FINAL REPORT
TO THE
CENTER FOR MULTIMODAL SOLUTIONS FOR CONGESTION MITIGATION (CMS)

CMS PROJECT NUMBER: 2011-009
CMS PROJECT TITLE: PRIVACY-PRESERVING METHODS TO RETRIEVE ORIGIN-DESTINATION INFORMATION FROM CONNECTED VEHICLES

FOR PERIOD APRIL 2011 TO DECEMBER 2012

FROM MAHMOOD ZANGUI, YIAN ZHOU, YAFENG YIN* AND SHIGANG CHEN
*DEPARTMENT OF CIVIL AND COASTAL ENGINEERING
UNIVERSITY OF FLORIDA
EMAIL: YAFENG@CE.UFL.EDU (YAFENG YIN)

January 23, 2013

Disclaimer and Acknowledgment

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

This work was sponsored by a grant from the Center for Multimodal Solutions for Congestion Mitigation, a U.S. DOT Tier-1 grant-funded University Transportation Center.

Contents

List of Tables

List of Figures

Abstract

This report investigates technical approaches to address privacy concerns associated with two innovative applications enabled by connected vehicle systems, i.e., origin-destination (OD) flow measurement and differentiated congestion pricing. The former is to retrieve the OD information from connected vehicles while the latter charges congestion tolls with respect to travel characteristics of connected vehicles, e.g., origins, destinations or paths that they traverse between their origins and destinations. Since both applications require tracking vehicles, they may violate the "anonymity by design" principle adopted by connected vehicle systems. For OD flow measurement, a novel measurement scheme is developed to collect aggregate OD flow data without compromising motorists' privacy. For differentiated congestion pricing, an incentive program is designed to encourage motorists to voluntarily reveal their private information and create a win-win situation for both motorists and the society.

Executive Summary

Because of concerns regarding the privacy of motorists, the current "anonymity by design" principle adopted by connected vehicle systems does not allow vehicles to be tracked over a long distance, and thus may restrict the development of innovative applications otherwise enabled by connected vehicle technologies. This report examines two such applications, i.e., origin-destination (OD) flow measurement and differentiated congestion pricing, and develops technical approaches to address privacy concerns associated with them.

The OD data provides information on flows of vehicles traveling from one specific geographical area to another. It plays a crucial role in planning, design and management of transportation systems. The advent of connected-vehicle systems provides the potential for a fundamental shift in the way how OD data are collected. Under the currently envisioned implementation, connected vehicles routinely communicate with each other and the roadside equipments (RSEs) in real time via, e.g., Dedicated Short Range Communications (DSRC). In principle, whenever a connected vehicle passes by an RSE, it can transmit to the RSE its unique identification number (ID). Subsequently, the OD demand pattern of a network can be obtained by comparing those IDs stored in the RSEs across the network. Unfortunately, such a straightforward application is not supported by the "anonymity by design" principle, which must ensure anonymity and untraceability to protect motorists privacy.

The objective of our work is to allow transportation authorities to collect aggregate OD flow data without learning information about individual vehicles. In our scheme, vehicle IDs should be preprocessed and protected by keys before transmission. In other words, RSEs will only be able to collect Keyed signatures of vehicles' IDs (referred to as KIDs). To compute the size of an OD flow between two locations based on KIDs, we introduce a family of commutative one-way hash functions. This family of hash functions, as its name suggests, has the unique properties of both commutativity and one-wayness. One crucial benefit of utilizing this hash function family is that vehicles can transmit their KIDs by hashing their IDs under totally different keys and be sure that no one is able to get their IDs, even knowing the keys (because of one-wayness), while it still allows us to compute the OD flow size as demanded (through commutativity). We further adopt statistical techniques and use sampling to achieve better efficiency without significantly degrading measurement accuracy. We perform simulations to demonstrate the feasibility and scalability of our scheme.

The second application this report investigates is differentiated congestion pricing, an innovative market-based instrument for traffic management and congestion mitigation. Connected vehicle systems are capable of tracking vehicles in real time and thus provide opportunities of charging vehicles with respect to the paths they traverse, their origins or destinations. In contrast, the literature on congestion pricing primarily focuses on anonymous link tolls. We perform computational experiments to compare differentiated pricing with

traditional link-based anonymous tolls. The experiments show that in a first-best network condition where all the links are tollable, differentiated pricing can substantially reduce motorists' financial burden; in a second-best environment where only some links are tollable, it helps to achieve a lower level of congestion.

One of the major implementation difficulties for differentiated pricing is potential violation of motorists' location privacy. The traditional way of manually collecting toll preserves location privacy almost completely. Electronic toll collection (ETC) systems have been built to make toll collection more efficient, but the way they currently operate may compromise motorists' privacy rights. The systems often link motorists' accounts and record locations and times of transactions. If toll gantries are ubiquitous, the recorded transaction information may impinge on the privacy rights of motorists. However, those who are concerned about their location privacy have the option to pay the toll by cash and avoid risk of privacy disclosure. Moreover, for anonymous link-based tolling, it is possible to design a privacy-preserving ETC system. Unfortunately, it is difficult, if not impossible, to design a privacy-preserving differentiated pricing system, because the system requires the knowledge of motorists' trip characteristics such as path.

There have been some indications that motorists, some at a price, are willing to provide private information with the understanding that it will not be published and/or misused. Recognizing that some may benefit from differentiated schemes while others with higher value of privacy may be better off under anonymous tolling, we develop an incentive program for travelers to opt in to differentiated pricing. More specifically, a hybrid of anonymous and differentiated pricing schemes will be implemented on the network. Travelers who choose to reveal their private information will pay differentiated tolls while those who remain anonymous will pay uniform tolls. Since travel costs (time plus toll) in differentiated schemes are generally less than those in the anonymous scheme, the cost savings can be viewed as incentives for drivers to participate in differentiated pricing. Although other incentives, such as subsidies or credits, can be provided, this report focuses on designing anonymous and differentiated tolls in the hybrid scheme and allowing for the cost savings as incentives. The overarching goal of this hybrid scheme is to create a win-win situation for both users and society.

Chapter 1

Introduction

1.1 Background

The advent of connected vehicle systems provides the potential for a fundamental shift in the way how traffic systems can be managed and operated. Connected vehicle (formerly known as IntelliDrive or Vehicle-Infrastructure Integration) is an initiative from US Department of Transportation (USDOT) to combine cutting edge technologies, including advanced wireless communications, on-board computer processing, advanced vehicle-sensors, GPS navigation and others, to produce safety, mobility and environmental benefits. USDOT envisions a nationwide system in which connected vehicles routinely communicate with each other and the roadside equipments (RSEs) in real time via, e.g., Dedicated Short Range Communications (DSRC). Major connected-vehicle testbeds have been initiated in the states of California, Michigan, and Arizona [7]. USDOT intends to decide in 2013 whether connected vehicle applications show enough promise to merit a nationwide deployment [56].

Privacy protection has been an important issue since the onset of Vehicle-Infrastructure Integration or VII. There was considerable concern that the program would substantially compromise the privacy of drivers, vehicle owners or passengers, since they may be tracked along their driving paths or detected violating traffic regulations and law. In 2007, the National VII Coalition established a VII privacy policies framework that defines nine privacy principles to govern basic privacy protections in the program [29]. Currently, connected vehicle systems require "anonymity by design" for privacy protection. In this approach, the personal identifiable information is not collected or revealed, in contrast to traditional mobile data collection practices where data are collected first and then processed to prevent the release of personal identifiable information [9]. To ensure trusted communications between parties, connected vehicles are issued a bundle of anonymous certificates and can use any of them, for a pre-defined period, to establish trusted anonymous communication. The certificate bundle will be refreshed frequently. Since no personally identifiable information is attached to messages being transmitted, the anonymity is largely maintained. However, the "anonymity by design" approach may hinder the development of innovative applications that would be otherwise enabled by connected vehicle technologies, e.g., origin-destination (OD) flow collection. In principle, whenever a connected vehicle passes by an RSE, it can transmit to the RSE its unique identification number (ID), e.g., its vehicle identification number. Subsequently, the OD flow pattern of a network can be obtained by comparing those IDs stored in the RSEs across the network. Unfortunately, such a straightforward appli-

cation is not supported by the connected-vehicle architecture, which is currently being developed to assure anonymity and untraceability [47].

1.2 Overview

The goal of this report is to investigate technical approaches to address privacy concerns associated with innovative applications enabled by connected vehicle systems. More specifically, the reports examines two distinctive applications, i.e., OD flow measurement and differentiated congestion pricing. The former is to retrieve the OD information from connected vehicles and the latter charges congestion tolls with respect to travel characteristics of connected vehicles, e.g., origins, destinations or paths that they traverse between their origins and destinations. Both applications require tracking vehicles, thereby creating privacy concerns and potentially violating the "anonymity by design" principle. This report thus explores two approaches to address the privacy issues. For OD flow measurement, we develop a novel measurement scheme that utilizes nice properties of a family of commutative one-way hash functions. The proposed scheme allows transportation authorities to collect aggregate OD flow data without learning information about individual vehicles. Furthermore, we adopt statistical methodology and use sampling to achieve far better efficiency without having to significantly degrade measurement accuracy. For differentiated congestion pricing, we propose an incentive program that allows motorists to opt in. The program is designed to encourage motorists to voluntarily reveal their private information and create a win-win situation for both motorists and the society.

The remainder of the report is organized as follows. Chapter 2 discusses the application of OD flow measurement while Chapter 3 proposes differentiated congestion pricing and addresses privacy concerns associated with it. Lastly, Chapter 4 concludes the report.

Chapter 2

Secure Origin-Destination Flow Measurement in Connected Vehicle Systems

2.1 Introduction

Traffic volume measurement is one of the most basic functions of road planning and management. Today the most widely used traffic volume statistic is the annual average daily traffic (AADT) [55], which describes the number of vehicles that traverse a specific *point* in the road system annually. Although AADT is very useful, it is only *"point"* information. It cannot tell us the traffic volume from one location to another in a city, and it cannot tell us the traffic volume on an arbitrary road segment. To gain better understanding of the road usage, we need *"point-to-point"* statistics that measure traffic volumes between distinct locations. Prior research has made steady advance in estimation of *"point"* statistics like AADT (e.g. Bozic et al. [12], Mohamad et al. [45], Lam and Xu [34], McCord et al. [43], Zhao and Park [67], Eom et al. [18], Neto et al. [46]). However, little work has been done on *"point-to-point"* traffic volume measurement.

In this chapter, we investigate the problem of secure *point-to-point* traffic volume measurement. We formalize point-to-point traffic as an origin-destination (OD) flow, whose size is the number of vehicles traveling from one geographical location (origin) to another (destination). Like AADT, OD flow data is an essential input to a variety of studies including estimation of transportation link flow distribution as part of investment planning, calculation of road exposure rates as part of safety analysis, and characterization of turning movements at intersections for signal timing determination, etc. However, very few techniques have been developed to collect OD data, not to mention doing so securely. Two commonly applied methods, household interviews and road surveys, are both time consuming and labor intensive. In general, issues about securely obtaining OD flow data have not been adequately addressed, and remain a major obstacle to a wide range of transportation studies.

Vehicular cyber-physical systems (VCPS) utilize the latest technologies in wireless communications, on-board computer processing, vehicle sensors, GPS navigation, etc., to improve safety, efficiency, resiliency, and environmental compatibility of transportation systems [19, 39]. There are several worldwide VCPS initiatives including connected vehicle systems, formerly known as IntelliDrive [4] or Vehicle Infrastructure Integration, from the US Department of Transportation (USDOT) [3], which envisions a nationwide system where connected vehicles routinely communicate with each other and roadside equipments (RSE) in real

time via Dedicated Short Range Communications (DSRC) or other wireless communications technologies. The advent of VCPS provides the potential for a fundamental shift in the way how OD data are collected. In principle, since vehicles are equipped with computing and communication capabilities, when a vehicle passes by an RSE, it can transmit its unique ID (e.g., its vehicle identification number or VIN) to the RSE. Subsequently, the OD flow between two RSEs can be easily recognized by comparing the two sets of IDs stored in them — there must be a vehicle traveling between them if both RSEs record a common ID. However, this straightforward approach leads to serious privacy breaching as it also tracks the entire moving history of each vehicle, which is against the "anonymity by design" principle for privacy protection required by connected vehicle systems. Hence, the challenge is to allow the collection of statistical OD flow data, yet protect information about each individual vehicle.

The objective of our work is to allow transportation authorities to collect aggregate OD flow data from VCPS without learning information about individual vehicles. First of all, globally unique IDs like VINs are identity information of vehicles. The leakage of such IDs will enable others to track the vehicles. Other permanent or temporary numbers that are transmitted repeatedly by a vehicle can also be exploited for the tracking purpose. Therefore, IDs (or other fixed numbers) should be preprocessed and protected by keys before transmission. In other words, RSEs will only be able to collect Keyed signatures of vehicles' IDs (referred to as KIDs). To compute the size of an OD flow between two locations based on KIDs, we introduce a family of commutative one-way hash functions. This family of hash functions, as its name suggests, has the unique properties of both commutativity and one-wayness. One crucial benefit of utilizing this hash function family is that vehicles can transmit their KIDs by hashing their IDs under totally different keys and be sure that no one is able to get their IDs, even knowing the keys (because of one-wayness), while it still allows us to compute the OD flow size as demanded (through commutativity).

Utilizing the family of commutative one-way hash functions, one straightforward solution for secure OD flow measurement is to have the RSEs be responsible for key generation, and they will broadcast their keys to the passing vehicles. When a vehicle passes by an RSE, it always reports a KID to the RSE, which is the hash result of its ID and the RSE's key, unless the key is equal to any previously received key. This scheme preserves vehicle privacy. It is also efficient in terms of computation overhead and space requirement. However, as we will demonstrate later, it is vulnerable to an identical-key attack. To address this problem, we propose an enhanced scheme for secure OD flow measurement. Instead of using the keys generated by the RSEs, the vehicles choose their own keys to protect their IDs. The second mechanism prevents the identical-key attack at the cost of increased computation overhead. To make this scheme practical, we adopt statistical method with sampling to construct a maximum likelihood estimation formula for the OD flow size. We summarize our contributions below:

1. We introduce the new problem of secure OD flow measurement, which measure point-to-point traffic in a privacy-preserving way. It has significantly practical impact in the context of a future connected vehicles system. We observe that VCPS have the potential for a fundamental shift in the way how OD flow data can be securely collected.

2. We adopt a family of commutative one-way hash functions and propose a secure scheme for OD flow measurement in VCPS. Not only will the new scheme correctly measure the OD flow sizes between all pairs of RSEs, but also it protects the identities of vehicles throughout the measurement process.

3. We further adopt statistical methods to improve the measurement efficiency at the cost of graceful

degradation in measurement accuracy. The tradeoff between computation efficiency and measurement accuracy can be controlled. The simulation results demonstrate the feasibility and scalability of our scheme.

2.2 Preliminaries

2.2.1 System Model

We consider a VCPS model involving three different entities: vehicles, RSEs, and a central server. Each vehicle has an unique ID, for example, VIN or other number chosen permanently or temporarily. The set of RSEs is denoted as $S = \{s_1, s_2, ..., s_N\}$, where N is the total number of RSEs of interest in the system. Both vehicles and RSEs are equipped with computing and communication capabilities, such as on-board computer chips and communication modules. Vehicles communicate with RSEs in real time via DSRC [3]. RSEs are connected to a central server through wired or wireless means. They collect information from vehicles and transfer the information to the server at the end of each measurement period (such as a day) to the server for further processing.

2.2.2 Problem Statement

We define an OD flow as the set of vehicles traveling from one RSE-equipped location (origin) to another RSE-equipped location (destination) during a measurement period. The size of an OD flow is the number of vehicles in the set. The problem is to design a secure OD flow measurement scheme that measures the sizes of OD flows between all pairs of origin/destination locations in a road system where RSEs are installed. By "secure", we mean no identity information of vehicles will be revealed during the whole measurement process. The proposed scheme should preserve the privacy of individual vehicles while allowing for aggregate OD flow measurement.

One easy way to measure OD flows is for each vehicle to transmit its ID whenever it passes by an RSE, which broadcasts queries in pre-set intervals (e.g., once a second), ensuring that each passing vehicle receives at least one query and in the meantime giving enough time for the vehicle to reply. But as we have explained in the introduction, this approach allows the authority to keep track of every individual vehicle. We want to design a solution in which a vehicle never transmits its ID or any fixed number that may be used for tracking purpose. Ideally, the information transmitted by a vehicle to any RSE is different each time and looks totally random, as a car that keeps transmitting the same number is more vulnerable of being tracked.

We assume that a special MAC protocol is used to support privacy preservation such that the MAC address of a vehicle is not fixed. For instance, when responding to an RSE, the vehicle may pick an MAC address randomly from a large space for one-time use. Since the number of vehicles in the vicinity of the RSE is limited, the probability for two vehicles to choose the same MAC address can be made negligibly small when the address space is sufficiently large.

2.2.3 Threat Model

We use a semi-trust model for the RSEs. We assume that the RSEs are all from the trustworthy authorities. This assumption can be enforced through authentication based on PKI. Each vehicle is pre-installed with

the public keys of the trusted third parties. Each RSE must have a public-key certificate from one of those third parties. It broadcasts the certificate in each query that it sends out. When receiving a query, the vehicle verifies the certificate, and then uses the RSE's public key to authenticate the RSE. We also assume that the authorities may exploit the information collected by RSEs to track individual vehicles when they need to do so. For instance, as we discuss previously, if a vehicle transmits any fixed number upon each query, that number can be exploited for tracking purpose.

It is important to note that there are many other ways to track a vehicle, for example, tailgating the vehicle, or setting cameras near RSEs to take photos and then using image processing to recognize vehicles. These methods are beyond the scope of this chapter. We focus on preventing automatic real-time tracking caused by the leakage of vehicle identity via RSEs.

2.2.4 Design Goals

To enable secure OD flow measurement under the aforementioned model, our scheme should achieve the following design goals.

1. Correctness: the proposed scheme should be able to correctly measure the OD flow size for arbitrary pair of RSEs, or with a measurement error that is probabilistically bounded.

2. Privacy guarantee: the proposed scheme should be able to protect the identity information of vehicles from unauthorized leakage and inference.

3. Efficiency: the proposed scheme should have means to control its overhead for scaling to a large road system.

2.3 Related Work

2.3.1 Traffic Volume Measurement

Much of existing work on traffic volume measurement is to design efficient protocols for estimation of the "point" traffic volume statistic, i.e., AADT (e.g. Bozic et al. [12], Mohamad et al. [45], Lam and Xu [34], McCord et al. [43], Zhao and Park [67], Eom et al. [18], Neto et al. [46]). To obtain the AADT values, automatic traffic recorders (ATR) are installed at road sections, whose major use is to count the number of vehicles passing by, and a prediction model, such as multiple linear regression (MLR) or artificial neural network (ANN), is then chosen to estimate AADT based on recorded data. The key issues are to choose the most appropriate prediction model under different circumstances, and to determine significant variables to be predictors. Challenges arise when there are limited ATRs and many independent variables to monitor for the prediction model. To address these issues and challenges, many approaches have been proposed over the passed few decades. For example, Mohamad et al. [45] propose to estimate AADT for county roads, where the scarce of APRs is the major concern, in an MLR model by using aggregated data at the county level. Lam and Xu [34] use both ANN and MLR to estimate AADT based on short-period counts of traffic in Hong Kong, and claim that ANN is more accurate than MLR. Another prediction model, geographically weighted regression (GWR), is presented in Zhao and Park [67]. Compared with MLR and ANN, GWR is more accurate and useful for studying the effects of the regressors at different locations.

Eom et al. [18] designed a spatial regression model (SRM), especially for nonfreeway facilities, which takes advantages of spatial dependency (i.e. the traffic volume at one monitoring station is correlated with the volumes at neighboring stations). Support vector machine for regression (SVR), a modified version of a pattern recognition technique, is introduced to forecast AADT in Neto et al. [46], which computes the SVR prediction parameters based on the distribution of the training data and achieves better performance than MLR. Besides these ground-based methodologies, McCord et al. also consider using high-resolution satellite imagery to identify vehicles for counting purpose [43]. As they observe through empirical results, combining satellite-based data with traditional ground-based data could reduce both AADT estimation errors and ground-based sampling efforts.

The aforementioned solutions, although elegant, are not appropriate for "point-to-point" traffic volume measurement (OD flow measurement). More complicated recoding techniques and computations are required to measure the traffic volume between a pair of locations than that of a single one. The problem also becomes more challenging when the security factor is involved as more and more people concern about their traveling privacy.

2.3.2 Privacy Preserving Data Mining

Another branch of research that relates to (but is also significantly different from) ours is privacy preserving data mining (PPDM), where researchers study various techniques to find simple rules or models that summarize the data (patterns) by examining data in large databases, while protecting the sensitive information about individuals whose information are the subject of the patterns [2, 24]. Several approaches have been proposed for different PPDM tasks over the past few decades, and the suggested solutions can be briefly summarized into two categories. One approach is to "randomly" perturb the data by adding "noise" before the data mining process, and mitigate the impact of the noise afterwards by reconstruction techniques. For example, Agrawal and Srikant are among the first who incorporate privacy concerns into data mining techniques, when they show the technical feasibility of PPDM by perturbing the original data using a randomizing function and reconstructing the distribution rather than individual records [6]. The work of Evfimievski et al. [20] and Evfimievski et al. [21] follow the same approach. However, some question about the usage of the randomization techniques for PPDM by showing that original data could be retrieved from the randomized dataset [31]. The alternative approach is to use cryptographic techniques to preserve privacy. For instance, Clifton et al. [13], propose a set of tools for privacy preserving mining of distributed data using encryption schemes. Others have also addressed association rule mining (e.g., Zhang et al. [66], Vaidya and Clifton [57]) and classification (e.g. Yu et al. [64]) following the same approach.

Although they are motivated by the same need to both protect privileged information and enable its use for research, industry or other purposes, directly applying PPDM techniques to secure OD flow measurement can still be problematic. The major reason is that PPDM does not address the security concerns of data collection, i.e., all these schemes in the context of PPDM require that all data collectors have collected "untouched" data in the first place. However, in our context of secure OD flow measurement, the fact that no one (including data collectors) should know the real vehicle ID ("untouched" data) except the vehicle itself, imposes a huge demand for customized design of a novel secure measurement scheme which targets at individual vehicle level starting from the very first step of data collection. Another reason is the different view of data during the data examining process. In PPDM tasks, since "untouched" data are distributed among a limited number of data collectors, efficient protocols can be designed under a unified database view.

However, given the nature of secure traffic volume measurement, data processing needs to be performed under an indistinguishable individual view (to protect vehicle identity throughout the process), which incurs inevitable higher computation overheads, motivating us to seek statistical methods to improve the overall efficiency without much degradation of measurement accuracy.

2.4 Solution Using Commutative One-way Hash Function

In this section, we propose a solution for secure OD flow measurement based on a family of commutative one-way hash functions (COHF). A common COHF is deployed to all RSEs and vehicles. RSEs are responsible for generating and distributing hash keys, and vehicles apply the hash function to produce Keyed signatures of their IDs (referred to as KIDs) using the keys obtained from RSEs that they pass by. The KIDs, instead of real IDs, are then reported to RSEs for OD flow measurement. Before describing the full solution, we first introduce the family of commutative one-way hash functions.

2.4.1 Commutative One-Way Hash Functions

Consider a hash function $h : A \times B \to C$, where the two arguments are a hash input and a hash key, respectively. A commutative one-way hash function, as its name suggests, satisfies both one-wayness and commutativity. The definitions of the properties below are collated from [44] and [10].

Definition 1 *A family of one-way hash functions (OHF) is a set of functions $h_\ell : X_\ell \times Y_\ell \to Z_\ell$ which satisfy the following three properties:*

- *Ease of computation: there exists a polynomial P such that for each integer ℓ, $h_\ell(x, y)$ is computable in time $P(\ell, |x|, |y|)$ for all $x \in X_\ell$ and all $y \in Y_\ell$.*
- *Preimage resistance: there is no polynomial P such that there exists a probabilistic polynomial time algorithm which can, given ℓ, a value $y \in Y_\ell$, and a value $z \in Z_\ell$, find $x \in X_\ell$ such that $h_\ell(x, y) = z$ with probability greater than $1/P(\ell)$ for all sufficiently large ℓ, when y is chosen uniformly from Y_ℓ and z is chosen uniformly from Z_ℓ.*
- *2nd-preimage resistance: there is no polynomial P such that there exists a probabilistic polynomial time algorithm which can, given ℓ, a pair $(x, y) \in X_\ell \times Y_\ell$, and a value $y' \in Y_\ell$, find a value $x' \in X_\ell$ such that $h_\ell(x, y) = h_\ell(x', y')$ with probability greater than $1/P(\ell)$ for all sufficiently large ℓ, when (x, y) is chosen uniformly among all elements of $X_\ell \times Y_\ell$ and y' is chosen uniformly from Y_ℓ.*

h_ℓ is said to have the one-wayness property if it satisfies the three properties above.

In Definition 1, the first property tells that for a known function h_ℓ, given an input x and a key y, it is relatively easy to compute $h_\ell(x, y)$ (in polynomial time). The second property tells that it is computationally infeasible to find any input which hashes under a given key to the arbitrarily pre-specified output. And the third property tells that it is computationally infeasible to find a second input that can be hashed under a given key to the same output as the arbitrarily pre-specified input and key.

Definition 2 *A commutative hash function (CHF) is a hash function $h_\ell : X_\ell \times Y_\ell \to X_\ell$ which satisfies the following property:*

- *Commutativity: for all $x \in X_\ell$ and for all $y_1, y_2 \in Y_\ell$, $h_\ell(h_\ell(x, y_1), y_2) = h_\ell(h_\ell(x, y_2), y_1)$ holds.*

One can see that the commutativity lies in the keys. In other words, given an input and two keys, commutativity tells that changing the order in which the two keys are applied to the input won't change the hash result. Further observations gives that, if the range of the one-way hash function is equal to the domain of its first argument, then we can exploit a new family of commutative one-way hash functions which shall satisfy both one-wayness and commutativity.

Definition 3 *A family of commutative one-way hash functions (COHF) is a family of hash functions which have both one-wayness property and commutativity property.*

We will see shortly one crucial benefit of utilizing this hash function family: Vehicles can transmit their KIDs by hashing their IDs under totally different keys, and be sure that no one will be able to get their IDs, even knowing the keys that the vehicles have used (because of one-wayness). Yet the KIDs allow OD flow measurement as demanded (through commutativity).

2.4.2 The Proposed Scheme

Using the COHFs, we propose the following scheme for secure OD flow measurement. Each measurement period consists of three phases: initialization, online reporting, and offline measurement. The initialization phase establishes the keys for RSEs. Information for OD flow measurement are securely collected during the online reporting phase. Finally, the offline measurement phase computes the sizes of OD flows between all pairs of RSEs.

Construction of Commutative One-Way Hash Functions

Before describing the three measurement phases, we construct the COHFs. According to Definition 3, a COHF is a hash function that satisfies both one-wayness and commutativity. There can be different constructions of COHFs given different types of hash functions, and the one that we adopt is based on the exponentiation modulo n function, $h_n(x, y) = x^y \ mod \ n$. We claim that h_n is a COHF with some restrictions on n.

Definition 4 *A prime p is defined to be safe if $p = 2p' + 1$ where p' is an odd prime.*

Definition 5 *n is defined to be a rigid integer if $n = pq$ where p and q are distinct large safe primes.*

Theorem 1 *The function $h_n(x, y) = x^y \ mod \ n$ is a commutative one-way hash function if n is a rigid integer.*

Note that the selection of n and h_n in Theorem 1 follows the RSA algorithm [50]. Through two lemmas, we prove Theorem 1 by showing that h_n satisfies both one-wayness and commutativity, which qualifies it as a COHF.

Lemma 1 *The function $h_n(x, y) = x^y \bmod n$ satisfies commutativity for any given n.*

Proof: For any given n, and for any integer x, y_1 and y_2, $h_n(h_n(x, y_1), y_2)$ is the result of multiplying x by itself modulo n for $y_1 + y_2$ times, while $h_n(h_n(x, y_2), y_1)$ is the result of multiplying x by itself modulo n for $y_2 + y_1$ times. Therefore, $h_n(h_n(x, y_1), y_2) = h_n(h_n(x, y_2), y_1)$ holds for all integers x and for all integers y_1 and y_2, which establishes that $h_n(x, y)$ satisfies commutativity according to Definition 2. □

Lemma 2 *The function $h_n(x, y) = x^y \bmod n$ satisfies one-wayness if n is a rigid integer.*

Proof: We prove the one-wayness of h_n by showing that h_n satisfies all three properties in Definition 1. First, h_n satisfies *ease-of-computation*. There are efficient methods to perform exponentiation of a base to an exponent. One is to start with an output of 1 and a value set to the base, read the exponent in binary bit by bit from low-order bit to high-order bit, square the value each time from the second bit on, and if the bit is 1, also multiply the output by the value. Modular reduction is performed after each operation to keep the intermediate results bounded by n. Applying this method, according to Kaufman et al. [32], the number of multiplications rises linearly with the length of the exponent in bits rather than with the value of the exponent itself.

Second, the *preimage resistance* of h_n follows the cryptographic security of the RSA cryptosystem, which is equivalent to the difficulty of taking roots modulo n. Root finding modulo n is shown in [53] to be very difficult (cannot be done in polynomial time) when n is a large composite number. Also, the commonly accepted RSA assumption [50] tells that, for these "appropriately" chosen n, computing x from $h_n(x, y)$, y, and n cannot be done in polynomial time for all but an exponentially small number of cases.

Third, the *2nd-preimage resistance* of h_n is given by the characteristics of rigid integers. It is demonstrated in [10] that if n is a rigid integer, then finding collisions with specific constraints (2nd-preimage) cannot done in polynomial time unless a set of roots can be obtained such that the product R of their indices is a multiple of the desired root index y'. However, the number of known roots which would have to be provided in order to have a non-negligible probability of their product being a multiple of a random number selected later (y') would be prohibitively large. In other words, to find collisions in the form of $h_n(x, y) = h_n(x', y')$ for given x, y, and y' is extremely unlikely. This completes the proof. □

Initialization

A common commutative one-way hash function h_n must be pre-distributed to all vehicles and RSEs. The hash function is determined by a large rigid integer n. Algorithm 1 provides a practical method to construct a rigid integer. The basic idea is that for $n = pq$ to be a rigid integer, each of p, q, $\frac{(p-1)}{2}$ and $\frac{(q-1)}{2}$ must be primes congruent to 5 modulo 6. Therefore, the process is to first select a "random" integer p' that is congruent to 5 modulo 6 until one is found such that p' and $2p' + 1$ are both prime, and then choose

Algorithm 1 Algorithm for Constructing Rigid Integers

INPUT: an upper bound U and a lower bound L.
OUTPUT: a rigid integer n such that $L \leq n \leq U$, or return -1 if such n does not exist.
findN $\leftarrow$ FALSE, $n \leftarrow -1$, $i \leftarrow -1$
repeat
  findP $\leftarrow$ FALSE
  repeat
    $i \leftarrow i + 1$, $p' \leftarrow 6i + 5$, $p \leftarrow 2p' + 1$
    if p' and p are both prime **then**
      findP $\leftarrow$ TRUE
    end if
  until findP = TRUE
  $j \leftarrow i$
  repeat
    $j \leftarrow j + 1$, $q' \leftarrow 6j + 5$, $q \leftarrow 2q' + 1$
    if q' and q are both prime **then**
      $n \leftarrow pq$
      if $L \leq n \leq U$ **then**
        findN $\leftarrow$ TRUE
      end if
    end if
  until $n > U \parallel$ findN = TRUE
until $n > U \parallel$ findN = TRUE
if findN = FALSE **then**
  $n \leftarrow -1$
end if
return n

a suitable q' similarly. After we find two distinct safe integers p' and q', n can be easily constructed by $n = pq = (2p' + 1)(2q' + 1)$.

All RSEs and the vehicles are pre-configured with a suitable n, and the clocks of RSEs are loosely synchronized as they are all connected to a central server through wired or wireless means. In the initialization phase of each measurement period, every RSE generates a random number as its hash key for the current period. With the assistance of the server, all hash keys are unique: Let k_i be the hash key generated by RSE s_i. We require that, for any two RSEs s_i and s_j, their keys k_i and k_j be different, i.e., $k_i \neq k_j$. If the server finds that two hash keys reported from RSEs are the same, it will inform one of them to regenerate a key. The key uniqueness requirement serves an important purpose, which will be explained later.

Online Reporting

The online reporting phase securely collects information for OD flow measurement. The RSEs broadcast queries in pre-set intervals (e.g., once a second), ensuring that each passing vehicle receives at least one query and in the meantime giving enough time for the vehicle to reply. Collisions can be resolved through well-established CSMA or TDMA protocols, which are not the focus of this chapter. Every query that an RSE sends out includes the RSE's ID, public-key certificate, as well as its current hash key. When a vehicle, whose ID is v_j, receives a query from an RSE s_i, it first verifies the certificate, and then uses the RSE's public key to authenticate the RSE. After verifying that s_i is from the trustworthy authority, the vehicle generates a KID based on its ID v_j and the RSE's key k_i by computing a hash $c = h_n(v_j, k_i) = v_j^{k_i} \ mod \ n$. After that, it reports this KID c to the RSE, which then stores c in its local storage.

Offline Measurement

At the end of each measurement period, the OD flow sizes between pairs of RSEs will be computed based on the KIDs collected by RSEs during the online reporting phase. More specifically, every RSE will send its key as well as the collected KID set to a central server, and the central server will be in charge of the offline OD flow size computation.

Thanks to the commutativity property of h_n, given two sets of KIDs, $\{h_n(\cdot, k_x)\}$ and $\{h_n(\cdot, k_y)\}$, collected by two RSEs s_x and s_y respectively, and the two corresponding keys, k_x and k_y, it is easy for the central server to determine the OD flow size between s_x and s_y. In principle, changing the order in which two keys are applied to the same vehicle ID using commutative one-way hash functions won't change the final hash result. Therefore, to find vehicles that pass both s_x and s_y, the central server will further hash each RSE's KID set by the other RSE's key, and then compare the common values of the two double-hashed sets. The process is as follows.

I. The central server performs an element-wise hash over the KID set $H_x = \{h_n(\cdot, k_x)\}$ collected by s_x using s_y's key, k_y, to obtain a double-hashed set $H_{x,y} = \{h_n(h_n(\cdot, k_x), k_y)\}$.

II. The central server performs an element-wise hash over the KID set $H_y = \{h_n(\cdot, k_y)\}$ collected by s_y using s_x's key, k_x, to obtain a double-hashed set $H_{y,x} = \{h_n(h_n(\cdot, k_y), k_x)\}$.

III. The central server finds the common elements in $H_{y,x}$ and $H_{x,y}$. According to Theorem 2 below, the OD flow size between s_x and s_y is equal to the number of common elements in the two double-hashed

sets $H_{y,x}$ and $H_{x,y}$. If we take the timestamps of the KIDs into consideration, we can easily determine the size of an directional OD flow for vehicles that appear at s_x first and then appear at s_y at a later time.

Theorem 2 *Given a commutative one-way hash function $h_n(v,k) = v^k \bmod n$, for arbitrary vehicle IDs v_1 and v_2, and arbitrary keys k_1 and k_2, $h_n(h_n(v_1,k_1),k_2) = h_n(h_n(v_2,k_2),k_1)$ holds if and only if $v_1 = v_2$ holds.*

Proof: ("$\Leftarrow$") Suppose $v_1 = v_2$.

$$h_n(h_n(v_1,k_1),k_2) = h_n(h_n(v_2,k_1),k_2). \tag{2.1}$$

Since h_n is a commutative one-way hash function, it is by definition commutative. Therefore, for any two keys k_1 and k_2, we have

$$h_n(h_n(v_2,k_1),k_2) = h_n(h_n(v_2,k_2),k_1). \tag{2.2}$$

From equation 2.1 and 2.2, we have

$$h_n(h_n(v_1,k_1),k_2) = h_n(h_n(v_2,k_2),k_1). \tag{2.3}$$

("$\Rightarrow$") Suppose $h_n(h_n(v_1,k_1),k_2) = h_n(h_n(v_2,k_2),k_1)$. Again, h_n is commutative. By definition, for any two keys k_1 and k_2, we have

$$h_n(h_n(v_2,k_2),k_1) = h_n(h_n(v_2,k_1),k_2). \tag{2.4}$$

From equation 2.4 and the assumption, we obtain

$$h_n(h_n(v_1,k_1),k_2) = h_n(h_n(v_2,k_1),k_2). \tag{2.5}$$

Since the number of vehicles in the vicinity of two RSEs is limited, and the hash space is sufficiently large, the probability for two distinct vehicle IDs to be hashed under the same key to the same value is negligibly small. Therefore, from Equation 2.5, we can conclude $v_1 = v_2$, with exceedingly high probability. This completes the proof. □

2.4.3 Scheme Analysis

The proposed scheme preserves vehicle privacy throughout the measurement process. The initialization phase is clearly privacy-preserving because no information of any vehicle is transmitted. During the online reporting phase, vehicles only transmit their KIDs to RSEs. Since a COHF h_n is applied, no one will be able to obtain real vehicle IDs from their KIDs because of h_n's one-wayness property. In addition, vehicles are protected from being tracked because no fixed information of any vehicle is allowed to be transmitted according to the key uniqueness requirement. As for the offline measurement phase, due to the one-wayness property of h_n, the central server cannot obtain any vehicle ID from its KID, either.

The proposed scheme is also efficient in terms of computation time and space requirement. To measure OD flow sizes, each vehicle only needs to compute one hash for each passing RSE, and each RSE only needs to store one KID value for each passing vehicle. Therefore, the overall time complexity for each vehicle is linear to the number of distinct RSEs that it passes by, which is bounded by $O(N)$, where N is the total RSEs in the whole system. The overall space requirement for each RSE is linear to the number of distinct vehicles that pass by, which is bounded by $O(M)$, where M is the total number of vehicles in the system. In addition, for the central server to compute an OD flow size between two RSEs, it takes an additional hash for each KID value from the two KID sets. So the number of hash operations is bounded by $O(M)$. To find the common double-hashed values, the central server needs to sort the two double-hashed sets, which takes $O(M \log M)$. Therefore, the overall time complexity for the central server to measure an OD flow is $O(M \log M)$.

2.4.4 Identical-key Attack

The above analysis assumes the transportation authority (who owns RSEs and the central server) is trustworthy. But this assumption also allows the transportation authority an easy way of tracking vehicles. It may simply set all or a portion of RSEs with the same key. When a vehicle passes these RSEs, its KID stays the same and therefore may be exploited for tracking purpose. Even if the authority uses different keys for different RSEs, it may use the same key for an RSE over multiple measurement periods. Any vehicle that passes the RSE will repeatedly transmit the same KID over this time. Even within the same measurement period (when the hash key is definitely unchanged), a vehicle may pass a RSE for two or more times, and it will transmit the same KID, which is still undesirable from the privacy-preserving point of view. Ideally, the vehicle should transmit a different value each time, and the value should appear totally random and unpredictable.

To avoid transmitting the same number (KID) in the above scenarios, a vehicle may keep record of the RSE keys that it has seen before, and will not respond to an RSE with its KID if the key from that RSE is already in the vehicle's record.

This solution however causes an underestimation problem. Suppose during a measurement period (e.g., a day), a vehicle v_j passes by an RSE s_i for two or more times. This is not uncommon in reality. For example, people driving to work are likely to follow the same route back home. On its way to work, vehicle v_j reports its KID to RSE s_i. But when it comes back to home on the same route, v_j receives the same hash key from the same RSE, if it does not respond with the same KID, it will be counted only in the OD flow from home to workplace, but will not be counted in the flow from workplace to home. In other words, while the vehicle contributes twice to traffic volume between home and workplace, it is actually counted only once (if the vehicle does not respond to the same key). This will result in an underestimation of the OD flow size.

To fully address the above concerns, we need to make a fundamental shift in who is responsible for key generation. We shall move that responsibility from RSEs to the vehicles in order to ensure that the key uniqueness requirement is met.

2.5 Enhanced Scheme for Secure OD Flow Measurement

Instead of using the keys generated by RSEs, our second scheme lets vehicles choose their own keys to protect their IDs. Still, RSEs will collect KIDs from vehicles, and a central server will then compute pairwise OD flow sizes based on the collected KID sets. The enhanced scheme has two phases: online reporting and offline measurement.

Similar to the previous scheme, vehicles and RSEs are pre-configured with a common commutative one-way hash function h_n, which is determined by a suitable n computed from Algorithm 1. Unlike the previous scheme, each RSE not only stores the KIDs of the passing vehicles, but also saves the corresponding keys that are used by the vehicles to compute the KIDs. Essentially, each RSE will store a set of $\langle$key, KID$\rangle$ pairs obtained from the passing vehicles in the online reporting phase, which is then used in the offline measurement phase to determine OD flow sizes.

2.5.1 The Enhanced Scheme

Online Reporting

During the online reporting phase, $\langle$key, KID$\rangle$ pairs are securely collected by RSEs from passing vehicles in preparation for offline OD flow measurement. More specifically, when a vehicle v_i passes by an RSE s_j, the vehicle will first verify that the RSE comes from trusted authorities based on the public-key certificate received from the RSE's periodic broadcast. Then the vehicle will randomly choose a hash key k, and compute a hash $c = h_n(v_i, k) = v_i^k \ mod \ n$, which serves as a KID of v_i. After that, the vehicle reports the KID c and the key k to s_j, which stores this $\langle$key, KID$\rangle$ pair in its local storage.

Offline Measurement

At the end of each measurement period, all RSEs will send their collected data to the central server, which computes the sizes of the OD flows between all pairs of RSEs. Given two sets of $\langle$key, KID$\rangle$ pairs collected by two RSEs s_x and s_y, the central server can compute the size of the corresponding OD flow based on the hash function h_n's commutativity. The process is to go through these two sets, and for each pair collected by s_x, check if it shares a common double-hashed value with any pair collected by s_y. If such a pair exists, a vehicle is found to pass both RSEs.

Definition 6 *Two $\langle$key, KID$\rangle$ pairs, $\langle k_x, c_x\rangle$ and $\langle k_y, c_y\rangle$, are said to share a common double-hashed value if $h_n(c_y, k_x) = h_n(c_x, k_y)$ holds. Note that c_x and c_y are hash values themselves.*

Algorithm 2 summarizes the offline measurement process. We give the basic idea of the algorithm as follows: Suppose $\langle k_x, c_x\rangle$ from s_x and $\langle k_y, c_y\rangle$ from s_y share a common double-hashed value, i.e., $h_n(c_y, k_x) = h_n(c_x, k_y)$. By definition, $c_x = h_n(v_x, k_x)$, and $c_y = h_n(v_y, k_y)$, where $v_x(v_y)$ is the ID of the vehicle passing by RSE $s_x(s_y)$. Thus, $h_n(c_y, k_x) = h_n(h_n(v_y, k_y), k_x) = h_n(h_n(v_x, k_x), k_y) = h_n(c_x, k_y)$, which means $v_x = v_y$ according to Theorem 2. Hence, we know that a common vehicle has passed both s_x and s_y.

Algorithm 2 Algorithm for Offline Measurement

INPUT: a commutative one-way hash function h_n; two sets of ⟨key, KID⟩ pairs collected by two RSEs s_x and s_y, $D_x = \{\langle k_{ix}, c_{ix}\rangle\}_{i=1}^{n_x}$, $D_y = \{\langle k_{iy}, c_{iy}\rangle\}_{i=1}^{n_y}$.
OUTPUT: the OD flow size between RSE s_x and s_y, i.e., the number n_{xy} of pairs in D_x and D_y that share common double-hashed values.
$n_{xy} \leftarrow 0$
for i from 1 to n_x **do**
  $k_x \leftarrow k_{ix}$, $c_x \leftarrow c_{ix}$
  for j from 1 to n_y **do**
    $k_y \leftarrow k_{jy}$, $c_y \leftarrow c_{jy}$
    $p_{xy} \leftarrow h_n(c_x, k_y)$
    $p_{yx} \leftarrow h_n(c_y, k_x)$
    if $p_{xy} = p_{yx}$ **then**
      $n_{xy} \leftarrow n_{xy} + 1$
    end if
  end for
end for
return n_{xy}

Scheme Analysis

The enhanced scheme eliminates the underestimation problem that is encountered by the previous scheme. Under the new scheme, even though a vehicle v_i may pass an RSE s_j for several times, each time it uses a different key to protect its ID. No matter how many times a vehicle passes by an RSE, each time a different ⟨key, KID⟩ pair will be recorded and be counted towards the final measurement result. Therefore, the measured OD flow sizes should always be equal to the real OD flow sizes.

The enhanced scheme improves the measurement accuracy at the cost of increased computation overhead. In order to compute the OD flow size between two RSEs, s_x and s_y, the central server needs to perform a re-hash for each pair collected by s_x under every key from s_y, and do the same thing for s_y. Suppose the two RSEs have collected n_x and n_y pairs of ⟨key, KID⟩, respectively. The time complexity for the central server to compute the OD flow size between the two RSEs will be $O(n_x \cdot n_y)$.

2.5.2 Down Sampling

To address the efficiency problem, we propose to use down sampling to estimate the OD flow size during the offline measurement phase. Given two sets of ⟨key, KID⟩ pairs collected by two RSEs s_x and s_y, $D_x = \{\langle k_{ix}, c_{ix}\rangle\}_{i=1}^{n_x}$, $D_y = \{\langle k_{iy}, c_{iy}\rangle\}_{i=1}^{n_y}$, the OD flow size between s_x and s_y is equal to the number n_{xy} of pairs in D_x and D_y that share common double-hashed values. The time required for the central server to calculate the OD flow size is $O(n_x \cdot n_y)$. To reduce computation overhead, we randomly select n_p elements from D_x and n_q elements from D_y, denoting them as D'_x and D'_y respectively, and calculate the number n'_{xy} of pairs in D'_x and D'_y that share common double-hashed values. It takes $O(n_p \cdot n_q)$ time to compute the OD flow size from such a sample. Based on n'_{xy} and the sampling probability, we can construct

the maximum likelihood estimate of n_{xy} as

$$N_{xy} = n'_{xy} \times \frac{n_x}{n_p} \times \frac{n_y}{n_q}, \tag{2.6}$$

which is derives as follows: If we define two pairs from D_x and D_y as a common element in these two sets if they share a common double-hashed value, then we have the following equivalent problem of set intersection size estimation: Let X and Y be two sets having cardinalities $|X| = a$, $|Y| = b$, $|X \cap Y| = c$. We randomly choose two subsets of elements, X_k and Y_k, with cardinalities a_k and b_k, from X and Y, respectively. We calculate the number of common elements in the two sets X_k and Y_k, denoted by c'. The problem is to construct the maximum likelihood estimate of c based on c', a, b, a_k, and b_k.

For a randomly selected element $x \in X$, the probability for $x \in Y$ equals the probability for $x \in X \cap Y$, namely, $P(x \in Y) = \frac{c}{a}$. Similarly, for a randomly selected element $y \in Y$, the probability for $y \in X$ equals the probability for $y \in X \cap Y$, namely, $P(y \in X) = \frac{c}{b}$. Partition X and Y into two subsets each at random, $X = X_k \cup X_u$, $Y = Y_k \cup Y_u$, $|X_k| = a_k \leq a$, $|Y_k| = b_k \leq b$, where X_k and Y_k represent the "known" elements that we choose from X and Y, while X_u and Y_u represent the remaining "unknown" elements. Then $|X_k \cap Y|$ is binomially distributed according to $B(n,p) = B(a_k, \frac{c}{a})$. Therefore, for each $x \in X_k$, $P(x \in Y) = \frac{c}{a}$. Since each element is chosen uniformly random from the original set, $P(x \in Y_k | x \in Y) = \frac{b_k}{b}$. Combining the probabilities of the two dependent events, for each $x \in X_k$,

$$P(x \in Y_k) = P(x \in Y) \cdot P(x \in Y_k | x \in Y) = \frac{c}{a} \times \frac{b_k}{b}.$$

The size of the subset intersection $c_k = |X_k \cap Y_k|$ is again binomially distributed according to $B(n,p) = B(a_k, \frac{c}{a} \times \frac{b_k}{b})$, with expectation value

$$\begin{aligned} E(c_k) &= |X_k| \times P(x \in Y_k | x \in X_k) \\ &= a_k \times \frac{c}{a} \times \frac{b_k}{b} \\ &= c \times \frac{a_k}{a} \times \frac{b_k}{b}. \end{aligned}$$

Now consider the above process in reverse. The sampling gives an estimate of c_k, which is c', we want to find the value of $c = |X \cap Y|$ that is *most likely* to have given rise to c', which in this case means finding c such that $E(c_k) = c'$. This is the maximum likelihood estimate of c, c_{mle}. From the equation just derived, $c_{mle} = c' \times \frac{a}{a_k} \times \frac{b}{b_k}$.

Given the equivalence of our problem and the set intersection estimation problem above, it is clear that $N_{xy} = n'_{xy} \times \frac{n_x}{n_p} \times \frac{n_y}{n_q}$ is the maximum likelihood estimate of n_{xy}. By adopting the down sampling method, the overall computation overhead for the central server to compute the OD flow size between s_x and s_y is reduced from $O(n_x \cdot n_y)$ to $O(n_p \cdot n_q)$.

2.6 Simulation Results

In this section, we evaluate the performance of our two schemes through simulation experiments. The programs are written in Matlab, and the commutative one-way hash function is realized through Java's

BigInteger class [1]. The experimental platform is a PC featured with an Intel Core 2 E8400 CPU and 4GB RAM, running Windows XP. However, we expect the central server in practice to be much more powerful than our PC. Moreover, the offline measurement task may be divided and distributed to several servers, which saves time by performing smaller pieces of work in parallel; the computation may also be outsource to cloud servers.

In the experiments, we consider two performance metrics, measurement accuracy and computation overhead. The measurement accuracy is represented by error ratio r, which is defined as

$$r = \frac{|N_{xy} - n_{xy}|}{n_{xy}} \times 100\%, \tag{2.7}$$

where N_{xy} is the measured OD flow size between RSEs s_x and s_y, and n_{xy} is the actual OD flow size. From the definition, smaller error ratio r represents more accurate measurement result, and in turn, reflects better performance of the scheme, and vice versa. The computation overhead is measured by time consumed by the central server to perform the offline measurement of an OD flow size between two RSEs, which dominates the overall measurement process. It is also obvious that lower computation overhead means higher efficiency of the scheme, and vice versa.

The datasets used in the experiments are generated such that each vehicle ID or key is a 64-bit number, and two RSEs, s_x and s_y, each store 10,000 vehicle records. There are 1,000 vehicles that pass both s_x and s_y, i.e., the actual OD flow size between s_x and s_y is 1,000.

Our first scheme has an error ratio of 0% unless it does not respond to any key that it has seen before (for privacy purpose as we have discussed in Section 2.4.4). Hence, our experiment only measures the time cost. The enhanced scheme solves the problem of identical-key attack at the cost of higher computation overhead. It has an error ratio of 0% only when the sampling probability is 100%. In our experiments, we vary the sampling probability p from 0.1 to 1, with a step size of 0.1. For each sample probability, we randomly draw a fraction p of all records from s_x and do the same for s_y. The offline measurement is performed over the sampled subsets and the OD flow sizes are estimated by Equation 2.6. The time cost is measured and the error ratio is computed from Equation 2.7. The process is repeated 10 times to show the statistic effect.

Table 1 and Figures 1-2 present our simulation results. Table 1 shows that the computation overhead of the first scheme is around $\frac{1}{4}$ of the second scheme when the sampling probability is 100%. The reason is that the latter performs more hash operations. The two figures are drawn from the simulation results of the second scheme. Figure 2.1 shows the mean and the standard deviation of the error ratio for OD flow measurement under varied sampling probabilities. The length of each error bar is two times the standard deviation of the error ratio, whose mean is at the center of the bar. We see that both mean and standard deviation of the error ratio decrease with the increment of the sampling probability. Intuitively, when we increase the sample size, the measurement result is likely to be more accurate. When the sampling probability equals 1, the error ratio is 0% (as shown in the rightmost of the figure), which agrees with our theoretical prediction.

Figure 2.2 shows the average time taken by the central server to measure the OD flow size under each sampling probability. It is clear that the computation overhead increases quadratically with the sampling probability, which is also consistent to our analysis in Section 2.5.2.

Table 2.1: Average computation overhead for the two proposed schemes.

	First Scheme	**Second Scheme with Different Sampling Probabilities**									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
time ($\times 10^4$**secs)**	0.8575	0.0350	0.1453	0.3152	0.5813	0.8761	1.3059	1.7146	2.3234	2.8343	3.6370

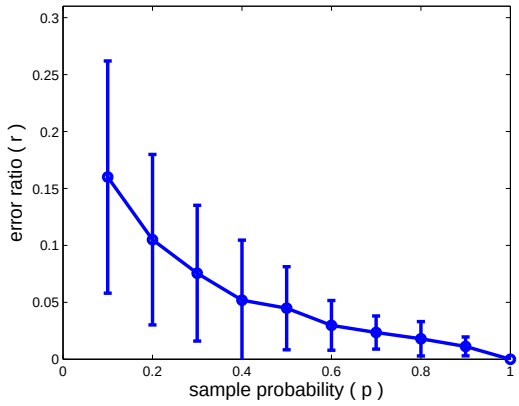

Figure 2.1: Mean and standard deviation of error ratios for OD flow measurement under different sampling probabilities.

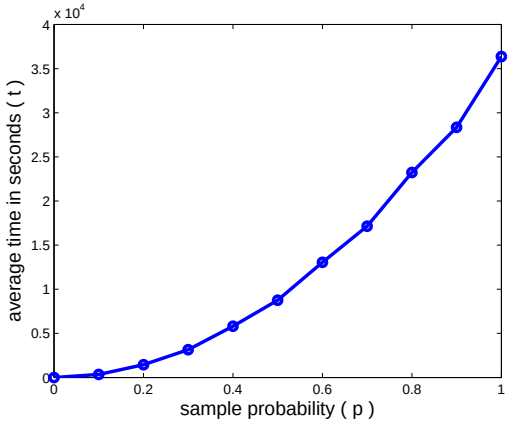

Figure 2.2: Average time overhead for offline measurement under different sampling probabilities.

Chapter 3

Differentiated Congestion Pricing of Urban Transportation Networks via Connected Vehicle Systems

3.1 Introduction

Price discrimination or differentiation is an economic concept defined by Dupuit [17] as a situation where identical products are sold for different prices [58]. Pigou [48] later classified price discrimination into three categories. First-degree price discrimination is the case where everyone pays his or her maximum willingness-to-pay for the product. If the price of a unit of product depends on the number of units of product being purchased, it is classified as second degree. Lastly, third-degree discrimination means that the price of a unit of product can be different for different type of users.

Price discrimination is not uncommon in the transportation market. A good example for second-degree discrimination is transit fare, when, e.g., a two-way ticket is cheaper than two one-way tickets, or the price of a daily pass is independent of the number of rides taken by a passenger within one day. Moreover, some transit agencies differentiate travelers by age and collect different fares for kids, students, adults and elder people, which is an example of third-degree discrimination. Previous studies have discussed price discrimination in the context of congestion pricing. Wang et al. [59] and Lawphongpanich and Yin [37] investigated nonlinear pricing, which is essentially an instance of second-degree discrimination where the amount of toll depends on, not strictly proportional to, the distance traveled inside a tolling area. A case of third-degree discrimination is investigated in [28], which differentiated users based on their vehicle type. Others, e.g., Small and Yan [54], Yang and Zhang [61], Yang and Huang [60], Yin and Yang [62], differentiated users based on their values of travel time. de Palma and Lindsey [15] compared the effect of toll differentiation based on the value of time and vehicle type on welfare. As pointed out by Pigou [48], third-degree price discrimination generally requires an ability to distinguish different customer groups, i.e., there must be some observable attributes associated with each group, unless the pricing scheme possesses a self-selection mechanism. Given that the value of time is not directly observable, it is not surprising to find little practice of price differentiation with respect to the value of time.

This chapter discusses another third-degree differentiated pricing scheme that differentiates travelers

with respect to their travel characteristics, i.e., origins, destinations, or paths that they traverse between their origins and destinations. Although similar schemes may have been implemented in closed networks, e.g., tolled freeways, to our best knowledge, it has not been explored in an open, urban network environment for the purpose of congestion mitigation. We note that the advancements of connected vehicle technologies have technically enabled such price differentiation.

The contributions of this chapter are threefold. First, we use numerical examples to demonstrate the potentials of price differentiation with respect to origin, origin-destination (OD) pair or path. The examples show that in a first-best network condition where all the links are tollable, differentiated pricing can substantially reduce travelers' financial burden; in a second-best environment where only some links are tollable, it helps achieve a lower level of congestion. Second, we formulate optimization models to determine optimal differentiated pricing schemes for general networks. Third and more importantly, recognizing that price differentiation with respect to travel characteristics may compromise travelers' location privacy, we propose an approach of modeling privacy, and then design an incentive program to provide incentives for travelers to reveal their travel information and voluntarily participate in differentiated pricing. Such an opt-in program is designed to create a win-win situation for both travelers and society.

The remainder of this chapter is organized as follows. Section 3.2 discusses different types of differentiated pricing and their formulations, and presents numerical examples to make a case for differentiated pricing. Section 3.3 discusses the location privacy issue associated with differentiated schemes, and proposes an approach of modeling privacy. Section 3.4 is dedicated to the development of an incentive program for differentiated pricing. Lastly, Section 3.5 concludes the chapter and discuss another way to mitigate travelers' privacy concerns.

3.2 Differentiated Pricing Schemes

Differentiated pricing schemes we discuss in this chapter include origin-specific, OD-specific and path-based. As their names suggest, travelers on the same link will be charged differently, with respect to their respective origin, OD pair or path. Intuitively, these schemes are more flexible than traditional anonymous tolling. Mathematically, they can be viewed as different levels of relaxation to anonymous schemes.

To facilitate the presentation, we label the differentiation level of anonymous pricing as zero, and subsequently the levels of differentiation for origin-specific, OD-specific and path-based pricing as one, two and three, respectively.

3.2.1 Notation

Let $G(N, A)$ denote a transportation network, where N is the set of nodes and A is the set of directed links. Index a is used to denote a link, which is also represented by its end nodes $i, j \in N$, i.e., $(i, j) = a$. For link a, x_a and γ_a are its aggregate flow and toll, respectively. The latter is expressed in the unit of time for the sake of simplicity. Let $W \subseteq N \times N$ be the set of OD pairs with strictly positive demand, w be the index of its elements and d_w be the demand of OD pair w. For every OD pair $w \in W$, $o(w)$ represents its origin. The set of all paths connecting OD pair w is denoted by P_w with its elements being indexed by p. A binary paramete δ represents the link-path incidence, i.e., if link a is on path p, then δ_{ap} is one; otherwise zero. For every path p, f_p and π_p denote its flow and toll, respectively. Also, $t_p(.)$ and $t_a(.)$ are the travel time

for path p and link a, respectively. For second-best pricing, the set of tollable links is denoted by Ψ, and its complement set $\bar{\Psi}$ includes all the untollable links.

3.2.2 Formulations

As aforementioned, path-based pricing has the highest level of price differentiation, because the origin or destination of a trip can be easily determined from the path utilized by the trip. Hence, a general path-based formulation is used in this chapter to describe all three different schemes. Notice that origin-specific and OD-specific pricing are link-based schemes, and thus the toll of each path is the sum of tolls on links comprising the path. In contrast, path-based tolls may not be link-wise additive, because they are determined for specific paths and may not be decomposable into link-based tolls.

We first discuss a first-best network condition where all links are tollable. In such an environment, even with the lowest level of price differentiation, i.e., anonymous tolling, congestion pricing is able to induce system optimum and replicate system optimum link flows (e.g., Hearn and Ramana [27]). Consequently, the benefit of price differentiation can only be reflected on a secondary objective. In this chapter, we choose revenue minimization as the secondary objective because it represents a financial burden to the traveling public. Below, we formulate a program for finding a first-best path-based pricing scheme to minimize the total toll revenue:

$$\min \quad \sum_{w \in W} \sum_{p \in P_w} \pi_p f_p \tag{3.1}$$

$s.t.$

$$\sum_{p \in P_w} f_p = d_w \qquad \forall w \in W \tag{3.2}$$

$$f_p \left(t_p \left(f \right) + \pi_p - \lambda_w \right) = 0 \qquad \forall p \in P_w, w \in W \tag{3.3}$$

$$t_p \left(f \right) + \pi_p - \lambda_w \geq 0 \qquad \forall p \in P_w, w \in W \tag{3.4}$$

$$f_p \geq 0 \qquad \forall p \in P_w, w \in W \tag{3.5}$$

$$\pi_p \geq 0 \qquad \forall p \in P_w, w \in W \tag{3.6}$$

$$\sum_{w \in W} \sum_{p \in P_w} \delta_{ap} f_p = \bar{x}_a \qquad \forall a \in A \tag{3.7}$$

where $\bar{x}_a$ is the system optimum link flow on link a. In the above, the objective function is to minimize total toll revenue. Equation (3.2) is to ensure flow conservation; Equations (3.3) and (3.4) are tolled user equilibrium conditions; Equations (3.5) and (3.6) specify non-negative path flow and toll, and the last constraint requires link flows to replicate system optimum link flows.

The above formulation can be easily modified for the other two differentiated schemes. In origin-specific and OD-specific schemes, tolls are imposed on links, but can be different for different origins or OD pairs. In our formulation, we associate a superscript to toll variables, γ, to differentiate tolls. Subsequently, adding

the following constraints to the above model yields a formulation for origin-specific pricing:

$$\pi_p = \sum_{a \in A} \delta_{ap} \gamma_a^{o(w)} \qquad \forall p \in P_w, w \in W \tag{3.8}$$

$$\gamma_a^{o(w)} \geq 0 \qquad \forall a \in A, w \in W \tag{3.9}$$

where $\gamma_a^{o(w)}$ is the toll on link a for users from origin $o(w)$.

Similarly, the formulation for OD-specific pricing can be obtained by adding the following constraints:

$$\pi_p = \sum_{a \in A} \delta_{ap} \gamma_a^{w} \qquad \forall p \in P_w, w \in W$$

$$\gamma_a^{w} \geq 0 \qquad \forall a \in A, w \in W$$

where γ_a^w is the toll on link a for users of OD-pair w.

We now consider a second-best network condition where not all the links are tollable. In this case, anonymous tolling may not induce system optimum and thus price differentiation provides additional flexibility to further reduce system travel time. Below we present a formulation to obtain a second-best origin-specific pricing scheme that minimizes system travel time:

$$\min \quad \sum_{w \in W} \sum_{p \in P_w} t_p(f) f_p \tag{3.10}$$

$$s.t.$$

$$(3.2), (3.3), (3.4), (3.5), (3.8), (3.9)$$

$$\gamma_a^{o(w)} = 0 \qquad \forall w \in W, a \in \bar{\Psi} \tag{3.11}$$

The OD-specific formulation can be developed similarly. Notice that because path-based pricing does not impose tolls on links, it becomes irrelevant here.

Comparing the above with the first-best formulations, Equation (3.7) is no longer included because system optimum link flows may not be achievable. In addition, because only specific links can be tolled, Constraint (3.11) is added. We further note that link-based formulations for the origin-specific and OD-specific schemes exist, but we do not present them to keep the chapter concise.

Because of Constraints (3.3)-(3.5), the formulations presented above all belong to the class of mathematical programs with complementarity constraints (MPCC). These problems are non-convex and standard stationary conditions, i.e., KKT conditions, may not hold for them because they do not satisfy Magasarian-Fromovitz constraint qualification [52]. Many solution algorithms have been proposed for MPCC (see, e.g. Luo et al. [42] and references cited therein). However, some only work well for small and medium problems while others, especially those based on solving equivalent nonlinear programs (e.g., Lawphongpanich and Yin [35]), can handle larger problems. More efficient algorithms may be developed to solve the above formulations by exploring special properties or structures that they may possess. For example, Zangui et al. [65] reformulated the first-best path-based pricing problem as a concave minimization problem and developed an efficient algorithm to solve it.

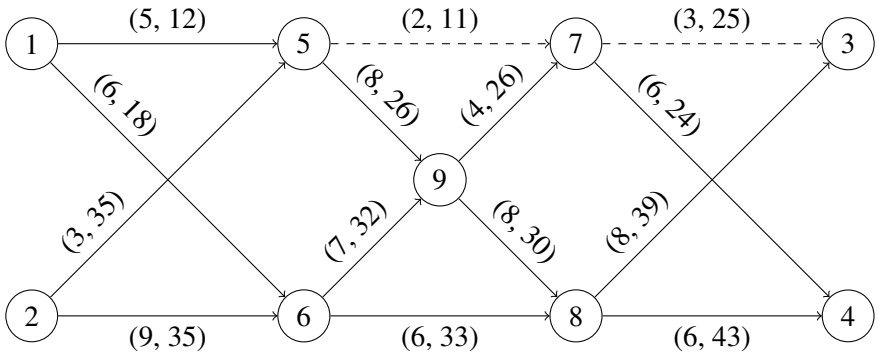

Figure 3.1: Nine-node network

3.2.3 Illustrative Examples

We now demonstrate the potentials of differentiated pricing schemes on a nine-node network. Figure 3.1 shows the network with four OD pairs $[1, 3]$, $[1, 4]$, $[2, 3]$, and $[2, 4]$, whose demands are 10, 20, 30 and 40, respectively. The link performance functions are of the following form:

$$t_a(x_a) = T_a \left(1 + 0.15 \left(\frac{x_a}{b_a}\right)^4\right)$$

where T_a and b_a are provided in Figure 3.1 as (T_a, b_a) near each link .

Table 3.1: Differentiated pricing with all links tollable

Tolling Scheme	Toll Revenue		OD Generalized Travel Cost			
	Amount	Reduction	[1, 3]	[1, 4]	[2, 3]	[2, 4]
User Equilibrium	-	-	24.9	23.8	24.3	25.1
Anonymous	887.6	0%	30.6	29.2	33.0	31.6
Origin-specific	311.6	65%	23.4	29.3	25.8	24.4
OD-specific	295.6	67%	23.4	22.0	25.8	27.6
Path-based	263.6	70%	23.4	22.0	29.0	24.4

Table 3.1 presents the results[1] of different levels of differentiation when all links are tollable. The second and third columns show the minimum toll revenue of each scheme, and the percent reduction as compared to the anonymous scheme. It can be observed that the toll revenue for all differentiated schemes are substantially lower than that of anonymous pricing. Moreover, as the level of differentiation increases, the revenue decreases. Particularly, price differentiation with respect to path yields a 70% reduction in revenue. The last four columns present the equilibrium travel cost for each OD pair. Observe that other than

[1]Results are the best obtained ones, but likely local optima. This note applies to other tables in this chapter.

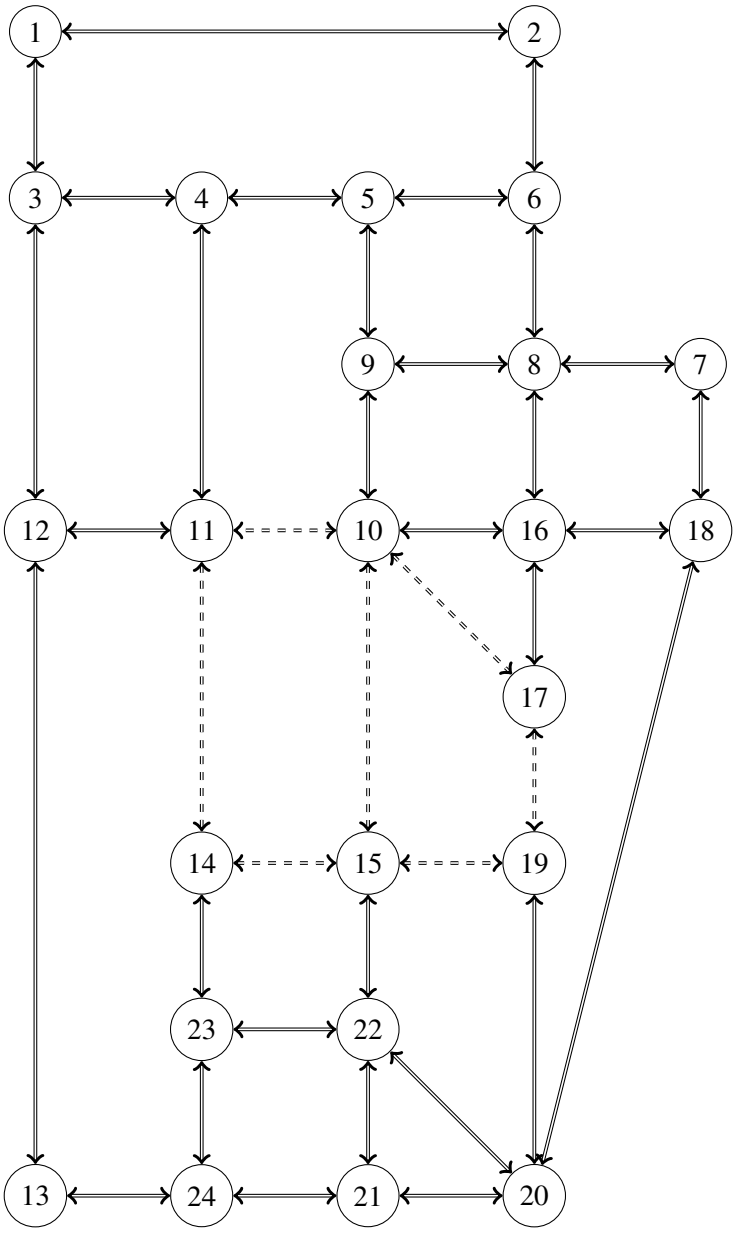

Figure 3.2: Sioux Falls network

OD-pair [1, 4] under origin-specific scheme, the travel costs under differentiated schemes are less than those under the anonymous scheme, suggesting that differentiated pricing may be more appealing to individual travelers in this network.

Table 3.2: Second-best differentiated pricing for nine-node network

Tolling Scheme	Total Travel Time		OD Generalized Travel Cost			
	Amount	Saving	[1, 3]	[1, 4]	[2, 3]	[2, 4]
User Equilibrium	2455.9	0%	24.9	23.8	24.3	25.1
Anonymous	2361.2	46.9%	25.8	24.9	25.1	25.9
Origin-specific	2306.1	74.2%	24.3	24.2	27.1	25.7
OD-specific	2281.7	86.2%	24.4	22.9	26.8	25.3
System Optimum	2253.9	100%	-	-	-	-

Table 3.2 presents the results of solving second-best differentiated pricing for the nine-node network, when only links (5,7) and (7,3) are tollable. In this table, the second column shows the total system travel time under each tolling scheme. Knowing that system optimum yields the smallest system travel time, we present the third column as the ratio between travel time reduction from user equilibrium and the maximum possible reduction, i.e., the difference in travel times of user equilibrium and system optimum. It is evident that price differentiation leads to additional travel time reduction. Specifically, even with only two links being tollable, the OD-specific tolling scheme achieves 86.2% of the maximum possible reduction, a reduction achieved by a first-best pricing scheme that may toll all links.

We also solved for differentiated schemes on the Sioux Falls network [38] as shown in Figure 3.2 and the results are presented in Table 3.3. For second-best pricing, only the dashed links in Figure 3.2 are assumed to be tollable. Table 3.3 shows that first-best differentiated pricing yields a substantial reduction in toll revenue, while minimizing system travel time. Similarly, the second-best pricing schemes offer promising results. Using system optimum as the benchmark, the additional travel time under user equilibrium is 2.859. The origin-specific scheme can reduce the additional time to 1.117, which is equivalent to a 60.93% reduction. Compared to anonymous tolling, origin-specific tolls can achieve twice the travel time saving.

Table 3.3: Differentiated pricing for Sioux Falls network

TollingScheme	First-best		Second-best	
	Min. Revenue	Reduction	Travel Time	Saving
Anonymous	23.441	0.00%	74.043	26.55%
Origin-specific	0.750	96.80%	73.060	60.93%
OD-specific	0.616	97.37%	72.997	63.13%
Path-based	0.182	99.23%	-	-
User Equilibrium	-	-	74.802	0.00%
System Optimum	-	-	71.943	100.00%

In general, for both the first-best and second-best conditions, higher levels of differentiation lead to more favorable results. On the other hand, differentiated pricing schemes are more difficult to implement. Such a trade-off needs to be investigated for a network of interest to determine whether a higher level of

differentiation is worth implementing or not.

3.3 Location Privacy

One of the major reasons for the implementation difficulty of differentiated pricing is potential violation of motorists' location privacy. Location privacy is defined as the ability to prevent other parties from learning one's current or past location [11]. The issue of location privacy commonly arises when offering a service requires some sort of location data. The issue has been mostly studied for situations where mobile applications or computer programs need to know the location of a user (e.g., Cvrcek et al. [14]).

The traditional way of manually collecting toll preserves location privacy almost completely. Needless to say, it is not an efficient way to collect toll, as vehicles have to stop and pay. Electronic toll collection (ETC) systems have been built to make toll collection more efficient, but the way they currently operate may compromise motorists' privacy rights [51]. The systems often link motorists' accounts and record locations and times of transactions (e.g., the Sunpass prepaid toll program in Florida [22]). If toll gantries are ubiquitous, the recorded transaction information may impinge on the privacy rights of motorists. However, those who are concerned about their location privacy have the option to pay the toll by cash and avoid risk of privacy disclosure. Moreover, for anonymous link-based tolling, it is possible to design a privacy-preserving ETC system(e.g., Balasch et al. [8]) .

Unfortunately, it is difficult, if not impossible, to design a privacy-preserving differentiated pricing system, because the system requires the knowledge of travelers' travel characteristics such as the origin and destination of each trip for an OD-specific scheme. Golle and Partridge [25] and Krumm [33] pointed out that the home/work location data, even if they are anonymous, can be used to identify individuals. In addition to this, the sole fact of being tracked by the tolling system can cause inconvenience or discomfort. All these privacy concerns need to be addressed.

On the other hand, there have been some indications that motorists, some at a price, are willing to provide private information with the understanding that it will not be published and/or misused. For example, in the Travel Choices Study completed by the Puget Sound Regional Council [49], each participant was given a $1016 debit account with a GPS-based on-board unit installed on his car. This unit tracks and records when and where the participants drive and deducts tolls from the account. The money remaining in each account at the end of the study was given to the study participant. In this example, private information was collected for the purpose of tolling and with full knowledge of study participants. We surmise that the participants may be attracted to the $1016 incentive when joining the study.

Empirical experiments in the literature have proved that individuals value their location privacy differently. They can be grouped into categories of privacy unconcerneds, privacy pragmatists, and privacy fundamentalists [33]. The first group do not care about location privacy and are insensitive to the negative consequences of location leak. The second group are willing to reveal their location for a, sometimes very small, price, while the last group highly value and strive to protect their location privacy. Mathematically, we can use a distribution to represent different individual valuations of privacy across the population. Acquisti et al. [5] suggested a U-shaped distribution, but cautioned that the value of privacy can be very malleable and many non-normative factors may affect this distribution (also see Cvrcek et al. [14]). Hence, we base our models not on any specific, but on a general, distribution for value of privacy. Nevertheless, it is important to understand the implication of a proposed distribution. As an example, Figure 3.3 shows the probability

density function of logistic distributions with different parameters, which offer computational advantages as we have seen in the choice modeling literature. However, a logistic distribution implies that some users will have a negative value of privacy, an unjustifiable assumption.

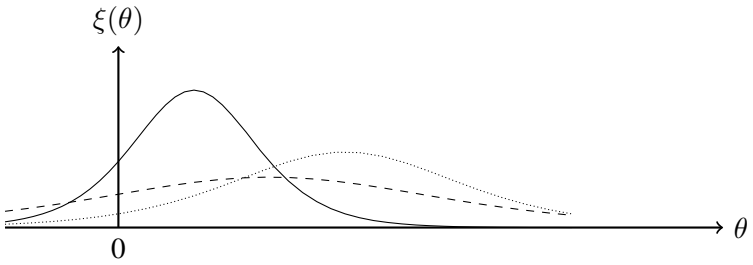

Figure 3.3: Logistic distributions with different parameters

Figure 3.4 illustrates more reasonable uniform and exponential distributions, both with a mean of two. In this figure, $U(0,4)$ denotes a uniform distribution between 0 and 4, and $EXP(0.5)$ is an exponential distribution with a parameter of 0.5. Notice that the exponential distribution is more clustered around smaller values, which implies that more users value their privacy less. But, as illustrated in Figure 3.5, exponential distributions with higher mean values become more evenly distributed. Also notice that the span for an exponential distribution is all nonnegative real numbers, while the uniform distribution is bounded on both sides. So, uniform distribution implies that the value of privacy of travelers is evenly distributed and has an upper bound. On the other hand, the exponential distribution suggests that some travelers are extreme privacy fundamentalists and will not disclose their locations at any price.

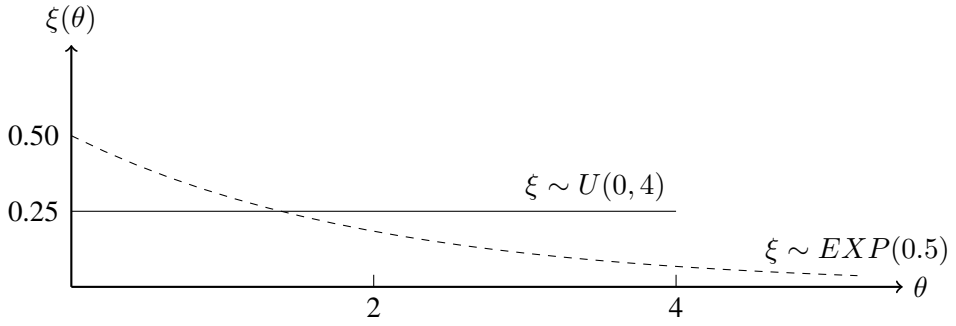

Figure 3.4: Uniform and exponential distributions with same mean, $E(\theta) = 2$

3.3.1 Modeling Privacy

We now use origin-specific pricing as an example for modeling privacy. Denote the travel cost between OD pair w under the scheme as $\lambda_{w,1}$, which consists of travel time and toll. Since they are being tracked, motorists incur additional cost for the loss of their location privacy, which we call privacy cost. Mathematically, the full cost for a traveler between OD pair w under origin-specific pricing is $\lambda_{w,1} + \beta$, where β is a random

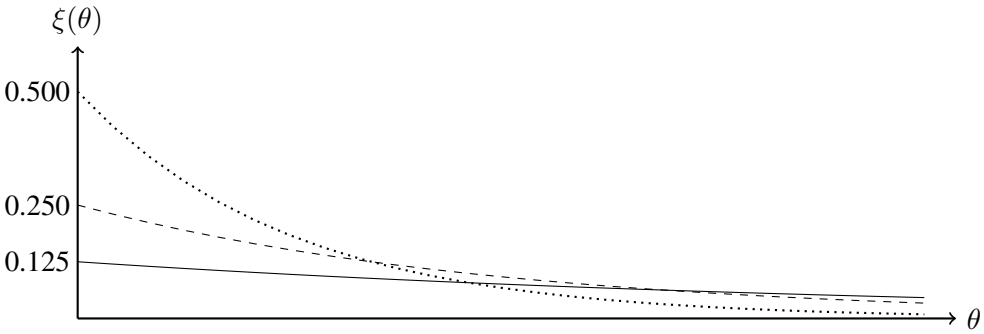

Figure 3.5: Exponential distributions with different means

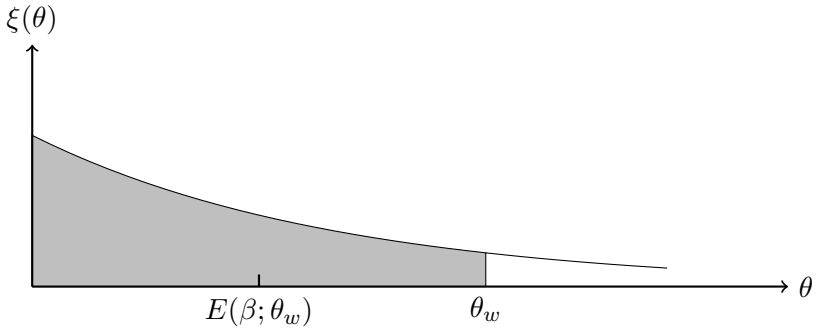

Figure 3.6: Expected privacy cost

variable representing the value of privacy, which is also expressed in the unit of time for simplicity. In this chapter, we assume the distribution of value of privacy is the same for travelers from each OD pair, but this assumption can easily be relaxed.

Suppose the value of privacy follows a known distribution ξ, i.e., $\beta \sim \xi$. Define $\Xi(\theta) = \int_0^\theta \xi(z)dz$ as the cumulative distribution function associated with the value of privacy, i.e., Prob$(\beta \leq \theta) = \Xi(\theta)$. Denote the travel cost between OD pair w under anonymous tolling as $\lambda_{w,0}$ and let $\theta_w = \lambda_{w,0} - \lambda_{w,1}$. If θ_w is negative, no one prefers the origin-specific scheme. As we observe from Tables 1 and 2, θ_w is likely positive. In this case, travelers who value their privacy less than θ_w would prefer the origin-specific scheme to the anonymous one, while those with higher value of privacy would prefer anonymous tolling. The percentage of the former is $\Xi(\theta_w)$ while it is $1-\Xi(\theta_w)$ for the latter. Figure 3.6 illustrates this situation for a hypothetical distribution of the value of privacy, where the shaded area represents the percentage of travelers who will be better off and thus prefer the origin-specific scheme. Their total privacy cost, denoted as $PC_w(\theta_w)$, can be computed as follows: $PC_w(\theta_w) = \int_0^{\theta_w} d_w \xi(z) z dz$, where d_w is the total demand between OD pair w. Define $E(\beta;\theta_w) = \int_0^{\theta_w} z\xi(z)dz$, and the equation is thus written as $PC_w(\theta_w) = d_w E(\beta;\theta_w)$. We will use this in Section 3.4.

3.3.2 Note on General Distributions

The calculations of $\Xi(\theta)$ and $E(\beta;\theta)$ involve integration. If the value of privacy follows a uniform or an exponential distribution, the integrals will have a closed form as shown in Table 3.4.

Table 3.4: Descriptors for uniform and exponential distributions of value of privacy

	$U(0,\alpha)$	$EXP(\alpha)$
$\xi(\theta)$	$1/\alpha$	$\alpha e^{-\alpha\theta}$
$\Xi(\theta)$	θ/α	$1-e^{-\alpha\theta}$
$E(\beta;\theta)$	θ^2/α	$(1-e^{-\alpha\theta}(\alpha\theta+1))/\alpha$

In a general case where the integrals do not have a closed form, we need to compute them via numerical integration methods. One of these methods is Riemann sum, which approximates the area under a curve by vertical bars as illustrated in Figure 3.7. Here, we use the following Riemann sums to approximate the value of integrals:

$$\Xi(\theta) = \int_0^\theta \xi(z)dz = \frac{1}{n}\theta\sum_{i=1}^{n}\xi(\frac{i\theta}{n})$$

$$E(\beta;\theta) = \int_0^\theta z\xi(z)dz = \frac{1}{n}\theta\sum_{i=1}^{n}\left(\frac{i\theta_w}{n}\xi(\frac{i\theta}{n})\right)$$

where n is the number of bars. Choosing a larger n would result in a higher precision.

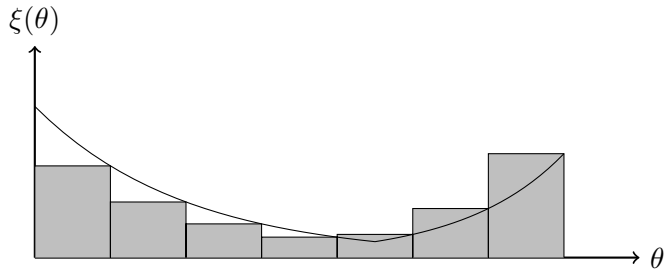

Figure 3.7: Riemann approximation for integral (n=7)

3.3.3 Privacy Analysis of Differentiated Schemes

In this section, we examine the results of the nine-node network in Section 3.2.3 from a privacy perspective. Table 3.5 calculates the percentages of travelers between each OD pair who would benefit from origin-specific pricing after considering privacy cost, i.e., $100\Xi(\theta_w)$, under different hypothetical distributions for the value of privacy.

The third row of Table 3.5 shows the saving of time and toll for travelers between each OD pair, i.e. $\theta_w = \lambda_{w,0} - \lambda_{w,1}$. It can be observed that, even if the average value of privacy is high, some travelers still benefit

Table 3.5: Percentage of travelers who benefit from origin-specific pricing on nine-node network

Network Condition	First-best				Second-best			
OD pair	[1,3]	[1,4]	[2,3]	[2,4]	[1,3]	[1,4]	[2,3]	[2,4]
Travel Cost Saving	7.2	-0.1	7.2	7.2	1.5	0.7	-2.0	0.2
$\beta \sim U(0,4)$	100.00	0.00	100.00	100.00	37.50	17.50	0.00	5.00
$\beta \sim U(0,8)$	90.00	0.00	90.00	90.00	18.75	8.75	0.00	2.50
$\beta \sim U(0,16)$	45.00	0.00	45.00	45.00	9.37	4.38	0.00	1.25
$\beta \sim EXP(0.500)$	97.27	0.00	97.27	97.27	52.76	29.53	0.00	9.52
$\beta \sim EXP(0.250)$	83.47	0.00	83.47	83.47	31.27	16.05	0.00	4.88
$\beta \sim EXP(0.125)$	59.34	0.00	59.34	59.34	17.10	8.38	0.00	2.47

from differentiated schemes. However, the percentage decreases as the average value of privacy increases. Also observe that when travel cost saving is small, exponential distributions predict higher percentages of users who will benefit from differentiated schemes, because the distributions are more clustered around smaller values.

Apparently, the savings of time and toll that some travelers enjoy from differentiated schemes are offset by the loss of their privacy. Section 3.4 presents a way to take advantage of the potentials of differentiated pricing, while allowing those concerned travelers to maintain their privacy.

3.4 Addressing Privacy Concerns with an Incentive Program

Recognizing that some may benefit from differentiated schemes while others with higher value of privacy may be better off under anonymous tolling, we propose to develop an incentive program for travelers to opt in to differentiated pricing. More specifically, a hybrid of anonymous and differentiated pricing schemes will be implemented on the network. Travelers who choose to reveal their private information will pay differentiated tolls while those who remain anonymous will pay uniform tolls.

Since travel costs (time plus toll) in differentiated schemes are generally less than those in the anonymous scheme, the cost savings can be viewed as incentives for drivers to participate in differentiated pricing. Although other incentives, such as subsidies or credits, can be provided, below we focus on designing anonymous and differentiated tolls in the hybrid scheme and allowing for the cost savings as incentives. The overall goal of this hybrid scheme is to create a win-win situation for both users and society.

3.4.1 Design of Incentive Program

As an example, we design the incentive program for a hybrid of origin-specific and anonymous tolls. The formulations for other hybrid schemes can be developed with some straightforward modifications and we do not present them to keep the chapter concise.

It is reasonable to assume all the motorists who are better off under an origin-specific scheme will opt in to this scheme. Thus, the number of these motorists will be $d_{w,1} = \Xi(\lambda_{w,0} - \lambda_{w,1})d_w$. Travelers who choose the anonymous scheme will not incur any privacy cost. Hence, the total privacy cost for travelers between OD pair w is equal to $PC_w(\lambda_{w,0} - \lambda_{w,1})$.

The following constraints define the feasible region of the problem:

$$d_{w,0} + d_{w,1} = d_w \qquad \forall w \in W \tag{3.12}$$

$$\sum_{p \in P_w} f_{p,c} = d_{w,c} \qquad \forall w \in W, c \in C \tag{3.13}$$

$$d_{w,1} = \Xi(\lambda_{w,0} - \lambda_{w,1}) d_w \qquad \forall w \in W \tag{3.14}$$

$$f_{p,c}\left(t_p\left(f\right) + \pi_{p,c} - \lambda_{w,c}\right) = 0 \qquad \forall p \in P_w, w \in W, c \in C \tag{3.15}$$

$$t_p\left(f\right) + \pi_{p,c} - \lambda_{w,c} \geq 0 \qquad \forall p \in P_w, w \in W, c \in C \tag{3.16}$$

$$\lambda_{w,0} \geq \lambda_{w,1} \qquad \forall w \in W \tag{3.17}$$

$$f_{p,c} \geq 0 \qquad \forall p \in P_w, w \in W, c \in C \tag{3.18}$$

$$\pi_{p,c} \geq 0 \qquad \forall p \in P_w, w \in W, c \in C \tag{3.19}$$

$$\pi_{p,0} = \sum_{a \in A} \delta_{ap} \gamma_a \qquad \forall p \in P_w, w \in W \tag{3.20}$$

$$\gamma_a \geq 0 \qquad \forall a \in A \tag{3.21}$$

$$\pi_{p,1} = \sum_{a \in A} \delta_{ap} \gamma_a^{o(w)} \qquad \forall p \in P_w, w \in W \tag{3.22}$$

$$\gamma_a^{o(w)} \geq 0 \qquad \forall a \in A, w \in W \tag{3.23}$$

where $C = \{0, 1\}$. Constraints (3.12) and (3.14) split the demand for each OD pair. Constraint (3.13) ensures flow balance. The tolled user equilibrium is guaranteed by Constraints (3.15) and (3.16). Constraint (3.17) requires travel cost (time plus toll) in the origin-specific scheme to be less than that in the anonymous scheme. Constraints (3.20) and (3.22) make the toll on each path to be equal to the sum of link tolls. Denote the feasible region defined by the above constraints as Φ.

We first discuss the problem of finding the optimal hybrid scheme in the first-best network setting where all the links are tollable. In this situation, we are interested in replicating the flow distribution with minimum system travel time as well as minimizing the sum of toll revenue and privacy cost as a secondary objective. The following is the total (full) user cost:

$$\sum_{w \in W} \left(PC_w(\lambda_{w,0} - \lambda_{w,1}) + \sum_{p \in P_w} \left(\pi_{p,0} f_{p,0} + \pi_{p,1} f_{p,1}\right) \right) + \sum_{a \in A} x_a t_a(x_a)$$

Since $x_a = \bar{x_a}$ has to be achieved in first-best pricing, the last term is a constant and can be omitted from the optimization. Consequently, we have the following formulation for finding an optimal hybrid scheme in

a network with all links being tollable:

$$\min \quad \sum_{w \in W} \left(PC_w(\lambda_{w,0} - \lambda_{w,1}) + \sum_{p \in P_w} (\pi_{p,0} f_{p,0} + \pi_{p,1} f_{p,1}) \right) \tag{3.24}$$

$$s.t.$$

$$(f, d, \pi, \lambda) \in \Phi$$

$$\sum_{w \in W} \sum_{p \in P_w} \delta_{a,p} \left(f_{p,0} + f_{p,1} \right) = \bar{x}_a \qquad \forall a \in A$$

where the last constraint is to ensure the link flows to be the least-system-time flows.

We now consider a second-best situation when only some links are tollable. In this situation, we attempt to minimize total system cost, which differs from the above total (full) user cost by the toll revenue, because the revenue is not a cost for the system but a transfer from travelers to the government. The problem of finding an optimal hybrid scheme can be formulated as follows:

$$\min \quad \sum_{w \in W} \left(PC_w(\lambda_{w,0} - \lambda_{w,1}) + \sum_{p \in P_w} t_p(f) f_p \right) \tag{3.25}$$

$$s.t.$$

$$(f, d, \pi, \lambda) \in \Phi$$

$$\gamma_a^{o(w)} = 0 \qquad \forall w \in W, a \in \bar{\Psi}$$

The last constraint ensures that only tollable links can have positive amount of toll.

Notice that the above formulations are path-based, and solving them requires path enumeration. However, it is possible to formulate them as link-based models. We use the above path-based formulations to facilitate the presentation.

3.4.2 Numerical Examples

The proposed models for designing the incentive program of the origin-specific scheme were implemented on the nine-node and Sioux Falls networks. Each model was solved for both uniform and exponential distributions of value of privacy, each with three different expected values.

Table 3.6 presents the results on the nine-node network of Figure 3.1 with all the links being tollable. The performances of the hybrid schemes are also compared with those of the anonymous and origin-specific tolls when implemented separately.

As pointed out previously, origin-specific pricing can reduce the toll revenue significantly in a first-best network condition. However, this reduction comes with a price of violating travelers' privacy. Since origin-specific schemes require all the users to reveal their origin information, the privacy cost is equal to the expected value of privacy times the total demand. The privacy cost increases as travelers value their privacy more, eventually causing the total user cost under origin-specific pricing to be larger than that under anonymous tolling, when the expected value of privacy is equal to 8. In contrast, the hybrid scheme offers

Table 3.6: Comparison of different schemes on nine-node network (all links tollable)

Pricing Scheme	Distribution of β	$E(\beta)$	Toll Rev.	Privacy Cost	Total User Cost
Anonymous	-	-	887.60	0.00	887.60
Origin-specific	-	2	311.60	200.00	511.60
	-	4	311.60	400.00	711.60
	-	8	311.60	800.00	1111.60
Hybrid	$U(0,4)$	2	247.82	28.46	276.28
	$U(0,8)$	4	235.25	58.47	293.72
	$U(0,16)$	8	220.76	116.96	337.72
	$EXP(0.500)$	2	249.84	17.49	267.33
	$EXP(0.250)$	4	237.43	35.52	272.95
	$EXP(0.125)$	8	213.08	71.02	284.10

an option for travelers of high value of privacy to remain anonymous. Such a self-selection mechanism leads to much less loss of privacy and subsequently less total user cost. Interestingly, in this example, the hybrid schemes also lead to less amount of toll revenue than their origin-specific counterparts. However, this observation need not be generally true.

Table 3.7: Comparison of different schemes on nine-node network (two tollable links)

Pricing Scheme	Distribution of β	$E(\beta)$	Travel Time	Privacy Cost	Total System Cost
Anonymous	-	-	2361.16	0.00	2361.16
Origin-specific	-	2	2306.10	200.00	2506.10
	-	4	2306.10	400.00	2706.10
	-	8	2306.10	800.00	3106.10
Hybrid	$U(0,4)$	2	2291.79	9.13	2300.92
	$U(0,8)$	4	2296.76	13.08	2309.84
	$U(0,16)$	8	2304.63	17.57	2322.20
	$EXP(0.500)$	2	2291.45	5.82	2297.27
	$EXP(0.250)$	4	2293.47	9.56	2303.04
	$EXP(0.125)$	8	2299.10	13.30	2312.40

We solved for the anonymous, origin-specific and hybrid schemes on the nine-node network when only two specific links, (5,7) and (7,3), are tollable. Table 3.7 displays the results for each scheme. As expected, in every case, the total cost under the hybrid scheme is less than those in the anonymous and origin-specific schemes. Interestingly, the hybrid schemes also yield even less total travel time than the origin-specific scheme, even though the latter is to minimize total travel time while the former is to minimize the total travel time plus privacy cost. This is not surprising because mathematically, the feasible regions of these two models are not the same; and intuitively, a hybrid scheme provides an additional dimension of flexibility. However, this observation does not necessarily apply to other networks. Also, notice that the total privacy cost associated with the uniform distribution is higher than exponential distribution with the same expected value.

Table 3.8: Second-best hybrid schemes on Sioux Falls network

Pricing Scheme	Distribution of β	$E(\beta)$	Travel Time	Privacy Cost	Total System Cost
Anonymous	-	-	74.043	0.000	74.043
Origin-specific	-	0.02	73.060	7.212	80.272
	-	0.04	73.060	14.424	87.474
	-	0.08	73.060	28.848	101.908
Hybrid	$U(0, 0.04)$	0.02	73.294	0.118	73.412
	$U(0, 0.08)$	0.04	73.421	0.138	73.421
	$U(0, 0.16)$	0.08	73.591	0.163	73.753
	$EXP(50.0)$	0.02	73.272	0.086	73.357
	$EXP(25.0)$	0.04	73.355	0.106	73.461
	$EXP(12.5)$	0.08	73.455	0.163	73.618

To demonstrate the models on a more realistic network, we solved them on the Sioux Falls network where the tollable links are the dashed ones in Figure 3.2. The obtained results are presented in Table 3.8. Similar to the nine-node network, the privacy cost and total cost increase as the expected value of privacy increases. Also, the privacy cost and total cost under the exponential distributions is less than those associated with the uniform distributions.

The results in this section illustrate the potentials of the incentive program for origin-specific pricing. For two extreme cases, with the value of privacy being zero or infinity, the hybrid scheme yields the same results as differentiated or anonymous scheme, respectively. But, in the real world, this value should be finite and positive. Our results indicate that the performance of the incentive program is much better when the expected value of privacy is relatively low, i.e., more users are willing to reveal their information for a small amount of money (previous empirical studies seem suggest so). The incentive program also demonstrates promising results when the expected value of privacy is relatively higher. While this section only focuses on a hybrid of origin-specific and anonymous tolls, we expect other hybrids to perform favorably in a similar fashion.

Chapter 4

Conclusion and Discussion

In this report, we have discussed "point-to-point" traffic volume monitoring under the context of connected vehicle systems. To protect drivers' location privacy, we have formalized the problem of secure OD flow measurement with a realistic threat model. We have proposed a novel secure OD flow measurement scheme that exploits the nice properties of commutative one-way hash functions to protect drivers' privacy. Our solution allows the authority to collect "statistical" OD flow information, but prevents it from learning identities of "individual" vehicles. We have used sampling to improve computation efficiency without significantly degrading measurement accuracy. Simulations have demonstrated the feasibility and scalability of our scheme.

This report has also explored a new class of tolling schemes that charge different amount of toll for users with different origins, destinations, or paths. These schemes provide more flexibility than traditional anonymous pricing and the numerical examples in this report have demonstrated that they can reduce the financial burden on motorists in a first-best network condition or lead to more travel time saving in a second-best condition.

Recognizing that the differentiated pricing may compromise travelers' privacy, we have proposed an incentive program to allow travelers to opt in to the differentiated pricing, if they find the amount of incentive to worth disclosing their location information. This self-selection mechanism allows the tolling agency to take (potential) advantages of differentiated pricing without doing harm to travelers' privacy rights.

Other approaches can be explored to mitigate privacy concerns associated with differentiated pricing. For instance, instead of charging users based on their true origins, the tolling agency can designate a tolling area and then charge users based on where they enter the area. Because the true origins are not revealed, this scheme may partially mitigate travelers' privacy concern. Note that this scheme is different from the traditional cordon pricing in which motorists pay a uniform toll to cross the cordon. In the refined scheme, motorists on a link within the tolling area will pay different amount of toll depending on where they enter the tolling area. We call this scheme as a sub-network origin-specific pricing scheme. To illustrate the concept, consider the network in Figure 4.1 where the tolling area consists of the dashed links. Consider three different paths, $p_1 : 2 \to 6 \to 5 \to 7 \to 3$, $p_2 : 2 \to 1 \to 5 \to 7 \to 3$, and $p_3 : 1 \to 5 \to 7 \to 3$. While in the original origin-specific scheme, travelers on p_1 and p_2 will pay the same amount of toll on link (5,7), they may pay different amount of toll for traversing the link under the refined scheme, because they enter the tolled sub-network from different nodes (Nodes 6 and 5 respectively). Also, unlike in the original

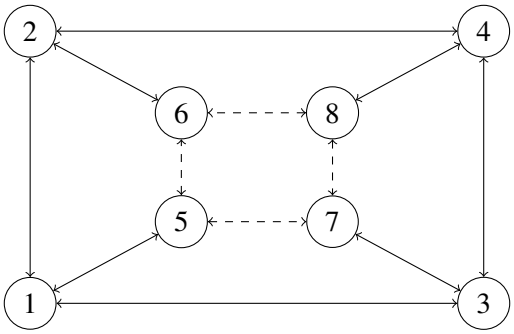

Figure 4.1: An illustrative network

origin-specific pricing, motorists on p_2 and p_3 will pay the same amount of toll for link (5,7) because they both enter the sub-network from Node 5.

We implemented the sub-network origin-specific scheme for the Sioux Falls network (Fig. 3.2), where the tolling area consists of dashed links and nodes 10, 11, 14, 15, 17 and 19. The best design yields a system travel time of 73.215, which is slightly greater than the system travel time of 73.060 under the true origin-specific scheme. In this case, the refined sub-network origin-specific pricing is very promising. An interesting future study can be conducted to explore how to select the tolling area to achieve a similar performance as the true origin-specific tolling.

Bibliography

[1] URL http://docs.oracle.com/javase/1.4.2/docs/api/java/math/BigInteger.html.

[2] URL http://dataprivacylab.org/dataprivacy/papers/ppdm/.

[3] URL http://www.dot.gov/.

[4] URL http://www.its.dot.gov/press/2010/vii2intellidrive.

[5] A. Acquisti, L. John, and G. Loewenstein. What is Privacy Worth? In *Twenty First Workshop on Information Systems and Economics (WISE)*, December 2000.

[6] R. Agrawal and R. Srikant. Privacy-Preserving Data Mining. *Proc. of the 2000 ACM SIGMOD International Conference on Management of Data*, 29:439–450, June 2000.

[7] A. Amanna. Overview of IntelliDrive/Vehicle Infrastructure Integration (VII). Technical report, Virginia Tech Transportation Institute, 2009.

[8] J. Balasch, I. Verbauwhede, and B. Preneel. An Embedded Platform for Privacy-friendly Road Charging Applications. In *DATE'10*, pages 867–872, 2010.

[9] X. Ban and M. Gruteser. Mobile Sensors as Traffic Probes: Addressing Transportation Modeling and Privacy Protection in an Integrated Framework. In *Proceedings of the 7th International Conference on Traffic and Transportation Studies*, pages 750–767, Kunming, China, 2010.

[10] J. Benaloh and M. D. Mare. One-way Accumulators: a Decentralized Alternative to Digital Signatures. *Proc. of EUROCRYPT '93 Workshop on the theory and application of cryptographic techniques on Advances in cryptology*, pages 274–285, 1993.

[11] A. Beresford and F. Stajano. Location Privacy in Pervasive Computing. *Pervasive Computing, IEEE*, 2(1):46 – 55, jan-mar 2003.

[12] M. Bozic, D. Kopic, F. Mihoci, and I. Kamber. Traffic counting on the roadways of Croatia in 2009 - digest. *HRVATSKE CESTE d.o.o. Management, construction and maintenance of State roadways*, 2010.

[13] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu. Tools for Privacy Preserving Distributed Data Mining. *ACM SIGKDD Explorations Newsletter*, 4:28–34, December 2002.

[14] D. Cvrcek, M. Kumpost, V. Matyas, and G. Danezis. A Study on the Value of Location Privacy. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, WPES '06, pages 109–118, New York, NY, USA, 2006. ACM.

[15] A. de Palma and R. Lindsey. Congestion Pricing with Heterogeneous Travelers: A General-Equilibrium Welfare Analysis. *Networks and Spatial Economics*, 4:135–160, 2004.

[16] F. Dion and R. Robinson. Sweden/Michigan Naturalistic Field Operational Test-Phase 1: Benefits of Origin and Destination Information in IntelliDrive Data Sets. Technical Report June, University of Michigan Transportation Research Institute, 2009.

[17] J. Dupuit. On Tolls and Transport Charges. *Annales des Pont et Chaussees*, 17:445–467, 1894. Also translated in International Economic Papers, Macmillan, London, 1952.

[18] J. K. Eom, M. S. Park, T. Heo, and L. F. Huntsinger. Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method. *Journal of the Transportation Research Board*, 1968/2006:20–29, 2006.

[19] J. Eriksson, H. Balakrishnan, and S. Madden. Cabernet: Vehicular Content Delivery Using WiFi. *Proc. of the 14th ACM MOBICOM*, pages 199–210, March 2008.

[20] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy Preserving Mining of Association Rules. *Proc. of the 8th ACM SIGKDD*, pages 217–228, 2002.

[21] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting Privacy Breaches in Privacy Preserving Data Mining. *Proc. of the 22nd ACM Symposium on Principles of database systems*, pages 211–222, June 2003.

[22] Florida Department of Transportation. SunPass Prepaid Toll Program, January 2012. URL `http://www.sunpass.com`.

[23] M. J. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. *ADVANCES IN CRYPTOLOGY - EUROCRYPT*, 3027/2004(10.1007/978-3-540-24676-3_1):1–19, 2004.

[24] H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall, 2 edition, 2008. ISBN 0131873253.

[25] P. Golle and K. Partridge. On the Anonymity of Home/Work Location Pairs. In *Proceedings of the 7th International Conference on Pervasive Computing*, Pervasive '09, pages 390–397, Berlin, Heidelberg, 2009. Springer-Verlag.

[26] C. Hazay and Y. Lindell. EFFICIENT SECURE TWO-PARTY PROTOCOLS: Semi-honest Adversaries. *Information Security and Cryptography*, (10.1007/978-3-642-14303-8_3):53–80, 2010.

[27] D. W. Hearn and M. V. Ramana. Solving Congestion Toll Pricing Models. In P. Marcotte and S. Nguyen, editors, *Equilibrium and Advanced Transportation Modeling*, pages 109–124. Kluwer Academic, Boston, 1998.

[28] J. Holguin-Veras and M. Cetin. Optimal Tolls for Multi-class Traffic: Analytical Formulations and Policy Implications. *Transportation Research Part A: Policy and Practice*, 43(4):445 – 467, 2009.

[29] L. Jacobson. VII Privacy Policies Framework, Version 1.0.2. In *National VII Coalition*, 2007.

[30] W. Jiang, C. Clifton, and M. Kantarcioglu. Transforming semi-honest protocols to ensure accountability. *Data and Knowledge Engineering*, 65:57–74, April 2008.

[31] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the Privacy Preserving Properties of Random Data Perturbation Techniques. *Proc. of the 3rd IEEE ICDM*, pages 99–106, November 2003.

[32] C. Kaufman, R. Perlman, and M. Speciner. *Network Security, Private Communication in a Public World*. Prentice Hall, 2 edition, 2002. ISBN 0-13-046019-2.

[33] J. Krumm. A Survey of Computational Location Privacy. *Personal Ubiquitous Comput.*, 13(6):391–399, Aug. 2009.

[34] W. H. K. Lam and J. Xu. Estimation of AADT from Short Period Counts in Hong Kong – A Comparison Between Neural Network Method and Regression Analysis. *Journal of Advanced Transportation*, 34:249–268, 2000.

[35] S. Lawphongpanich and Y. Yin. Solving the Pareto-improving toll problem via manifold suboptimization. *Transportation Research Part C: Emerging Technologies*, 18(2):234 – 246, 2010.

[36] S. Lawphongpanich and Y. Yin. Solving the Pareto-improving toll problem via manifold suboptimization. *Transportation Research Part C: Emerging Technologies*, 18(2):234 – 246, 2010.

[37] S. Lawphongpanich and Y. Yin. Nonlinear Pricing on Transportation Networks. *Transportation Research Part C: Emerging Technologies*, 20(1):218 – 235, 2012.

[38] L. J. LeBlanc, E. K. Morlok, and W. P. Pierskalla. An Efficient Approach to Solving the Road Network Equilibrium Traffic Assignment Problem. *Transportation Research*, 9:309–318.

[39] U. Lee, J. Lee, J. Park, and M. Gerla. FleaNet: A Virtual Market Place on Vehicular Networks. *IEEE Trans. on Vehicular Technology*, 59(1):344–355, January 2010.

[40] T. Li, S. Chen, and Y. Ling. Fast and Compact Per-flow Traffic Measurement Through Randomized Counter Sharing. In *INFOCOM, 2011 Proceedings IEEE*, Shanghai, China, 2011.

[41] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. *Journal of Cryptology*, 15:177–206, April 2002.

[42] Z. Q. Luo, J. S. Pang, and D. Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, Cambridge ; New York :, 1996.

[43] M. R. McCord, Y. Yang, Z. Jiang, B. Coifman, and P. K. Goel. Estimating annual average daily traffic from satellite imagery and air photos: Empirical results. *Journal of the Transportation Research Board*, 1855/2003:136–142, 2003.

[44] A. J. Menezes, P. C. Oorschot, and S. A. Vanstone. *Handbook of Applied Cryptography*. CRC Press, 1996. ISBN 0-8493-8523-7.

[45] D. Mohamad, K. C. Sinha, T. Kuczek, and C. F. Scholer. Annual Average Daily Traffic Prediction Model for County Roads. *Journal of the Transportation Research Board*, 1617/1998:69–77, 1998.

[46] M. C. Neto, Y. Jeong, M. K. Jeong, and L. D. Han. AADT prediction using support vector regression with data-dependent parameters. *Expert Systems with Applications*, 36:2979–2986, March 2009.

[47] Noblis. Anonymity and IntelliDrive: Pre-Decisional Discussion Document, 2009. URL `http://www.its.dot.gov/research_docs/pdf/6Anonymity.pdf`.

[48] A. C. Pigou. *The Economics of Welfare*. London: Macmillan and Co., 4th edition, 1932.

[49] Puget Sound Regional Council. Travel choices study - a summary report, April 2008.

[50] R. Rivest, A. Shamir, and L. Adleman. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Communications of the ACM*, 21:120–126, 1978.

[51] T. Sager. Privacy as a Planning Problem: Some Transport-related Examples. *Scandinavian Housing and Planning Research*, 15(1):37–52, 1998.

[52] H. Scheel and S. Scholtes. Mathematical Programs with Complementarity Constraints: Stationarity, Optimality, and Sensitivity. *Math. Oper. Res.*, 25(1):1–22, Feb. 2000. ISSN 0364-765X.

[53] A. Shamir. On the Generation of Cryptographically Strong Pseudorandom Sequences. *ACM Transactions on Computer Systems*, 1:38–44, February 1983.

[54] K. A. Small and J. Yan. The Value of Value Pricing of Roads: Second-Best Pricing and Product Differentiation. *Journal of Urban Economics*, 49(2):310 – 336, 2001.

[55] U.S. Department of Transportation. Traffic Monitoring Guide. *Section 3, Traffic Volume Monitoring, http://www.fhwa.dot.gov/ohim/tmguide/tmg3.htm*, 2001.

[56] U.S. Department of Transportation. ITS Strategic Research Plan 2010-2014, 2010.

[57] J. Vaidya and C. Clifton. Secure Set Intersection Cardinality with Application to Association Rule Mining. *Journal of Computer Security*, 13:593–622, July 2005.

[58] H. R. Varian. *Microeconomic Analysis*. W. W. Norton & Company, 3rd edition, feb 1992.

[59] J. Y. Wang, R. Lindsey, and H. Yang. Nonlinear Pricing on Private Roads with Congestion and Toll Collection Costs. *Transportation Research Part B: Methodological*, 45(1):9 – 40, 2011.

[60] H. Yang and H.-J. Huang. The Multi-class, Multi-criteria Traffic Network Equilibrium and Systems Optimum Problem. *Transportation Research Part B: Methodological*, 38(1):1 – 15, 2004.

[61] H. Yang and X. Zhang. Multiclass Network Toll Design Problem with Social and Spatial Equity Constraints. *Journal of Transportation Engineering*, 128(5):420–428, 2002.

[62] Y. Yin and H. Yang. Optimal Tolls with a Multiclass, Bicriterion Traffic Network Equilibrium. *Transportation Network Modeling 2004*, 1882(1):45–52, 2004.

[63] M. Yoon, T. Li, S. Chen, and J.-K. Peir. Fit a Spread Estimator in Small Memory. In *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*, pages 504–512, Apr. 2009.

[64] H. Yu, X. Jiang, and J. Vaidya. Privacy Preserving SVM using Nonlinear Kernels on Horizontally Partitioned Data. *Proc. of the 2006 ACM symposium on Applied Computing*, pages 603–610, 2006.

[65] M. Zangui, H. Z. Aashtiani, S. Lawphongpanich, and Y. Yin. Path Tolls and the Price of Anonymity. *Submited to Transportation Research Part B*.

[66] N. Zhang, S. Wang, and W. Zhao. A New Scheme on Privacy Preserving Association Rule Mining. *Proc. of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 484–495, 2004.

[67] F. Zhao and N. Park. Using geographically weighted regression models to estimate annual average daily traffic. *Journal of the Transportation Research Board*, 1879/2004:99–107, 2004.