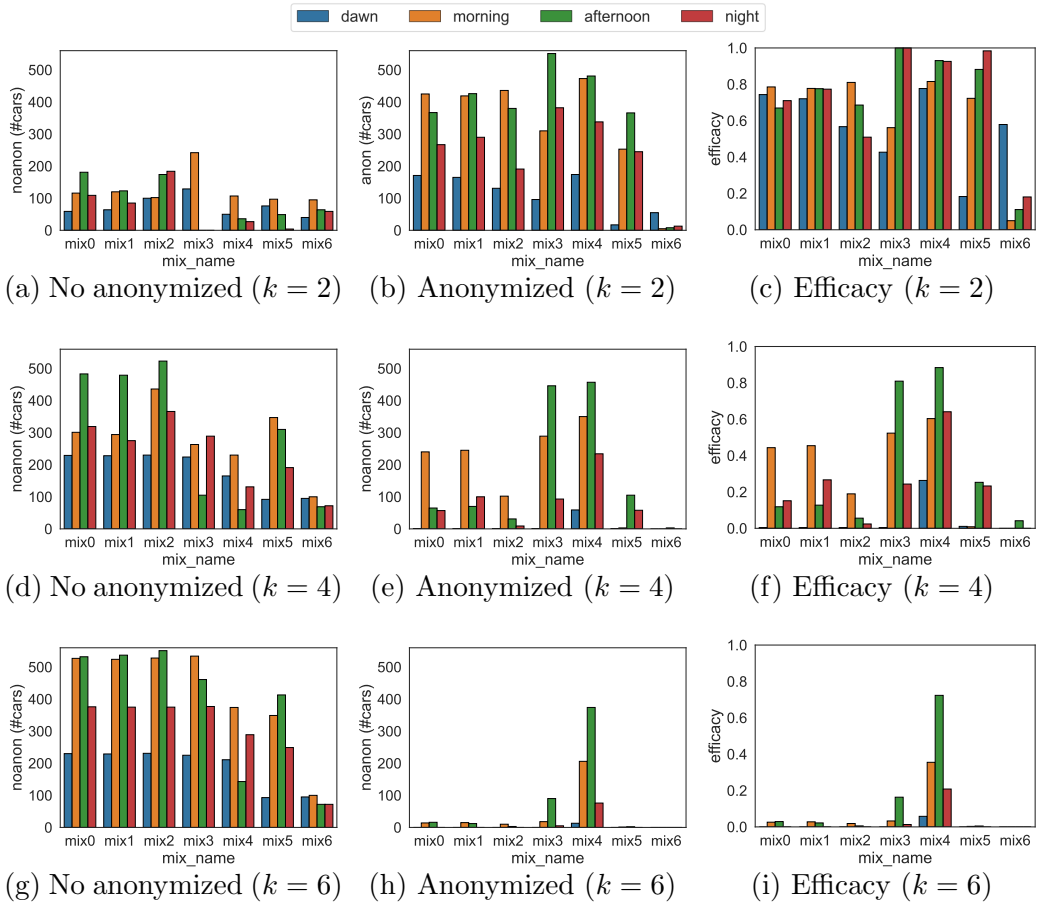
coefficient of variation (CV) measures the dispersion of a probability distribution used to identify the percentage of variation about the central mean of a sample. Table 8.1 shows the coefficient of variation[3] of vehicles for each mix-zones. We can see that mix6, mix2, and mix1 had higher variation in the number of vehicles about the central average than the rest of the mix-zones, with highlight to mix6 with the highest coefficient of variation. This means that these mix-zones had a larger oscillation of vehicle density over time, which can negatively impact their anonymization rate, as can be seen in the AR metric for setups $k = 2$, 4, and 6 in Figures 6.4b, 6.4e, and 6.4h, respectively.

Figure 6.6 depicts the mix-zones metrics of mix-zones positioned with DBSP. We note that the mix-zones metrics behaved similarly to those from the mix-zones positioned with FPMT in which k is inversely proportional to the AR. When $k = 6$, we have the highest privacy level but low AR. Additionally, the AR is lower when $k = 6$ in comparison to $k = 2$ and $k = 4$ setups. Mix2, mix3, and mix5 had a higher variation in the number of vehicles w.r.t. the central average than all the mix-zones (see Table 6.3). Highlighting for mix3 and mix5 that the variation of vehicles over time in these mix-zones had a negative impact on the AR metric in relation to the other mix-zones, as seen in Figures 6.6b, 6.6e, and 6.6h for the three setups of k.

6.5.3.1 Number of Cars in Mix-zone (NCM)

For each mix-zone, we extracted the NCM metric, which reflected the AR over time (Figure 6.5a). During the dawn, few taxis traveled in the mix-zones generating lower AR values (Figure 6.4b). Over time, the volume of vehicles and AR increase during the morning, afternoon, and night. This behavior is more accentuated for $k = 4$ (Figure 6.4e). Mix1, mix2, and mix4 had the highest volume of simultaneous vehicle traffic at noon, with NCM up to 6, 6, and 12, respectively, producing the highest anonymization spikes. For mix3, the peak of NCM was at night. However, the AR prevailed in the afternoon. Mix4 has a particular discrepancy of NCM from the others as it is located in the airport region. The mix-zones mix0, mix3, and mix5 had a higher flow at night, with NCM values of 4, 7, and 5, respectively, and AR in the evening. Mix6, located in the central region of San Francisco, had little traffic volume with an NCM of 3 vehicles during the dawn, which is its period of most significant AR. The factors that led the mix-zones to have different anonymization performances can be related to their positioning along the city, implying traffic flow and the high number of vehicle trips that start or end within them.

[3]The coefficient of variation (CV) measures the dispersion of a probability distribution used to identify the percentage of variation about the central mean of a sample.



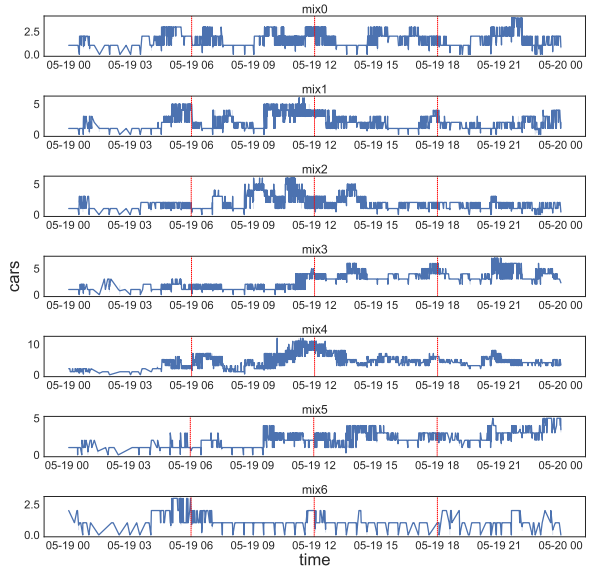
(a) No anonymized ($k = 2$) (b) Anonymized ($k = 2$) (c) Efficacy ($k = 2$)

(d) No anonymized ($k = 4$) (e) Anonymized ($k = 4$) (f) Efficacy ($k = 4$)

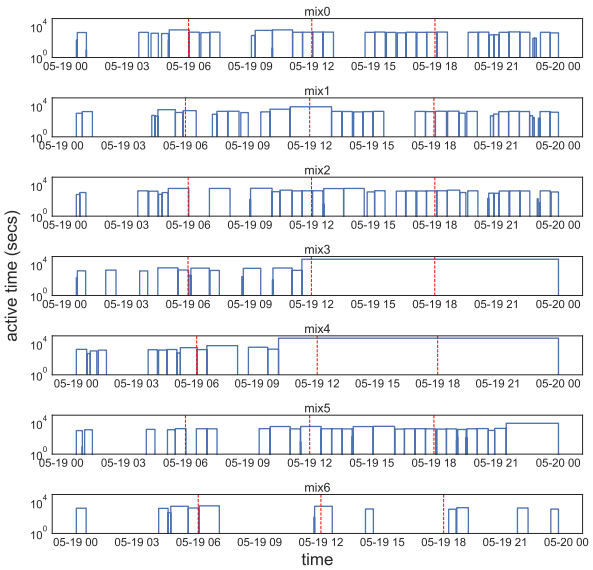
(g) No anonymized ($k = 6$) (h) Anonymized ($k = 6$) (i) Efficacy ($k = 6$)

Figure 6.4: (Non) Anonymization and Efficacy of the mix-zones for $k = [2, 4, 6]$ positioned with FPMT algorithm.

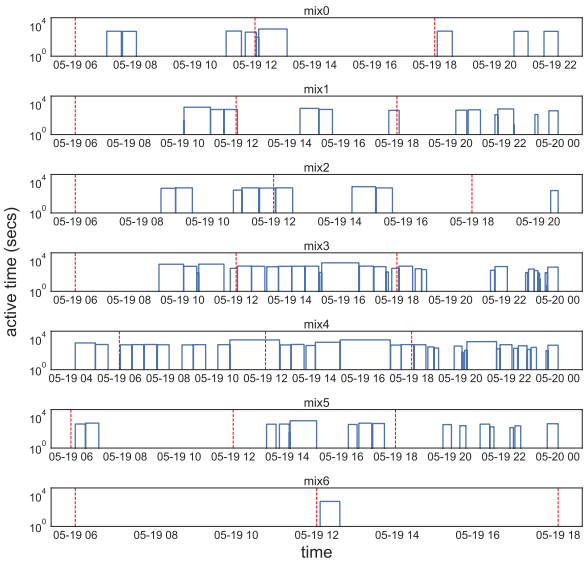
The NCM metric of mix-zones positioned with DBSP (Figure 6.7a) presents behavior similar to the FPMT algorithm. During the dawn, the traffic flow of taxis in the mix-zones generated lower NCM and AR values in any k setup (Figures 6.6b, 6.6e, and 6.6h). The NCM and AR tend to increase during other day periods. For instance, the periods morning and afternoon of mix2 configured with $k = 2$ had more AR than other periods, with AR of 525 and 468 vehicles (see Figure 6.6b) and the NCM equal to 12 and 11 cars in these respective periods. Mix0, mix2, mix3, mix4, and mix6 had the highest volume of simultaneous vehicle traffic at noon, with NCM reaching 11, 12, 4, 5, and 6, respectively, producing the highest anonymization spikes. Except for mix1, which had the highest AR in the morning, it got around 468 vehicles and NCM equal to 6. Mix2 stood out from the others and is positioned on the 101 expressway near the docks tourist region.



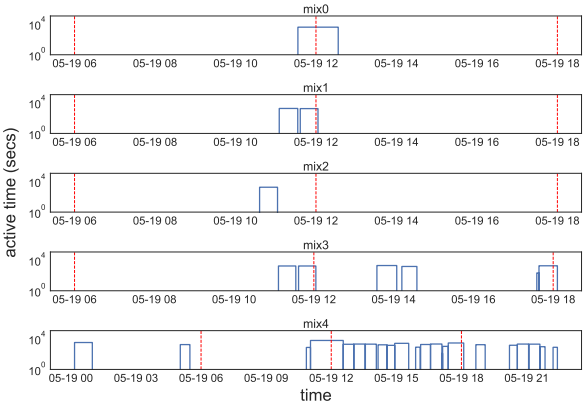
(a) NCM



(b) ATM for $k = 2$



(c) ATM for $k = 4$

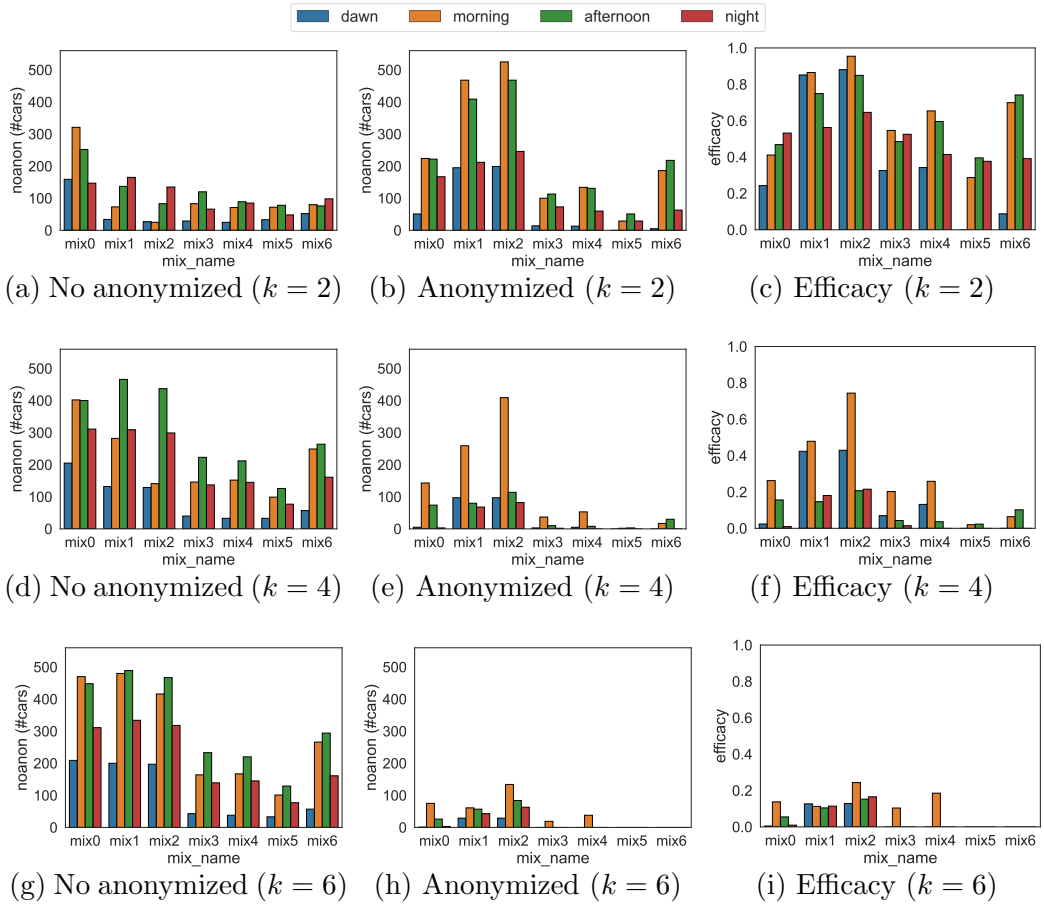


(d) ATM for $k = 6$

Figure 6.5: NCM and ATM for mix-zones positioned with FPMT algorithm.

6.5.3.2 Flow Quality Metrics in Mix-zones

The flow quality metrics in mix-zones, are composed of the Interval of Arrival Time between Cars on Mix-zones (ITM), the Interval of Departure Time between Cars on Mix-zones (IDM), and the Number of Trips Completed within the Mix-zone (NTC). Regarding ITM and IDM metrics for mix-zones positioned with FPMT, it is possible to capture the anonymization in the mix-zones over time. In the dawn, the ITM and IDM averages (Figures 6.8a and 6.8b) and NAR (Figure 6.4a) are significant in all mix-zones, while the AR (Figure 6.4b) and ME (Figure 6.4c) are low. For the rest of the day, the ITM and also IDM average is twice lower than dawn, and the curve is crescent per period,



(a) No anonymized ($k = 2$) (b) Anonymized ($k = 2$) (c) Efficacy ($k = 2$)

(d) No anonymized ($k = 4$) (e) Anonymized ($k = 4$) (f) Efficacy ($k = 4$)

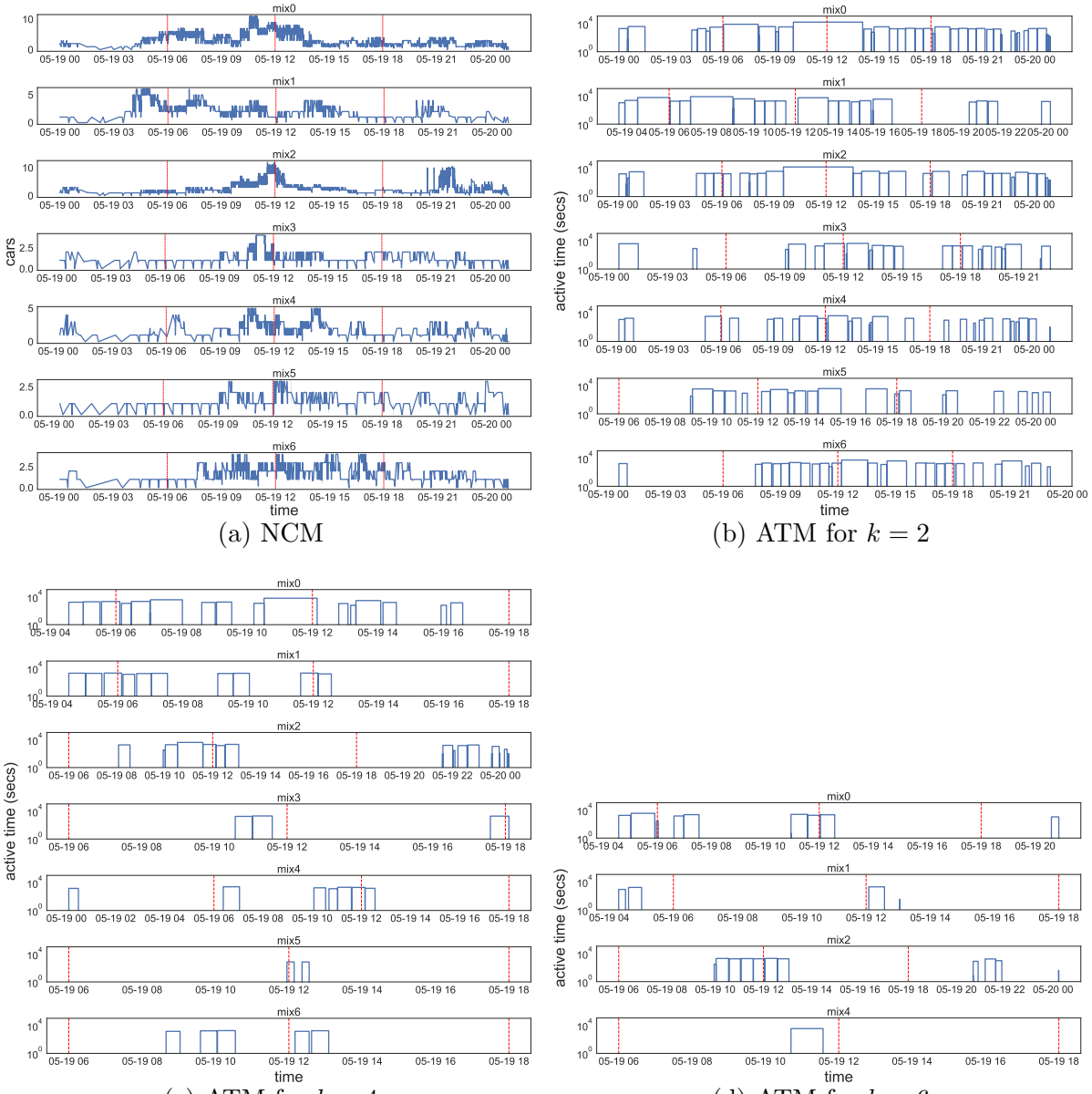
(g) No anonymized ($k = 6$) (h) Anonymized ($k = 6$) (i) Efficacy ($k = 6$)

Figure 6.6: (Non) Anonymization and Efficacy of the mix-zones for $k = [2, 4, 6]$ positioned with DBSP algorithm.

representing a significant probability of anonymization. For instance, for mix0, the ITM average per period is 92.25, 33.89, 39.00, and 54.98, and the IDM is 105.98, 35.77, 36.56, and 52.84 while the ARs are 171, 425, 367, and 267 for dawn, morning, afternoon, and night, respectively. The mix6 is less likely to anonymize the vehicle because it has low and sparse vehicle traffic than the other mix-zones and performs better during dawn.

Figure 6.8c presents the relation between IDM and ITM, $\mathcal{I}_s$. The results were normalized with *min_max* for better visualization. We can note that the dawn period, rather than other day periods, can represent that more vehicles enter and leave the mix-zones. Then it can affect the anonymization performance of the mix-zone. For instance, mix0, mix3, mix5, and mix6 had $\mathcal{I}_s$ above 0.7 in the dawn periods; we can note that for these periods was low to AR, ME (Figure 6.4) and NCM (Figure 6.10). In contrast, in the periods of $\mathcal{I}_s$ near zero, we had a high AR, such as the afternoon period for mix0, mix1, mix2, mix3, mix4, and mix6.

The ITM and IDM of mix-zones deployed with DBSP presented the same behavior that mix-zones positioned with the algorithm of FPMT (Figures 6.8e and6.8f). Specifically, the ITM and IDM rates in the dawn are at least twice as high as the other periods, indicating low vehicle flow, AR, and NCM (Figure 6.7a). For instance, for mix3, the



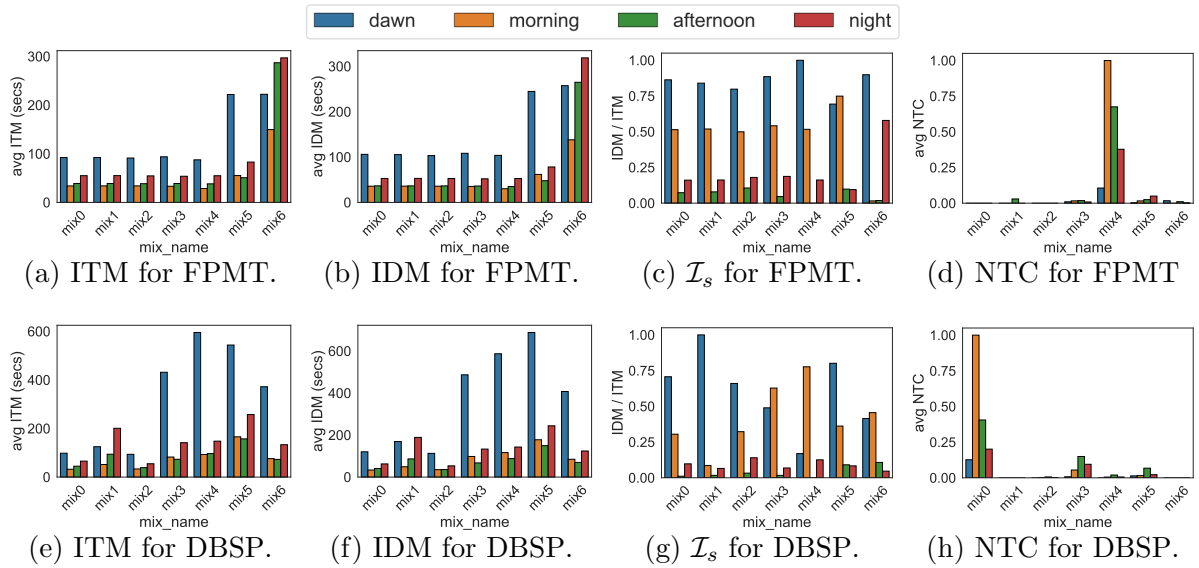
(a) NCM (b) ATM for $k = 2$

(c) ATM for $k = 4$ (d) ATM for $k = 6$

Figure 6.7: NCM and ATM of mix-zones positioned with DBSP algorithm.

ITM average per period is 431.63, 82.46, 73.33, and 139.91, and the IDM is 486.82, 98.05, 67.36, and 133.25, while, the ARs with $k = 2$ are 14, 100, 113, and 73 for dawn, morning, afternoon, and night (Figure 6.6a), respectively.

The ratio between IDM and ITM, called $\mathcal{I}_s$, showed behaviors of the mix-zones that the coverage metrics cannot detect. For instance, the possibility of vehicle concentration and traffic jams within the mix-zone over time. The $\mathcal{I}_s$ near to one may suggest a large concentration of vehicles in the mix-zone region, an indication of a traffic jam. Or even the vehicles are parked and are not leaving the mix-zone, which are likely not to be anonymized, such as in the dawn to mix3 and mix5 positioned with FPMT (Figure 6.8c) and mix0 placed with DBSP (Figure 6.8g). Particularly, mix0 is positioned at a point of taxi inside the airport and has greater inflow than outflow of vehicles within the mix-zones



(a) ITM for FPMT. (b) IDM for FPMT. (c) $\mathcal{I}_s$ for FPMT. (d) NTC for FPMT

(e) ITM for DBSP. (f) IDM for DBSP. (g) $\mathcal{I}_s$ for DBSP. (h) NTC for DBSP.

Figure 6.8: ITM, IDM, its ration normalized, and NTC of mix-zones positioned with FPMT and DBSP algorithms.

in dawn and morning. Taxis tend to park, and consequently, they are not anonymized. The high NAR in the mix0 identifies this about the other mix-zones (Figure 6.6). In this case, the $\mathcal{I}_s$ of mix0 is 70%, 30 % at dawn and in the morning, indicating this behavior and showing the low AR and ME about the other mix-zones. The hypothesis of vehicle concentration can be evidenced with the NTC metric.

The NTC identifies mix-zones that vehicles terminate their trips inside of the mix-zones and are not anonymized. For instance, mix0 positioned with DBSP inside the airport had the highest NTC in all periods than the others, with a peak at the noon period (see Figure 6.8h). Similarly, the mix4 positioned with FPMT had the highest NTC in all periods than the other mix-zones, with a peak at the noon period (see Figure 6.8d); however, mix4 had the best performance of AR and ME equal 481 and 93% with $k = 2$ setup, including compared with mix0 that had NTC equals to 0. A point to be taken into account is that two mix-zones close to each other, positioned by different positioning algorithms, can have different performances. For instance, the mix4 positioned with FPMT and the mix0 placed with DBSP stay in the international airport and have different AR and ME. This discrepancy is because mix4 was positioned in the airport portal, while mix0 was positioned inside the taxi stop point. The mix4 has the advantage that vehicles cross it and thus generate anonymity. As for mix0, the vehicles end their trips within the mix-zone, so they are not anonymized, affecting the performance of AR and ME, as seen in Figures 6.4 and 6.6. This fact is related to the mix-zone's anonymization policy. Mix-zones can be designed to anonymize when the vehicle enters or goes out of the mix-zones. In our algorithm, the mix-zones anonymize when the vehicles go out of them, generating this discrepancy between AR and ME.

6.5.3.3 Activation Time of the Mix-zone (ATM)

We perform two analyses in the ATM metric for each positioning algorithm to understand the anonymization quality. The first analysis explores the relationship between ATM and anonymization over time (ATM vs. Anonymization). In the second analysis, we verify the relationship between ATM and different privacy levels for choosing good-quality mix-zones on anonymization (ATM vs. Privacy).

The ATM vs. Anonymization analysis naturally can raise questions such as:

- What is the ATM over the time in which the anonymization is performed?
- Are longer-duration ATMs more efficient than shorter-duration ATMs in terms of anonymization?
- Or are short ATMs close to each other better for anonymization?

To answer these questions, we observe mix-zones set up with the same k and verify that ATM follows the anonymization behavior over time through the ME metric.

For mix-zones positioned with FPMT algorithm, setup with $k = 2$ (Figure 6.5b) presented the best anonymization performance, with mix0, mix1, mix2, mix3, and mix4 anonymized more than 400 vehicles in at least one period of the day (Figure 6.4), with ME rates above 90% and high ATM, except for the mix6 that has low traffic. Mix0, mix1, mix3, and mix4 had the highest values of ATM with 3512 secs (morning), 7207 secs (morning, afternoon), 44916 secs (morning, afternoon, night), and 50181 secs (morning, afternoon, night). In these periods, the AR and ME peaks were, respectively, 425 and 78% to mix0; 426 and 77% to mix1; 551 and 100% to mix3; 481 and 93% to mix4. The morning and afternoon periods got the highest AR of trajectories and their ME. Different from mix3, mix6 had anonymization peaks in the periods (afternoon, night) and (morning, night). We can observe that the number of ATMs per period tends to follow the ME and the AR. For example, the highest ME per period for mix1 and mix2 are the morning and afternoon periods, with more ATMs. Another note is that long ATMs are more efficient than short ATMs close to each other. In particular, mix3 and mix4 present 100% ME during periods with an ATM of 44916 and 50181 secs of duration, involving the morning, afternoon, and night periods. Even mix5 is more effective in the night period, in which the longest ATM prevails (9083 secs), compared to other short and adjacent periods, such as ATMs in the afternoon or morning.

In ATM vs. Privacy analysis, we focused on the variation of the k parameter to understand the behavior of ATM with different privacy levels. We expect that the anonymization quality can be measured when a mix-zone has long ATMs or a mix-zone has ATMs, even with the upward variation of k. For this, we perform two investigations in ATM vs. Privacy analysis. In the first one (**ATM vs Priv1**), we verify the relation

between ATM, AR, and ME metrics to understand if it is possible to identify mix-zones that yield good anonymization with the variation of k. For $k = 4$ setup, the AR, ME, and ATM were lower than $k = 2$ setup (Figure 6.4). However, the privacy levels are higher than the $k = 2$ mix-zones setup. Likewise, when $k = 2$, the ATM $k = 4$, the ATM tends to be larger in periods with high efficacy, e.g., in the afternoon at mix3 and mix5, respectively. Additionally, the privacy level is inversely proportional to the amount of ATM (see Figures 6.5b, 6.5c). For $k = 6$ setup, mix3 and mix4 are highlighted more than others, with trajectories anonymized mainly in the afternoon (Figure 6.4h). Also, the peaks of the ATMs were reduced to 2000 and 5000 secs to mix3 and mix4, respectively. Also, mix5 and mix6 disappeared because no trajectory was anonymized (Figure 6.5d).

In the second analysis of ATM vs. Privacy (**ATM vs Priv2**), we explore ATM behavior in re-identification attacks as defined in Section 6.4.4. We used the sampling of the 19th day of mobility data that was anonymized by mix-zones varying k with 2, 4, and 6 (Figure 6.11a). We show that mix-zones with high k (such $k = 6$) and some ATM values present low TMA, e.g., mix2, mix3, and mix4. In contrast, the mix-zones that do not have ATM present high TMA, such as mix5 and mix6. Another point to note is that mix5 and mix6 have a higher TMA for $k = 4$, $k = 6$ than for $k = 2$. This occurs because the vehicle flow is more sparse in these mix-zones, as evidenced by the high ITM average (Figure 6.8a). Furthermore, as k increases, AR and ME decrease significantly. Under these conditions, for $k = 6$, the chances of anonymization are lower than $k = 2$, which results in a higher TMA for $k = 6$.

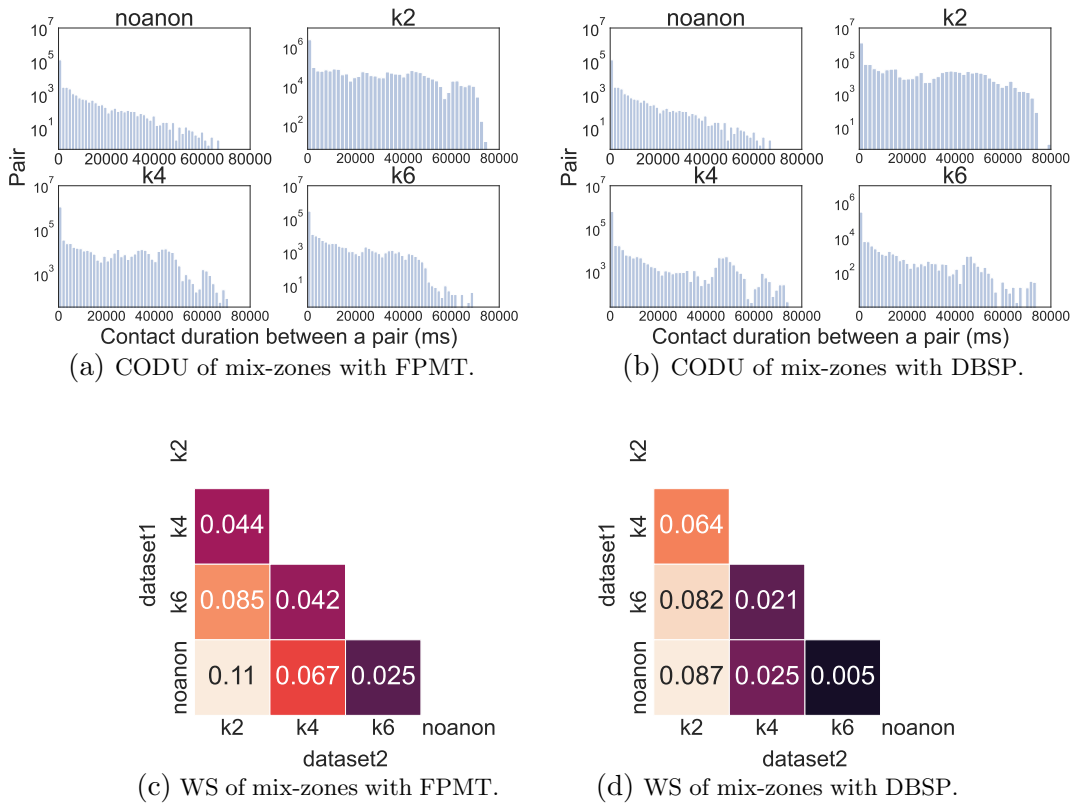
The ATM metric for mix-zones positioned with DBSP presented a similar behavior in the periods as FPMT, e.g., the dawn had few ATMs and had low AR and ME, and along the day periods, the ATM, AR, and ME tended to increase. However, mix-zones positioned with DBSP had an inferior performance than FPMT in AR and ME. For instance, $k = 2$ (Figure 6.7b) presented the best anonymization AR performance than other setups, with just mix0, mix1, mix2, and mix6 anonymized more than 200 vehicles in at least one period of the day (Figure 6.6), with ME rates above 53%, 86%, 95%, and 74%, high ATM, except for the mix3, mix4, mix5 that has low traffic 598, 608, and 340 vehicles in all day, respectively. Mix0, mix1, and mix2 had the highest values of ATM with 14390 secs, 7253 secs, and 14447 secs (all starting in the morning), and AR equal to 224, 468, 525, and ME equivalent to 41%, 86%, 95%, respectively in corresponding periods. One fact is that a long ATM does not always mean better AR performance. For example, the longest ATM of mix4 is at noon with 4053 (Figure 6.7b), and the peak of the AR occurs in the morning with 134 vehicles (Figure 6.6). On the other hand, many ATMs with less time and close to each other are slightly lower than long ATMs, as at night for mix0 and mix3 protected with $k = 2$, which had this behavior and AR of 167 and 73.

The **ATM vs. Priv1** analysis for mix-zones positioned with DBSP also had be-

havior similar to FPMT case. Although, $k = 4$ setup, the AR, ME, and ATM were lower than $k = 2$ setup (Figure 6.6), the privacy with $k = 4$ is higher than $k = 2$. Like this $k = 2$, the ATM with $k = 4$ tends to be large in periods where efficacy is high, such as in the morning at mix2. Additionally, the privacy level is inversely proportional to the number of ATM (see Figures 6.7b, 6.7c, and 6.7d).

The **ATM vs. Priv2** analysis to mix-zones positioned with DBSP also presented a similar behavior to the mix-zones positioned with FPMT. A high k, such as $k = 6$, and some ATM values present lower TMA than $k = 2$, e.g., mix0, mix1, and mix2 (Figure 6.11d). In contrast, the mix-zones that do not have ATM present high TMA, such as mix5 and mix6 configured with $k = 6$ (Figure 6.7). Different from FPMT case, the mix-zones deployed with DBSP and configured with $k = 4$ had the best performance against attacks than other k setups, as we can see for mix0, mix1, mix2, mix3, and mix4.

Finally, with k variation and with ATM, it is possible to elect the mix-zones not just considering traffic flow but also activation time, yielding anonymization quality. Zones mix4, mix3, mix1, and mix0 present better anonymization quality than other mix-zones candidates.



(a) CODU of mix-zones with FPMT.

(b) CODU of mix-zones with DBSP.

(c) WS of mix-zones with FPMT.

(d) WS of mix-zones with DBSP.

Figure 6.9: CODU metric and WS of the (non)anonymized data for mix-zones with $k = (2, 4, 6)$, positioned with FPMT and DBSP algorithm.

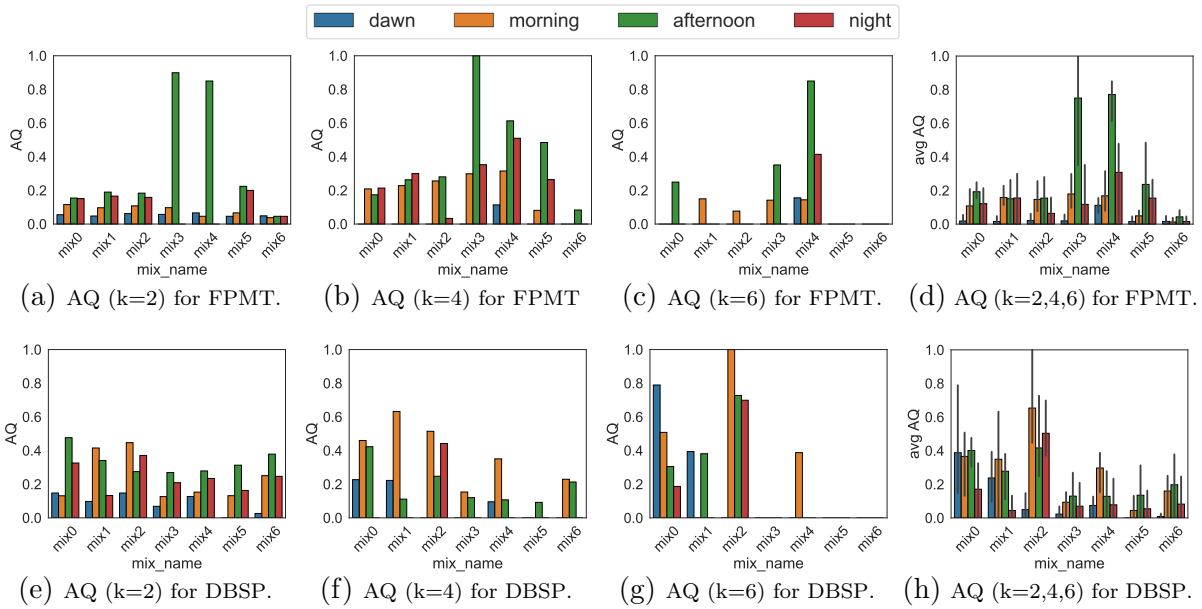
6.5.4 The Utility of Anonymized Data

The results for CODU metric of anonymized data with mix-zone positioned with FPMT, seen in Figure 6.9a, show the collateral effects of anonymization. In all privacy settings, the number of user pairs that made contacts with a duration close to zero increased from 10^5 to 10^6. Also, when $k = 2$, there is a higher distortion to the original dataset than other setups pair, in which the contact duration of 60000 ms had an up from 10^2 to 10^5. The $k4$ and $k6$ had a high distortion but in fewer pairs of contacts, from 10^3 to 10^4.

The CODU metric for mix-zone positioned with DBSP also had a higher distortion than FPMT, case as shown in Figure 6.9b. In all privacy settings, the number of user pairs that made contacts with a duration close to zero increased from 10^5 to 10^6. Furthermore, for $k2$, there was a significant increase in the number of pairs, from 10^3 to 10^5, with a contact duration of up to 70%. This behavior can also be observed in $k4$, but in smaller pairs of contacts, from 10^3 to 10^4. Finally, in the $k6$ the curve follows the noanon more than other anonymization setups, starting with 10^4 pairs of contacts until 10000 secs and finishing with 10^2 pairs of contacts with duration up to 70000 secs.

For the WS metric for the CODU, the non-anonymized trajectories were far from all k value anonymization settings in FPMT algorithm (Figure 6.9c). As a result, the utility had a high distortion of the anonymized dataset about the original trajectories. The significant distortion from the original dataset was $k = 2$ setup, with WS equal to 0.11 (pair $< noanon, k2 >$). The $k = 2$ was the second with significant distortion, and the WS got 0.067. The $k = 6$ setup had less than all setups with WS 0.025. The pair $< k4, k6 >$ had a minor divergence regarding protected data. Both distributions smoothly tended to have many pairs of contact, few duration.

In contrast to FPMT case, CODU for mix-zones positioned with DBSP had less distortion of the anonymized dataset from the original dataset (Figure 6.9d). The significant variation was $k = 2$, with WS 0.087 (pair $< noanon, k2 >$). On the other hand, the $k = 6$ had minor distortion, possibly indicating high utility with a WS of 0.005. Regarding WS between the protected datasets, the pairs $< k2, k6 >$ and $< k4, k6 >$ had significant and slight divergence, with WS equal to 0.082 and 0.021, respectively.



(a) AQ (k=2) for FPMT. (b) AQ (k=4) for FPMT (c) AQ (k=6) for FPMT. (d) AQ (k=2,4,6) for FPMT.

(e) AQ (k=2) for DBSP. (f) AQ (k=4) for DBSP. (g) AQ (k=6) for DBSP. (h) AQ (k=2,4,6) for DBSP.

Figure 6.10: Anonymization Quality (AQ) of mix-zones positioned with FPMT and DBSP.

6.5.5 Anonymization Quality (AQ)

This section presents the results obtained from the analysis of the AQ for the mix-zones positioned with the FPMT and DBSP algorithms. In particular, we present an analysis of the behavior of AQ regarding mobility data protection, privacy attack, and the utility of anonymizing data.

6.5.5.1 AQ Analysis of Mix-zones Positioned with FPMT

The highest AQ of mix-zones positioned with FPMT algorithm for $k = 2$ were the mix3 and mix4, mainly at noon (see Figure 6.10a). They had high AR, ME, and low NAR during these periods. This behavior is reflected in the same NCM and ATM metrics periods, where mix3 and mix4 are highlighted more than others, particularly mix4 got NCM and ATM up to 12 and 10^4 secs, respectively. Additionally, the $\mathcal{I}_s$ is lower than the other periods, indicating more vehicles out of the mix-zones than entering it. Although the mix4 performed better anonymously than other mix-zones, its NTC was high, indicating that many vehicles concluded their trips within the mix-zone. Some of these vehicles may not have been anonymized for this; their AQ was penalized and was lower than mix3.

The AQ ranking of mix-zones #TOP3 for $k = 2$ setups are mix3, mix4, and mix5. Mix3 is on the expressway close to a coastal region with a high tourist flow, and mix4 is

located at the entrance to the airport. Mix5 on the downtown area and cover the 80th expressway that connects to Oakland.

Concerning AQ for $k = 4$, the mix-zones mix3 and mix4 also remained in the afternoon, reaching AQ with a value of 100% and 61%, respectively (Figure 6.10b). The AQ for these mix-zones converged to the ME, with a value of 80% and 88% for mix3 and mix4, respectively. The mix3 and mix4 in the afternoon had longer duration ATMs, around 5000 and 7400 secs, than the other mix-zones with 2500 secs (see Figure 6.10). The ITM and IDM for mix3 in the afternoon were 38.91 and 35.60 secs (see Figures 6.8a and 6.8b), respectively. While for mix4, the ITM and IDM were 38.22 and 34.47. For both mix-zones, they had an $\mathcal{I}_s$ near to zero representing good vehicle flow in these mix-zones (see Figure 6.8c). The other mix-zones obtained an AQ above 20%, highlighting the mix5 with a peak AQ of 48% in the afternoon.

In the $k = 6$ setup, the afternoon period had the best performance, in this time, mix4 performed better than the other mix-zones due to the high volume of traffic at the airport with AQ equal to 85%, followed by mix3 and mix0 with AQ of 35% and 25%, respectively (see Figure 6.10c). Following the result for the mix-zones metrics in the same period, the best ME scores for mix4 and mix3 were 73% and 16% in the afternoon. For the quality metrics, mix4 and mix3 had ATMs of 5433 and 1790 secs. On the other hand, mix5 and mix6 had the AQ zero because these mix-zones did not have the ATMs even with $\mathcal{I}_s$ and NTC near to zero (see Figures 6.8c and 6.8d).

One of the issues with mix-zones is their resilience: The mix-zones behavior in the face of different levels of privacy. So, naturally, the questions arise:

I. What are the mix-zones that present better stability with the variation of k?
II. What mix-zones present better stability in periods of the day?
III. Which mix-zones perform best at all possible k setups?

These are pertinent questions linked to the mix-zones' performance aspects over time. We made the analysis 1 (**An1**) to answer questions I and II. Specifically, we verify the mix-zones performance concern k setup variation per period. Then, for each mix-zone, we calculate the average per period between setups $k = 2, 4$, and 6, $avgAQ_{k=2,4,6}$. Finally, the highest $avgAQ_{k=2,4,6}$ is selected as the mix-zones and period with more performance against the k variation.

To answer question III, we made the analysis 2 (**An2**), in which we calculate the weighted average for each mix-zones, where each weight is a k value as

$$avgAQ_{all_k} = \frac{\sum_{i\in\{2,4,6\}} W_{ki} avgAQ_{ki}}{\sum_{i\in\{2,4,6\}} W_{ki}},$$

where $avgAQ_{ki}$ is the AQ average of mix-zones of all periods for a ki setup $i \in \{2, 4, 6\}$, and W_{k2}, W_{k4}, and W_{k6} are the weights of $k = 2$, 4, and 6, respectively. The high $avgAQ_{all_k}$ is the mix-zones that perform best at all possible k setups.

Table 6.4: AQ weighted average ($avgAQ_{all_k}$) for mix-zones positioned with FPMT and DBSP.

Name	$avgAQ_{all_k}$ **for FPMT**	$avgAQ_{all_k}$ **for DBSP**
mix0	0.1007253185	**0.3610248048**
mix1	**0.1054977890**	**0.2183576170**
mix2	0.0784483852	**0.4551782232**
mix3	**0.2431063586**	0.0508220354
mix4	**0.3648428612**	0.1275244962
mix5	0.0915499554	0.0329825509
mix6	0.0143782553	0.0744508244
avg$[AQ_{mix}]$	**0.1426498462**	**0.1886200789**

In the **An1**, denoted in Figure 6.10d, the AQ for the afternoon prevailed for all mix-zones and showed significant AQ variation between periods. Mix4, mix3, and mix5 had the highest AQ averages in the afternoon with values of 0.79., 0.75, and 0.25. A point to be noted is that mix1 presented the lowest AQ variation in three periods, suggesting a linear activation flow for the morning, afternoon, and night periods. So mix1 tends to perform better than mix5. This linearity is also noticed in the ME metric (Figure 6.4c) for these periods in the $k = 2$ setup. Finally, mix4, mix3, and mix1 perform better per period and k variation. In the **An2** (see Table 6.4), mix4, mix3, and mix1 were highlighted more than others with more performance over the mean all possible k setups, with $avgAQ_{all_k}$ of 0.36, 0.24, and 0.10, respectively.

6.5.5.2 AQ Analysis of Mix-zones Positioned with DBSP

The four mix-zones positioned with DBSP algorithm for $k = 2$ with the highest AQ peak in a period where the mix0, mix2, mix1, and mix6 with AQ equal 47%, 44%, 41%, and 37% in the periods afternoon, morning, morning, and afternoon, respectively (Figure 6.10e). These periods had high AR, ME, and low NAR (Figure 6.6). For instance, mix0, mix2, mix1, and mix6 had ME peaks to 46%, 95%, 86%, and 74%.

Concerning the quality metrics, AQ also follows the NCM and ATM curve for all periods (see Figure 6.7). For instance, the AQ for mix2 in the dawn is 14%, we have a lower value of NCM and ATMs values of 3 vehicles and 3046 secs for the same period, while in the morning, the metrics curve tends up to AQ equal 44%, NCM equal 12 and ATM equal 3593 secs, having climbed in the afternoon with ATM equal 14447 secs, NCM similar 12, and AQ equal 27%. At night, AQ is up to 37%, following NCM and ATM equal 11 and 3476 secs. We can also observe this AQ behavior in specific periods in all mix-zones.

A point to note is that although the mix0 is positioned in the taxi stop inside the

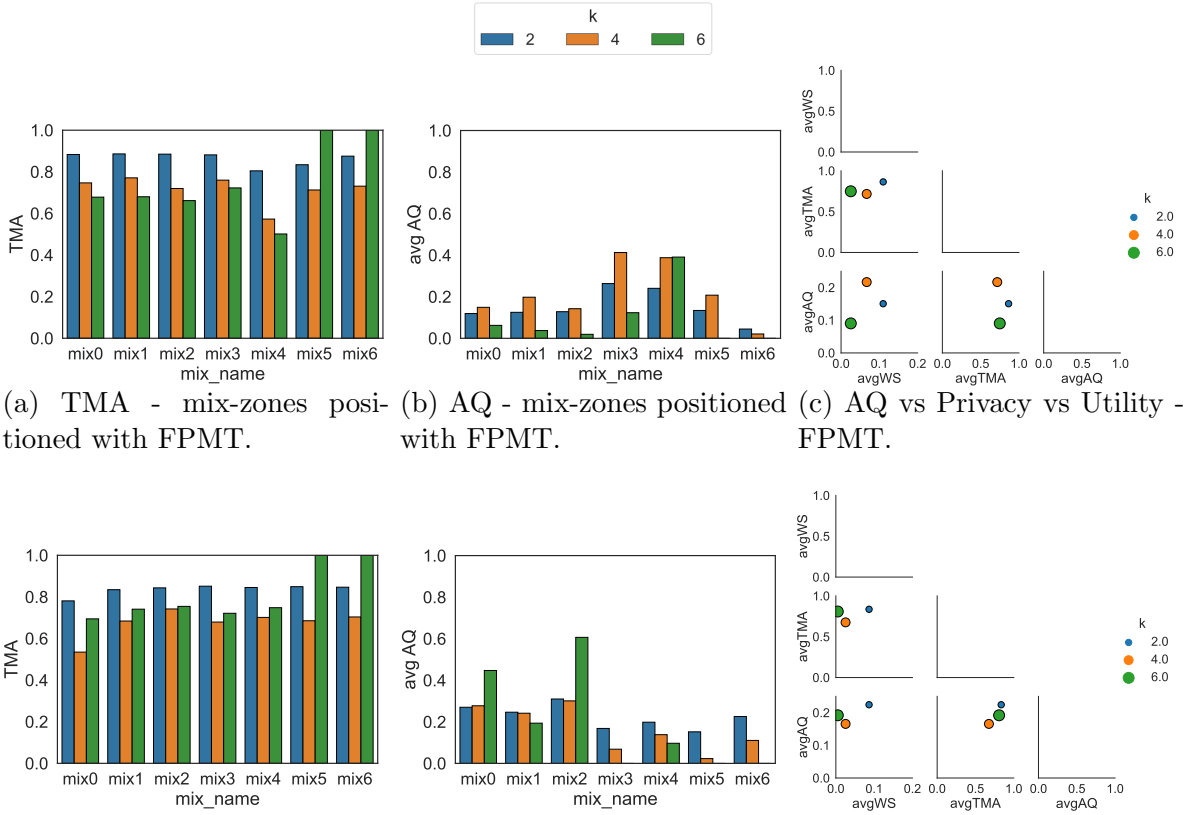
SF airport, a region with vehicle volume, and has the number and duration of ATMs (represented by ϕ in function AQ) higher than all mix-zones, it had no so good performance with lower AR, ME, and a high NAR than mix1 and mix2. AQ captured this behavior, which showed this degradation. The mix0 is justified because it had a NTC higher than all mix-zones (see Figure 6.8h), with a peak of 100 in the morning, which means a significative number of vehicles have completed their trips inside mix-zones and could not be anonymized, so the AQ score was penalized.

The ranking of mix-zones #TOP3 that had more AQ average for $k = 2$ setups were mix2, mix0, and mix1 with average AQ of 30%, 27%, and 24%. The average AR and ME for mix2 were 359.5 and 83.2%, mix0 were 166 and 41.3%, and mix1 had 321 and 75.7%. Mix2 is on the expressway close to a coastal region with a high tourist flow. Mix0 is located inside the SF airport. Mix1 is located in a return of Expressway 101 near the downtown.

In the AQ analysis for $k = 4$, the #TOP3 ranking as $k = 2$ was maintained to mix2, mix0, and mix1 with average AQ equal 30%, 27%, and 24.6%, respectively, with peaks at morning AQ = 51%, 45.9%, and 63% for each mix-zone (Figure 6.10f). Also, the AQ followed the coverage metric, such as the ME for mix2, mix0, and mix1 were 39.8%, 30.7%, and 11%. Concerning the AQ vs. ATM, the mix0 had a higher number and duration, which favored the first position of mix2 at the top, with average ATMs with 1920.58 secs, against 1759.58 secs of mix0 and 1708.4 secs of mix1. Mix3, mix5, and mix6 had lower AQ and ME, with AQ peaks of 12%, 9%, and 23%, respectively (see Figure 6.10f and 6.6f).

In the $k = 6$ setup, just four mix-zones had AQ mix2, mix0, mix1, and mix4 with AQ peak of 100% in the morning, 79% in the dawn, 39.5% in the dawn, and 38.7% in the morning and average AQ of 60.5%, 44.7%, 19%, 9.6%, respectively (see Figure 6.10g). Mix3, mix5, and mix6 did not provide anonymization, and AQ was equal to 0. AQ also follows the quality metrics, electing the #TOP3 as the three mix-zones like $k = 2$ and $k = 4$ setups. In the mix-zones election with ATM metric, we achieved the same result of AQ, where mix2 had six contiguous ATMs between 1600 and 1800 secs, mix0 one ATM of 3145 secs, and mix1 two ATMs of 1469 (dawn) and 1763 secs (afternoon) as seen in Figure 6.7.

In the **An1**, Figure 6.10h, all mix-zones had high AQ averages during the day (morning and afternoon). For the morning, it prevailed for mix2, mix1, and mix4 with AQ averages of 67%, 37%, and 30%, while mix0, mix3, mix5, and mix6 had the high AQ for the afternoon, with AQ averages of 40%, 15%, 17%, and 20%, respectively. Mix0 showed more significant AQ variation between the periods because mix0 deployed inside the airport; it is a region with a high traffic variation over time that affects quality metrics (see Figure 6.7b). In the **An2** (see Table 6.4), mix2, mix0, and mix1 were highlighted with better performance than other mix-zones, with AQ values of 0.45, 0.36, and 0.21.



(a) TMA - mix-zones positioned with FPMT.

(b) AQ - mix-zones positioned with FPMT.

(c) AQ vs Privacy vs Utility - FPMT.

(d) TMA - mix-zones positioned with DBSP.

(e) AQ - mix-zones positioned with DBSP.

(f) AQ vs Privacy vs Utility - DBSP.

Figure 6.11: AQ, TMA, and WS for mix-zones protected with $k = [2, 4, 6]$ and positioned with FPMT and DBSP algorithms.

6.5.6 AQ vs. Privacy vs. Utility

This section presents the relation of AQ vs. Privacy vs. Utility for mix-zones positioned with the FPMT and DBSP algorithms. In this analysis, we explore the capacity of identifying the best k setup to maximize quality and privacy without affecting the utility of the trajectory data. Recall that TMA $\in [0, 1]$ is average of Trajectory Matching Accuracy (avgTMA), an attack score where avgTMA equal to 1 is the max attack success; avgAQ $\in [0, 1]$ represents the average of Anonymization Quality, where avgAQ equal to 1 means the max quality; avgWS is average of Wasserstein metric, an utility level of the dataset where avgWS near to zero is maxing utility.

6.5.6.1 Anonymization Quality and Privacy

In the AQ vs. Privacy analysis, we explore AQ behavior in re-identification attacks against mobility data protected with different privacy levels. The goal is to verify its capacity to identify the best mix-zones against re-identification attacks varying the k privacy level, see Section 6.4.4.

Figure 6.11a represents the analysis for mix-zones positioned with FPMT. We show that mix-zones with high AQ, k (such $k = 6$), and some ATM values present the lowest TMA of all configurations , e.g., mix0, mix1, mix2, and mix4 with 67%, 68%, 66%, and 50% (see Figure 6.11b). In contrast, the mix-zones with AQ equal to zero present TMA score of 100%, such as mix5 and mix6. High TMA also can be identified by the absence of ATM (see Figure 6.10).

Another point to note is that mix5 and mix6 have a higher TMA for $k = 4$, $k = 6$ than for $k = 2$. TMA difference between setups occurs because the vehicle flow is more sparse in these mix-zones, as evidenced by the high ITM and IDM averages (Figures 6.8a and 6.8b). Furthermore, as k increases, AR, ME decreases significantly. Under these conditions, for $k = 6$, the chances of anonymization are lower than $k = 2$, which results in a higher TMA for $k = 6$.

The dataset protected with mix-zones positioned with DBSP had similar behavior to mix-zones placed with FPMT in terms of AQ and privacy (Figure 6.11d). In which mix-zones with high AQ (Figure 6.11b) and some ATM present low TMA, e.g., AQ of $k = 4$ setup for mix0, mix1, and mix3 have TMA average of 0.53, 0.68, and 0.67, respectively. For the mix-zones with AQ equal to zero present high TMA, such as mix5 and mix6 with $k = 6$. In both cases of protection that use positioning algorithms, we can observe that the AQ behavior is inverse to the attack score.

6.5.6.2 AQ vs. Privacy vs. Utility for mix-zones positioned with FPMT

Regarding privacy vs. utility, in the cell [avgWS, avgTMA] in Figure 6.11c, the dataset protected with $k = 2$ had the highest TMA (avgTMA = 0.86) and data distortion (avgWS=0.10) of the dataset than $k = 4$ and $k = 6$. However, in the $k = 4$ setup, avgTMA decays to 0.71, and the utility improved, decaying to 0.06. For $k = 6$, the TMA rises to 0.74, indicating a greater privacy risk; however, there was less data distortion than $k = 2$ and $k = 4$ with avgWS=0.02. We must consider the coverage for $k = 6$, in which few vehicles are anonymized (see Figure 6.6i). In conclusion, avgWS vs. avgTMA, the best

choice is the $k = 4$ setup.

In the quality vs. utility analysis, the cell [avgWS, avgAQ], the $k = 2$ setup had an avgAQ intermediary between the setups around 0.15 and a high distortion level possible with avgWS equal to 0.10. In contrast, the $k = 4$ had the highest avgAQ=0.21, and the distortion level decays to avgWS=0.06. $k = 6$ had minimal data distortion (avgWS=0.02) but had a lower AQ of all setups with avgAQ equal to 0.09. Regarding AQ vs. utility, the best result is $k = 4$ again.

In the privacy vs. quality analysis ([avgTMA, avgAQ] cell), the $k = 2$ setup presented worse performance than the other k setups regarding privacy. The worst performance is because the TMA reached 0.86 and the AQ intermediary of 0.15, but it represents that the dataset is vulnerable to re-identification attacks. Furthermore, the $k = 6$ was down avgAQ to 0.09, but TMA had reduced to 0.75. On the other hand, the $k = 4$ setup had a better setup performance - presented good anonymization quality, avgAQ=0.21 against 0.15 of $k = 2$ and 0.09 of $k = 6$, and privacy level with avgTMA of 0.71 against 0.86 and 0.75 of $k = 2$, $k = 6$, respectively.

6.5.6.3 AQ vs. Privacy vs. Utility for mix-zones positioned with DBSP

In the DBSP algorithm case, privacy vs. utility, depicted by the cell [avgTMA, avgWS] in Figure 6.11f, the dataset protected with $k = 2$ had the highest TMA (avgTMA=0.83) and data distortion (avgWS=0.08) of the dataset than $k = 4$ and $k = 6$. Nevertheless, in the $k = 4$ setup, avgTMA got 0.67, and the utility improved, decaying to 0.02. For $k = 6$, the TMA increased to 0.80, indicating a greater privacy risk; however, there is a better utility for the data with an avgWS equal to 0.005, meaning little distortion on the dataset. However, like the FPMT case, the $k = 6$ setup yields low coverage, and few vehicles are anonymized (see Figure 6.6). In conclusion, for avgWS vs. avgTMA, the $k = 4$ setup is the best choice.

The quality vs. utility, the cell [avgWS, avgAQ], the $k = 2$ setup had the best avgAQ equal to 0.22, but a high distortion level of 0.06. For $k = 4$, the avgAQ slightly fell concerning $k = 2$, with an avgAQ equivalent to 0.16. After that, however, the distortion level degenerated to avgWS=0.02. The setup $k = 6$ had the minor data distortion possible but equally low AR and ME (see Figure 6.6); consequently, with the worst AQ of the three setups, the avgAQ equal 0.19. Regarding quality vs. utility, the best setup is $k = 4$.

Finally, for the privacy vs. quality analysis (the cell [avgTMA and avgAQ]), the $k = 2$ setup presented better performance than the other k setups regarding quality, with avgAQ equal to 0.22, but not a good privacy level. The high TMA value of 0.83 indicates

vulnerability to re-identification attacks. The $k = 6$ got the lower avgAQ of all setups, but TMA reduced to 0.80 about $k = 2$. The setup $k = 4$, although it had presented a lower anonymization quality with avgAQ=0.16 than other setups, had the best protection, which got a privacy level with an avgTMA of 0.67 against 0.83 and 0.80 of $k = 2$ and $k = 6$, respectively. In conclusion, the $k = 4$ setup is the best option for privacy vs. quality analysis.

6.5.6.4 What is the best Positioning Algorithm?

Choosing a positioning algorithm from a set, based on the trade-off between the requirements: AQ, privacy, and utility, is a challenging task because it depends on the user's objectives who will consume the protected dataset. For example, there are cases where the priority is the privacy of mobility data, even if it compromises its utility or vice versa. However, to answer this question (question **Q5** denoted in Section 6.3), we will consider an open data scenario where a mobility dataset must be published to design dissemination protocols in VANETs, in which it is necessary to maximize the three requirements: privacy, utility, and anonymization quality [261, 269]. In our case, these requirements are represented by the minimum of avgTMA, avgWS, and the maximum of avgAQ, respectively, and it expects that when comparing the two positioning algorithms, the winner should have the best result between the corresponding cells of Figures 6.11c and 6.11f. Specifically, for cell [avgWS, avgTMA], we expect the values [minimum, minimum]. For cell [avgWS, avgAQ], we expect the pair [minimum, maximum], and for [avgTMA, avgAQ], we expect the tuple [minimum, maximum].

For $k = 2$, we can see that DBSP stands out in the three variable analysis cells, where [avgWS, avgTMA] in relation to FPMT. For the $k = 4$, the DBSP performs better than FPMT in utility and privacy but has lower performance in AQ (see cells [avgWS, avgAQ] and [avgTMA, avgAQ] of Figures 6.11c and 6.11f). Finally, for $k = 6$, DBSP performs better than FPMT in terms of AQ and utility (see cell [avgTMA, avgAQ]) but performs worse in privacy (see cell [avgWS, avgTMA]). Finally, in this open data scenario, the results suggest that the DBSP algorithm performed slightly better than FPMT in the three k setups in at least two guidelines, for example, privacy and utility. However, this is not a rule, and it is up to the data consumer to define the main guidelines to be adopted.

6.5.6.5 Expected Behavior of the AQ with Packet Loss and Mix-zone Radius Tuning

Packet loss in VANETs is a critical aspect that we must consider and will undoubtedly impact the behavior of AQM. Many aspects of the architectural viewpoint must be analyzed. In this section, we discussed the expected behavior of AQM in a VANET environment with packet loss in a V2X communication. Specifically, the communication between a vehicle and the RSU (see Figure 2.17b) in a Dedicated Short Range Communication (DSRC) Protocol architecture [270].

Based on the layered architecture of DSRC protocol, the AQM stays in the application layer having the TCP or UDP as the subjacent transport layer. Let a scenario of packet loss in a V2X communication between a vehicle and RSU. If AQM is over TCP, the TCP itself can handle the packet loss with retransmissions of the location points. However, AQM is over UDP; certainly, we have packet loss, and we must consider three scenarios. The first and worst scenario is when the mix-zone does not detect the cars through for it. In this case, the ATM metric tends to zero, so it is expected that AQ tends to zero too, although having cars crossing mix-zone M. The second scenario occurs when cars send packets to RSU when entering M, and packets are lost when cars are out of M. In this case, although they have high ATM and NCM, in contrast, it expects that AQ will be penalized because $\mathcal{I}_s$ and NTC can be high too. After all, the behavior is conducted by the cars that end their trips at M. Third scenario is when cars send packets to RSU when staying inside at M. In this case, the M will detect the cars and work normally; ATM and NCM are high, but $\mathcal{I}_s$ and NTC are low; consequently, AQ will be higher than in previous scenarios.

Another factor that can impact the AQ is the mix-zones radius. Likewise, the privacy level k, the radius, is a factor for coverage of the mix-zones and can also impact AQ. For example, for the ATM metrics, part of the AQ function, with a mix-zone set up with little radius, is more likely that ATM is low too, the reason of coverage area is reduced. So, we will expect to have a smaller AQ. In contrast, with a significative radius coverage, we expect to be a high ATM.

6.5.7 Lessons Learned

We learned some lessons from the analysis of Anonymization Quality in finding the answer to questions of Section 6.1. The answer to question Q1 is that more than the

traffic flow is needed to elect mix-zones. The positioning algorithms analyzed here have the principle of selecting mix-zones by higher traffic in descending order. So technically, mix0 has more traffic than mix6. However, we have shown that certain mix-zones that were not the first can perform better than the one with the highest traffic, and this one has a high AQ. For example, mix3 and mix4 positioned with FPMT performed better than their predecessors mix0, mix1, and mix2. This fact is also evidenced in the case of DBSP in which mix2 obtained a better performance and AQ than mix0 and mix1.

The answer to question Q2 is the nearby mix-zones may have different behaviors and AQ. For instance, mix4 deployed with FPMT, and mix0 positioned with DBSP. Although both are in the airport region, they perform differently. Mix4 is located on the access road to the airport where taxis cross and has high AR and ME. Mix0 is located inside the departure and arrival areas where taxis end their trips, and consequently, they are not anonymized and have a lower AR and ME than the other mix-zones. AQ could detect and consider such situations, including performance variation over time in both cases.

In response to question Q3, AQM and the analysis of its quality metrics identified subtle behaviors of the mix-zones over time, situations in which conventional metrics cannot measure. As is the case with the mix-zones close to the San Francisco airport positioned with FPMT and DBSP.

Concerning question Q4, to perform anonymization quality, mix-zones need a good traffic flow, an Inflow equal to or less than the outflow of vehicles, with little retention of vehicles inside it, and that anonymizes when activated over time. Mix4, mix3, and mix1 positioned with FPMT are examples of AQ with the best performance per period and k variation, as evidenced by the AQM.

Regarding question Q5, the AQM framework could identify the positioning algorithm that had a better performance from the trade-off between the AQ vs. Privacy vs. Utility requirements, such as having elected the DBSP as the more acceptable positioning algorithm. However, we must consider the dataset's customer goals and, from this, deliberate about the main requirements to be adopted of him.

About question Q6, the Anonymization Quality is a new perspective on how we see anonymization. Here we explored their potentialities with positioning algorithms and the election of potential mix-zones. However, many other applications can use the AQM. For instance, in the smart mobility context, the AQM can be used for online monitoring of mix-zones helping the operator to discover and enable mix-zones with good performance and disable mix-zones with not a good performance at certain times of the day. Also, verify if a mix-zone is working well. Furthermore, in the open data context, AQM can verify the data quality regarding anonymization quality, privacy, and utility in publishing trajectory data.

Finally, the lessons we learned from the analysis with mix-zones metrics were:

- NCM is proportional to the AR and ME, as k increases. Because if k is low, more chances of NCM over time are higher than k. Thus, more vehicles pass by the mix-zone and are anonymized, yielding high AR and ME rates. In contrast, if k is high, it is likely to have a low AR and ME because it has more chances that NCM is less than k. Nevertheless, the data are anonymized and have a high privacy level.

- ITM, IDM, and NTC are inversely proportional to the anonymization likelihood. If we have an ITM and IDM high, it indicates that the vehicles are coming in and out of the mix-zone in the sparse time of each other. This can mean vehicle traffic (NCM) is lower than k; consequently, no anonymization occurs. A high NTC means a high vehicle concentration inside the mix-zone, and they are not outing it. So if the pseudonym change policy is used when a vehicle goes out of the mix-zone, no anonymization occurs, too.

- For AR, the longer ATMs are better than shorter and adjacent ATMs, which in turn is better than short, scattered ATMs. Longer ATMs are the best for anonymization because they attend the condition NCM $\geq k$ for a high time interval that enables anonymization. The adjacent ATMs mean NCM $\geq k$ but shorter than longer ATMs periods indicating less anonymization likelihood. Finally, scattered ATMs is the situation in which we have NCM $\geq k$ of less period of all ATM types indicating the lowest anonymization likelihood.

- How much increase value of k, the value of ATM decrease. Because the higher the k, the lower are chances of satisfying the condition NCM $\geq k$ that impacts the ATM, having less likelihood of enabling the mix-zone for anonymization.

- Given the k vs. NCM, ITM, IDM, NTC, and ATM analysis, it is possible to elect mix-zones with good privacy level, anonymization rate, and activation time about the candidates. These metrics are a fundamental pillar of anonymization quality.

- AQ showed a new perspective on anonymization based on the operating principles of mix-zones, reflecting its performance over time in terms of quality, privacy, and utility.

6.6 Concluding Remarks

In this chapter, we investigated the anonymization quality in the context of mix-zones. We raised gaps in mix-zones approaches and questions about how the literature

addresses these issues. We noticed that mix-zones that use just the traffic flow information are insufficient to achieve the anonymization quality. For this purpose, we proposed an anonymization quality framework and the quality metrics Number of Cars in Mix-zone, Interval of Arrival Time between Cars on Mix-zones, Interval of Departure Time between Cars on Mix-zones, Number of Trips Completed within the Mix-zone, and Activation Time of the Mix-zone. We conducted experiments with a cab mobility dataset and two positioning algorithms to explore the potentialities of the anonymization quality to elect mix-zones that do not consider only the traffic but its operating requirements, too. Furthermore, we explored the relationship of anonymization quality with coverage, privacy, and utility in a cab mobility dataset. The results enabled identifying particular anonymization concerns that coverage metrics cannot identify. Additionally, the framework enabled the selection of mix-zones that yield data anonymization considering the quality, privacy, and utility analysis. To our knowledge, this is the first study that analyzes mix-zone coverage and quality metrics to observe the anonymization quality. Anonymization quality will serve as a base for developing dynamic and more robust anonymization approaches family. For instance, for mix-zones positioning and selection algorithms, LPPMs based on k-anonymity and pseudonyms swap do not just consider the traffic flow.

In the next chapter, we explore a new perspective on the utility of anonymized mobility data for smart cities. To this end, we propose a framework for identifying the best smart city domains, applications, and services that best leverage anonymized mobility data.

Chapter 7

Utility Analysis of Anonymized Mobility Data for Smart City Applications

When designing smart cities' building blocks, mobility data plays a fundamental role in applications and services. However, mobility data usually comes with unrestricted location of its corresponding entities (e.g., citizens and vehicles) and poses privacy concerns, among them recovering the identity of those entities with linking attacks. LPPMs based on anonymization, such as mix-zones, have been proposed to address the privacy of users' identities. Once the data is protected, a comprehensive discussion about the trade-off between privacy and utility happens. However, issues still arise about the application of anonymized data to smart city development: what are the smart cities applications and services can best leverage mobility data anonymized by mix-zones? In this chapter, we propose the Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT) to answer this question. This characterization framework measures the utility through twelve metrics related to privacy, mobility, and social, including mix-zones performance metrics from anonymized trajectories produced by mix-zones. This framework aims to identify applications and services where the anonymized data will provide more or less utility in various aspects. The results evaluated with cabs and privacy cars datasets showed that further characterizing it by distortion level, UAFAT ranked the smart cities application domains that best leverage mobility data anonymized by mix-zones. Also, it identified which one of the four case studies of smart city applications had more utility. Additionally, different datasets present different behaviors in terms of utility. These insights can contribute significantly to the utility of both open and private data markets for smart cities.

This chapter is organized as follows. Section 7.1 brings an introduction and motivation to discuss the utility of anonymized mobility data by mix-zones for smart city applications and services. Sections 7.2 and 7.3 present the related studies and describe concepts essential to the utility of anonymized mobility data, respectively. Section 7.4 introduces the framework of utility analysis and details of their metrics. Section 7.5

presents the results of this work. Finally, in Section 7.6, we present the final remarks of this chapter.

7.1 Introduction

In the development of smart cities, mobility data is a substantial aspect of designing applications and services. However, this data has massive and unrestricted location data of citizens and vehicles that pose a privacy concern [173]. By mining such data, it is possible to identify latent information about places of interest, individual and collective habits, and even the identity of citizens. The privacy issues of mobility data, including the identity of citizens, have become a concern even more with the substantial adoption of data protection standards such as Data Protection Regulation (GDPR) [271]. To address the privacy of users' identities, LPPMs based on anonymization have been proposed, such as mix-zones. Specifically, mix-zones are designated anonymization areas defined by a radius r where entities change their pseudonyms according to a trigger function, such as when reaching a minimum of k entities simultaneously inside it [170, 167].

Once protected the data, a comprehensive discussion has been done about the trade-off between privacy and utility in the broad sense. Nevertheless, there are open questions about the anonymized data utility for developing smart cities in both open data and privacy market contexts, such as: what are the smart city applications and services that can best leverage mobility data anonymized by mix-zones? The answer to this complex question is not easy because the utility of anonymized datasets must be analyzed in various contexts, domains, and applications, which can be characterized as a decision-making problem that leads us to more questions:

Q1 What are the contexts of use of the dataset anonymized by mix-zones that are most and least affected in terms of utility?

Q2 Can different versions of a dataset, each anonymized by a different privacy level (k), differ in ranking metrics at the distortion level?

Q3 Which metrics have more and less distortion between the original and anonymized data with a variation of k?

Q4 Do datasets of different types present the same behavior in terms of utility?

Q5 What is the ranking of smart city application domains with high utility in the anonymized mobility data by mix-zones?

Q6 What is the utility of anonymized mobility data for a specific application of a smart city domain?

To answer the questions above, we investigate the utility aspects of anonymization data for smart cities. For this, we propose the Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT). This characterization framework analyzes the utility through twelve metrics related to privacy, mobility, and social, including mix-zones performance metrics from anonymized trajectories produced by mix-zones. Specifically, UAFAT aims to identify applications and services where the anonymized data will provide more or less utility in various aspects, such as:

- To identify the utility level of the anonymized datasets for specific smart city applications and compare each other;
- To compare the utility of the mobility data based on privacy level for applications and domains;
- Generating a ranking of smart city application domains that best leverage mobility data anonymized by mix-zones with our proposal of a Multi-Criteria Decision-Making algorithm based on utility metrics for automatic criteria weight;
- It assists in deciding which anonymized mobility dataset is most useful for an application.

Another point to note is that the utility analysis considers data distortion to be a utility, and the mix-zones performance metrics, such as efficacy. Thus, the utility calculation becomes more robust. It considers both the levels of distortion and coverage of anonymized data, even allowing the performance of anonymization mechanisms, such as mix-zones, to be evaluated.

We comprehensively evaluated the UAFAT in utility terms with real taxicabs and private cars datasets protected with three privacy levels. The results suggest that anonymized trajectories from the cabs dataset have more utility for applications associated with social metrics, such as dissemination protocol design, than urban planning and POI mining. Also, the framework identified social networks, opportunistic networks, and statistical analysis as the top three smart cities application domains that best leverage the anonymized cabs dataset with a privacy level. In another experiment that applied the anonymized dataset to four specific applications, the application of model mobility simulation from the opportunistic networks domain got the best leverage of the anonymized trajectory cabs dataset. Finally, the cabs dataset had a significant utility compared to the private cars dataset.

To the best of our knowledge, this work is the first proposal to analyze the utility of anonymized trajectories by mix-zones and identify the smart city application segments according to privacy, mobility, and social metrics. The main contributions of this work are:

Table 7.1: Contributions of each proposal, their application domains, and utility analysis.

			Utility Analysis			
Author	**LPPM**	**Contribution**	**Utility-related metrics** [a]	**Application Domains** [b]	**Ranking Multi-domain**	**Specific Application**
[18]	anon.	Framework to shape the beginning of the data monetization decision-making process for vehicle manufacturers.	POIs, TDSL	TM	No	No
[16]	anon.	Framework to link data-driven human mobility research with the potential implementation of smart city developments.	Cost vs utility	SA, UP, DB, SN, TM.	No	No
[17]	anon.	Issues about sharing and protecting sensitive personal information, focusing on mobile phone signaling big data in smart city planning.	POIs	SA, UP, DB, TM	No	No
[272]	anon.,obf., mix-zones	Survey about the impacts of LPPMs, includinig mix-zones, on vehicular applications.	POIs, CODU, INCO	PS	No	No
[15]	obf.	Local differential privacy approach that preserves the data for statistics purposes.	POIs	SA	No	Yes
[273]	obf.	Individual fairness metrics of mobility models.	MPA	UP	No	No
[167]	mix-zones	Social Mix-zones, a mix-zone architecture designed to protect contact tracing data.	SPC	SA, UP	No	No
[173] [20]	mix-zones	Anonymization quality framework for mix-zones enabling evaluating the impacts of anonymization over time and space in mobility data, considering coverage, privacy, quality, and utility metrics	CODU	ON, PS	No	No
our work	mix-zones	Utility analysis framework of anonymized trajectories for smart cities-domains based on privacy, mobility, and social metrics.	PSEU, RRET, SPC, SPD, TDSL, TRIPTIME, NUMVIS, MAXCON, CODU, INCO, CONEN, RGYR	SA, UP, DB, SN, ON, TM, PS	Yes	Yes

[b] Statistical Analysis (SA), Urban Planning (UP), Driver Behavior (DB), Social Networks (SN), Opportunistic Networks (ON), Targeted Marketing (TM), Privacy & Safety (PS).
[a] POI, Pseudonyms per User (PSEU), RRET, SPC, SPD, Travel's Distance Straight Line (TDSL), Trip Time (TRIPTIME), Number of Visits per User (NUMVIS), Maximum of Connections between a User Pairs (MAXCON), CODU, Inter-contact Time (INCO), Contact Entropy (CONEN), Mobility Prediction Accuracy (MPA).

- A framework for the utility characterization for mobility data anonymized by mix-zones;
- An evaluation of smart city applications and services with privacy, mobility, and social metrics to identify utility levels for these applications;
- An extensive evaluation of the proposed framework and their metrics extracted from anonymized mobility data by mix-zones from taxicabs and private cars.

7.2 Related Studies

Analyzing the utility properties of smart cities' mobility data can reveal various insights for data usage, both to open data and monetization. Thus, an effort exists to link the data and their practice. Following, we highlight some relevant literature proposals that pursue these issues.

Ridder et al. [18] investigated the potential business models that can effectively monetize IoT data in smart mobility. Specifically, it discussed the advantages and challenges of different business models, such as data aggregation, data brokerage, and monetization through value-added services. Also, it showed the importance of using public data marketplaces to accommodate most possible use cases and data business models.

Wang et al. [16] proposed a research framework to link data-driven human mobility research with the potential implementation of smart city developments. The framework comprises a systematic review of human mobility with big data, a policy review, and an analysis of smart city development. The framework is applied to various governmental

departments in Hong Kong. The results showed some insights into data-driven research and the development of the smart city.

Lin et al. [17] discussed issues that might arise over sharing and protecting sensitive personal information, focusing on mobile phone signaling big data in smart city planning. They analyzed the de-anonymization technology of mobile phones signaling big data and concluded that only anonymization techniques are insufficient to protect mobile data.

Kim et al. [274] did a comprehensive survey about mechanisms for protection. They divided the location protection mechanisms into three categories depending on the nature of their algorithm and compared them from the viewpoints of architecture, privacy, computational overhead, and utility. However, the utility analyzed is limited in distortions of location points with obfuscation techniques.

Concerning the utility analysis of protected data in the VANETs context, Mdee et al. [272] focused on the effects of LPPMs on vehicular applications. They identified that data protected by online LPPMs-based anonymization, like mix-zones, could affect the Quality of Services (QoS), including communications, computational, and storage overheads caused by frequently changing pseudonyms of vehicles.

Regarding the trade-off between privacy and utility using obfuscation mechanisms, Alvim et al. [15] proposed a variant of local differential privacy based on the notion of d$\mathcal{X}$-privacy, which not only can be used for real-time punctual applications; it could also be utilized to protect privacy when collecting data for statistical purposes, such as POI mining. Also, they claimed that if the statistics are distance-sensitive, then d$\mathcal{X}$-privacy preserves the utility of the data better than the standard Local differential privacy methods.

Zhan et al. [273] extended from the notion of fairness in broader machine learning literature. They measured and evaluated the individual fairness of privacy-utility in the location privacy-preserving algorithms applied to mobility traces. Specifically, they proposed a set of fairness metrics designed explicitly for human mobility, based on structural similarity and entropy of the trajectories and evaluated with two state-of-art privacy-preserving models that rely on GAN and representation learning with two real mobility datasets. They concluded that neither privacy algorithm guarantees individual fairness.

Our previous work [173, 167, 20] showed a practice view of the collateral effects of anonymized data by mix-zones and its derivation. In [167], we proposed the Social Mix-zones, a mix-zone architecture designed to protect contact tracing data. As utility analysis, we used the metric where users stayed for some time, called Stay Point Count (SPC). In [173], we did initial studies focused on the quality of the trajectory data protected by mix-zones. For this, we characterized and evaluated the impacts of anonymization quality over space and time in mobility data with metrics related to mix-zones operation, privacy, and coverage. Then, we extended it, and we proposed an anonymization quality framework in which we analyzed the relationship between privacy, quality, and utility of

anonymized data [20].

Despite the vast literature addressing the utility of trajectory data in smart cities, only a few investigate the impact of anonymization-based LPPMs – like mix-zones – regarding utility in both open data and monetization contexts. There are significant contributions of surveys and research frameworks to study data-driven human mobility and the trade-off between privacy and utility as proposed in [16, 17, 272, 15, 274]. Our previous proposals brought new directions about the utility of anonymized data focused on the application and anonymization quality but are limited to only one utility metric and one application domain. In this way, a practical view of data privacy and security must be explored.

In this chapter, unlike previous studies, we advance the state-of-the-art w.r.t. practical aspects of utility. We propose the UAFAT for characterizing and evaluating mix-zones' impacts on anonymization data and their use by different applications and smart cities application domains. Thus, it is possible to understand and identify applications where the anonymized data by mix-zones can be more useful, contributing to monetization and open data context. The UAFAT was built with twelve metrics based on privacy, mobility, and social metrics. We conducted a comprehensive evaluation of UAFAT of the capability of ranking the utility data by application domains and utility analysis for specific applications with two algorithms and using mobility datasets of real taxicabs and private cars. The results show that the proposed framework identifies and ranks the application domains of smart cities in which anonymized data can have more or less utility. Additionally, identifying datasets of different types presents different behaviors in terms of utility. To the best of our knowledge, no previous proposal analyzed the utility of anonymized trajectories in many perspectives, such as different metrics, applications, and domains-driven. Table 7.1 summarizes this discussion.

7.3 Mix-zones and Utility Problem

This section describes the anonymization problem regarding data utility protected by mix-zones.

Mix-zones is a k-anonymity-based approach widely used in VANETs to pseudonym changing that yields anonymization of users (more details, see Section 2.5.3). When a moving vehicle goes inside a mix-zone with radius r, its trajectory will be sliced into two sub-trajectories delimited by different pseudonyms – one corresponding to the part before the mix-zone and the other corresponding to the region after the mix-zone. Vehicles change their pseudonyms inside the mix-zone if at least k vehicles are present simul-

taneously [164, 172]. A vehicle can travel through multiple mix-zones in its path and, consequently, has its pseudonym changed various times, resulting in several sub-paths delimited by different pseudonyms. In the data publishing context, mix-zones present collateral effects that can compromise the data utility. First, the mix-zones can slice the trajectory data in half in the worst case. Second, mix-zones have a silent period in which the vehicles inside them can not generate location registers, yielding data gaps in those regions. Both effects can significantly influence the utility and depend on the mix-zones parameters' calibration.

However, analyzing the utility of anonymized data goes beyond examining the parameter tuning of a technique or using anonymized data in a generalized form. The utility of anonymized mobility data must be explored in a dedicated way to application classes that will use this data, as shown in Figure 7.1. Based on data analysis and its owner decisions, it identifies aspects of the dataset with metrics capable of measuring utility, such as data distortions. From metrics analysis, it defines which application domains will best take advantage of this data. Concerning the utility of anonymization mobility data, the questions as mentioned in Section 7.1 arise.

To answer these questions, we need to understand the utility of anonymization data. For this, we offer a characterization framework that identifies and measures the utility type with privacy, mobility, and social metrics in anonymized trajectories produced by mix-zones. This framework identifies which use application domains the anonymized data may have the most and least utility. This proposal is a precursor study that analyzes the utility of anonymized trajectories by mix-zones and identifies the application segments with privacy, mobility, and social metrics.

7.4 Data Utility Analysis Framework for Smart City Applications

This section introduces the Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT). We present its metrics and algorithms for utility analysis of anonymized mobility data for smart city applications and their domains.

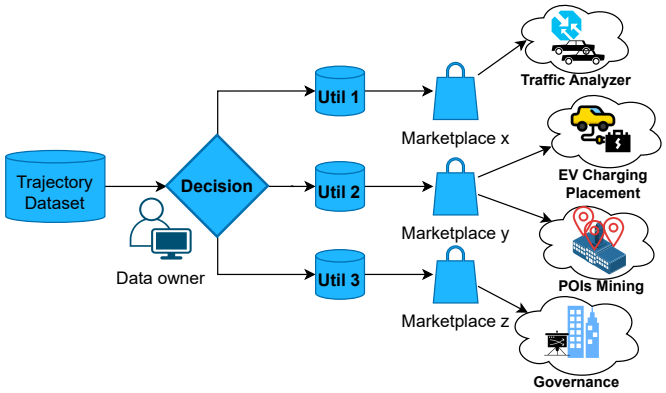


Figure 7.1: The relation between utility types of a trajectory dataset and consumption of this data.

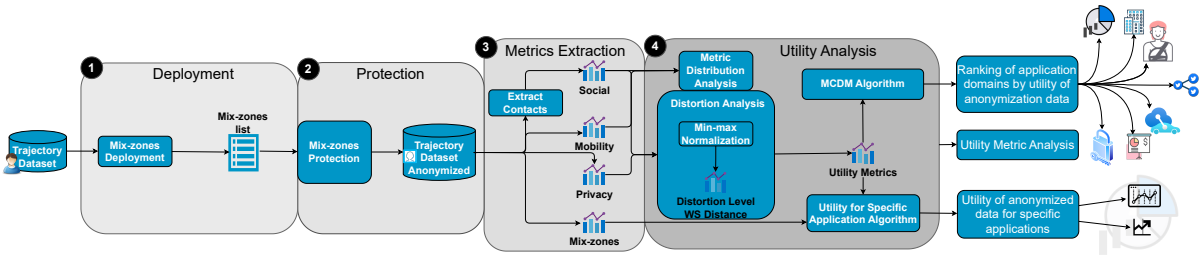


Figure 7.2: Utility framework for anonymized mobility data.

7.4.1 Framework

UAFAT aims to measure the utility of anonymized mobility data for Smart City domains, including their applications. Specifically, it enables the identification of a ranking of smart city application domains sorted by utility level of anonymized mobility data. Also, it's possible to identify the utility level of the anonymized datasets for specific smart city applications. Additionally, it is possible to compare the utility of mobility data based on privacy levels for applications and domains. Furthermore, it assists in deciding which anonymized mobility dataset is more useful from a set of anonymized mobility datasets in utility terms for an application.

The first step in Figure 7.2, Deployment, involves positioning the mix-zones with a deployment algorithm. A sample of the trajectory dataset $\mathcal{D}$ is the algorithm's input, and the output is a list of mix-zones M. In the second step, Protection, the dataset is protected using the previously deployed mix-zones M list, set up with radius r and k values. After the anonymization step, we have the anonymized dataset $\mathcal{D}'$ and mix-zones' metrics, such as (no) anonymization rate and efficacy of each mix-zone. Note that $\mathcal{D}$ can have many versions, each with an anonymization level, simply varying the privacy level k, for example, 2, 4, and 6.

In the third step, Metrics Extraction, we extract the mobility, privacy, and social metrics from the original dataset $\mathcal{D}$ and its anonymized versions $\mathcal{D}'_{k=2}$, $\mathcal{D}'_{k=4}$, and $\mathcal{D}'_{k=6}$.

In the fourth step, Utility Analysis, we normalize each metric with min-max normalization, transforming the data into a scale from 0 to 1, enabling us to analyze and identify which metric had more distortion than the original one. So, we check the distortion level Z_D, with distortion function WS, of the metrics between the original dataset and the anonymized dataset, for example, $Z_D = WS(\mathcal{D}, \mathcal{D}'_{k=2})$. Next, we calculate the utility level of each metric, denoted as $U = 1 - Z_D$ (Details about this step, see Subsection 7.4.3). The utility U and distortion Z_D metrics are used for various analyses in utility terms. For instance, identify a ranking of metrics that had more distortion in a protected dataset. It can be used as input for two algorithms: MCDM and Utility for Specific Applications. The first one is Multiple-criteria decision-making capable of ranking smart cities application domains with more utility. The second one can identify the utility level of a protected dataset with k level for specific applications. Further, it enables the Decision-maker (DM) to decide which dataset is more appropriate for a specific application, given various datasets protected by different k levels.

Regarding the third step, we select nine metrics of privacy, mobility, and social aspects closely associated with classes of applications that could use anonymized data. Thus, this analysis enables us to identify inherent utility characteristics in the applications class level to understand their behavior more deeply and the relation between anonymization data usage and utility. Next, we will present the metrics group used in this work and detail each step of the framework.

7.4.2 Mobility, Privacy, and Performance Metrics

In the context of open and target market data, mobility data anonymized should have high utility, privacy, and anonymization performance (regarding coverage). However, achieving these controversial requirements is complex, and a utility analysis framework must be able to measure the utility, considering privacy and coverage metrics. UAFAT measures the utility considering the anonymization performance with the mix-zones' efficacy metric Eff_M. The Eff_M is possible to measure if the vehicles that crossed the mix-zones were anonymized or not, enabling the identity of the anonymization scheme to have a good performance. Considering the mix-zones set M, the mix-zones efficacy is calculated as follows: $Eff_M = AR_M/(AR_M + NAR_M)$, where AR_M is the number of trajectories that passed and were anonymized by M; and NAR_M is the number of trajectories that passed and were not anonymized by M.

In human mobility, mobility metrics extract inherent characteristics of the mobility models that can be useful in areas like transport planning, the private market, and sta-

Table 7.2: Mapping of the smart cities domains and applications with privacy, mobility, and social metrics.

Domain	Applications	Metrics	Ref
Statistical Analysis	• Traffic monitoring; • Route plan; • Vehicle trajectories mining; • Multimodal transport integration; • Urban population migration.	TDSL, TRIPTIME NUMVIS, RGYR, PSEU	[16] [18] [17] [19] [15]
Urban Planning	• Multimodal integration: integration of transport modes to optimize the transport of people or goods.	TDSL, TRIPTIME PSEU	[19]
	• POI mining: Correlation between mobility patterns and station functions with subway transaction records and POIs; • Clusters identification: Understanding of polycentric urban with city hubs, centers, and borders to enhance the quality of services; • Strategical places identification: Identify places in the city for deploying services e.g. Electric Vehicles (EV) charging, ride and vehicle sharing, smart park lot, and mix-zones deployment.	SPC, SPD, NUMVIS RGYR,PSEU	[16] [18] [17] [19] [15]
	• Testbeds for smart mobility systems: ride and vehicle sharing; traffic accident detection; smart traffic lights; route plan;	TDSL, TRIPTIME PSEU, SPC, SPD	[16] [19]
Driver Behavior	• Driver/passenger behavior clustered by trip purpose or pattern; • Bus route plan, electric cabs; • Mine the temporal and spatial characteristics of behavioral trajectories.	TDSL, TRIPTIME PSEU, SPC, SPD	[16] [18] [17]
Social Networks	• Understand the differences in location-based social network usage and the types of places visited.	MAXCON, CODU CONEN, INCO, RGYR	[16] [167] [20]
Opportunistic Network	• Design of data dissemination protocols for V2I and V2V.	MAXCON, CODU, CONEN, INCO, RGYR	[18] [167] [20]
Targeted Marketing	• Public data marketplace: promote sales campaigns for targeted marketing. • Selling data: Collecting, analyzing, and re-selling big data to third parties.	TDSL, TRIPTIME PSEU, SPC, SPD	[16] [18] [17]
Privacy & Safety	• Analysis to safeguard user data and location privacy while ensuring safety by preventing tracking, re-identification, and unauthorized access.	TDSL, TRIPTIME PSEU, RRET, SPC SPD	[19] [15] [273] [167] [20]

tistical analysis. Mobility metrics enable analyzing spatial, temporal, and social aspects, such as the encounter between user groups and the time and frequency of users' visits to specific places. Also, these metrics could be used as effective utility metrics and reach perspectives on utility that other metrics could not bring—for example, identifying temporal, social, and spatial distortions between original and anonymized dataset versions. Next, we present the mobility characteristics metrics and then show how these metrics can be used to measure utility in the smart cities application domain.

Mobility characteristics metrics can be categorized by spatial, temporal, and social metrics [265]. Spatial metrics concern identifying spatial and temporal characteristics of mobility. In smart cities, domains like urban planning and statistical analysis use spatial metrics to identify strategic points for the building and identify traffic monitoring, respectively.

Travel's Distance Straight Line (TDSL) is a spatial metric that computes the distance (in kilometers) traveled straight line by a set of individuals in a trajectory. The straight line distance traveled by an individual is calculated as the sum of the distances traveled [275]. It is denoted by the formulae $TDSL_{o,d} = distance(o, d)$.

Number of Visits per User (NUMVIS) is another spatial metric that computes the number of visits per user (i.e., data points). It is expected that when a trajectory is anonymized by mix-zones the NUMVIS tends to be lower than the original dataset.

Another kind of spatial mobility metrics group is based on Stay Point (SP). Stay Point is a region where an entity stays for a minimum interval [233]. The parameters of a SP are the radius r in meters of the region and the minimum time to stay there t in minutes. These points are relevant for detecting many mobility characteristics, such as traffic lights and even traffic jams. Stay points are commonly used as a substrate for

many privacy mechanisms in the location privacy context. In LPPM design, stay points are typically used to detect POIs and apply obfuscation methods. Additionally, stay points can be used for mix-zones placement [32]. In location attacks, stay point mining enables identifying and characterizing behaviors in the users' trajectory, revealing sensitive information, such as social preferences, since victims regularly go to those places [32, 197]. In this way, stay points bring valuable information w.r.t. location privacy [261]. We use the metrics related to the stay points: SPC per user and SPD.

Stay Point Count (SPC) per user refers to the different locations one visits. This metric can be used to understand the different mobility characteristics of users. Its distribution contains the regions visited by users in which some may have only a few places while others may have a large collection. That can be used for POI extraction, routing algorithms, and contagion models.

Stay Point Duration (SPD) refers to the time a user spends at a location. The stay time algorithm's parameter defines the lower bound, and the upper bound has no limit. Understanding the time users (or population) spend on average at a location can be a good indicator of its capacity regarding data offloading, helping to design handoff solutions.

Temporal metrics evaluate the temporal aspects of the mobile entities. In the smart cities domain, like target marketing, temporal aspects of mobility can measure the duration of hot points for a targeted market. Or the development of smart semaphores in the urban planning domain.

Trip Time (TRIPTIME) is a temporal metric that measures time spent moving between two places. Given two locations of the user u: l_a origin and l_b destiny. They are represented by formulae $TRIPTIME_{u,l_a,l_b} = arrival_{u,l_b} - departure_{u,l_a}$.

Social metrics measure the relationship between mobile entities to generate data dissemination. These metrics can be used to design protocol data dissemination in VANETs, particularly in V2I and V2V communication models. It also can be used to analyze social networks to improve and develop location-based services in the city.

Inter-contact Time (INCO) is a social metric measuring the interval between two consecutive encounters between a pair of users [265]. In opportunistic scenarios, INCO represents the frequency of encounters between each pair of users, representing an opportunity for message transmission.

Contact Duration between a Pair (CODU) represents the time two users spend inside each other's transmission range without interruptions. The CODU also represents an opportunity to transmit a message in an opportunistic network. However, instead of showing the frequency of encounters, the CODU shows the amount of data that can be delivered in each encounter.

Another social metric is **Contact Entropy (CONEN)**, which is an user that describes how distributed the contacts of a user u among his set of contacted users are.

Users with small entropy values have most of their contacts with few other users, who can be seen as friends. In contrast, users with high entropy values have their encounters well-distributed among their peers. CONEN can help identify users with high entropy being likely possible for routing messages in opportunistic networks. Given a user u and its set of contact peers $C = c_1, c_2, \ldots, c_n$, its contact entropy C_{Eu} can be defined as the Shannon's Entropy of the probabilities $P(c_i)$ of user u contacting user c_i.

Maximum of Connections between a User Pairs (MAXCON) represents the maximum number of connections in an interval d between a pair of users. In developing data dissemination protocols for opportunistic networks, this metric can be used to identify connection peaks over time.

The **Radius of Gyration (RGYR)** of a user u is a spatial metric that measures the maximal Euclidean distance between a user's home and the other places they have visited. We consider a user's home as their most visited location l, which favors the idea of a recurrent place [265]. The **RGYR** can also be considered a social metric for measuring social interactions and, consequently, be used to design opportunistic protocols [276]. Individuals with similar RGYR patterns tend to be a strong iteration between each other, enabling the dissemination of information. Thus, this work considers the RGYR as a social metric. The RGYR of a user u is denoted by $RGYR_u = \max(distance(u_{home}, l)\ \forall l \in L)$ where L is a location set of u.

Pseudonyms per User (PSEU) is a privacy metric representing the number of pseudonyms a trajectory has when crossing the mix-zones. The higher PSEU value means this trajectory has more privacy. A trajectory with $PSEU = 1$ means that the mix-zones did not protect this trajectory.

Finally, the **Re-identification Risk in entire Trip (RRET)** is a privacy metric for measuring a user's risk u in an entire trip. It is expected that a user in his trajectory can cross many mix-zones, and being anonymized by them has a lower re-identification risk than a user that has passed no one mix-zone. In the same way, it is expected that a user u who crosses mix-zones with a high k has **RRET** lower than mix-zones set with a low k. Consider a mix-zones set $M = \{M_1, M_2, \ldots, M_i, \ldots, M_m\}$ and a privacy level set $K = \{k_1, k_2, \ldots, k_i, \ldots, k_m\}$, in which k_i represents the privacy level of respective M_i. Consider a user u crossing mix-zones and getting anonymization yielding pseudonyms in their trip. So, the RRET of a user u is denoted by RRET= $\prod_{i=1}^{(PSEU-1)} \frac{1}{k_i}$, where $PSEU$ is a Pseudonyms per User (PSEU) of a user u and k_i is the privacy level k setup in M_i. Note that the mix-zones can have different privacy levels from each other.

Table 7.2 summarizes the discussion on mobility characterization metrics and their potential association with the application domain in smart cities. The associations identified in this work were based on analyzing the proposals found in Section 7.2.

7.4.3 Utility Metrics

Utility analysis of the anonymized data is a pillar of publishing data. Anonymized data with enough utility can be used for the public good or monetized in the data marketplace and consumed by smart city applications [18]. However, it is necessary to identify the context of the application in which these data will be used. Because a protected dataset is more useful for one class of applications than another, this depends on how this dataset was protected. Also, the trade-off between privacy and utility can be greatly improved by exploiting the concept of approximation intrinsic in metrics [15]. Following this intuition, our strategy is to identify metrics closely related to the applications class. Each metric identifies data distortions with a distance between the mobility metrics measured before and after applying an LPPM, such as mix-zones, in dataset $\mathcal{D}$. Thus, the distortion of the statistical distance of mobility metrics can be used as a utility metric in the mobility context.

Statistical distance is the approach to identifying the distance between two probability distributions. Here, we used the WS, which measures the difference between two distributions (distortion level) by the optimal cost of rearranging one distribution into the other[1].

Definition: The distortion associated with an random variable X is the result over the original dataset $D \in \mathcal{D}$ and its anonymized version $D'_{k=i}$ as $WS(D, D'_{k=i}) = Wasserstein[X(D), X(D'_{k=i})]$ where $i \in K = \{2, 4, 6\}$. The smaller the WS value is, the less effort is needed to transform one distribution into another, and consequently, the two distributions show high similarity and low data distortion. The Wasserstein distance is asymmetric, (weakly) continuous, and ideal for analyzing corrupted data, unlike common distribution divergence approaches, such as Kullback-Leibler or Jensen-Shannon [256].

In this paper, we verify the utility of cabs and private cars datasets in three sets of metrics – mobility, social, and privacy – closely associated with classes of applications.

Specifically, for each metric, we normalize each metric with min-max normalization, transforming the data into a scale from 0 to 1. With all metrics normalized, it's possible to analyze distortion, such as identifying which metric had more distortion than the original one. Next, we calculate the WS of each metric before and after the D being submitted to mix-zones, previously configured, e.g., $WS[MAXCON(D), MAXCON(D'_{k=2})]$ that represents the distortion level between $MAXCON$ metric extracted from original and anonymized dataset by mix-zone with $k = 2$. In the utility analysis, we expect to identify which application domain the anonymization process had a major and less impact in data distortion, enabling direct what application domain of a smart city, being market-

[1]For more details, please refer to C. Villani, "Topics in Optimal Transportation". American Mathematical Soc., 2003, no.58

place or open data, the dataset anonymized by mix-zones can be more useful. Applying this process to all of the metrics, we obtain a distortion level matrix Z from a dataset D, Z_D (Equation 7.1), whose rows and columns represent the metrics and privacy level, respectively. Next, we calculate the utility level of each metric, that represents the utility metrics U of the dataset D, denoted as $U = 1 - Z_D$.

$$Z_D = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_n \end{array} \begin{array}{c} \begin{array}{cccc} k=2 & k=4 & \dots & k=s \end{array} \\ \begin{pmatrix} WS[X_1(D), X_1(D'_{k=2})] & WS[X_1(D), X_1(D'_{k=4})] & \cdots & WS[X_1(D), X_1(D'_{k=s})] \\ WS[X_2(D), X_2(D'_{k=2})] & WS[X_2(D), X_2(D'_{k=4})] & \cdots & WS[X_2(D), X_2(D'_{k=s})] \\ \vdots & \vdots & \ddots & \vdots \\ WS[X_n(D), X_n(D'_{k=2})] & WS[X_n(D), X_n(D'_{k=4})] & \cdots & WS[X_n(D), X_n(D'_{k=s})] \end{pmatrix} \end{array} \tag{7.1}$$

Suppose privacy, mobility, or social metrics are used in an application; for example, by analyzing the distortion of these metrics Z_D or their complement U, we can obtain the utility level of a metric. In that case, it is possible to check how much this application or smart city domain is affected by anonymization. Also, it is possible to rank the metrics sorted by distortion level in descending order for each k setup.

7.4.4 Utility Ranking

Deciding the usefulness of protected mobility data is a complex task. The utility depends on many aspects and criteria. The data utility is also closely related to the application that will consume this data. For example, protected mobility data may have higher utility for statistical analysis than data for testing electric vehicle driving systems. The relationship between the utility and application is dynamic and varies according to the LPPM that protects the mobility dataset.

7.4.4.1 Multi-Criteria Decision-Making (MCDM)

The ranking of smart city domains sorted by anonymized mobility data can be modeled as a decision-making problem. Multi-Criteria Decision-Making (MCDM) is a sub-discipline of operations research that explicitly evaluates multiple conflicting criteria and alternatives in decision-making [277]. These techniques enable DMs to make rational

decisions and use various quantitative criteria to assess and select the optimal alternatives from qualitative judgments. The Equation 7.2 represents the MCDM problem, where $\mathcal{V} = \{a_1, a_2, \ldots, a_m\}$ represents a set of feasible alternatives and $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ is a set of decision-making criteria, and p_{ij}, where $1 \leq i \leq m$ and $1 \leq j \leq n$, is the score of alternative i with respect to criterion j. The goal is to select the most desirable or important alternative based on criteria. The overall value of alternative i, V_i can be obtained using various methods. In general, is assigning weight W_j $(w_j, \sum_j w_j = 1)$ to criterion j, then V_i is obtained with a simple additive weighted value function, which is the underlying model for most MCDM methods, $V_i = \sum_{j=1}^{n} w_j p_{ij}$ [278].

$$A = \begin{array}{c} \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \ldots & c_n \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{ij} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{pmatrix} \end{array} \tag{7.2}$$

In the MCDM methods alternatives, criteria, and weights are attributed by the DM. Specifically, the DMs assign relative importance to the criteria, also called weights, which are the set of preferences given to each criterion by the DMs. The weights are important to the problem because they influence the outcome of the decision-making process. Thus, selecting the weighing method is an essential part of any MCDM problem as it quantitatively determines the relative importance of criteria and greatly impacts the reliability and accuracy of the decision results.

These concepts can be applied to deciding the usefulness of protected mobility data for application domains. The set of application domains $\mathcal{V}$ = {Statistical Analysis, Urban Planning,..., Privacy & Safety} represents the alternatives in which a dataset will have different levels of utility, i.e., the row a_i in A in Equation 7.2. A dataset contains a set of metrics $\mathcal{C}$ = {TDSL, NUMVIS, SPD,..., CONEN}, denoted as column c_j in A. These metrics may be valuable for certain application domains while others may not. For example, TDSL and NUMVIS metrics are more relevant to the Statistical Analysis domain than Privacy & Safety. Therefore, these metrics will have more weight for Statistical Analysis than Privacy & Safety. In this way, $\mathcal{C}$ can be used as criteria for selecting application domains that best leverage protected data. For each cell $\langle a_i, c_j \rangle$ in A, i.e., p_{ij} receives a weight regarding alternative i regarding criterion j.

There are many MCDM proposals, the methods widespread including Weighted Product Method (WPM), Analytic Network Process (ANP), Analytic Hierarchy Process (AHP), Best Worst Method (BWM). The selection of an appropriate weighing method for solving an MCDM problem is one of the critical issues.

7.4.4.2 Best Worst Method (BWM)

The BWM is weighing method for a MCDM problem. BWM highlights from others approaches by the efficiency of this method in reducing the times of pairwise comparisons and the good performance in maintaining consistency between judgments, which is important where we have many alternatives and criteria [279, 280]. However, the BWM can result in multiple optimal solutions for a simple decision-making problem with more than three criteria/alternatives. It implies that for not fully consistent comparisons, having more than three criteria/alternatives, the interval weights can be engaged to derive the weight of factors in the multi-optimality problems. So, in most cases, a single solution is preferred compared to multiple optimal solutions. The linear BWM is a BWM approach that always results in a unique solution, desirable when no higher-level information needs to be considered [278, 280]. Figure 7.3 details the BWM steps. In steps 3 and 4, the preference between criteria is defined with Saaty's scale, which is a well-known scale in decision problems that aims to support defining the relative importance between criteria [3]. Saaty's scale has a bounder 1 to 9, lowest and highest importance, respectively. For instance, if criterion A's relative importance is more significant than criterion B's in 8, A is eight times more important than B [277].

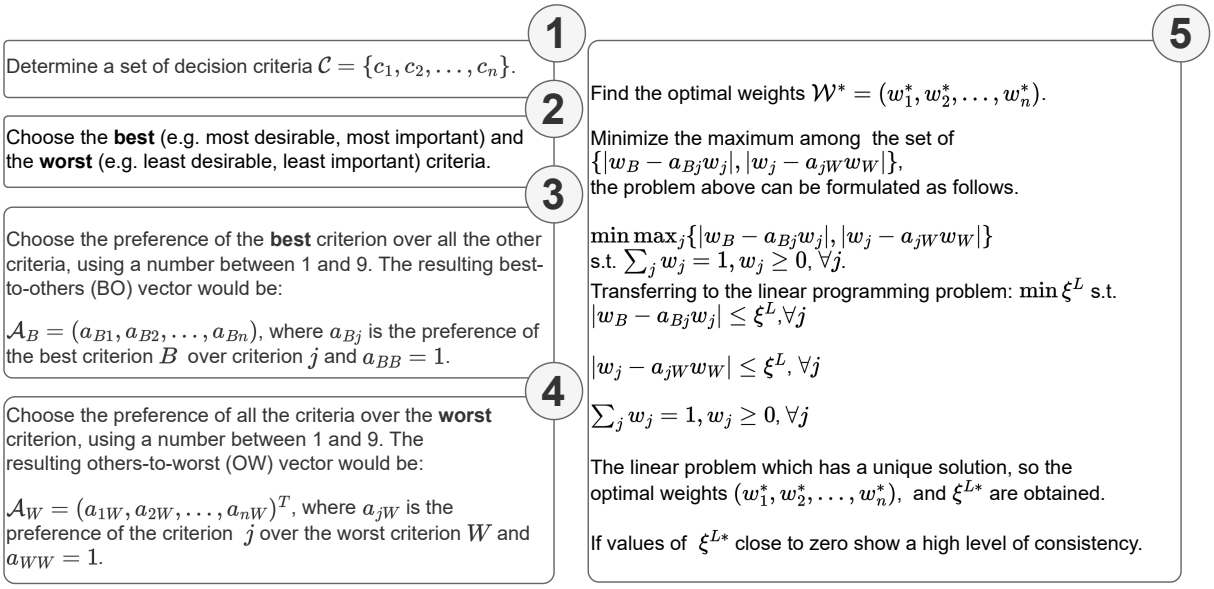


Figure 7.3: Linear BWM steps, proposed in [3].

Algorithm 5: BWM for Data Utility Ranking in Smart Cities Application Domains.

Data: Application domains set $\mathcal{V}$ as alternatives; Utility metrics set with x privacy level $U_{:,k=x}$ as criteria; Threshold of consistency index λ
Result: Ranking of application domains V

$\xi^{L*} \leftarrow 1$
$J \leftarrow$ defineCriteriaWeights($U_{:,k=x}$)
$J \leftarrow$ changeScale(J) `// Change the matrix` J `to Saaty's scale.`
$B \leftarrow$ getIndex(max($U_{:,k=x}$)) `// Gets an index in` $U_{:,k=x}$ `with the greatest cell value.`
$W \leftarrow$ getIndex(min($U_{:,k=x}$)) `// Gets an index in` $U_{:,k=x}$ `with the lowest cell value.`
$\mathcal{A}_B \leftarrow J_B$
$\mathcal{A}_W \leftarrow J_W$
while ($\xi^{L*} > \lambda$) **do**
  $\mathcal{W}^*, \xi^{L*} \leftarrow$ calculeOptimalWeights($\mathcal{A}_B, \mathcal{A}_W$)
  if ($\xi^{L*} > \lambda$) **then**
    $\mathcal{A}_B, \mathcal{A}_W \leftarrow$ adjustCriteriaPreference($\mathcal{A}_B, \mathcal{A}_W$))
  end
end
$A \leftarrow$ definePriorityAlternativesCriteria($\mathcal{V}, U_{:,k=x}$)
$A \leftarrow$ changeScale(A)
$A \leftarrow$ NormalizeMax(A)
$V \leftarrow A\mathcal{W}^{*T}$
sort(V)

7.4.4.3 BWM Applied in Data Utility for Ranking Smart Cities Application Domains

The selection of suitable MCDM methods is based on the structure of problems [279]. After defining the problem, boundary conditions, criteria, establishing a decision matrix, and determining criteria weights are significant for any MCDM techniques.

So, we must define boundary conditions in our scenario to define the suitable MCDM method for ranking the smart city application domains sorted by the utility of anonymized mobility data. In our context, we have metrics that identify distortion levels between original and anonymized mobility data. Because these metrics are intimately related to the smart city application domains, some metrics have more importance to certain application domains than others. In this way, metrics with more distortion levels than others mean that certain application domains are more affected than others when using such anonymized data. Thus, it's possible to determine to which application domain the anonymized dataset has more utility. In this way, the metrics can be seen as criteria

$\mathcal{C}$ that can be analyzed to rank the alternatives represented by the application domains $\mathcal{V}$ (see Figure 7.4). Algorithm 5 ranks Smart Cities' application domains based on utility metrics. We detail each step as follows, but first, we present some preliminary definitions.

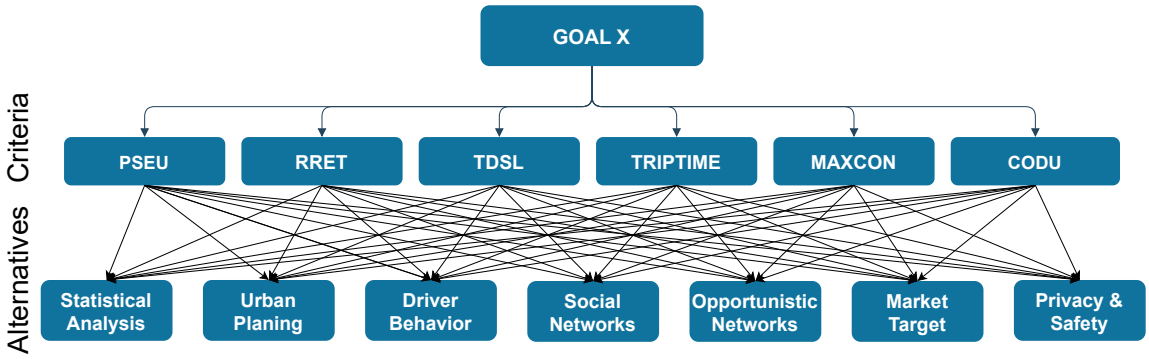


Figure 7.4: MCDM method, Goal X: Ranking smart city domains where anonymized data is most useful.

An important step of MCDM methods is the attribution of the weights between the criteria. Specifically, we determine the preference of the best/worst criterion over all the other criteria using Saaty's scale with numbers between 1 and 9. Thus, background knowledge of DM about the criteria is necessary to define the weights.

Unlike the conventional BWM model, in which the DM participates in the criterion judgment, the criterion weight definition is automatic in our model. We apply the linear BWM for the ranking, considering each utility metric as a criterion that must be analyzed and evaluated to elect the application domain. Recall that the utility metrics are distortion level between calculating the WS of each metric before and after the D being submitted to mix-zones, previously configured, e.g., $WS[CODU(D), CODU(D'_{k=2})]$ that represents the distortion level between CODU metric extracted from the original and anonymized dataset by mix-zone with $k = 2$ and Z_D is the distortion matrix produced for all metrics and privacy levels k applied in D (see Section 7.4.3). In this way, we define the utility function that defines the utility level of each metric, denoted as $U = 1 - Z_D$. Thus, the criteria set of BWM method $\mathcal{C} = (c_1, c_2, \ldots, c_x, \ldots, c_n)$ can be represented as a utility metric set from dataset anonymized, e.g., with $k = 2$, $U_{:,k=2} = (u_1, u_2, \ldots, u_x, \ldots, u_n)$, where $U_{:,k=2} \in U$ is the utility metrics column in which applied the privacy level $k = 2$ in dataset D and u_x is the utility metric x. The Algorithm 5 requires as input the alternatives $\mathcal{V}$, criteria set $U_{:,k=x}$, and threshold of consistency index λ, where $\lambda \in (0, 1)$.

In the context of data utility, it is desirable that a dataset is protected and, at the same time, has a utility for specific application domains. Thus, the level of utility obtained by a metric is an essential criterion for selecting an application domain. In BWM steps, in which one needs to choose the preference of the best criterion over all the other criteria and also the preference of all criteria over the worst criterion, steps 3 and 4 of Figure 7.3, respectively, the DM must enter the value of the criteria weights on a scale of

1 to 9. In our model, we use the utility metric as criteria weights. Specifically, we define the criteria weight a_{ij} as the ratio between $a_{ij} = \frac{u_i}{u_j}$, where u_i and u_j are utility metrics of criterion i and criterion j, respectively. The ratio $\frac{u_i}{u_j}$ is used as a penalty and reward in the criteria. For instance, if criterion i is much greater than criterion j, i.e., $u_i >> u_j$, it means that criterion i has higher utility and consequently importance for the dataset than criterion j. Therefore, a_{ij} tends to be high. In contrast, if $u_i << u_j$, the a_{ij} tends to be zero, and criterion i has low utility compared to criterion j. From the definition, we construct a matrix of criteria judgments J that shows the relative preference of criterion i to criterion j (see Equation 7.3), represented in the line 1 of Algorithm 5.

$$J = \begin{array}{c} \\ c_1 \\ c_2 \\ \vdots \\ c_n \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \dots & c_n \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{ij} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \end{array} \tag{7.3}$$

The a_{ij} in J can generate a different scale from the Saaty's scale, which has numbers 1 to 9 ([278]) (line 3 of Algorithm 5). In this way, we must transfer from the scale of values in J to Saaty's scale [278]. For this, we use a linear normalization with approximation denoted by Equation 7.4, where ns is the new score, os is the original value you want to normalize, olb is the lower limit of the original range, oub is the upper limit of the original range, nlb is the lower limit of the new desired range, and nub is the upper limit of the new desired range. The original ranges are obtained as $olb = \min(J)$ and $oub = \max(J)$ and new desired range are $nlb = 1$ and $nub = 9$, corresponding the Saaty's scale.

$$ns = nlb + \frac{nub - nlb}{oub - olb}(os - olb) \tag{7.4}$$

We must select the best and worst criterion. In utility data terms, we consider the best and worst criterion as the utility metrics with great and low utility levels, respectively. Lines 4 and 5 of Algorithm 5 identify the great and low utility metric values and get the index corresponding in the matrix J.

Next, we choose the preference of the best criterion B over all the other criteria. For this, selects the line B in J, i.e., $\mathcal{A}_B = J_B$, $\mathcal{A}_B = (a_{B1}, a_{B2}, \dots, a_{Bn})$, where a_{Bj} is the preference of the best criterion B over criterion j and $a_{BB} = 1$. After that, we choose the preference of all criteria over the worst criterion W. For this, selects the column W in J, i.e., $\mathcal{A}_W = J_W$, $\mathcal{A}_W = (a_{1W}, a_{2W}, \dots, a_{nW})^T$, where a_{jW} is the preference of the criterion j over the worst criterion W and $a_{WW} = 1$ (lines 6 and 7 of Algorithm 5).

The loop, in lines 8-13, is used to ensure that the consistency index of the optimal weights (ξ^{L*}) is lower than the threshold λ. The lower the ξ^{L*}, the more consistent is the optimal weights $\mathcal{W}^*$.

After calculating the optimal weights $\mathcal{W}^* = (w_1^*, w_2^*, \ldots, w_n^*)$ and ξ^{L*}, line 9, we verify the consistency of the weight comparing the ξ^{L*} with the λ, if the weights are not consistent, we must adjust of priority of the best and worst criteria (line 11).

In the next step, we apply the optimal weights in our decision-making problem for ranking the alternatives (line 14). It is done by comparing alternatives against each criterion. Specifically, each alternative is a smart city application domain i concerning criterion j, represented by metrics detailed in Section 7.4, i.e., $p_{ij} \in$ matrix A (see Equation 7.2). So, the DM must define a score of 1 to 9, which is the importance of the metrics in each domain and application. This score can be obtained from the prevalence of such metrics in related work to the context of each application domain (see Table 7.2).

Defined each alternative value i concerning criterion j (p_{ij}) by DM and calculating the optimal weights, we multiply the optimal weight represented by utility metric u_j for each p_{ij}, i.e., $p_{ij} = p_{ij}u_j$. The reason is to consider the level of importance of each metric, in terms of usefulness, for the application domains in which these metrics are most important.

After calculating the $p_{ij}u_j$, we changed the matrix A values to a scale of 1 to 9 with Equation 7.4 (line 15). Following, we normalize each element p_{ij} of matrix A with max value of column j, i.e., $p_{ij} = \frac{p_{ij}}{\max(p_j)}$. Then, we multiply the optimal weights w_j^* metric by p_{ij}. Finally, we calculate the overall score of alternative i as $V_i = \sum_{j=1}^{n} w_j^* p_{ij}$ and sorting the values of $V_i \forall i$, the best alternative is identified (lines 17 and 18).

7.4.5 Utility of Anonymized Mobility Data for a Specific Application

An essential issue regards measuring the utility of anonymized mobility data for a specific application of a smart city domain (question **Q6**). We can answer this question with the analysis of utility metrics. Specifically, we must identify the utility metrics intimately related to an application analyzed and calculate the weighted average. For instance, let's use the scenario that uses anonymized mobility data, protected with a privacy level, for positioning new smart parking based on clustering. These clustering approaches require mobility metrics (such as SPC, SPD, and TDSL) rather than social metrics (CODU, CONEN, and MAXCON). In this case, the mobility metrics are more relevant than social metrics and consequently have more weight. In contrast, social metrics can be more appropriate than privacy and mobility when developing a data dissemination protocol. After defining the weight of each utility metric, we calculate the weighted average to identify the utility level of the anonymized data protected with a dedicated

privacy level. The Algorithm 6 present details about the utility level.

The Algorithm 6 calculates the utility level of anonymized mobility data for a specific technique i of an application Υ_i from a smart cities application domain $\mathcal{V}_i$. The algorithm receives as parameters the application Υ_i and utility metrics extracted from the dataset anonymized with privacy level x, $U_{:,k=x}$. The first step is to identify the technique used in the application's design. One application can be developed with many techniques, so we must select a specific technique for which the application was developed (line 1 of Algorithm 6). For instance, an application of mix-zones deployment can be designed with many strategies, such as clustering approaches, flow traffic, and map context, so it's important to identify what approach to use in design. Next, we must identify which metrics are intimately related to the development of the current technique (line 2). For instance, metrics based on POIs, stay points, and vehicle traffic flow can be relevant for the positioning of EV charging. Next is, to assign the metrics ($\mathcal{M}$) extracted from Υ_i with utility metrics $U_{:,k=x}$ by similarity (line 3). The next step is to define the weights for the utility metrics because an approach has metrics that are more relevant than others. We use Saaty's scale, which denotes less and greater relevance at 1 to 9 (line 4). Finally, we calculate the utility level for a specific technique with the weighted average having the weights and utility metrics as parameters (line 5).

Algorithm 6: Utility of Mobility Data Anonymized for a Specific Application.

Data: Application Υ from an application domain i, $\Upsilon \in \mathcal{V}_i$; utility metrics of the dataset anonymized with privacy level x, $U_{:,k=x}$; Mix-zones efficacy Eff_M calculated from mix-zones set M.

Result: Utility level χ for specific smart cities application.

1 Identify a technique i of interest related to Υ, i.e. Υ_i.
2 Identify the related metrics $\mathcal{M} = \{m_1, m_2, ..., m_f\}$ in Υ_i.
3 assign($\mathcal{M}, U_{:,k=x}$)
4 $\mathcal{W}^* \leftarrow$ defineMetricsWeight($U_{:,k=x}$)
5 $\bar{x} \leftarrow$ weightedAvg($\mathcal{W}^*, U_{:,k=x}$)
6 $\chi \leftarrow \bar{x} \times \mathit{Eff}_M$

7.5 Results and Discussion

7.5.1 Experiments Setup

In this study, we used two trajectory datasets. The first is the Cabspotting dataset containing mobility traces of San Francisco, USA, taxicabs, with data from 500 unique taxis collected over 30 days [126]. We extract a sample D of high traffic from the Cabspotting dataset corresponding to the 19th day of the collection period with 417,781 registers, 454 users, and 2036 trips. The second is a private car dataset from Shenzhen, China [281]. We extract a sample D of 13 days with 379,551 registers, 8985 users, and 8985 trips. For the utility characterization, a privacy budget of seven mix-zones was defined for each mix-zone positioning algorithm proposed in [56]. We empirically selected a budget of seven mix-zones based on to limit computational costs and facilitate mix-zone analysis. Another condition is that the mix-zones should anonymize trajectories in at least two of the three k setups. Additionally, the mix-zones could not overlap areas between the mix-zones already selected. Given the anonymization condition above and the restricted size of the San Francisco and Shenzhen regions, it resulted in a privacy budget of seven mix-zones. The mix-zones parameters radius r and k were also selected by empirical study. The radius r was obtained by a coverage empirical analysis ranging the radius threshold r from 100 to 600 m ($r = \{100, 200, 250, 300, 400, 500, 600\}$). We had mix-zones overlapping from $r = 600$ m, significantly reducing the privacy budget. Thus, the r of 500 meters was defined for all mix-zones [261]. The k values were chosen based on the criteria that the k values produced the three highest anonymization coverage rates for each dataset. This way, for Cabspotting, the k values were 2, 4, and 6; the Shenzhen dataset was anonymized with k equal to 2, 3, and 4 because the dataset is less dense than Cabspotting. Table 7.3a represents the mix-zones positioning for both datasets.

Table 7.3b presents the performance metrics for the two datasets. For Cabspotting, a dataset with a considered traffic density, we can observe that the AR_M and Eff_M metrics of the mix-zones set M present better performance than Shenzhen, which is a less dense dataset. We can also notice that with the variation in the value of k, we have a significant degradation in Eff_M. But why Shenzhen had low Eff_M? The answer to this question is that we can attribute this result to several factors. Among them, the positioning of the mix-zones may need to be improved. Therefore, poorly positioned mix-zones have a high probability of low AR_M that directly affects the Eff_M. Another factor that affects mix-zone performance is the calibration of mix-zone parameters, such as k value and radius [261]. If the value of k is high and the radius is low, this is the worst-case scenario

for AR_M and Eff_M. For the Shenzhen dataset, varying the value of k for a single unit can considerably affect performance. When none of these factors mentioned are the problem, the very nature of the dataset in terms of vehicle density can affect performance. In other words, if the number of vehicles is less than the value of k, even if it is the minimum value such as $k = 2$, this may affect the AR_M and Eff_M metrics. The latter is a critical factor since mix-zones are inefficient for datasets with little traffic, so generating dummy trajectory approaches must be considered to perform anonymization and increase traffic volume. It is important to emphasize that despite the low value of Eff_M for values of $k = 3$, 4, and 6, these will be important for analyzing the sensitivity of the Framework in detecting datasets with low utility, not only due to the data distortion caused by the use of an LPPM but also due to the low Eff_M and having a high NAR_M.

Table 7.3: Mix-zones deployed in San Francisco and Shenzhen cities a and their performance metrics b.

(a) Mix-zones deployed in San Francisco and Shenzhen.

	Cabspotting		Shenzhen	
Mix-zones	Latitude	Longitude	Latitude	Longitude
mix0	37.714801	-122.397982	22.534184	114.075663
mix1	37.724830	-122.400157	22.531598	114.063069
mix2	37.735133	-122.404532	22.569104	114.067704
mix3	37.676005	-122.391491	22.572283	114.082865
mix4	37.615315	-122.393566	22.531354	114.044641
mix5	37.774378	-122.401540	22.542688	114.047722
mix6	37.768990	-122.419450	22.567239	114.054169

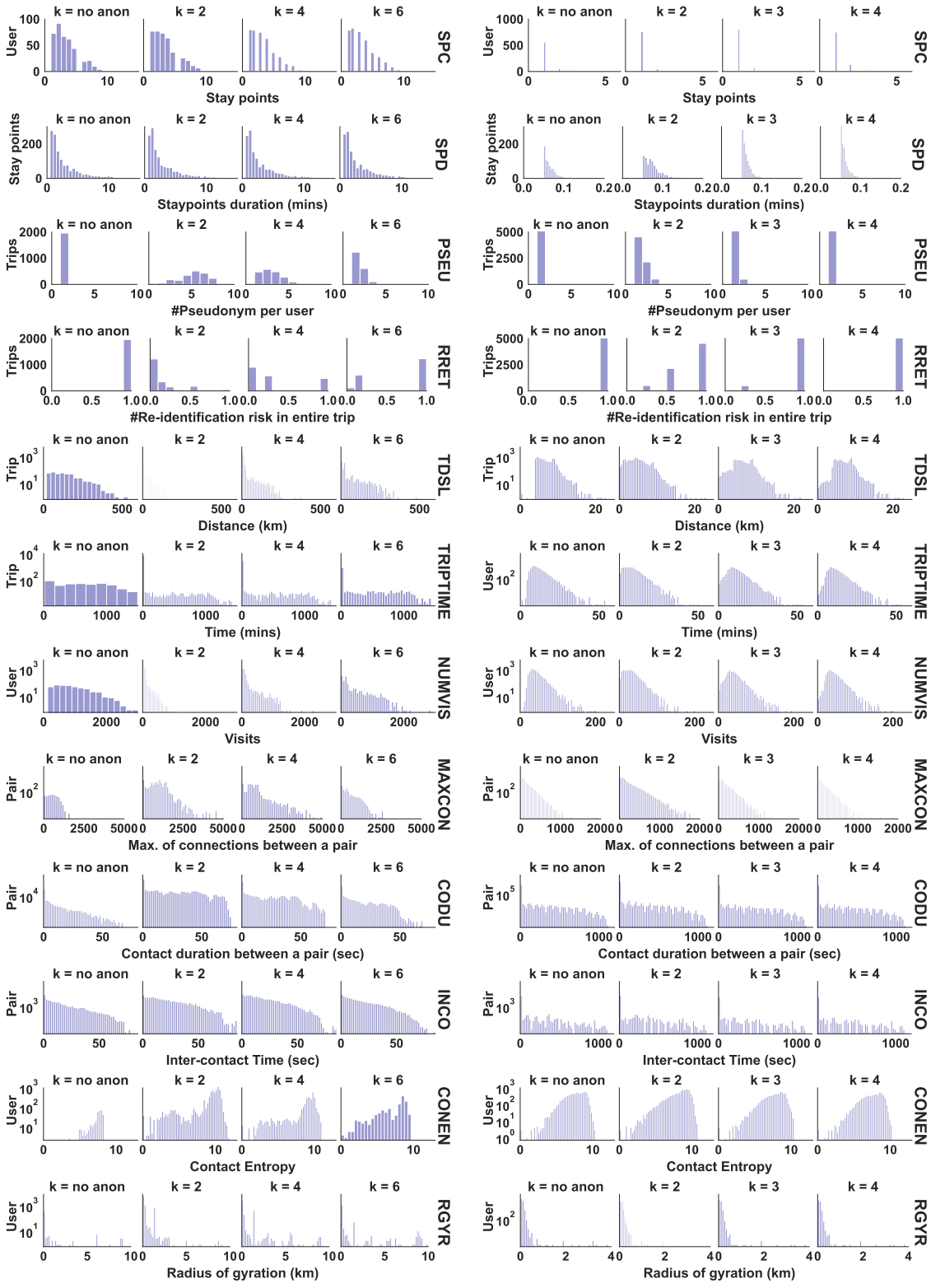
(b) Anonymization (AR_M), No Anonymization NAR_M, and Efficacy Eff_M.

Cabspotting				Shenzhen			
Privacy	AR_M	NAR_M	Eff_M	Privacy	AR_M	NAR_M	Eff_M
$k = 2$	7435	2492	0.749	$k = 2$	3372	12531	0.212
$k = 4$	3021	6906	0.304	$k = 3$	568	15335	0.036
$k = 6$	855	9072	0.086	$k = 4$	68	15835	0.004

7.5.2 Metrics Distribution Analysis

In the metrics distribution analysis, we investigated the spatial, temporal, social, and privacy metrics for the Cabspotting and Shenzhen datasets, depicted in Figures 7.5a and 7.5b, respectively.

In SPC for the Cabspotting dataset, we can note the collateral effect of anonymization on the trajectory dataset. For instance, the number of users (or trajectories) with only one stay point increased from 73 to 77, 80, and 79 for $k = 2$, $k = 4$, and $k = 6$, respectively (see Figure 7.5a). In contrast, the number of trajectories with exactly two stay points decayed from 92 to 77, 79, and 83 for $k = 2$, $k = 4$, and $k = 6$, respectively. The private cars dataset of Shenzhen city, depicted in Figure 7.5b, also increased the number of trajectories with only one stay point, which had an increase of 562 users to 761, 810, and 754 with only one stay point for $k = 2$, $k = 3$, and $k = 4$, respectively. This fact occurs because, with the anonymization, the trajectories are sliced into sub-trajectories minors. Consequently, stay points can be separated in these trajectories resulting in tra-



(a) San Fran. anonymized with k=2,4,6. (b) Shenzhen anonymized with k=2,3,4.

Figure 7.5: Spatial, temporal, social, and privacy metrics extracted from the San Francisco (cabs) and Shenzhen (private cars) datasets.

jectories with few stay points. Alternatively, with the slicing of trajectories, the number of their points can be insufficient to identify the stay points.

Regarding the PSEU of Cabspotting dataset (Figure 7.5a), the $k = 2$ setup is

the lower privacy level but is responsible for the highest data distortion, with 62% of the trajectories having several pseudonyms greater than 5. Specifically, 515 (26% of trajectories), 433 (22%), and 234 (11%) trajectories changed their pseudonyms 5, 6, and 7 times, respectively. This means these trajectories passed and were anonymized at least four mix-zones or passed at least four times through a mix-zone. Therefore, the probability of re-identification (P_{Rind}) for trajectories with five pseudonyms is $P_{Rind} = 1/16$, that is, $P_{Rind} = 1/2 \times 1/2 \times 1/2 \times 1/2$, for six and seven pseudonyms, the P_{Rind} decay to 1/32 and 1/64, but the utility is compromised. As for $k = 4$, trajectories with at least 2 and 3 pseudonyms prevailed, totaling 54% of anonymized trajectories. Finally, for $k = 6$, 62% of the trajectories around 1232 were not anonymized due to the high mix-zones activation level to anonymize. Similar to the Cabspotting dataset, the Shenzhen dataset (Figure 7.5b), protected with $k = 2$, had more slice trajectories resulting in 2153 trajectories with two pseudonyms, 519 trajectories with three pseudonyms, 70 trajectories with four pseudonyms, and six trajectories with five pseudonyms. Few vehicles were anonymized for the $k = 4$ setup, and the pseudonym rate was minimal.

The RRET metric for the Cabspotting dataset corroborates with PSEU analysis. We can observe a significant risk increase as k grows. Initially, the RRET for the original dataset has a value of 1. This value means that the trajectories are entirely unprotected. For $k = 2$, the RRET drops significantly, with 97.3% of the trajectories having an RRET equal to or less than 50%. However, for $k = 2$ and $k = 6$, the RRET tends to increase (see Figure 7.5a), and 75.5% and 37.2% of the trajectories for $k = 2$ and $k = 6$ have an RRET equal to or less than 50%. The increase in RRET is because the higher the value of k, the lower the probability of anonymizing a quantity equal to or greater than k vehicles. The increase in RRET is even more evident for the Shenzhen dataset, which has a lower density than Cabspotting and affects anonymization performance. For example, for $k = 2$, $k = 3$, and $k = 4$, only 37.6%, 7.3%, and 0.8% of the trajectories have an RRET equal to or less than 50%.

Regarding the TDSL metric, the anonymization dataset of Cabspotting had significant distortion from the original dataset (Figure 7.5a). Cab trips are sliced into smaller portions, creating trips over shorter distances from 400 km to 180, 200, and 300 for $k = 2$, $k = 4$, and $k = 6$ setups, respectively. Additionally, the number of trips grew from the original trips. For example, $k = 2$ and $k = 4$ have values above 10^3 trips, while for the original trajectories, they are in the order of 10^2. Also, in the anonymized data, there was a reduction in the size of the trips. For the original dataset, most trips have up to 30% of the length of the longest trip, while for $k = 2$ and $k = 4$, both are 20% of the length.

The TDSL of the Shenzhen dataset (Figure 7.5b) also had significant distortion between original and anonymized trajectories. For instance, the number of trips grew up to 5 km in length from zero to 10^3, 10^2, and ten trips for $k = 2$, $k = 3$, and $k = 4$ setups. For trips around 10 km, the number of trips decreased from 10^3 to 10^2 for the $k = 2$.

In addition, it reduced trips higher than 15 km in the $k = 2$ compared with the noanon dataset.

In TRIPTIME in the Cabspotting dataset (Figure 7.5a), the anonymization effect produced a much larger number of trajectories with a duration close to zero than the original trajectories. For example, for $k = 2$ and $k = 4$ setups, the number of trajectories near 10 minutes peaked to 10^4 and 10^3 compared to original trajectories on the order of 10^2. The Shenzhen dataset had a similar behavior in which the number of trips until 10 minutes of duration increased from zero to 10^3 trips for three setups (Figure 7.5b).

Anonymizing Cabspotting trajectories also significantly affected the NUMVIS metric. For the original dataset, around 96% of trajectories had more than 200 visits. Meanwhile, for $k = 4$ and $k = 6$, this percentage decreased to 12% and 34%, respectively. The most significant discrepancy was for $k = 2$, where only 3% of the trajectories had a NUMVIS value equal to or greater than 200 visits. Although the Shenzhen dataset presents only around 10% of the NUMVIS compared to Cabspotting, it is still possible to identify discrepancies about the original dataset. For example, around 75.8% of trajectories have a NUMVIS of no more than 50. However, for $k = 2$ and $k = 3$, the number of trajectories increases to 88.8%, and 78% have NUMVIS up to 50, for the $k = 4$ maintained with a distribution similar to the original dataset.

The social metrics group also was affected by anonymization. For instance, the MAXCON for the Cabspotting dataset, the anonymization with $k = 2$ was the setup that caused the most data distortion, with an increase in connections in the scale from 10^2 to 10^3 pairs, and the maximum number of relationships between pairs increased from 2000 to 3000 connections. For $k = 4$, there was distortion between the original data but with a lower intensity than $k = 2$, reaching the order of 10^2 user pairs with the number of connections around 3000. The $k = 6$ had less distortion than the original data; even so, it obtained several connections around 2000. The Shenzhen dataset behaved similarly to the Cabspotting dataset regarding anonymization, in which the $k = 2$ prevailed in distortion terms. The MAXCON increased from 1000 to 1500, for $k = 4$ and $k = 6$ had a similar distribution to the noanon dataset.

Regarding the CODU metric for Cabspotting (Figure 7.5a), the number of user pairs that made contacts with a duration close to zero increased from 10^5 to 10^6 in all privacy settings. Furthermore, for $k = 2$, there was a significant increase in the number of pairs, from 10^3 to 10^5, with contact durations of up to 70 secs. This behavior can also be observed in $k = 4$ and $k = 6$, but in smaller pairs of contacts, from 10^3 to 10^4. Although the metric does not present a level of distortion like Cabspotting, the distributions of the anonymized versions for Shenzhen' CODU metric followed the same behavior: wherein $k = 2$ had greater distortion than $k = 3$ and $k = 4$ (see Figure 7.5b). Also, this behavior can be observed in other metrics, such as INCO and CONEN.

Finally, the RGYR metric had a smooth distortion with the anonymization. For

the Cabspotting dataset, we can observe that $k = 2$ and $k = 4$ had an increase of trajectories that had an RGYR lower or equal to 3 km than the original trajectories (see Figure 7.5a). Specifically, it increased 96.4% of original trajectories to 99.8% and 98.5% for k=2 and k=4, respectively, with RGYR lower or equal to 3 km. The $k = 6$ was maintained to be equal to the original trajectories, around 96.3%. The Shenzhen dataset had a lower RGYR value than Cabpotting, limited up to 97% of trajectories with RGYR not more than 200 meters. Also, the anonymized datasets did not differ significantly from the original dataset.

7.5.3 Distortion Level Analysis

Table 7.4 represents the distortion level (WS) of the metrics between original and anonymized datasets of Cabspotting and Shenzhen varying the k value. Overall, the low, intermediary, and high privacy levels had descending distortion for all metrics in both datasets, as shown by an average of k value ($\text{Avg}(k = i)$), where $i \in K = \{2, 3, 4, 6\}$ in Table 7.4. For instance, the WS for the SPC of the Cabspotting dataset, in which the distribution of the pair $\langle noanon, k = 2\rangle$ had more data distortion than $k = 4$ and $k = 6$ setups. In other words, $\langle noanon, k = 2\rangle$ had less utility with 0.0720, but $\langle noanon, k = 6\rangle$ had low distortion and high utility with 0.0088. The reason is that the $k = 6$ has lower chances of anonymization. Thus, many trajectories pass through the mix-zones without being anonymized, which makes the distribution of $k = 6$ metrics similar to the original dataset. Identical to the Cabspotting, the SPC of Shenzhen dataset, the pair $\langle noanon, k = 2\rangle$ had more distortion WS equals 0.036. But, with the slightest distortion, was the pair $\langle noanon, k = 3\rangle$ with WS equals 0.016.

Concerning the privacy metrics, RRET got the highest distortion between anonymized and original datasets about all metrics in both datasets. Highlight for Cabspotting pair $\langle noanon, k2\rangle$ with 0.86 of distortion. Then, the distortion level decreases in the pairs $\langle noanon, k4\rangle$ and $\langle noanon, k6\rangle$ to 0.66 and 0.32 but maintains the highest in all setups. The same behavior is observed in the Shenzhen dataset, in which the distortion level is high to low in the pairs $\langle noanon, k2\rangle$, $\langle noanon, k3\rangle$, and $\langle noanon, k4\rangle$, respectively. The high distortion is because, in $k = 2$, we have more anonymization probability than in other setups. For instance, for $k = 2$ in Cabspotting, the RRET is not more than 0.5 for 97% of trajectories, while for $k = 6$, we have just 37% of trajectories with RRET equal to 0.5 (see Figure 7.5a).

For the PSEU distributions of the Cabspotting dataset, the less similarity between original and anonymized data was the pair $\langle noanon, k = 2\rangle$ with WS of 0.23, which

indicates less utility in the open data context, since most trajectories for $k = 2$ had 5 or 6 pseudonyms (see Figure 7.5a). In contrast, trajectories anonymized with $k = 4$ presented greater utility with a prevalence of trajectories with 2 and 3 pseudonyms and *WS* of 0.19. The Shenzhen dataset had a similar behavior to Cabspotting, which had a higher anonymization setup and little data distortion. The pair $\langle noanon, k = 2\rangle$ had the major distortion, and the pair $\langle noanon, k = 4\rangle$ had the less distortion with *WS* equals 0.0046.

Regarding the TDSL for both Cabspotting and Shenzhen datasets, the higher distortion was with $k = 2$, with *WS* to 0.2 and 0.083, respectively. In contrast, the smallest data distortion was with $k = 6$ and $k = 4$ with *WS* equal to 0.14 and 0.014, respectively. Among all metrics, the TDSL had the highest distortion for Shenzhen.

The NUMVIS metric also had similar behavior to spatial metrics, with the greatest to most minor distortion about the original data, $k = 2$, $k = 4$, and $k = 6$, respectively. Mainly, Cabspotting had more accentuated than Shenzhen, such as the pair $\langle noanon, k2\rangle$ with 0.23 against 0.057.

In the TRIPTIME metric for the Cabspotting dataset, the $\langle noanon, k = 2\rangle$ pair had more distortion above all metrics with *WS* equal to 0.38, and the $\langle noanon, k = 6\rangle$ pair had less distortion with *WS* similar to 0.26 (see Table 7.4). Likewise, the pair $\langle anon, k = 2\rangle$ in the Shenzhen dataset had a high data distortion, and low distortion was the pair $\langle noanon, k = 3\rangle$ pair (intermediary privacy level) with *WS* equal to 0.039 and 0.0044, respectively.

Concerning the social metrics, the MAXCON in the Cabspotting dataset, $k = 4$ setup had more distortion than the original trajectory from the social category, with a *WS* of 0.21. The $k = 2$ and $k = 6$ had the same *WS* equal to 0.15. Unlike Cabspotting, the Shenzhen dataset had more data distortion on the $k = 2$ setup with a *WS* of 0.015 and less on the $k = 4$ setup with 0.003.

The CODU and INCO metrics for Cabspotting had high and low distortions with $\langle noanon, k = 2\rangle$ and $\langle noanon, k = 6\rangle$ pairs, respectively (see Table 7.4). But in the Shenzhen dataset, the high distortion for these metrics was $\langle noanon, k = 2\rangle$ and $\langle noanon, k = 3\rangle$ pairs with *WS* equal to 0.0140 and 0.0056. And the smallest distortion was $\langle noanon, k = 4\rangle$ for both metrics. The RGYR was the metric with lower distortion than all metrics in both datasets. For the Cabspotting, the pairs $\langle noanon, k2\rangle$ and $\langle noanon, k4\rangle$ got 0.029 with 0.028. RGYR was also lower for the Shenzhen dataset, with 0.013 and 0.0009 for $\langle noanon, k2\rangle$ and $\langle noanon, k4\rangle$.

Table 7.4: Distortion level (WS) of the metrics between original and anonymized datasets of Cabspotting and Shenzhen.

	Cabspotting				Shenzhen			
Metric	$k=2$	$k=4$	$k=6$	Avg($k=\langle 2,4,6\rangle$)	$k=2$	$k=3$	$k=4$	$Avg(k=\langle 2,3,4\rangle)$
PSEU	**0.2300**	0.1900	0.1100	**0.1767**	**0.1200**	0.0200	0.0046	**0.0482**
RRET	**0.8600**	0.6600	0.3200	**0.6133**	**0.2100**	0.0560	0.0072	**0.0911**
SPC	**0.0720**	0.0490	0.0088	0.0433	**0.0360**	0.0160	0.0300	0.0273
SPD	0.0028	**0.0035**	0.0025	**0.0029**	**0.0830**	0.0330	0.0280	**0.0480**
TDSL	**0.2000**	0.1900	0.1400	0.1767	**0.0830**	0.0190	0.0140	0.0387
TRIPTIME	**0.3800**	0.3400	0.2600	**0.3267**	**0.0390**	0.0044	0.0200	**0.0211**
NUMVIS	**0.2300**	0.2100	0.1500	0.1967	**0.0570**	0.0082	0.0100	0.0251
MAXCON	0.1500	**0.2100**	0.1500	**0.1700**	**0.0150**	0.0064	0.0033	0.0082
CODU	**0.1100**	0.0670	0.0250	0.0673	**0.0140**	0.0089	0.0010	0.0080
INCO	**0.0820**	0.0610	0.0330	0.0587	0.0019	**0.0056**	0.0044	**0.0040**
CONEN	0.1100	0.0830	**0.1200**	0.1043	**0.0250**	0.0082	0.0054	**0.0129**
RGYR	**0.0290**	0.0280	0.0083	**0.0218**	**0.0130**	0.0012	0.0009	0.0050
Avg($k=x$)	**0.2047**	0.1743	**0.1106**		**0.0581**	0.0156	**0.0107**	

Table 7.5: Matrix of Criteria J- Utility Metrics of Cabspotting dataset.

Criterion	PSEU	RRET	TDSL	TRIPTIME	MAXCON	CODU
PSEU	1.000	5.500	0.963	1.242	0.906	0.865
RRET	0.182	1.000	0.175	0.226	0.165	0.157
TDSL	1.039	5.714	1.000	1.290	0.941	0.899
TRIPTIME	0.805	4.429	0.775	1.000	0.729	0.697
MAXCON	1.104	6.071	1.063	1.371	1.000	0.955
CODU	1.156	6.357	1.113	1.435	1.047	1.000

Table 7.6: Matrix of criteria J - Utility metrics of Cabspotting dataset normalized to Saaty's scale. The CODU line (green) is the best-to-others criteria and the RRET column (yellow) is the others-to-worst criterion.

Criterion	PSEU	RRET	TDSL	TRIPTIME	MAXCON	CODU
PSEU	1.000	8.000	2.000	2.000	2.000	2.000
RRET	1.000	1.000	1.000	1.000	1.000	1.000
TDSL	2.000	8.000	1.000	2.000	2.000	2.000
TRIPTIME	2.000	7.000	2.000	1.000	2.000	2.000
MAXCON	2.000	9.000	2.000	3.000	1.000	2.000
CODU	2.000	9.000	2.000	3.000	2.000	1.000
w_j^*	**0.194**	**0.032**	**0.161**	**0.129**	**0.290**	**0.194**
ξ^{L*}	**0.167**					

Table 7.7: App. Domains (alternatives) vs. utility metrics (criteria) as weights.

App. Domain	PSEU	RRET	TDSL	TRIPTIME	MAXCON	CODU
Statistical Analysis	$5.000\times u_1$	$2.000\times u_2$	$9.000\times u_3$	$8.000\times u_4$	$4.000\times u_5$	$2.000\times u_6$
Urban Planning	$3.000\times u_1$	$3.000\times u_2$	$9.000\times u_3$	$8.000\times u_4$	$4.000\times u_5$	$2.000\times u_6$
Driver Behavior	$6.000\times u_1$	$7.000\times u_2$	$8.000\times u_3$	$6.000\times u_4$	$3.000\times u_5$	$4.000\times u_6$
Social Networks	$4.000\times u_1$	$2.000\times u_2$	$3.000\times u_3$	$4.000\times u_4$	$9.000\times u_5$	$8.000\times u_6$
Opportunistic Network	$3.000\times u_1$	$1.000\times u_2$	$4.000\times u_3$	$2.000\times u_4$	$9.000\times u_5$	$9.000\times u_6$
Targeted Market	$7.000\times u_1$	$5.000\times u_2$	$6.000\times u_3$	$2.000\times u_4$	$4.000\times u_5$	$3.000\times u_6$
Cybersecurity	$9.000\times u_1$	$9.000\times u_2$	$7.000\times u_3$	$6.000\times u_4$	$1.000\times u_5$	$2.000\times u_6$
u_j	**0.77**	**0.14**	**0.8**	**0.62**	**0.85**	**0.89**

Table 7.8: App. Domains vs. utility metrics scaled to Saaty's scale [1 to 9].

App. Domain	PSEU	RRET	TDSL	TRIPTIME	MAXCON	CODU
Statistical Analaysis	5.000	2.000	9.000	6.000	5.000	3.000
Urban Planning	4.000	2.000	9.000	6.000	5.000	3.000
Driver Behavior	6.000	2.000	8.000	5.000	4.000	5.000
Social Networks	4.000	2.000	4.000	4.000	9.000	9.000
Opportunistic Network	4.000	1.000	5.000	3.000	9.000	9.000
Targeted Market	7.000	2.000	6.000	3.000	5.000	4.000
Cybersecurity	8.000	3.000	7.000	5.000	2.000	3.000

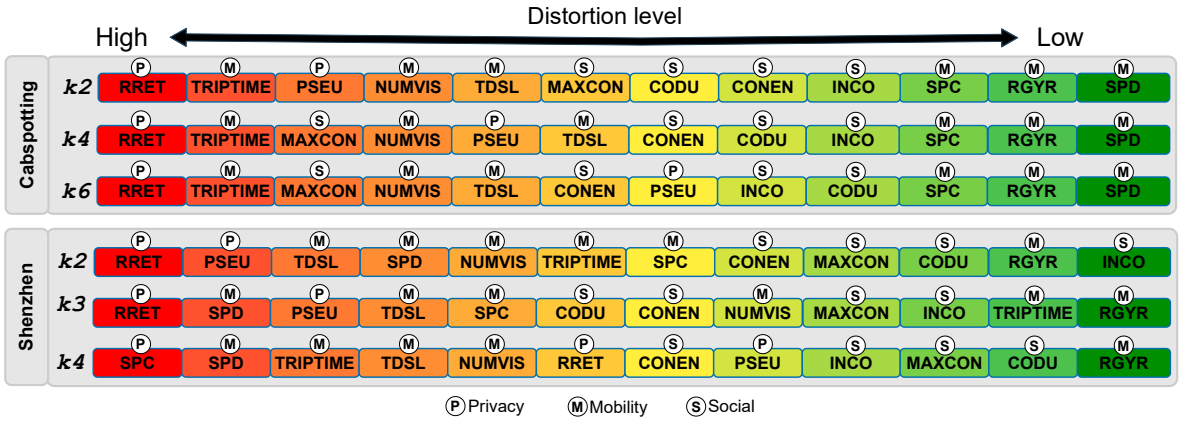


Figure 7.6: Distortion metrics ranking for trajectory datasets.

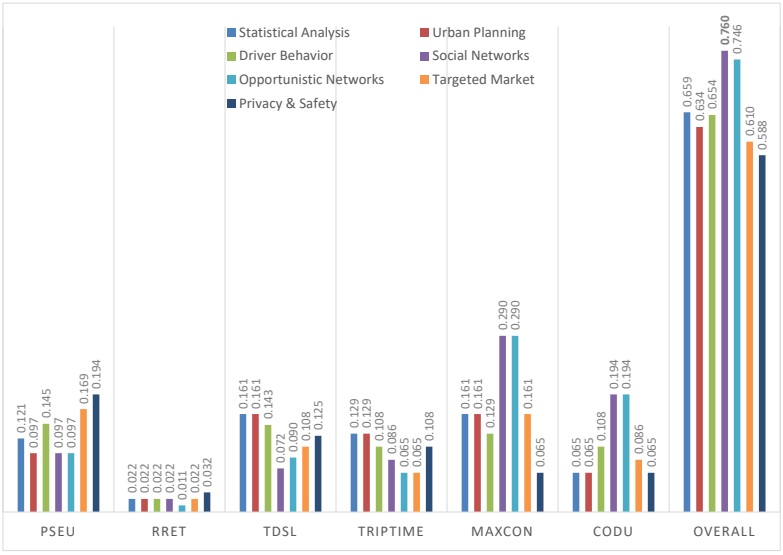


Figure 7.7: Ranking of smart cities application domains generated with BWM method for Cabspotting dataset. The metrics are the criteria, and the application domains are the alternatives.

7.5.4 Data Utility Driven to Smart Cities Domains

To answer the questions in Section 7.1, we must analyze the metrics by groups and verify which groups had the greatest data distortion. This way, we can identify which metrics are closely related to an application context that has been more or less affected.

In general, the $k = 2$ setup had more prevalence than other privacy setups regarding data distortion for the metrics. For both datasets, the metrics groups that were most affected were privacy, mobility, and social. Specifically, in the $k = 2$ setup for the Cabspotting dataset, the metrics with more distortion of the original trajectories were RRET, TRIPTIME, PSEU, and TDSL (see Figure 7.6), which are mobility metrics intimately related to applications of statistical analysis, urban planning, route plan, and smart mobility testbed (see Table 7.2). The results suggest that the anonymized trajectories by mix-zones with $k = 2$ setup have less utility for the applications above. In contrast, the

SPC, RGYR, and SPD metrics had less distortion than other metrics. It means that the anonymized data by mix-zones can be useful in the application domain related to these metrics. The application contexts related to these metrics are urban planning, including POI mining, EV charging deployment, and statistical analysis (see Figure 7.6). The social metrics - MAXCON, CODU, CONEN, and INCO - had intermediary distortion levels with $k = 2$, i.e., applications such as data dissemination protocols, including developing communication protocols in V2V and V2I networks, can be partially compromised when using data protected by mix-zones with $k = 2$ setup.

The Shenzhen dataset anonymized with $k = 2$ setup also had privacy, mobility, and social metrics with more data distortion than the original trajectories. The metrics with high distortion were RRET, PSEU, TDSL, SPD, NUMVIS, and TRIPTIME metrics, respectively (see Figure 7.6). It suggests that applications for urban planning and driver behavior may have degradation of service quality in using the dataset anonymized by mix-zones, as represented in Table 7.2. On the other hand, social metrics - CONEN, MAXCON, CODU, and INCO - had fewer data distortion, which suggested that the Shenzhen dataset protected with $k = 2$ can be more useful to applications, such as social networks and dissemination protocols for VANETs. The above discussion summarizes the answer to question **Q1**.

The answer to question **Q2** is to sort the order of the metrics by the distortion level for each k setup of Table 7.4. The order of the metrics in the ranking may differ, but the order of the group of metrics varies very little (see Figure 7.6). The ranking order can change a few for dense datasets regarding traffic volume, which are more likely to be anonymized. For example, with the k variation in the Cabspotting dataset, the order was not changed from the first and last positions in the ranking, TRIPTIME and SPD. However, for less dense datasets, such as Shenzhen, there was a change in the positions of the metrics in the ranking, but there was little change in the order of the group of ranking metrics. The first in the ranking for the $k = 2$, $k = 3$, and $k = 4$ were RRET, RRET, and SPC, respectively. However, we can observe a clear separation between the mobility and social groups. The mobility metrics group distorted more than the social metrics group for all three setups.

Regarding question **Q3**, which metrics have significant and little distortion with the k variation? The *WS* average between k setups in Cabspotting, the mobility metrics RRET and SPD are highlighted with the highest and least average distortion about all metrics, with $\text{Avg}(k = \langle 2, 4, 6 \rangle)$ of 0.6133 and 0.0029, respectively (see Table 7.4). These results suggest that using an anonymized dataset for statistical analysis like traffic monitoring and route planning can be compromised. In contrast, using anonymized datasets in the urban planning domain, e.g., POI mining can be more helpful, even with k variation. In the Shenzhen dataset, applications that RRET is a sensitive feature, such as multimodal integration in urban planning, can be less useful because RRET got the highest

distortion with Avg($k = \langle 2, 3, 4 \rangle$) of 0.0911. However, application segments based on social or opportunistic networks can be attractive because social metrics had less distortion, including INCO with the least Avg($k = \langle 2, 3, 4 \rangle$) of 0.0040.

Another point we highlighted was that datasets of different types present different behaviors in terms of utility, in answer to question **Q4**. For example, regarding the distortion ranking, Cabspotting's mobility metrics group concentrated on the extremes of the ranking, and social metrics remained in the middle. While for the Shenzhen dataset, the mobility metrics had more significant distortion than the social ones. In this way, the anonymized versions of Cabspotting may be more useful in the application of dissemination protocols, statistical analysis, and drive behavior. At the same time, the Shenzhen dataset may be more relevant for developing applications related to social networks and for developing data dissemination protocols even more than Cabspotting.

7.5.5 Ranking of Smart City Application Domains

This section presents the ranking smart cities application domains analysis with our proposal of BWM, in answer to question **Q5**. For validation, we analyzed the dataset Cabspoting protected with the low privacy level $k = 2$, but with the most significant coverage in anonymization terms. To study collateral effects caused by the anonymization process regarding privacy, spatial, and social terms, we selected two metrics of each aspect that had more distortion: PSEU, RRET, TDSL, TRIPTIME, MAXCON, and CODU. The variation of these metrics enables identifying which application domain can be more and less affected. The threshold of consistency index defined for this analysis is $\lambda = 1$, representing the upper bound, as suggested in [3].

Table 7.5 presents the result of the criteria weights definition of the metrics made of our algorithm. We can note that the J cells $\langle CODU, RRET \rangle$ and $\langle RRET, CODU \rangle$ had the highest and lowest weight levels, with 6.357 and 0.157, respectively. In the scale changing, denoted in Table 7.6, these cells were altered to Saaty's scale with values 9 and 1, respectively.

The best and worst criteria selection occurs based on the utility (or distortion) level identified in the utility metrics. In the context of the utility of anonymized data collection, in an ideal situation, the metrics should have the highest possible utility (i.e., the complement of the distortion level of the original data). In this way, we selected the best and worst metrics and those with the greatest and least utility within the set of metrics. When observing the level of distortion shown in Table 7.4, we can conclude that within the set of selected metrics, the best and worst metrics were those with the

lowest and highest levels of distortion represented by the CODU and RRET metrics, respectively. Table 7.6 shows the choice of the preference of the best criterion CODU (best criterion) over all the other criteria, represented by CODU row (in green). We can see that CODU is more relevant to RRET with a value of 9 and slightly greater importance to the MAXCON, TDSL, and PSEU with a value of 2. Additionally, Table 7.6 shows the preference of all the criteria over the RRET (worst criterion) in the RRET column (in yellow). We can note that MAXCON and TDSL are more relevant than RRET, with values of 9 and 8, respectively.

The w_j^* and ξ^{L*} depicted in Table 7.6 represent the optimal weights and indicator of consistency calculated with the best and worst criteria of the BWM. ξ^{L*} has a value of 0.167, close to zero, indicating that the pairwise comparison consistency level is acceptable. For the optimal weights, MAXCON was highlighted with 0.290, followed by CODU and PSEU metrics with 0.194. In contrast, RRET had a lower optimal weight of 0.032. The weight prevalence of social metrics indicates that an application domain that uses an anonymized dataset that uses social metrics can be more useful than one that uses privacy metrics.

After calculating the optimal weights, the next step is to define the priority of alternatives considering the criteria. Particularly, each alternative is a smart city application domain i concerning criterion j, represented by metrics, i.e. (p_{ij}). Thus, the DM must define a score of 1 to 9, which is the importance of the metrics in each domain and application. The DM can determine the score from the prevalence of such metrics in related work to the context of each application domain. In each application domain of smart cities, comprehensive research indicates the uses of spatial, social, and privacy metrics, as shown in Table 7.2. So, the DM can infer what metric can be more or less critical for an application domain than another and assign a score for each one, as denoted in Table 7.7. Additionally, each criterion is considered as an additional weight, so each cell p_{ij} of the matrix A is multiplied by the utility level of criterion j, i.e. u_j in the Table 7.7. After applying the utility metrics as weights, the scale change to Saaty's scale is applied (see Table 7.8). Then, each column from the matrix A is normalized by dividing the cell p_{ij} by the maximum value of each criterion found in its respective column, $\max(p_{ij})$.

The ranking of application domains is shown in Figure 7.7, which is obtained by multiplying the columns of the matrix A by the optimal weights w_j^* of their respective criteria (defined in Table 7.6). In Figure 7.7, the utility metrics group Smart city application domains. On average, social metrics had the most utility in application domains, followed by spatial and privacy metrics. The metric with the lowest utility among the application domains was RRET, highlighted on Opportunistic Networks with a value of 0.011. On the other hand, RRET has the highest utility for Privacy & Safety with 0.032. The MAXCON metric represents the most utility for application domains, highlighting the Social Networks and Opportunistic Network domains, with a value of 0.290. The least

useful domains would be the Privacy & Safety domains, which have a value of 0.065.

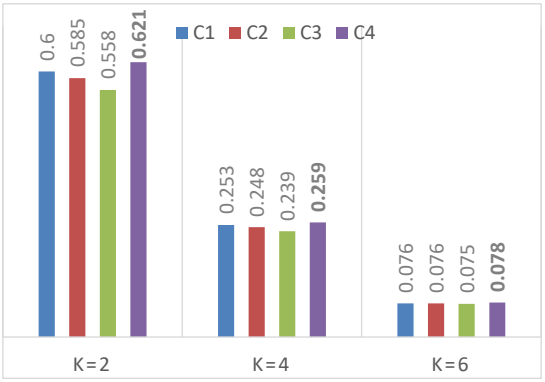
Overall, the application domain ranking in terms of usefulness was as follows: Social Networks ≻ Opportunistic Networks ≻ Statistical Analysis ≻ Driver Behavior ≻ Urban Planning ≻ Targeted Market ≻ Privacy & Safety. Highlighting the domains that prevailed in social metrics, Social Networks, and Opportunistic Networks, with 0.760 and 0.746, respectively. In conclusion, applying mix-zones as a form of protection was less harmful for social and spatial domains than for the Targeted Market and Privacy & Safety domains, which had values of 0.610 and 0.588, respectively.

7.5.6 Utility of Anonymized Data for Specific Applications

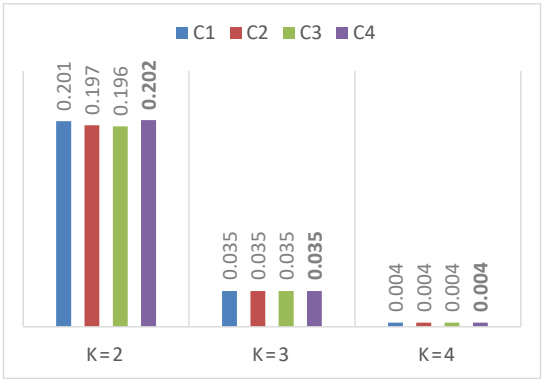
Table 7.9: The utility of the anonymized datasets of Cabspotting and Shenzhen for specific applications.

				Association between the proposal and utility metrics	
Case	Ref.	**Smart City Domain**	**Application/Technique**	**Proposal Metrics**	**Utility Metrics and Saaty's Scale**
C1	[282]	Urban Planning and Opportunistic Networks	EV charging planning and data dissemiination protocol	number of traffic lights average road speed distance to destination distance between the entities connection time between the entities –	– TDSL (9), TRIPTIME (9) TRIPTIME (9) TDSL (9), GlsSPC (5), SPD (5), RGYR (3) CODU (9), CONEN (8), INCO (5), MAXCON (5) PSEU (1), RRET (1), NUMVIS (1)
C2	[283]	Urban Planning	EV charging planning/ postioninng	pseudonym trajectory timestamps of travel departure time interval total travel time origin zone destination zone total travel distance vehicle type –	PSEU (4), RRET (4) TRIPTIME (9) TRIPTIME (9) TRIPTIME (9) SPC (8), SPD (8) SPC (8), SPD (8) TDSL (9), RGYR (3) – NUMVIS (1), MAXCON (1), CODU (1), INCO (1), CONEN (1)
C3	[284]	Cybersecurity	Location privacy-preserving Spatial crowdsourcing (SC) for IoV based on blockchain	pseudonym stay points number gyration radius trajectory features –	PSEU (8), RRET (8) SPC (6), SPD (6) RGYR (8) TDSL (9), TRIPTIME (9) NUMVIS (1), MAXCON (1), CODU (1), INCO (1) CONEN (1)
C4	[285]	Opportunistic Networks	Model mobility simulator	pseudonym trajectory features POIs contact duration inter-contact time repeated contacts –	PSEU (4), RRET (4) TDSL (9), TRIPTIME (5), RGYR (5) SPC (9), SPD (8) CODU (9) INCO (9), MAXCON (8), CONEN (8) NUMVIS (1)

One of the critical points about utility is obtaining the utility of protected data for a specific application (**Q6**). To answer this question, we analyze the utility of the dataset Cabspotting and Shenzhen anonymized with privacy levels $[k = 2, 4, 6]$ and $[k = 2, 3, 4]$, respectively, applying the Algorithm 6 to four study cases, that represent a specific application of a distinct application domain: C1 is a managing EV charging planning; C2 is an EV charging infrastructure planning; C3 is a decentralized location privacy-preserving SC for IoV based on a blockchain scheme; and C4 is a mobility model in terms of number of nodes and simulation time. First, we detail each proposal, associate their



(a) Cabspotting dataset, protected with mix-zones configured with $k = \{2, 4, 6\}$.



(b) Shenzhen dataset, protected with mix-zones configured with $k = \{2, 3, 4\}$.

Figure 7.8: The utility of anonymized trajectories for the applications C1, C2, C3, and C4.

metrics with utility metrics, and define the weights of each utility metric. Next, we discuss the utility results of each case, generated by Algorithm 6.

In the case study **C1**, Bautista et al.[282] introduced a Charging station (CS) scheme that aims to minimize the total trip time of the user. They reduce the total trip duration considering traffic conditions on the EV's route toward its destination, assuming an intermediate recharging stop at the selected CS. Additionally, to aid the CS management scheme, they proposed a communication framework among EVs and RSUs to interchange charging service messages. When anonymized mobility data by mix-zones is used to test the proposal, we can assign the proposal metrics with utility metrics as follows. The metrics of managing the EV charging planning scheme are intimately related to traffic conditions, EV's trip time, and communication delay of the VANET routing protocol. Specifically, the number of traffic lights, average road speed, and driving distance are associated with the utility metrics: TDSL, TRIPTIME defined in Table 7.9 and receive high weight. Regarding the message dissemination protocol scheme, it considers the vehicles' traffic, distance to destination, and connection time between the entities (RSU and vehicles), so the utility metrics TDSL, CODU, CONEN, MAXCON, and INCO receive high-weight values. In contrast, the other utility metrics that don't have a direct association receive a low weight.

The second case study (**C2**) is an EV charging planning proposed by Kavianipour et al. [283]. They presented a framework for urban charging infrastructure positioning that considers fast charging to address the range anxiety issue of EVs in the urban environment. Specifically, the framework tackles a lack of a dataset that includes daily chains of trips for all travelers and the availability of level 2 chargers at each intermediate destination. For this, the framework proposed a simulation tool to generate trip trajectories. Next, the trajectories generated are input for building a charging behavior simulation tool to assign the stochastic initial state of charge for each vehicle trajectory according to the departure time, trip purpose, and land use characteristics at the origin. Then, charging

demand is the input to a mixed-integer nonlinear program that seeks to charge station configuration to minimize the total system cost.

In the case study **C2**, using the anonymized mobility dataset protected by mix-zones can directly affect the output of traffic simulation that generates trajectories once this component uses the Origin-Destination (OD) demand features and road network properties as inputs. So, the trajectories' slices made by mix-zones can affect OD features because part of the anonymized trajectory will have a pseudonym different from the original, producing a different destiny. Additionally, mix-zones can affect other key features, including traveled paths and travel time stamps along each vehicle's path; the latter is the average zone-to-zone travel distances and times. Notably, these features are used in the posterior components of the framework. Also, using the anonymized dataset to test intermediary framework steps, such as the charging behavior simulation component, can be potentially compromised using sliced trajectories. The Table 7.9 summarizes the key features of the proposal above associated with utility metrics SPC, SPD, TDSL, and TRIPTIME that receive a high weight. Although PSEU and RRET are not directly related to proposal metrics, they receive a medium weight because of their impacts on the framework.

The third case study (**C3**) is a decentralized location privacy-preserving SC for IoV based on a blockchain scheme proposed by Zhang et al. [284]. The approach allows vehicle users to participate in SC, ensuring the task's location policy privacy and providing multi-level privacy preservation for workers' locations. The proposal has the following steps: Location Record, Task Submission, Solution Submission, and Location Verification.

The purpose of the utility of the anonymized dataset is to test the proposal components, such as in the solution submission step by the workers to get the reward. Notably, in the solution submission step, the requester, responsible for proposing a task, sends a task request that contains the task's location policy to workers who decide whether or not to participate in the task. If the workers obey the location policy, they send the solution with its privacy-preserving location proof to the blockchain via the nearby RSU in the solution submission phase. The workers inform the obfuscated location based on the map grid size. The worker privacy level is proportional to the grid size and inversely proportional to the payment. A worker offered a more accurate location deserves a higher payment. However, the accurate location can bring privacy threats. In this way, the actual location of a worker j is obfuscated inside a cell of a grid of size n. It means that the GPS's registers and stay points of a trajectory of worker j must be obfuscated. An issue regarding the anonymizing trajectories, particularly by mix-zones, is which will slice the trajectories, yielding information loss when a worker j moves in the grid's cells. Considering this step of the framework proposed, the metrics that can impact its performance are pseudonym, stay, points number, gyration radius, and trajectory features, which are related to the utility metrics: PSEU, RRET, SPC, SPD, RGYR, TDSL, and TRIPTIME

that receive a high weight (see Table 7.9).

In the last case study **C4**, we explore using anonymized data in the context of opportunistic networks. Specifically, Sarmiento and Förster [285] presented a scalable mobility model in terms of number of nodes and simulation time, called TRAILS (TRAce-based ProbabILiStic Mobility Model), capable of imitating spatial dependence, geographic restrictions, and temporal dependence from real scenarios. As validation, they compared mobility metrics of TRAILS simulations (e.g., contact duration, inter-contact time, and repeated contacts), real traces, and another synthetic mobility model. In the TRAILS workflow, users travel to popular places extracted from real scenarios using real paths. A user selects a POI as its destination based on the POI's popularity from a POI list. A graph generator imports the POIs and paths extracted from traces to create a mobility graph imported to the simulator to control the users.

When using an anonymized dataset to validate TRAILS, such as extracting POIs and paths from an anonymized dataset, it may also suffer from the effect of anonymization on the original dataset. That is, slicing trajectories using mix-zones could significantly increase the number of trajectories and users in the dataset and reduce the size of a trajectory TDSL. Consequently, there will be an increase in the number of SPC, which can be used as a substrate for POI identification. Furthermore, with shorter trajectories, some links between two POIs in the graph cannot be reached when compared with the original data. Regarding TRAILS validation, mobility metrics related to contacts such as CODU, INCO, CONEN, and MAXCON may also be affected. Therefore, the metrics that receive greater weight are TDSL, SPC, SPD, CODU, INCO, CONEN, and MAXCON, as shown in Table 7.9.

After defining the weights of each utility metric, the Algorithm 6 is executed for each application for Cabspotting and Shenzhen datasets, and the utility results (χ) are shown in Figure 7.8. Regarding the level of utility of a dataset applied in different specific applications, application C4 stands out among the others, which obtained for $k = 2$, $k = 4$, and $k = 6$ a utility of 0.621, 0.259, and 0.078 about the original dataset with a utility value of 1, respectively. In other words, for the C4 application, with the effect of anonymization, the Cabspotting dataset degraded by approximately 37.9%, 74.1%, and 92.2% to the original dataset. For the Shenzhen dataset, the C4 application also stood out in utility and maintained the ascending order of k values with the utility of 0.202, 0.035, and 0.004 for $k = 2$, $k = 3$, and $k = 4$. This fact is also observed for all applications in which the greatest utility in the two anonymized datasets, which follow the increasing order of privacy setup with $k = 2$, $k = 4$, and $k = 6$ for Cabspotting and $k = 2$, $k = 3$, and $k = 4$ for Shenzhen.

Furthermore, we can observe that for both datasets, the utility ranking was respectively for applications C4 (opportunistic networks), C1 (urban planning and opportunistic networks), C2 (urban planning), and C3 (Privacy & Safety), shown in Figures 7.8a and

7.8b. This ranking corroborates the results obtained from the utility analysis for the application domains in Sub-section 7.5.5.

An important point to note is that, as the value of k increases (k =2, 4, and 6; k = 2, 3, and 4), the level of distortion decreases (see Sub-section 7.5.3) and also the anonymization utility decreases, which is a critical factor that affects the result of utility calculation. In this way, Cabspotting had a significant result about Shenzhen, which, despite having a low distortion with the variation of k, had a low utility, producing low utility values. Finally, the results suggest that for the C4 application, the Cabspotting dataset protected with $k = 2$ is the most appropriate, concerning all other setups, when considering the privacy factor and data utility.

7.5.7 Lessons Learned

We learned some lessons from the utility analysis of anonymized mobility data with UAFAT in finding the answer to questions of Section 7.1. In answer to question Q1, we analyzed the data utility with distortion of the metrics. We concluded that in both datasets, the metrics groups that were most affected were privacy, mobility, and social, respectively. For instance, the Cabspotting dataset, which was protected with $k = 2$, affected metrics related to applications such as statistical analysis, urban planning, route planning, and smart mobility testbed. These results suggest that the anonymized trajectories by mix-zones with $k = 2$ setup have less utility for the applications. In contrast, some mobility metrics, e.g., the SPC, RGYR, and SPD, had less distortion than others. So, application contexts related to these metrics include urban planning, POI mining, EV charging deployment, and statistical analysis.

Based on the analysis of distortion metrics with k variation, the answer to question Q2 is that the order of the metrics in the ranking may differ. Still, the order of the group of metrics varies very little. The order of distortion ranking can change less for dense datasets (Cabspotting) than for less dense datasets (Shenzhen).

The answer to question Q3 depends on the dataset. For instance, the WS average between k setups in Cabspotting, the mobility metrics RRET and SPD are highlighted with the highest and least average distortion about all metrics. Thus, using anonymized datasets in the urban planning domain, e.g., POI mining, can be more helpful than statistical analysis, even with k variation. The Shenzhen dataset had the same behavior, in which applications in which RRET is a sensitive feature, such as multimodal integration in urban planning, can be less useful. However, application segments based on social or opportunistic networks can be attractive because social metrics have less distortion,

including INCO.

Concerning question Q4, the answer is that datasets of different types present different utility behaviors. For instance, Cabspotting's mobility metrics group concentrated on the extremes of the distortion ranking, while social metrics remained in the middle. While Shenzhen's mobility metrics had more significant distortion than the social ones, there was a more defined separation between these groups of metrics.

In answer to question Q5, we ranked the smart cities application domains with our proposal of BWM, which uses the utility metrics related to domains and automatically defines criterion weight. It is worth emphasizing that the criterion weight is defined by a DM in the conventional MCDM technique. We applied our algorithm in the Cabspotting dataset protected with $k = 2$. It ranked the domains in terms of usefulness from high to low as follows: Social Networks $\succ$ Opportunistic Networks $\succ$ Statistical Analysis $\succ$ Driver Behavior $\succ$ Urban Planning $\succ$ Targeted Market $\succ$ Privacy & Safety.

Regarding question Q6, we proposed obtaining the utility of protected data for a specific application with an algorithm that considers the utility and mix-zones efficacy. The validation using the anonymized Cabspotting and Shenzhen datasets in four specific applications concluded that applications in opportunistic networks could be more useful, followed by applications in urban planning and Privacy & Safety domains. Note that as k increases, the distortion of data and anonymization efficacy decrease, which affects utility calculation. Another observation point is that the algorithm identified Cabspotting as more useful than the Shenzhen dataset for the case studies.

We summarize the impacts of anonymization on privacy, mobility, and social metrics that can affect utility. For datasets with low density, as k increases, the re-identification risk increases too because the higher the value of k, the lower the probability of anonymizing a quantity equal to or greater than k vehicles. In general, when k increases, there is an increase in the number of trajectories with: SPC close to one; SPD with duration less than SPD of the original dataset; RRET increases due to the anonymization performance; TDSL with sliced in smaller portions; TRIPTIME with a duration close to zero than the original trajectories; MAXCON with a high pair with connections, CODU with contacts with a duration close to zero; INCO with duration and CONEN in which entropy increases. It also increases the RGYR lower than 3 km. In contrast, PSEU and NUMVIS decreased as k increases. Overall, $k = 2$ significantly impacts the metrics, but as k increases, the metrics tend to the original dataset value due to low anonymization.

From the point of view of security of transmitted and shared information, our solution can have a positive impact with two layers of protection when encryption algorithms further encrypt anonymized data. A layer provided by symmetric or asymmetric encryption algorithms that encrypt anonymized trajectories. Another layer is provided by mix-zones, in which anonymized trajectories have a re-identification probability of $1/2$

in the worst case. However, the probability of re-identification tends to decrease substantially when a trajectory passes through more mix zones and when set up with k greater than 2, as discussed in subsection 7.5.2. Furthermore, designing a two-layer protection model could significantly mitigate surveillance and undue monitoring by unauthorized third parties, as transmitted data is anonymized and combined with other communications. Finally, it can substantially contribute to compliance with privacy regulations, such as GDPR.

In conclusion, although there is a discussion about the privacy and utility of open mobility data in a broad sense, more needs to be focused on the notion of utility in practical terms and how specific applications and domains can use anonymized data. Unlike previous works, our solution, in addition to providing privacy in terms of anonymization, considers the usefulness of this data in a practical way, such as directing it to specific domains and applications that could best use this data. Therefore, consuming such data for specific applications becomes possible, positively impacting security. Specific communication channels and publication platforms could be implemented to transmit this data to certified consumers in accordance with GDPR.

7.6 Concluding Remarks

This chapter explored the utility of data anonymization for smart cities. To achieve this, we introduced UAFAT, a framework designed to measure the utility aspects through twelve privacy, mobility, and social metrics, including mix-zone performance metrics from anonymized trajectories from cabs and private car datasets protected by mix-zones. The UAFAT enabled us to analyze which metrics are more and less affected by dataset versions anonymized by mix-zones configured in different privacy levels. It also allowed us to identify subtleties regarding the utility of these datasets. The UAFAT allowed us to solve the complex Multi-Criteria Decision-Making problem that ranks smart city application domains that best leverage mobility data anonymized by mix-zones. Also, it was possible to identify the utility level of the anonymized datasets for specific smart city applications and compare each other. Finally, the framework identified which one between cabs and private cars datasets has more utility for specific applications.

The results suggest that anonymized trajectories from the cabs dataset have more utility for applications associated with social metrics, such as dissemination protocol design, than urban planning and POI mining. Another insight is that with the privacy level variation, the utility type of a protected dataset can change, but it depends on the density of the dataset. Another fact is that the utility of anonymized trajectories must be eval-

uated together with mix-zones performance metrics, such as efficacy, to maximize both utility (distortion level) and coverage of anonymization. Concerning domain ranking, social networks, opportunistic networks, and statistical analysis as the top 3 smart cities application domains for an anonymized cabs dataset. Regarding the anonymized dataset applied in the specific applications, the model mobility simulation from the opportunistic networks domain got the best leverage of the anonymized trajectory cabs dataset, followed by applications in urban planning and privacy & safety domains. Specifically, for mobility simulation, the cabs dataset degraded in utility by approximately 37.9% ($k = 2$), 74.1% ($k = 4$), and 92.2% ($k = 6$) compared to the original dataset due to the low anonymization performance. Thus, the cabs dataset protected with $k = 2$ is the most appropriate concerning all other setups. Recall that the utility calculus considers mix-zones efficacy. Finally, the cabs dataset had a significant utility to the private cars dataset.

As far as we know, this proposal is the first study that analyzes the utility of anonymized trajectories by mix-zones with privacy, mobility, and social metrics, allowing the identification of smart city applications and services in which anonymized data will provide more or less utility. We believe that the utility characterization of anonymized data by mix-zones will present a new perspective on the utility of open data as a private data market for smart city design.

The next chapter presents a dynamic anonymization-based LPPM. Particularly, a dynamic mix-zone that considers the traffic conditions to tune the privacy level over time will be proposed, which outperformed the classical mix-zone schemes regarding privacy and quality of their anonymized data.

Chapter 8

k-DynMix: A Dynamic Privacy Protection in Mix-zones

This chapter presents our contributions regarding the anonymization-based LPPM design. We propose the k-DynMix, a dynamic mix-zone that tunes the privacy level over time in an online model with linear complexity, according to vehicles' traffic fluctuations, to achieve higher anonymization. The k-DynMix had presented an efficacy, anonymization rate, and AQ similar to the highest result of classical mix-zones. Furthermore, it maximized the privacy level to the best possible with the lowest re-identification rate, outperforming the classical mix-zones.

The remaining of this chapter is organized as follows. Section 8.1 brings an introduction and motivation regarding the importance of designing new dynamic mix-zones to improve mix-zones' performance in the Smart Mobility context. Section 8.2 discusses the related studies about mix-zones. Section 8.3 presents essential concepts and the problem of static mix-zones. Section 8.4 details the adversary model. Section 8.5 presents the k-DynMix. The metrics definition and experimental evaluation of the k-DynMix are in Section 8.6. Section 8.7 shows and analyzes the results from this work. Finally, Section 8.8 presents the final concluding remarks.

8.1 Introduction

In the age of pervasive computing, technologies like the IoT and IoV have facilitated the interconnection of objects and the sharing of location services across extensive environments such as smart cities, delivering numerous advantages to citizens [231]. Nonetheless, these services generate massive and unrestricted location data concerning citizens, raising significant privacy concerns. Through the analysis of location data, it becomes feasible to uncover latent insights regarding points of interest, individual and collective behavior from their mobility, and potentially even the identities of citizens [261]. To address users'

identity approaches such as LPPMs have been proposed as anonymization techniques like mix-zones. Mix-zones are designated anonymization areas defined by a radius r, within which entities change their pseudonyms based on privacy level k as a trigger function (e.g., when reaching a minimum of k entities simultaneously inside it) [170, 167].

Despite mix-zones being widely used in many areas, including VANETs, they still have some limitations. Notably, mix-zones heavily rely on factors such as positioning, geometry, mobility patterns, vehicle density, and arrival rates, which can impact their performance. For instance, misconfigured parameters of mix-zones can degrade the anonymization rate or protect with a low privacy level, enabling linking and inference attacks [261, 262]. Some proposals address these issues, but only a few focus on privacy levels to improve the mix-zones' performance and quality behavior.

This chapter proposes the k-Dynamic Mix-zone (k-DynMix), a dynamic mix-zones that tunes the privacy level k over time in an online model, with linear complexity, according to events like vehicles' traffic fluctuations to achieve the notion of higher anonymization. We inspired the k-DynMix on mix-zones quality metrics and some concepts of TCP congestion control mechanisms.

We conducted experiments, analyzing two datasets, one containing actual data and another containing synthetic data, comparing k-DynMix with two prediction mechanisms to explore the potentialities of estimating privacy levels over time and with classic mix-zones regarding mix-zone coverage, privacy metrics, and AQ. The results showed that k-DynMix outperformed the prediction mechanisms in predicting privacy level. Additionally, the k-DynMix got efficacy similar to the highest result of classical mix-zones. They maximized the privacy level to the best possible, having the lowest re-identification rate, outperforming the classical mix-zones. Unlike static mix-zones, k-DynMix got AQ for all mix-zones, showing better mix-zone behavior, including for low-traffic mix-zones. To the best of our knowledge, this is the first approach of dynamic privacy level over time in mix-zones that considers events vehicle traffic fluctuations to maximize the mix-zones privacy level, efficacy, and anonymization quality.

8.2 Related Studies

Notably, many proposals address critical issues on which mix-zones heavily rely, such as positioning, geometry, mobility patterns, vehicle density, and arrival rates [261, 167, 262]. Following, we present some relevant literature proposals that pursue these issues. Vehicular mix-zones are different from traditional mix-zones due to their numerous spatial and temporal constraints, including trajectories exclusively on physical roads,

directional headings, adherence to traffic regulations, traffic conditions, and road conditions [167, 200].

The initial studies into spatiotemporal considerations were pioneered by [170], who introduced the CMIX protocol for vehicular networks. CMIX employs mix-zone encryption at road intersections, guaranteeing dynamic pseudonym changes. Palanisamy and Liu [190] proposed a solution to address limitations in users' movement patterns and statistical behaviors within Mix-zones geometry. They proposed a non-rectangular-shaped, adaptive mix-zones, where the length is determined by factors like the average speed of the road segment, the time window, and the minimum pairwise entropy threshold.

To mitigate inference attacks facilitated by speed limitations at intersections and traffic lights where mix-zones are situated, Zhoun and Zang [100] introduced an LPPM strategy. This approach incorporates pseudonym swapping and adds trajectory noise to trajectories passing through mix-zones. About the static mix-zones infrastructure and their positioning, Yamazaki et al. [286] proposed a mix-zone scheme-centric vehicle that uses vehicles for anonymizing data instead of RSUs. They proposed resolving the communication delay between vehicles with two-stage mix-zones by removing the vehicles that will cause communication delay due to their dynamic mobility from pseudonym change processes.

Efforts to determine the value of k have also been made in k-anonymity approaches aimed at the IoT when applied to health Internet of Health Things (IoHT) to protect health data. Coelho et al. [174] proposed a dynamic anonymization method using separatrices to define k and dynamically group data for anonymization. Their approach applies the Elbow method, commonly used in clustering, to identify the optimal k for anonymizing numeric attributes.

Efforts have been made to understand mix-zones' behavior and the utility of anonymized data. In our previous work [261], we evidenced that mobility can impact location privacy approaches, like mix-zones, in the context of smart mobility. We showed the privacy and data utility side effects when mix-zone parameters are misconfigured in a multimodal environment. Additionally, we studied the mix-zones behavior that reflects anonymization quality for the posterior design and selection of robust LPPM with a notion of AQ – is a concept related to the protection, efficacy, and internal functioning of an LPPM, enabling an understanding of how privacy occurs and analyzing its performance over time [173, 20]. Regarding the trade-off between privacy and utility, we address how to identify smart city applications and services that can best leverage mobility data anonymized by mix-zones. For this, we proposed a methodology that evaluates the utility in many aspects with metrics related to privacy, mobility, and anonymized trajectories produced by mix-zones [287].

To the best of our knowledge, no previous mix-zones proposals focus on adjusting particularly of privacy levels k over time in online mode. Unlike previous works – that

address geometry, radius size, positioning, and behavior of the mix-zones – we advance the state-of-the-art proposing the k-DynMix: A dynamic mix-zones that tunes the privacy level k over time according to vehicles' traffic fluctuations to get higher anonymization.

8.3 Mix-zones and Problem Statement

This section defines the essential concepts of this work. First, we introduce the mix-zones scheme. Next, we describe the mix-zones problem related to privacy level.

Mix-zone M is a geographical area of k-anonymity that vehicles change their pseudonyms inside the mix-zone if there is a set of vehicles A, denoted as the anonymity set, present in it simultaneously and obey the condition $|A| \geq k$ called of Mix-zone Activation (MA), where k is anonymity mix-zone parameter [164, 172].

In mix-zones, k must be previously defined because the pseudonym change should occur within the mix-zone region [164]. A k previously configured, the pseudonym change process can be executed immediately when a vehicle enters M, which includes the use of the vehicular network infrastructure when vehicles are still in M. Furthermore, with a predefined k, it is possible to guarantee the minimum level of privacy required when using certain applications and services. For example, when mix-zones are used to protect the identity of users while using LBS. A user may determine that having an anonymity set $|A|$ of at least five individuals is enough to ensure their pseudonyms can not be linked between different application areas. This way, the k can be set up to $k = 5$.

Setting up mix-zone parameters such as privacy level k is a critical factor that affects the mix-zones performance, [261, 262]. A higher value of k corresponds to a greater level of privacy, whereas a lower k results in reduced privacy. For both cases, the anonymization occurs if it has MA in the mix-zone M. That is, NCM is greater than or equal to k, i.e., $MA \leftarrow NCM \geq k$.

If M does not achieve MA, the vehicles are not anonymized, which can degrade the anonymization rate and enable linking and inference attacks with a high success rate. The mix-zone efficacy degradation is more evident when using a static setup of k, particularly at a high privacy level [20]. The mix-zones can vary vehicle density over time and not achieve MA in low-traffic scenarios. Besides, in a high-traffic scenario, the privacy level cannot be efficient once the mix-zones are set with low k, particularly mix-zones based on pseudonym swap approaches in which the vehicles change the pseudonyms of each other. In ideal conditions, the best anonymization scenario for M is the – *Optimal Anonymization (OA)* – when k distribution is equal to NCM over time t, i.e., $NCM_t = k_t$. However, predicting k to achieve OA over time is complex and requires the prediction mechanism

ideal. Thus, we present a relaxed definition of OA called – *Higher Anonymization (HA)* – is k distribution follows the NCM having it as the lower bound of NCM over time, which satisfies the MA anonymizing as close to traffic flow. In other words, as near k approximates the NCM by an inferior limit, the higher the privacy level possible. Having trajectories with HA makes the re-identification problem more difficult because users' re-identification is an intricately combinatorial problem, where an attacker tries linking users' trajectories sliced by at least k vehicles. Therefore, the higher the k, the higher the combination to be made by the attacker to re-identify the user's identities. We summarize this discussion with higher anonymization definitions:

▶ **Def. 1** *Higher Anonymization (HA).* Let $M_{k(t)} \in \mathcal{M}$ be a mix-zone set up with privacy level k at time t. For $M_{k(t)}$ to get a higher anonymization, $(NCM_t \geq k_t \wedge k_t \approx NCM_t) \implies MA_t^H$, where k_t, NCM_t, and MA_t^H are privacy level, Number of Cars in Mix-zone (NCM) is a random variable, and Mix-zone Activation (MA) in a HA at time t, respectively.

8.4 Adversary Model

The adversary model for validating our proposal is based on Local Passive Adversary (LPAd) that applies the Semantic Linking Attack (SeLA), particularly the Multi-target tracking (MTT) for re-identifying users' trajectory, defined as follows.

Consider the vehicle set $\mathcal{V} = \{V_1, V_2, ..., V_i, ..., V_m\}$ where $1 \leq i \leq m$. Each V_i that makes a trip T_a composed of GPS points and $T_a \in \mathcal{T} = \{T_1, T_2, ..., T_a, ..., T_n\}$ where $1 \leq a \leq n$. $\mathcal{T}$ can across through multiple mix-zones $\mathcal{M} = \{M_1, M_2, ..., M_i, ..., M_n\}$ where $[1 \leq b \leq n]$ that anonymize them yielding the anonymized trajectory dataset D' and can eventually published as open data.

The Local Passive Adversary (LPAd) strategically deploys low-cost receivers over local regions of the road network near some mix-zones, for instance, a mix-zone M_i, and eavesdrops on exchanged messages, collecting sub-trajectories composed by GPS points of each vehicle V_j before and after in each M_i. This attack is sufficient for only two points before and after the mix-zone. It also assumes that the Local Passive Adversary (LPAd) collets D' in a repository. The GPS points and D' information represent the adversary's background knowledge about the users $\mathcal{B} = (B_1, \ldots, B_b, \ldots, B_m)$, where $[1 \leq b \leq m]$ and b represents the number of elements in $\mathcal{B}$ known by the adversary, enabling the adversary to execute a Tracking Attack $\mathcal{Z}$, and consequently perform the Multi-target tracking (MTT) for multi-users. In a Tracking Attack (TA), the adversary aims to determine the whole sequence (or a partial sub-sequence) of events in a user's trace. Given an

anonymized dataset $\mathcal{D}'$ composed of users and background B_u, a tracking attack is defined as $T_u \leftarrow \mathcal{Z}(\mathcal{D}', B_u)$, where T_u represents the re-constructed trajectory of user u.

8.5 k-DynMix Mechanism

This section details k-DynMix, an algorithm for predicting a privacy level over time for mix-zones.

Algorithm 7: K-DynMix

```
   Data: k_t; k_b; NCM; t; τ_lst
   Result: κ
 1 η, τ_lst ← timeoutArrCar(t, τ_lst, NCM)
 2 if (η) ∨ (NCM < k_t) then
 3 |   ρ ← k_t / 2
 4 |   κ ← k_b
 5 end
 6 else
 7 |   if k_t < ρ then
 8 |   |   κ ←min(2k_t, ρ)
 9 |   end
10 |   else
11 |   |   κ ← k_t + 1
12 |   end
13 |   if (κ > NCM) ∧ (κ > k_b) then
14 |   |   κ ← κ − c
15 |   end
16 end
```

Algorithm 8: TAC alg.

```
   Data: t; τ_lst; NCM
   Result: η; τ
 1 η ← False
 2 δ ← computeITM(t, NCM)
 3 if (δ > τ_lst) ∨ (|τ_lst − δ| > τ_thsh) then
 4 |   δ_est ← (1 − α)δ_est + αδ
 5 |   δ_dev ← (1 − β)δ_dev + β|δ − δ_est|
 6 |   τ ← δ_est + uδ_dev
 7 |   if δ > τ_lst then
 8 |   |   η ← True
 9 |   end
10 end
```

The main idea behind our approach is to control the level of privacy over time-based on events that occur in mix-zones. From these events, the strategy is to tune k as a lower bound but close to NCM as fast as possible to attend the higher anonymization (Def.1). k-DynMix must have strategies for decrement and increasing exponentially and linearly the value of k according to the events that occur in M. We identified events related to vehicular traffic flow that occurs in the M and actions that affect the privacy level k. The mix-zone events are:

- Mix-zone Activation (MA): signals that M is active, that is, $NCM \geq k$, indicating the presence of vehicle traffic flow in M, thus it is possible to anonymize vehicles.
- Mix-zone Deactivation (MD): a complement of MA, signaling that M is deactivated, that is, $NCM < k$, thus anonymizing vehicles is not possible. Mix-zone Deactivation (MD) may indicate a low flow of vehicles or that k is misconfigured. This way, it can occur at any time of the day.

- Timeout of Arriving Cars in Mix-zones (TAC): an event when vehicles do not arrive at M at the time limit τ defined. It can indicate an absence of cars in M to achieve the MA, consequently not yielding vehicle anonymization. TAC events can occur in periods of low traffic with low car arrival rates, such as dawn. TAC must follow the changes in vehicle traffic over time in M [20] and consider more recent vehicle entry rates that reflect the current vehicle flow in M.

Several empirical experiments have been carried out regarding the growth of k. We observed that the time to meet Def. 1 is greater in linear growth than in exponential growth. However, using purely exponential growth of k, it soon reaches the NCM limit and produces many MD events. k-DynMix's prediction strategy is that in each MD or TAC event in M the privacy value decreases significantly to initial privacy value k_b to reach MA. Setting the value of k to k_b is an approach to meet situations of low vehicle traffic in k_b and also to achieve MA. Then, it defines a privacy threshold ρ as half of k. The ρ orchestrates the growth of k. When MA occurs, k grows fast, exponentially, until reaching ρ. When k is great or equal to ρ, k increases linearly – to avoid loss of privacy with MD – until achieving NCM. After the increment actions, the k can be higher than NCM; this way, k is adjusted as a decrement to tend to NCM, so we have a fine adjustment. This approach follows the same principle as the TCP congestion control. Next, we further detail the k-DynMix algorithm.

8.5.1 Definition of k-DynMix

Let M be a mix-zone deployed in road intersection inside on RSU enabled with a vehicular sensor device that periodically senses the Number of Cars in Mix-zone (NCM) on M region. At each iteration, it invokes k-DynMix (Algo. 7). Its inputs are privacy level k_t and NCM at the time t; k_b is the initial privacy level. τ_{lst} represents the value of the last timeout. The output is the next privacy level predicted, denoted by κ. We first compute (Line 1) the timeout of arriving cars estimation function *timeoutArrCar* (Algo. 8) to verify if timeout event η happened and compute the timeout from the last timeout τ_{lst}. Lines 2-5 compute the κ decrement in timeout and MD events. These events means low traffic in M and κ must be set to k_b, ρ set to half k_t. Lines 6-12 are about the MA and compute κ increment being linear and exponential. If k_t is less than ρ indicate that k_t is far from NCM value than that κ can increases exponentially until achieve ρ (lines 7-9). In contrast, when k_t is great or equal to ρ means that k_t is near NCM, the κ must increase linearly, i.e., increases in one unit until it arrives at the NCM. That is the best scenario for anonymization, with the maximum privacy level equal to NCM. In the same iteration of the algorithm, κ can be upper bound but near than NCM, yielding MA is False events. Then, adjusts κ decreasing it minus a constant c, where $\kappa \geq c$ (lines 13-15).

The TAC algorithm (Algo. 8) is inspired by the TCP Retransmission Timer mechanism, which computes and manages the retransmission timer (timeouts) on the host sender [288]. Basically, TAC algorithm favors recent events of ITM over old to compute vehicles timeout in M. ITM is a mix-zone quality metric that measures the time interval in seconds between vehicles entering the mix-zone [173, 20].

The input of the TAC algorithm is τ_{lst} is last TAC and NCM at current time t. The algorithm's output returns two values: η represents if had timeout event; τ is the TAC for the next iteration. η is initialized with False (line 1), indicating that it did get a timeout. Line 2 computes the ITM quality metric, represented by δ. In each vehicle entry input event at M, the ITM is calculated. Lines 3-10 estimate the τ and if occurred a timeout event η in M. The timeout will be estimated in two cases (line 3). In the first case, if ITM is higher than the last timeout estimated τ_{lst}. The second case, for prevention of timeout too high, controlled for if the difference between τ_{lst} and ITM is higher than timeout threshold τ_{thsh}. Line 4 computes the estimated ITM (δ_{est})) that is a weighted combination between previous values of ITMs (δ_{est})) and δ, α is the weight to favor one of both variables. The (δ_{est})) favors the recent ITMs values. Line 5 computes the deviation between ITMs (δ_{est})) and δ to smooth the timeout calculate. This equation also uses the WEMA for weighting the recent δ_{dev}. Finally, the timeout on line 6 with δ_{est} plus a margin of error that is high when existing high variation of δ_{dev} and low in otherwise, with a margin factor constant u, where $u > 0$. Lines 7-9 return if had timeout event η.

▶ *Complexity Time Analysis.* Considering that M has a loop of size n, which senses NCM and then run k-DynMix has a complexity of $\Theta(n)$. The Algorithm 7 have $\Theta(1) + 2\Theta(1)$, being the Algorithm 8 has $\Theta(1)$. Hence, the complete time complexity is: $T(k\text{-}DynMix) = \Theta(n).(\Theta(1) + 2\Theta(1)) = \Theta(n)$

8.6 Experimental Evaluation

For validation of predicting a fair k, we compared the k-DynMix with two widely used approaches of moving average. The is *Simple Moving Average (SMA)* computes the average of the preceding n data points where each data point holds equal importance. The *Weighted Exponential Moving Average (WEMA)* allows for customized weighting schemes through specific weighting functions, such as attributing high weight to recent points. For details about these approaches, see the Appendix A.

8.6.1 Mix-zones Performance Metrics and Privacy Attack

This analysis aims to evaluate the k-DynMix in terms of Accuracy, Coverage, and Quality Mix-zones metrics. In the first analysis, we investigate the ability of an engine to predict values of k to achieve MA. For this use the **Accuracy in k's Predictions ($\mathbf{ACC}_{pred}$)**. It is denoted by $ACC_{pred} = |MA|/(|MA| + |MD|)$, where $|MA|$ is the number of attempts that an engine estimated k and got MA by the total of attempts, represented by the sum of $|MA|$ and $|MD|$.

In the second analysis, we perform mix-zones coverage metrics extracted from static mix-zones and k-DynMix. Let mix-zone M be, the mix-zone coverage metrics are:

- **Anonymization Rate (AR)**: AR_M is the number of trajectories that passed and were anonymized by$M$;
- **Non-Anonymization Rate (NAR)**: NAR_M is the number of trajectories that passed and were not anonymized by M;
- **Mix-zone Efficacy (ME)**: $textitME_M$ refers to the ratio between a number of users anonymized by M, denoted by AR_M, and the population P_M that crossed it, i.e., $ME_M = |AR_M|/|P_M|$.

We analyzed the k-DynMix in anonymization quality terms in the third analysis. Mix-zones with anonymization quality yield high efficacy and considerable privacy levels, anonymizing the mobility data at the moment they are activated, which can thus lead to energy savings. The mix-zones quality metrics are:

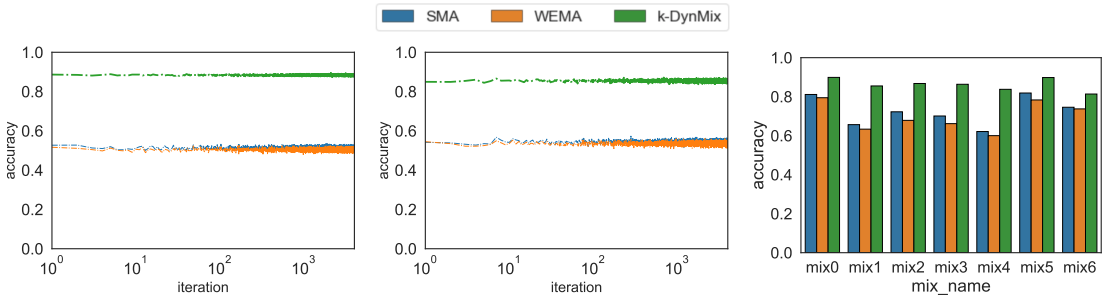
- **Number of Cars in Mix-zone (NCM)**: Number of cars crossing the mix-zone over time;
- **Interval of Arrival Time between Cars on Mix-zones (ITM)**: Measures the time interval in seconds between vehicles entering the mix-zone, enabling the measuring of traffic volume inside the mix-zone;
- **Interval of Departure Time between Cars on Mix-zones (IDM)**: Measures the time interval, in seconds, between vehicles going out of the mix-zone, enabling the measurement of traffic volume inside the mix-zone;
- **Number of Trips Completed within the Mix-zone (NTC)**: Measures the number of vehicles that terminate their trips within the mix-zones;
- **Activation Time of the Mix-zone (ATM)**: Time period in which a mix-zone is activated for anonymization when the number of cars inside it is greater or equal than the k parameter.High ATM denotes mix-zones with anonymization for a long period of time.

Based on quality metrics, we compute the AQ function – describes the considered aspects of privacy and behavior of the mix-zone over time [20]. The AQ of a day period

s is denoted by Boltzmann distribution, represented by $AQ_s = H_s(u')\phi_s\epsilon^{-(\mathcal{I}_s+\mu)/\lambda}$. The $H_s(u')$, ϕ_s, $\mathcal{I}_s$, and μ represent privacy weight, the total number of ATMs, throughput, and NTC in mix-zone M, in the day period s, respectively. The λ is the Boltzmann constant. The $\epsilon^{-(\mathcal{I}_s+\mu)/\lambda}$ is used to penalize the result when many vehicles enter the mix-zones and do not leave it. Consequently, these vehicles are not anonymized. Specifically for the relation $\mathcal{I}_s + \mu > 1$ then $\epsilon^{-(\mathcal{I}_s+\mu)/\lambda}$ tends to zero. Otherwise, if $\mathcal{I}_s + \mu \leq 1$ then $\epsilon^{-(\mathcal{I}_s+\mu)/\lambda}$ tends to one. $AQ \in [0, 1]$ where 1 is the maximum value for the anonymization quality.

Finally, we analyzed the privacy potentialities of k-DynMix and static mix-zones setups. Particularly, we consider the adversary model definition in Section 8.4 in which we applied a re-identification attack on protected mobility data by k-DynMix and mix-zones setups $k = 2$, $k = 4$, and $k = 6$. For re-identification of trajectories, we used the attack proposed in [56] that uses only two location points of trajectories as knowledge. The attack hypothesis is that most vehicles, especially taxis, choose minimal paths to complete their routes. Thus, it is possible to re-identify anonymous trajectories by comparing candidates' trajectories with a minimal path built between two points. The attack efficacy denoted by **Trajectory Matching Accuracy (TMA)** is a privacy metric to measure the effectiveness of re-identification attacks [54], which is defined as $TMA = N_{\text{reid}}/|AR_M|$, where $N_{\text{reid}} \in [0, |AR_{MZ_x}|]$ and $|AR_M|$ represent the total of re-identified trajectories and the total of anonymized trajectories that have passed in the mix-zone M, respectively. TMA $\in [0, 1]$ where TMA equals 1 is the max attack success. In contrast, TMA equals 0, representing no success, and the trajectory is protected against the attack.

8.7 Results and Discussion



(a) Gaussian distr. D_c (b) Discrete Unif. distr. D_d (c) Cabs distr. D_b

Figure 8.1: Accuracy (ACC_{pred}) of k-prediction techniques using of NCM of Gaussian, Discrete Uniform, and actual cabs distributions. In Figure 8.1c the NCM was extracted from mix-zones positioned with the FPMT algorithm from the Cabspotting dataset.

This section presents the results and discussion of k-DynMix performance on predicting privacy against two prediction mechanisms using NCM. Also, we evaluated mix-zone coverage metrics and AQ extracted from datasets protected with k-DynMix and statics setups.

8.7.1 Experiments Setup

In this study, two datasets were analyzed, one containing real data and another being a synthetic dataset. The former is the *Cabspotting* dataset, which comprises mobility data from approximately 500 distinct taxicabs in San Francisco, USA, collected over 25 days [126]. This dataset was collected in 2008 and included information on the taxis' locations, sampled periodically by a built-in GPS sensor. The dataset contains around 440,000 trips, averaging about 17,600 daily. This extensive mobility data covers over 70% of the city's road segments, with approximately 400,000 contacts between vehicles per day, highlighting the dataset's significant potential for our analysis. Two samples of high traffic were extracted from the *Cabspotting* dataset for analysis. The first sampling D_a was on Sunday, May 18, 2008, with 366,951 registers, 422 users, and 1770 trips, for mix-zone metrics and anonymization quality analysis. The second sampling D_b was on Monday, May 19, 2008 corresponding to 417,781 registers, 454 users, and 2036 trips was used to select k-prediction approaches based on NCM of mix-zones positioned in the city.

Also, to validate k-prediction approaches for mix-zones, we used two synthetic temporal series datasets that simulate the traffic flow in a mix-zone. Specifically, we generate random NCM with Gaussian (D_c) and Discrete Uniform (D_d) distributions of size to 2000 and time variation between 1s to 30 minutes. Regarding the Gaussian distribution (D_c), we used the mean and standard deviation of a NCM distribution equal to 10 and 2.5, respectively. The NCM in the lower and upper bounds used for Discrete Uniform distribution were 0 and 21.

The mix-zones were chosen with a mix-zones deployment algorithm known as FPMT [56]. This algorithm generated a list of mix-zone candidates, with each list being sorted by vehicle frequency at intersections and centroids in descending order. From each candidate list, only seven were chosen based on two criteria. Firstly, the mix-zones had to anonymize trajectories in at least two of the three k setups. Secondly, the selected mix-zones could not overlap with each other. Due to these conditions and the San Francisco region's limited size, seven mix-zones were determined as the privacy budget.

8.7.2 Privacy Level Prediction Mechanisms Analysis

The first step in designing dynamic mix-zones with the tuning of k over time is to identify an efficient prediction mechanism for estimating the fair value of k. Particularly, we analyzed the ACC_{pred} of k-DynMix and prediction approaches defined in Section 8.6.1 using two synthetic datasets of NCM with Gaussian (D_c) and Discrete Uniform (D_d) distributions. We used Bootstrap, which randomly generated both datasets by resampling 4000 iterations. For each iteration, we calculated ACC_{pred}. Discrete Uniform distribution generated an average of 9.2% of NCM in the samples, which NCM is lower than minimal k, indicating a more realistic than Gaussian distribution that got near zero. Figure 8.1a shows the ACC for the NCM Gaussian Distribution, where k-DynMix had better performance than other prediction mechanisms, with ACC_{pred} average ($avg(ACC_{pred})$) equal to 88.3%. The SMA and EWMA had the worst results, with an $avg(ACC_{pred})$ of 51.4% and 50.4%, respectively. The same behavior is observed for Discrete Uniform distribution (see Figure 8.1b), in which k-DynMix had the best performance with $avg(ACC_{pred})$ equal to 85.5%, followed by SMA, and EWMA with $avg(ACC_{pred})$ equals 54.3%, and 53.5%, respectively.

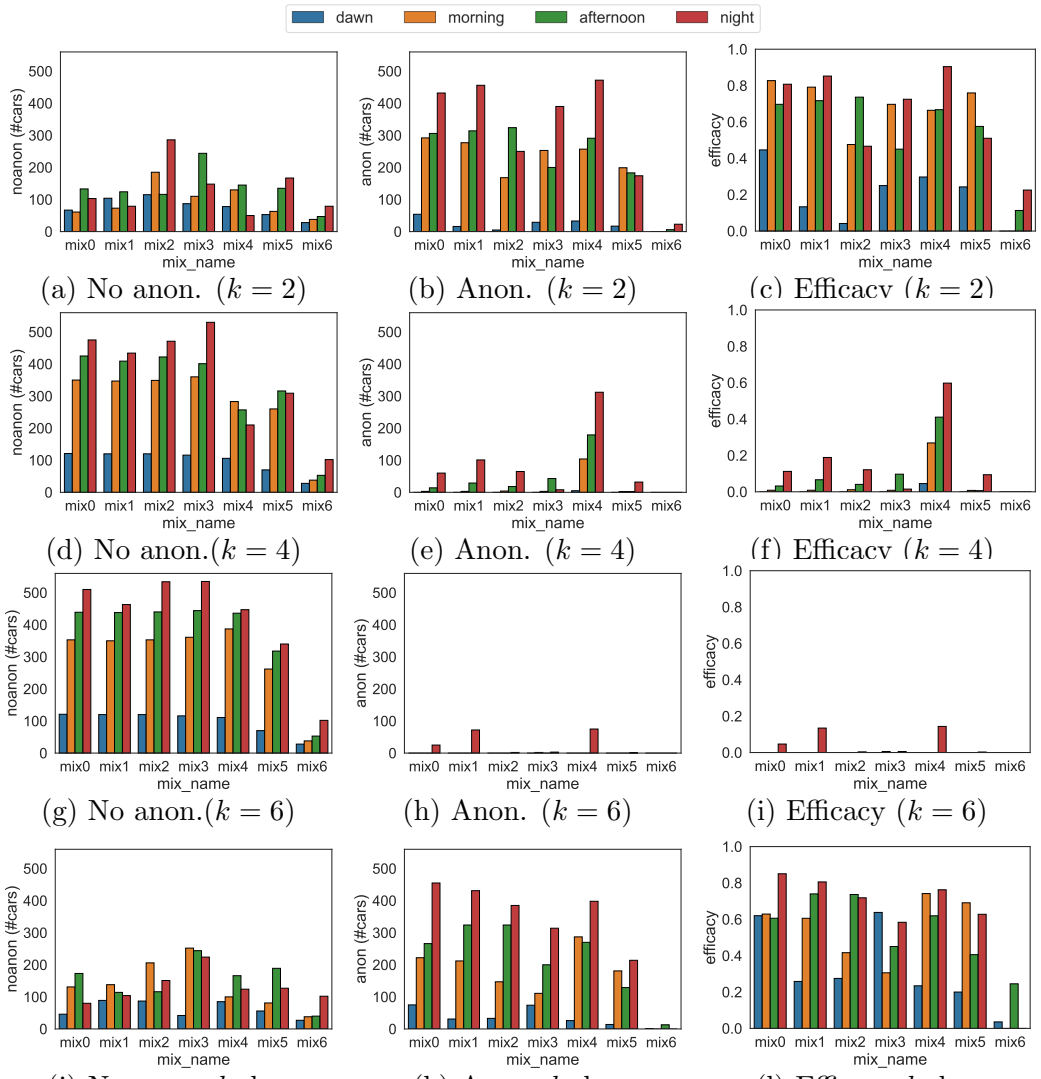
Regarding NCM analysis with the actual traffic flow of cabs (D_b), we analyzed the NCM of seven mix-zones previously deployed by a positioning algorithm [56]. Table 8.1 details the location of these mix-zones. This analysis presented a different behavior than synthetic datasets (see Figure 8.1c). For D_b had around 44.5% of NCM in the samples, which NCM is lower than minimal k, indicating a more realistic than synthetic datasets D_c and D_d. The mobile average approaches performed better than the analysis with synthetic datasets but were lower than k-DynMix, who performed best over prediction approaches in all mix-zones. Particularly, SMA and EWMA had $avg(ACC_{pred})$ equal to 72.5% and 69.8%, but k-DynMix got 86.2%, highlighting a peak on mix0 and mix5 with ACC_{pred} up to 89.8%. The results suggest that k-DynMix performed better than other k-prediction approaches.

Table 8.1: Mix-zones deployed with FPMT, their locations and coef. variation.

Name	Latitude	Longitude	Coef. Var. (%)	Location	Places Covered by Mix-zone
mix0	37.714801	-122.397982	74.13	101 express way, Visitacion Valley.	cafes, subways station, restaurants.
mix1	37.724830	-122.400157	74.43	101 express way, Bay View.	supermarkets, restaurants, clinics.
mix2	37.735133	-122.404532	82.49	101 express way, Bernal Heights.	road interchange, gas station, supermarket.
mix3	37.676005	-122.391491	75.95	101 express way, Firth Park.	tourist area, hotels, docks.
mix4	37.615315	-122.393566	65.68	entrance road to the airport.	cabs park, bus stop, restaurant, garages
mix5	37.774378	-122.401540	72.10	downtown, near 101 express way.	exit to Oakland city, shopping, church.
mix6	37.768990	-122.419450	**93.11**	downtown, Mission District.	tourist area, museums, subways station.

Table 8.2: General (Non) Anon. and Efficacy: static mix-zones and k-DynMix.

	No Anon	Anon	Total	Efficacy
$k = 2$	3048	5421	8469	0.640
$k = 4$	7482	987	8469	0.117
$k = 6$	8289	180	8469	0.021
K-DynMix	3332	5137	8469	**0.607**



(a) No anon. ($k = 2$) (b) Anon. ($k = 2$) (c) Efficacv ($k = 2$)

(d) No anon.($k = 4$) (e) Anon. ($k = 4$) (f) Efficacv ($k = 4$)

(g) No anon.($k = 6$) (h) Anon. ($k = 6$) (i) Efficacy ($k = 6$)

(j) No anon. k dyn (k) Anon. k dyn (l) Efficacy k dyn

Figure 8.2: (Non) Anonymization and Efficacy of the mix-zones for $k = [2, 4, 6]$ and k-DynMix positioned with FPMT algorithm.

8.7.3 Mix-zone Characterization with Coverage Metrics

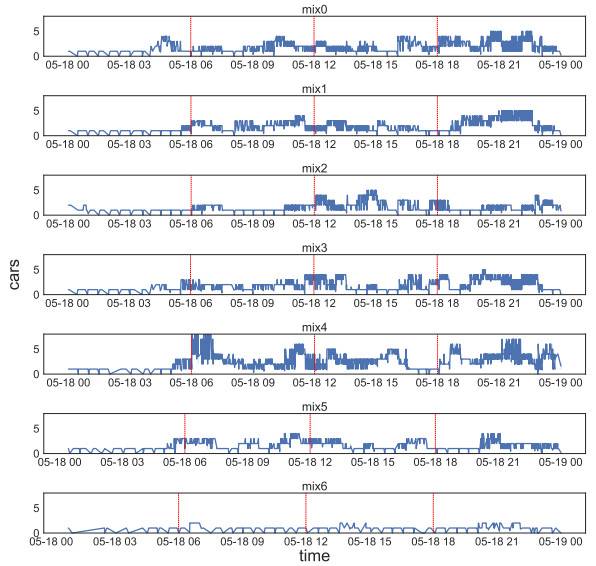
After mix-zones positioned with the positioning algorithm defined in [56], we anonymized the dataset D_a with mix-zones static setups $k = 2$, 4, 6 and k-DynMix with a radius equal to 500 meters, as proposed in [20]. Table 8.1 shows the location of mix-zones deployed and the label of geolocations and places that these mix-zones cover.

The Coefficient of variation (CV)[1] of vehicles for each mix-zones (i.e., NCM). We can see that mix6, mix2, and mix1 had a higher variation in the number of vehicles about the central average than the others, highlighting mix6 with the CV of 93.11. Thus, these mix-zones had a larger oscillation of vehicle density over time, which can negatively impact performance metrics, such as AR and ME.

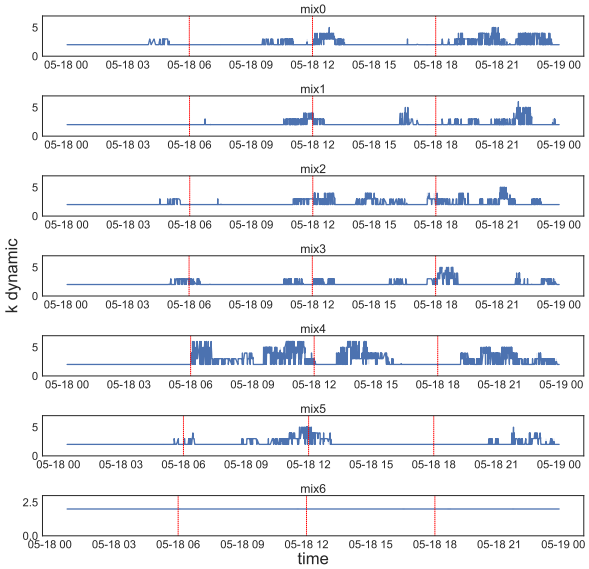
Table 8.2 shows the synthesis of a total of trips of Anonymized, No Anonymized, and General Efficacy of all mix-zones of each setup. For static mix-zones, k is inversely proportional to the anonymization rate and general efficacy and directly proportional to the no anonymization rate. That is, as k increases, the anonymization rate and general efficacy decrease, and the protected dataset becomes vulnerable to tracking attacks. The best efficacy was mix-zones setup $k = 2$ with 64% but had low privacy. k-DynMix had an anonymization rate equal to 5421 trips and an efficacy of 60.7% near the setup $k = 2$ but a higher privacy level than $k = 2$.

The mix-zones statics and k-DynMix behaviors can also be seen in Figure 8.2 that details about NAR, AR, and ME of the mix-zones protected with static k setup and k-DynMix. To understand the subtilizes of the behavior of mix-zones setups, we define a time window W of four periods: dawn with a period between 00:00:00 and 05:59:59, morning with a period of 06:00:00 to 11:59:59, afternoon from 12:00:00 to 17:59:59, and night from 18:00:00 to 23:59:59, as proposed in [20]. Overall, at night and dawn were when mix-zones got the high and low AR, respectively. Particularly, at night, mix4 got the highest AR and ME than other mix-zones for all setups, indicating a great traffic flow (a high NCM) deployed at San Francisco Airport. For setups $k = 4$ and $k = 6$ got a low NCM insufficient to activate the mix-zones, causing a high NAR and ME to zero. In a comparison of static mix-zones setups vs. k-DynMix, the ME of k-DynMix performed near to the $k = 2$ setup and better than $k = 4$ and $k = 6$ setups, but with the possibility to have more privacy level than $k = 2$. For instance, at night for mix0, mix1, and mix4, the ME for k-DynMix were respectively 0.85, 0.80, and 0.76 against 0.11, 0.18, and 0.59 for $k = 4$, and $k = 6$ the ME were 0.04, 0.13, 0.14. Also, k-DynMix improves the mix-zones performance that had low traffic, as shown in the $k = 6$ setup with mix2 and mix3 in which the AR were 2 and 5 respectively, and with k-DynMix had an improvement to 998 and 889.

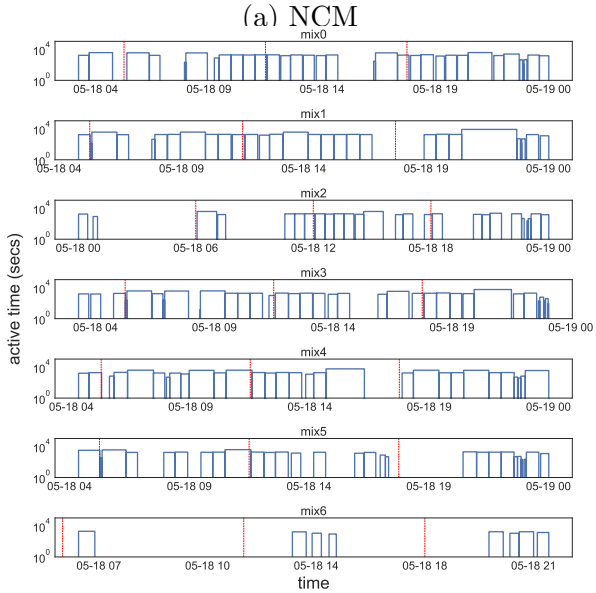
[1]The coefficient of variation (CV) measures the dispersion of a probability distribution used to identify the percentage of variation about the central mean of a sample.



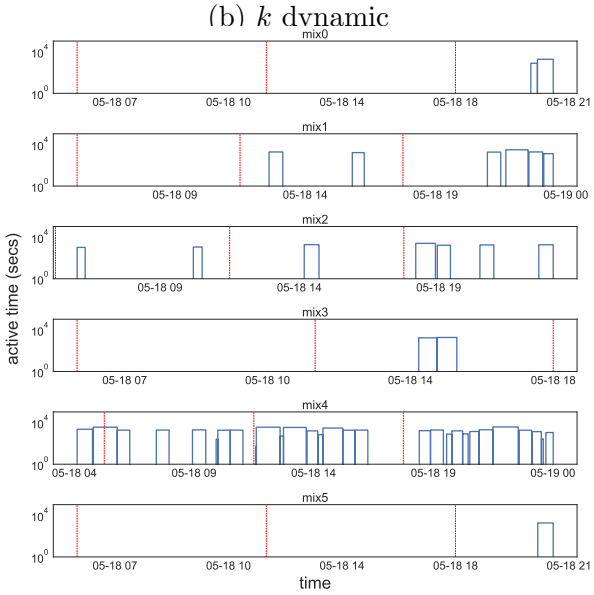
(a) NCM



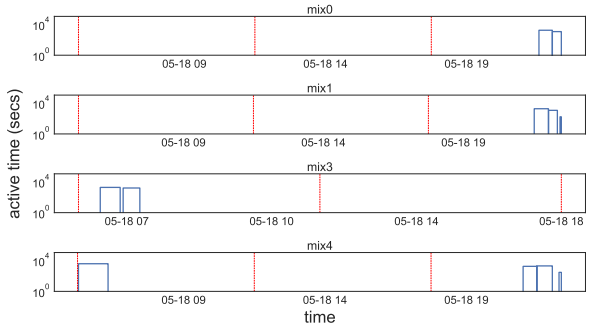
(b) k dynamic



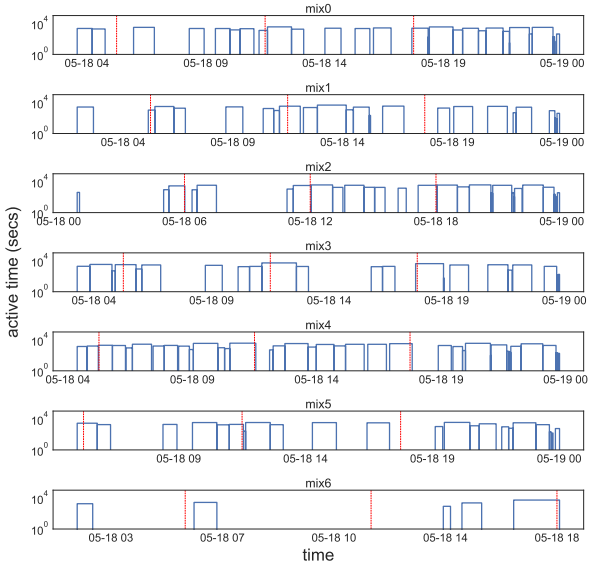
(c) ATM for $k = 2$



(d) ATM for $k = 4$



(e) ATM for $k = 6$



(f) ATM for k dynamic

Figure 8.3: NCM, ATM, and k analysis for static mix-zones and k dynamic.

8.7.4 Mix-zones with k Static vs. k-DynMix in AQ metrics terms

8.7.4.1 Number of Cars in Mix-zone (NCM) and k Dynamic Over Time

We calculated the NCM metric for each mix-zone to assess the AR trends over time (see Figure 8.3a). In the dawn, few taxis passed through the mix-zones, resulting in lower AR values (Figures. 8.2b and 8.2k). The volume of vehicles and AR increased gradually during the morning, afternoon, and evening. This trend was highlighted for $k = 4$ as shown in Figures. 8.2e and 8.2k. At night, mix0, mix1, and mix2 experienced the highest levels of simultaneous vehicle traffic, with NCM values peaking at five, leading to significant spikes in anonymization. But the mix4 got peaks of NCM in the morning and night with NCM equal to 8 and 7, respectively (see Figure 8.3a). Notably, mix4, situated in the airport region, exhibited a distinct NCM pattern compared to the other mix-zones. The mix6, located in the central area of San Francisco, had lower traffic volume, registering an NCM of 2 vehicles at noon, coinciding with its peak AR period.

Figure 8.3b details the k over time produced by k-DynMix for the seven mix-zones. All mix-zones got a peak of privacy level upper to 2, in which mix0 to mix5 got a k of 5, 6, 5, 5, 6, and 5, respectively. Except mix6 that got ($k = 2$) for all periods because of a low NCM. We also note that k distribution follows the variation of NCM over time for each mix-zone. For instance, minimal k occurs in dawn periods, and k peaks occur during morning, afternoon, and night (Figure 8.3b). This is the case of mix4, which had more NCM in the morning, the afternoon got peaks of 6, and at night, the prevalence of the k equal to 5 (Figure 8.3a). Mix0, mix1, and mix2 had k peak at night with k of 5, 6, and 5, respectively. This analysis suggests that k dynamic had better performance than static k in terms of privacy level.

8.7.4.2 Activation Time of the Mix-zone (ATM) and k-Dynamic

A high k means a high privacy level, but anonymization with a high k occurs if and only if $NCM \geq k$, which implies that mix-zones are active (MA). For AR with a high k, the longer ATMs are better than shorter and adjacent ATMs, which in turn is better than short, scattered ATMs [20]. We can note peaks of AR in periods of longer ATMs, such as the $k = 2$ at night for mix1, mix3, and mix4 had longer ATMs (see Figure 8.3c) and consequently high AR (see Figure 8.2b). However, the shorter and adjacent ATMs,

as generated by k-DynMix in Figure 8.3f, can have AR with better privacy level than static approaches. For instance, mix4 for $k = 2$, which has a probability of re-identifying a trajectory of 1/2 in the morning, afternoon, and night, had AR of with AR 257, 291, and 472, respectively. Nevertheless, for k-DynMix, in the same period had trips that were anonymized with peaks of k up to 6, i.e., the probability of re-identification is 1/6, got AR performance close to $k = 2$ equal to 287, 270, and 398. The same behavior had mix0, mix1, and mix2, with shorter and adjacent ATMs in the afternoon and night. k-DynMix also performed better than the $k = 4$ and $k = 6$ settings because some mix-zones did not achieve anonymization for these settings, such as mix6 for $k = 4$ and mix2, mix5, and mix6 for $k = 6$ (see Figures. 8.3d and 8.3e).

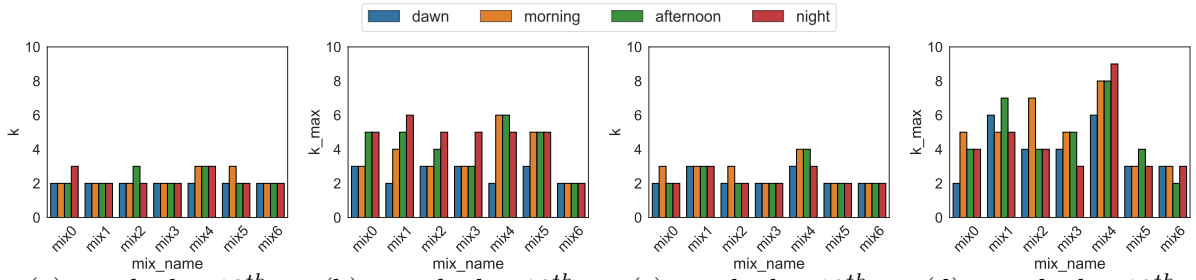
8.7.4.3 Average of k and k Maximum per Period

This analysis investigated the ability of the k-DynMix to estimate k. We analyze the average and maximum value of k per period with two scenarios. One with lower vehicle flow corresponds to samples of D_a trajectories from Sunday, 05/18/2008. The other with higher vehicle flow corresponds to samples of D_b trajectories from Monday, 05/19/2008 (see subsection 8.7.1 for details about the samples D_a and D_b). For the D_a, k-DynMix produced an average k equal to 3 for at least one period for the mix-zones mix0, mix2, mix4, and mix5 (see Figure 8.4a). We highlighted mix4, got an average k of 3 for three periods. For D_b sampling, the average k was equal to 3 for mix0, mix1, mix2, and mix4. We highlighted mix4, which obtained an average k of 4 for the morning and afternoon periods (see Figure 8.4c).

We also observed the variation in anonymization regarding the maximum k estimated per period value. For D_a, of the seven mix-zones, six of them reached peaks equal to 5 or greater than 5 in the afternoon and evening periods (Figure 8.4b), being mix0, mix1, mix2 mix3, mix5, and mix4; highlighting mix4 which got a k equal to 6 during the afternoon and evening. About D_b, all mix-zones had k peaks equal to or greater than three and five mix-zones equal to or greater than 4 (Figure 8.4d). Once more the mix4 had a maximum k equal to or greater than 8 in the morning, afternoon, and night. mix4 is located in the San Francisco International Airport region, where many taxis remain to carry out their trips.

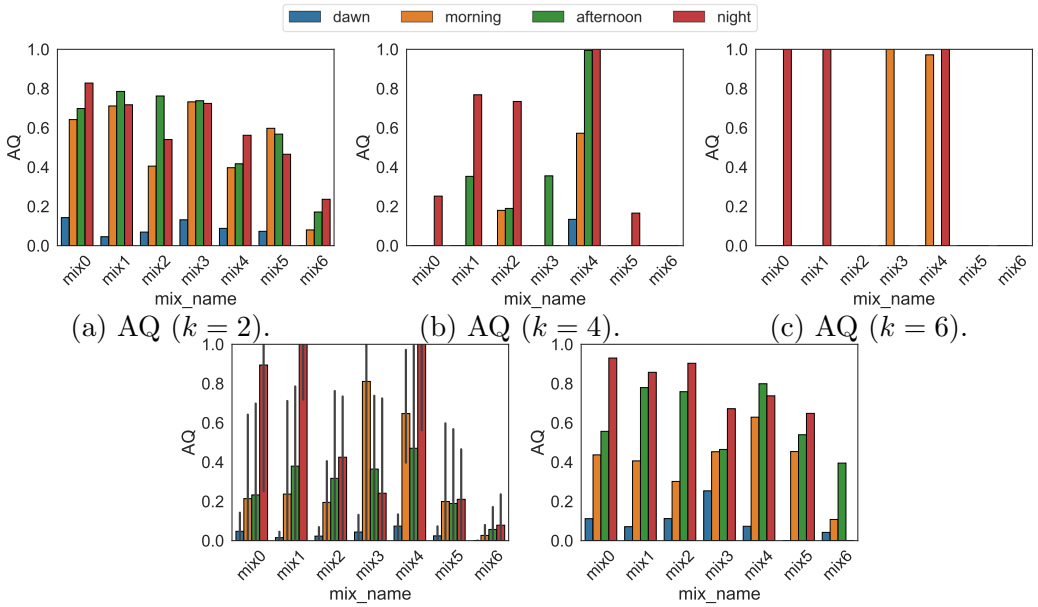
8.7.4.4 Anonymization Quality (AQ) and k-Dynimic

In the mix-zones with a static setup, the highest AQ of mix-zones for $k = 2$ were the mix0, mix1, mix3, mainly in the afternoon and evening (see Figure 8.5a). They had



(a) avg k, day 18^{th} (b) max k, day 18^{th} (c) avg k, day 19^{th} (d) max k, day 19^{th}

Figure 8.4: Average of k and maximum k per period calculated by k-DynMix. Sampling D_a: Sunday, May 18, 2008 - Figures. 8.4a and 8.4b. Sampling D_b: Monday, May 19, 2008 - Figures. 8.4c and 8.4d.



(a) AQ ($k = 2$). (b) AQ ($k = 4$). (c) AQ ($k = 6$).

(d) $AVG(AQ_{W_i}(k))$ (e) AQ k dyn.

Figure 8.5: Anonymization Quality (AQ) of mix-zones protected with $k = [2, 4, 6]$ (Figures. 8.5a, 8.5b, 8.5c), average AQ per period of the privacy levels (Figure 8.5d), and k-DynMix (Figure 8.5e).

high AR, ME, and low NAR during these periods. This behavior is reflected in the same NCM and ATM metrics periods, where mix0 and mix1 are highlighted more than others, particularly mix1 got NCM and ATM up to 5 and 1832 secs, respectively. With the increase of k, the AR and ME decrease, and the mix4 highlights above the mix-zones due to its high traffic flow, resulting in an AQ close to the max value. $k = 6$ setup is an example where the AQ is 0.97 and 1 in the morning and night. But, the AQ equals zero in all periods for mix2, mix5, and mix6 (Figure 8.5c).

Unlike static k's setups, the k-DynMix got AQ for all mix zones. Particularly in the morning, afternoon, and night on mix4, k-DynMix had AQ higher than $k = 2$, in which k-DynMix got AQ of 0.62, 0.80, 0.73, for these periods, respectively, and $k = 2$ got AQ 0.39, 0.41, 0.56. The performance of k-DynMix was caused by the k's variations over time, achieving an average of k equal to three in these periods (see Figure 8.4a). Also, about the $k = 2$, the AQ of k-DynMix at night for mix0, mix1, mix2, and mix4

was superior to $k = 2$ where k-DynMix, respectively, got AQ equal to 0.93, 0.85, 0.90, and 0.73 against $k = 2$ in which AQ were 0.82, 0.71, 0.54, and 0.56 (see Figures. 8.5e and 8.5a).

Regarding the comparison between the average AQ per period W_i of the privacy levels $k = \{2, 4, 6\}$, $AVG(AQ_{W_i}(k))$, and AQ estimated by the k-DynMix also performed well (see Figure 8.5d). For instance, the $AVG(AQ_{W_i}(k))$ of mix0 at dawn, morning, afternoon, and night were 0.048, 0.21, 0.23, and 0.89, respectively, and the AQ of k-DynMix were 0.11, 0.43, 0.55, and 0.93. Other mix-zones presented the same behavior, e.g., mix0, including mix-zones with low traffic, such as mix6, in which AQ of k-DynMix outperformed $AVG(AQ_{W_i}(k))$.

8.7.5 Mix-zones under Tracking Attack: k Static vs. k-DynMix

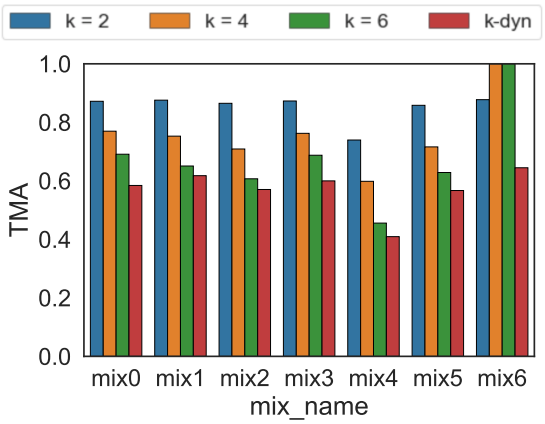


Figure 8.6: Trajectory Matching Accuracy (TMA) of mix-zones with static setups $k = [2, 4, 6]$ and k-DynMix.

Figure 8.6 represents the results of the tracking attack on dataset D_a – 18th day of the Cabspotting dataset – protected with the mix-zones set up with static setups and k-DynMix. The k-DynMix outperformed all setups in protection terms, with a TMA average of 57.03%, while the $k = 2$, $k = 4$, $k = 6$ setups had 85.15%, 75.83%, and 67.43%, respectively.

In privacy terms, the k-DynMix performed better than the static mix-zones setups in mix-zones with high vehicle traffic, like the mix4 positioned at the airport. Particularly, the k-DynMix got a TMA of 40.92% against the 73.95%, 59.83%, and 45.56% for $k = 2$, $k = 4$, and $k = 6$. This performance is due to the peaks of k in various day periods that resulted in the anonymization of trajectories with high k as shown in Figures 8.4b and 8.2k.

Regarding the static setups, the TMA was low in mix-zones with a high setup, e.g., mix0, mix4, and mix5 with $k = 6$. Except for mix6, which setups $k = 4$ and $k = 6$ got TMA equal 100%. The motive is that mix6 had low traffic and did not achieve the minimum k value to activate it (MA), and consequently, the vehicles were not anonymized (see anon rate in Figure 8.6). However, in mix6, the k-DynMix also outperformed all mix-zones with the lower TMA of 64.46%.

Although k-DynMix did not protect the dataset completely, it mitigated the level of re-identification by up to 33%, 18.9%, and 4.6% compared to the $k = 2$, $k = 4$, and $k = 6$ setups, respectively. Also, the k-DynMix outperformed the static mix-zones setups in privacy terms in different traffic scenarios, i.e., in mix-zones with low and high traffic, even in datasets with no considerable traffic, such as dataset D_a.

8.8 Concluding Remarks

In this chapter, we proposed the k-Dynamic Mix-zone (k-DynMix), a dynamic mix-zones that tunes the privacy level k over time in an online model, with linear complexity, according to events like vehicles' traffic fluctuations to obtain optimal anonymization. We designed k-DynMix inspired by Anonymization Quality metrics and TCP congestion control mechanisms. We conducted experiments, analyzing real and synthetic datasets comparing k-DynMix with two prediction mechanisms to estimate privacy levels over time and with classic mix-zones regarding mix-zone coverage, privacy metrics, and AQ. The results showed that k-DynMix outperformed the prediction mechanisms in predicting privacy level. Furthermore, it got efficacy, anonymization rate, and AQ similar to the highest result of classical mix-zones (setup $k = 2$). It maximized the privacy level to the best possible, with the lowest re-identification rate outperforming the classical mix-zones.Further, it improved the performance of mix-zones with low traffic. Unlike classical mix-zones, the k-DynMix got AQ for all mix zones. To the best of our knowledge, this is the first approach of dynamic privacy level over time in mix-zones that considers events vehicle traffic fluctuations to maximize the mix-zones privacy level, efficacy, and anonymization quality.

The next chapter presents the concluding remarks of this thesis and future directions concerning smart mobility open data.

Chapter 9

Conclusion and Future Work

This chapter presents the final remarks of this thesis as well as the future work to guide the investigation of further research. Therefore, Section 9.1 presents our concluding remarks, summarizing the contributions addressed in this thesis. Section 9.2 presents future research directions based on the challenges highlighted in this thesis.

9.1 Concluding Remarks

This thesis studied location privacy based on anonymization applied to smart mobility open data in smart cities, considering the particular characteristics of this environment. The open data, in its essence, presents requirements that are often contradictory, such as privacy, utility, and data quality, which require discussion and exploration of the location privacy mechanisms in the literature, even more so for dynamic environments such as smart mobility.

Guiding this study, we have introduced a novel anonymization-based framework for smart mobility open data. This framework, which takes into account the privacy, utility, and anonymization quality requirements, is organized into four key areas: Mobility Impacts on Location Privacy; Attack and Protection of Trajectories; Anonymization Quality; and Applications and Domains-driven Data Utility. Its novelty lies in its comprehensive approach and its potential to address the unique challenges of location privacy in dynamic environments like smart mobility.

As a starting point, we surveyed the current location privacy on mobile networks. We categorized the major groups of location privacy attacks and protection schemes that can mitigate these threats. However, we identified that little has been explored about these mechanisms in the context of open mobility data for smart cities. Particularly, attacks and defense mechanisms that consider the characteristics present in mobility, such as dynamically configuring the parameters of an anonymization-based LPPM over time according to vehicle traffic fluctuations to maximize privacy, need to be developed.

From this study, we address the issue of the impacts of mobility on privacy represented by Mobility Impacts on Location Privacy front. We proposed a framework to characterize and find similarities in the statistical distributions extracted from two Stay Point (SP) metrics. Since SPs serve as fundamental elements in the construction of location privacy protection mechanisms, we propose the hypothesis that their analysis could bring essential insights into the impact of mobility on privacy, including the analysis of different mobility datasets, such as smart mobility. As a result, we identified high similarity between datasets of the same modal type. We have evidenced the hypothesis that mobility can affect privacy. The proof of this hypothesis directed the design of the location privacy mechanisms to consider the mobility aspect, enabling two other contributions represented by the Attack and Protection of Trajectories front: trajectory re-identification attack and dynamic mix-zone.

For Attack and Protection of Trajectories front we introduced a straightforward and efficient trajectory re-identification attack technique that requires only two geo-referenced points as input. This attack leverages the mobility characteristic, where the adversary has knowledge about drivers' behaviors in a city—such as their route preferences—to re-identify trajectories anonymized by mix-zones. Results demonstrated that this approach successfully re-identified up to 95% of anonymized taxi trajectories.

k-DynMix is another contribution to the Attack and Protection of Trajectories front. The k-DynMix is a dynamic mix-zone that adjusts the level of privacy over time in online mode and linear complexity, according to the flow of vehicles, to achieve higher anonymization. We validated our approach using two prediction engines that assess privacy based on accuracy. Additionally, we compared k-DynMix with classic mix-zones by evaluating mix-zone coverage, Anonymization Quality (AQ) metrics, and the AQM framework. Lastly, we examined k-DynMix's effectiveness in anonymizing trajectories against the trajectory re-identification attack contribution. The k-DynMix simultaneously maximized privacy over time and matched the top performance of traditional mix-zones in coverage and AQ metrics. Also, the k-DynMix outperformed the classic mix-zones against the trajectory re-identification attack.

Regarding the Anonymization Quality front, we proposed the Anonymization Quality Framework for Mix-zones (AQM). The Anonymization Quality (AQ) is a concept focused on the protection, efficacy, and internal operation of a concept related to the protection, efficacy, and internal functioning of an LPPM, allowing for an understanding of privacy dynamics and performance assessment over time. The AQM provides a means to characterize and evaluate the impact of anonymization over time and space in mobility data. The AQM enabled the selection of mix-zones based not solely on traffic but also on operational requirements such as LPPM quality, coverage, privacy, and utility metrics. Additionally, with AQM, one can select a positioning algorithm from two available approaches. These aspects impacted the quality of anonymized data, enabling the

measurement of it.

Finally, for the Applications and Domains-driven Data Utility front, we presented the Utility Utility Analysis Framework of Anonymized Trajectories for Smart Cities-Application Domains (UAFAT), which aims to identify domains, applications, and services where the anonymized data will provide more or less utility in various aspects. Moreover, it measures the utility through twelve metrics related to privacy, mobility, and social, including mix-zones performance metrics from anonymized trajectories produced by mix-zones. The results showed that the UAFAT ranked the smart cities application domains that best leverage mobility data anonymized by mix-zones. Moreover, UAFAT identified which of the four smart city application case studies could best use these anonymized data.

In conclusion, the proposed mechanisms, organized on four fronts, presented answers to the research questions raised in this thesis. These studies contributed to significant advancements regarding the design of LPPMs in real-world smart city environments, considering the privacy, utility, and anonymization quality of smart mobility open data that substantially impact the development of smart cities.

9.2 Future Research Directions

This thesis addressed smart mobility open data, in which we proposed an anonymization-based framework that considers the requirements of utility, anonymization quality, and privacy for publishing data. We contributed to each requirement as a framework component and opened new challenges properly described in each chapter's final remarks. Thus, these research opportunities point to new directions and trends in location privacy driven to open data in smart cities. These future directions are summarized as follows:

- Regarding the impact of mobility on privacy, a direction is to extend the proposed framework with theoretical analysis. To pave the hypothesis sustained in this work, it is possible to also extend the framework to other mobility metrics, such as collective mobility metrics, and add more mobility datasets, such as private cars, contact patterns, call details records, and multimodal datasets.

- In future work on trajectory re-identification attack, we plan to compare this solution with additional approaches from the literature, focusing on factors such as result quality, complexity, and execution time. Furthermore, we aim to test our solution with other datasets available in the literature to assess the extent to which the assumptions made here are sufficient for the algorithm. Another proposal is

to combine this attack with other location privacy attacks, such as attacks based in POIs, to improve the re-identification rate. Finally, we propose a compositional attack that combines mobility datasets with other kinds of datasets from different sources to re-identify users.

- For future studies of k-dynmix, we intend to perform a utility analysis of anonymized data with UAFAT, including validating it with datasets from different transportation modes. Another contribution is to compare the distribution of privacy levels produced by k-dynmix with prediction techniques based on time series in terms of correlation, MAE (Mean absolute error), and prediction accuracy.

- Concerning Anonymization Quality, a future direction is to extend the AQM in a simulated VANET environment to better understand AQ behavior in different aspects, such as packet loss scenarios and mix-zone parameter tuning. Another potential proposal is to apply AQM in other anonymization-based LPPMs and multimodal datasets to achieve the results addressed in AQM.

- In the Applications and Domains-driven Data Utility, with the UAFAT, we plan to explore more smart city application domains and metrics, such as privacy, complex networking, and collective mobility metrics in the smart cities context, to gain more accuracy regarding different utility perspectives. Additionally, we plan to add more dense mobility datasets and multimodal datasets to evaluate the sensitivity of the framework. Finally, we intend to explore other anonymization-based LPPMs to pave the results addressed in this work.

In a panoramic vision of this thesis, we focused on anonymization-based LPPMs, particularly the mix-zones. In the anonymization sphere, as future work is to explore other anonymization-based LPPMs that consider smart mobility open data requirements. Also, to explore and add more LPAs to solidify the evaluation of the privacy of the anonymized data. Another important study is to adapt the framework proposed in this thesis to obfuscation-based LPPMs and hybrid-based LPPMs that combine anonymization and obfuscation mechanisms.

Bibliography

[1] David Eckhoff and Isabel Wagner. Privacy in the smart city? Applications, technologies, challenges, and solutions. *IEEE Communications Surveys & Tutorials*, 20(1):489–516, 2017.

[2] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. *arXiv preprint arXiv:1212.1984*, 2012.

[3] Jafar Rezaei. Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega*, 64:126–130, 2016.

[4] Mehdi Sookhak, Helen Tang, and F Richard Yu. Security and Privacy of Smart Cities: Issues and Challenge. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1350–1357. IEEE, 2018.

[5] Zhaolong Ning, Feng Xia, Noor Ullah, Xiangjie Kong, and Xiping Hu. Vehicular social networks: Enabling smart mobility. *IEEE ComMag*, 55(5):16–55, 2017.

[6] Amardeep Das, Sumanta Chandra Mishra Sharma, and Bikram Kesari Ratha. The New Era of Smart Cities, From the Perspective of the Internet of Things. In *Smart Cities Cybersecurity and Privacy*, pages 1–9. Elsevier, 2019.

[7] Iain Docherty, Greg Marsden, and Jillian Anable. The governance of smart mobility. *Transportation Research Part A: Policy and Practice*, 115:114–125, 2018.

[8] Robert Bodenheimer, Alexej Brauer, David Eckhoff, and Reinhard German. Enabling GLOSA for adaptive traffic lights. In *2014 IEEE Vehicular Networking Conference (VNC)*, pages 167–174. IEEE, 2014.

[9] Antoni Martínez-Ballesté, Pablo A Pérez-Martínez, and Agusti Solanas. The pursuit of citizens' privacy: a privacy-aware smart city is possible. *IEEE Communications Magazine*, 51(6):136–141, 2013.

[10] Volker Buscher and Léan Doody. Global innovators: International case studies on smart cities. *BIS Research Paper*, 135, 2013.

[11] Michela Longo, Chowdhury Akram Hossain, and Mariacristina Roscia. Smart mobility for green university campus. In *2013 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pages 1–6. IEEE, 2013.

[12] Michela Longo and Mariacristina Roscia. Sustainable transportation application for smart mobility. In *2014 International Symposium on Power Electronics, Electrical Drives, Automation and Motion*, pages 1054–1059. IEEE, 2014.

[13] Azzedine Boukerche, Antonio AF Loureiro, Eduardo F Nakamura, Horacio ABF Oliveira, Heitor S Ramos, and Leandro A Villas. Cloud-assisted computing for event-driven mobile services. *Mobile Networks and Applications*, 19(2):161–170, 2014.

[14] Francisco R Soriano, Juan Javier Samper-Zapater, Juan Jose Martinez-Dura, Ramon V Cirilo-Gimeno, and Javier Martinez Plume. Smart mobility trends: Open data and other tools. *IEEE Intelligent Transportation Systems Magazine*, 10(2):6–16, 2018.

[15] Mário Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Local Differential Privacy on Metric Spaces: optimizing the trade-off with utility. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 262–267. IEEE, 2018.

[16] Anqi Wang, Anshu Zhang, Edwin HW Chan, Wenzhong Shi, Xiaolin Zhou, and Zhewei Liu. A review of human mobility research based on big data and its implication for smart city development. *ISPRS International Journal of Geo-Information*, 10(1):13, 2020.

[17] Yong Lin, Zhenjiang Shen, and Xiao Teng. Review on Data Sharing in Smart City Planning Based on Mobile Phone Signaling Big Data From the Perspective of China Experience: Anonymization VS De-anonymization. *International Review for Spatial Planning and Sustainable Development*, 9(2):76–93, 2021.

[18] RK Ridder. Business Models for Smart Mobility IoT Data. B.S. thesis, University of Twente, 2019.

[19] Paulo Antonio Maldonado Silveira Alonso Munhoz, Fabricio da Costa Dias, Christine Kowal Chinelli, André Luis Azevedo Guedes, João Alberto Neves dos Santos, Wainer da Silveira e Silva, and Carlos Alberto Pereira Soares. Smart mobility: The main drivers for increasing the intelligence of urban mobility. *Sustainability*, 12(24):10675, 2020.

[20] Ekler Paulino de Mattos, Augusto CSA Domingues, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Slicing who slices: Anonymization quality

evaluation on deployment, privacy, and utility in mix-zones. *Computer Networks*, 236:110007, 2023.

[21] GroundTruth Toyota's Case. How GroundTruth helped Toyota steer high-intent customers to their dealerships: case study— groundtruth results, June 2019. [Online; accessed 01-November-2024].

[22] Jonathan E Spinney. Mobile positioning and LBS applications. *Geography*, pages 256–265, 2003.

[23] National household travel. National household travel survey compendium of uses — january 2017 - december 2017, January 2019. [Online; accessed 01-November-2024].

[24] Anne Immonen, Marko Palviainen, and Eila Ovaska. Requirements of an open data based business ecosystem. *IEEE access*, 2:88–103, 2014.

[25] Chun Sing Lai, Youwei Jia, Zhekang Dong, Dongxiao Wang, Yingshan Tao, Qi Hong Lai, Richard TK Wong, Ahmed F Zobaa, Ruiheng Wu, and Loi Lei Lai. A review of technical standards for smart cities. *Clean Technologies*, 2(3):290–310, 2020.

[26] Mohamed Maouche, Sonia Ben Mokhtar, and Sara Bouchenak. HMC: Robust Privacy Protection of Mobility Data against Multiple Re-Identification Attacks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):124, 2018.

[27] Amy Maxmen. Can tracking people through phone-call data improve lives?, May 2019. [Online; accessed 01-November-2024].

[28] Bin Zan, Zhanbo Sun, Macro Gruteser, and Xuegang Ban. Linking anonymous location traces through driving characteristics. In *Proceedings of the 3rd ACM conference on Data and application security and privacy*, pages 293–300. ACM, 2013.

[29] Yu Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29, 2015.

[30] Zhanbo Sun and Xuegang (Jeff) Ban. Identifying multiclass vehicles using global positioning system data. *Journal of Intelligent Transportation Systems*, 22(1):1–9, 2018.

[31] Marius Wernke, Pavel Skvortsov, Frank Dürr, and Kurt Rothermel. A classification of location privacy attacks and approaches. *Personal and ubiquitous computing*, 18(1):163–175, 2014.

[32] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. The long road to computational location privacy: A survey. *IEEE Communications Surveys & Tutorials*, 2018.

[33] Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, and Yong Xiang. Location privacy and its applications: A systematic study. *IEEE Access*, 6:17606–17624, 2018.

[34] Nazanin Takbiri, Amir Houmansadr, Dennis L Goeckel, and Hossein Pishro-Nik. Limits of location privacy under anonymization and obfuscation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 764–768. IEEE, 2017.

[35] Gonzalo M Vazquez-Prokopec, Donal Bisanzio, Steven T Stoddard, Valerie Paz-Soldan, Amy C Morrison, John P Elder, Jhon Ramirez-Paredes, Eric S Halsey, Tadeusz J Kochel, Thomas W Scott, et al. Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *PloS one*, 8(4):e58802, 2013.

[36] Yuu Nakajima, Hironori Shiina, Shohei Yamane, Toru Ishida, and Hirofumi Yamaki. Disaster evacuation guide: Using a massively multiagent server and GPS mobile phones. In *2007 International Symposium on Applications and the Internet*, pages 2–2. IEEE, 2007.

[37] Y Matsu, M Okazak, and T Sakak. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW 2010: Proceedings of the 19th World Wide Web Conference*, 2010.

[38] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[39] Andrew J Blumberg and Peter Eckersley. On locational privacy, and how to avoid losing it forever. *Electronic frontier foundation*, 10(11), 2009.

[40] John Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

[41] Chris YT Ma, David KY Yau, Nung Kwan Yip, and Nageswara SV Rao. Privacy vulnerability of published anonymous mobility traces. *IEEE/ACM transactions on networking (TON)*, 21(3):720–733, 2013.

[42] Dan Calacci, Alex Berke, Kent Larson, et al. The tradeoff between the utility and risk of location data and implications for public good. *arXiv preprint arXiv:1905.09350*, 2019.

[43] Andreas Pfitzmann and Marit Hansen. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management, 2010. [Online; accessed 02-October-2020].

[44] Jonathan Petit, Florian Schaub, Michael Feiri, and Frank Kargl. Pseudonym schemes in vehicular networks: A survey. *IEEE communications surveys & tutorials*, 17(1):228–255, 2014.

[45] John Krumm. Inference attacks on location tracks. In *International Conference on Pervasive Computing*, pages 127–143. Springer, 2007.

[46] Claudio Bettini. Privacy Protection in Location-Based Services: A Survey. In *Handbook of Mobile Data Privacy*, pages 73–96. Springer, 2018.

[47] Claudia Diaz, Stefaan Seys, Joris Claessens, and Bart Preneel. Towards measuring anonymity. In *International Workshop on Privacy Enhancing Technologies*, pages 54–68. Springer, 2002.

[48] Claudia Diaz. Anonymity metrics revisited. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2006.

[49] Isabel Wagner and David Eckhoff. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):57, 2018.

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

[51] Antoine Boutet, Sonia Ben Mokhtar, and Vincent Primault. *Uniqueness Assessment of Human Mobility on Multi-Sensor Datasets*. PhD thesis, LIRIS UMR CNRS 5205, 2016.

[52] Dionysis Manousakas, Cecilia Mascolo, Alastair R Beresford, Dennis Chan, and Nikhil Sharma. Quantifying privacy loss of human mobility graph topology. *Proceedings on Privacy Enhancing Technologies*, 2018(3):5–21, 2018.

[53] Huandong Wang, Chen Gao, Yong Li, Gang Wang, Depeng Jin, and Jingbo Sun. De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice. In *The 25th Annual Network & Distributed System Security Symposium (NDSS'18)*, 2018.

[54] Shan Chang, Chao Li, Hongzi Zhu, Ting Lu, and Qiang Li. Revealing privacy vulnerabilities of anonymous trajectories. *IEEE TVT*, 2018.

[55] Takao Murakami. A succinct model for re-identification of mobility traces based on small training data. *Training*, 2(x2):x4, 2018.

[56] Ekler P. Mattos, Augusto C.S.A. Domingues, and Antonio A. F. Loureiro. Give Me Two Points and I'll Tell You Who You Are. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV'19)*. IEEE, 2019.

[57] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset). *University of Texas at Austin*, 2008.

[58] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614, 2014.

[59] Xuan Ding, Lan Zhang, Zhiguo Wan, and Ming Gu. A brief survey on de-anonymization attacks in online social networks. In *2010 international conference on computational aspects of social networks*, pages 611–615. IEEE, 2010.

[60] Rahul Deb Das and Stephan Winter. A fuzzy logic based transport mode detection framework in urban environment. *Journal of Intelligent Transportation Systems*, 22(6):478–489, 2018.

[61] Vedran Sekara, Enys Mones, and Håkan Jonsson. Temporal limits of privacy in human behavior. *arXiv preprint arXiv:1806.03615*, 2018.

[62] Arsalan Mosenia, Xiaoliang Dai, Prateek Mittal, and Niraj K Jha. PinMe: Tracking a smartphone user around the world. *IEEE Transactions on Multi-Scale Computing Systems*, 4(3):420–435, 2017.

[63] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.

[64] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24. IEEE, 2006.

[65] Huaxin Li, Haojin Zhu, Suguo Du, Xiaohui Liang, and Xuemin Sherman Shen. Privacy leakage of location sharing in mobile social networks: Attacks and defense. *IEEE Transactions on Dependable and Secure Computing*, 15(4):646–660, 2016.

[66] Lorenzo Franceschi-Bicchierai. Redditor cracks anonymous data trove to pinpoint muslim cab drivers. *Online at: http://mashable. com/2015/01/28/redditor-muslim-cab-drivers*, 2015.

[67] Anthony Tockar. Riding with the stars: Passenger privacy in the nyc taxicab dataset. *Neustar Research, September*, 15, 2014.

[68] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):47, 2014.

[69] Yulong Gu, Yuan Yao, Weidong Liu, and Jiaxing Song. We know where you are: Home location identification in location-based social networks. In *2016 25th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9. IEEE, 2016.

[70] Carmen Ruiz Vicente, Dario Freni, Claudio Bettini, and Christian S Jensen. Location-related privacy in geo-social networks. *IEEE Internet Computing*, 15(3):20–27, 2011.

[71] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying location privacy. In *2011 IEEE symposium on security and privacy*, pages 247–262. IEEE, 2011.

[72] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? Membership inference on aggregate location data. *NDSS Symposium*, 2018.

[73] Gilbert Wondracek, Thorsten Holz, Engin Kirda, and Christopher Kruegel. A practical attack to de-anonymize social network users. In *2010 IEEE Symposium on Security and Privacy*, pages 223–238. IEEE, 2010.

[74] Ge Zhong, Ian Goldberg, and Urs Hengartner. Louis, lester and pierre: Three protocols for location privacy. In *International Workshop on Privacy Enhancing Technologies*, pages 62–76. Springer, 2007.

[75] Igor Bilogrevic, Kévin Huguenin, Murtuza Jadliwala, Florent Lopez, Jean-Pierre Hubaux, Philip Ginzboorg, and Valtteri Niemi. Inferring social ties in academic networks using short-range wireless communications. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 179–188. ACM, 2013.

[76] Chen Wang, Chuyu Wang, Yingying Chen, Lei Xie, and Sanglu Lu. Smartphone privacy leakage of social relationships and demographics from surrounding access points. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 678–688. IEEE, 2017.

[77] Fan Zhou, Bangying Wu, Yi Yang, Goce Trajcevski, Kunpeng Zhang, and Ting Zhong. Vec2Link: Unifying heterogeneous data for social link prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1843–1846. ACM, 2018.

[78] Michael Backes, Mathias Humbert, Jun Pang, and Yang Zhang. walk2friends: Inferring social links from mobility profiles. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1943–1957. ACM, 2017.

[79] Yanwei Yu, Hongjian Wang, and Zhenhui Li. Inferring Mobility Relationship via Graph Embedding. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):147, 2018.

[80] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[81] Huy Pham, Cyrus Shahabi, and Yan Liu. EBM: an entropy-based model to infer social strength from spatiotemporal data. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 265–276. ACM, 2013.

[82] Mohamed Maouche, Sonia Ben Mokhtar, and Sara Bouchenak. Ap-attack: a novel user re-identification attack on mobility datasets. In *Proceedings of the 14th EAI Int. Conf. on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 48–57. ACM, 2017.

[83] Shivendra Tiwari and Saroj Kaushik. Extracting region of interest (roi) details using lbs infrastructure and web-databases. In *2012 IEEE 13th International Conference on Mobile Data Management*, pages 376–379. IEEE, 2012.

[84] Hoa Ngo and Jong Kim. Location privacy via differential private perturbation of cloaking area. In *2015 IEEE 28th Computer Security Foundations Symposium*, pages 63–74. IEEE, 2015.

[85] Peipei Sui, Tianyu Wo, Zhangle Tianyu, and Xianxian Li. Privacy-preserving trajectory publication against parking point attacks. *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing*, 2013.

[86] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. Discovering personal gazetteers: an interactive clustering approach. In

Proceedings of the 12th annual ACM international workshop on Geographic information systems, pages 266–273. ACM, 2004.

[87] Matt Duckham and Lars Kulik. Simulation of obfuscation and negotiation for location privacy. In *International conference on spatial information theory*, pages 31–48. Springer, 2005.

[88] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *2012 IEEE 12th international conference on data mining*, pages 1038–1043. IEEE, 2012.

[89] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

[90] John Krumm and Dany Rouhana. Placer: semantic place labels from diary data. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 163–172. ACM, 2013.

[91] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show me how you move and I will tell you who you are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, pages 34–41. ACM, 2010.

[92] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Time distortion anonymization for the publication of mobility data with high utility. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 539–546. IEEE, 2015.

[93] Osman Abul, Francesco Bonchi, and Mirco Nanni. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884–910, 2010.

[94] Emre Kaplan, Mehmet Emre Gürsoy, Mehmet Ercan Nergiz, and Yücel Saygin. Location disclosure risks of releasing trajectory distances. *Data & Knowledge Engineering*, 113:43–63, 2018.

[95] Ekler P. Mattos, Augusto C.S.A. Domingues, and Antonio A. F. Loureiro. Reidentificação de trajetórias de veículos baseada na caracterização das preferências de caminho. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, 2019.

[96] Nikos Pelekis, Aris Gkoulalas-Divanis, Marios Vodas, Despina Kopanaki, and Yannis Theodoridis. Privacy-aware querying over sensitive trajectory data. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 895–904. ACM, 2011.

[97] Gabriel Ghinita, Maria Luisa Damiani, Claudio Silvestri, and Elisa Bertino. Preventing velocity-based linkage attacks in location-aware applications. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 246–255. ACM, 2009.

[98] Marco Gruteser and Dirk Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 31–42. ACM, 2003.

[99] Ikram Ullah. Location Privacy Schemes in Vehicular Networks: Taxonomy, Comparative Analysis, Design Challenges, and Future Opportunities. *Authorea Preprints*, 2023.

[100] Yuye Zhou and Dongmei Zhang. Double Mix-Zone for Location Privacy in VANET. In *7th Int'l Conf. on Info. Tech.: IoT and Smart City*, pages 322–327, 2019.

[101] Ikjot Saini, Sherif Saad Ahmed, and Arunita Jakel. Attacker placement for detecting vulnerabilities of pseudonym change strategies in VANET. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, pages 1–5. IEEE, 2018.

[102] Shibin Wang and Nianmin Yao. LIAP: A local identity-based anonymous message authentication protocol in VANETs. *Computer Communications*, 112:154–164, 2017.

[103] Abdelwahab Boualouache and Samira Moussaoui. Urban pseudonym changing strategy for location privacy in VANETs. *International Journal of Ad Hoc and Ubiquitous Computing*, 24(1-2):49–64, 2017.

[104] Abdelwahab Boualouache, Sidi-Mohammed Senouci, and Samira Moussaoui. A survey on pseudonym changing strategies for vehicular ad-hoc networks. *IEEE Communications Surveys & Tutorials*, 20(1):770–790, 2017.

[105] Abdelwahab Boualouache and Samira Moussaoui. TAPCS: Traffic-aware pseudonym changing strategy for VANETs. *Peer-to-Peer networking and Applications*, 10:1008–1020, 2017.

[106] Bastian Bloessl, Christoph Sommer, Falko Dressler, and David Eckhoff. The scrambler attack: A robust physical layer attack on location privacy in vehicular networks. In *2015 International Conference on Computing, Networking and Communications (ICNC)*, pages 395–400. IEEE, 2015.

[107] Ines Khacheba, Mohamed B Yagoubi, Nasreddine Lagraa, and Abderrahmane Lakas. Location privacy scheme for VANETs. In *2017 International Conference on*

Selected Topics in Mobile and Wireless Networking (MoWNeT), pages 1–6. IEEE, 2017.

[108] Yuan Yao, Bin Xiao, Gaofei Wu, Xue Liu, Zhiwen Yu, Kailong Zhang, and Xingshe Zhou. Multi-channel based Sybil attack detection in vehicular ad hoc networks using RSSI. *IEEE Transactions on Mobile Computing*, 18(2):362–375, 2018.

[109] Mahesh Pawar and Jitendra Agarwal. A literature survey on security issues of WSN and different types of attacks in network. *Indian J. Comput. Sci. Eng*, 8(2):80–83, 2017.

[110] Rubina S Zuberi and Syed N Ahmad. Secure Mix-Zones for Privacy Protection of Road Network Location Based Services Users. *Journal of Computer Networks and Communications*, 2016(1):3821593, 2016.

[111] Mohammed Ali Hezam Al Junaid, AA Syed, Mohd Nazri Mohd Warip, Ku Nurul Fazira Ku Azir, and Nurul Hidayah Romli. Classification of security attacks in VANET: A review of requirements and perspectives. In *MATEC web of conferences*, volume 150, page 06038. EDP Sciences, 2018.

[112] Amit Kumar Tyagi and N Sreenath. Location privacy preserving techniques for location based services over road networks. In *2015 International Conference on Communications and Signal Processing (ICCSP)*, pages 1319–1326. IEEE, 2015.

[113] Karim Emara. Safety-aware location privacy in VANET: Evaluation and comparison. *IEEE Transactions on Vehicular Technology*, 66(12):10718–10731, 2017.

[114] Qasim Ali Arain, Zhongliang Deng, Imran Memon, Asma Zubedi, and Farman Ali Mangi. Map services based on multiple mix-zones with location privacy protection over road network. *Wireless Personal Communications*, 97:2617–2632, 2017.

[115] Qasim Ali Arain, Imran Memon, Zhongliang Deng, Muhammad Hammad Memon, Farman Ali Mangi, and Asma Zubedi. Location monitoring approach: multiple mix-zones with location privacy protection based on traffic flow over road networks. *Multimedia Tools and Applications*, 77:5563–5607, 2018.

[116] Karim Emara, Wolfgang Woerndl, and Johann Schlichter. Context-based pseudonym changing scheme for vehicular adhoc networks. *arXiv preprint arXiv:1607.07656*, 2016.

[117] Ines Khacheba, Mohamed B Yagoubi, Nasreddine Lagraa, and Abderrahmane Lakas. CLPS: context-based location privacy scheme for VANETs. *International Journal of Ad Hoc and Ubiquitous Computing*, 29(1-2):141–159, 2018.

[118] Marco Gramaglia and Marco Fiore. Hiding mobile traffic fingerprints with glove. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, page 26. ACM, 2015.

[119] Jonathan Mayer, Patrick Mutchler, and John C Mitchell. Evaluating the privacy properties of telephone metadata. *Proceedings of the National Academy of Sciences*, 113(20):5536–5541, 2016.

[120] Luca Rossi, James Walker, and Mirco Musolesi. Spatio-temporal techniques for user identification by means of gps mobility data. *EPJ Data Science*, 4(1):11, 2015.

[121] Zheng Tan, Cheng Wang, Xiaoling Fu, Jipeng Cui, Changjun Jiang, and Weili Han. Re-identification of vehicular location-based metadata. *ICST Trans. Security Safety*, 4(11):e1, 2017.

[122] Alket Cecaj, Marco Mamei, and Nicola Bicocchi. Re-identification of anonymized CDR datasets using social network data. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, pages 237–242. IEEE, 2014.

[123] Takao Murakami, Atsunori Kanemura, and Hideitsu Hino. Group sparsity tensor factorization for de-anonymization of mobility traces. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 621–629. IEEE, 2015.

[124] Takao Murakami and Hajime Watanabe. Localization attacks using matrix and tensor factorization. *IEEE Transactions on Information Forensics and Security*, 11(8):1647–1660, 2016.

[125] Microsoft Research. geolife trajectories (v. 1.3). Downloaded from `research.microsoft.com/jump/131675`, August 2012.

[126] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. CRAWDAD dataset epfl/mobility (v. 2009-02-24). Downloaded from `https://crawdad.org/epfl/mobility/20090224`, February 2009.

[127] Lorenzo Bracciale, Marco Bonola, Pierpaolo Loreti, Giuseppe Bianchi, Raul Amici, and Antonello Rabuffi. CRAWDAD dataset roma/taxi (v. 2014-07-17). Downloaded from `https://crawdad.org/roma/taxi/20140717`, July 2014. [Online; accessed 01-June-2020].

[128] Changrong Li and Zhenfu Li. Trajectory tracking attack for vehicular ad-hoc networks. *Security and Privacy*, 7(6):e433, 2024.

[129] Mirco Musolesi, Kristof Fodor, Mattia Piraccini, Antonio Corradi, and Andrew Campbell. CRAWDAD dataset dartmouth/cenceme. Downloaded from `https://crawdad.org/dartmouth/cenceme/20080813`, August 2008.

[130] M Johanson. How burglars use Facebook to target vacationing homeowners. *International Business Times*, 2013.

[131] Jacob Bellatti, Andrew Brunner, Joseph Lewis, Prasad Annadata, Wisam Eltarjaman, Rinku Dewri, and Ramakrishna Thurimella. Driving habits data: Location privacy implications and solutions. *IEEE Security & Privacy*, 15(1):12–20, 2017.

[132] Nilothpal Talukder and Sheikh Iqbal Ahamed. Preventing multi-query attack in location-based services. In *Proceedings of the third ACM conference on Wireless network security*, pages 25–36. ACM, 2010.

[133] Kazuhiro Minami and Nikita Borisov. Protecting location privacy against inference attacks. In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*, pages 123–126. ACM, 2010.

[134] Adam Sadilek and John Krumm. Far out: Predicting long-term human mobility. In *Twenty-sixth AAAI conference on artificial intelligence*, 2012.

[135] Berker Ağır, Kévin Huguenin, Urs Hengartner, and Jean-Pierre Hubaux. On the privacy implications of location semantics. *Proceedings on Privacy Enhancing Technologies*, 2016(4):165–183, 2016.

[136] John Krumm and Eric Horvitz. Predestination: Inferring destinations from partial trajectories. In *International Conference on Ubiquitous Computing*, pages 243–260. Springer, 2006.

[137] John Krumm, Eric Horvitz, and Julie Letchner. Map matching with travel time constraints. Technical report, SAE Technical Paper, 2007.

[138] Jon Froehlich and John Krumm. Route prediction from trip observations. Technical report, SAE Technical Paper, 2008.

[139] John Krumm. A Markov Model for Driver Turn Prediction. *Society of Automotive Engineers (SAE) 2008 World Congress*, 2008.

[140] Donald J Patterson, Lin Liao, Dieter Fox, and Henry Kautz. Inferring high-level behavior from low-level sensors. In *International Conference on Ubiquitous Computing*, pages 73–89. Springer, 2003.

[141] Hongtao Zhang and Lingcheng Dai. Mobility Prediction: A Survey on State-of-the-Art Schemes and Future Applications. *IEEE Access*, 7:802–822, 2018.

[142] Azzedine Boukerche, Horacio ABF Oliveira, Eduardo F Nakamura, and Antonio AF Loureiro. Vehicular ad hoc networks: A new challenge for localization-based systems. *Computer communications*, 31(12):2838–2849, 2008.

[143] Nandish P Kuruvatti, Wenxiao Zhou, and Hans D Schotten. Mobility prediction of diurnal users for enabling context aware resource allocation. In *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pages 1–5. IEEE, 2016.

[144] Mehammed Daoui, Malika Belkadi, Lynda Chamek, Mustapha Lalam, Sofiane Hamrioui, and Amine Berqia. Mobility prediction and location management based on data mining. In *2012 Next Generation Networks and Services (NGNS)*, pages 137–140. IEEE, 2012.

[145] Yantao Jia, Yuanzhuo Wang, Xiaolong Jin, and Xueqi Cheng. TSBM: The temporal-spatial Bayesian model for location prediction in social networks. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 194–201. IEEE, 2014.

[146] Youssef Khazbak and Guohong Cao. Deanonymizing mobility traces with co-location information. In *2017 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9. IEEE, 2017.

[147] Kaidi Meng, Haojie Li, Zhihui Wang, Xin Fan, Fuming Sun, and Zhongxuan Luo. A Deep Multi-Modal Fusion Approach for Semantic Place Prediction in Social Media. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, pages 31–37. ACM, 2017.

[148] Tarlis T Portela, Francisco Vicenzi, and Vania Bogorny. Trajectory Data Privacy: Research Challenges and Opportunities. In *Brazilian Symposium on GeoInformatics*, pages 99–110, 2019.

[149] Alessandra De Paola, Andrea Giammanco, Giuseppe lo Re, and Giuseppe Anastasi. Detection of Points of Interest in a Smart Campus. In *2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI)*, pages 155–160. IEEE, 2019.

[150] Christopher Moore. Enhancing map data based on points of interest, July 4 2019. US Patent App. 16/236,155.

[151] Andreas Gutscher. Coordinate transformation-a solution for the privacy problem of location based services? In *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, pages 7–pp. IEEE, 2006.

[152] Prateek Mittal, Charalampos Papamanthou, and Dawn Song. Preserving link privacy in social network based systems. *arXiv preprint arXiv:1208.6189*, 2012.

[153] Cynthia Dwork. Differential Privacy. *in Automata, Languages and Programming*, 4052:1–12, 2006.

[154] Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pages 39–54. Springer, 2011.

[155] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer, 2013.

[156] Hidetoshi Kido, Yutaka Yanagisawa, and Tetsuji Satoh. Protection of location privacy using dummies for location-based services. In *21st International Conference on Data Engineering Workshops (ICDEW'05)*, pages 1248–1248. IEEE, 2005.

[157] Hidetoshi Kido, Yutaka Yanagisawa, and Tetsuji Satoh. An anonymous communication technique using dummies for location-based services. In *ICPS'05. Proceedings. International Conference on Pervasive Services, 2005.*, pages 88–97. IEEE, 2005.

[158] Bugra Gedik and Ling Liu. Location privacy in mobile systems: A personalized anonymization model. In *25th IEEE International Conference on Distributed Computing Systems (ICDCS'05)*, pages 620–629. IEEE, 2005.

[159] Claudio Agostino Ardagna, Marco Cremonini, Ernesto Damiani, S De Capitani Di Vimercati, and Pierangela Samarati. Location privacy protection through obfuscation-based techniques. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 47–60. Springer, 2007.

[160] Claudio A Ardagna, Marco Cremonini, Sabrina De Capitani di Vimercati, and Pierangela Samarati. An obfuscation-based approach for protecting location privacy. *IEEE Transactions on Dependable and Secure Computing*, 8(1):13–27, 2009.

[161] Hongbo Jiang, Jie Li, Ping Zhao, Fanzi Zeng, Zhu Xiao, and Arun Iyengar. Location privacy-preserving mechanisms in location-based services: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(1):1–36, 2021.

[162] TS ETSI. 102 867 v1. 1.1-Intelligent Transport Systems (ITS). *Security; Stage 3 Mapping for IEEE*, 1609, 2012.

[163] SAE Standard. Sae j2735 v1.1.1—dedicated short range communications (dsrc) message set dictionary, 2009.

[164] Alastair R Beresford and Frank Stajano. Location privacy in pervasive computing. *IEEE Pervasive computing*, 2(1):46–55, 2003.

[165] Balaji Palanisamy and Ling Liu. Mobimix: Protecting location privacy with mix-zones over road networks. In *2011 IEEE 27th International Conference on Data Engineering*, pages 494–505. IEEE, 2011.

[166] Jiawen Kang, Rong Yu, Xumin Huang, Magnus Jonsson, Hanna Bogucka, Stein Gjessing, and Yan Zhang. Location privacy attacks and defenses in cloud-enabled internet of vehicles. *IEEE Wireless Communications*, 23(5):52–59, 2016.

[167] Augusto CSA Domingues, Ekler Paulino de Mattos, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Social Mix-zones: Anonymizing Personal Information on Contact Tracing Data. In *Proceedings of the 18th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, pages 81–88, 2021.

[168] Zhikai Xu, Hongli Zhang, and Xiangzhan Yu. Multiple mix-zones deployment for continuous location privacy protection. In *2016 IEEE Trustcom/BigDataSE/ISPA*, pages 760–766. IEEE, 2016.

[169] Alisson R Svaigen, Azzedine Boukerche, Linnyer B Ruiz, and Antonio AF Loureiro. Mixdrones: A mix zones-based location privacy protection mechanism for the internet of drones. In *Proceedings of the 24th International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 181–188, 2021.

[170] F Julien, M Raya, M Felegyhazi, and P Papadimitratos. Mixzones for location privacy in vehicular networks. In *Association for Computing Machinery (ACM) Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS)*, 2007.

[171] Zhenyu Chen, Yanyan Fu, Min Zhang, Zhenfeng Zhang, and Hao Li. A Flexible Mix-Zone Selection Scheme Towards Trajectory Privacy Protection. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 1180–1186. IEEE, 2018.

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

[173] Ekler Paulino de Mattos, Augusto CSA Domingues, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Behind the Mix-Zones Scenes: On the Evaluation of the Anonymization Quality. In *Proceedings of the 19th ACM International Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, pages 133–140, 2022.

[174] Kristtopher K Coelho, Maurício M Okuyama, Michele Nogueira, Alex B Vieira, Edelberto F Silva, and José Augusto Miranda Nacif. A Dynamic Approach to Health Data Anonymization by Separatrices. In *2024 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2024.

[175] Alisson Renan Svaigen, Heitor S Ramos, Linnyer B Ruiz, and Antonio AF Loureiro. Dynamic Temporal Mix-Zone Placement Approach for Location-Based Services Privacy. In *2019 IEEE Latin-American Conference on Communications (LATINCOM)*, pages 1–6. IEEE, 2019.

[176] Nirupama Ravi, Mani C Krishna, and Israel Koren. Privacy and Traffic Efficiency under Dynamic Conditions in ITS. *Available at SSRN 4891518*, 2023.

[177] C Kalaiarasy and N Sreenath. An incentive-based co-operation motivating pseudonym changing strategy for privacy preservation in mixed zones in vehicular networks. *Journal of King Saud University-Computer and Information Sciences*, 34(1):1510–1520, 2022.

[178] Xinyang Deng, Tianhan Gao, Nan Guo, Cong Zhao, and Jiayu Qi. PCP: A pseudonym change scheme for location privacy preserving in VANETs. *Entropy*, 24(5):648, 2022.

[179] Richard Gilles Engoulou, Martine Bellaïche, Samuel Pierre, and Alejandro Quintero. Vanet security surveys. *Computer Communications*, 44:1–13, 2014.

[180] Mohamed Amine Ferrag, Leandros Maglaras, and Ahmed Ahmim. Privacy-preserving schemes for ad hoc social networks: A survey. *IEEE Communications Surveys & Tutorials*, 19(4):3015–3045, 2017.

[181] Messaoud Babaghayou, Nabila Labraoui, Ado Adamou Abba Ari, Nasreddine Lagraa, and Mohamed Amine Ferrag. Pseudonym change-based privacy-preserving schemes in vehicular ad-hoc networks: A survey. *Journal of Information Security and Applications*, 55:102618, 2020.

[182] Hesiri Weerasinghe, Huirong Fu, Supeng Leng, and Ye Zhu. Enhancing unlinkability in vehicular ad hoc networks. In *Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics*, pages 161–166. IEEE, 2011.

[183] Rong Yu, Jiawen Kang, Xumin Huang, Shengli Xie, Yan Zhang, and Stein Gjessing. MixGroup: Accumulative pseudonym exchanging for location privacy enhancement in vehicular social networks. *IEEE Transactions on Dependable and Secure Computing*, 13(1):93–105, 2015.

[184] Yuanyuan Pan and Jianqing Li. Cooperative pseudonym change scheme based on the number of neighbors in VANETs. *Journal of Network and Computer Applications*, 36(6):1599–1609, 2013.

[185] Shibin Wang, Nianmin Yao, Ning Gong, and Zhenguo Gao. A trigger-based pseudonym exchange scheme for location privacy preserving in VANETs. *Peer-to-Peer Networking and applications*, 11:548–560, 2018.

[186] Imran Memon, Ling Chen, Qasim Ali Arain, Hina Memon, and Gencai Chen. Pseudonym changing strategy with multiple mix zones for trajectory privacy protection in road networks. *International journal of communication systems*, 31(1):e3437, 2018.

[187] Xinghua Li, Huijuan Zhang, Yanbing Ren, Siqi Ma, Bin Luo, Jian Weng, Jianfeng Ma, and Xiaoming Huang. PAPU: Pseudonym swap with provable unlinkability based on differential privacy in VANETs. *IEEE Internet of Things Journal*, 7(12):11789–11802, 2020.

[188] Bidi Ying and Dimitrios Makrakis. Reputation-based pseudonym change for location privacy in vehicular networks. In *2015 IEEE International Conference on Communications (ICC)*, pages 7041–7046. IEEE, 2015.

[189] Imran Memon, Qasim Ali, Asma Zubedi, and Farman Ali Mangi. DPMM: dynamic pseudonym-based multiple mix-zones generation for mobile traveler. *Multimedia Tools and Applications*, 76:24359–24388, 2017.

[190] Balaji Palanisamy and Ling Liu. Attack-resilient mix-zones over road networks: architecture and algorithms. *IEEE Transactions on Mobile Computing*, 14(3):495–508, 2014.

[191] Abdelwahab Boualouache and Samira Moussaoui. S2si: A practical pseudonym changing strategy for location privacy in vanets. In *2014 International Conference on Advanced Networking Distributed Systems and Applications*, pages 70–75. IEEE, 2014.

[192] Albert Wasef and Xuemin Shen. REP: Location privacy for VANETs using random encryption periods. *Mobile Networks and Applications*, 15:172–185, 2010.

[193] Antonio M Carianha, Luciano Porto Barreto, and George Lima. Improving location privacy in mix-zones for VANETs. In *30th IEEE International Performance Computing and Communications Conference*, pages 1–6. IEEE, 2011.

[194] Levente Buttyán, Tamás Holczer, and István Vajda. On the effectiveness of changing pseudonyms to provide location privacy in VANETs. In *Security and Privacy in Ad-hoc and Sensor Networks: 4th European Workshop, ESAS 2007, Cambridge, UK, July 2-3, 2007. Proceedings 4*, pages 129–141. Springer, 2007.

[195] Rongxing Lu, Xiaodong Lin, Tom H Luan, Xiaohui Liang, and Xuemin Shen. Anonymity analysis on social spot based pseudonym changing for location privacy in VANETs. In *2011 IEEE International Conference on Communications (ICC)*, pages 1–5. IEEE, 2011.

[196] Youhuizi Li, Yuyu Yin, Xu Chen, Jian Wan, Gangyong Jia, and Kewei Sha. A secure dynamic mix zone pseudonym changing scheme based on traffic context prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9492–9505, 2021.

[197] Xinxin Liu, Han Zhao, Miao Pan, Hao Yue, Xiaolin Li, and Yuguang Fang. Traffic-aware multiple mix zone placement for protecting location privacy. In *2012 Proceedings IEEE INFOCOM*, pages 972–980. IEEE, 2012.

[198] Zeng Mengjia and Xu Huibin. Mix-Context-Based Pseudonym Changing Privacy Preserving Authentication in VANETs [J]. *Mobile Information Systems*, 2019:34, 2019.

[199] Nan Guo, Linya Ma, and Tianhan Gao. Independent mix zone for location privacy in vehicular networks. *IEEE Access*, 6:16842–16850, 2018.

[200] Chi-Yin Chow and Mohamed F Mokbel. Trajectory privacy in location-based services and data publication. *ACM Sigkdd Explorations Newsletter*, 13(1):19–29, 2011.

[201] Bin Yang, Chenjuan Guo, and Christian S Jensen. Travel cost inference from sparse, spatio temporally correlated time series using Markov models. *Proceedings of the VLDB Endowment*, 6(9):769–780, 2013.

[202] Zhenyu Chen, Yanyan Fu, Min Zhang, Zhenfeng Zhang, and Hao Li. The De-anonymization method based on user spatio-temporal mobility trace. In *International Conference on Information and Communications Security*, pages 459–471. Springer, 2017.

[203] C Kalaiarasy, N Sreenath, and A Amuthan. An effective variant ring signature-based pseudonym changing mechanism for privacy preservation in mixed zones of

vehicular networks. *Journal of Ambient Intelligence and Humanized Computing*, 11:1669–1681, 2020.

[204] Nirupama Ravi, C Mani Krishna, and Israel Koren. Mix-Zones as an Effective Privacy Enhancing Technique in Mobile and Vehicular Ad-hoc Networks. *ACM Computing Surveys*, 56(12):1–33, 2024.

[205] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. On the optimal placement of mix zones. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 216–234. Springer, 2009.

[206] Yipin Sun, Bofeng Zhang, Baokang Zhao, Xiangyu Su, and Jinshu Su. Mix-zones optimal deployment for protecting location privacy in VANET. *Peer-to-Peer Networking and Applications*, 8(6):1108–1121, 2015.

[207] Imran Memon and Qasim Ali Arain. Optimal placement of mix zones in road networks. *arXiv preprint arXiv:1705.11104*, 2017.

[208] Nirupama Ravi, C Mani Krishna, and Israel Koren. Enhancing Vehicular Anonymity in ITS: A New Scheme for Mix Zones and their Placement. *IEEE Transactions on Vehicular Technology*, 2019.

[209] Reza Shokri, Julien Freudiger, Murtuza Jadliwala, and Jean-Pierre Hubaux. A distortion-based metric for location privacy. In *Proceedings of the 8th ACM workshop on Privacy in the electronic society*, pages 21–30. ACM, 2009.

[210] Simon Oya, Carmela Troncoso, and Fernando Pérez-González. Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1959–1972. ACM, 2017.

[211] Jian Wang, Yameng Shao, Jianqi Zhu, and Yuming Ge. Spatio-temporal location privacy quantification for vehicular networks. *IEEE Access*, 6:62963–62974, 2018.

[212] Darakhshan J Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N Wright. Dp-where: Differentially private modeling of human mobility. In *2013 IEEE international conference on big data*, pages 580–588. IEEE, 2013.

[213] Cynthia Dwork. Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340, 2011.

[214] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. Adaptive location privacy with alp. In *2016 IEEE 35th Symposium on Reliable Distributed Systems (SRDS)*, pages 269–278. IEEE, 2016.

[215] Sophie Cerf, Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, Robert Birke, Sara Bouchenak, Lydia Y Chen, Nicolas Marchand, and Bogdan Robu. PULP: achieving privacy and utility trade-off in user mobility data. In *2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS)*, pages 164–173. IEEE, 2017.

[216] Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In *International Workshop on Privacy Enhancing Technologies*, pages 41–53. Springer, 2002.

[217] Zhendong Ma, Frank Kargl, and Michael Weber. Measuring long-term location privacy in vehicular communication systems. *Computer Communications*, 33(12):1414–1427, 2010.

[218] Baik Hoh and Marco Gruteser. Protecting location privacy through path confusion. In *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM'05)*, pages 194–205. IEEE, 2005.

[219] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2575–2582. IEEE, 2018.

[220] Christoph Sommer, David Eckhoff, Alexander Brummer, Dominik S Buse, Florian Hagenauer, Stefan Joerer, and Michele Segata. Veins: The open source vehicular network simulation framework. *Recent advances in network simulation: the OMNeT++ environment and its ecosystem*, pages 215–252, 2019.

[221] Seyed Morteza Mousavi, Hamid R Rabiee, M Moshref, and Ali Dabirmoghaddam. Mobisim: A framework for simulation of mobility models in mobile ad-hoc networks. In *Third IEEE international conference on wireless and mobile computing, networking and communications (WiMob 2007)*, pages 82–82. IEEE, 2007.

[222] András Varga and Rudolf Hornig. An overview of the OMNeT++ simulation environment. In *1st International ICST Conference on Simulation Tools and Techniques for Communications, Networks and Systems*, 2010.

[223] Thomas R Henderson, Mathieu Lacage, George F Riley, Craig Dowell, and Joseph Kopena. Network simulations with the ns-3 simulator. *SIGCOMM demonstration*, 14(14):527, 2008.

[224] Yeong-Sheng Chen, Tang-Te Lo, Chiu-Hua Lee, and Ai-Chun Pang. Efficient pseudonym changing schemes for location privacy protection in VANETs. In *2013*

International Conference on Connected Vehicles and Expo (ICCVE), pages 937–938. IEEE, 2013.

[225] Murtuza Jadliwala, Igor Bilogrevic, and Jean-Pierre Hubaux. Optimizing mix-zone coverage in pervasive wireless networks. *Journal of computer security*, 21(3):317–346, 2013.

[226] Bidi Ying, Dimitrios Makrakis, and Hussein T Mouftah. Dynamic mix-zone for location privacy in vehicular networks. *IEEE Communications Letters*, 17(8):1524–1527, 2013.

[227] Bidi Ying and Dimitrios Makrakis. Pseudonym changes scheme based on candidate-location-list in vehicular networks. In *2015 IEEE International Conference on Communications (ICC)*, pages 7292–7297. IEEE, 2015.

[228] Abdelwahab Boualouache, Sidi-Mohammed Senouci, and Samira Moussaoui. Vlpz: The vehicular location privacy zone. *Procedia Computer Science*, 83:369–376, 2016.

[229] Hugo Barbosa et al. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[230] Shixiong Jiang, Wei Guan, Wenyi Zhang, Xu Chen, and Liu Yang. Human mobility in space from three modes of public transportation. *Physica A: Statistical Mechanics and its Applications*, 483:227–238, 2017.

[231] Sara Paiva, Mohd Abdul Ahad, Sherin Zafar, Gautami Tripathi, Aqeel Khalique, and Imran Hussain. Privacy and security challenges in smart and sustainable mobility. *SN Applied Sciences*, 2:1–10, 2020.

[232] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914, 2013.

[233] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800, 2009.

[234] Qingying Yu, Yonglong Luo, Chuanming Chen, and Xiaoyao Zheng. Road Congestion Detection Based on Trajectory Stay-Place Clustering. *ISPRS International Journal of Geo-Information*, 8(6):264, 2019.

[235] Diego Minatel, Vinícius Ferreira, and Alneu Andrade Lopes. A multilevel approach for building location-based social network by using stay points. In *8th Brazilian Conf. on Intelligent Systems*, pages 359–364. IEEE, 2019.

[236] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[237] Kai Zhao, Mirco Musolesi, Pan Hui, Weixiong Rao, and Sasu Tarkoma. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Scientific reports*, 5(1):1–7, 2015.

[238] Feng Xia, Jinzhong Wang, Xiangjie Kong, Zhibo Wang, Jianxin Li, and Chengfei Liu. Exploring human mobility patterns in urban scenarios: A trajectory data perspective. *IEEE ComMag*, 56(3):142–149, 2018.

[239] Subhayan Bhattacharya, Sankhamita Sinha, Sarbani Roy, and Amarnath Gupta. Towards finding the best-fit distribution for OSN data. *The Journal of Supercomputing*, pages 1–19, 2020.

[240] Juan Li, Hui Zhang, Yanru Zhang, and Xuan Zhang. Modeling Drivers' Stopping Behaviors during Yellow Intervals at Intersections considering Group Heterogeneity. *Journal of advanced transportation*, 2020, 2020.

[241] Laura Alessandretti, Piotr Sapiezynski, Sune Lehmann, and Andrea Baronchelli. Multi-scale spatio-temporal analysis of human mobility. *PloS one*, 12(2):e0171686, 2017.

[242] Michela Papandrea, Karim Keramat Jahromi, Matteo Zignani, Sabrina Gaito, Silvia Giordano, and Gian Paolo Rossi. On the properties of human mobility. *Computer Communications*, 87:19–36, 2016.

[243] Zhu Xiao, Hui Xiao, Wenjie Chen, Hongyang Chen, Amelia Regan, and Hongbo Jiang. Exploring Human Mobility Patterns and Travel Behavior: A Focus on Private Cars. *IEEE Intelligent Transportation Systems Magazine*, pages 2–19, 2021.

[244] Jie Zhang, WANG Junhua, and FANG Shou'en. Prediction of urban expressway total traffic accident duration based on multiple linear regression and artificial neural network. In *2019 5th Int. Conf. on Transportation Information and Safety (ICTIS)*, pages 503–510. IEEE, 2019.

[245] Joahannes B D. da Costa, Allan M de Souza, Denis Rosário, Eduardo Cerqueira, and Leandro A Villas. Efficient data dissemination protocol based on complex networks' metrics for urban vehicular networks. *Journal of Internet Services and Applications*, 10(1):15, 2019.

[246] Pierre Barthelemy, Jacopo Bertolotti, and Diederik S Wiersma. A Lévy flight for light. *Nature*, 453(7194):495–498, 2008.

[247] B Clifford Neuman and Theodore Ts'o. Kerberos: An authentication service for computer networks. *IEEE Communications magazine*, 32(9):33–38, 1994.

[248] Yuan Zhang, Chunxiang Xu, Hongwei Li, Kan Yang, Nan Cheng, and Xuemin Shen. PROTECT: efficient password-based threshold single-sign-on authentication for mobile users against perpetual leakage. *IEEE Transactions on Mobile Computing*, 20(6):2297–2312, 2020.

[249] Yuan Zhang, Chunxiang Xu, Nan Cheng, and Xuemin Sherman Shen. Secure Password-Protected Encryption Key for Deduplicated Cloud Storage Systems. *IEEE Transactions on Dependable and Secure Computing*, 2021.

[250] Karim Emara, Wolfgang Woerndl, and Johann Schlichter. CAPS: Context-aware privacy scheme for VANET safety applications. In *Proc. of the 8th ACM Conf. on security & privacy in wireless and mobile networks*, pages 1–12, 2015.

[251] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[252] Crawdad. Crawdad: A Community Resource for Archiving Wireless Data At Dartmouth. `https://crawdad.org/keyword-vehicular-network.html`, August 2020.

[253] Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 316–324, 2011.

[254] Dublin City Council. Dublin Bus GPS sample data from Dublin City Council (Insight Project), June 2013. [Online; accessed 01-Jun-2020].

[255] Jie Chen, Zhu Xiao, Dong Wang, Wangchen Long, Jing Bai, and Vincent Havyarimana. Stay time prediction for individual stay behavior. *IEEE Access*, 7:130085–130100, 2019.

[256] Jung Hun Oh, Maryam Pouryahya, Aditi Iyer, Aditya P. Apte, Joseph O. Deasy, and Allen Tannenbaum. A novel kernel Wasserstein distance on Gaussian measures: An application of identifying dental artifacts in head and neck computed tomography. *Computers in Biology and Medicine*, 120:103731, 2020.

[257] Artur Souza, Luigi Nardi, Leonardo B Oliveira, Kunle Olukotun, Marius Lindauer, and Frank Hutter. Bayesian optimization with a prior for the optimum. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 265–296. Springer, 2021.

[258] Rob Matheson. The privacy risks of compiling mobility data, 2018.

[259] Augusto CSA Domingues, Fabrício A Silva, and Antonio AF Loureiro. Space and time matter: An analysis about route selection in mobility traces. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00958–00963. IEEE, 2018.

[260] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[261] Ekler P de Mattos, Augusto CSA Domingues, Bruno P Santos, Heitor S Ramos, and Antonio AF Loureiro. The impact of mobility on location privacy: A perspective on smart mobility. *IEEE Systems Journal*, 16(4):5509–5520, 2022.

[262] Mohammad Khodaei and Panos Papadimitratos. Cooperative location privacy in vehicular networks: why simple mix zones are not enough. *IEEE Internet of Things Journal*, 8(10):7985–8004, 2020.

[263] Christian Vaas, Mohammad Khodaei, Panos Papadimitratos, and Ivan Martinovic. Nowhere to hide? Mix-zones for private pseudonym change using chaff vehicles. In *2018 IEEE Vehicular Networking Conference (VNC)*, pages 1–8. IEEE, 2018.

[264] Luca Pappalardo and Filippo Simini. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, 32(3):787–829, 2018.

[265] Fabrício R de Souza, Augusto CSA Domingues, Pedro OS Vaz de Melo, and Antonio AF Loureiro. Mocha: A tool for mobility characterization. In *Proceedings of the 21st ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 281–288, 2018.

[266] Pedro OS Vaz de Melo, Aline C Viana, Marco Fiore, Katia Jaffrès-Runser, Frédéric Le Mouël, and Antonio AF Loureiro. Recast: Telling apart social and random relationships in dynamic networks. In *Proceedings of the 16th ACM international conference on Modeling, analysis & simulation of wireless and mobile systems*, pages 327–334, 2013.

[267] Alisson R Svaigen, Azzedine Boukerche, Linnyer B Ruiz, and Antonio AF Loureiro. BioMixD: A bio-inspired and traffic-aware mix zone placement strategy for location privacy on the internet of drones. *Computer Communications*, 195:111–123, 2022.

[268] Anna Izabel J Tostes, Fátima de LP Duarte-Figueiredo, Renato Assunção, Juliana Salles, and Antonio AF Loureiro. From data to knowledge: City-wide traffic flows

analysis and prediction using bing maps. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pages 1–8, 2013.

[269] Ruizhi Liao. Smart mobility: challenges and trends. *Toward Sustainable And Economic Smart Mobility: Shaping The Future Of Smart Cities*, 10:1, 2020.

[270] A Costandoiu and M Leba. Convergence of V2X communication systems and next generation networks. In *IOP Conference Series: Materials Science and Engineering*, volume 477, page 012052. IOP Publishing, 2019.

[271] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[272] Abdueli Paulo Mdee, Malik Muhammad Saad, Murad Khan, Muhammad Toaha Raza Khan, and Dongkyun Kim. Impacts of location-privacy preserving schemes on vehicular applications. *Vehicular Communications*, page 100499, 2022.

[273] Yuting Zhan, Hamed Haddadi, and Afra Mashhadi. Analysing Fairness of Privacy-Utility Mobility Models. *arXiv preprint arXiv:2304.06469*, 2023.

[274] Jong Wook Kim, Kennedy Edemacu, and Beakcheol Jang. Privacy-preserving mechanisms for location privacy in mobile crowdsensing: A survey. *Journal of Network and Computer Applications*, page 103315, 2022.

[275] Nathalie E Williams, Timothy A Thomas, Matthew Dunbar, Nathan Eagle, and Adrian Dobra. Measures of human mobility using mobile phone records enhanced with GIS data. *PloS one*, 10(7):e0133630, 2015.

[276] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6(1):8166, 2015.

[277] Rafael Verão Françozo, Luiz Sérgio Velasquez Urquiza Junior, Elana Souza Carrapateira, Bruna Cristine Scarduelli Pacheco, Márcio Teixeira Oliveira, Guilherme Botega Torsoni, and Jiyan Yari. A web-based software for group decision with analytic hierarchy process. *MethodsX*, 11:102277, 2023.

[278] Jafar Rezaei. Best-worst multi-criteria decision-making method. *Omega*, 53:49–57, 2015.

[279] Xiaomei Mi, Ming Tang, Huchang Liao, Wenjing Shen, and Benjamin Lev. The state-of-the-art survey on integrations and applications of the best worst method in decision making: Why, what, what for and what's next? *Omega*, 87:205–225, 2019.

[280] Meenu Singh and Millie Pant. A review of selected weighing methods in MCDM with a case study. *International Journal of System Assurance Engineering and Management*, 12:126–144, 2021.

[281] Yourong Huang, Zhu Xiao, Xiaoyou Yu, Dong Wang, Vincent Havyarimana, and Jing Bai. Road network construction with complex intersections based on sparsely sampled private car trajectory data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(3):1–28, 2019.

[282] Pablo Barbecho Bautista, Leticia Lemus Cárdenas, Luis Urquiza Aguiar, and Mónica Aguilar Igartua. A traffic-aware electric vehicle charging management system for smart cities. *Vehicular Communications*, 20:100188, 2019.

[283] Mohammadreza Kavianipour, Fatemeh Fakhrmoosavi, Harprinderjot Singh, Mehrnaz Ghamami, Ali Zockaie, Yanfeng Ouyang, and Robert Jackson. Electric vehicle fast charging infrastructure planning in urban networks considering daily travel and charging behavior. *Transportation Research Part D: Transport and Environment*, 93:102769, 2021.

[284] Junwei Zhang, Fan Yang, Zhuo Ma, Zhuzhu Wang, Ximeng Liu, and Jianfeng Ma. A decentralized location privacy-preserving spatial crowdsourcing for internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 22(4):2299–2313, 2020.

[285] Leonardo Sarmiento and Anna Förster. TRAILS mobility model. *Simulation*, 99(4):385–402, 2023.

[286] Rei Yamazaki, Masashi Yoshida, and Hiroshi Shigeno. A Dynamic Mix-zone Scheme Considering Communication Delay for Location Privacy in Vehicular Networks. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 245–250. IEEE, 2021.

[287] Ekler Paulino de Mattos, Augusto CSA Domingues, Fabrício A Silva, Heitor S Ramos, and Antonio AF Loureiro. Protect your Data and I'll Show Its Utility: A Practical View about Mix-zones Impacts on Mobility Data for Smart City Applications. In *Proceedings of the Int'l ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, & Ubiquitous Networks*, pages 45–52, 2023.

[288] Vern Paxson, Mark Allman, Jerry Chu, and Matt Sargent. Computing TCP's retransmission timer. Technical report, 2011.

[289] Seng Hansun. A new approach of moving average method in time series analysis. In *2013 conference on new media studies (CoNMedia)*, pages 1–4. IEEE, 2013.

Appendix A

Privacy Level Prediction Approaches

The Simple Moving Average (SMA) is a widely used method to compute the average of the preceding n data points in a time series dataset. Each data point holds equal importance in SMA without any applied weighting factors, ensuring an unbiased calculation. The SMA is defined as $SMA = \frac{P_t + P_{t-1} + \cdots + P_{t-(n-1)}}{n}$, where P_t is the point at time t and n stands for the numbers of data points used in the calculation.

The Weighted Exponential Moving Average (WEMA) is a variation of the Exponential Moving Average (EMA) that assigns varying levels of importance to data points. Unlike the EMA, which treats all data points equally, WEMA allows for customized weighting schemes through specific weighting functions. This enables the emphasis of recent or significant data points over others, providing more flexibility in analysis. The WEMA is denoted by $WEMA_t = \alpha P_t + (1 - \alpha)\, WMA_t$, where P_t is value at period t, α represents the the degree of weighting decrease as in equation $\alpha = \frac{2}{(n+1)}$, and WMA_t is Weighted Moving Average (WMA) at time t [289].