PrivateHunt: Multi-Source Data-Driven Dispatching in For-Hire Vehicle Systems

XIAOYANG XIE, Dept. of Computer Science, Rutgers University, USA
FAN ZHANG, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
DESHENG ZHANG, Dept. of Computer Science, Rutgers University, USA

Recently, for-hire vehicle services (FHV, e.g., Uber and Lyft) have become essential to people's daily transportation. Similar to taxis, how to effectively dispatch these FHV based on demand and supply is important for both FHV passengers and drivers. Based on real-world multi-source data, we identify two new challenges for FHV dispatching: (i) diverse demand: FHV passengers are a mix of passengers previously using taxis, buses, subways, or private vehicles; (ii) uncertain supply: FHV drivers join and leave the FHV system with spatiotemporal dynamics. As a result, the state-of-the-art taxi dispatching techniques cannot be applied to FHV systems. In this paper, we design the first FHV dispatching system PrivateHunt based on extremely large-scale urban transportation data from New York City and Shenzhen in China. In particular, we present (i) a passenger demand model based on taxi, bus, subway, and private vehicle data; (ii) a driver supply model based on small-scale FHV data; (iii) a dispatching technique for FHV vehicles based on proposed demand/supply models to reduce idle driving time. We implement PrivateHunt based on 14 thousand taxis, 13 thousand buses, and 8-line subway system and 10 thousand private vehicles. The experimental results show that our data-driven dispatching strategy significantly outperforms the state-of-the-art dispatching strategies without data-driven FHV insights.

CCS Concepts: • **Networks** → **Sensor networks**; • **Information systems** → *Location based services*;

Additional Key Words and Phrases: Design, Transportation, Urban Computing, Mobility Models, Multi-modal Interaction

ACM Reference Format:
Xiaoyang Xie, Fan Zhang, and Desheng Zhang. 2018. PrivateHunt: Multi-Source Data-Driven Dispatching in For-Hire Vehicle Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 45 (March 2018), 26 pages. https://doi.org/10.1145/3191777

1 INTRODUCTION

In recent years, novel urban services based on sharing economy have been experiencing a significant growth, most notably in For-Hire Vehicle Service (FHV), e.g., Uber [41], Lyft [40], and DiDi [39]. By the end of 2015[42][24], Uber has 8 million users in 300 cities in 60 countries and more than 400,000 drivers in U.S.; Lyft has 2.8 million users, 315 thousand drivers in 60 large cities in U.S., and DiDi, i.e., Uber's local competitor in China, has more than one million drivers in 360 Chinese cities. The FVH becomes increasingly essential to people's daily transportation by exploring traditionally under-utilized private vehicles to complement urban public transportation. In particular, compared to all-day running expensive taxi services, FHV has flexible operating time for drivers and

Authors' addresses: Xiaoyang Xie, Dept. of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, Piscataway, NJ, 08854, USA, xx88@scarletmail.rutgers.edu; Fan Zhang, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guang Dong, 518055, China, zhangfan@siat.ac.cn; Desheng Zhang, Dept. of Computer Science, Rutgers University, 110 Frelinghuysen Road, Piscataway, Piscataway, NJ, 08854, USA, desheng.zhang@cs.rutgers.edu.

2474-9567/2018/3-ART45 $15.00
https://doi.org/10.1145/3191777

less expensive fares for passengers; compared to bus and subway services, FHV provides door-to-door services with an on-demand nature based on smartphone apps.

Similar to other transportation services, e.g., taxi, how to effectively dispatch idle FHVs for potential passengers is very important to satisfy passenger demand in time by bringing FHVs closer to potential passengers and increase driver income by reducing idle driving time. Dispatching FHVs seems straightforward and similar to taxi dispatching [22] because based on FHV locations, we can intuitively suggest FHV drivers to some areas with high historical demand. But we identify two new challenges for FHV dispatching (details are in Section 2) based on real-world large-scale data. (i) Diverse Demand: different from passenger demand for taxi, bus, and subway, FHV passengers are a mix of passengers previously using taxi, bus, subway or private vehicle due to low costs and flexible service coverage of FHVs; (ii) Uncertain Supply: different from full-time taxi drivers without service preference, FHV drivers often work in part time and join or leave the system at any time from any place and often they prefer some services near their home or work [32].

However, the state-of-the-art work on transportation demand/supply modeling and resultant dispatching strategies have been focused on traditional transportation systems (e.g., taxis [3], buses [4], subways [21] and bikes [46]), instead of FHV services. Those dispatching strategies are inappropriate for FHV. On one hand, each traditional transportation system has a single source of passengers and they are used for specified usage. For example, buses and subways are more used for commutes and taxis are more used for errands and recreation [6]. Contrary, the source of passengers for FHV comes from multiple transportation systems. Thus, the usage of FHV is more diverse. In addition, dispatch systems for traditional transportation systems barely consider the supply side since they usually have fixed working schedule. However, transportation network companies (e.g., Uber and Lyft) allow FHV drivers work as a part-time driver with the dynamic working schedule. On the other hand, even though both taxi and FHV provide door-to-door services, the mechanisms of taxis and FHVs are different. A taxi driver chooses a passenger when he/she requests stop within a short range while an FHV driver accepts a request from the network within long range. This difference indicates two aspects: (i) the scale of demand for FHVs can be larger, (ii) a taxi driver picks up a passenger within a short driving distance while an FHV driver may pick up a passenger within a long driving distance. Moreover, a taxicab driver must shift in a fixed time and fixed location making the supply model simpler while an FHV driver is free to choose his/her own schedule. Thus, the dispatching strategies for traditional transportation systems cannot be directly applied to FHV services because of new challenges of diverse demand and uncertain supply for FHV. To the best of our knowledge, little work, if any, has been conducted to investigate demand/supply models and dispatching strategies for FHV based on real-world heterogeneous data. This is because the detailed FHV data and related urban transportation data (e.g., taxi, bus, subway and private vehicles) are often inaccessible due to real-world competition to model diverse demand and uncertain supply of FHV.

In this paper, by collaborating with a few transportation companies and agencies, we conduct the first comprehensive study on demand/supply modeling and resultant dispatching for FHV services. Through the study of the origin-destination patterns of users from multiple transportation systems, we build a demand model to address the diverse demand issue of FHV. Through the study of the daily mobility pattern of FHVs, we build a supply model to address the uncertain supply issue of FHV. In particular, we motivate, design, and evaluate a service called PrivateHunt to dispatch FHV to find their potential passengers based on demand/supply models tailored to FHV and driven by multi-source data. Specifically, the key contributions of this paper are as follows:

- We conduct the first systematic study on For-Hire Vehicle services based on multi-source urban transportation data. To our knowledge, our study on FHV has two key unique features for urban mobility modeling and resultant services: (i) it is based on the most comprehensive long-term multi-modal transportation data,

i.e., including data from the taxicab, bus, subway, private vehicles and FHV, for the same city, covering major transportation modalities; (ii) it is targeted at more than 8 million urban passengers and 37 thousand vehicles in an aggregated fashion. More importantly, we will share such valuable data for the benefit of the research community.

- We identify the key demand/supply features of FHV systems based on real-world data. In particular, from the demand side, we quantify impacts of FHV demand on traditional transportation demand (e.g., taxis) with 3-year data to compare ridesharing during the time periods without and with FHV services; from the supply side, we investigate the low bound of FHV supply and validate its uncertainty due to its flexible operating model. These two features provide essential data-driven insights for our demand/supply modeling.

- We design a data-driven framework PrivateHunt for FHV dispatching with three key components: (i) a predictive demand model to capture potential FHV passenger demand based on FHV and major urban transportation data including taxis, buses, subways, and private vehicles; (ii) a predictive supply model to infer FHV supply based on FHV and traffic data, which infers where and when FHV drivers join and leave the system by analyzing their daily routines with a Bayesian model; (iii) a dispatching strategy built upon both the above demand and supply models to balance the relationship between potential passenger demand and FHV supply. The dispatching is based on modeling predictive control techniques with a comprehensive objective function to minimize percentages of idle cruising mileage, i.e., the percentages of driving distances without passengers. Our dispatching strategy also has the potential to reduce passenger waiting time since it brings more FHVs to regions with more potential passengers.

- We implement PrivateHunt along with its demand/supply and dispatching models in the Chinese city Shenzhen and New York City based on comprehensive data from 8-million smartcards, 14-thousand taxis, 13-thousand buses, 10-thousand private vehicles, and 2-thousand FHVs.

- We evaluate PrivateHunt by comparing it to the state-of-the-art dispatching strategies without FHV data-driven insights. The experimental results show that our data-driven dispatching strategy in PrivateHunt significantly outperforms other strategies.

The rest of the paper is organized as follows. Section 2 shows the motivation we found from the fluctuation between demand and supply of taxicabs and FHVs in the city of Shenzhen and New York. Section 3 presents PrivateHunt overview Section 4 shows urban infrastructures for building PrivateHunt. Section 5 provides our demand and supply models. Section 6 gives our data-driven dispatching strategy. Section 7 provides evaluation results. Section 8 reviews related works. Finally, Section 9 concludes the paper.

2 MOTIVATION

We investigate demand and supply for FHV services based on real-world data from Shenzhen and New York City.

2.1 FHV Demand

Based on three years of Shenzhen taxi data, we study the impact of FHV on the demand of taxis to obtain the potential demand decrease for taxis and demand increase for FHV. Figure 1 gives taxi demand, i.e., the total number of pickups, for 3 years from March 2014 to February 2017. FHV first entered the Shenzhen transportation market around March 2015. Based on the data we found that compared with the demand of the first year, i.e., from March 2014 to February 2015, the taxi demand reduced 24% for the second year, i.e., from March 2015 to February 2016. This decrease of the taxi demand continues in the third year, i.e., March 2016 to Feb 2017. Based

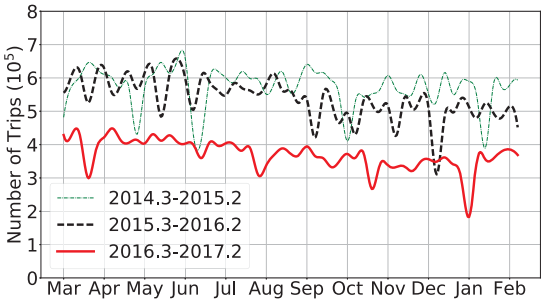

Fig. 1. Passenger Demand of Taxi in Shenzhen

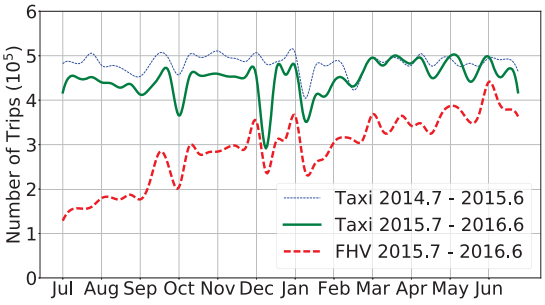

Fig. 2. Passenger Demand of Taxi and FHV in NYC

on the comparison between the same time period from three different years, we found the taxi demand reduced 21% from March 2015, right after the FHV was introduced. It suggested the decreased taxi demand may be taken by FHVs due to their low fares.

Based on a two-year dataset of taxis and a one-year dataset of FHVs in New York City (NYC), we show the increase of FHV demand and the decrease of taxi demand in Figure 2. We found that the FHV demand has been increasing significantly from July 2015 to June 2016 . In contrast, the taxi demand has been decreasing significantly from July 2015 to June 2016 compared with the same time period last year, i.e., from July 2014 to June 2015. These data also suggested a correlation between the increase of FHV demand and the decrease of taxi demand in NYC. Some studies provide evidence to support this claim. For example, through using the NYC open data about taxicab and information from Google Trends, [37] found that controlling for underlying trends and weather conditions that might affect taxi service, Uber's increasing popularity is associated with a decline in consumer complaints per trip about taxis in New York and provided a figure to show the decline in the traditional taxi industry since Uber's entry.

Beyond taxicabs, there is also evidence that FHVs appeals to passengers who typically use the public transportation systems. For example, a report [6] from American Public Transportation Association (APTA) argues that person who uses FHVs is associated with less car ownership and more use of public transportation systems. Some studies [31][30] support this argument by showing that FHV is the competitor and also the complement of public transportation systems. In addition, [6] and [30] also suggest that FHVs replace some private automobile usage. [2] finds that the entry of FHVs is associated with a decline in bus service and shows that some passengers from subway also use FHVs for the complementary trips. Based on our studies and some studies in the social economy and transportation, it is reasonable to expect that passengers from FHVs have the similar demand to the passengers from other transit systems. Hence, we assume that the passengers of FHVs are the mix of the previous passengers from other transportation systems.

2.2 FHV Supply

In this section, we study the feature of FHV supply compared to traditional taxis. Figure 3 gives the lower bound of FHV and taxi supply in Shenzhen, i.e., the number of FHV and taxi available for pickup, on 24 hours of a day and their variance. We found that the supply variance of FHV is very high during the morning and evening rush hour because of their dynamic natures. We also found that during the evening rush hour, the taxi supply is low because of shift changes; whereas the FHV does not have such an issue, which indicates the FHV supply and taxi supply have different features.

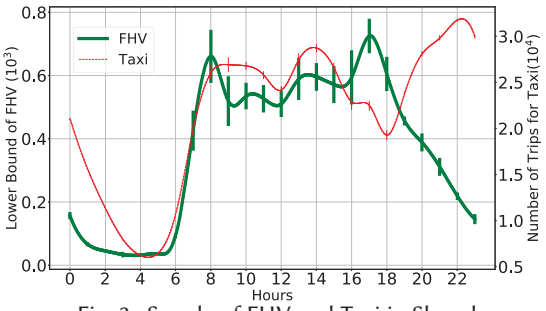

Fig. 3. Supply of FHV and Taxi in Shenzhen

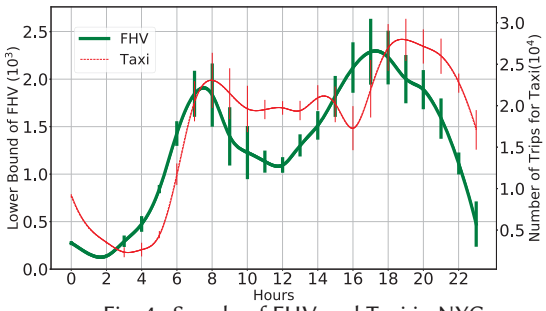

Fig. 4. Supply of FHV and Taxi in NYC

In Figure 4, we study a small fleet of FHV in NYC based on their detailed GPS traces. We found the FHV supply in NYC also has a different pattern with the taxi supply. Their daily supply variance is also very high during the rush hours. From 3 PM to 5 PM, the supply of taxi is also lower than that of FHV due to shift changes.

2.3 Summary

In short, based on long-term data from two cities, we investigate two key features of FHV services, in terms of demand and supply. For the demand side, the passenger demand for the FHV service increases with the decrease of passenger demand in taxis, which suggested taxi passengers may be part of the FHV service. In addition, some studies about the impact of FHVs on transportation systems also reveal passengers from public transit such as bus and subway, and private transit, may also be the source of passengers for FHV service. For the supply side, the FHV supply has a high variance than the taxi supply on 24 hours a day, which suggested that FHV drivers have different working patterns from taxi drivers. These two new features related to demand and supply indicate demand/supply-driven dispatching systems for traditional taxi systems may not work well for FHV services. It motivates us to design FHV demand/supply models and resultant dispatching techniques to improve the efficiency of FHV services.

3 PRIVATEHUNT OVERVIEW

In this section, we first present the architecture of PrivateHunt, and then concisely introduce the three layers in our system.

In PrivateHunt, we utilize a set of urban-scale systems for passenger demand and transit supply modeling, along with dispatching design for FHV services. From a broad perspective, we treat individual components, i.e., smartcards and vehicles, in these systems as probing sensors in PrivateHunt to sense passenger demand and transit supply at urban scale in real time. Built upon an integration of multiple urban-scale systems, i.e., an urban transportation system and a private vehicle system, as an FHV system, PrivateHunt provides unseen demand/supply dynamics under extremely fine spatiotemporal resolutions to support real-world FHV services, which cannot be achieved by either of the systems alone in isolation.

In Figure. 5, we outline PrivateHunt with three layers, *i.e.*, urban physical systems, demand/supply modeling, and FHV vehicle dispatching, providing a road map for the rest of the paper. (i) In Section 4, we introduce four urban-scale transportation networks as concrete frontend systems along with their data. (ii) In Section 5.1, we explore multi-source data from urban public transportation systems to model potential passenger demand for FHV services based on entropy maximizing model; in Section 5.2, we explore GPS and status data from FHV system. (iii) In Section 6, based on these demand and supply models, we provide our dispatching strategy for private vehicle drivers to find their passengers with minimized idle driving distance in FHV services. Finally, our

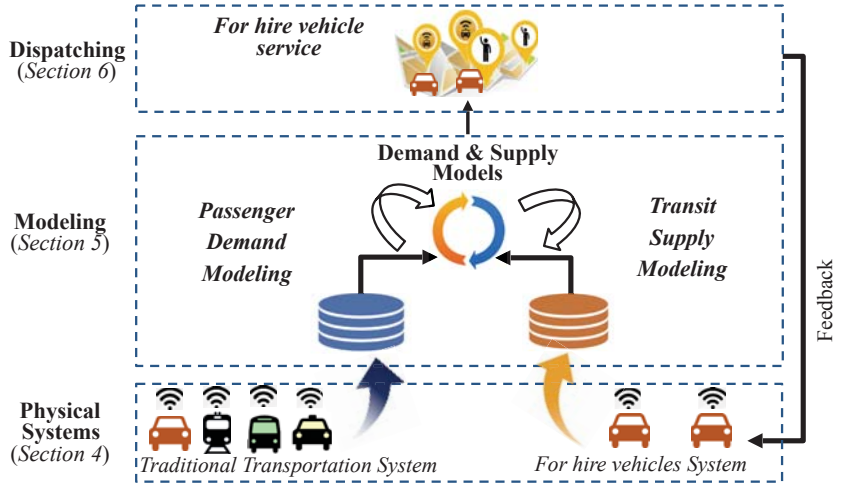

Fig. 5. PrivateHunt Architecture

dispatching strategy in FHV services provides positive feedback to the frontend private vehicle hiring system, which closes our control loop.

4 DATA COLLECTION

Based on the previous collaboration [50] [49], we have been collaborating with several companies and Shenzhen Transport Committee (hereafter STC) for both public transportation networks and private vehicle networks and accessing their data feeds for our research. In detail, the data are collected by an insurance company and a navigation company, and all participants have signed the consent agreement for the reduced monthly premium. In this version of PrivateHunt implementation, we consider two kinds of transportation networks including four systems as follows to model dynamic passenger demand and transit supply for FHV services from complementary perspectives as in Figure 5.

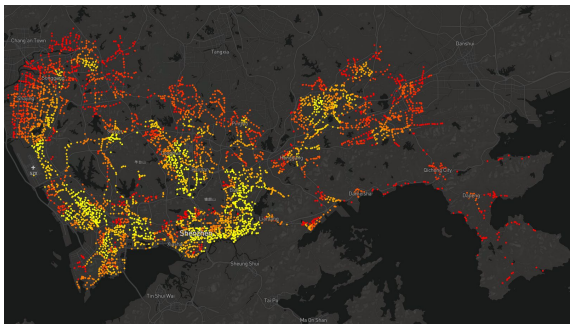

Fig. 6. Bus Potential Passenger Density

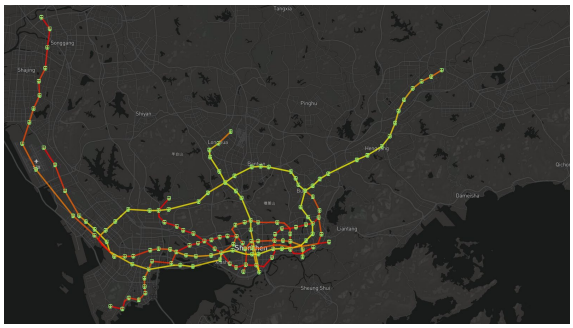

Fig. 7. Subway Potential Passenger Density

- **Buses** are used to detect real-time bus passengers' locations by cross-referencing data of onboard smartcard readers for fare payments. We study bus data through STC to which bus companies upload their bus status in real time, accounting for all 6 thousand buses generating 2 GPS records/min. In detail, we study two months of bus data, containing March of 2014 and March of 2017.

- **Smartcards for Subway and Buses** are used to detect locations of a total of 8 million smartcards used to pay bus and subway fares. These readers capture more than 10 million rides and 6 million passengers per day. We study reader data from STC, which accesses real-time data feeds of a company that operates the smartcard business. In particular, there are two kinds of the readers: (i) a total of 14,270 onboard mobile reader sensors in all buses capturing 138 thousand bus passengers per hour, and (ii) a total of 2,570 fixed reader sensors in 127 subway stations capturing 60 thousand subway passengers per hour. As same as bus data, we study two months of smartcards data, containing March of 2014 and March of 2017.

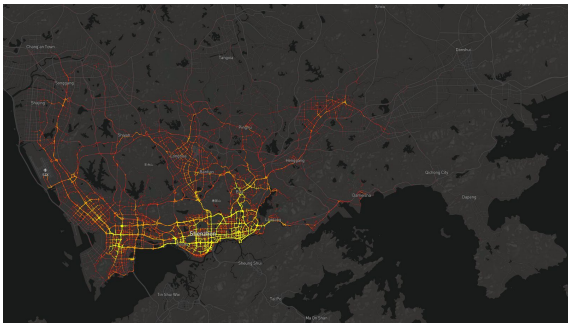

Fig. 8. Taxi Potential Passenger Density

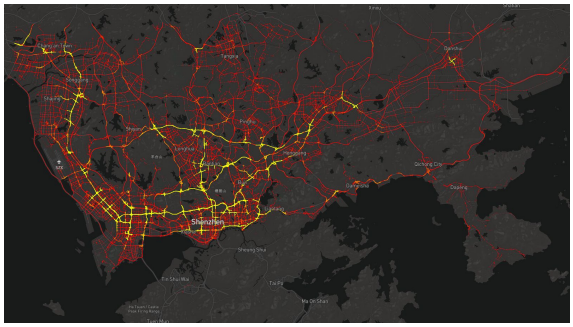

Fig. 9. Private Potential Passenger Density

- **Taxicabs** are used to detect real-time taxicab passengers' locations based on taxicab status (*i.e.*, GPS and occupancy) to obtain taxi passenger demand. We study taxicab data through STC to which taxicab companies upload their taxicab status in real time, accounting for all 14 thousand taxicabs generating 2 GPS records/min with a 400 thousand ridership per day. The daily size of taxicab sensor data is 2 GB. Besides, we also study the monthly passenger data, including the total number of passengers and total incomes. In detail, we study two months of taxicabs data, containing March 2014 and March 2017. In addition, we study the monthly fare data of taxicabs started from January 2014 to February 2017.

- **Private Vehicles** are used in both the demand and supply model. On one hand, private vehicles are used to detect potential passengers for the demand model. On the other hand, we study the mobility pattern of taxicabs and private vehicles to discover which private vehicles are the potential FHVs and then based on this we train a supply model for dispatching. For the private vehicle network, it has 294,845 vehicles, among which 10,450 are in Shenzhen. Their data are collected through onboard devices installed inside vehicles, which are mainly used for navigation purposes. In this project, we access these data through a navigation service provider to which every involved vehicle uploads their real-time status to a cloud server by a cellular network. One day data collected from all private vehicles in the network are about 9 GB with an average uploading interval of 10 seconds when devices are turned on. The private vehicle users can choose to opt out this optional data uploading service, but most users still upload their data in order to reduce their monthly premium. In particular, we study the data of private vehicles in Shenzhen from January 2016 to February 2016 and from April 2017 to May 2017.

Based on the above physical systems, we model passenger demand and transit supply with a real-time large-scale data-driven fashion. A visualization for passenger density in Shenzhen downtown areas based on three public transit systems is given in Figures 6 7, and 8 and a private vehicle network Figure 9 where subway and bus

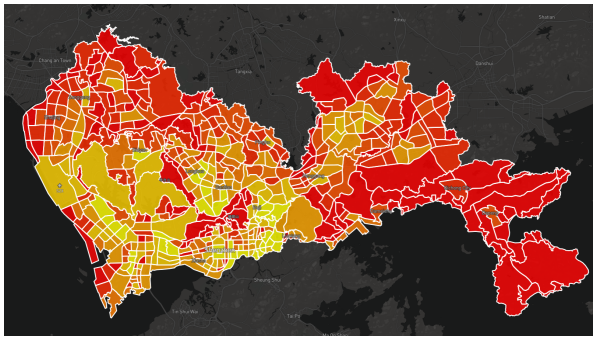

Fig. 10. Shenzhen Urban Partition

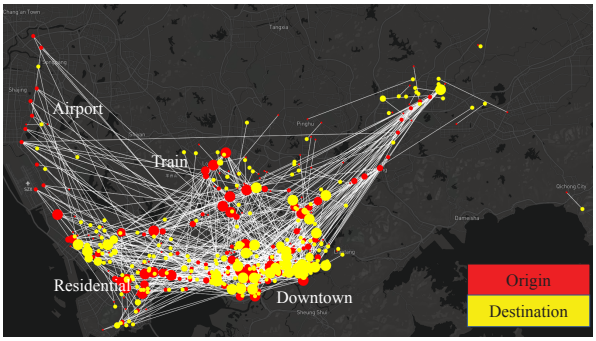

Fig. 11. Passenger Mobility Demand

passengers' densities are given at station levels and taxi passenger and private vehicle densities are given at GPS levels. The lighter the icon; the higher the demand. Based on the above systems, we also visualize pattern mobility demand. We divide entire Shenzhen urban area into 491 urban regions based on Shenzhen Government Census Blocks as shown in Figure 10, and then we visualize passenger mobility patterns at region levels in Figure 11 where we found that there are strong mobility demands between downtown regions, the airport region, and the residential region.

Table 1. Transportation DataSets

Bus GPS Dataset		Bus Fare Dataset		Taxi Dataset	
# of records	192,722,381	# of records	23,226,357	# of records	290,802,066
# of bus	6,234	# of passenger	4,017,901	# of taxicab	14,863
Size	27.58 G	Size	2.19 G	Size	18.83 G
Format		Format		Format	
plate id	date&time	smartcard id	date&time	plate id	date&time
GPS	line	line	plate id	GPS	status

Subway Dataset		Private Vehicle Dataset	
# of records	9,174,402	# of records	454,018
# of passengers	2,508,770	# of car	10,443
Size	2.23 G	Size	8.67 G
Format		Format	
plate id	date&time	plate id	date&time
check in/out	location	GPS	speed

Since our paper concentrates on modeling and dispatching aspects, we briefly introduce our data management related issues due to space limitation. We establish a secure and reliable transmission mechanism, which feeds our server the above data collected by service providers with a wired connection. As in Table 1, we have been storing a large amount of data for demand/supply modeling as well as dispatching strategy design. We utilize a 34 TB Hadoop Distributed File System (HDFS) on a cluster consisting of 11 nodes, each of which is equipped with 48 cores and 512 GB RAM. For daily management and processing, we use the MapReduce-based Spark. Due to the extremely large size of our data, we have been finding several kinds of errant data, *e.g.*, duplicated data, missing data and data with logical errors, and thus we have been conducting a detailed cleaning process to filter out errant data on a daily basis. Besides using HDFS and Spark for data cleaning, we use those tools for some preprocessing. For example, (i) We match bus GPS data with bus fare data to obtain the information of origin and destination

of passengers of the bus. In general, the bus passengers do not need to swipe their cards when they dropped off from bus. Therefore, to obtain a complete trip record, we predict the destination from the later record since the origin station of a passenger on the later hours in one day, has a high probability of being the destination station of his/her last trip record on the earlier time based on the previous survey [48]. The probability also is relevant with the distance between two stations. (ii) We analyze GPS data of taxicabs to recognize their pickup and drop off locations. (iii) Based on GPS records only, it is hard to define trips of private vehicles among the entire trajectories in one day. Therefore, we process the data of private vehicles as [12] to infer the complete trip data of private vehicles. (iv) We obtain the origin and destination information of passengers of the subway by analyzing every two continuous records with the same id. The data preprocessing transfers raw records of each trip with extra information to obtain the origin information and destination information, which will not harm the accuracy of our model. We protect the privacy of residents by anonymizing all data and presenting models in aggregation. In short, our endeavor of consolidating the above data enables extremely large-scale fine-grained passenger demand/supply modeling along with dispatching strategy for FHV services, which is unprecedented in terms of both quantity and quality shown as follows.

5 DEMAND/SUPPLY MODELING

In this section, we first present our demand modeling in section 4.1 and supply modeling in section 4.2 for FHV services, which are used for efficient dispatching strategies to maximize mileage with passengers for all vehicles for FHV services.

5.1 Passenger Demand Modeling

Since FHV passengers previously come from taxicabs, buses, subways and private vehicles, in this paper we infer that FHV services have a reducing impact on the number of passengers in these traditional transportation systems. Therefore, we integrate the data of taxicabs, buses, subway and private vehicles within the date before and after the FHV service started in a city, study the difference of the demands inferred for these two periods, and then use this difference to infer demand for FHV as follows.

5.1.1 Demand Transform Estimation. Given a spatial partition of a city, e.g., the spatial partition in Figure 10, the passenger demand of a transportation system is defined as a K dimension vector $\bar{X}$, where K is the total number of partitioned regions in the city. Therefore, for the transportation system of taxicabs, buses, subways, and private vehicles in a city, the transformation between those transportation systems and FHV is presented as

$$\begin{array}{c} \\ r_1 \\ r_2 \\ \vdots \\ r_k \end{array} \begin{array}{c} \begin{array}{cccc} X_t & X_b & X_s & X_p \end{array} \\ \begin{pmatrix} X_{t1} & X_{b1} & X_{s1} & X_{p1} \\ X_{t2} & X_{b2} & X_{s2} & X_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{tk} & X_{bk} & X_{sk} & X_{pk} \end{pmatrix} \end{array} \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \pi \end{pmatrix} = \begin{array}{c} \\ r_1 \\ r_2 \\ \vdots \\ r_k \end{array} \begin{array}{c} X_f \\ \begin{pmatrix} X_{f1} \\ X_{f2} \\ \vdots \\ X_{fk} \end{pmatrix} \end{array}$$

Where X_t, X_b, X_s, X_p, X_f denote the demand vector of taxicabs, buses, subways, private vehicles and FHV respectively and r_k denotes the region k. In the equation, we use 4 parameters α, β, γ and π to represent the rate of passenger demand transferred from the traditional transportation systems to FHV services. To study these four parameters, we investigate the change of the amount of passengers in the taxicabs, buses, subways and private vehicles one by one, and then set a reasonable range of values for these four parameters.

5.1.2 Demand Model Design. To learn the detail of passenger demand in a given spatiotemporal partition of a city, we study the mobility pattern of passengers and present their mobility pattern as a set of origin-destination

matrices (OD matrices). The passenger demand of a spatiotemporal combination is defined as the number of passengers leaving from this region going to another region during this time slot. We represent passenger demand D by OD matrices at urban region levels in which there are K rows and columns, respectively. An element $D(i, j) = n$ represents n passengers moving from i_{th} region to j_{th} region. Based on the observed data from four urban transportation systems, e.g., taxis, buses, subways and private vehicles, it is straightforward to obtain the historical passenger demand by completing an OD matrix for a particular time window.

As follows, we obtain the future passenger demand model **D** based on the difference between two observed historical passenger demands in different datasets. Given the operating nature of FHV services and sizes of urban regions, we use a time interval of 5 minutes to generate OD matrices. In this case, the passenger demand at time t minutes is the total number of passengers moving out in the period $[5i, 5i + 5)$ where $t \in [5i, 5i + 5)$ and $i \in N$, e.g., $[0, 5), [5, 10), [10, 15), ..., [1435, 1440)$. In total, we have 288 OD matrices for different periods of one day. Based on these regions, we design and implement our model **D** based on entropy maximizing model [36] to predict passenger demand from multi-source data. Although entropy maximizing model has been used before in the context of transportation modeling [35], it has not been used in the context of multi-source passenger demand modeling. Therefore, we use the parameter vector $\bar{w} = [\alpha, \beta, \gamma, \pi]^T$ to vary the number of passengers transferred from other transportation systems to FHV services as the input of our model.

In our model, we use centroids of the urban regions to represent these urban regions. The distance d_{ij} between two regions i and j is defined by the traveling distance between the two centroids. Taking the real road network information and vehicle types into consideration, we utilize our supply model proposed in section 4.2 to obtain traveling time between two regions. We denote the probability of a trip starting from the ith region R_i as g_i; the probability of a trip ending in jth region R_j as a_j; the probability of a trip from R_i to R_j as t_{ij}. If the total number of trips starting from R_i is G_i, the total number of trips ending in R_j is A_j, the total number of trips from R_i to R_j is X_{ij}, and the total count of trips is X, then the probability is calculated as follows. $g_i = G_i/X, \quad a_j = A_j/X, \quad t_{ij} = X_{ij}/G_i$ Where X_{ij}, G_i and A_j are defined as $G_i = \sum_{j=0}^{n} X_{ij}, \quad A_j = \sum_{i=0}^{n} X_{ij}$ We calculate prior probability q_{ij} as $q_{ij} = \alpha g_i a_j d_{ij}^{-\gamma}$, where d_ij is the travel distance, α and γ are fitting parameters. According to the entropy maximizing model, the objective function is defined as

$$\max L = -\sum_{i}^{n}\sum_{j}^{n} g_i t_{ij} \ln(t_{ij}) - \gamma \sum_{i}^{n}\sum_{j}^{n} g_i t_{ij} \ln(d_{ij}) \qquad S.T. \begin{cases} \sum_{j}^{n} t_{ij} = 1, \\ \sum_{i}^{n} g_i t_{ij} = a_j, \end{cases} \tag{1}$$

By solving this problem with a numerical method, t_{ij} is obtained as

$$\begin{cases} t_{ij} = e^{-1} k_i m_j d_{ij}^{-\gamma} \\ \sum_{j}^{n} t_{ij} = 1 \\ \sum_{i}^{n} g_i t_{ij} = a_j \end{cases} \tag{2}$$

where

$$\begin{cases} k_i = exp(\mu_i / g_i) \\ m_j = exp(\lambda_j) \end{cases} \tag{3}$$

μ_i and λ_j are the Lagrange coefficients related to the subject conditions. Algorithm 1 shows the calculation

algorithm.

With the entropy maximizing model, we obtain predicted 288 OD matrices for 288 5-mins time windows. To feed the demand vector of FHV into these 288 matrices, we normalize each row in these matrices individually and obtain the distribution of destinations for each region. Then we apply the demand vector of FHV into this model to get the OD matrices of FHV.

We implement our demand model based on our multi-source transportation data including taxi, bus, and subway in Shenzhen. In detail, we quantify transformation of passengers after FHV entered the market in Shenzhen for α, β, γ and π by leveraging (i) the data in 2014 March and 2017 March of the three public transportation systems, namely taxi, bus, subway, (ii) the transformation of private vehicle drivers to passengers of FHV by leveraging the data in 2016 January 2016 February 2017 April and 2017 May, (iii) the proportions [34] of passengers of each transportation systems in total number of passengers in Shenzhen. We compute the distribution of the travel demand in 491 regions of Shenzhen for each transportation systems and then feed the transferred pattern of passengers into our demand model to obtain the OD demand matrix for FHV.

ALGORITHM 1: Entropy Maximizing Model Calculation

Input: a_j, g_i, d_{ij}
Output: X_{ij}, t_{ij}
Algorithm `main`(a_j, g_i, d_{ij})
  estimate γ by q_{ij} as $q_{ij} = \alpha g_i a_j d_{ij}^{-\gamma}$;
  initial value of μ_i, λ_i;
  repeat
    $\mu_{i+1} = -g_i \ln[\sum_j^n e^{(-1+\lambda_j) d_{ij}^{-\gamma}}]$;
    i = i + 1 ;
    $\lambda_{j+1} = \ln a_j - \ln[\sum_i^n g_i e^{(-1+\mu_i/g_i) d_{ij}^{-\gamma}}]$;
    j = j + 1;
  until $|\mu_i - \mu_{i-1}| < \epsilon$ *and* $|\lambda_i - \lambda_{i-1}| < \epsilon$;
  calculate $t_i j$ in equations 3 ;
  $X_{ij} = t_{ij} X_i$;
  return;

5.2 Supply Modeling

In this section, we provide a data-driven supply model S to capture the supply of private vehicles for FHV services.

5.2.1 FHV detection. We study the mobility pattern of taxicabs and private vehicles based on real entropy proposed in [33] for months and find out the private vehicles whose daily mobility pattern is similar to taxicabs as the FHVs. Fig.12 gives the distribution of the entropy of Private Vehicles and taxicabs. We found there is a small group of private vehicles whose entropy is overlapped with taxicabs. we study the long term historical trips data of those private vehicles and taxicabs to learn the distribution of durations for the gaps between two trips, e.g. the distribution of Fig. 13 We found their mobility is similar to that of taxicabs. Therefore, we assume those private vehicles have a high probability as FHVs.

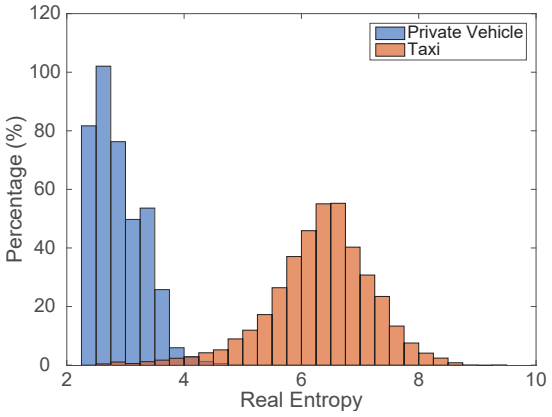

Fig. 12. Entropy of Taxi and Private Vehicle

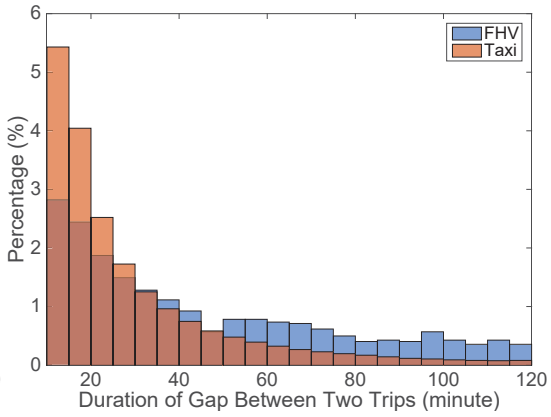

Fig. 13. Duration Gaps Distribution

5.2.2 Supply Model Design. Given a pair of time slots in one day, a working schedule is defined as a time period from a time one driver starts to work to the time the driver stop to work. Different from other transportation supply, e.g., taxis, buses, and subways, the vehicle supply for FHV is highly dynamic because most drivers have the freedom to choose their own working time for FHV companies. Therefore, they typically join and leave the system without noticing the dispatching center in advance. However, it is a challenge to distinguish if a driver is seeking a passenger or stop working in the gap of two trips. To address this challenge, we analyze result on Fig. 13 and found that (i) the gaps of FHVs and taxicabs are almost overlapped on the regions which are less than 50 minutes, (ii) as the minutes increasing, the difference between FHVs and taxicabs becomes larger. This is because of the dynamic natures of FHVs. Since taxicabs and FHV have the similar mobility pattern when they are working, we assume an FHV is possibly inactive during large gaps. Based on this assumption, we use the gap data of taxicab and FHVs to learn a Gaussian function and apply this function to the data of FHV to predict the complete working schedules for FHV drivers. With this Gaussian model, we study the complete working schedules of FHV drivers for a long-term historical data and design a generic Bayesian model to infer the start time and end time of each working schedules for FHV drivers given an initial region. For an FHV driver, given his/her first working region, we predict the start working time, the end working region and time individually.

5.2.3 Supply Tensors. We model the supply of FHV service in a spatiotemporal combination using two Tensors $\mathcal{O} \in R^{K \times N \times M}$ and $\mathcal{D} \in R^{K \times N \times M}$, with three dimensions denoting K regions, N time slots and M FHV drivers respectively for the start and end of the working schedules.

- Region dimension: This first dimension denotes regions $r = [r_1, r_2, \ldots, r_K]$ defined in section 5.1.
- Time slot dimension: We divide a day into equal slots $t = [t_1, t_2, \ldots, t_N]$ defined in section 5.1.
- FHV driver dimension: The dimension denotes FHV drivers $d = [d_1, d_2, \ldots, d_M]$ of the FHV service in a city.
- An entry: An entry $\mathcal{O}(i, j, k)$ or $\mathcal{D}(i, j, k)$ stores a working schedule of a driver k start or stop working on region i at time slot j.

5.3 Summary

With demand model, we learn the probabilities of passenger requests in every region and every time slots and distribution of destinations of those requests. With supply model, we obtain the numbers of FHVs in every region and every time slots. When dispatching an FHV in a given time slot, PrivateHunt considers the distances between the FHV and regions, the probabilities of passenger requests, and the numbers of FHVs in regions spatiotemporally. In the next section, we will introduce the detail of the dispatching strategy of PrivateHunt.

6 FHV DISPATCHING

PrivateHunt contains two parts, a data-driven dispatching strategy and an assignment algorithm when it dispatches an FHV. The data-driven dispatching strategy integrates demand model and supply model and makes a cruising policy for the FHV to improve its profitability. The assignment algorithm matches FHVs with passenger requests in given regions. In PrivateHunt, the data-driven strategy can be integrated with different assignment algorithms. In this paper, we implement six different assignment algorithms to show the robustness of PrivateHunt. In addition, our overall dispatching goal is to minimize the total idle cruising mileage of all FHVs in the system. In the following, we introduce the scenario of the dispatch problem and the design of our dispatch strategy.

6.1 Dispatch Problem Scenario

Different from taxicabs, FHVs accept a passenger request with smartphone apps. Once a driver accepts a request, he/she cannot accept other requests during the way to pick up his/her accepted passenger, even though the new request is closer to him/her. Therefore, FHVs prefer to stay in the region with dense passenger requests because they have more options of accepting requests.

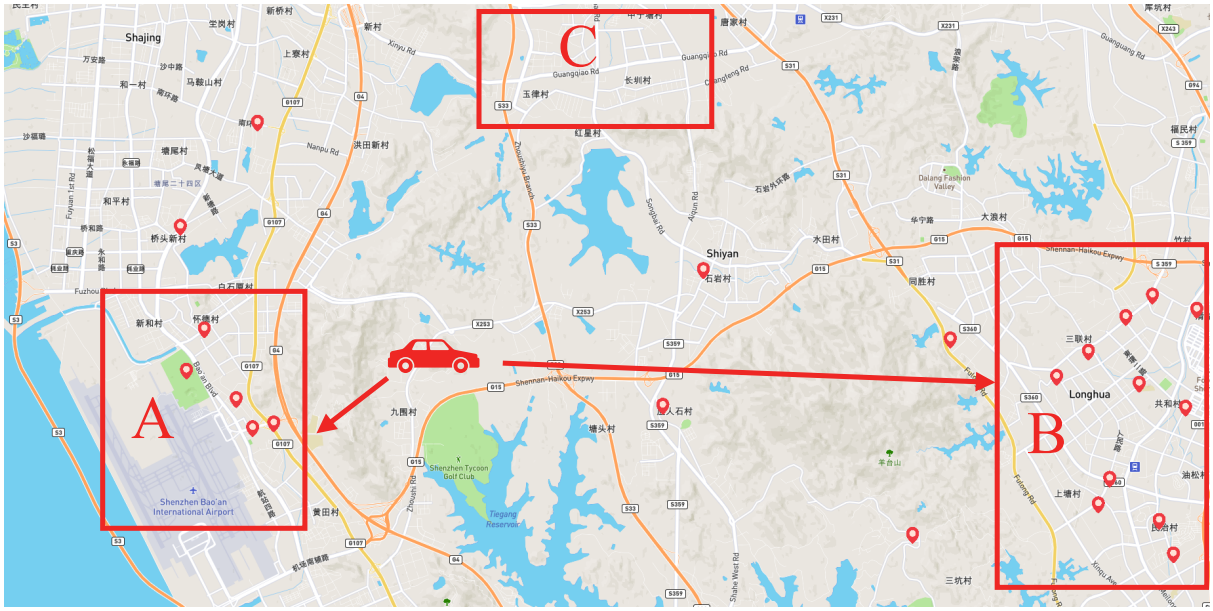

Fig. 14. Dispatch Scenario

Fig. 14 gives a possible scenario of a dispatch problem. This figure is the road map of the northwest part of the Shenzhen City. Red points represent pickup locations, which are sampled based on passenger demand

in Shenzhen. The car represents an FHV and the three boxes represent three main possible directions for the FHV to cruise. In the three regions, Region B has the most points but is very far to the current location of the FHV; Region A has the fewest points but is closer to the FHV; Region C has a very low probability of passenger requests since it has no red point. Without data support, old drivers will cruise based on experience and new drivers will cruise more based on a random strategy. As the map shows, all the requests are within Region A and Region B. The other places are similar to Region C with a very small number of requests. FHVs are easier go to those places with a random strategy. Even though they still can accept the requests from apps, they will spend a lot of gas during the way to pickup their customers. In contrast, the data-driven strategy will recommend the FHV go to Region A or Region B. With closer locations, they are more competitive and have lower gas consumption.

ALGORITHM 2: Policy Iteration Algorithm

Input: S,A,R,λ
Output: π^*
Algorithm `main(`S,A,R,λ`)`

```
t = 0, k = 0;
initial π_t(s_i) = s_i for all s_i ∈ S;
initial V_k(s_i) = 0 for all s_i ∈ S;
repeat
    repeat
        for s_i ∈ S do
            V_{k+1}(s_i) = R(s_i, a_i, π_t(s_i)) + λV_k(π_t(s_i)) ;
        end
        k = k + 1;
    until |V_k(s) − V_{k−1}(s)| < ε, ∀s_i ∈ S or reach limited iteration;
    t = t + 1;
until π_t = π_{t−1} or reach limited iteration;
π* = π_t ;
return;
```

6.2 Dispatch Problem Formulation

We define the dispatch problem as a Markov Decision Process (MDP) and use Bellman equation to solve it. Formally, a MDP is a 5-tuple (S,A,T,R,λ).

- S is a set of states. We define the location of pickup as a state. To make our algorithm simpler, we just use the states who are close to current state and whose percentage of passenger is not zero.
- A is a set of actions. Here, it represents a FHV run from a location to another location.
- T is the transition matrix. Here $T(s_i, a_i, s_j)$ represents the transition probability between s_i and s_j, where s_i and a_i is current state and action of FHV and s_j is the next state. Since FHVs has fully control of their vehicle, we set it $T(s_i, a_i, s_j) = 1$.
- R is a reward function. We define it as

$$R(s_i, a_i) = (1 - \frac{D(s_i, s_j)}{\sum_{s_k}^{S} D(s_i, s_k)})/(N-1) * \frac{p_i}{q_i} \tag{4}$$

where $D(s_i, s_j)$ represents the travel distance between s_i and s_j, p_i is the percentage of passenger requests, derived from the demand model and q_i is the percentage of FHVs serving in the future time slot, derived from the supply model, N is the number of states. This reward function consider the factors including distances, demand and supply.

- $\lambda \in [0, 1]$ is the discount factor, indicates the difference in importance between future rewards and present rewards.

The goal of data-driven strategy is to obtain a policy π, satisfies that

$$V^{\pi}(s) = \max_{a \in A}[R(s, a) + \lambda \sum_{s' \in S} T(s, a, s')V^{*}(s')] \tag{5}$$

Therefore, we use the policy iteration algorithm to solve this problem.
Based on algorithm 2, we build a data-driven dispatching strategy. The strategy contains two steps. Step 1, dispatch FHVs based on the result of Bellman equation. Step 2, after dispatching, in a region, we regard FHVs and passengers as vertexes in a graph and implement assignment algorithm on it.

6.3 Assignment Algorithm

In a given time slot, passengers send requests to FHVs. In addition, once an FHV accept a request, he/she will keep traveling to the assigned passenger. In this time slot, we represent the assignment problem as a directed graph G. In this graph, there is a set U of passenger vertexes u, a set V of FHV vertexes v, and a set E of edges $e = (u, v)$, connects a passenger vertex and an FHV vertex. The weight of e is the current travel distance from a passenger to his/her assigned FHV. The objective of the assignment algorithm in this time slot is to find a subset from E for bipartite matching, whose sum of weights is minimal. The unassigned passengers and FHVs will be reserved. We implement six assignment algorithms into our data-driven framework. We will compare their data-driven version with their normal version. In the following part, we will introduce the basic idea of each algorithm.

- **Passenger Greedy method** The most normal algorithm for assignment problem is the greedy algorithm. Considering from the passenger side, one greedy method allows the passenger to request the nearest unassigned FHV, referred as Passenger Greedy method(Greedy-P). In a given time slot, we choose passengers randomly and find pairs for them. For each passenger, Passenger Greedy method will pair him/her with the nearest unassigned FHV from FHVs in the limited range. We set the range as 3 kilometers since it is the searching range in the FHV APP in China. The potential FHV will be checked if it can accept the request before it is assigned. If an edge $e = (u, v)$ is matched, the passenger u and the FHV u are removed from G. Otherwise, Passenger Greedy method will check the next closest FHV.

- **FHV Greedy method** Another greedy method is considered from the FHV side. We refer it as FHV Greedy method(Greedy-V). It allows FHVs to accept the nearest passenger request. In a given time slot, we choose FHVs in random order and find viable pairs for them. For each FHV, FHV Greedy method will pairs it with the nearest unassigned FHV from FHVs in the limited range. For the potential passenger, FHV Greedy method will check if the predicted arrival time of his/her destination is acceptable for the current FHV. If an edge $e = (u, v)$ is matched, the passenger u and the FHV u are removed from G. Otherwise, FHV Greedy method will check the next closest passenger.

- **Hungarian method** The Hungarian method(HUN) is a combinatorial optimization algorithm that solves the assignment problem in polynomial time. Since Hungarian method requires a perfect matching and in the general dispatch scenario, $|U| \neq |V|$, we add some dummy vertexes, u_d or v_d, into the smaller set. For all edges of $e = (u, v_d)$ or $e = (u_d, v)$, we set the weight of e as 0. For the edges of $e = (u, d)$, if the weight of e is larger than 3 kilometers, we set the weight as infinite. Then we run Hungarian method on the new graph. We refer the assignment problem as a matrix of weights of edges between U and V. The Hungarian method uses 4 steps to solve this problem. In step 1, for each row, Hungarian method finds the lowest element and subtract it from each element in that row. In step 2, for each column, Hungarian method finds the lowest element and subtract it from each element in that column. In step 3, Hungarian method covers all zeros in the resulting matrix using a minimum number of horizontal and vertical lines. The algorithm stops when it needs $max(|U|, |V|)$ lines. Otherwise, Hungarian method continues with Step 4. In step 4, Hungarian method finds the smallest element (call it k) that is not covered by a line in Step 3. Then it subtracts k from all uncovered elements, and add k to all elements that are covered twice. Hungarian method will repeat step 3 and step 4 until all elements can be covered by $max(|U|, |V|)$ lines. In our implementation, we divide a city into a grid and set the length of the side of cells as 3 kilometers. For each cell, we implement Hungarian method on all FHVs and passengers in it.

- **SCRAM method** Hungarian does not consider the waiting time of passengers after a request of the passenger is accepted. Therefore, some passengers may wait for a long time after their requests are accepted. To solve this problem, Josiah P.Hanna et al in [18] implement SCRAM algorithm on the dispatching of taxicabs. Similar to Hungarian method, we add some dummy vertexes into U or V and generate a new graph as same as that of Hungarian method. SCRAM method first finds refers the assignment problem as bottleneck assignment problem to find a minimal maximum edge e_{max} in a perfect matching. In this paper, we also use Hungarian algorithm to find the minimal maximum edge. Then SCRAM method remove all the edge e if the weight of e is larger than e_{max}. Here we remove edge by setting the weight of it to infinite. Finally, run Hungarian method again on the new graph.

- **NSTD-P** Previous algorithms are designed naturally for taxicabs. However, for the FHV transportation system, passengers and FHVs both have their own interest in accepting an assignment. Huanyang Zheng et al in [51] provides new methods for FHV by considering both interests of passengers and FHVs. The first method is named as NSTD-P. This method mainly considers interest from the side of passengers. Each passenger u keeps a preference order of FHVs and each FHV v also keeps a preference order of FHVs. The preference is calculated by the distance between the passenger and the FHV. For a passenger, we add a dummy entry in his/her preference order to represent that all the FHVs in the preference order after this dummy entry are out of the searching range. For an FHV, we add a dummy entry to its preference order with the same meaning. For $\forall u \in U$, NSTD-P will propose requests to FHVs orderly based on the preference of u iteratively until an FHV accepts his/her request or u prefers no match over the rest of FHVs. If an FHV v has no temporal proposed request or the order from u over dummy request, v will accept the request of u. If v has temporal request from other passenger u' and the order of u' is over that of u, v will reject u, otherwise, NSTD-P propose new request for u'.

- **NSTD-T** NSTD-T mainly considers the interest from the FHV side. The key idea is to start with NSTD-P and try to break the assignment to obtain other solution. NSTD-T first run NSTD-P to build an initial assignment. Then for all u_i in U, run break step on it. On break step, NSTD-T first breaks its current proposed partner and proposes the new request for u from next entry in its preference. If new match is successful, call break step on u_j which $j > i$ recursively.

6.4 Summary

The core idea of the strategy is to dispatch FHVs into regions with the higher probability of passenger requests in advance. It is likely to add an additive, e.g demand model, and supply model, to the existed dispatch algorithms. Different with the assignment algorithms proposed previously for taxicabs, PrivateHunt dispatch FHVs into regions with the higher numbers of passenger requests before matching requests to FHVs by assignment algorithms. It matches FHVs to passengers with easy success and reduces the idle distances from FHVs to passengers since FHVs accept requests from the network, which is different with way taxicabs do. Therefore, PrivateHunt improves the performance of existed dispatch algorithms by adding demand/supply models.

7 EVALUATION

7.1 Metrics

In order to evaluate the performance of our strategy, two metrics are implemented in our evaluation. The first one is the waiting time for passengers. This metric evaluates the fairness of the dispatching algorithm. The second one is the idle mileage rate with the total mileage of FHVs.

- **FHV Idle Mileage Rate:** In this paper, the percentage of mileage without passengers (idle mileage rate) is used as a metric in the evaluation. This metric is used in previous studies[25] and to evaluate the performance of the dispatching strategy for FHVs. With a lower idle mileage rate, the FHV spends less money on seeking a passenger. Assume t_i represent a time slot, I_i is the idle mileages of an FHV in time slot t_i, M_i is the total mileage of an FHV in time slot t_i and T is the number of time slots. In our evaluation, an FHV's idle mileage rate is $\frac{\sum_i^T I_i}{\sum_i^T M_i}$
- **Passenger Waiting Time:** Waiting time is the delay from the time that a passenger requests a trip to the time that an FHV is dispatched to this passenger request. The FHV still needs some travel time before picking up the passenger. In our evaluation, we generate a passenger request based on demand model and this request will cancel as the waiting time increasing. The passenger waiting time for a passenger request is the time slots from the time slot it requests to the time slot it cancels.

7.2 Baseline Approach

We regard the two strategies as Baselines,

- Random Strategy: FHVs will go to the near regions randomly before the implemented assignment algorithm calls.
- RHC Strategy from [25]: It dispatches FHVs by considering the distribution of potential passengers estimated from demand model and limit the number of vehicles to the same destination.

We implement the data-driven strategy in PrivateHunt with 6 assignment algorithms and compare the results of waiting time for passengers and idle mileage rate for FHVs with that of Baselines.

7.3 Impact of Assignment Algorithms

Results of the performance of PrivateHunt for datasets in Shenzhen City are shown in the following.

7.3.1 Idle mileage rate. We start with idle mileage rates for FHVs in Figure.15. Figure.15 shows Cumulative Distribution Function (CDF) of the idle mileage of FHVs. The X-axis is the idle mileage rates of FHVs and the

Y-axis is the percentage in CDF. The performance of a method is better if its curve is on the upper side compared with that of another one.

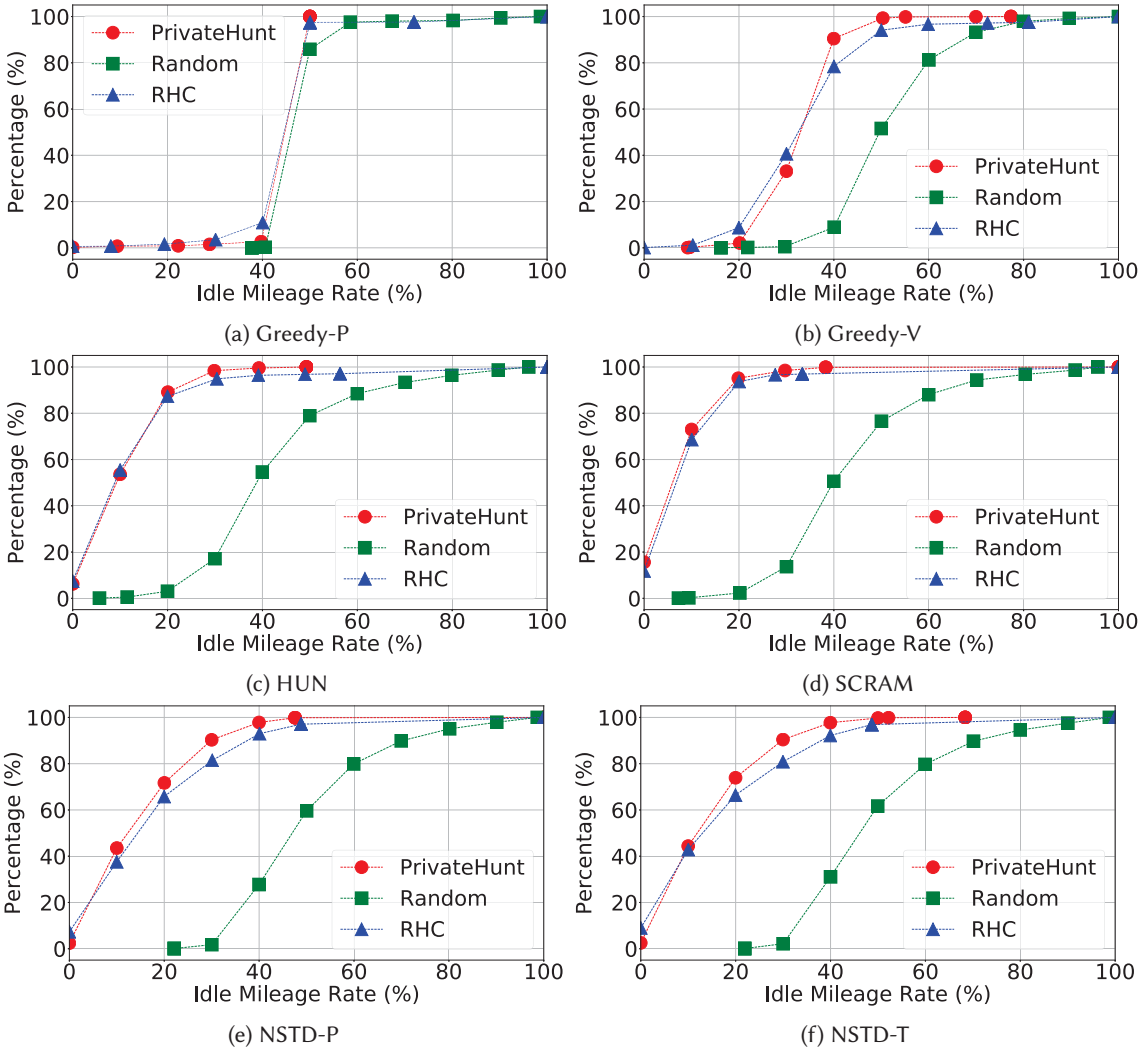

Fig. 15. Idle Mileage Rate distribution

The results of comparisons between PrivateHunt and the two baselines show PrivateHunt has a better performance than the two baselines for all the algorithms. In details, the curves of the Greedy-P algorithm focus on the idle mileage rates from 40%to 60%. PrivateHunt is worse than RHC in the idle mileage rate of 40% but better than RHC after idle mileage rate of 50%. While it is always better than Random. In the curves of Greedy-V, the gaps between PrivateHunt and the two baselines are larger. 95% of FHVs have an idle mileage rate smaller than 40% with PrivateHunt. In contrast, with RHC, the percentage of FHVs whose idle mileage rate is smaller than 40% is about 80%. And that with Random is only about 10%. In the curves of HUN, 85% of FHVs with PrivateHunt have an idle mileage rate smaller than 20%. Compared with this, the random curve has more than 90% of FHVs having an idle mileage rate larger than 20%. RHC has a closed performance in HUN method before the idle mileage of

80% but becomes worse after that region. Similar to HUN, PrivateHunt on SCRAM also performs better than that of the two Baselines. In the result of NSTD-P and NSTD-T, about 80% of FHVs with PrivateHunt have an idle mileage rate smaller than 20%. In these two methods, RHC becomes worse after about idle mileage rate of 10%. While the random curve has more than 95% of FHVs having an idle mileage rate larger than 20%. In short, PrivateHunt has a slightly better performance than RHC and much better than Random in these 6 methods. From the study of the performance of PrivateHunt on these 6 algorithms, we see that it reduces the idle mileage rate enormously, which improves the profit of the drivers of FHVs significantly.

7.3.2 Waiting time. Fig.16(a) to Fig.16(f) show Cumulative Distribution Function of the passenger waiting time. The X-axis is the waiting time of passenger and the Y-axis is the percentage in CDF. The performance of a method is better if its curve is on the upper side compared with that of another one.

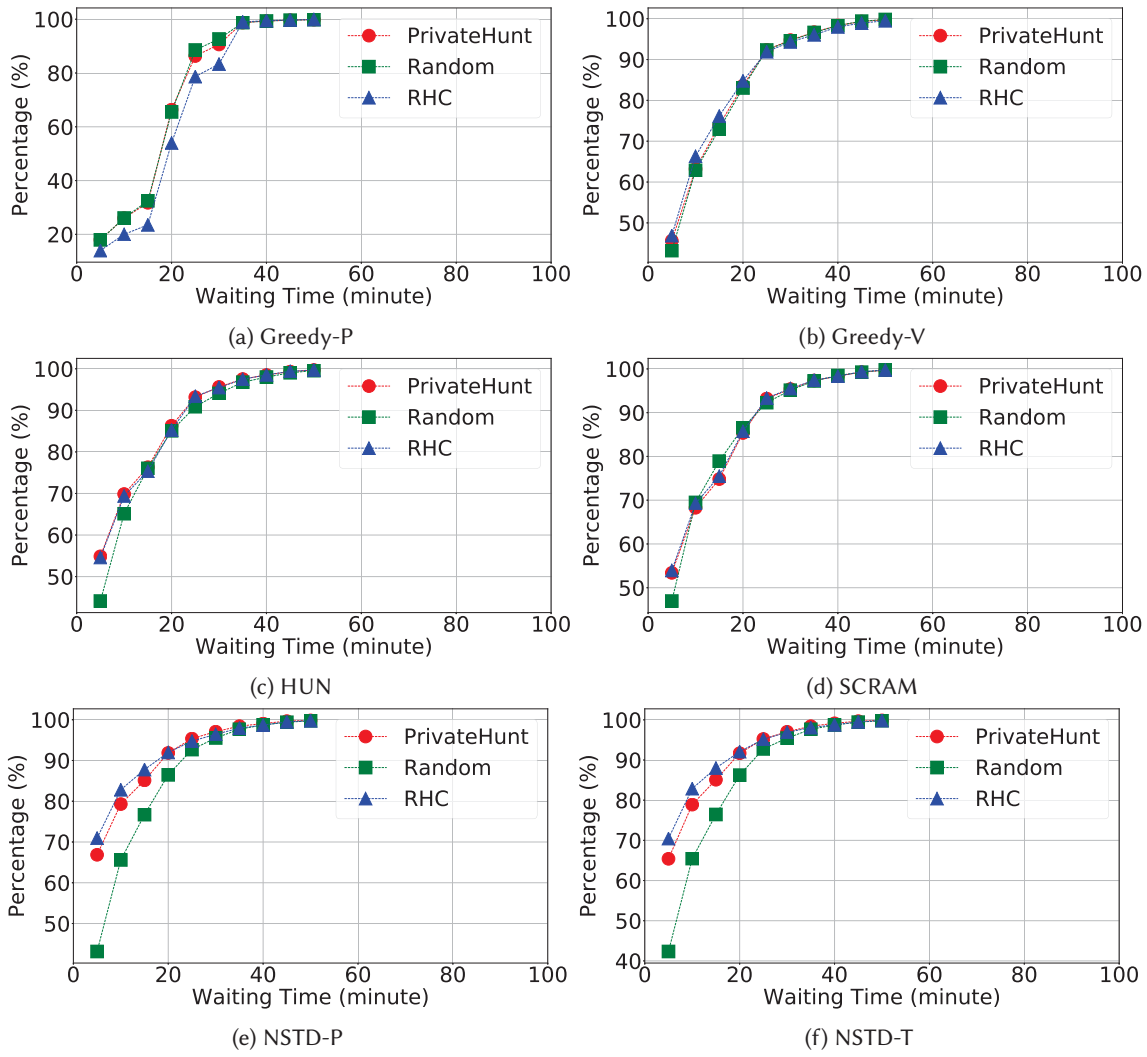

Fig. 16. Waiting time distribution

In the curves of Greedy-P and Greedy-T, the performances of PrivateHunt and Random are almost the same. The RHC curve is lower than other two curves. In the curves of Hun, PrivateHunt has a very close performance to RHC and outperforms Random with a waiting time from 5 minutes to 15 minutes. For SCRAM, with a waiting time from 10 minutes to 20 minutes, the performance of Random is better than that of PrivateHunt. In this method, PrivateHunt also has the same performance with RHC. In the results of NSTD-P and NSTD-T, RHC outperforms PrivateHunt by 2% before the waiting time of 20 minutes. PrivateHunt has smaller waiting time than Random and 90% of passengers have a waiting time smaller than 20 minutes, while that of Random is about 85%. The result in the comparison of waiting time is caused by that, PrivateHunt FHVs focus on the places with a higher probability with passenger requests, leading to the passengers in places with lower probability waiting longer. Even so, in the comparisons of waiting time, PrivateHunt still has better performance on NSTD-P and NSTD-T algorithms than Random and has a very close performance in HUN and SCRAM and better performance in Greedy-P compared with RHC. In summary, PrivateHunt has a slightly better performance in the reduction of waiting time than that of Random and has similar performance with RHC.

7.4 Impact of Cities

To study the impact of different cities on the performance of PrivateHunt, we implement it on the dataset of New York City. Fig. 17 gives the comparison of idle mileage rate and waiting time of PrivateHunt, RHC and Random by using Greedy-V assignment algorithm. The result shows PrivateHunt is robust. It also reduces the idle mileage rates for the FHVs on NYC, especially in the interval from 45% to 50% of idle mileage rates. For the waiting time, PrivateHunt still has similar performance compared with that of the two baselines.

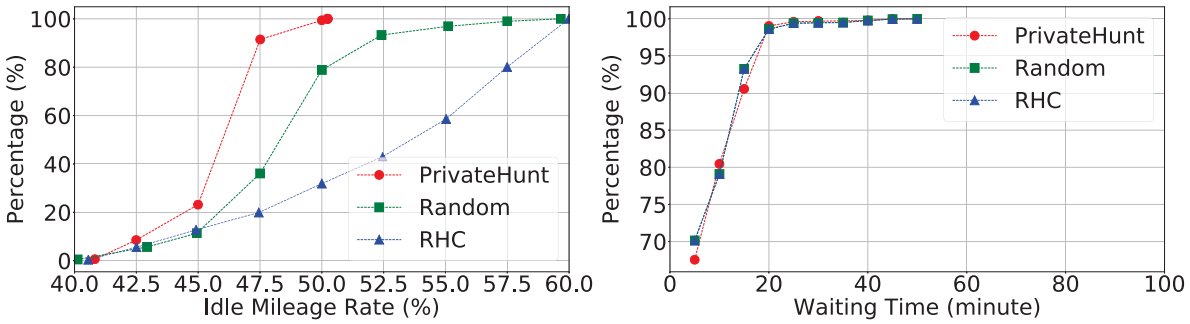

Fig. 17. Implementing On NYC

7.5 Impact of Days of Week

To study the impact of different days in one week, we also implement PrivateHunt with 6 algorithms on one-week datasets in Shenzhen City. Similar to the last section, we first show the result of idle mileage rate, followed with waiting time for passengers.

7.5.1 Idle mileage rate. Fig.18 shows the result of comparison of average idle mileage rates of PrivateHunt and Baseline. The X-axis is the days in one week and Y-axis is the average idle mileage rates. The lower the average idle mileage, the better the performance.

The performance of PrivateHunt on idle mileage rates is better than RHC and Random in the algorithms of Greedy-T, HUN, SCRAM, NSTD-P, and NSTD-T. In Greedy-P, RHC is better than PrivateHunt by having about

idle mileage rate of 1%. However, in the rest of algorithms, PrivateHunt is better. In the curves of the later four algorithms, PrivateHunt reduces average idle mileage rates of days in a week about 30% compared with Random and about 5% compared with RHC. Therefore, PrivateHunt performances well on different days in the metric of idle mileage rate.

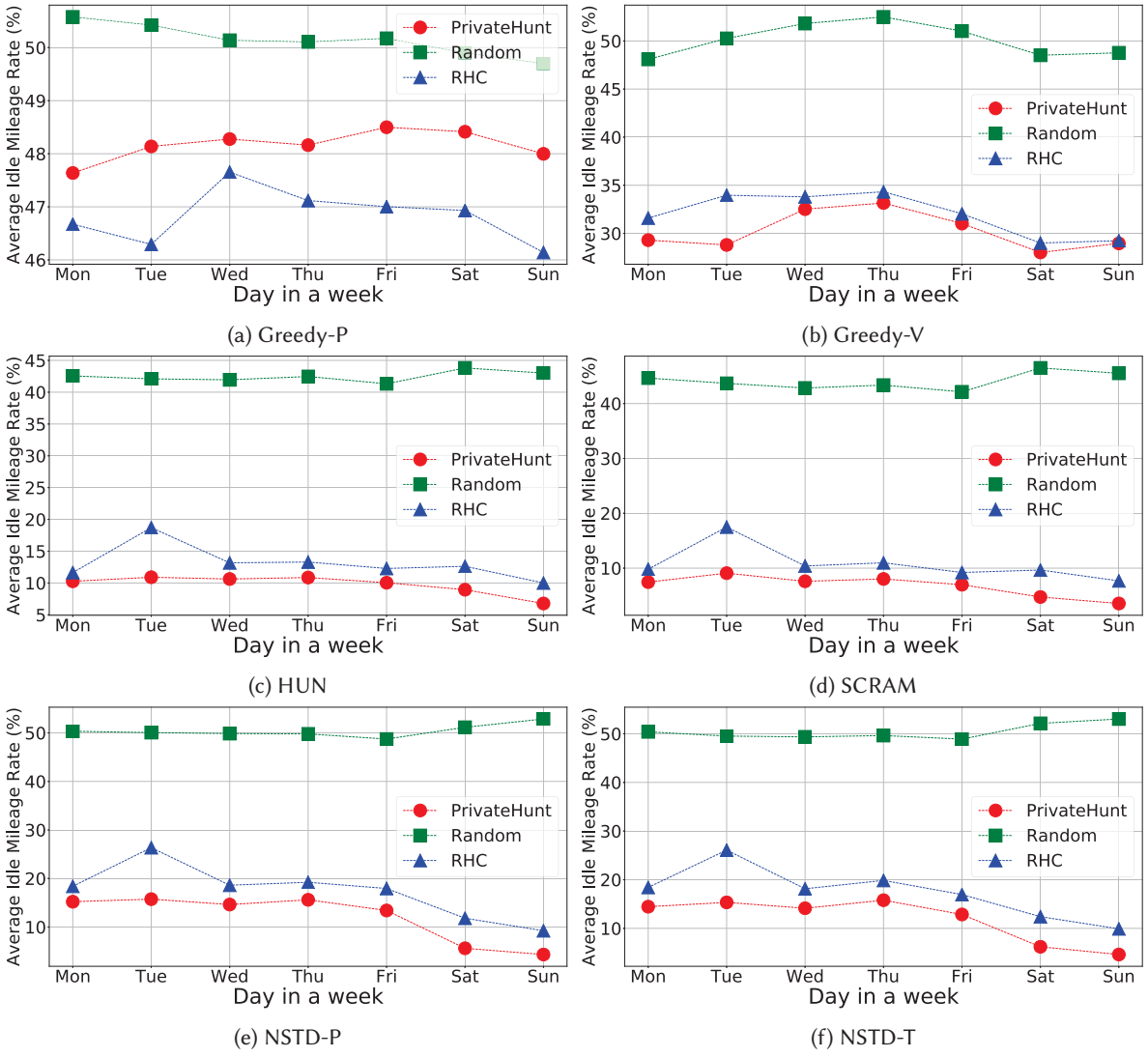

Fig. 18. Idle mileage rate distribution in one week

7.5.2 Waiting time. Fig.19 shows the average waiting time of passengers during one week for PrivateHunt and Baseline on these 6 assignment algorithms. The X-axis is the days in one week and Y-axis is the average waiting time. The lower the average waiting time, the better the performance. The results show that RHC has a much worse performance in the average waiting time compared with the other two strategies. In the results of

Greedy-P and Greedy-T, the differences between PrivateHunt and Random are very small. In contrast, in the algorithms of HUN, SCRAM, NSTD-P, and NSTD-T, PrivateHunt reduces waiting time about 2 minutes in average during the weekday and in the weekend, PrivateHunt is still better than Random. In summary, PrivateHunt has a better performance on passenger waiting time than RHC and Random with the algorithms we implemented.

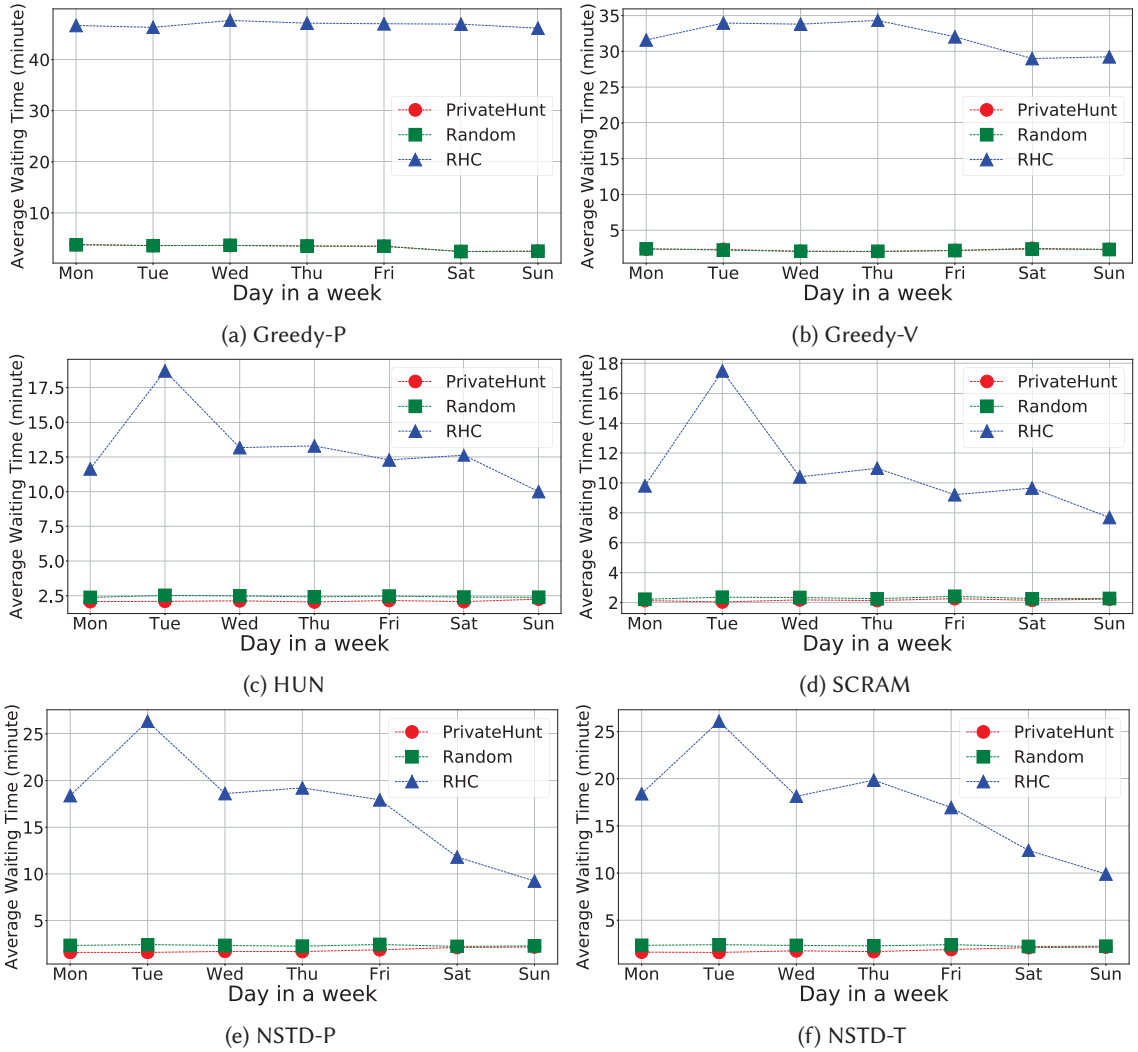

Fig. 19. Waiting time distribution in one week

7.6 Summary

In summary, PrivateHunt has a significant improvement in reducing the idle mileage rate. With PrivateHunt, the drivers of FHVs will spend less time on driving without any income. Compared to the outstanding effect on reduction of idle mileage, PrivateHunt has limited improvement in reducing the passenger waiting time. This is because (i) by using PrivateHunt, FHVs will focus on some places, while people who are not in the place will spend more time to wait for an FHV, (ii) the number of FHVs is much smaller than that of passengers, which means the places with a higher probability of passenger requests will consume most of the FHVs.

8 RELATED WORK

Vehicle dispatching systems have been proposed along with and wide adoption of real-time GPS devices [53] and the development of intelligent transportation systems [22]. The most related work to our PrivateHunt system is taxi dispatching where passenger demand and taxi supply are modeled by taxicab GPS data. The related work can be classified into two categories.

8.1 Modeling Demand and Supply

Based on GPS data from a metropolitan area, Zheng *et al.* [52] [47] present several novel models to capture urban transportation demand and supply. Ge *et al.* [15] [14] present models to recommend a taxicab driver with a sequence of pick-up points so as to maximize the profit, via a centralized solution. Phithakkitnukoon *et al.* [28] employ the naive Bayesian classifier with an error-based learning approach, which can obtain the number of vacant taxicabs at a given time and location to enhance the dispatching system. Balan *et al.* [3] provide a system allowing taxicab passengers to query the expected duration and fare of a planned trip based on previous trips. Yang *et al.* [45] propose a model for urban taxicab services, which indicates the vacant and occupied taxicab movements as well as the relationship between passengers and taxicab waiting time. Yamamoto *et al.* [44] present an adaptive routing scheme and a clustering scheme to enhance dispatching system via assigning vacant taxicabs to the locations with a high expected potential passengers number adaptively. Chang *et al.* [7] propose a model that can predict taxicab demand distribution based on weather condition, time and locations.

Besides the literature on data mining community, there are some studies about the demand and supply of FHV from other communities. [26] explores the impact of FHV on the taxi industry by leveraging the dataset of taxicabs of Shenzhen and claims that the taxi industry in Shenzhen has experienced a significant loss. [30] states that besides taxicabs at least half of FHV trips replaced modes including public transit and driving. [31] studies the MTA data in NYC and found that FHVs drag public transit use down. [9] found that Uber drivers drive more at times when earnings are high, and flexibly adjust to drive more at high surge times. The above literature from transportation and economic communities suggest that despite similar, the dispatch systems for taxicabs is inappropriate for FHVs.

Compared to the above data-driven systems, PrivateHunt is different in the objective. Our objective is to provide a dispatching service for FHV whereas the most existing work has been focused on commercial vehicle networks. Technically, our dispatching is based on (i) a demand model driven by data from three urban transportation systems, e.g., taxis, buses, and subway, whereas the most existing work is focused on single modalities; (ii) a supply model is driven by data from a large-scale FHV network; whereas the most existing work is focused on commercial vehicle networks with different mobility patterns.

8.2 Dispatch Strategies

Powell *et al.* [29] propose an approach to suggest profitable grid-based locations for taxicab drivers by constructing a profitability map where according to the potential profit calculated by the historical data, the nearby regions of the driver are scored serving as a metric for a taxicab driver decision making process. Gonzalez *et al.* [17] compute the fastest route by taking into account the speed and driving patterns of taxicabs, which are obtained from historical GPS trajectories. Ziebart *et al.* [54] utilize GPS trajectories obtained from 25 taxicabs, instead of providing the fastest route for drivers, aiming to predict the destination of drivers. Aslam *et al.* [1] design a method to model city-scale traffic based on a large-scale vehicular sensor network. Li *et al.* [23] study how to find passengers via several strategies for taxicab drivers in Hangzhou. Wu *et al.* [43] propose a system to assist mobile users in making transportation decisions based on demand and supply, such as taking a taxicab or not. In

addition to taxicab systems, some work on mobility demand and supply has also been focused on other transportation modalities, e.g., buses [4], subways [21], and private vehicles [16], as well as cell phone networks [19] [11] [20].

Most of the previous work [50] [49] is designed for taxicabs, whereas the mobility features of FHVs are significantly different with taxicabs. Even though some work [51] provide dispatch algorithms by considering features from the FHVs aspect, they lack the support from both the demand side and supply side. In spite of there are several papers about analysis on FHV data, they mostly are based on small-scale partial data and lack an integration of urban public and private transportation data [27] [8] [38]. Experimentally, our system is at least one or two orders of magnitude larger than existing academic experimental systems (e.g., GreenGPS [13], BusNet [10] and EasyTracker [5]), outperforming the existing systems in both system variety and scale.

9 DISCUSSION

Due to the high sensitivity of the data for human mobility study, we took two steps for privacy protections. (i) Anonymization: Before providing to us, all data with identifiable IDs, such as SIM card IDs, plate number or smart card IDs, are anonymized by the service providers who are not involved in this project. They replace the IDs with a serial identifier during the analyses. (ii) Aggregation: All the data obtained by PrivateHunt are given with the origin and destination information, ignoring the intermediate trajectory data.

10 CONCLUSION

In this paper, we introduce PrivateHunt, a cruising system for FHV services based on multi-source urban systems in real time. Our efforts lead to a few valuable insights for fellow researchers to design and implement real-time data-driven models for private vehicle hiring services. Specifically, these insights are that (i) heterogeneous urban transportation systems provide complementary data, which can be intelligently integrated together to model potential passenger demand for FHV services; (ii) an individual-based supply model capture the driving habit of private vehicle drivers to model transit supply for private vehicle hiring services; (iii) a dispatching strategy based on demand/supply modeling driven by multi-source urban data has potential to significantly reduce idle mileage for private drivers to find their passengers; (iv) while it is challenging to model and calibrate national-scale traffic models, it is more challenging to negotiate with service providers for data. We will share our data for the benefit of IMWUT community for urban-scale demand and supply modeling.

ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their valuable comments. This work is partially supported by the by Rutgers Global Center, China 973 Program (2015CB352400), National Natural Science Foundation of China (41401470), Research Program of Shenzhen under Grants JSGG20150512145714248, KQCX 2015040111035011, CYZZ 20150403111012661 and KJYY 20160331162313860.

REFERENCES

[1] Javed Aslam, Sejoon Lim, Xinghao Pan, and Daniela Rus. 2012. City-scale traffic estimation from a roving sensor network. In *Proceedings of 10th ACM Conference on Embedded Network Sensor Systems (SenSys '12)*.

[2] Yash Babar and Gordon Burtch. 2017. Examining the Impact of Ridehailing Services on Public Transit Use. (2017).

[3] Rajesh Krishna Balan, Khoa Xuan Nguyen, and Lingxiao Jiang. 2011. Real-time trip information service for a large taxi fleet. In *Proceedings of the international conference on Mobile systems, applications, and services (MobiSys '11)*.

[4] Sourav Bhattacharya, Santi Phithakkitnukoon, Petteri Nurmi, Arto Klami, Marco Veloso, and Carlos Bento. [n. d.]. Gaussian Process-based Predictive Modeling for Bus Ridership *(UbiComp '13)*.

[5] James Biagioni, Tomas Gerlich, Timothy Merrifield, and Jakob Eriksson. 2011. EasyTracker: Automatic Transit Tracking, Mapping, and Arrival Time Prediction Using Smartphones. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems (SenSys '11)*. ACM, New York, NY, USA, 68–81. https://doi.org/10.1145/2070942.2070950

[6] Shared-Use Mobility Center. 2016. Shared Mobility and the Transformation of Public Transit. *American Public Transportation Association* (2016).
[7] Han-wen Chang, Yu;chin Tai, and Jane Yung;jen Hsu. 2010. Context-aware taxi demand hotspots prediction. *Int. J. Bus. Intell. Data Min.* 5, 1 (Dec. 2010).
[8] Le Chen, Alan Mislove, and Christo Wilson. 2015. Peeking Beneath the Hood of Uber. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference (IMC '15)*. ACM, New York, NY, USA, 495–508. https://doi.org/10.1145/2815675.2815681
[9] M Keith Chen and Michael Sheldon. 2016. Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Platform.. In *EC*. 455.
[10] Kasun De Zoysa, Chamath Keppitiyagama, Gihan P. Seneviratne, and W. W. A. T. Shihan. 2007. A Public Transport System Based Sensor Network for Road Surface Condition Monitoring. In *Proceedings of the 2007 Workshop on Networked Systems for Developing Regions (NSDR '07)*. ACM, New York, NY, USA, Article 9, 6 pages. https://doi.org/10.1145/1326571.1326585
[11] Kateřina Dufková, Jean-Yves Le Boudec, Lukáš Kencl, and Milan Bjelica. [n. d.]. Predicting User-cell Association in Cellular Networks *(MELT'09)*.
[12] Raghu Ganti, Mudhakar Srivatsa, Anand Ranganathan, and Jiawei Han. 2013. Inferring human mobility patterns from taxicab location traces. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 459–468.
[13] Raghu K. Ganti, Nam Pham, Hossein Ahmadi, Saurabh Nangia, and Tarek F. Abdelzaher. 2010. GreenGPS: A Participatory Sensing Fuel-efficient Maps Application. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys '10)*. ACM, New York, NY, USA, 151–164. https://doi.org/10.1145/1814433.1814450
[14] Yong Ge, Chuanren Liu, Hui Xiong, and Jian Chen. [n. d.]. A taxi business intelligence system. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*.
[15] Yong Ge, Hui Xiong, Alexander Tuzhilin, Keli Xiao, Marco Gruteser, and Michael Pazzani. 2010. An energy-efficient mobile recommender system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10)*.
[16] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. 2011. Unveiling the Complexity of Human Mobility by Querying and Mining Massive Trajectory Data. *The VLDB Journal* (2011).
[17] Hector Gonzalez, Jiawei Han, Xiaolei Li, Margaret Myslinska, and John Paul Sondag. 2007. Adaptive fastest path computation on a road network: a traffic mining approach. In *Proceedings of the 33rd international conference on Very large data bases (VLDB '07)*.
[18] Josiah P Hanna, Michael Albert, Donna Chen, and Peter Stone. 2016. Minimum cost matching for autonomous carsharing. *IFAC-PapersOnLine* 49, 15 (2016), 254–259.
[19] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. [n. d.]. Human Mobility Modeling at Metropolitan Scales *(MobiSys '12)*.
[20] Sibren Isaacman, Richard A. Becker, Ramón Cáceres, Stephen G. Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. Ranges of human mobility in Los Angeles and New York. In *PerCom Workshops*.
[21] Neal Lathia and Licia Capra. [n. d.]. How Smart is Your Smartcard?: Measuring Travel Behaviours, Perceptions, and Incentives *(UbiComp '11)*.
[22] D. Lee, H. Wang, R. Cheu, and S. Teo. 2004. Taxi dispatch system based on current demands and real-time traffic conditions. In *Journal of the Transportation Research Board*.
[23] Bin Li, Daqing Zhang, Lin Sun, Chao Chen, Shijian Li, Guande Qi, and Qiang Yang. 2011. Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset. In *PerCom Workshops*.
[24] Tracey Lien. 2016. Lyft defies predictions by continuing to grow as a rival to Uber. (2016). http://www.latimes.com/business/technology/la-fi-0105-lyft-growth-20160105-story.html [Online; accessed 5-January-2016].
[25] Fei Miao, Shuo Han, Shan Lin, John A Stankovic, Desheng Zhang, Sirajum Munir, Hua Huang, Tian He, and George J Pappas. 2016. Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach. *IEEE Transactions on Automation Science and Engineering* 13, 2 (2016), 463–478.
[26] Yu Marco Nie. 2017. How can the taxi industry survive the tide of ridesourcing? Evidence from Shenzhen, China. *Transportation Research Part C: Emerging Technologies* 79 (2017), 242–256.
[27] Jiyong Park, Junetae Kim, and Byungtae Lee. 2016. Are Uber Really to Blame for Sexual Assault?: Evidence from New York City. In *Proceedings of the 18th Annual International Conference on Electronic Commerce: E-Commerce in Smart Connected World (ICEC '16)*. ACM, New York, NY, USA, Article 12, 7 pages. https://doi.org/10.1145/2971603.2971615
[28] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti. 2011. Taxi-aware map: Identifying and predicting vacant taxis in the city. In *Proc. AMI*.
[29] J. Powell, Y. Huang, F. Bastani, and M. Ji. 2011. Towards reducing taxicab cruising time using spatio-temporal profitability maps. In *Proceedings of the 12th International Symposium on Advances in Spatial and Temporal Databases*.
[30] Lisa Rayle, Danielle Dai, Nelson Chan, Robert Cervero, and Susan Shaheen. 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. *Transport Policy* 45 (2016), 168–178.

[31] Nicole Sadowsky. [n. d.]. Estimating the Impact of Ride-Hailing App Services on Public Transportation Use in Major US Urban Areas. ([n. d.]).
[32] Sarah. 2016. Why people choose to drive with Uber. (2016). https://newsroom.uber.com/uk/why-people-choose-to-drive-with-uber/ [Online; accessed 3-June-2016].
[33] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.
[34] Jiyuan Tan, Yibin Huang, Zhengxi Li, Li Wang, Weiwei Guo, and Honghai Li. 2017. The characteristics on commuting travel mode split and origin destination (OD) distribution in Shenzhen, China. In *Data Driven Control and Learning Systems (DDCLS), 2017 6th*. IEEE, 716–720.
[35] Jinjun Tang, Fang Liu, Yinhai Wang, and Hua Wang. 2015. Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and its Applications* 438 (2015), 140–153.
[36] Henk J Van Zuylen and Luis G Willumsen. 1980. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological* 14, 3 (1980), 281–293.
[37] Scott Wallsten. 2015. The competitive effects of the sharing economy: how is Uber changing taxis? *Technology Policy Institute* (2015).
[38] Hua Wei, Yuandong Wang, Tianyu Wo, Yaxiao Liu, and Jie Xu. 2016. ZEST: A Hybrid Model on Predicting Passenger Demand for Chauffeured Car Service. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. ACM, New York, NY, USA, 2203–2208. https://doi.org/10.1145/2983323.2983667
[39] Wikipedia. 2016. Didi Chuxing — Wikipedia, The Free Encyclopedia. (2016). https://en.wikipedia.org/w/index.php?title=Didi_Chuxing&oldid=749544968 [Online; accessed 14-November-2016].
[40] Wikipedia. 2016. Lyft — Wikipedia, The Free Encyclopedia. (2016). https://en.wikipedia.org/w/index.php?title=Lyft&oldid=748891300 [Online; accessed 11-November-2016].
[41] Wikipedia. 2016. Uber (company) — Wikipedia, The Free Encyclopedia. (2016). https://en.wikipedia.org/w/index.php?title=Uber_(company)&oldid=754324190 [Online; accessed 12-December-2016].
[42] Jochen Wirtz and Christopher Tang. 2016. Uber: Competing as Market Leader in the US versus Being a Distant Second in China. In *SERVICES MARKETING: People Technology Strategy*. 626–632.
[43] Wei Wu, Wee Siong Ng, Shonali Krishnaswamy, and Abhijat Sinha. 2012. To Taxi or Not to Taxi? - Enabling Personalised and Real-Time Transportation Decisions for Mobile Users. In *Proceedings of the 2012 IEEE 13th International Conference on Mobile Data Management (MDM '12)*.
[44] K. Yamamoto, K. Uesugi, and T. Watanabe. 2010. Adaptive routing of cruising taxis by mutual exchange of pathways. In *Knowledge-Based Intelligent Information and Engineering Systems*.
[45] Hai Yang, C.S. Fung, K.I. Wong, and S.C. Wong. 2010. Nonlinear pricing of taxi services. In *Transportation Research Part A: Policy and Practic*.
[46] Zidong Yang, Ji Hu, Yuanchao Shu, Peng Cheng, Jiming Chen, and Thomas Moscibroda. 2016. Mobility Modeling and Prediction in Bike-Sharing Systems. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '16)*. ACM, New York, NY, USA, 165–178. https://doi.org/10.1145/2906388.2906408
[47] Jing Yuan, Yu Zheng, and Xing Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*.
[48] Nicholas Jing Yuan, Yingzi Wang, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2013. Reconstructing individual mobility from smart card transactions: A space alignment approach. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 877–886.
[49] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. [n. d.]. Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales. In *the 20th ACM International Conference on Mobile Computing and Networking (MobiCom'14)*.
[50] Desheng Zhang, Ye Li, Fan Zhang, Mingming Lu, Yunhuai Liu, and Tian He. [n. d.]. coRide: Carpool Service with a Win-Win Fare Model for Taxicab Networks. In *the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys'13)*.
[51] Huanyang Zheng and Jie Wu. 2017. Online to Offline Business: Urban Taxi Dispatching with Passenger-Driver Matching Stability. In *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 816–825.
[52] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. 2010. Understanding transportation modes based on GPS data for web applications. *ACM Trans. Web* 4, 1 (Jan. 2010).
[53] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. 2008. Understanding mobility based on GPS data. In *Proceedings of the 10th international conference on Ubiquitous computing (UbiComp '08)*.
[54] Brian D. Ziebart, Andrew L. Maas, Anind K. Dey, and J. Andrew Bagnell. 2008. Navigate like a cabbie: probabilistic reasoning from observed context-aware behavior. In *Proceedings of the 10th international conference on Ubiquitous computing (UbiComp '08)*.

Received August 2017; revised November 2017; accepted January 2018