


4-10-2018

User-Centric Privacy Preservation in Mobile and Location-Aware Applications

Mingming Guo
mguo001@fiu.edu

DOI: 10.25148/etd.FIDC006533

Follow this and additional works at: <https://digitalcommons.fiu.edu/etd>

 Part of the [Artificial Intelligence and Robotics Commons](#), [Databases and Information Systems Commons](#), [Digital Communications and Networking Commons](#), [Information Security Commons](#), [Probability Commons](#), and the [Theory and Algorithms Commons](#)

Recommended Citation

Guo, Mingming, "User-Centric Privacy Preservation in Mobile and Location-Aware Applications" (2018). *FIU Electronic Theses and Dissertations*. 3674.

<https://digitalcommons.fiu.edu/etd/3674>

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

USER-CENTRIC PRIVACY PRESERVATION IN MOBILE AND LOCATION-
AWARE APPLICATIONS

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Mingming Guo

2018

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Mingming Guo, and entitled User-Centric Privacy Preservation in Mobile and Location-Aware Applications, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Deng Pan

Bogdan Carbunar

Kang Yen

S.S. Iyengar, Co-Major Professor

Niki Pissinou, Co-Major Professor

Date of Defense: April 10, 2018

The dissertation of Mingming Guo is approved.

Dean John L. Volakis
College of Engineering and Computing

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2018

© Copyright 2018 by Mingming Guo

All rights reserved.

DEDICATION

To my beloved family.

ACKNOWLEDGMENTS

First, I would like to express my great gratitude to my advisor and mentor, Dr. Niki Pissinou, whose encouragement, guidance and support from the initial to the final level enabled me to complete this work. Her encouraging advice and strict training have been guiding me to conduct ethical and independent research.

I would like to thank my co-advisor, the Director of School of Computing and Information Sciences, Dr. S.S. Iyengar for his support and encouragement from a higher level for my doctoral studies and research. Without Dr. Iyengar's enormous support to me on my publications, this research would have not been possible.

Many thanks to Dr. Deng Pan, who has been served as my dissertation committee as well as my qualifier committee. His attitude and hard-working style for research has set a wonderful example for me to follow. Thank Dr. Bogdan Carbutar, who has been guiding me on research problems and encouraging me to resolve them independently. I also appreciate Dr. Kang Yen for serving on my dissertation committee and giving me valuable comments.

I would like to acknowledge the interactive research environment provided by our research laboratory. Special thanks go to Mr. Jerry Miller, the research coordinator in Discovery Lab, for his effort on coordinating our research projects. Thank all the colleagues, including Dr. Xinyu Jin, Dr. Sitthapon Pumpichet, Dr. Hao Jin, Dr. Charles Kamhoua, Concepción Sánchez Alemán, Samia Tasnim, Georges A. Kamhoua, Abdur Rahman Bin Shahid, and Sanjeev Kaushik Ramani.

Thanks to Florida International University, U.S. National Science Foundation - RET and REU Sites, Air Force Office of Scientific Research, and Army Research Office for

supporting my research. Thanks to University Graduate School for supporting my research with a Dissertation Year Fellowship in the final stage of my doctoral study.

Finally, I would like to show my deepest gratitude to my entire families for giving me unconditional love, support and inspiration to pursue my dreams in the United States of America.

ABSTRACT OF THE DISSERTATION

USER-CENTRIC PRIVACY PRESERVATION IN MOBILE AND LOCATION-
AWARE APPLICATIONS

by

Mingming Guo

Florida International University, 2018

Miami, Florida

Professor Niki Pissinou, Co-Major Professor

Professor S.S. Iyengar, Co-Major Professor

The mobile and wireless community has brought a significant growth of location-aware devices including smart phones, connected vehicles and IoT devices. The combination of location-aware sensing, data processing and wireless communication in these devices leads to the rapid development of mobile and location-aware applications. Meanwhile, user privacy is becoming an indispensable concern. These mobile and location-aware applications, which collect data from mobile sensors carried by users or vehicles, return valuable data collection services (e.g., health condition monitoring, traffic monitoring, and natural disaster forecasting) in real time. The sequential spatial-temporal data queries sent by users provide their location trajectory information. The location trajectory information not only contains users' movement patterns, but also reveals sensitive attributes such as users' personal habits, preferences, as well as home and work addresses. By exploring this type of information, the attackers can extract and sell user profile data, decrease subscribed data services, and even jeopardize personal safety.

This research spans from the realization that user privacy is lost along with the popular usage of emerging location-aware applications. The outcome seeks to relive user location and trajectory privacy problems. First, we develop a pseudonym-based anonymity zone generation scheme against a strong adversary model in continuous location-based services. Based on a geometric transformation algorithm, this scheme generates distributed anonymity zones with personalized privacy parameters to conceal users' real location trajectories. Second, based on the historical query data analysis, we introduce a query-feature-based probabilistic inference attack, and propose query-aware randomized algorithms to preserve user privacy by distorting the probabilistic inference conducted by attackers. Finally, we develop a privacy-aware mobile sensing mechanism to help vehicular users reduce the number of queries to be sent to the adversarial servers. In this mechanism, mobile vehicular users can selectively query nearby nodes in a peer-to-peer way for privacy protection in vehicular networks.

TABLE OF CONTENTS

CHAPTER	PAGE
1. INTRODUCTION	1
1.1 Background and Motivating Applications	2
1.2 Research Challenges	3
1.3 Research Objectives	5
1.4 Research Contributions	7
1.5 Dissertation Outline	10
2. RELATED WORK	11
2.1 Location and Trajectory Privacy Protection in LBS	12
2.1.1 Trusted Third-Party Based Techniques	12
2.1.2 Semi-Trusted Third-Party Techniques	15
2.1.3 Trusted Third-Party Free Techniques	16
2.2 Location and Trajectory Privacy Preservation in MSN	23
2.2.1 Location Trajectory Privacy of Mobile Sinks	24
2.2.2 Location Trajectory Privacy of Mobile Sensors	25
2.3 Location and Trajectory Privacy in P2P, MANET and VANET	27
2.4 Query Privacy Preservation in LBS	30
2.5 Summary	35
3. PSEUDONYM-BASED ANONYMITY ZONE GENERATION FOR LBS	36
3.1 Introduction	36
3.2 System Model	37
3.2.1 System Architecture	37
3.2.2 Attack Model	37
3.3 Pseudonym-based Mechanism Design	38
3.3.1 The Basic Geometric Transformation Process	38
3.3.2 The Dynamic Pseudonym Changing Process	40
3.3.3 The Personalized Dummies Generation Process	40
3.3.4 Results Filter Deployed on User Devices	41
3.4 Analysis and Simulation	43
3.4.1 Simulation Setup	44
3.4.2 Privacy Level	44
3.4.3 Communication Cost	45
3.4.4 Performance Evaluation	49
3.5 Summary	55
4. QUERY-FEATURE-BASED ATTACKS AND PRIVACY PROTECTION	56
4.1 Introduction	56
4.2 System Model	57
4.2.1 System Architecture	57
4.2.2 Feature-based Inference Attack Model	57
4.2.3 Problem Statement	61

4.3 Query-Aware Privacy Algorithm Design	62
4.4 Analysis and Performance Evaluation	67
4.4.1 Privacy Level by Entropy	67
4.4.2 Utility Analysis	70
4.4.3 Approximation Ratio	70
4.4.4 Computation Complexity	73
4.4.5 Communication Cost	74
4.5 Summary	77
5. PRIVACY-AWARE MOBILE SENSING MECHANISM FOR VANET	78
5.1 Introduction	78
5.2 System Model	80
5.2.1 System Architecture	80
5.2.2 Attack Model	80
5.2.3 Problem Statement	81
5.3 Privacy-Aware Mobile Sensing Mechanism for VANET	81
5.4 Comparative Analysis	84
5.4.1 Privacy-Level Analysis	85
5.4.2 Service-Quality Analysis	86
5.5 Performance Evaluation	87
5.5.1 Simulation Setup	87
5.5.2 Performance Evaluation	88
5.6 Summary	94
6. LIMITATIONS, FUTURE WORK AND CONCLUSION	95
6.1 Limitations	95
6.1.1 Pseudonym-based Distributed Anonymity Zone Generation Scheme	95
6.1.2 Query-Feature-based Inference Attacks and Defense Randomized Algorithms	96
6.1.3 Cache-based Mobile Sensing Mechanism for Mobile Vehicles	97
6.2 Future Work	98
6.2.1 A Comprehensive Understanding of the Attack Space from Advanced Attackers	98
6.2.2 An Intelligent and Adaptive Privacy-Preserving and Secure Framework	100
6.3 Conclusion	101
BIBLIOGRAPHY	103
VITA	113

LIST OF FIGURES

FIGURE	PAGE
3.1 LBS system architecture	37
3.2 Communication cost when radius of ROIs $R = 2, 3, 4, 5, 6, 7, 8, 9$ and 10 ($\times 100$ meters).....	46
3.3 Communication cost when anonymity parameter $k = 2, 3, 4, 5, 6, 7, 8, 9$ and 10	48
3.4 Privacy level when anonymity parameter $k = 2, 3, 4, 5, 6, 7, 8, 9$ and 10	50
3.5 Privacy level when radius of ROIs $R = 2, 3, 4, 5, 6, 7, 8, 9$ and 10 ($\times 100$ meters)....	51
3.6 Overall Anonymity Zone vs. Anonymity Parameter k	52
3.7 Overall Anonymity Zone vs. Radius of ROIs R	53
4.1 System service architecture.....	57
4.2 The probabilities for the top 10 girds.....	59
4.3 Location probability distribution example.....	64
4.4 Entropy E vs. Privacy Parameter k	69
4.5 Entropy E vs. Iteration Parameter r	69
4.6 Approximation Ratio R vs. Iteration Parameter r	71
4.7 Approximation Ratio R vs. Privacy Parameter k	72
4.8 Average Communication Costs vs. Algorithms.....	76
5.1 Mobile system architecture	80
5.2 Mobile sensing privacy mechanism logic flow	82
5.3 The statistical information about the trace length for all users.....	88
5.4 Privacy Gain G vs. Percentage of Sensitive Contexts P	89
5.5 Privacy Gain G vs. Percentage of Reduced Query Numbers N	90

5.6 Quality of Service Q vs. Percentage of Sensitive Contexts P.....	91
5.7 Quality of Service Q vs. Percentage of Reduced Query Numbers N	93

CHAPTER 1

INTRODUCTION

With the advancement of mobile and wireless networks, as well as the recent development of affordable mobile sensors, location-aware mobile services and applications are rapidly growing. Meanwhile, privacy is becoming an indispensable concern. These services and applications, which collect data from mobile sensors carried by users or vehicles, return valuable data collection services (e.g., health condition monitoring, traffic monitoring, and natural disaster forecasting) in real time. Mobile nodes, such as users or vehicles, continuously communicate with the base station or peer nodes to exchange data through wireless media. The sequential spatial-temporal data queries sent by users or vehicles form their location trajectory information. The location trajectory information not only contains the movement patterns of the mobile sensors, but also reveals the sensitive attributes like users' personal habits, preferences as well as other valuable information such as home and work addresses. By exploring this type of location trajectory information, the attackers can extract and sell user profile data, decrease subscribed data services, and even jeopardize personal safety.

User privacy issues stem from the fact that many mobile service providers aggressively collect sensitive data without clear statements about how to use and share user data beyond the regular services. A study shows that half of the popular Android applications they investigated built user profiles for other purposes, such as targeted advertising and selling to third parties for revenue without user agreements [EGC⁺10]. In addition, peer nodes can launch privacy attacks within the network when a mobile node decides to transmit data

through them. The location data sharing experience can lead to location and trajectory privacy leakage and even serious safety vulnerabilities. There is limited effort to solve this problem and more research participation is needed.

1.1 Background and Motivating Applications

This research aims to develop practical privacy preservation techniques for a variety of location-aware applications that can use location data for providing valuable services. For example, location-aware mobile applications, such as Google Maps [GM18] and Foursquare [Fou18], can use a user's current location to provide different Point-of-Interests (POIs). A restaurant or a gas station finder app can use users' real-time coordinates to find the closest restaurants or gas stations in the search region. A car navigator app can utilize users' current location to provide navigation services. These location-aware apps collect users' location data either in a one-shot way or a continuous way.

The installation of mobile sensors or wearable devices can provide users with valuable monitoring services, such as healthcare applications, etc. Sensor data, including location (where), time stamp (when), and sensed value (what), can be transmitted to the base station, the internet, and finally reach the third-parties which can provide instant insights and actions based on what they received. For example, mobile sensors that are carried by human beings, such as heart beat sensor [Mon17], can be used for monitoring real-time heart conditions, critical for instant care. An online doctor service can also send alerts or suggested actions to the patients themselves or to other people to help the patients in real time.

Mobile sensors that are installed on mobile vehicles can be used to report road conditions, traffic monitoring and driving habit detecting. For example, a new driver wants to buy insurance for his or her vehicle. The insurance companies may require the driver to install a mobile sensor to report his/her driving habit in a certain time period to determine the insurance price [Ins16]. Thus, the driver's location trajectories are recorded and may even be analyzed by the insurance companies [Tra14].

However, those location data sharing behaviors can pose serious threats to users' location and trajectory privacy. For example, the insurance companies can find more sensitive and personal information by analyzing the driver's trajectories [DRK16]. They may find the driver prefers to stop at a bar for a certain time before reaching home which may not a good signal for the driver. Thus, service providers that provide data services to users can compromise user privacy by tracking and analyzing users' location and trajectory in real time. In addition, the eavesdroppers, who can monitor the communication packets through wireless media, can also obtain users' location data from the communication packets. Furthermore, malicious peer nodes can also launch privacy attacks within the network when a mobile node or user decides to transmit data through them.

1.2 Research Challenges

While the research community has made a lot of progress trying to solve the one-shot location privacy problem, little work has been focused on location trajectory privacy and query privacy. There are different types of privacy preservation techniques in the earlier research. Location anonymization is a popular technique. k -anonymity has been widely used to prevent potential malicious application service providers and/or any curious

eavesdroppers from obtaining mobile subscriber locations [GG03, NLZ⁺14]. Researchers attempt to make the user's real location indistinguishable from other users'. However, these approaches with k-anonymity have a strong drawback because they need a trusted third-party who may not always be available, creating a bottleneck within the service system. Another approach is called dummy user generation which can also achieve k-anonymity without the involvement of trusted third-parties. The work [LLG⁺13] proposed a game theoretic dummy user generation scheme for a set of non-cooperative users to achieve k-anonymity when the number of real users is less than k. However, the drawback is that the incentives for other users to participate in the game may not be strong enough, which makes the scheme fail to work.

There are also some other techniques, and Mix Zone is a famous one among them. Mix Zone is like a black box. Multiple users enter into the zone, and at a certain time they must all be in it. When they are all in the zone, they change their pseudonyms to break the links between the entry point and exit point to avoid tracking activities. Researchers proposed optimal Mix-Zone placement schemes to protect user location privacy [FSH09, PL11, PL15]. However, a centralized trusted authority is needed for coordinating the process which may not be available. Cryptography is another popular category. The work in [GKK⁺08] proposed a cryptography protocol to defend against strong attackers that know the system and the temporal relations among the query updates. However, this work is only applied to these specific applications such as nearest neighbor queries. The cryptography methods cannot be directly applied to general location-aware applications.

Recently, the geometric approach is becoming pervasive. This approach is where a mobile device performs geometric computations after obtaining its coordinate, and then

sends the translated information to the application service server. It has no effect on the existing system model and does not require any other user involvement. The work [LSTL13] proposed a smart geometric transforming mechanism. However, the drawbacks for this mechanism are that it only considers a user's one-shot location privacy; furthermore, this mechanism is also vulnerable to a strong adversary model named inference attack, which uses side-information to identify the real user among a set of dummy users through the information collected by the application service provider. Obviously, more research needs to be done to address the research challenges.

1.3 Research Objectives

The aim of this research is to develop privacy-preserving algorithms and mechanisms for preserving location and trajectory privacy of mobile objects with spatial-temporal data communications. This research needs to explore and model the knowledge and capacity of potential attackers. Based on the attackers' knowledge, this research tries to develop distributed real-time and privacy-aware algorithms for defending mobile users and/or vehicles in a robust way. Furthermore, this research intends to improve user experiences by considering the diverse privacy requirements from different users or vehicles.

This dissertation involves the design, development and performance evaluation of lightweight, novel algorithms and mechanisms in mobile and location-aware applications. This dissertation uses metrics like k-anonymity, information entropy, etc., for comparative analysis and performance evaluation. To be specific, we made the investigations in the following topics:

1. Prevent the location trajectory privacy leakage of mobile users in continuous location-based services from third-parties and eavesdroppers

The modeling of attackers' knowledge is the first step for designing privacy-preserving solutions. In the location-based service system, users need to report location data to the service providers, compromising user privacy. The eavesdroppers can also monitor the communication packets to obtain or infer users' location. Previous work mainly focuses on one-shot location privacy without further investigating the trajectory privacy for mobile users. For protecting users' trajectory privacy, our first objective is to design anonymization scheme to let users conceal their real location trajectories so that the attackers have strong difficulties to effectively track the users.

2. Use a query-aware privacy model to prevent query-feature-based inference attack from adversarial servers

In location-aware applications, user query features or interests can leak a user's private information. Previous research has mainly focused on how to prevent location privacy leakage from location data itself in a user query. However, query keywords or query interests can also leak users' location information. Query keywords can have certain relationship with the location information in user queries. We should investigate on a new user privacy problem from a historical query point of view, considering query search features as new tools for the attackers to locate users in a specific region. Our second objective is to analyze how query-feature-based probabilistic inference attack works, and design novel algorithms to countermeasure the query-based privacy attacks.

3. Prevent the privacy leakage of mobile vehicles in the mobile sensing process in a peer-to-peer way

In vehicular networks, mobile vehicular users' sensing privacy is a serious issue. The main research in vehicular networks has been focused on authentication and security for mobile vehicles without user privacy consideration. Mobile vehicles are moving on road networks. In order to reduce the tracking risk from vehicular service providers, we should let mobile vehicular users reduce the number of queries to be sent to online servers. Mobile vehicles can collaborate with other nearby vehicles to achieve this purpose. There is a need to reuse cache data in each vehicle. Our third objective is to design a mechanism for a group of mobile vehicles to share cache data to reduce the number of queries to be sent to the vehicular service providers. This mechanism shall allow vehicular users to query nearby mobile vehicles in a peer-to-peer way for privacy protection purposes.

1.4 Research Contributions

Following the research goals, we first proposed a pseudonym-based anonymity zone generation approach to address the location and trajectory privacy problem in location-based services. Second, we investigated the query privacy problem caused by query search features and designed randomized algorithms to prevent user privacy leakage from query search features. Last but not least, we focused on mobile vehicular user privacy in vehicular networks. We developed a privacy-aware mobile sensing mechanism with peer-to-peer cooperation and evaluated the performance of the proposed approach with intensive software simulations. To be specific, we have made the following contributions:

1. Design a pseudonym-based distributed anonymity zone generation scheme for protecting location and trajectory privacy of location-based service users [GPI15]

- We improve a geometric algorithm by adding a periodic pseudonym changing mechanism. This mechanism allows the mobile user to stay anonymous to the application service providers as well as eavesdroppers. This novel design breaks the linkage among the query updates from mobile users so that the attackers cannot directly track the continuous query behaviors for the targeted users.
- We propose a personalized dummies generation with user-controlled parameters for defending against side information aided inference attack. The privacy requirement can be different for different users. This fact calls for a customized solution for mobile users. With our method, users can adjust the privacy parameter based on their needs. This dummy generation process further enlarges the anonymity region to maximize user privacy.
- With the combination of the geometric method and personalized dummy generation approach, we achieve a new level of trajectory privacy preservation for mobile users. This design breaks the linkage among query updates, and enlarges the anonymity zone for users in each query update.

2. Design a query-aware randomized algorithm for location-aware service users to prevent the query-feature-based probabilistic inference attacks [GBP⁺18]

- We propose a novel attack model named query-feature-based location regional probability distribution inference attack that can be launched by attackers based on the historical query reported location data.

- We model the user query privacy protection problem as a linearly-constrained convex optimization problem.
- We propose randomized approximated searching algorithms to solve the optimization problem, and perform probabilistic k-anonymity with grid obfuscation to stop the probabilistic inference attack.
- We comprehensively evaluate the proposed solution by privacy level, utility analysis, approximation ratio, computation complexity and communication cost.
- We conduct simulation and analysis based on a real word location-based social network dataset, showing the effectiveness and efficiency of the inference attack model and the proposed randomized algorithm.

3. Design a privacy-aware mobile sensing mechanism with peer-to-peer collaboration for privacy protection in vehicular networks [GPI16]

- We develop a peer-to-peer collaborative mechanism for mobile vehicular users to query nearby peers for privacy preserving purposes when they are in certain sensitive contexts (e.g. location such as work site).
- We use data caching technique to let mobile vehicular users store query results from context-aware vehicular applications. Cache data can be reused to reduce the query numbers to be sent by vehicular users since the previous query results can be used to answer the repeated search queries.
- We identify the quality of query result reduction problem in existing sensing privacy solutions especially when users are in sensitive contexts (e.g. locations). When users are in sensitive contexts, the system transformed the query data into obfuscated data so the attackers cannot learn more from the released data. However,

this process adds noise data into the real query data, thus reducing the accuracy of the query results. This process certainly decreases the users' experiences.

- We propose a mechanism for improving user experience by considering both privacy preservation and service quality. We combine a peer-to-peer collaboration process with local and nearby query caching capability to better preserve mobile vehicular users' privacy with service quality consideration.

1.5 Dissertation Outline

The rest of this dissertation is organized as follows. In chapter 2, we survey the related work and provide the literature review. Chapter 3 proposes the pseudonym-based distributed anonymity zone generation scheme. It demonstrates the advantage of our scheme compared with the state-of-the-art result. Chapter 4 presents a query-feature-based attack model and shows how our query-aware randomized algorithms can stop the probabilistic inference attacks. In Chapter 5, we propose a cache-based mobile sensing mechanism for achieving privacy protection in a peer-to-peer way for mobile vehicular users. Finally, we summarize the limitations of our existing work, point out research directions for our future works, and provide the conclusion in Chapter 6.

CHAPTER 2

RELATED WORK

The practical privacy protection techniques are widely needed to address the location and trajectory privacy issues of emerging location-aware applications, such as Location-Based Services (LBS), Mobile Sensor Network (MSN) and Vehicular Networks (VANET). Location-aware applications offer service providers a unique opportunity to make use of users' real-time location data to provide them with valuable services. For example, LBS is the key technology that enables next-generation content consumption such as POI discovery, car navigation, and tourist city guides. However, these valuable data services often come at the cost of compromised user privacy, potentially allowing mobile users' location and trajectory to be tracked. Significant progress has been made in the past several years concerning user privacy. Most of the works have been focused on preserving users' one-shot location privacy in LBS. However, there are only a few solutions that attempted to address the serious trajectory privacy and query-related privacy issues. In MSN and VANET, mobile sensors and mobile vehicles are also generating location and trajectory data for different purposes. The location and trajectory data contain valuable information, and thus become the target of adversaries. The recent development of privacy-preserving works in the two areas draw our attention.

In this chapter, we provide a brief literature review of those related areas, including LBS, MSN and VANET. We also survey the query privacy, an emerging topic in LBS. We categorize and organize each area based on its own features and structures. We present a clean category of the related works in each field.

2.1 Location and Trajectory Privacy Protection in LBS

LBS can provide mobile users with convenient functions and services with respect to their locations. In LBS, mobile users need to report their coordinates, obtained from GPS, Cell-ID or Wi-Fi connections, to the LBS server for accessing the data services. However, in this way, a user's real-time location and/or trajectory is revealed to the LBS service provider which may compromise the user's privacy. Since the LBS systems rely heavily on mobile applications running on smart phones, smart watch or other smart mobile devices, users' trajectory privacy protection should consider the users' mobility pattern. Based on the types of LBS system architectures, we provide the categories of the techniques in the literature. The first LBS architecture has a trusted third-party involved, the second LBS architecture has a semi-trusted third-party involved, and the last one has no trusted third-party involved. The specific categories for each architecture are organized and surveyed.

2.1.1 Trusted Third-Party Based Techniques

Anonymity Based Techniques

Anonymity-based techniques consist of many cloaking region based solutions. The idea is to hide a user's location into a location set, so that the attackers have difficulties to figure out which location is the user's real location. To create a location set, the user's query needs to be combined with other users' queries to send to the service provider. The majority of the pervious works are using k-anonymity. k-anonymity in [Swe02] is originally developed in the area of data publication. The purpose is to keep a data record's identifier, which is

also called quasi-identifier, indistinguishable with other $k-1$ data records. Consequently, the k data records cannot be identified by their identifiers. In the work [GG03], the concept of k -anonymity is firstly applied to the LBS area, used to hide a user's location with other $k-1$ users' when sending a query set to the service provider. The work in [PR08, PR09] focused on the personal privacy requirement by allowing users to predefine their personal profiles for privacy protection purposes.

To achieve k -anonymity property, the work [Ghi09] presented an anonymizer or a cloaking agent, which serves as a trusted third party, to process the user request and generate k -anonymous query requests. When a user sends a query request to the anonymizer, the anonymizer collects other $k-1$ users' query request and aggregates them in a query area to be sent to the server provider. The query requests with the location data form a cloaking region to hide the original user's location. The anonymizer receives the query results from the service provider and sends the specific query results to the original user. By dividing the query region into different hierarchical levels, a data structure with cell and pyramid structure is used to generate the cloaking regions. The solutions in [GKS07, YJHL08, DBS08] tried to generate the cloaking regions using different techniques, including k Nearest Neighbors (kNN) queries and Hilbert area decomposition. [GL08] proposed an adaptive solution so that users can increase or decrease the privacy degree based on k -anonymity, considering both time and space.

In [WXH⁺12], Wang et al. proposed several location-aware algorithms to protect location privacy of mobile users. In this work, a user can adjust the privacy level according to his/her preference along with the movement. Based on the surrounding conditions and users' density, the user's location privacy can be protected by modifying certain parameters

in several ways. Masoumzadeh and Joshi proposed an alternative anonymity concept named LBS (k, T)-anonymity to defend against location attacks in a time window [MJ11]. The main idea is to make sure that the user population should achieve k in a time period T, which is not always available in other main methods based on k-anonymity. In this work, the problem is formed as an optimization problem related to spatiotemporal dimensions. A greedy algorithm is proposed to ensure that all queries have k, the coverage value.

Gong et al. designed a framework that only needs the server to handle the incremental nearest-neighbor queries, and then it guarantees that a user cannot be distinguished from other k-1 users [GSX10]. Instead of sending all the points to the server, this method only sends the center of a k-anonymizing spatial region. The authors also proposed an anonymizer-side kNN algorithm to process the query for the LBS server. Consequently, the LBS server sends POIs back to the anonymizer. These methods only need kNN query processes without more complex computation at the server side. Hwang et al. introduced an r-anonymity concept to blur the user's trajectory by preprocessing some similar trajectories R [HHC12]. This approach provides a time-obfuscation method which can break the list of time issuances of the users' queries. When the users are traveling, the anonymity server applies the time-obfuscated method to break the normal sequence of the queries, and sends them randomly to the LBS server. The approach can prevent the attacker from learning the user's trajectory information with the direction. Moreover, this work considers the s-segment together with k-anonymity to enhance the privacy level when a user sends out a query request. The key aspect is that when the users are moving, the anonymity server applies the time-obfuscated method to break the normal sequence of queries and sends them to the service provider in a random way. The related information

can be cached in the anonymity server and the query results can be sent back to the users. In [SVAC10], Shin et al. tried to divide the whole request trajectory into many shorter trajectories in an optimal way. This work introduced the concept of trajectory k-anonymity, meaning that a user's trajectory should be anonymized by at least k-1 other trajectories.

Pseudonym Based Techniques

Mix zone is one of the famous methods in this category. Liu et al. proposed a traffic-aware mix zone scheme to set multiple mix zones along with the movement of mobile users [LZP⁺12]. By utilizing a graph theory, the problem becomes an optimization problem with certain constraints. The placement of mix zone is also affected by the traffic conditions. By computing the entropy, the best mix zone locations are selected. In [PL11], the authors also proposed a mix-zone approach that considers multiple factors to protect users' location privacy. These factors include the zone's geometric shape, the user population's statistical behavior, and the spatiotemporal resolution of the location's exposure. By devising a suite of construction methods to build a mix zone, this work provides a lower bound on the anonymity level and a higher level for the attack resilience.

However, the assumption that an anonymity server is always available is not a realistic one. The major drawback of the trusted third-party based techniques is that the trusted third-party becomes the bottleneck of the LBS system, as well as the single targeted point, for the attacks.

2.1.2 Semi-Trusted Third-Party Techniques

For semi-trusted third-party techniques, the system assumes that a semi-trusted third-party exists between the service provider and mobile users. Schlegel et al. proposed a user-

defined privacy system to provide the privacy-preserving LBS [SCHW15]. This approach deploys a semi-trusted party in the system. A user sends a query request including the encrypted query and the encrypted identifiers to the semi-trusted party. The semi-trusted party stores all the identifiers, and only forwards the encrypted query to the service provider. When the query result returns, the semi-trusted party stores the whole set of encrypted results, and only returns to the user with a subset of encrypted results whose corresponding identifiers match any one initially sent by the user. After this process, the user can decrypt the received data to obtain the query results. However, the drawback for this type of solution is that the encryption process creates too much overhead and consumes more energy. In addition, users may refuse to encrypt and decrypt data on their devices.

Peng et al. proposed an entity, named Function Generator, to be deployed in the LBS system [PLW17]. This entity can distribute the spatial transformation parameters to the users and the service providers. Both users and the service providers use the parameters to perform the mutual transformation between a real location and a pseudo-location. Without the transformation parameters, the semi-anonymizer has no knowledge about a user's real location. However, this design incurs too much communication overheads to exchange the transformation parameters.

2.1.3 Trusted Third-Party Free Techniques

Pseudonym Based Techniques

Montazeri et al. introduced an information-theoretic notion named perfect location privacy [MHP17]. By changing a user's pseudonym before a certain number of observations made by attackers, the user can preserve perfect location privacy with the assumption that the

user's current location is independent from his or her past locations. Next, the authors considered the user's movement as a Markov chain. The experiment results showed that if the user changed his or her pseudonym before a new number of observations collected by attackers, the user can also preserve his or her perfect location privacy. The new number of observations is a function of the number of edges in the Markov chain model of the LBS user.

Obfuscation Based Techniques

Suzuki et al. tried to generate scatter locations closing to a user based on the former fake locations and the user's real location [SIC⁺10]. According to the real road conditions, this work can adjust the process to obtain a better effect. Ma et al. explored an effective tool named Gaussian Process Regression (GPR) to preserve the trajectory privacy [MLSK11]. The idea is to re-construct the trajectory information of mobile users by exposing selected locations. The GPR is used for inference of the possible direction for the trajectory, and then provides the estimated information by feeding the GPR tool the location samples. By carefully selecting the locations to be exposed, the exposure rate can be controlled within a certain level. This work allows mobile users to send necessary information to LBS servers while controlling the trajectory privacy levels. The drawback is that the service quality heavily depends on the selected exposing locations.

Zhu et al. designed a system to allow mobile users to report fake positions to the LBS server to obtain the query results [ZC11]. In this system, mobile devices generate location proofs, and the location proof server can be used by the devices to verify the trust level of each location proof. A mobile device can also be protected by changing the pseudonyms

statistically. The location privacy model with user concentration evaluates the user privacy level periodically, and makes the decision to accept the request of a location proof. This approach focuses on the combination of location proof and location privacy to improve users' privacy level. In [ACVS11], the authors devised some basic obfuscation operators to transform a location measurement by changing the center or radius. In addition, the basic operators can be used together to be executed in a sequence to effectively protect users' location. Feng et al. argued to generate fake paths along with users' true trajectories [FLZ12]. A noisy location is considered as a real location if it is reachable at the generated time with map information. A message containing a user's true position and the noise data should be sent to the server with two scheduling strategies. The first one is the normal scheduling strategy, and the other one is the disordered scheduling strategy. Both are used to confuse the attackers.

Anonymity Based Techniques

In [CML11], Chow et al. proposed a spatial cloaking algorithm for LBS based on the peer-to-peer environment we mentioned in the previous section. This algorithm has several functions, including an information sharing scheme, a historical location scheme, and a cloaked adjustment scheme. For the first two schemes, the algorithm is divided into a peer search step and a cloaked area building step. For the cloaked area adjustment scheme, the algorithm is divided into a center adjustment step and an area adjustment step. This algorithm satisfies k -anonymity as well as the privacy requirement of the least area specified by users. Jia and Zhang proposed two anonymity algorithms for the LBS users in a mobile peer-to-peer environment to preserve their location privacy [JZ13]. The first

algorithm generates grid areas, allowing the users to judge the areas, and then sends the grid area IDs instead of the real coordinates to the LBS server. The second algorithm allows a proxy peer to generate an anonymized spatial region for the query user. However, this work cannot resist the attack with respect to continuous queries.

Pingley et al. proposed a context-aware privacy scheme that has two components, named location perturbing and anonymous routing components, to eliminate the disclosure of private information [PYZ⁺09]. The perturbing component utilizes the Hilbert curve mapping to produce the perturbed location. The anonymous routing component attempts to reduce users' network identities by relaying the LBS queries to other nodes in an anonymous network. Nussbaum et al. introduced an (i, j) -privacy method, which uses the information of the travel time to distribute the probabilities and to assign less likely travelled locations to the users [NOS12]. Each user's location in one area should be available with at least i locations in another area, and each user's location in the latter area should have at least j locations with the first area. By this way, the anonymized degree of a user becomes higher along with increase of the value i and j .

Liu et al. adopted game theory to achieve k -anonymity for the LBS users [LLG⁺13]. The method is to allow users to generate fake positions according to the privacy level they need, especially when the preferred privacy level is less than k . The work proposed two Bayesian games in both static and time-aware contexts to model users' behaviors, helping them achieve the optimal payoffs. In [ZKMC13], Zhu et al. proposed a two-tier adaptive location privacy-preserving system. The separation tier introduces artificial perturbations into the location data. However, an attacker could perform outlier filtering techniques to

deduce which points have been modified. The conformation tier smoothens these anomalies to reduce the appearance that the location information has been tampered.

Protocol/Encryption Based Techniques

A protocol means that all the participants in a system should follow the same set of rules. In the protocol based techniques for LBS, all the participants in a mobile system should collaborate and follow a protocol to protect the location and/or trajectory information of mobile users. Without the assistance from an anonymizer, an offline phase to map POI locations in the service regions into indexes is required in [GKK⁺08]. During the query process, the users encrypt the queries with redundant information, and then filter the redundancy in the response. A similar idea proposed in [RB12] is to simply generate dummy queries to confuse attackers. In [KSS11], Khoshgozaran et al. introduced an approach to handle range and kNN queries based on the principle of Private Information Retrieval (PIR). This approach places trust on a secure coprocessor used for initiating PIR requests inside the LBS server. The range queries are handled by a sweeping algorithm. The kNN queries are privately evaluated by three algorithms, which are Hierarchical, Progressive and Hilbert-based algorithms. This approach prevents the server from learning users' location information, and even the content of the users' queries. By applying PIR, the user privacy can be guaranteed better than the cloaking and anonymity based techniques.

Buchanan et al. proposed an encryption approach to protect users' location and trajectory privacy based on private equality primitive. By creating a single encrypted table of identities, the users can match their identities with their location privately by checking

the table [BKE13]. The protocol also allows users to privately select the interesting records provided by the server. Ardagna et al. proposed a protocol that allows the local users of a Wi-Fi network to form a group to defeat a global adversary [AJSS13]. They also provide an incentive to stimulate the peers to cooperate in the process. The peers who participate in the protection process will be anonymously rewarded by a micropayment scheme. In addition, this protocol tried to minimize the probability of the fake reward in hybrid scenarios.

Multiple Shares Techniques

The idea of location sharing is to divide the original position information into a set of imprecise location shares, distributed to many different LBS servers. As the result, a single LBS server cannot reveal the accurate location of a user. In [SVAC10], Shin et al. proposed a share generation algorithm to protect user location privacy in non-trusted systems. The algorithm reduces the location predictability so that it is harder for attackers to obtain more accurate locations of the users. With a map-aware position sharing approach, the size of the obfuscation area that is defined by the map information shares can be adapted accordingly.

Xue et al. proposed a sub-trajectory synthesis algorithm to predict the destination of a LBS user [XZZ⁺13]. The algorithm uses a Markov model to compute the posterior probability for an online given query trajectory. Subsequently, the author tried to use a grid graph for abstracting the map and to use Bayer's rule to predict the destination. A user can remove some critical locations in the query trajectory so that the destination of the query trajectory cannot be predicted in a higher level than a given threshold with a probability.

Wernke et al. devised a location sharing method to manage users' private location information. This method divides users' location data into location shares, and then distributes the location shares to different location servers that are used by multiple LBS providers [WDR13]. As a result, the malicious LBS providers can only discover some locations with a lower degree of precision. This approach is powerful as it can defend against the maximum velocity attack, as well as the mapping attack, to a certain degree.

Shokri et al. introduced a method to store a user-side profile, representative of the user's preference on the user client, and only push a subset of that profile on the server side [SPTH09]. The users can contact each other and update their offline profiles by introducing a subset of their peers' profiles. The server that receives the aggregated profiles can report back to the users with a set of recommendations. Moreover, the user client removes redundant or irrelevant recommendations based on its privileged information. This work is further refined in [SPTH11, STP⁺14] with the use of MobiCrowd. MobiCrowd establishes a mobile transparent proxy among nearby users through an ad-hoc network. The LBS queries are checked against nearby devices by this proxy. Only if there are no nearby devices having the request cached, the request should be sent to the LBS provider. By hiding in the crowd, the users can minimize their location leakage to the application service providers. However, the minimal number of nearby peers cannot be always guaranteed when the users query targeted services in the local area.

Geometric Based Techniques

In [LSTL13], Li et al. proposed a geometric approach to solve the location privacy problem for mobile users. The main idea is to send queries with multiple center and radii pairs to

the application service provider instead of sending a user's real location and his/her real interesting scanning radii. By devising a geometric computation algorithm, the query set can cover the user's original interesting area. As a consequence, an attacker can only derive the anonymity zone from these multiple queries. The attacker knows that users are located in the anonymity zones without learning their exact positions. The drawbacks for this method are that it is still vulnerable to trajectory attack for continuous LBS users and it mainly focuses on POIs related applications. Guo et al. extended this method by proposing a pseudonym changing process and a dummy generation mechanism to defend against trajectory privacy attacks [GPI15]. By generating distributed anonymity zones and breaking the linkages among query updates on the fly, users' location and trajectory privacy leakage can be reduced dramatically.

Furthermore, privacy is also an issue in trajectory data publications. Researchers attempted to preserve user privacy when publishing a data set for public research purposes. For example, the work [TZX⁺17] proposed a trajectory data processing scheme to protect trajectory privacy against semantic attacks. It considers k-Anonymity [Swe02], l-Diversity [MKG07], and t-Closeness [LLV07] properties for privacy protection. However, this work only considers the offline data in databases. It does not apply to the online data processing scenario when user data arrives the LBS server on the fly. The reason is that the work needs the whole data set to apply the scheme.

2.2 Location and Trajectory Privacy Preservation in MSN

There are only a few works that studied trajectory privacy issues specifically focusing on MSN. We categorize the existing works based on the main techniques developed in the

algorithms and their studied objects. Based on the mobility mode for the sensor nodes and the base station, we categorize the literature into two main parts, one is location trajectory privacy for mobile sinks, and the other one is location trajectory privacy for mobile nodes. Depending on the specific applications, either the nodes are mobile or the base station/sinks are mobile, or both the sensor node and the base station are mobile. For each assumption or setup, the solutions are different.

2.2.1 Location Trajectory Privacy of Mobile Sinks

Ngai and Rodhe proposed a random data collection scheme to preserve the mobile sink's location privacy in sensor networks [NR09]. The scheme is composed of two stages. The first stage is the data forwarding and storage process. The source node forwards its data randomly to its neighbors, and continues the forwarding process for several hops until the data are stored. The second stage is that the mobile sink moves randomly in the area, requests the data from its neighbors periodically, and filters out the data that have been received. Due to the randomness of the data storage and the movement of the mobile sink, it is very difficult for the attackers to track and attack the mobile sink. The drawbacks are that the delivery latency is larger than the non-random methods, and the message loss rate is also high. To improve these drawbacks, Yao adopted an improved random walk scheme [Yao10]. The first step is called local flooding where the source node broadcasts the packets to all its neighbors. Subsequently, the mobile sink takes a greedy random walk in the sensor region from the start point to the unreached area, and continues moving to the areas that have nodes with less pass-time counters.

Rios et al. proposed two protocols for the probabilistic receiver location privacy protection [RCL15]. The first protocol prevents traffic analysis attacks by probabilistically hiding the real traffic flow. The second protocol is a perturbation protocol that can modify the routing table of the nodes to inject uncertainty to prevent attackers from retrieving the routing information for the nodes. However, the major drawback for those methods is that they cannot supply an end-to-end solution to preserve trajectory privacy for both the sensor nodes and the sinks.

2.2.2 Location Trajectory Privacy of Mobile Sensors

One proposed technique to preserve mobile nodes' trajectory privacy is to reduce the location resolution to achieve a desired level of safety protection [XC09]. The authors consider an ad hoc network formed by a set of sensor nodes deployed in a hostile environment, where the communications among the nodes may be open to the attackers. This location cloaking technique allows the nodes to reveal their location information, yet make it practically infeasible for the attacker to locate them based on such information. To be more specific, each node will recursively compute a cloaking box by broadcasting its current locating region partition P and counting the number of neighbors within P . P is divided into equal halves until the number of nodes within P meets the desired safety level, and then P is set to be the cloaking box. This cloaking box is used as location information for reporting to the service providers. To compute the cloaking box in the presence of the node mobility, three types of messages need to be created to update the cloaking boxes for all the nodes in the corresponding partitions upon movements from the nodes. If a node M moves out of its partition P , it broadcasts a leaving message to notify the nodes inside of P

for them to compute the new safety level of P . This message contains the status of the node M who sends out the message. When M tries to join a new partition P' , it broadcasts a joining message to the nodes inside of P' to compute the new safety level. If the safety level is lower than a certain value, each node inside of P' will take the parent partition P'' of P' and broadcasts a merging message to the nodes inside of P'' . The other nodes who are receiving these messages can take certain actions toward improving the safety level of the current partition.

Another technique is proposed in [JPC⁺12]. The authors consider that the infrastructure of mobile sensor networks is under passive attacks. To hide the trajectory of the target node in an online manner, Basic Trajectory Privacy (BTPriv) and Secondary Trajectory Privacy (STPriv) preservation algorithms were developed. Both of BTPriv and STPriv employ the unique privacy-aware routing process, where each node selects the next-hop node according to the dynamic trajectory distance to hide its trajectory. The privacy-aware routing phase requests that each node should route its data packet through a privacy-aware path instead of the shortest path. In order to select the proper next-hop node that helps the target node hide its location at the time of data transmission, the next-hop node needs to collect limited trajectory information from its neighboring nodes. To avoid privacy invasion of the neighbors' trajectory privacy, the one-time pad virtual name is used to exchange messages. Using the trajectory information from the neighbors, the target node computes the dynamic trajectory distances to its neighbors. The dynamic trajectory distance indicates the irrelevance between two trajectories at a specific time. Finally, the target node selects the neighbor that has the highest probability to mislead invaders as the next hop.

Although both two techniques have strong limitations, such as frequent message exchanges for updating trajectory information for privacy preservation, which consume extra power and create traffic burden for the network, they are good beginnings as early works to address the trajectory privacy issues for MSN in an online manner.

2.3 Location and Trajectory Privacy in P2P, MANET and VANET

Given the highly restricted network resources and special mobile network topologies, the trajectory privacy issue in MSN has been a challenge, and only a few works have been developed. In this subsection, we briefly review some location and trajectory privacy preservation works in Peer-to-Peer (P2P) networks, Mobile Ad hoc Networks (MANET), and VANET in the hope to motivate new techniques.

In P2P networks, Chow et al. designed a peer-to-peer spatial cloaking scheme that can preserve the users' location privacy while using LBS [CML11]. The idea is to allow the mobile user to cooperate with other users to map his/her location into a cloaked region in order to increase the difficulty for the attacker to find the exact location of the user. The scheme allows mobile users to share their location information with each other in the local areas, and also cache the historical location of other peers that can be used for k-anonymity privacy computation. This scheme also avoids the situation that the target user might always be the center of the cloaked areas, thus providing a strong location protection against the attacker. Freudiger et al. provided a framework to evaluate the privacy gains from the mix zones [FMBH10]. In a mix zone, the mobile nodes can change their pseudonyms at the same time. The mix zone is a region that can be used by the mobile nodes in proximity of each other to protect their locations in a collaborative manner. The

attackers only know the location of the zone but not the targeted user's exact location. By utilizing multiple pseudonyms and proper age for each pseudonym, strong privacy protection for mobile nodes can be achieved.

Liang et al. proposed a message authentication scheme to protect user privacy in MANET [LLLS10]. There are two important aspects. First, the service provider can trace the users' identities while keeping them invisible from each other in a group of users. Second, the message receiver should not deliver the authenticity to other parties when a node delivers authenticity of the message to him or her. The user location privacy can be improved significantly due to the fact that the total number of authenticated messages is reduced, and the attackers cannot easily justify which one is true when they receive the messages. In [DS10], a protocol has been proposed to protect the source and destination privacy against a powerful global adversary in MANET. The first stage of the protocol is an initialization process, where all nodes will be initialized in a broadcast mode at a certain rate. The second stage is the route extrapolation process using a Dijkstra algorithm. Both the sender and the receiver use the algorithm to select nodes until the two selected nodes are facing each other. The third stage is to generate dummy traffic at a rate that should be the same as the source node to hide the real packet transmission pattern. This protocol can provide privacy for both the sender and receiver in a flexible level. However, there will be a high delivery latency to find a pair of nodes facing each other.

Hao et al. defined a uniform framework for privacy protection that combines perturbation-based privacy preservation and malicious node revocation [HLD10]. In this framework, a node protects its location privacy by controlling the distribution of its location information. The location servers cannot link a node's position to its real identity.

The second mechanism is to defend against the inner malicious nodes which may reveal the location of the sender. It defeats the malicious node by verifying the group signature for data transmission. A malicious node can be revoked according to the bad reputation level accumulated by its malicious behaviors. This framework can handle inner malicious nodes and achieve node-controlled privacy. However, the heavy traffic load generated by the excessive controlling activities reduces the node's lifetime dramatically.

For VANET, Yu et al. proposed an accumulative pseudonym exchanging mechanism for vehicular location privacy [YKH⁺16]. It devised an entropy-optimal negotiation procedure for the pseudonym exchanges among the moving vehicles. In the procedure, each vehicle evaluates the benefit and risk before participating in the pseudonym exchange process. The risk and benefit are quantified by the predefined pseudonym entropy. Moreover, mobile vehicles choose to have consecutively increased pseudonym entropy based on their calculations and evaluations. However, this method has more overhead to continuously calculate the entropy in each pseudonym exchange process. While the privacy protection method allows each user to become a new node to its neighbors after a pseudonym change, the reputation management requires a user to maintain a long term identity. Balancing the reputation and the privacy protection for the same vehicle remains a challenge.

Li et al. developed a localized model to enhance both privacy and reputation management for mobile vehicles [LC14]. Each node uses a neighbor-certified reputation label to demonstrate its reputation history. In face of the network topology's change caused by the changed pseudonym and the node's mobility, the reputation information of any node is maintained by its neighbors and itself. This work also allows the honest nodes to manifest

the same reputation in a privacy-preserving way. However, even if this work has a strong notation on maintaining a vehicle's reputation, it is still vulnerable to the external attackers who have more information about the network topologies.

2.4 Query Privacy Preservation in LBS

Query privacy is an emerging issue in the LBS system. Most past works have been focused on location and trajectory privacy without query privacy in mind. There are few works in this area. Query privacy mainly comes from the fact that users' query interests can reveal some personal and even sensitive information. For example, if a user always searches for "casino" in his or her query keyword, the LBS servers can learn that the user may have a bad habit. We summarize the related work in query privacy in this section.

The work [WLY⁺17] proposed a k-anonymity approach with cloaking region-based technique to preserve query privacy in LBS. When query interests have similar prior probability, a circle segment is used to provide effective query privacy. Two privacy metrics, named expected entropy and expected max-min ratio, are adopted by this work to evaluate the approach. However, this work lacks a study about how to conduct a probability generation process that can be performed by attackers. The measure of similarities and probabilities for two given query interests is not illustrated in the work. Another problem of this work is that it has no consideration of the location privacy, and only directly keeps the user's original location in the query.

This work [YPBV16] studied the location and query privacy problems in approximate kNN queries for LBS. The work proposed a public-key cryptosystem for protecting location and query privacy in approximate kNN queries. Without revealing to the LBS

service provider what type of PoIs that are retrieved, the user query privacy is preserved. The proposed PIR protocol is used to achieve the privacy preservation in the computation stage when the server provider is processing the kNN queries. The major drawback is the query overhead due to the heavy computation for the PIR protocol.

A distributed sensing framework is proposed for real-time queries of popular paths with user privacy protection in the work [DNZG17]. The work uses the MinHash data structure to record the query data of mobile entities. The MinHash signature provides privacy protection for the users with a differential privacy model. The major drawback of this work is that the user privacy is heavily dependent on where the MinHash is implemented. The entity that is carrying the MinHash becomes the bottleneck and single point of failure of the LBS system.

[PKYB14] presented a solution for location-based query content protection. The solution is divided into two steps. The first step is based on oblivious transfer stage, and the second approach is based on a PIR protocol. In the first stage, based on the oblivious transfer function, the user privately determines his or her location within an open and public cloaking area. The user ID and related symmetric key are contained in the data for his or her private location or area. In the second stage, the user tries to extract the data block in the private grid using a communication-efficient PIR protocol. The data block can be decrypted by the key generated in the first stage. The PIR protocol creates a serious problem because of the generated overhead in the primality test in the two stages. Another drawback is that the key management may create more security problems for the mobile users.

In order to prevent query privacy leakage from a user's profile attributes, [DACA16] proposed a collaborative approach considering users' profile attributes. This approach provides p -sensitive and k anonymity for a mobile user by using Ciphertext-Policy Attribute-Based Encryption (CP-ABE). In particular, this approach decides: 1) which attribute is identical to other attributes within the anonymity set; and 2) which attribute is sensitive in order to find p different attributes to protect it. The CP-ABE allows the user to apply attributed-based rules or policies on his or her queries, so that only the actions under the rules or policies can be performed on the user query. However, attribute-based encryption methods have higher overhead and lower execution speed, so it is not practical to be implemented in the real world.

[ABCP13] introduced the differential privacy concept to the location-based systems. Even if this work only consider a user's location privacy, it brings some hints for the possible query privacy protection. In this work, it proposed a perturbation technique for achieving geo-indistinguishability. The perturbation technique can add controlled noise to the user's location data to generate a new location, making the two locations indistinguishable. The noise can be drawn from a planar Laplace distribution. The differential privacy is a strong notion for privacy protection, because it provides a theoretical and statistical bound for hiding a data record in a dataset. This work adopts the differential privacy in LBS for hiding a user location in a location distribution set. However, this work have no consideration about the query feature's impact on the user's location privacy.

[EG16] provided a probabilistic model for obfuscation generation with utility constraints for differential privacy. The work also introduced the notion of (D, ϵ) -location

privacy which is an adaption of the standard ϵ -differential privacy for LBS. This notion is to prevent the attacker from observing the output of the different private obfuscation function, and then distinguishing the user's real location from other locations within distance D with parameter ϵ . This work has the merit that users have no need to worry about the background knowledge that the attackers may gather in unexpected ways. Thus, the attackers have strong difficulties to figure out the user's real location and the generated location. However, the users may still lose their privacy when they use the obfuscation mechanism frequently. New methods need to be developed for handling the cumulative privacy loss.

[HTXZ18] identified the problem that user privacy can still be compromised based on the previous differential privacy works, especially when the number of users' queries increases. The privacy consumption increases when a user sends many queries to a LBS server even if he or she is using differential privacy perturbation mechanisms. This work proposed an improved geo-indistinguishable approach to solve the problem. It divided the target space into cells for the LBS users. The users can have zero privacy consumption if they are not close to the cell borderlines. This work also proposed an improvement to help users reduce the privacy consumption when they are close to the cell borderlines. The idea is to adjust the cell size with respect to the user movement. When a user moves close to the borderlines, the algorithm can dynamically change the cell size to reduce the users' privacy consumption. However, this work still lacks the consideration of query privacy, but it provides insights about how to dynamically apply differential privacy in a timer manner in the LBS domain.

The work [CEP17] investigated the privacy and utility tradeoff problem. When a user's privacy is strongly protected by a privacy-preserving method, his or her service utility may not be guaranteed as optimal. This work argues that a Bayesian approach is better for solving the utility optimization problem with privacy constraints. That means, given a prior information, the reported location can be remapped to the optimal location, based on the loss function and the given prior information. This work found that users' utilities can be improved considerably when they use the best remapping function, with respect to the given prior information. For the prior information, this work proposed an approach to construct a global prior parameter for remapping purposes and a user-specific prior parameter for measuring the utilities. This approach shows improvements comparing with other standard planar Laplace mechanism. However, this work only considered the utility and privacy tradeoff based on the pure location-based privacy protection mechanisms.

[AB18] introduced that the location pattern privacy protection is an important issue beyond pure location privacy problems. This work argues that location pattern privacy protection should provide safer sanitized trajectories shared with the service providers. Therefore, the work presented a framework where users' sensitive movement patterns can be defined and sanitized in online and offline manners. Moreover, an efficient dynamic programming approach is adopted to find and prevent sensitive pattern disclosure. This work also explored the spatial-temporal relationship for the locations in a user's query history, and provided an online location pattern privacy algorithm to countermeasure the pattern privacy attacks. In addition, based on the privacy constraint, this work tried to find the minimum time delay to speed up the protection process. However, this work has some limitations. First, it assumes that the attackers may not know the specific sensitive patterns

that are drawn by the users. The attacker may have some background knowledge of the sensitive patterns even if the number is small. Second, this work did not consider the strong differential privacy property that can apply the query pattern privacy protection. In addition, the query pattern privacy protection in the work has no consideration about the query feature pattern protection, which may have strong correlation with the location pattern. Query feature pattern can be used by attackers to learn the users' preferences and sensitive information in their behavioral query data. Thus, more research needs to be done in this field.

2.5 Summary

In this chapter, we have reviewed the existing main stream techniques to protect location and trajectory privacy in the areas of LBS, MSN and VANET, etc. We also examined the special query privacy related work for LBS. These previous works are of high inspiration and have motivated our works. In the following chapters, we will present our research outcomes so far and point out the future directions of this research.

CHAPTER 3

PSEUDONYM-BASED ANONYMITY ZONE GENERATION FOR LBS

3.1 Introduction

The popularity of location-aware mobile devices and the advances of wireless networking have seriously pushed LBS into the market. However, moving users need to report their coordinates to an application service provider to utilize interesting services that may compromise their privacy. In this chapter, we propose a novel, personalized design to grant control to users themselves by considering their own privacy requirements in LBS. Based on a geometric transformation algorithm, we let users periodically change their identities using pseudonyms to break the trajectory linkage for the same user. Furthermore, we combine anonymity zone generation with personalized k-anonymity to achieve strong privacy preservation against common malicious service providers, eavesdroppers, as well as side information aided attack. To our best knowledge, this work is the first work to combine the geometric method with personalized k-anonymity for a better effect. In summary, our contributions are listed as follows:

- We improve a geometric transformation algorithm by adding a periodic pseudonym changing mechanism for mobile users.
- We propose personalized dummies generation with user-controlled parameters for defending against normal attacks and side information aided attacks.
- With the combination of the geometric method and the personalized approach, we achieve a new level of trajectory privacy preservation for mobile users.
- Our solution shows advantages compared with trajectory k-anonymity, the geometric transformation and the basic location obfuscation.

3.2 System Model

3.2.1 System Architecture

The system consists of a user with a mobile device, the positioning/communication network, and an application service provider shown in Figure 3.1. The user can obtain the coordinate from GPS, Cell-ID, WiFi or other available position technologies. The user sends the query request through the communication network, and it finally reaches the application service provider. The application service provider processes the query and returns the desired result to the requesting user. The mobile user conducts a random walk mobility pattern.

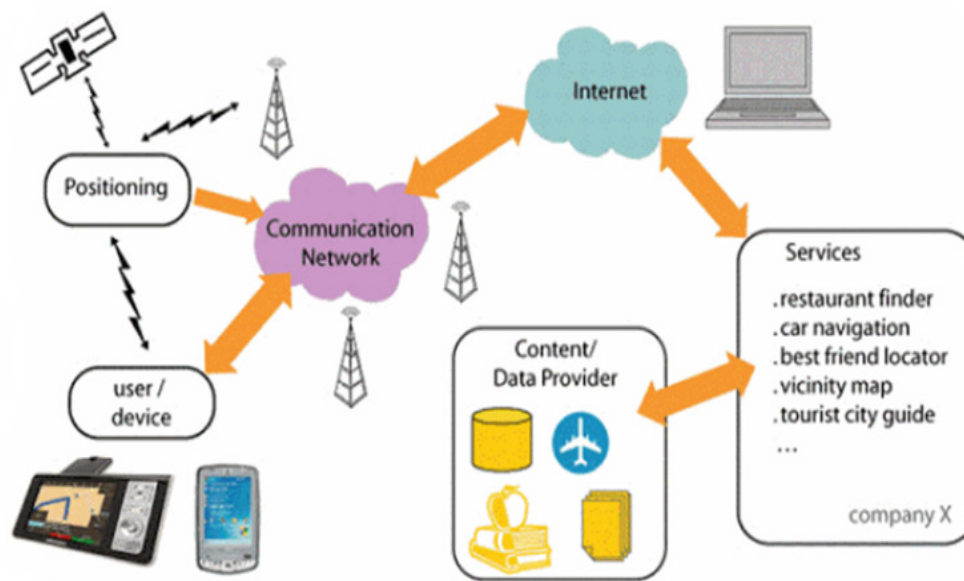


Figure 3.1: LBS system architecture

3.2.2 Attack Model

We make the assumption that the application service provider itself can be the attacker, who would like to collect the users' context information like location, identity, etc., for mobile advertisement purposes and/or for revenue by selling the context information to

third parties. The application service provider can also be compromised by the attackers. We also consider another type of attacker who can eavesdrop on a user's message in the communication network. In addition, we consider an attack where an attacker can use side information to identify the real user in a public place, while the adversary is static and will not follow the targeted user.

3.3 Pseudonym-based Mechanism Design

Our design improves upon previous research [LSTL13] by designing a scheme for generating anonymity zones with continuous changing identities for a moving object. We also complement this by proposing personalized dummies' generation to defend against a strong inference attack. For mobile applications retrieving the POIs (Point of Interests), we adopt the concept ROI (Region of Interests) because the users can rank the results according to their own criteria, such as price and rating. We first let users set radii for the interesting area and k value according to their own privacy requirements. The mobile device performs an anonymity zone generation algorithm and sends the generated centers and radii with dummies to the application service provider. The application service provider returns the results within that area to the users. Finally, the users choose to filter the results according to their own criteria.

3.3.1 The Basic Geometric Transformation Process

When a mobile user tries to query the interesting service, his/her mobile device generates a first-level query, such as $\{\mathbf{u}_{id}, [(x, y), R], T\}$, where \mathbf{u}_{id} is the user's identity, (x, y) is

the user's current coordinate obtained by the available position techniques on the device, R is the interesting radii for the targeted region, and T stands for the type of interesting application like shopping mall, restaurants, etc. $[(x, y), R]$ is named the term ROI. After this setting, the basic geometric transformation engine begins to transform the original ROI into a concealing space to fully cover the ROI. The algorithm for the transformation is: given the radii R and center (x, y) , it divides the cycle into four sectors (for example, set the sector number to 4). Each sector will be covered by a new cycle with a randomly chosen center within the sector and a calculated radii. The new cycles are created one by one until all of the sectors are covered. All of the center and radii pairs, such as $[(x_1, y_1), r_1]$, $[(x_2, y_2), r_2]$, $[(x_3, y_3), r_3]$, $[(x_4, y_4), r_4]$, are generated as a second-level query. In order to prevent the adversary from finding out the first cycle and performing the algorithm (note: the geometric algorithm is available to the attackers), the engine rotates the generated cycles by a random angle $\sigma \in (0^\circ, 360^\circ)$. Thus, the new third-level query is $\{u_{id}, [(x_1', y_1'), r_1'], [(x_2', y_2'), r_2'], [(x_3', y_3'), r_3'], [(x_4', y_4'), r_4'], T\}$. Usually, this third level query can be sent directly to the application service provider. After the service provider receives the query, it will return the results with the coordinates for all the POIs. The mobile users can set a Result Filter on the device to filter the returned results according to their own criteria, such as the review rating of restaurants or the price of food. However, we have additional processing steps on the third-level query considering a combined strong adversary model.

3.3.2 The Dynamic Pseudonyms Changing Process

From the basic geometric transformation, we know the third-level query is $\{u_{id}, [(x_1', y_1'), r_1'], [(x_2', y_2'), r_2'], [(x_3', y_3'), r_3'], [(x_4', y_4'), r_4'], T\}$. At this time, we will take care of the user identity which is one of the most sensitive pieces of information. We deploy a dynamic pseudonyms changing mechanism for protecting a user's identity. The mechanism generates a series of different identities $\{u_{id1}, u_{id2}, u_{id3}, u_{id4}, \dots, u_{idk}\}$. Every time a user has a query request for a target service, the geometric engine adopts one of the generated ids and integrates it with other parts of the query. Considering the adversary model (eavesdroppers, application service provider and/or the attacker who compromised it), with the continuous request updates, this mechanism prevents the attackers from linking snapshot location updates for the same user and reduces the trajectory leakage risks. From the series of user generated queries like $\{\{u_{id1}, \dots, T_a\}, \{u_{id2}, \dots, T_b\}, \dots, \{u_{idk}, \dots, T_x\}\}$, the adversaries will have difficulties in identifying the direct relations among them. Here, we consider multiple users will have access to the LBS system, and the timestamp will not be easily linked for the same user. We also assume that different types of LBS providers will not collaborate together to perform the attack.

3.3.3. The Personalized Dummies Generation Process

In this chapter, we introduce a new strong attacker model named inference attacker, that can use side information to identify the real user in a public place, which is beyond the normal adversary model. Based on the side information collected by the strong adversary, together with the information collected by the normal adversary, the attacker will conclude

that a user, who is in the approximate same position reported by the pseudonym, is the same user in the real world found by the side information reporter controlled by the attacker. However, if there are multiple queries with different pseudonyms, the adversary will find it difficult to match the real users with their pseudonyms in use.

We design a personalized dummy generation approach to prevent this inference attack. A user can set the parameter k according to his/her own privacy requirement in a real-time scenario. The dummy generation mechanism complements the basic geometric transformation with $k - 1$ generated dummies. It will choose the random coordinates other than the user's real coordinate and utilize the result of the initial geometric computation. The query set is demonstrated as:

$$\left\{ \begin{array}{l} q_1 : \{u_{id1}, \dots\} \\ q_2 : \{u_{id2}, \dots\} \\ \dots \\ q_k : \{u_{idk}, \dots\} \end{array} \right\}$$

The $k - 1$ generated dummy queries together with the real query consist of the whole query set, which will be sent to the application service provider. The probability for the inference adversary to link the observed user with the user id in the queries received by the application service provider will be greatly reduced.

3.3.4 Results Filter Deployed on User Devices

Since we send out a query set to the application service provider, a certain amount of the returned POIs will not fall into the user's real interesting region. Therefore, a Results Filter (RF) is necessary to accomplish this filtering task.

Pseudonym-based Mechanism Design (Pseudo Code)

```
//Input: User request with (radii:  $R$ , type of app:  $T$ ) for services with
personalized parameter  $k$ .
//Output: The POIs of the interesting app in the requested region (ROI).
// Let  $r$  be the radii,  $t$  be the type,  $(x, y)$  be the coordinate. The mechanism
engine divides the circle with center  $(x, y)$  and radii  $r$  into multiple circles (set
4, for example):
for  $i < 1$  to 4 do
Divide  $\{(x, y), r\}$  into  $[(x_1, y_1), r_1], [(x_2, y_2), r_2], [(x_3, y_3), r_3], [(x_4, y_4), r_4]$ 
Rotate  $\sigma \in (0^\circ, 360^\circ)$  angel
//Pseudonyms will be added to the query for the generated center and radii
pairs with type  $T$ 
 $q := \{u_{id}, [(x_1', y_1'), r_1'], [(x_2', y_2'), r_2'], [(x_3', y_3'), r_3'], [(x_4', y_4'), r_4'], T\}$ 
//Dummies generation with personalized  $k$  setting for achieving  $k$ -anonymity
for  $j < 1$  to  $k-1$  do
Randomly choose a coordinate  $(x_k, y_k)$  with  $u_k$ 
Randomly match the result of the basic dividing algorithm
//Get the query set  $S$  and send to the service provider which will retrieve POIs.
for  $m < 1$  to  $n(n = 4k)$  do
Retrieve POIs in  $\{(x_m, y_m), r_n\}$ 
//Results Filter filters the POIs falling into the real cycle.
 $D := \sqrt{(x_m - x)^2 + (y_m - y)^2}$ 
//User ranks the POIs according to the preference like price  $P$ , or distance  $D$ ,
etc.
```

Algorithm PAZ: Pseudonym-based anonymity zone generation

The RF will perform a simple distance computation to make the judgement if the POIs are falling into the right region. We know the user's coordinate is (x, y) and the interesting radii is R . We assume that we receive the coordinate of the returned POI is (x_m, y_m) . Thus, the user device can use the following calculation to get the distance.

$$D = \sqrt{(x_m - x)^2 + (y_m - y)^2}$$

According to the radii R , which is set by the user, we compare the distance D with the radii R resulting in:

- 1) If $D > R$, the RF will discard that POI and will not store it on the user's mobile device, so that the storage space will be reduced.
- 2) If $D \leq R$, the RF will save that POI and store it on the user's device in order for some further process.

Based on the set of saved POIs, the second function of the RF is to rank the POIs based on its preference, such as Price P and distance D . The Algorithm is shown above as Algorithm PAZ.

3.4 Analysis and Simulation

In this section, we analyze the privacy level and communication cost for the proposed scheme. We also conduct simulations in Matlab R2013a to evaluate our solution. The users can adjust the anonymity parameter to achieve k-anonymity to satisfy their desired privacy level. Our work PAZ compares with the state-of-the-art CD scheme in [LSTL13] for trajectory privacy protection. In addition, we compare our solution with trajectory k-anonymity [ZSXP15, FLZ12] and the baseline location obfuscation [ACVS11]. We show

that even if there is a slight increase of communication cost in our algorithm, a user's trajectory privacy can be strongly protected comparing with the trajectory k-anonymity scheme, while the CD scheme and the baseline location obfuscation have completely no privacy guarantee on the user's trajectory preservation.

3.4.1 Simulation Setup

We conduct the simulation using the Matlab R2013a version from the Mathworks Software. We set the network area 2 mile \times 2 mile, similar to the size of downtown Miami. There are 2000 POIs randomly distributed in the area. We consider that the users conduct random walks in this area, and send query requests periodically with an interval of about 5~10 minutes. In order to simplify the process, we set the $t = 10$, $t' = 50$ (Normally, the returned POI contains much more information than the original center/radii pair.)

3.4.2 Privacy Level

The basic geometric algorithm generates several cycles to collaboratively cover a user's interesting region. The attacker who obtained these center/radii pairs can find the overlap of the cycles and know that the user is located in the overlap region, which we call the anonymity zone. The larger the anonymity zone, the higher the privacy level. With only basic geometric computation, the privacy level does not greatly change. Our proposed solution gives two levels of privacy preservation:

- 1) One Snapshot Privacy Level c : for the snapshot privacy, we propose a distributed anonymity zone generation that a user can set k to personalize his/her privacy requirement. The anonymity zones are distributed, and the privacy level should be

$k \times \tau = k\tau$, where τ is the privacy level of the basic single anonymity zone generation.

- 2) Trajectory Privacy Level C: for the trajectory privacy, we propose the pseudonym changing process for continuous query updates. The privacy level is determined by the ability of the adversary to infer the real trace for the same user. The inference attacker can only get sparse side information to correlate the real user's identity with the pseudonyms or dummies. The probability of a user's overall trajectory

leakage should be:
$$l = \frac{1}{k_1^\tau} \times \frac{1}{k_2^\tau} \times \frac{1}{k_3^\tau} \dots = \frac{1}{\tau^n \times k_1 \times k_2 \times k_3 \dots}$$
 where n is the

number of query updates. The lower trajectory leakage, the higher privacy level. Similar to work [GMS13], we use information entropy as the privacy metric to calculate the trajectory privacy level C, considering the trajectory leakage probability l .

When the distributed anonymity zone generation and periodically changing pseudonyms combine, the overall trajectory privacy level becomes strong enough against the attackers. The reason is that the attackers can neither make simple connection between two different zones, nor can they easily link two different pseudonyms.

3.4.3 Communication Cost

Communication cost is another important metric for performance evaluation. We see that for our proposed solution, we need to report a query set including all generated cycles (each cycle is described by a center/radii pair) to the application service provider. Hence, the uplink communication cost is determined by N , which is the number of center/radii pair.

The downlink communication cost is determined by M , which is the number of returned POIs.

The users can define the k parameter and, for comparison convenience, we set the dividing function $n = 4$ (we will explore the case that n changes in the future). Thus, the number of the generated cycles $N = 4k$. We assume each center/radii pair contains t byte information. Totally, the uplink contains $4k \times t = 4kt$ byte information. We assume each of POI contains t' byte information. Totally, the downlink contains Mt' byte information. Considering both the uplink and downlink, the communication overhead is $(4kt + Mt')$ for our proposed solution for one query process. The overall communication cost P should be $p \times (4kt + Mt')$, where p is the number of query updates that can be calculated by experience time T dividing time interval s . Thus, $P = T \times (4kt + Mt') / s$.

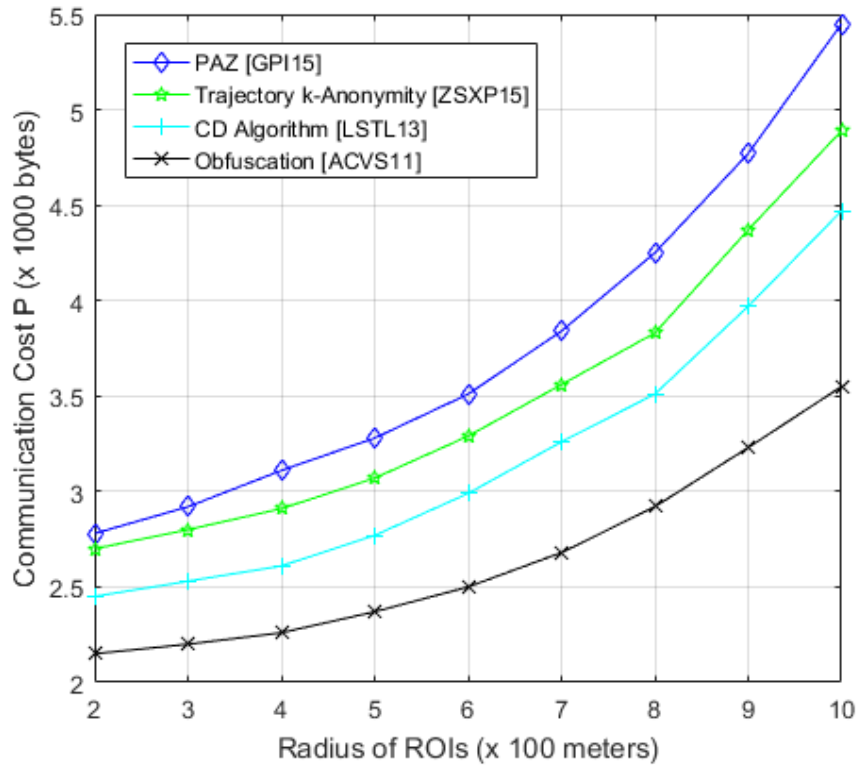


Figure 3.2: Communication cost when radius of ROIs $R = 2, 3, 4, 5, 6, 7, 8, 9$ and 10 ($\times 100$ meters)

First, we evaluate the relationship between the communication cost P and the radius of the ROIs (which is set by the users) with a fixed anonymity parameter. From Figure 3.2, we see that the communication cost increases quickly along with the positive change of the radius for all methods. When the radius becomes larger there are more POIs in the ROI, because the region covered by the ROI is becoming large at a quick rate. On average, we observe that the communication cost of PAZ is slightly higher than CD, about 0.8 KB. Furthermore, the PAZ has about 0.5 KB overhead more than trajectory k -anonymity, and about 1.45 KB more than the basic obfuscation method. This is because we sacrifice some communication cost for achieving the trajectory level privacy protection. Specifically, the CD scheme has the communication overhead due to the fact that the query area formed by the circles covers more than the original query area. Compared with the baseline, the CD has extra 0.65 KB communication overhead. The trajectory k -anonymity needs to form a tube that covers the k trajectories. The query area in the cube at a specific time is larger than CD, so that the communication overhead is also higher than CD. However, the trajectory k -anonymity does not provide anonymity zones. Our PAZ has a slightly higher overhead than k -anonymity because of the distributed anonymity zone generation. The combined query area of the distributed query areas in PAZ is larger than the query area of trajectory k -anonymity. In other words, we sacrifice the extra cost 0.5 KB to achieve trajectory k -anonymity as well as distributed anonymity zones for privacy enhancement. In addition, we can see that the communication costs for all methods increase faster when the radius of the ROIs increments by 100 meters at a time. This is because the area covered by the radius increased with 100 meters is becoming larger at a quick rate. Thus, the communication costs increase at a quick rate.

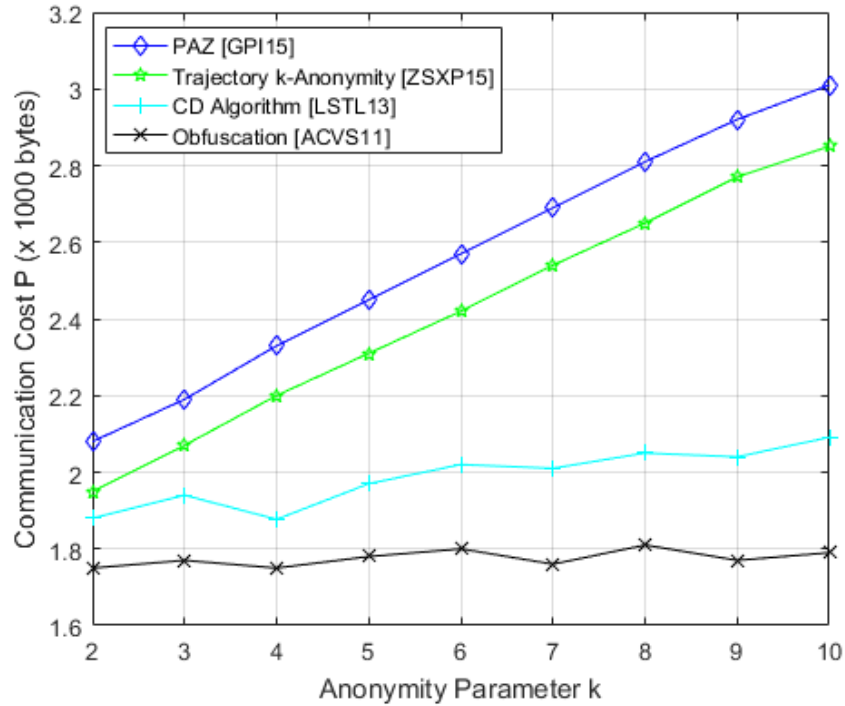


Figure 3.3: Communication cost when anonymity parameter $k = 2, 3, 4, 5, 6, 7, 8, 9$ and 10

Second, we evaluate the communication cost P based on the privacy parameter k with a fixed radius of the ROIs. We allow the user to set k to be $2, 3, 4, 5, 6, 7, 8, 9$ and 10 . We try to see how the communication cost change along with the increase of privacy parameters. From Figure 3.3, we see that our PAZ has a little higher communication costs but similar to the trajectory k -anonymity, when the anonymity parameter increases. The reason is because both methods are based on k -anonymity. Thus, we can see both of them are near linear to the anonymity parameter k . When k is smaller, the communication cost is higher for PAZ because of the non-overlapped distributed anonymity area. However, the communication costs are getting close to one another for two algorithms when the privacy requirement becomes stronger. On the other hand, both of the CD scheme and the baseline location obfuscation method are not sensitive to the anonymity parameter k , because they

only focus on one-shot method without the anonymity parameter involved. Thus, both of them show a random pattern with respect to the anonymity parameter k . However, the CD still has extra communication cost due to its larger query area than the original query area covered by the baseline. The users need to pay for communication consumption if they want to deeply protect their privacy using PAZ to enlarge their anonymity zones. The tradeoff between the privacy benefit and the payoff of the communications is an interesting topic for further investigation.

3.4.4 Performance Evaluation

For privacy-related performance evaluation, we first evaluate the relationship between the anonymity parameter k and the trajectory privacy level C . We also evaluate the relationship between the radius of ROIs R and the trajectory privacy level C . We allow the user to set k and R all to be 2, 3, 4, 5, 6, 7, 8, 9 and 10. We can see that k has 1 as a unit and R has 100 meters as a unit. In order to simplify the process, we assume that the anonymity parameter or radius will not change for the same user in one run. At the same time, we evaluate the privacy level of CD based on its nature performance.

First, we compare our work PAZ with trajectory k -anonymity, CD and location obfuscation for privacy level with anonymity parameter k . From Figure 3.4, we see that when k increases, the privacy levels based on the PAZ and trajectory k -anonymity rise at the same time, the privacy levels of CD and the baseline are close to the bottom. We also observe that along with the increase of k value, the increasing speed of privacy level based on PAZ becomes slow and will approach only a certain level. The trend of privacy level based on trajectory k -anonymity is similar. Our PAZ shows obvious advantage than

trajectory k-anonymity when k increases. The trajectory k-anonymity still shows some merit, since its anonymity level is becoming higher in a sub-linear way when the anonymity parameter k increases linearly. We also observe that the CD algorithm is not stable with a low privacy level. Thus, we find that PAZ has improved the user privacy level significantly compared to the trajectory k-anonymity, the CD scheme, and the baseline whose privacy level almost approaches zero with a non-stable status. The reason behind the observation is that the CD and the basic scheme have no consideration of continuous query requests, so that the trajectory level privacy is very weak and changes randomly. In addition, the speed of privacy level based on PAZ is increasing at a slower pace, because when k becomes larger, the anonymity zones will overlap. Therefore, the privacy level will barely increase and will remain close to a certain level.

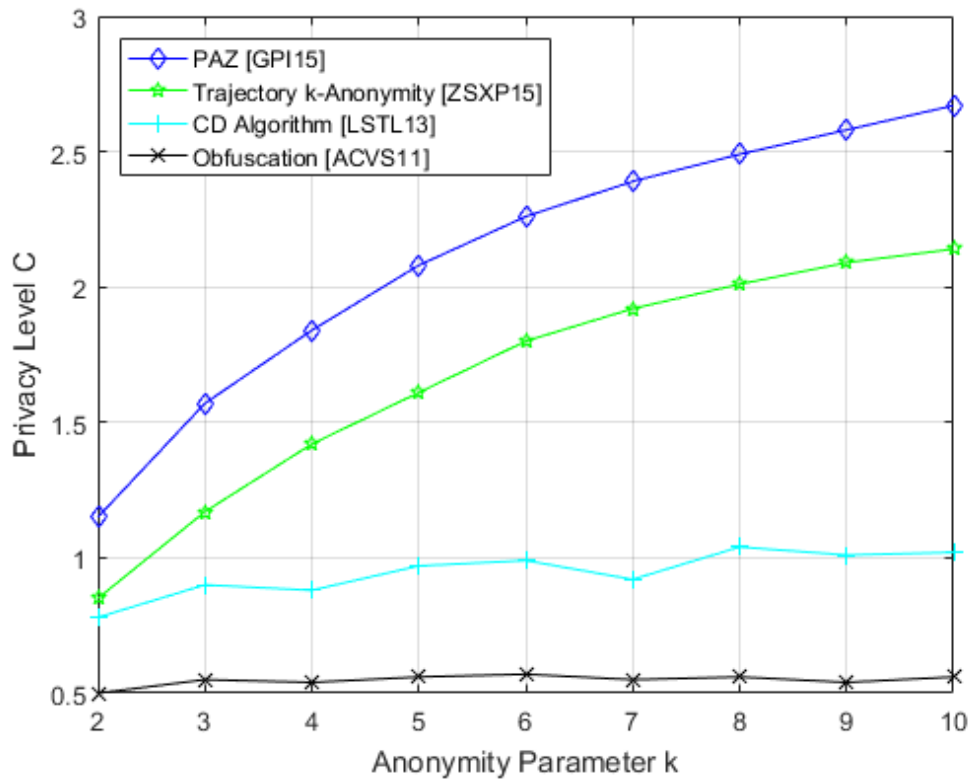


Figure 3.4: Privacy level when anonymity parameter k = 2, 3, 4, 5, 6, 7, 8, 9 and 10

Second, we evaluate the privacy level based on the radius of the ROIs with fixed anonymity parameter k . From Figure 3.5, we observe that the privacy level is becoming higher for our PAZ algorithm along with the increase of the radius of ROIs. This trend is similar for other methods. Our PAZ algorithm shows obvious advantage than the other three methods. The trajectory k -anonymity improves the privacy level and perform better than CD, while CD performs more satisfactory than the baseline. On the contrary, the CD and the baseline have a lower privacy level, and the privacy level is still very low even if the radius increases to the maximum value. The anonymity zone is become larger when the radius of ROIs increases. Thus, the PAZ algorithm generates larger anonymity zone than the CD algorithm. The distributed anonymity zones are larger than a single anonymity

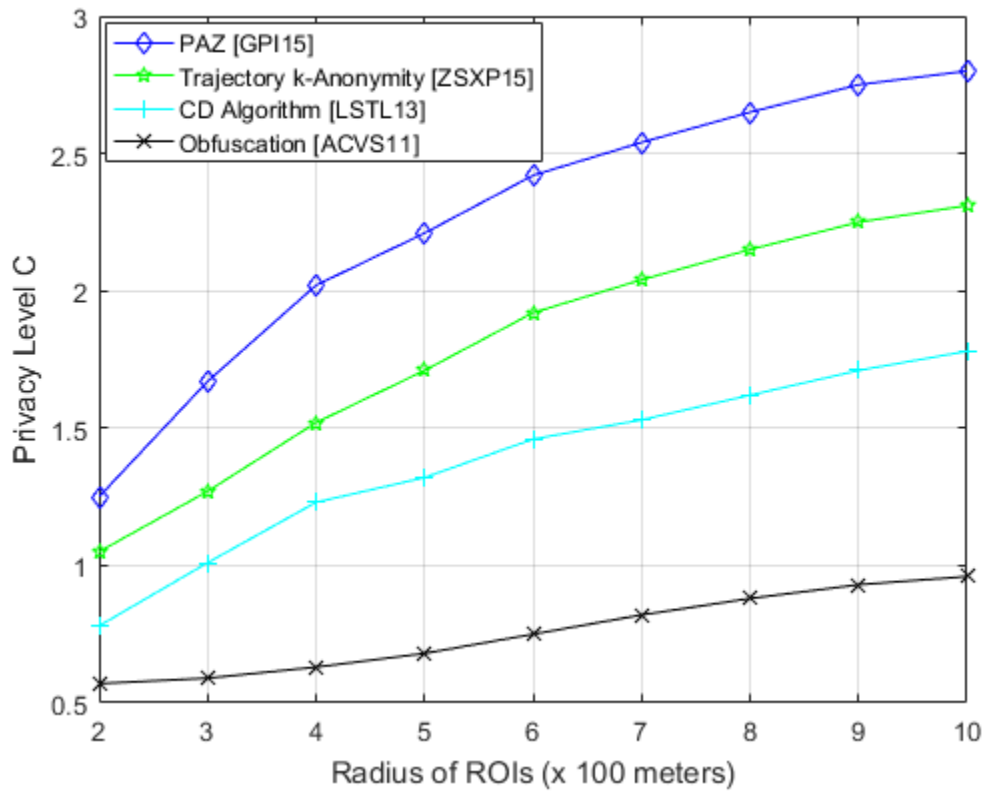


Figure 3.5: Privacy level when radius of ROIs $R = 2, 3, 4, 5, 6, 7, 8, 9$ and 10 ($\times 100$ meters)

zone, when the radius of the ROIs increased by 100 meters. Our PAZ algorithm outperforms the trajectory k-anonymity method due to PAZ’s distributed anonymity zones with k-anonymity property providing a better anonymity level than the single k-anonymity property. In addition, we observe that the CD algorithm increases its privacy level when the area of the ROIs becomes larger, for the related anonymity zone is also created in a larger size. The baseline increases its privacy level in some degree due to the fact that the obfuscation transformation can be more efficient when the query area becomes larger. Overall, our PAZ algorithm performs better than the trajectory k-anonymity, the CD algorithm and the baseline, even if the privacy parameter is fixed and only the radius of ROIs is changed.

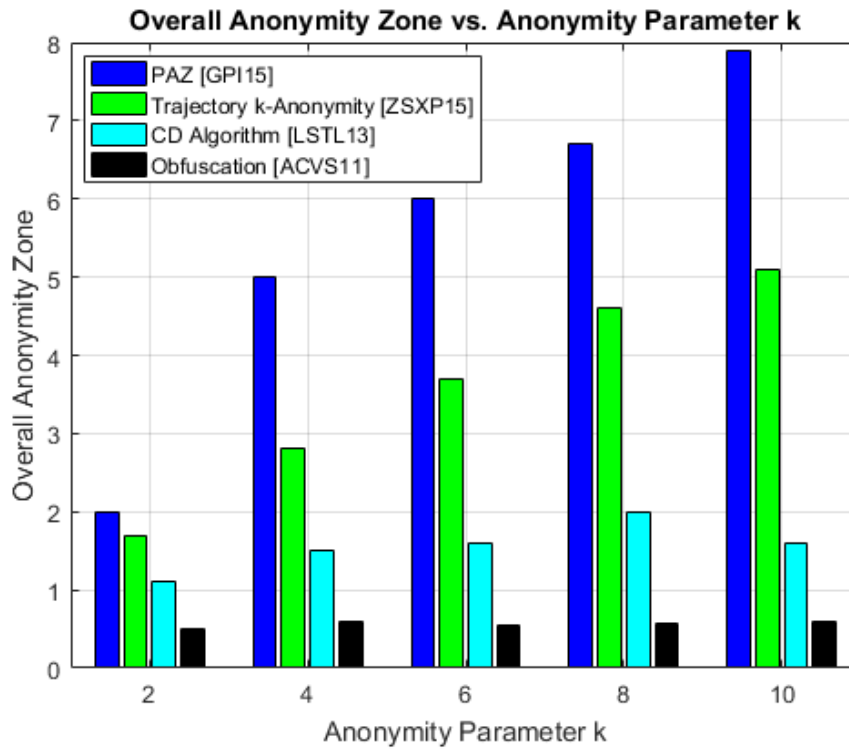


Figure 3.6: Overall Anonymity Zone vs. Anonymity Parameter k

Third, we show how the overall anonymity zone size changes according to the anonymity parameter k for dummies’ generation. From Figure 3.6, we can see that the size

of the anonymity zone is becoming larger for PAZ and trajectory k-anonymity, while the CD has a random pattern because of the irrelevance of parameter k. All three algorithms outperform the baseline. The trajectory k-anonymity increases the size of the anonymity zone when it needs to achieve k-anonymity with a larger value of k. Because the anonymity zone generated by CD is not correlated to the parameter k, the trajectory k-anonymity performs better than the CD algorithm. In other words, the algorithms that perform better have a strong dependency on parameter k. The PAZ generates more distributed anonymity zones when parameter k increases. Thus, the overall size of the anonymity zone becomes sublinear due to the overlap of the distributed anonymity zones. The PAZ algorithm shows its obvious advantage over all other methods to generate distributed anonymity zones protecting mobile users.

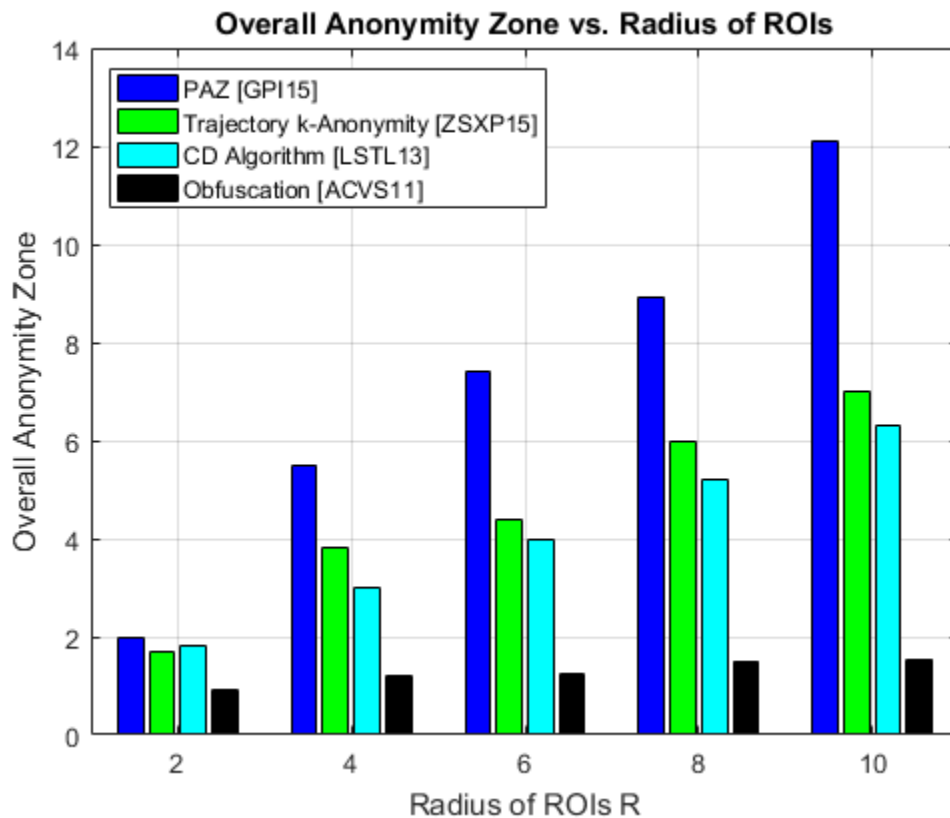


Figure 3.7: Overall Anonymity Zone vs. Radius of ROIs R

Furthermore, we visualize the size of the overall anonymity zone according to the changes of the radius of ROIs. Figure 3.7 presents that the PAZ, the trajectory k-anonymity and the CD are increasing their zones when the radius of ROIs increases. The baseline randomly changes the size of the anonymity zone. Based on the observation, trajectory k-anonymity has a slightly higher anonymity zone size than CD. However, the difference of the anonymity zone size between the two algorithms is smaller when comparing with their difference in Figure 3.6. For each value of radius of ROIs, the difference ranges from 1 to 3 units. We also see that the anonymity zone of the CD becomes larger along with the increase of the radius of ROIs. The reason for the increase of the zone size is that the generated anonymity zone directly depends on the value of radius for CD algorithm. Moreover, both the CD and trajectory k-anonymity perform much better than the baseline. The PAZ generates the larger size of the anonymity zone compared to the CD and the trajectory k-anonymity, while the baseline has the lowest size of the zone with a given k value. The query area becomes larger when the query radii increase. The anonymity zone covered by circles generated by PAZ is enlarged at a quick rate. The rate is larger than the trajectory k-Anonymity and the CD. Thus, the PAZ gains the largest size of the anonymity zones to benefit the LBS users if they adjust the query radius to be larger.

Last but not least, in our future research, we will try to improve the communication costs under the optimal privacy level constraints. We would like to find the optimal anonymity parameter with a radius for users to choose. We will explore how the function parameter n changes, affecting the trajectory privacy level. Furthermore, we will consider data de-anonymization techniques as advanced tools for attackers to launch more serious attacks, and find solutions to countermeasure such attacks.

3.5 Summary

In this chapter, we proposed a novel pseudonym-based anonymity zone generation mechanism for location and trajectory privacy protection. Based on a geometric transformation algorithm, we deploy a pseudonym changing process to periodically change the users' identities. Furthermore, we design a personalized dummies generation process to let users generate distributed anonymity zones to further enlarge the size of overall anonymity region. We also introduce a strong adversary model named inference attack that the attacker can collect side information to launch such an attack. Our solution can reduce the attack risk and preserve a high level of trajectory privacy for mobile users in the real world.

CHAPTER 4

QUERY-FEATURE-BASED ATTACKS AND PRIVACY PROTECTION

4.1 Introduction

In this chapter, we propose location regional probabilistic inference attacks based on historical query features, and present novel algorithms to protect user privacy against such attacks. By adding randomness in probabilistic k-anonymity augmented with grid obfuscation, we simply make it hard for the attacker to guess the user's location region even when the historical location probability distribution data is known to the attacker. To the best of our knowledge, this work is the first work to really investigate the query features as the tool needed to make inferences about the user's location region based on historical query data, and also design algorithms to prevent such attacks. In summary, our contributions are listed as follows:

- We propose a novel attack model named query-feature-based location regional probability distribution inference attack that can be launched by adversaries based on the historical query reported location data.
- We model the user privacy protection problem as a linearly-constrained convex optimization problem.
- We propose randomized approximated searching algorithms to solve the optimization problem, and perform probabilistic k-anonymity with grid obfuscation to stop the probabilistic inference attack.
- We comprehensively evaluate the proposed solution by privacy level, utility analysis, approximation ratio, computation complexity and communication cost.

- We conduct simulation and analysis based on a real word dataset to show the effectiveness of the attack model and the proposed solution.

4.2 System Model

4.2.1 System Architecture

The LBS system consists of a set of users with mobile devices, the localization network, internet cloud and LBS server in Figure 4.1. The users' devices can obtain the coordinates from GPS, Base Station or Wi-Fi. The mobile users send the LBS queries with location data through the internet cloud, and they finally reach the LBS server. The LBS server retrieves POI data based on users' queries and sends back the related POIs to the users.

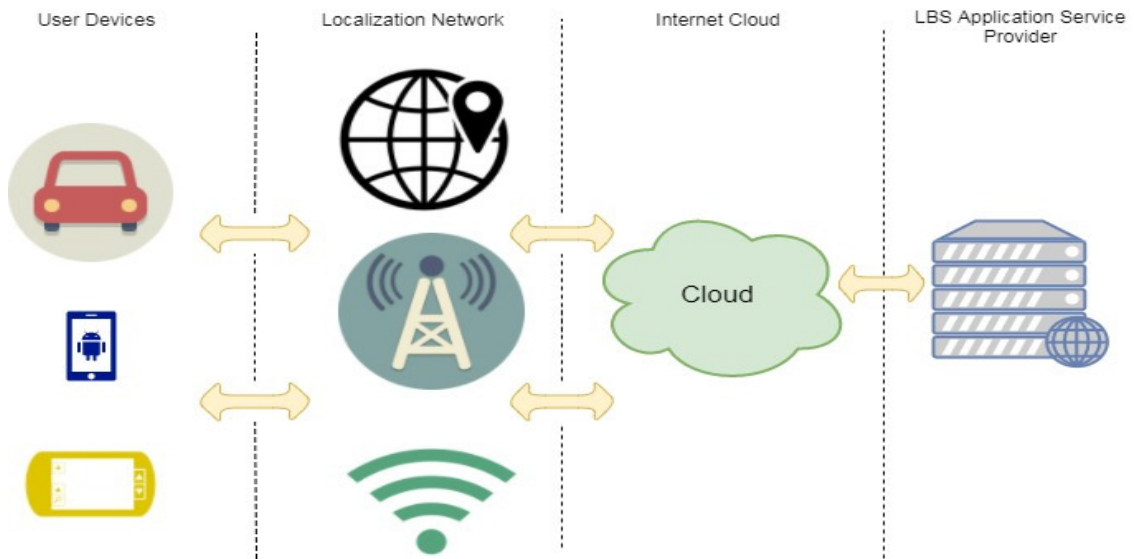


Figure 4.1: System service architecture

4.2.2 Feature-based Inference Attack Model

We proposed a new attack model named query-feature-based probabilistic inference attack, which means an attacker can monitor the historical queries of the users. Consequently, the

attacker can make an inference about the user location region probability distribution based on query features. We assume attackers can be LBS service providers themselves or external attackers who have compromised the LBS servers [ZSZZ15] [LLG⁺13].

Definition of a Query Feature: a query feature is a search keyword or category that is provided by the LBS applications which can be used by the user to find the most relevant information related to his or her location. In other words, a feature is a keyword that users are interested in searching for with the given location. For example, if a user is searching for a bank, then “bank” is a query keyword or feature in the query.

Illustration of a Query-Feature-based Inference Attack: The attacker may obtain the spatial probability distribution based on historical query features for users. The Gowalla dataset [LLAM13] is collected from a location-based social network called Gowalla. In this dataset, we have 36,001,959 check-ins made by 319,063 users over 2,844,076 locations. The locations were grouped into different categories. We use the categories, e.g. “Casino”, as the query features. We select a subset of the dataset for illustration purpose. The test network area is 10 miles \times 10 miles which is similar to the size of Miami city. We set the query feature to be “Casino.” From the Gowalla dataset, there are about 100,000 locations with more than 2,000 users in this area. Therefore, the attacker can retrieve the information about those users and their related location information. From the information, the attacker can build the mapping between the query feature and the query location. Specifically, the way to derive the probability distribution for a specific feature by the attacker is as follows:

- Step 1: Discretize the query area into unit sized grids $\{g_i\}_n$ with 100×100 in total.

Note, the words “grid” and “region” are exchangeable with one another.

- Step 2: Calculate location probability distribution over the grids, i.e. find $P: \{g_i\}_n \rightarrow [0, 1]$ for a query feature $f_i \in \{f_i\}_m$, e.g. “Casino”, based on a historical query set Q for the 100×100 grids.
- Step 3: The attacker makes an inference of the user location when a new query with the feature “Casino” released by a user. The attacker infers that the user has a higher probability located in the grid 1. In addition, if there are only grid 2, 3 and 5 in the user query, the attacker can infer that the user is more likely located in grid 2 since grid 2 has a higher probability than grid 3 and 5.

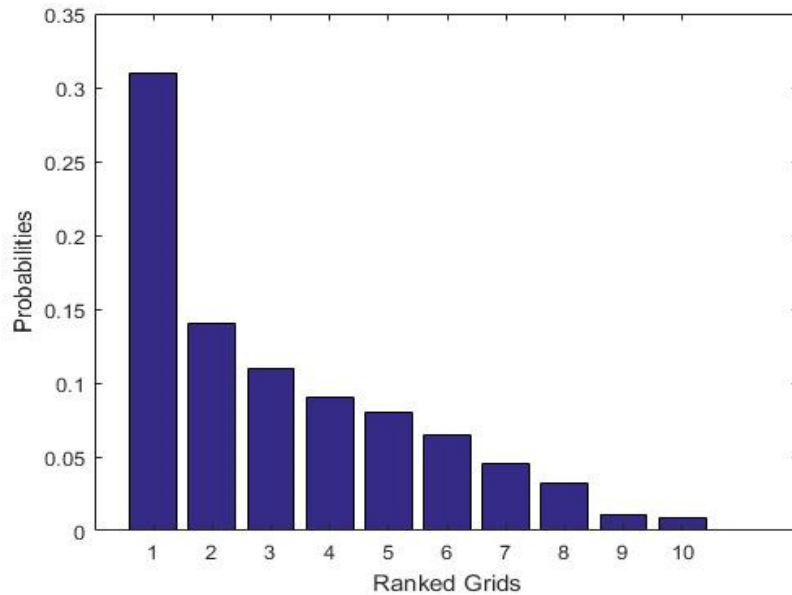


Figure 4.2: The probability for the top 10 grids

Experiment results: Figure 4.2 shows the top 10 grids or regions that have the highest location probability based on the query feature “Casino.” From the Figure 4.2, an attacker can infer that a user who sends a new query containing the “Casino” keyword has a high probability located in the first grid. Thus, the user location regional privacy can be compromised. Even if the user sends an obfuscated location, the attacker can still get some

information from the query feature. This is a strong attack model even without knowing the location data in the query itself. The purpose of the experiment is to build a probabilistic mapping between a specific query feature and location probability distribution. When a new query comes, the attacker can quickly figure out the possible user location region based on the query feature in the query.

Possible Attacks: based on the query area 10×10 miles and the grid number 100×100 in the experiment, we can approximately get the grid size, which is 160×160 square meters. The attacker can launch a privacy attack or even locate a user in the real world in this region. For example, after obtaining user grid 1 based on the inference process, attackers can send location-based advertisings to grid 1 based on the query keyword “Casino” to reduce their ad costs to other non-related grids. More seriously, attackers can also recursively apply steps 1-3 to reduce the grid size to find more fine-grained grids about where the user locates if they collect more queries from grid 1.

We explored the relationship between the grid size and the possible location probability distribution. We found that, when the grid size is too large, there is no useful information for attackers; when the grid size is too small, there is no general pattern for finding the relationship between the query feature and the location distribution. From the user side, if the grid size is small, the user privacy may be compromised; while if the grid size is too large, the LBS query cost is too high. We plan to quantify about the impact of grid size in our future work.

Note: Different query features have different probability matrixes. The attackers can build the probability matrixes for all query features or for query features with high-query frequencies using the above steps. The Q set in step 2 can be a historical query set for all

users, which is released periodically by the service provider for public research purpose. The Q set can also be a historical query set for a specific user and is already known by the user. The users can also build feature-based probability matrixes based on the public released query set or his/her own historical query set Q .

We agree that discretizing the query region into grids is a common method in many existing works. However, we have not seen any existing works that propose the attacks based on LBS query features. Most of the existing works focus on using discretizing the region into grids so that users have no need to send their real locations to the LBS server. However, in this chapter, we propose an attack model that the attacker can find the grid where a user is located based on the query feature information in the user query. They are totally different things.

4.2.3 Problem Statement

LBS users can send queries with spatial-temporal information to the LBS server for retrieving related POIs with certain query features. We assume LBS users perform a simple random walk mobility model. In the query, it has the format $\{UID, t, loc, feature\}$. UID is the user identity, t is the query time, loc is the query location and feature is an interesting keyword that users want to retrieve from the LBS server. The problem is how to defend against such attackers with obtained spatial probability distribution knowledge. The mobility model does not affect the problem formation since this problem is more focused on location privacy but not trajectory privacy. We will explore how the proposed attack model works for inferring a user's trajectory in our later work.

4.3 Query-Aware Privacy Algorithm Design

In this session, we formulize the user privacy problem and design query-aware privacy-preserving algorithms to distort the attackers' knowledge. The idea is to find $k-1$ different grids with close probabilities to the grid where the user is located in a random way so that the attackers cannot figure out in which grid the user is located as well as how the grids are selected. Our design considers that the attacker conduct location inference using historical query data based on query features. For a query $Q_i = \{UID_i, t_i, loc_i, feature_i\}$, in order to protect the user location loc_i privacy for a given query feature $feature_i$, we propose unifying the query-feature-based probabilistic k -anonymity algorithms with grid obfuscation to hide user location, which is different from previous k -anonymity works, such as [ZKMC13], [SCH⁺16] and [GPI15]. Note: at this time, we are not focusing on concealing query features, which is our next work for query privacy protection.

Definition 1: Probabilistic k -Effectiveness. A query-feature based privacy mechanism M provides the Probabilistic k -Effectiveness, if for $\forall l \in L(M)$, and $\forall l_1, l_2 \dots l_{k-1} \in l$, the following absolute difference of the conditional probabilities is smaller than threshold t holds true, given the query feature f :

$$\|P(l_q|f) - P(l_0/f)\| < t, \text{ for } \forall q \in [1, k - 1]$$

This property guarantees that for any of the grids in the grid set with k members, the attacker cannot remove any grid q from the received query, because its conditional probability is bounded to the user's true location or grid l_0 's probability. By this way, the probabilistic k -anonymity can be achieved with k -Effectiveness. In practice, we divide the

region into grids so a grid set will be sent to the server instead of users' real locations. The $feature_i$ still stays in the query while the loc_i becomes a grid set. We select the other $k - 1$ location grids with close probabilities with the grid in which the user is truly located. For any given query q with feature f and a location grid l :

$$l = l_0 \in \{g_i\}_n \quad (4.1)$$

Find the $(k - 1)$ grids:

$$l_1, l_2, l_3, \dots, l_{k-1} \in \{g_i\}_n \quad (4.2)$$

$$such\ that\ \| p\{l_i\} - p\{l_o\} \| < threshold\ t$$

where g_i stands for grid g with index i , l_i stands for the grid that the user is located in.

We use g_i and l_i exchangeable in the work. $p(l_i)$ stands for the probability distribution of l_i based on the feature $feature_i$. t is the difference threshold between the user grid probability distribution with the probability distribution of the grids that we are searching for protecting the user.

More specifically, to find the most similar grids (similar means grids with very close location probabilities), we formulate the problem as the following linearly-constrained convex optimization problem:

$$\max \sum_{i=0}^{k-1} P'(l_i) \log P'(l_i) \quad (4.3)$$

$$st. l_i \in \{g_i\}_n \text{ and } l_i \neq l_j, \forall i \neq j$$

$$where\ P'(l_i) = \frac{P(l_i)}{\sum_{i=0}^{k-1} P(l_i)}$$

where $P'(I_i)$ is the conditional probability of the grid I_i in the selected grid set. The purpose is to find a grid set within which the entropy is maximized so that the attacker will have the maximum uncertainty to guess the user's true location. The best case is a uniform distribution for all the grids in the grid set.

Figure 4.3 is an illustration of probability distribution in grids: $n \times n$ matrix by dividing a region into grids, and different colors stand for different location probabilities. For example, if a user located in the first top green grid, he/she can select the other two green grids to send together to the LBS server. Those grid has similar probabilities so the attacker has difficulty to figure out which grid the user is really locating in.

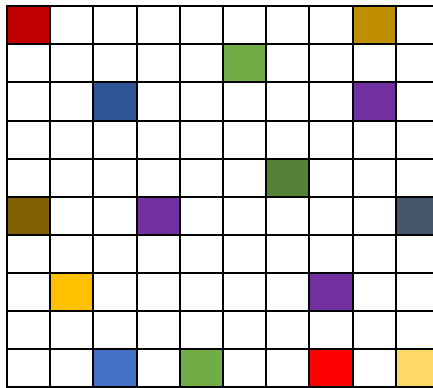


Figure 4.3: Location probability distribution example

In order to find a grid set for the user, we proposed a randomized search algorithm by adding randomness into the search process. Based on the random search, it would be hard for adversaries to find the patterns among the grids in the grid set. This algorithm is an approximation algorithm for solving the linearly-constrained optimization problem. We also set a time bound T for the algorithm to be able to respond the user request in a reasonable time. The proposed algorithm is shown in Algorithm 1.

Algorithm 1 Query-aware Randomized Search Algorithm (Q-RSA)

Input: (User U, Time t_0 , Location l_0 , Query Feature f)

Output: (Grid_set)

Dis_f \leftarrow Probability_Distribution_Matrix (f)

Grid_candidate \leftarrow {};

Grid_set \leftarrow { l_0 };

For grid l_i in Dis_f:

 If $|p(l_i) - p(l_0)| < \text{threshold } t$ and $l_i \neq l_0$:

 Add the grid l_i into Grid_candidate;

While sizeof (Grid_set) < k:

 Generate a random number index j;

 Select the grid l_j with index j in Grid_candidate;

 If l_j not in Grid_set:

 Add the grid l_j into Grid_set;

For l_i in Grid_set:

$$\text{Calculate } P'(l_i) = \frac{P(l_i)}{\sum_{i=0}^{k-1} P(l_i)};$$

Calculate $E \leftarrow \sum_{i=0}^{k-1} P'(l_i) \log P'(l_i)$;

Repeat sampling grids from Grid_candidate until E in Grid_set is maximized in a certain time bound T if necessary;

Return Grid_set;

End

This Q-RSA tries to find the k-1 grids with similar probabilities to the one in which the user is located in a random way. By conducting random search, there is no pattern that can be found about how to select those grids. The attackers have difficulties figuring out how to select the grids.

The second algorithm we propose is a baseline but fast algorithm called Top-K algorithm. This algorithm simply sorts the grids in the Grid_candidate by probabilities and selects the top k - 1 grids with higher or lower probabilities in the set. This algorithm works faster because it just quickly sorts without the need to iteratively find the related grids.

Algorithm 2 Top-K Algorithm (Top-K)

Input: (User U, Time t_0 , Location l_0 , Query Feature f)

Output: (Grid_set)

Dis_f \leftarrow Probability_Distribution_Matrix (f)

Grid_candidate \leftarrow {};

Grid_set \leftarrow { l_0 };

For grid l_i in Dis_f:

 If $|p(l_i) - p(l_0)| < \text{threshold } t$ and $l_i \neq l_0$:

 Add the grid l_i into Grid_candidate;

Sort (Grid_candidate)

Pick top k - 1 grids from Grid_candidate and add into Grid_set;

For l_i in Grid_set:

$$\text{Calculate } P'(l_i) = \frac{P(l_i)}{\sum_{i=0}^{k-1} P(l_i)};$$

Calculate E \leftarrow $\sum_{i=0}^{k-1} P'(l_i) \log P'(l_i)$;

Return Grid_set;

End

Moreover, we compare our work with Greedy Algorithm (GA) and Randomized Greedy Algorithm (RGA) in the work [CMBL11] to find related grids. They are different ways to find nearby grids with close probabilities to the user-located grid.

How do our algorithms (e.g. Q-RSA) stop the inference attack? We find a set of grids with very close probabilities for the given query feature, so the attacker cannot make an accurate inference about which grid the user is located in. Our algorithms try to find grids in a more random way, so no specific pattern exists in selecting these grids. For a query feature in a new query, the attacker cannot distinguish in which grid the user is located, thus our algorithms stop the inference attack. Furthermore, based on the availability of the historical query data to the users, there are two ways to apply the proposed algorithms: 1) a user can apply the proposed algorithm in a personalized way based on the Q set, which

is his or her own historical query data; 2) a user can apply the algorithm in a general way based on the Q set, which is a historical query set for all users. In our approaches, we allow users' devices to build the probability distribution matrix based their own historical query data for their used query features. The users' devices aware and store the query history when they are releasing the queries. Therefore, the matrix is highly personalized and being built automatically on the users' devices. The devices can select the grid set satisfying the privacy requirements with a self-chosen grid size. Thus, the grid size may not be expected as the same size by the attacker, and this mismatch produces more difficulty for the attacker to locate the user. The intuition is no matter how a user choose the grid size, the grids in his/her grid set have tightly bounded probabilities that make them indistinguishable.

4.4 Analysis and Performance Evaluation

In this section, we evaluated the proposed algorithms by analyzing the privacy level, utility analysis, approximation ratio, computation complexity, and communication cost. We implement the algorithms in Python on a Dell computer with Intel Core i7-6700 3.4 GHz and 16.0GB RAM. We use a subset of the Gowalla dataset with related POIs distributed in this selected area.

4.4.1 Privacy Level by Entropy

The privacy level is measured by the entropy metric. In the ideal and optimal case, the selected grids are uniformly distributed and no knowledge is leaked based on the location probability distribution. In our solution, the entropy is the summation of the individual entropy of conditional probabilities of the selected grids in the grid set. The intuition is that we expect the selected grids shall have the same conditional probabilities in the grid set so

attackers have no obvious evidence to infer which grid is the true grid in which the user is locating. For measuring the entropy metric, we use $\sum_{i=0}^{k-1} P'(J_i) \log P'(J_i)$ to record the results of the four algorithms. The larger the entropy, the better user privacy.

First, we measure the impact of privacy parameter k , which is selected by the user on the entropy metric. The result is shown in Figure 4.4. We can see the theoretical and optimal entropy for each privacy parameter k . The best algorithm should give the entropy that is closest to the optimal entropy boundary. The Q-RSA outperforms other algorithms based on entropy levels for all the values of parameter k . The RGA algorithm performs better than Top-K but worse than Q-RSA. The GA algorithm has the lowest performance. Q-RSA has the highest level of randomness in its grid search process and the probability for each grid to be selected is equal. On the other hand, the GA algorithm selects the grids in a greedy way which results in different probabilities for those grids to be selected. Thus, the GA algorithm performs worst in the experiments.

Second, we also measure the impact of the iteration parameter r on the entropy metric. The result is shown in Figure 4.5. From the results, we can see that our proposed Q-RSA algorithm is the best one approaching the optimal entropy than other algorithms. The reason is similar to the measurement of parameter k . In each iteration, the probability to select each grid is equal for Q-RSA but not equal for other algorithms. The GA algorithm still performs worst in this case. The Top-K and RGA algorithms perform better than GA but worse than Q-RSA. The RGA has a better performance than the Top-K. In the experiment, we set up the parameters for the iteration parameter r and privacy parameter k in Table I. Iteration parameter r is the iteration number for the algorithms to run in order to get better entropy. k is the privacy parameter that can be set by the user.

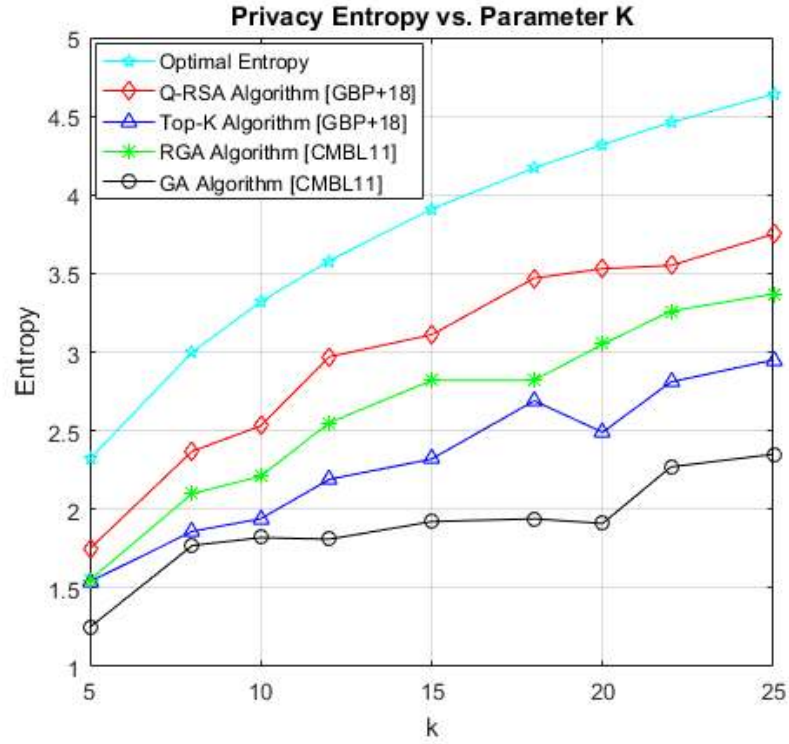


Figure 4.4: Entropy E vs. Privacy Parameter k

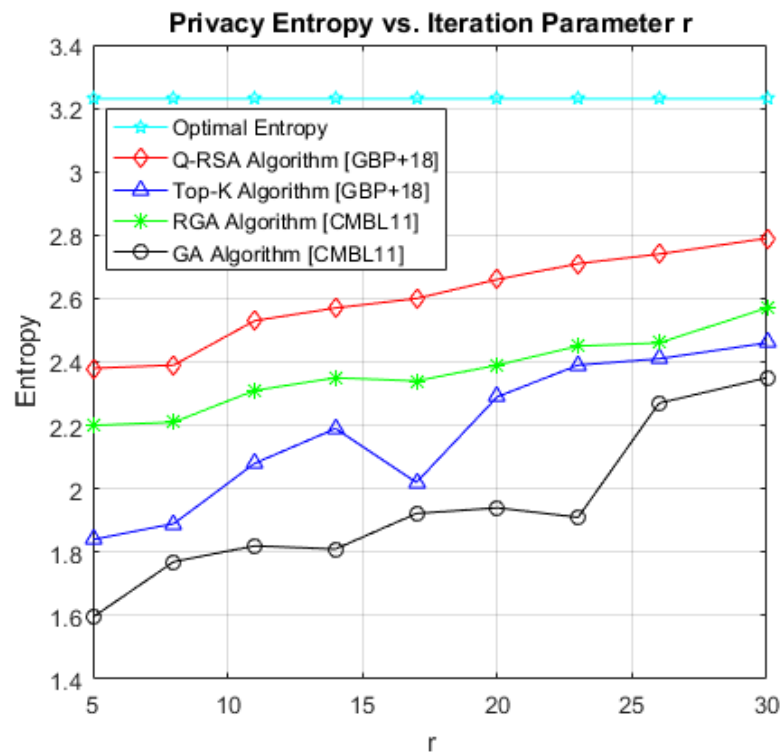


Figure 4.5: Entropy E vs. Iteration Parameter r

Parameter	Values
r	5, 8, 11, 14, 17, 20, 23, 26, 30
k	5, 8, 10, 12, 15, 18, 20, 22, 25

Table I: Parameters in the Experiments

4.4.2 Utility Analysis

What about the LBS utility after applying our algorithms? Since the user is located in one of the grids sending to the LBS servers, the LBS service utility can be guaranteed for the user query based on the user located grid information. The cost is the computation and communication costs in session 4.4.4 and 4.4.5 used to mislead the attackers by selecting and sending the grid set.

4.4.3 Approximation Ratio

From session 4.3, we formulated a convex optimization problem. The global value for the entropy is when the k grids have a uniform distribution. The maximum entropy for the uniform distribution is as follows:

$$P(J_0) = P(J_1) = \dots = P(J_{k-1}) = \frac{1}{k} \quad (4.4)$$

Then:

$$\begin{aligned} E_{optima} &= \sum_{i=0}^{k-1} P'(J_i) \log P'(J_i) \quad (4.5) \\ &= \frac{k \times 1}{k \times \log k} \\ &= \log k \end{aligned}$$

By computing the entropy of the selected k for Top-k, RSA and RGA, GA, we compute the approximation ratio for each algorithm:

$$\begin{aligned}
 \text{Ratio} &= \frac{E_{\text{optimal}}}{\sum_{i=0}^{k-1} P'(l_i) \log P'(l_i)} \quad (4.6) \\
 &= \frac{\log k}{\sum_{i=0}^{k-1} P'(l_i) \log P'(l_i)}
 \end{aligned}$$

where $P'(l_i)$ is the conditional probability for the selected grids in the grid set. From (4.6),

we know the ratio is smaller and the algorithm is better.

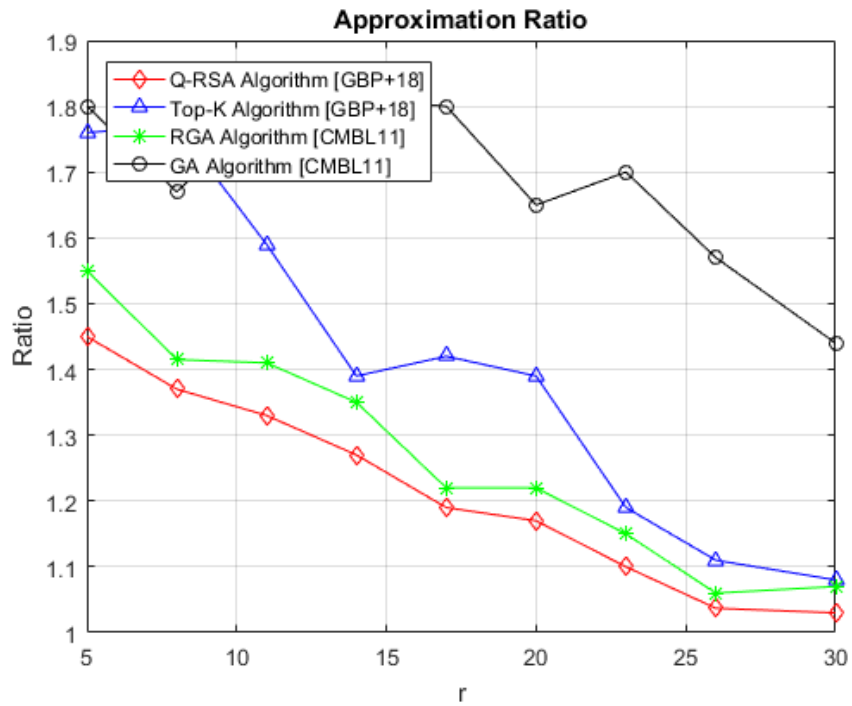


Figure 4.6: Approximation Ratio R vs. Iteration Parameter r

First, we measure the impact of iteration parameter r on the approximation ratio. We fix k value and run r in the selected parameters for the running number. Here, we also run multiple times for other algorithms in order to show their convergence process to the best approximation ratio. Using this ratio metric, we get the results in Figure 4.6. The experiment results show that our proposed Q-RSA is better than RGA and GA. The reason

is that the grids have equal probabilities to be selected in the random search process. Thus, the approximation ratio for the random search is lower than other search processes. For RGA and GA, they choose the grids in a semi-random and a grid way, respectively. Therefore, they have a higher approximation ratio.

Second, we also measure the impact of k , which is selected by the user on the approximation ratio. In the experiment, we fix r and record the results, as shown in Figure 4.7. The results also show that with the increase of privacy parameter k , our proposed Q-RSA algorithm has a better chance to get a better approximation ratio. The reason is similar to the first measurement with iteration number r . No matter how the privacy parameter k changes, the GA and RGA have a non-random way to select the grids. They release some information so their information entropy is lower. Thus, GA and RGA have higher approximation ratios than Q-RSA. The Top-K has a worse performance than Q-RSA and RGA but better than GA.

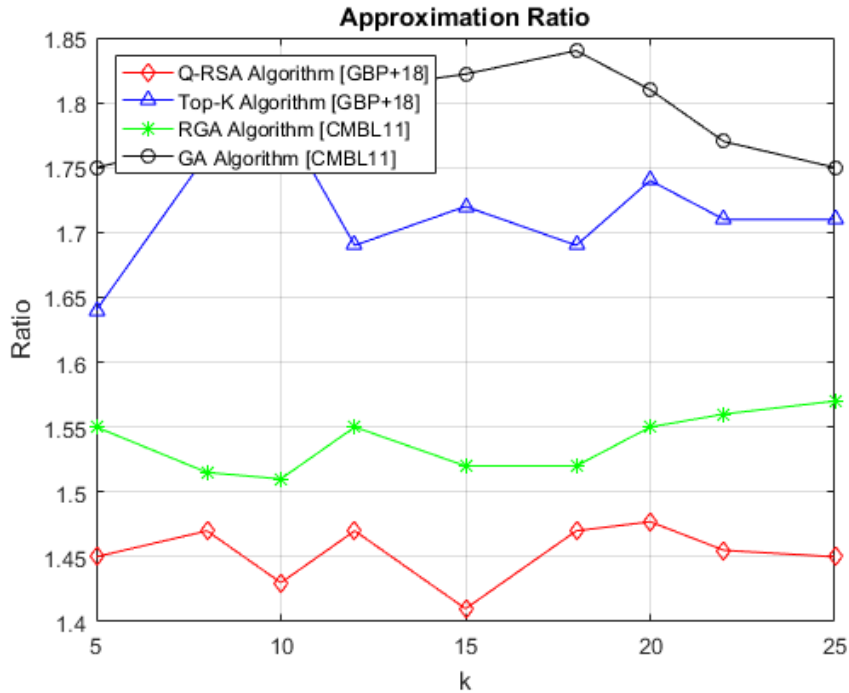


Figure 4.7: Approximation Ratio R vs. Privacy Parameter k

4.4.4 Computation Complexity

Computation cost is another important metric for evaluating the performance for our proposed algorithms. When generating a grid set for a LBS query, we need to consider the computation complexity. We mainly consider the time complexity of the proposed algorithms. We assume the probability matrix has $n \times n$ grids. The scan process for the probability matrix is the same for four algorithms, which is $O(n^2)$. The main difference among the algorithms is how to select the $k-1$ grids.

- Top-K: $O(\theta \times \log m)$, where m is the grid number in the Grid_candidate, the algorithm sort the grid in order to find the top k grids.
- Q-RSA: $O(\theta \times k)$, where k is the number of grids sent to the LBS service provider, θ is the number that the algorithm needs to loop in the time bound T .
- GA and RGA [11]: $O(n^2)$ is the upper bound for the Greedy Algorithm and the Randomized Greedy Algorithm.

1) For the comparison between Q-RSA and RGA, we see from the RGA algorithm that it has an addition step to select an adjacent grid when the random number is below the given threshold. That step needs to compare the probabilities of the neighboring grids with the user grid, which consumes more time than a random search which is $O(1)$. Therefore, the RGA has a higher time complexity than our Q-RSA algorithm. Similarly, the GA has a higher time complexity than Q-RSA. In sum, the Q-RSA algorithm performs better than GA and RGA in terms of time complexity.

2) For the comparison between Q-RSA and the faster Top-K algorithm. Their time complexities are as follows:

$$\theta \times k = m \times \log m \quad (4.7)$$

$$\theta = \frac{m \times \log m}{k} \quad (4.8)$$

We let the time complexity of Q-RSA equal Top-K, and get θ equal $m / k \times \log m$. If θ is smaller than $m / k \times \log m$, then Q-RSA is better than the baseline; otherwise it is slower than the Top-K. This gives a threshold to select θ in the Q-RSA algorithm from (4.9):

$$\theta \leq \frac{m \times \log m}{k} \quad (4.9)$$

4.4.5 Communication Cost

The communication cost mainly considers the grid size and number of grids that will be sent to the server, the number of queries, as well as the POIs data to be sent back from the server. We assume the POIs, which are located in the selected grids, will be sent back to the user. Each grid contains c byte information. The query number is n . Thus, the uplink communication cost is shown in (4.10). For n queries, the uplink contains nkc byte information. The downlink communication cost is determined by N_i , which is the number of replied POIs in each grid.

$$C_{uplink} = k \times c \quad (4.10)$$

$$C_{downlink} = \sum_{i=0}^{k-1} q N_i \quad (4.11)$$

We assume each of the POI contains q byte information. Totally, the downlink contains $\sum_{i=0}^{k-1} q N_i$ byte information for each query from (4.11). The overall communication cost is $(k \times c + \sum_{i=0}^{k-1} q N_i)$ byte information for a single query. For n queries, the overall communication

cost is $n \times (k \times c + \sum_{i=0}^{k-1} q N_i)$ byte. Since all the algorithms send the same k number of grids

to the server, the main difference is the number of POIs located in each grid. We evaluate the communication cost with the experiment setup and get the statistics by averaging the cost for each algorithm. We set the q to 64 bytes. The number of query is set to 1000. The average cost in unit KB per query for each algorithm is shown in Figure 4.8. On average, our algorithms show advantages in the experiment, even if the related POIs are randomly located in each grid.

Furthermore, the Q-RSA has the lowest average communication cost per query which is smaller than 100 bytes. The Top-K and Q-RSA algorithms have slightly higher average cost, which are 102 and 103 bytes, respectively. The RGA has the highest average communication cost per query which is around 120 bytes. Even if there is some randomness involved in the grid selection process, our Q-RSA and Top-K algorithms perform better than RGA and GA algorithms in the communication cost metric. In fact, along with increase of the query number, user privacy needs more effort to be protected. If the query number increases, attackers have a better chance to observe and even track the user behavior based on the historical query data with potential side information. The side information can be any external information that can be obtained by attackers. For example, online social networks and/or public profiles may provide some side information about the target user. Thus, researchers should focus more on releasing strict assumptions about what information is available to attackers, and develop new methods to satisfy stricter privacy measurements to better protect user privacy. The stricter privacy measurements should have a strong property that no matter how side information is used, user privacy can still be measured in an accurate way.

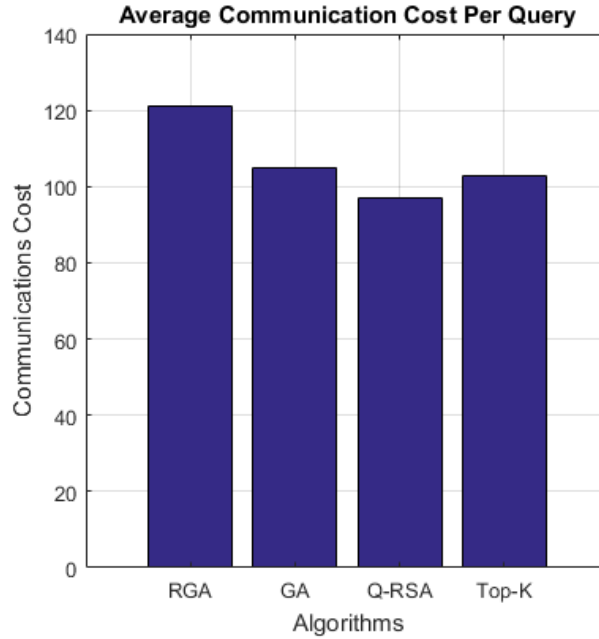


Figure 4.8: Average Communication Costs vs. Algorithms

Finally, after applying our proposed algorithms, LBS service providers can still use or release the newly obtained LBS query dataset for public research purposes. The grids with high probabilities for a specific query feature in the past can still have relevant high probabilities, and the grids with low query frequencies for a query feature may still have relevant low frequencies. Our proposed algorithms will not change the inner statistical structure of the existing dataset, and only affect the absolute query numbers for a specific query feature and query location. In addition, our solutions can have more benefit to hide the correlation between query location and query interest to avoid the possible privacy leakage for specific users in future data publications. The query data can be used by the public research to optimize the related infrastructure, such as road planning and base station replacement, based on the high-level statistical query information. The dataset obtained from LBS servers can still be utilized by researchers and scientists to discover new knowledge and benefit society.

4.5 Summary

In this chapter, we focus on a new location regional privacy problem from a historical query point of view, and consider query search features as new tools for the attackers to locate users in a specific region. We propose a new probability inference attack based on query features, and design novel randomized algorithms to countermeasure the privacy attacks. We comprehensively evaluate the proposed solution by privacy level, utility analysis, approximation ratio, computation complexity and communication cost. We conduct simulation and analysis based on a real-world dataset to show the effectiveness of the attack model and the proposed solution.

CHAPTER 5

PRIVACY-AWARE MOBILE SENSING MECHANISM FOR VANET

5.1 Introduction

Mobile sensing applications and services are becoming pervasive on the sensor-enabled mobile vehicles. This trend is accelerated by the fact that cheap sensors such as GPS, light and accelerometer sensors are increasingly used in mobile platforms. All of the sensor data consists of a vehicular user's context modeling the behavior and/or environment, including mobility mode, location, activities, etc. Context-aware applications can utilize the real-time context to conveniently provide personalized service to the mobile vehicles. For example, one of the applications named *SocialGroupon* delivers recommendations or coupons when a vehicular user is with a group of close friends, and *Saga* automatically captures details of people's daily adventures, learns about them and offers feedback right when they need, and so on.

However, context-aware applications are curious of vehicular users' personal information. Privacy issue is an important factor for adopting such applications for mobile vehicles. Location-related context privacy has already attracted attention by smartphone users [Mic17]. Those mobile applications may not only collect users' context data for personalized service, but also aggressively build users' profiles for other purposes: targeted advertising and selling to third parties for revenue. A study has been performed on multiple applications on Android platforms that have access to users' sensor data including microphone and location. The study shows that half of the applications collect more data than they need for returning services, and also send the data to remote analytics servers or advertisement providers [EGC⁺10]. More seriously, contextual data stream may contain

very sensitive information like medical condition, physical location, etc., and which can be used by attackers to harm the vehicular users' well-being. On the other hand, users may lose most of the interesting functions and services if they simply turn off or remove those applications.

To improve vehicular users' experience and enable the adoptability of context-aware services and applications, privacy preserving solutions need to be developed. The work [CRJS13] considers how to prevent mobile apps from accessing the raw sensor data and how to infer context from the sensor data. Caching techniques have been used to address the privacy issue in LBS [ZCN⁺13]. However, they have several limitations. The major drawback for these schemes is that they use data cache for every query update resulting in too much storage consumption on mobile vehicles. Our work only uses data cache at the sensitive contexts. The works that aimed to address the contextual privacy problem are limited. Gotz et al. proposed a middleware to control the release of context stream to mobile applications in a private way [GNG12]. Wang et al. enhanced this work with the consideration of a real-time adversary model [WZ14]. However, the drawback for both of them is that the service quality will be reduced dramatically when they suppress the sensitive contexts. The main contributions of our work are listed as follows:

- We are the first work to use data caching techniques for solving mobile sensing privacy problems for context-aware vehicular applications.
- We identify the quality of query result reduction problems in the existing contextual privacy solutions, especially when mobile users are in sensitive contexts.
- We propose a privacy-aware peer-to-peer collaborative mechanism for improving users' experience by considering both privacy preservation and service quality.

5.2 System Model

5.2.1 System Architecture

The system consists of a vehicular user with mobile vehicles, the cell tower and context-aware services as shown in Figure 5.1. The mobile vehicle learns the context information by the embedded sensors: one type is the physical context, the other logical context (high level context drawn from physical context). The mobile vehicle releases the context sensing data stream to context-aware services through the cell tower. After the query process, context-aware service providers return the query results to the vehicular user. The mobile vehicle moves on road networks.

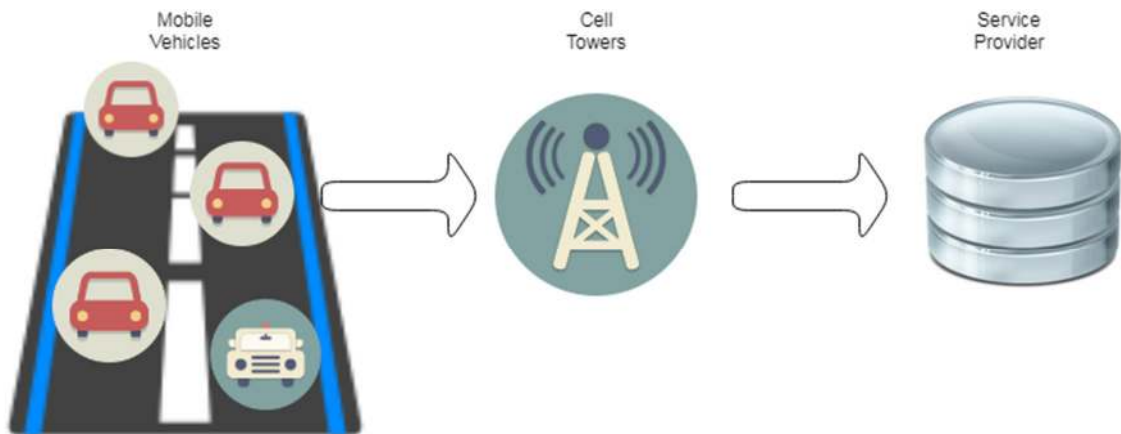


Figure 5.1: Mobile system architecture

5.2.2 Attack Model

We make the assumption that context-aware service providers themselves can be the attackers who would like to collect the vehicular users' context information, such as location and mobility mode, for mobile advertisement purposes, and/or for increasing revenue by selling the users' context information to third parties. The context-aware service providers can be compromised by attackers. We also consider another type of adversary

who can eavesdrop on data packets through the communication networks. In addition, we consider that attackers know the Markov chain of mobile vehicles, temporal relations among their query updates, and can launch attacks at the users' sensitive contexts.

5.2.3 Problem Statement

In previous work [GNG12] and [WZ14], the mobile users need to suppress sensitive contexts for preserving their privacy and defending against the intelligent adversaries. Even if the two works can conduct the optimal strategies for controlling the data granularity for the sensitive contexts, they lack the considerations of privacy gain and service quality guarantee, when the users are in such contexts. This is a serious problem that needs to be addressed to improve the mobile users' experiences, when they are using personalized vehicular applications. The goal of our work is to improve privacy levels and maintain the service qualities when users suppress the sensitive contexts, leading to poor quality of the query results.

5.3 Privacy-Aware Mobile Sensing Mechanism for VANET

The privacy mechanism on mobile vehicles consists of four procedures as shown in Figure 5.2. For the first procedure, we adopt a privacy-aware middleware to control the granularity of the released data for the embedded sensors. This procedure is responsible for identifying and suppressing the sensitive contexts for mobile vehicles. The second one is the local data cache query procedure, which is responsible for the query process, when the first procedure suppresses the sensitive context but users need a high quality of the query results. The third procedure is the query process for nearby vehicles which may have the interesting query results, when there is no local data available on the user's mobile platform. The last

procedure is to release the suppressed context if there is no local data available on nearby vehicles. The key difference between our work and other works is that we use the cached data to answer user queries only in a sensitive context, and this sensitive context cannot be drawn by a strong adversary model.

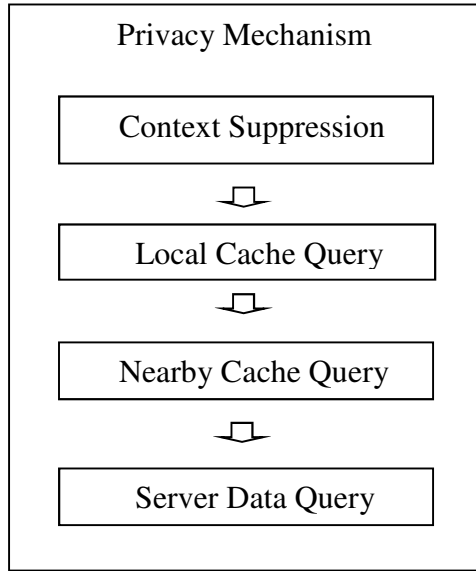


Figure 5.2: Mobile sensing privacy mechanism logic flow

Context Compression Procedure

We first allow users to set some sensitive contexts by using a tool at the very beginning [TCD⁺10]. This procedure is used for privacy check to decide whether to release or suppress the user's current context. The system periodically senses the user's context, such as C_1 and C_2 , for services. For a sensitive context, it will be suppressed and stored, and the workflow will proceed to the second procedure. For a non-sensitive context, it will be directly released to the service provider for the normal query process without worrying about the privacy leakage from the server query procedure. For example, if C_1 is a non-sensitive context inducted from the component, it will not be suppressed and directly

released to the fourth procedure for processing; if C_2 is a sensitive context, the system will suppress it and produce an outcome C_2' , and at the same time, C_2 will be sent to second procedure for processing. Here, we use the type of application T as the function for returning the interesting data, such as $T(C) = D$ or $T(C_2') = D$.

Local Cache Query Procedure

On a vehicle's mobile platform, we use a buffer B to store the cached data for a period of time based on the users' interests. The initialization of the buffer can be done by prefetching related data or simply setting to empty. After initialization, the buffer is ready for answering the queries from the default user or from other users. If the buffer is empty, it will be ready for updating by receiving query results from nearby users or service providers. Users can set the lifetime of the cached data based on its hit ratio. From the first process, we know that the context C is sensitive, so we need to search the local buffer B to see if there is a matching result $T(C)$ that passes the criteria A. If there is a match, the users can enjoy the result immediately. If there is no match, the workflow will go to the third procedure for further processing. Here we use C to search local buffer, but not C' . The system will only handle C' if there is no match results for C.

Nearby Cache Query Procedure

If the second process cannot answer a user's query by local buffer B, the user needs to send the query $T(C)$ to his or her nearby vehicular users to find interesting query results. Here, we make an important assumption that the sensitive contexts may be different for different users and the users are willing to share the results. For example, the context of a hospital

is sensitive for a patient but may be not sensitive for a doctor. Therefore, a patient may not be willing to share the results but a doctor may be willing to do so. If the received results $T(C)$ pass the criteria A , the local buffer B will be updated by the results.

Server Data Query Procedure

From the workflow in the mechanism, if a context C is determined as not sensitive by the first procedure, it can be sent directly by server data query procedure to the remote servers for retrieving the dataset $T(C)$. In addition, if the context C is sensitive and compressed by the first procedure to C' , then the compressed context C' needs to be sent to the remote service provider for accessing service. The application provider cannot learn the private information of the vehicular users. The result $T(C')$ will be retrieved and buffer B will be updated. Based on the criteria A , the vehicular users decide if they need to keep the data or clear them after the usage of that data.

5.4 Comparative Analysis

In this section, we analyze the privacy gain and query quality level for the proposed mechanism. Our work improves the privacy level compared to the baseline. The reason is that the basic work sends the compressed context to the server if the original context is sensitive. For our work, this step can be reduced if the local or nearby cache data is available to serve the user's request. Therefore, the server receives less query requests from users and learns less about their private information. Compared with the basic context compression scheme [WZ14], our work improves the query quality based on two points. First, the basic scheme compresses the sensitive context resulting in the reduction of query

quality, while the data cache query procedure returns the results in a local way. Second, the quality of the local or nearby cached data can be better than, or at least as good as, the query results retrieved from the server for sensitive contexts because of the setup of the criteria A.

5.4.1 Privacy-Level Analysis

At the beginning, the sensitive context set can be identified by the user. For the non-sensitive context, the first procedure decides if the context needs to be compressed based on its context sensitivity on the impact of the future context. The privacy loss is measured by the sensitivity of the context. If the adversary does not learn the sensitive context, the privacy loss is zero. The privacy can be defined as the sensitivity of the underlying context learned by the attacker minus the real sensitivity of that context. We define the privacy as follows:

$$\begin{aligned}
 P(\mathcal{C}_u^t) &= [\mathcal{S}_l(\mathcal{C}_u^t) - \mathcal{S}_r(\mathcal{C}_u^t)] \times \sigma \times \beta \\
 &= [\sum_0^t \mathcal{S}(\mathcal{C}_u^t) - \sum_0^t \mathcal{S}(\mathcal{C}_u^{t-1}) - \mathcal{S}_r(\mathcal{C}_u^t)] \times \sigma \times \beta \quad (5.1)
 \end{aligned}$$

where P is the privacy for the context c at time t. \mathcal{S}_l is the learned context sensitivity for context c for the attacker. S is the sum of the learned sensitivity of the past contexts. \mathcal{S}_r is the real sensitivity of the current context. σ is a discount parameter for the context privacy and β is a constant. If the local or nearby data satisfies the user's query request, then the privacy gain will be the reduction of current context sensitivity leakage. We define the privacy gain as follows:

$$P = -P(\mathcal{C}_u^t)$$

$$= [\sum_0^v S(c_u^{t-1}) + S_r(c_u^t) - \sum_0^v S(c_u^t)] \times \sigma \times \beta \quad (5.2)$$

In order to better quantify the privacy, we also use the sigmoid function for the privacy gain P' :

$$P' = \frac{1}{1 + e^{-p}}$$

$$= \frac{1}{1 + e^{-\{[\sum_0^v S(c_u^{t-1}) + S_r(c_u^t) - \sum_0^v S(c_u^t)] \times \sigma \times \beta\}}} \quad (5.3)$$

5.4.2 Service-Quality Analysis

The query quality of context-aware services and applications can be measured by the vehicular users' degree of satisfaction. It can be modelled by a sigmoid function of the accuracy of the context recognition. The sigmoid function is widely used to measure the approximated service quality or satisfaction for users [GJP⁺15]. The QoS for context c is shown as follows:

$$Q(c_u^t) = \frac{1}{1 + e^{-\theta(c_u^t - \eta)}} \quad (5.4)$$

where η is the satisfaction criteria or threshold, θ decides the steepness of the satisfaction curve of the QoS, u is the identity of the user, and t is the time stamp of the service request. We define the QoS loss as the server data satisfaction minus the local data satisfaction, which is shown as follows:

$$g = Q(c_u^t) - Q(c_u^t)'$$

$$= \frac{1}{1 + e^{-\theta(C_u^i - \eta)}} - \frac{1}{1 + e^{-\theta(C_u^i - \eta)'}} \quad (5.5)$$

5.5 Performance Evaluation

In this section, we conduct simulations to evaluate the performance of the proposed scheme compared with the baseline without caching capability. We study the privacy gain and the service quality based on a real-world mobility dataset, and we show the effectiveness and efficiency of our proposed mechanism.

5.5.1 Simulation Setup

Dataset: we evaluate the performance of our proposed mechanism by using the Reality Mining Datasets [RMD04]. The dataset contains the records of the continuous activities of more than 90 students and staffs at MIT with Nokia 6600 smartphone during the 2004-2005 academic year [EPL09]. The trace of each user includes features such as location granularity of cell tower, proximity to each other through Bluetooth, mobility mode such as walk or static, and activities such as calling or playing with applications. The combined length of all users' traces is 266,200 hours. Figure 5.3 shows the minimum, maximum and average trace length of all users based on the statistical information. We use location as the main feature to evaluate our solution as it is the most complete and fine-grained context in the data set. The minimum, maximum and average number of locations per user is 7, 40, and 19, respectively. We train a Markov chain for each user based on location traces. Based on the trained Markov chain, we simulate the users' behaviors. For sensitive contexts, we set a certain percentage P of the trace of a user to be sensitive. We also simulate the cache capability with vehicular users' behaviors.

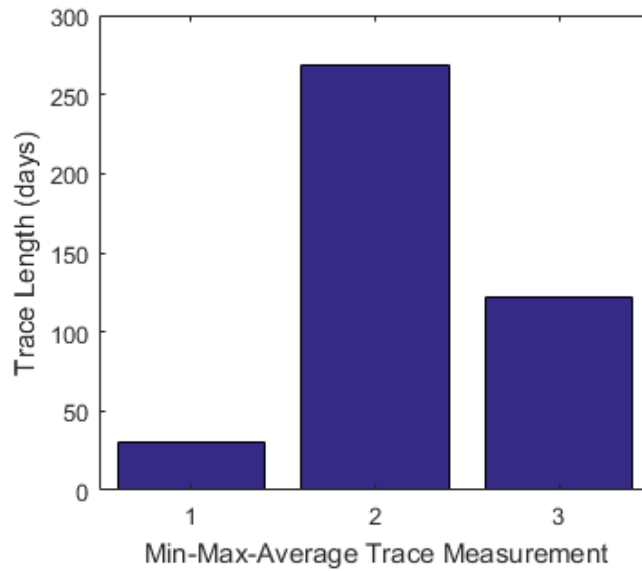


Figure 5.3: The statistical information about the trace length for all users

System parameters: we follow the default system parameters in our simulation. In addition, we set up the percentage of sensitive context P to be 0.4, QoS steepness Θ to be 12, the threshold of satisfaction degree to be 0.8, and the discount factor σ to be 0.7. We set that a user randomly chooses other users to query, and the local buffer can store the returned data with a single parameter.

Competitive work and baseline scheme: we compare the performance of the proposed work with the competitive stochastic scheme in the work [WZ14]. This performance of the query procedures on the dataset is evaluated. The baseline is the context releasing scheme in [TCD⁺10].

5.5.2 Performance Evaluation

- 1) Privacy Gain Measurement and Analysis. We show that the privacy gain is positive along with the increase of the percentage of sensitive context P . The highest point of privacy gain approaches 10% as shown in Figure 5.4.

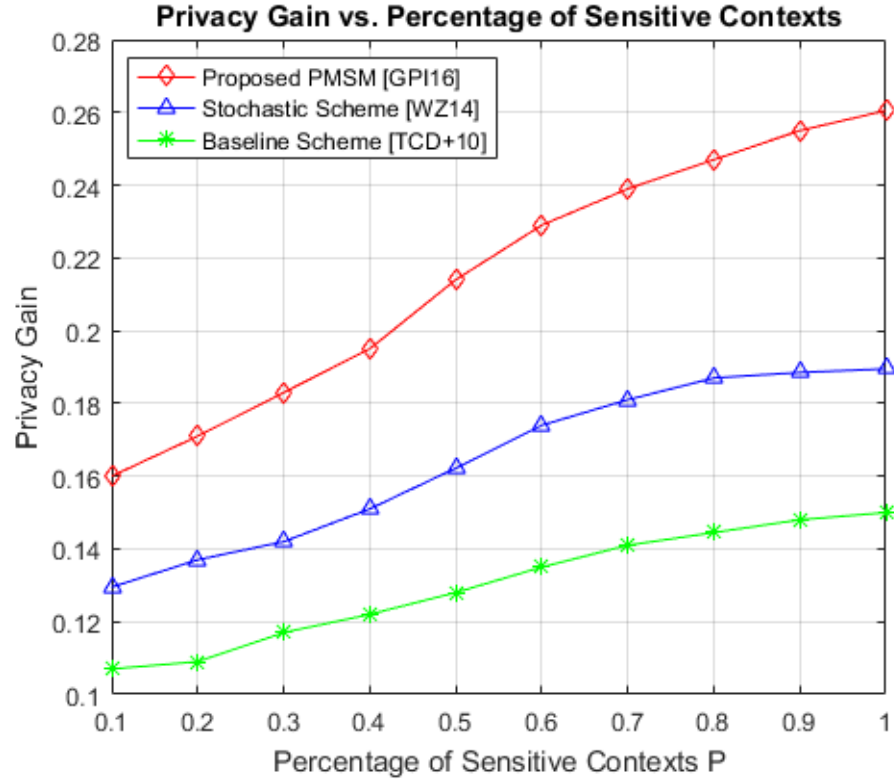


Figure 5.4: Privacy Gain G vs. Percentage of Sensitive Contexts P

At the beginning, sensitive contexts are not released or do not exist, so privacy gain for both the proposed work and the baseline are zero. Nevertheless, the proposed scheme outperforms the baseline with the increase of percentage of sensitive contexts. This is because more cached data is stored in local devices for sensitive contexts, and then the query number to be sent to the remote server is reduced. As a consequence, the server learns less about the vehicular user, and the privacy leakage of the user is reduced. As long as the user's context is sensitive, a better way to protect user privacy against malicious servers is to avoid releasing the query to servers. In fact, the key point here is that we sacrifice the local storage to improve user privacy and maintain service quality.

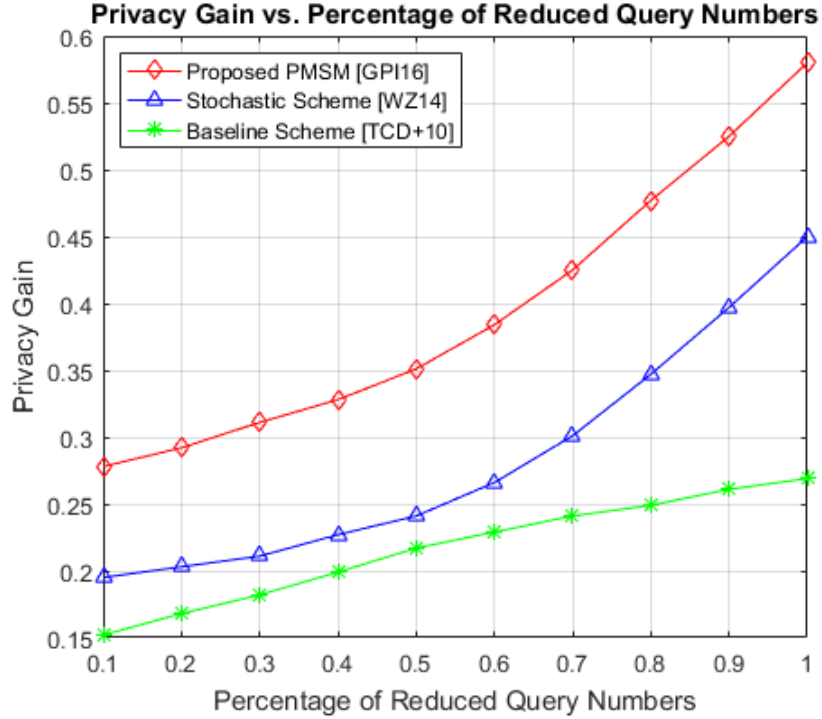


Figure 5.5: Privacy Gain G vs. Percentage of Reduced Query Numbers N

Moreover, we evaluate the privacy gain with respect to the percentage of reduced queries in overall queries sent to remote servers. We measure the privacy level based on the reduced number of queries to be sent to remote servers. The principal is that the less number of queries, the better privacy for mobile vehicular users. The result is shown as Figure 5.5. We can see that when the percentage of the reduced query number increases, the proposed method obtains more privacy gain against the stochastic scheme and the baseline. Even if the stochastic scheme has a trend to have a higher privacy gain with the increase of the reduced query number, it still performs worse than our method about 7.5% to 12.5% percentage. At the same time, our method outperforms the baseline about 15% to 30% percentage. The reason of this phenomenon is the cached data in local or nearby storage can be used to answer the query from the

original user. To decrease the number of queries needed to be sent to the LBS server, the risk of user privacy leakage is greatly reduced. In extreme cases, if there are no queries to be released by the user to the server, the malicious server learns nothing about the user. Thus, the vehicular user privacy is maximized. Here, we assume that user queries can be potentially addressed by the local and/or nearby storage, if the user chooses not to send the queries to the server due to the sensitivity of the query content.

- 2) QoS Measurement and Analysis. We show that along with the increase of the percentage P of sensitive contexts, the service quality based on our proposed solution is higher than the competitive stochastic scheme and the baseline. The maximum advantage of our solution reaches more than 20% than the other two works, as shown in Figure 5.6.

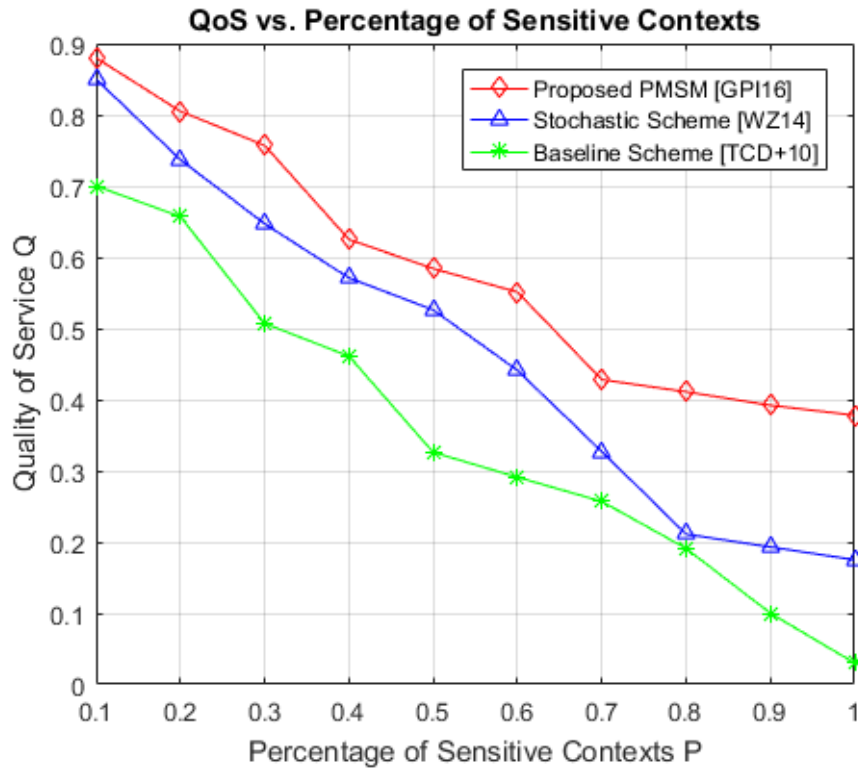


Figure 5.6: Quality of Service Q vs. Percentage of Sensitive Contexts P

From Figure 5.6, we can observe that when sensitive contexts are not released or do not exist, the QoS for the proposed method and the other two methods are all high. However, the proposed scheme beats the stochastic scheme and the baseline scheme along with the increase of percentage of sensitive contexts. The reason behind this phenomenon is that the data cache in local devices stores more related data for more sensitive contexts, thus reducing the quality loss of the query results along with the positive changes of sensitive contexts. On the contrary, the QoS loss L increases when the QoS Q decreases. We see that our approach has a higher QoS compared with the stochastic method and the baseline. Specifically, when all the contexts are sensitive, our approach impressively keeps about 40% of the QoS. The competitive method and the baseline have relative lower QoS scores. Moreover, our method can maintain a much higher QoS when the percentage of sensitive contexts is larger than 0.7. The stochastic scheme and the baseline decrease their performance dramatically after the specific 0.7 point in x-axis. The reason behind this phenomenon is that our method can help users address their queries in a local or nearby way, even if not all queries can be answered.

Furthermore, we measure the QoS score against the percentage of reduced query numbers. We see from Figure 5.7 that our approach has obvious advantages compared to the competitive method and the baseline. The QoS of the baseline decreases substantially and reaches zero when the queries are all reduced, and no queries are sent to remote servers. The stochastic method suffers from the reduced query numbers, and has a lower QoS score than our

proposed approach. We can see that our method still can maintain around 28% of the QoS even if there are no queries to be sent to the service provider. The stochastic scheme maintains about 18% when the user's queries are all reduced, while the baseline loses all the QoS due to the loss of the data service. In addition, we can see that the baseline has a near-linear decreasing trend along with the increase of the percentage of reduced query numbers. The proposed scheme and the stochastic scheme decrease their QoSs slower when N increases. On average, our method performs 10% better than the stochastic scheme, and outperforms 30% of the baseline. With the help of data caching, our approach maintains a better QoS level and shows its effectiveness when the query numbers are reduced.

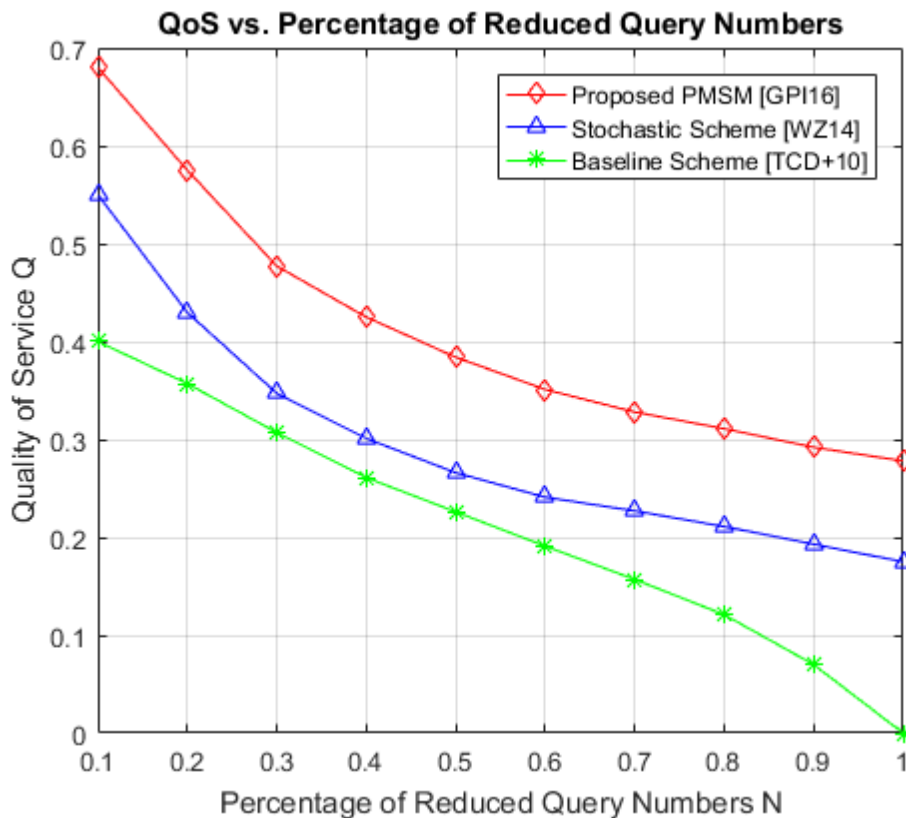


Figure 5.7: Quality of Service Q vs. Percentage of Reduced Query Numbers N

5.6 Summary

In this chapter, we propose a novel mechanism for privacy-aware mobile sensing on mobile vehicles. Multiple sensors with different functionalities on these vehicles are generating large amount of data in real time, consisting of vehicular users' instant contexts. Context-aware vehicular applications and services utilize users' contexts to provide personalized services in return. These collection behaviors can harm user privacy. To address this problem, we develop a privacy-aware mobile sensing mechanism to help vehicular users reduce the number of queries to be sent to the adversarial servers. In this mechanism, mobile vehicular users can selectively query nearby nodes in a peer-to-peer way for privacy protection in vehicular networks. We conduct simulations to demonstrate the efficiency and effectiveness of our approach.

CHAPTER 6

LIMITATIONS, FUTURE WORK AND CONCLUSION

This dissertation has been focused on location trajectory privacy, query and mobile sensing privacy in location-aware applications. For trajectory privacy, a pseudonym-based distributed anonymity zone generation scheme is designed to conceal users' real location trajectories. For query privacy, this dissertation proposes a query-feature-based inference attack model and develops novel randomized algorithms to stop the inference attack. For mobile sensing privacy, a cache-based mechanism is proposed to preserve the privacy of mobile vehicles. This chapter describes the limitations of this research, points out the future work and presents the conclusion to this research.

6.1 Limitations

6.1.1 Pseudonym-based Distributed Anonymity Zone Generation Scheme

We design the pseudonym-based distributed anonymity zone generation scheme to hide users' location trajectories. Based a geometric transformation algorithm, we present the distributed anonymity zone generation process to help users enlarge their anonymity zones. We also enhance the scheme with personalized parameters to enable users to customize their privacy requirements. Information entropy has been used as the main evaluation metric to measure the trajectory privacy level of the proposed method. The simulation results show that our scheme can reduce the attack risk and preserve a high level of trajectory privacy for mobile users in the real world.

There are several limitations of this scheme. First, we generate extra communication overhead to achieve the trajectory level privacy in the distributed anonymity zone

generation process. Second, the distributed anonymity zones are generated in a random manner. This may result in some unreachable anonymity zones in the real world, vulnerable to map-mapping attacks. In addition, we have not considered data de-anonymization techniques as advanced tools for attackers to launch more serious attacks. Data analytics is a rapidly developing field which can be used by the attackers to filter out the anonymity zones and rebuild users' real location trajectories.

6.1.2 Query-Feature-based Inference Attacks and Defense Randomized Algorithms

We propose a new attack model named query-feature-based inference attacks. In this attack model, attackers can analyze the historical query search data and build a probabilistic mapping between a specific query search feature and location probability distribution. This is a strong attack model even without including the location data in user queries. We design the randomized grid selection algorithms to stop the query-feature-based inference attacks. We comprehensively evaluate the algorithms by privacy level, utility analysis and costs of resources. The experiment results based on a real world location-based social network data show the effectiveness of the attack model and the proposed randomized algorithms.

There are also some limitations of this work. First, this work so far has been focused on the model that attackers can infer user location based on query search features. An advanced attack model which we have not considered that can be launched by the attackers is to use query search features to infer users' trajectories. Second, it is still not clear how a group of LBS servers working together can affect our attack model and proposed algorithms. Third, differential privacy is a strong privacy notion that has not been integrated in the work. How to preserve user privacy level in a differential privacy setting

is the next question to ask of our work. This question will challenge our models and algorithms to become more robust against advanced attacks. Finally, we need to be aware of the possible Distributed Denial of Services (DDoS) attacks that are based on the proposed privacy-preserving algorithms.

6.1.3 Cache-based Mobile Sensing Mechanism for Mobile Vehicles

To improve vehicular users' experience and enable mobile sensing process with privacy consideration, we design a cache-based mobile sensing mechanism for mobile vehicles. This work focuses on peer-to-peer collaborations by sharing query cache data to reduce the query numbers to be sent to the adversarial servers. The proposed mechanism allows mobile vehicular users to query local cache data, and/or nearby nodes' cache data, when they are in sensitive contexts. This process breaks the tracking activities from the service providers and reduces the possible privacy leakage, especially when mobile vehicles are in a sensitive context. Two metrics, including privacy level and quality of service level, have been used to evaluate the proposed mechanism on a real-world mobility dataset.

However, this mechanism has several drawbacks. First, this mechanism does not consider the motivations for nearby mobile vehicles to join in the collaborative querying process. A separate reward function or model needs to be developed to motivate the collaboration process. Second, users need to specify their sensitive context (e.g. location) which may not be the case, because users may not prefer to do that or even refuse to do that. In addition, malicious vehicular users, who are willing to participate in the query collaboration process, can also harm the privacy of the user who is sending out the query

request. Furthermore, malicious nearby mobile vehicles can even follow the querying users to learn their sensitive contexts in their moving trajectories, compromising their privacy.

6.2 Future Work

In the future, our research will focus on two main directions. The first main direction is to have a comprehensive understanding of the attack space from advanced attackers. We will investigate new techniques and tools that can be used by attackers to discover vulnerabilities in emerging location-aware applications, as well as new types of possible attacks targeted on these applications. The second main direction is to develop an intelligent, adaptive privacy-preserving and secure framework that can identify and stop possible advanced attacks discovered from our first direction.

6.2.1 A Comprehensive Understanding of the Attack Space from Advanced Attackers

- New types of location-aware applications (e.g. Location-based Augmented Reality) are developing rapidly. We must investigate those emerging applications to have a better understanding about possible vulnerabilities. This is an essential step to discover advanced attacks in the attack spaces.
- Advanced data analysis tools are now available to everyone. Cloud-based online analytic infrastructure (e.g. Microsoft Azure, AWS, etc.) can be even more powerful for attackers to find hidden relationships between users' query behaviors and users' location trajectories through analyzing large scale mobility datasets.
- Crowd-sourcing is another powerful way that can be used by attackers to collect side information about the targets that they are interested in. Attackers may even

pay certain rewards to the participants for collecting specific location information related to the target and use that information to conduct inference attacks.

- Data de-anonymization techniques can link two or more online open datasets which share certain attributes and user information. Data de-anonymization is a strategy to re-identify the anonymous data sources such as individual preferences, query search data, transaction records and so on [NS08]. The attackers are highly possible to collect different datasets from online sources or open APIs to identify the target users, discovering their sensitive information using data de-anonymization techniques.
- Location-aware applications can share their users' profiles with other location-aware applications. This sharing behavior can provide users more convenience when they are translating between different applications. However, this behavior poses a great privacy challenge to users. More personal information becomes available to those applications while less user privacy remains.
- Sensor-enabled Internet of Things is drawing more attention due to its popularity. This trend poses a whole new family of privacy challenges. Sensors are connected to the internet and their privacy can be compromised by attackers from anywhere in the world at any time.
- Mobile sensors that are carried by humans or mobile vehicles may be followed by the attackers based on their signal strength in the physical world. This type of chasing attack can be effective when attackers assign a group of observers around the targets.

- DDoS attacks are reported quite often in the cyber space. We need to assess the possible DDoS attacks and design practical policies when using privacy-preserving mechanisms to protect user privacy.

6.2.2 An Intelligent and Adaptive Privacy-Preserving and Secure Framework

- From the attack space, we see different tools and techniques that can be used to compromise user privacy. Thus, a strong privacy notion is needed to ensure user privacy whenever a new user query record is logged on the adversarial server. Differential privacy applies to this protection space due to its strong mathematical property for user privacy guarantee from a statistical point of view.
- We plan to combine location, trajectory and query privacy protection into the privacy-preserving and secure framework. A layer-wise approach is appropriate to deal with different attacks. For example, the first layer can generate distributed anonymity zones to satisfy users' anonymity zone requirement; the second layer can adjust and select the anonymity zones with close probabilities with respect to the zone where the user is locating; the final layer can apply different privacy metrics, such as differential privacy, to evaluate the effectiveness of the framework in a pre-defined privacy setting.
- In the privacy-preserving and secure framework, each layer should be a modular design so that different privacy-preserving algorithms can be tested in each layer without affecting other layers.

- Energy-efficient models and algorithms should be designed to optimize the computation and communication costs in the energy-limited mobile devices and/or mobile vehicles.
- In our future design, we need to consider security issues and techniques in the framework due to the high vulnerabilities of mobile sensors in an open environment. Encryption is highly needed to protect critical parts of a query, such as the user identity.
- For location and trajectory query streams, our framework should be able to deal with query updates with different frequencies, and adaptively change the underlying algorithm parameters to preserve the required user privacy level.
- Furthermore, localization problems can also be considered together within the privacy preservation framework. The framework can consider the existing localization errors as a natural way to mislead the adversaries. By integrating localization errors in the framework, we reduce the workload on privacy-preserving mechanisms and algorithms.
- Finally, our future framework should apply dynamic and evolutionary game theoretic approach to model the interaction between users and attackers, and derive optimal strategies to preserve user privacy with minimal energy consumptions.

6.3 Conclusion

This research spans from the realization that user privacy is lost along with the popular usage of location-aware applications. The outcome seeks to relive user privacy problems in the emerging location-aware applications and services. First, we develop a pseudonym-

based anonymity zone generation scheme against a strong adversary model in continuous LBS. Based on a geometric transformation algorithm, the scheme generates distributed anonymity zone with personalized parameters to conceal users' real location trajectories. Second, based on the analysis of historical query search data, we introduce a query-feature-based probabilistic inference attack, and propose query-aware randomized algorithms to preserve user privacy by distorting the probabilistic inference conducted by adversaries. Finally, we develop a privacy-aware mobile sensing mechanism to help vehicular users reduce the number of queries to be sent to the adversarial servers. In this mechanism, mobile vehicular users can selectively query nearby nodes in a peer-to-peer way for privacy protection in vehicular networks.

BIBLIOGRAPHY

- [AB18] O. Abul and C. Bayrak. From Location to Location Pattern Privacy in Location-based Services. *Journal of Knowledge Information System*, 08 January, 2018.
- [ABCP13] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential Privacy for Location-based Systems. In *Proceeding of CCS '13 Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp 901-914, November, 2013.
- [ACVS11] C. A. Ardagna, M. Cremonini, S. Vimercati, and P. Samarati. An Obfuscation-based Approach for Protecting Location Privacy. *IEEE Transactions on Dependable and Secure Computing*, Vol. 8, Issue 1, pp 13–27, 2011.
- [AJSS13] C. A. Ardagna, S. Jajodia, P. Samarati, and A. Stavrou. Providing Users' Anonymity in Mobile Hybrid Networks. *ACM Transactions on Internet Technology*, Vol. 12, Issue 3, Article 7, 33 pages, May 2013.
- [BKE13] W. J. Buchanan, Z. Kwecka, and E. Ekonomou. A Privacy Preserving Method using Privacy Enhancing Techniques for Location Based Services. *Mobile Networks and Applications*, Vol. 18, Issue 5, pp. 728–737, Oct. 2013.
- [CEP17] K. Chatzikokolakis, E. ElSalamouny, and C. Palamidessi. Efficient Utility Improvement of Location Privacy. In *Proceedings on Privacy Enhancing Technologies*, Vol. 4, pp. 308–328, 2017.
- [CMBL11] C.Y. Chow, M. F. Mokbel, J. Bao, and X. Liu. Query-Aware Location Anonymization for Road Networks. In *Journal of Geoinformatica*, Vol. 15 Issue 3, pp. 571-607. ACM, 2011.
- [CML11] C.Y. Chow, M. F. Mokbel, and X. Liu. Spatial Cloaking for Anonymous Location Based Services in Mobile Peer-to-peer Environments. *Geoinformatica*, Vol. 15, Issue 2, pp. 351–380, April 2011.
- [CRJS13] S. Chakraborty, K. R. Raghavan, M. P. Johnson, and M. B. Srivastava. A Framework for Context-aware Privacy of Sensor Data on Mobile Systems. *Hotmobile*, 2013.

- [DACA16] T. Dargahi, M. Ambrosin, M. Conti, and N. Asokan. ABAKA: A Novel Attribute-based k-anonymous Collaborative Solution for LBSs. *Elsevier Computer Communications*, pp 1- 13, 2016.
- [DBS08] M. Damiani, E. Bertino, and C. Silvestri. PROBE: An Obfuscation System for the Protection of Sensitive Location Information in LBS. *Purdue Technical Report TR2001-145*, CERIAS, 2008.
- [DNZG17] J. Ding, C.-C. Ni, M. Zhou, and J. Gao. MinHash Hierarchy for Privacy Preserving Trajectory Sensing and Query. In *Proceedings of The 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 12 pages, Pittsburgh, PA, April 2017.
- [DRK16] S. Derikx, M. Reuver, and M. Kroesen. Can Privacy Concerns for Insurance of Connected Cars be compensated? *Electronic Markets*, Vol 26, Issue 1, pp. 73-81. February 2016.
- [DS10] M. R. Doomun and K. M. S. Soyjaudah. Route Extrapolation for Source and Destination Camouflage in Wireless Ad Hoc Networks. *IEEE International Conference on Computer Communications Networks (ICCCN'10)*. IEEE Press, 1–7, 2010.
- [EG16] E. ElSalamouny and S. Gambs. Differential Privacy Models for Location Based Services. *Transactions on Data Privacy*. Vol 9, pp15-48, 2016.
- [EGC⁺10] W. Enck, P. Gilbert, B.G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones. In *OSDI'10 Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, pp. 393-407, 2010.
- [EPL09] N. Eagle, A. Pentland, and D. Lazer. Inferring Social Network Structure using Mobile Phone Data. In *PNAS*, Volume 106, No. 36, 2009.
- [FLZ12] Y. Feng, P. Liu, and J. Zhang. A Mobile Terminal Based Trajectory Preserving Strategy for Continuous Querying LBS Users. In *IEEE 8th International Conference on Distributed Computing in Sensor Systems (DCOSS'12)*. IEEE Press, 92–98, 2012.
- [FMBH10] J. Freudiger, M. H. Manshaei, J.Y. L. Boudec, and J.P. Hubaux. On the Age of Pseudonyms in Mobile Ad Hoc Networks. In *Proceedings of the 29th Conference on Information Communications (INFOCOM'10)*. IEEE Press, 1577–1585, 2010.

- [Fou18] Foursquare.com. <https://www.foursquare.com/>. Accessed 01/10/2018.
- [FSH09] J. Freudiger, R. Shokri, and J. Hubaux. On the Optimal Placement of Mix Zones. In *Proceedings of PETS*, 2009.
- [GBP⁺18] M. Guo, K. G. Boroojeni, N. Pissinou, K. Makki, J. Miller, and S.S. Iyengar. Query-Aware User Privacy Protection for LBS over Query-Feature-based Inference Attacks. *IEEE ISCC*, 2018.
- [GM18] Google Maps. <https://www.google.com/maps/>. Accessed 01/10/2018.
- [GG03] M. Gruteser and D. Grunwald. Anonymous Usage of Location-Based Services through Spatial and Temporal Cloaking. In *ACM Mobisys*, 2003.
- [Ghi09] G. Ghinita. Private Queries and Trajectory Anonymization: A Dual Perspective on Location Privacy. *Transactions on Data Privacy*, Vol. 2, Issue 1, pp 3–19, April 2009.
- [GL08] B. Gedik and L. Liu. 2008. Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms. *IEEE Transactions on Mobile Computing* 7, 1 (Jan. 2008), 1–18.
- [GJP⁺15] M. Guo, X. Jin, N. Pissinou, S. Zanlongo, B. Carbutar, and S.S. Iyengar. In-Network Trajectory Privacy Preservation. *Journal of ACM Computing Surveys*, Volume 48, Issue 2, No. 23, 2015.
- [GKK⁺08] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.L. Tan. Private Queries in Location Based Services: Anonymizers Are Not Necessary. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD'08)*. ACM, New York, NY, 121–132.
- [GKS07] G. Ghinita, P. Kalnis, and S. Skiadopoulos. MOBIHIDE: A Mobile Peer-to-peer System for Anonymous Location-based Queries. In *Proceedings of the 10th International Conference on Advances in Spatial and Temporal Databases (SSTD'07)*. Springer-Verlag, Berlin, 221–238.
- [GMS13] S. Gao, J. Ma, and W. Shi. TrPF: A Trajectory Privacy-Preserving Framework for Participatory Sensing. *IEEE Transactions on Information Forensics and Security*, Volume 8, No. 6, 2013.

- [GNG12] M. Gotz, S. Nath, and J. Gehrke. MaskIt: Privately Releasing User Context Streams for Personalized Mobile Applications. *ACM SIGMOD, 2012*.
- [GPI15] M. Guo, N. Pissinou, and S.S. Iyengar. Pseudonym-based Anonymity Zone Generation for Mobile Service with Strong Adversary Model. *IEEE CCNC, Security, Privacy and Content Protection*. January 9-12, Las Vegas, Nevada, 2015.
- [GPI16] M. Guo, N. Pissinou, and S.S. Iyengar. “Privacy-Aware Mobile Sensing in Vehicular Networks”, *IEEE ICNC Conference, Communications and Information Security*. February 15-18, Kauai, Hawaii, 2016.
- [GSX10] Z. Gong, G. Sun, and X. Xie. 2010. Protecting Privacy in Location-based Services using K-Anonymity without Cloaked Region. In *2010 11th International Conference on Mobile Data Management (MDM'10)*. IEEE Press, 366–371.
- [HHC12] R.H. Hwang, Y.L. Hsueh, and H.W. Chung. A Novel Time-obfuscated Algorithm for Trajectory Privacy. In *Proceedings of the 2012 12th International Symposium on Pervasive Systems, Algorithms and Networks (I-SPAN'12)*. IEEE Computer Society, 208–215, 2012.
- [HLD10] J. Hao, W. Liu, and Y. Dai. A Controllable Privacy Protection Framework in Position Based Routing for Suspicious MANETs. *IET International Conference on Wireless Sensor Network (IETWSN'10)*. 291–296, 2010.
- [HTXZ18] J. Hua, W. Tong, F. Xu, and S. Zhong. A Geo-Indistinguishable Location Perturbation Mechanism for Location-Based Services Supporting Frequent Queries. *IEEE Transactions on Information Forensics and Security*, Vol. 13, No. 5, May 2018.
- [Ins16] How Do Those Car Insurance Tracking Devices Work? <https://cars.usnews.com/cars-trucks/best-cars-blog/2016/10/how-do-those-car-insurance-tracking-devices-work>. Accessed 1/5/2018.
- [JPC⁺12] X. Jin, N. Pissinou, C. Chesneau, S. Pumpichet, and D. Pan. Hiding Trajectory on the Fly. In *Proceedings of the IEEE International Conference on Communications (ICC'12)*. IEEE Press, 403–407, 2012.

- [JZ13] J. Jia and F. Zhang. Twice Anonymity Algorithm for LBS in Mobile P2P Environment. *Journal of Computational Information Systems*, Vol. 9, Issue 9, pp. 3715–3722, 2013.
- [KSS11] A. Khoshgozaran, C. Shahabi, and H. Shirani-Mehr. Location Privacy: Going beyond K-anonymity, Cloaking and Anonymizers. *Knowledge and Information Systems* 26, 3 (March 2011), 435–465.
- [LC14] Z. Li and C. Chigan. On Joint Privacy and Reputation Assurance for Vehicular Ad Hoc Networks. In *IEEE Transactions on Mobile Computing*, Vol. 13, Issue 10, 2014.
- [LLAM13] X. Liu, Y. Liu, K. Aberer, and C. Miao. Personalized Point-of-Interest Recommendation by Mining Users’ Preference Transition. In *Proceedings of the 22nd ACM CIKM*, pp. 733-738. ACM, 2013.
- [LLLS10] X. Liang, R. Lu, X. Lin, and X. Shen. Message Authentication with Nontransferability for Location Privacy in Mobile Ad Hoc Networks. In *2010 IEEE Global Telecommunications Conference (GLOBECOM’10)*. IEEE Press, 1–5, 2010.
- [LLG⁺13] X. Liu, K. Liu, L. Guo, X. Li, and Y. Fang. A Game-Theoretic Approach for Achieving k-Anonymity in Location-Based Services. In *Proceedings of IEEE INFOCOM*, IEEE Press, 2985–2993, 2013.
- [LLV07] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *IEEE 23rd International Conference on Data Engineering (ICDE)*. April 15-20, 2007.
- [LSTL13] M. Li, S. Salinas, A. Thapa, and P. Li. n-CD: A Geometric Approach to Preserving Location Privacy in Location-Based Services. In *Proceedings of IEEE INFOCOM*, April, 2013.
- [LZP⁺12] X. Liu, H. Zhao, M. Pan, H. Yue, X. Li, and Y. Fang. Traffic-aware multiple mix zone placement for protecting location privacy. In *2012 Proceedings of IEEE INFOCOM. IEEE*, pp. 972–980, 2012.
- [MCA06] M. F. Mokbel, CY Chow, and W.G. Aref. The New Casper: Query Processing for Location Services without Compromising Privacy. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB’06)*. ACM, New York, NY, 763–774.
- [MHP17] Z. Montazeri, A. Houmansadr, H. Pishro-Nik. Achieving Perfect Location Privacy in Wireless Devices Using Anonymization. *IEEE Transactions on Information Forensics and Security*, Vol. 12, No. 11, November 2017.

- [Mic17] Microsoft. Location Based Services Usage and Perceptions Survey. <https://www.microsoft.com/en-us/download/details.aspx?id=3250>. Accessed 10/2/2017.
- [MJ11] A. Masoumzadeh and J. B. D. Joshi. An Alternative Approach to k-anonymity for Location Based Services. *Procedia Computer Science* 5 (2011), 522–530.
- [MKGv07] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. Vol. 1 Issue 1, Article No. 3, March 2007.
- [MLSK11] L. Ma, J. Liu, L. Sun, and O. B. Karimi. The Trajectory Exposure Problem in Location-aware Mobile Networking. In *Proceedings of the 2011 IEEE 8th International Conference on Mobile Ad-Hoc and Sensor Systems (MASS'11)*. IEEE Computer Society, 7–12, 2011.
- [Mon17] Heartbeat Sensor using Arduino (Heart Rate Monitor). <https://www.electronicshub.org/heartbeat-sensor-using-arduino-heart-rate-monitor/>. Accessed 01/10/2018.
- [NLZ⁺14] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li. Achieving k-anonymity in Privacy-Aware Location-Based Services. In *Proceedings of IEEE INFOCOM*, 2014.
- [NOS12] D. Nussbaum, M. T. Omran, and J. Sack. 2012. Techniques to Protect Privacy against Inference Attacks in Location Based Services. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS'12)*. ACM, New York, NY, 58–67.
- [NR09] E. C. H. Ngai and I. Rodhe. On Providing Location Privacy for Mobile Sinks in Wireless Sensor Networks. In *Proceedings of the 12th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM'09)*. ACM, New York, NY, 116–123, 2009.
- [NS08] Arvind Narayanan and Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *IEEE Symposium on Security and Privacy*. May, 2008.
- [PKYB14] R. Paulet, M. G. Kaosar, X. Yi, and E. Bertino. Privacy-Preserving and Content-Protecting Location Based Queries. In *IEEE Transactions on Knowledge and Data Engineering*. Vol. 26, No. 5, May 2014.

- [PL11] B. Palanisamy and L. Liu. 2011. MobiMix: Protecting location Privacy with Mix-zones over Road Networks. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE'11)*. IEEE Computer Society, 494–505, 2011.
- [PL15] B. Palanisamy and L. Liu. Attack-resilient Mix-zones over Road Networks: Architecture and Algorithms. *IEEE Transactions on Mobile Computing*, vol. 14, no. 3, pp. 495-508, 2015.
- [PLW17] T. Peng, Q. Liu, and G. Wang. Enhanced Location Privacy Preserving Scheme in Location-Based Services. *IEEE System Journal*. Vol. 11, Issue 1, pp. 219-230, 2017.
- [PR08] N. Poolsappasit and I. Ray. Towards A Scalable Model for Location Privacy. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS (SPRINGL'08)*. ACM, New York, NY.
- [PR09] N. Poolsappasit and I. Ray. Towards Achieving Personalized Privacy for Location-based Services. *Transactions on Data Privacy*, Vol. 2, Issue 1, pp. 77–99, April 2009.
- [PYZ⁺09] A. Pingley, W. Yu, N. Zhang, X. Fu, and W. Zhao. CAP: A Context-aware Privacy Protection System for Location-based Services. In *Proceedings of the 29th IEEE International Conference on Distributed Computing Systems (ICDCS'09)*. IEEE Press, 49–57, 2009.
- [RB12] D. Riboni and C. Bettini. Private Context-aware Recommendation of Points of Interest: An Initial Investigation. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE Press, 584–589, 2012.
- [RCL15] R. Rios, J. Cuellar, J. Lopez. Probabilistic Receiver-location Privacy Protection in Wireless Sensor Networks. In *Elsevier Journal of Information Sciences*. 2015.
- [RMD04] Reality Mining Dataset. <http://reality.media.mit.edu/dataset.php>. Accessed 09/10/2015.
- [SCH⁺16] Y. Sun, M. Chen, L. Hu, Y. Qian, and M. Hassan. ASA: Against Statistical Attacks for Privacy-aware Users in Location-based Service. In *Future Generation Computer Systems*, 2016.

- [SCHW15] R. Schlegel, C.Y. Chow, Q. Huang, and D. S. Wong. User-Defined Privacy Grid System for Continuous Location-Based Services. *IEEE Transactions on Mobile Computing*, 2015.
- [SIC⁺10] A. Suzuki, M. Iwata, Y. Arase, T. Hara, X. Xie, and S. Nishio. A User Location Anonymization Method for Location Based Services in a Real Environment. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS'10)*. ACM, New York, NY, 398–401.
- [SPTH09] R. Shokri, P. Pedarsani, G. Theodorakopoulos, and J.P. Hubaux. Preserving Privacy in Collaborative Filtering through Distributed Aggregation of Offline Profiles. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys'09)*. ACM, New York, NY, 157–164, 2009.
- [SPTH11] R. Shokri, P. Papadimitratos, G. Theodorakopoulos, and J.P. Hubaux. Collaborative Location Privacy. In *Proceedings of the IEEE 8th International Conference on Mobile Ad-Hoc and Sensor Systems (MASS'11)*. IEEE Press, 500–509, 2011.
- [STP⁺14] R. Shokri, G. Theodorakopoulos, P. Papadimitratos, E. Kazemi, and J.P. Hubaux. Hiding in the Mobile Crowd: Location Privacy through Collaboration. *IEEE Transactions on Dependable and Secure Computing, Special Issues on "Security and Privacy in Mobile Platforms"*, 2014.
- [SVAC10] H. Shin, J. Vaidya, V. Atluri, and S. Choi. 2010. Ensuring Privacy and Security for LBS through Trajectory Partitioning. In *Proceedings of the 2010 11th International Conference on Mobile Data Management (MDM'10)*. IEEE Computer Society, 224–226.
- [Swe02] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.
- [TCD⁺10] E. Toch, J. Cranshaw, P. H. Drielsma, J. Y. Tsai, P. G. Kelley, J. Springfield, L. Cranor, J. Hong, and N. Sadeh. Empirical Models of Privacy in Location Sharing. In *Proceedings of ACM Ubicomp*, pp. 129-138, 2010.
- [Tra14] Car insurance companies want to track your every move—and you're going to let them. <https://qz.com/230055/car-insurance-companies-want-to-track-your-every-move-and-youre-going-to-let-them/>. Accessed 1/10/2018.

- [TZX⁺17] Z. Tu, K. Zhao, F. Xu, Y. Li, L. Su, and D. Jin. Beyond K-Anonymity: Protect Your Trajectory from Semantic Attack. *The 10th Annual IEEE International Conference on Sensing, Communications and Networking (SECON'17)*. 12-14 June 2017.
- [WDR13] M. Wernke, F. DuRr, and K. Rothermel. PShare: Ensuring location privacy in non-trusted systems through multi-secret sharing. *Pervasive and Mobile Computing*. Vol. 9, Issue 3 (June 2013), 339–352.
- [WLY⁺17] J. Wang, Y. Li, D. Yang, H. Gao, G. Luo, and J. Li. Achieving Effective k-Anonymity for Query Privacy in Location-based Services. *IEEE Access*, Volume 4, 2017.
- [WXH⁺12] Y. Wang, D. Xu, X. He, C. Zhang, F. Li, and B. Xu. L2P2: Location-aware Location Privacy Protection for Location-based Services. In *Proceedings of the 29th Conference on Information Communications (INFOCOM'12)*. IEEE Press, 1996–2004.
- [WZ14] W. Wang and Q. Zhang. A Stochastic Game for Privacy Preserving Context Sensing on Mobile Phone. In *Proceedings of IEEE INFOCOM*, 2014.
- [XC09] T. Xu and Y. Cai. Location Safety Protection in Ad Hoc Networks. *Journal of Ad Hoc Networks*, Vol. 7, Issue 8 (Nov. 2009), 1551–1562, 2009.
- [XZZ⁺13] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu. Destination Prediction by Sub-Trajectory Synthesis and Privacy Protection against Such Prediction. In *Proceedings of the 2013 IEEE International Conference on Data Engineering (ICDE'13)*. IEEE Computer Society, 254–265.
- [Yao10] Jianbo Yao. Preserving Mobile-sink-location Privacy in Wireless Sensor Networks. In *2nd International Workshop on Database Technology and Applications*. IEEE Press, 1–3, 2010.
- [YJHL08] M. L. Yiu, C. S. Jensen, X. Huang, and H. Lu. 2008. SpaceTwist: Managing the Tradeoffs among Location Privacy, Query Performance, and Query Accuracy in Mobile Services. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering (ICDE'08)*. IEEE Computer Society, Washington, DC, USA, 366–375.
- [YKH⁺16] R. Yu, J. Kang, X. Huang, S. Xie, Y. Zhang, and S. Gjessing. MixGroup: Accumulative Pseudonym Exchanging for Location

- Privacy Preservation in Vehicular Social Networks. In *IEEE Transactions on Dependable and Secure Computing*, Vol. 13, No. 1, Jan. 2016.
- [YPBV16] X. Yi, R. Paulet, E. Bertino, and V. Varadharajan. Practical Approximate k Nearest Neighbor Queries with Location and Query Privacy. In *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 6, June 2016.
- [ZC11] Z. Zhu and G. Cao. APPLAUS: A Privacy-preserving Location Proof Updating System for Location-based Services. In *Proceedings of the 29th Conference on Information Communications (INFOCOM'11)*. IEEE Press, 1889–1897, 2011.
- [ZCN⁺13] X. Zhu, H. Chi, B. Niu, W. Zhang, Z. Li, and H. Li. MobiCache: When k-anonymity meets Cache. *IEEE Global Communications Conference (Globecom)*, 2013.
- [ZKMC13] J. Zhu, K. Kim, P. Mohapatra, and P. Congdon. An Adaptive Privacy-preserving Scheme for Location Tracking of a Mobile User. In *Proceedings of 10th Annual IEEE SECON Conference*, pp 140–148, 2013.
- [ZSXP15] Z. Zhang, Y. Sun, X. Xie, and H. Pan. An Efficient Method on Trajectory Privacy Preservation. *International Conference on Big Data Computing and Communications*, BigCom 2015, pp 230–240, 2015.
- [ZSZZ15] R. Zhang, J. Sun, Y. Zhang, and C. Zhang. Secure Spatial Top-k Query Processing via Untrusted Location-based Service Providers. *IEEE Transactions on Dependable and Secure Computing*, Vol. 12, No. 1, pp. 111-124, 2015.

VITA

MINGMING GUO

Born, Xingtai, Hebei, China

- 2009 B.S., Computer Science
North China Institute of Science and Technology
Beijing, China
- 2014 M.S., Computer Science
Florida International University
Miami, Florida
- 2018 Doctoral Candidate, Computer Science
Florida International University
Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Guo, M., Boroojeni, K. G., Pissinou, N., Makki, K., Miller, J., Iyengar, S.S. “*Query-Aware User Privacy Protection for LBS over Query-Feature-based Attacks*”, IEEE Symposium on Computers and Communications (ISCC 18). June 25-28, Natal, Brazil 2018.

Guo, M., Pissinou, N., Iyengar, S.S. “*Privacy-Aware Mobile Sensing in Vehicular Networks*”, IEEE International Conference on Computing, Networking and Communications (ICNC 16), Communications and Information Security. February 15-18, Kauai, Hawaii 2016.

Guo, M., Jin, X., Pissinou, N., Zanlongo, S., Carbutar, B., Iyengar, S.S. “*In-Network Trajectory Privacy Preservation*”, Journal of ACM CSUR, Volume 48 Issue 2, Article No. 23, November 2015.

Guo, M., Pissinou, N., Iyengar, S.S. “*Pseudonym-based Anonymity Zone Generation for Mobile Service with Strong Adversary Model*”, IEEE Consumer Communications and Networking Conference (CCNC 15), Security, Privacy and Content Protection. January 9-12, Las Vegas, Nevada 2015.