A Force-directed Approach to Seeking Route Recommendation in Ride-on-demand Service Using Multi-source Urban Data

Suiming Guo, Chao Chen, Jingyuan Wang, Yan Ding, Yaxiao Liu, Ke Xu, *Senior Member, IEEE,* Zhiwen Yu, *Senior Member, IEEE,* and Daqing Zhang, *Fellow, IEEE,*

Abstract—The rapidly-growing business of ride-on-demand (RoD) service such as Uber, Lyft and Didi proves the effectiveness of their new service model – using mobile apps and dynamic pricing to coordinate between drivers, passengers and the service provider, to manipulate the supply and demand, and to improve service responsiveness as well as quality. Despite its success, dynamic pricing creates a new problem for drivers: how to seek for passengers to maximize revenue under dynamic prices. Seeking route recommendation has already been studied extensively in traditional taxi service, but most studies do not consider the effects of taxis and passengers on the seeking taxi simultaneously. Further, in RoD service it is necessary to consider more factors such as dynamic prices, the status of other transportation services, etc. In this paper, we employ a force-directed approach to model, by analogy, the relationship between vacant cars and passengers as that between positive and negative charges in electrostatic field. We extract features from multi-source urban data to describe dynamic prices, the status of RoD, taxi and public transportation services, and incorporate them into our model. The model is then used in route recommendation in every intersection so that a driver in a vacant RoD car knows which road segment to take next. We conduct extensive experiments based on our multi-source urban data, including RoD service operational data, taxi GPS trajectory data and public transportation distribution data, and results not only show that our approach outperforms existing baselines, but also justify the need to incorporate multi-source urban data and dynamic prices.

Index Terms—Ride-on-demand, dynamic pricing, seeking route, driver revenue.

1 Introduction

The birth and success of ride-on-demand (RoD) service such as Uber, Lyft and Didi mark a change in city transportation. Compared to the traditional taxi service, RoD service offers a convenient, affordable and flexible experience for passengers; for drivers, it allows them to arrange working hours flexibly, and to enter the service without the hassle of applying for licenses or medallions. An increasing amount of passengers are now using RoD services as an everyday choice.

RoD service uses a new service model that distinguishes itself from taxi service, with two key features – **mobile-app-based** and **dynamic pricing**.

- *S. Guo is with the College of Information Science and Technology, Jinan University, Guangzhou, 510632, China. E-mail: guosuiming@email.jnu.edu.cn*
- *C. Chen is with Chongqing University, Chongqing, China. E-mail: cschaochen@cqu.edu.cn*
- *J. Wang is with the Beijing Advanced Innovation Center for Big Data and Brain Computing, and with School of Computer Science and Engineering, Beihang University, Beijing 100191, China. E-mail: jywang@buaa.edu.cn*
- *Y. Ding is with State University of New York, NY, US. E-mail: yding25@binghamton.edu*
- *Y. Liu and K. Xu are with Tsinghua University, Beijing, China. K. Xu is also with BNRist and Peng Cheng Laboratory. E-mail: rootliu@gmail.com, xuke@mail.tsinghua.edu.cn*
- *Z. Yu is with Northwestern Polytechnical University, Xi'an, China. E-mail: zhiwenyu@nwpu.edu.cn*
- *D. Zhang is with Institut Mines-Telecom/Telecom SudPais, Evry Cedex, France. E-mail: daqing.zhang@telecom-sudparis.eu*

Manuscript received January 30, 2020; revised XX XX, 2020.

Mobile-app-based. In a RoD service, both passengers and drivers rely on the specially designed mobile apps on their smart phones. For passengers, they use the *passenger app* to request for rides, to search for nearby drivers, to find out the distribution of dynamic prices nearby, and to pay for rides. For drivers, they use the *driver app* to view, accept or decline nearby passenger requests, to find out the distribution of dynamic prices across the city, and to manage and review recent rides. In fact, mobile apps serve as a bridge that connects drivers, passengers and the service provider – in such a way that the three parties can share information through mobile apps, to and from each other. It is true that in recent years the taxi services in some cities also resort to mobile apps to manage trips, but the number of trips created in this way is still relatively small, and in RoD service all trips are created through mobile apps.

In addition to information sharing, the *mobile-app-based* feature also enables drivers and passengers to match in advance. In taxi service, most non-reserved trips are created through street-hailing, in which the driver and the passenger have to be within sight of each other. Comparatively, in RoD service, the driver and the passenger can be matched when they are close enough (e.g., 1 or 2 km), and the service provider will then attempt to match the closest driver to the passenger. Matching-in-advance is possible in RoD service because the service provider has full knowledge of the location and status of both drivers and passengers. Matching-in-advance helps to increase matching probability, as it is no longer necessary for a driver to be in exactly the same location with a potential passenger to pick him up.

Dynamic pricing. In most RoD services, dynamic pricing mechanism is used to manipulate the supply (i.e., the number of cars on the road) and demand (i.e., the number of passenger requests). Basically, when demand exceeds supply, a higher price is used to attract more drivers to come and to defer requests from passengers not in a hurry; when supply exceeds demand, a lower price does just the opposite. In most cases, dynamic pricing is represented by a dynamic price multiplier, and the total trip fare is the product of a *dynamic* multiplier (based on the supply and demand condition) and a *fixed* normal price (based on trip time and distance).

The introduction of dynamic pricing indeed makes the service and prices more flexible, and improves the service's responsiveness to changes in supply or demand. However, it also creates a new problem for drivers: *how to seek for passengers to make more revenue under dynamic pricing?* Studies on this problem can be roughly divided into two directions – *seeking strategies analysis* and *seeking route recommendation*. Seeking strategies analysis has already been studied in [1] – it is on the macro-level, and focuses on mining general, profitable strategies that drivers should keep in mind in seeking for passengers. Seeking route recommendation is on the micro-level, and tries to guide a driver, at every intersection, to the right road segment that may lead to a higher profit. Here, we concentrate on seeking route recommendation.

Seeking route recommendation receives little attention in RoD service. In taxi service, it has been studied thoroughly using a number of heuristics and algorithms including recommending a driver to local or global hotspots [2], modelling a driver's behavior using a Markov decision process model [3], [4], simulating a driver's behavior using a force-directed approach [5], and etc. In RoD service, however, even the macro-level studies (i.e., seeking strategies analysis [1]) are rare, and are always in non-rigorous forms such as blogs or news stories. There are literally no, to the best of our knowledge, studies on seeking route recommendation that consider new features in RoD service such as mobile-app-based and dynamic pricing.

Because of RoD service's new features, seeking route recommendation in RoD service requires the consideration of three more factors than in taxi service:

- **Matching-in-advance**: As mentioned, matching-in-advance allows a driver and a passenger to match before they come to see each other, and it increases matching probability. We need to take into account this feature in driver behavior simulation.
- **Dynamic pricing**: As one of the key features, dynamic pricing should be considered in modelling a location's or region's attractiveness to a driver. In taxi service, a region with a higher demand is already good enough. But in RoD service, among two regions with the same level of demand, the region with a higher dynamic price multiplier maybe a better suggestion for a driver.
- **Status of other transportation services**: Though with some new features, RoD service is similar to taxi service, and hence there is a complex relationship between them. Also, profitable seeking locations in RoD service are related to other public transportation services such as bus or metro. [1], [6] conclude from real data that RoD and these services are complementary instead of competitive to each other.

In this paper, we employ a force-directed approach to tackle the seeking route recommendation problem in RoD service. The force-directed approach borrows the concept of physical interaction in a electrostatic field – opposites attract and likes repel. By analogy, if we regard vacant cars as positive point charges and potential passengers as negative point charges, then the relationship between vacant car and passenger is similar to charge interactions: potential passenger attracts vacant car, while vacant car repels each other. For a vacant car at a particular intersection, the aggregated force, including the repulsive and attractive forces from vacant cars and potential passengers nearby, is calculated and the road segment closest to its direction should be recommended to the driver. This approach has two advantages: (a) the effects of vacant cars and potential passengers on a seeking vacant car could be considered simultaneously; and (b) drivers at different locations have different aggregated forces, and hence they generally receive different road segment recommendations. This prevents the common problem of recommending the same route to many drivers in other approaches. In modelling such forces, we introduce multi-source urban datasets, from which features are extracted to describe not only the status of RoD service, but also dynamic prices and status of other transportation services (i.e., taxi, bus and metro). The matching between a vacant car and a potential passenger is achieved when the car arrives at an intersection and when there is at least one potential passenger close enough.

Our contributions are three-fold:

- Our study is one of the very few on seeking route recommendation in RoD service. Previous relevant studies either are confined to taxi service, or fail to consider new features such as dynamic pricing. Instead, we take into account three more factors in RoD service – matching-in-advance, dynamic pricing, and status of other transportation services. This helps us to describe the status of RoD service more accurately, and to improve the effectiveness of recommendation.
- We are the first, as we know, to introduce multi-source urban data into seeking route recommendation. This allows us to extract features to describe the status of RoD service, dynamic prices, and status of other transportation services, and these features are used in modelling the attractive and repulsive forces.
- We adopt and extend the force-directed approach in our study. It helps us to model drivers' seeking behavior, with the two advantages mentioned above. We extend this approach by introducing multi-source urban data and considering matching-in-advance, dynamic pricing as well as status of other transportation services. This approach is evaluated by extensive experiments based on real data.

The remainder of the paper is organized as follows. Section 2 reviews related work and Section 3 explains our multi-source urban data. In Section 4 we present some patterns in RoD service, including utilization rate, passenger

density, driving pattern and revenue efficiency. The force-directed approach is discussed with details in Section 5. Evaluations of our approach based on real data are conducted and presented in Section 6. Some discussions are shown in Section 7, and Section 8 concludes the paper.

2 Related Work

Seeking route recommendation has already been studied extensively in taxi service, but only receives very limited attention in emerging RoD services. We first review related work in RoD service, then discuss previous studies on seeking route recommendation. As to the methodology we use, we also review related work in force-directed approach.

RoD Service. RoD service, also known as on-demand ride hailing, is a relatively new transportation service compared to taxi service. There are thus much fewer studies on RoD service. Based on the similarities between RoD and taxi service, a number of studies compare the differences of the price, waiting time, incentives and service quality between them from a data statistical perspective. For examples, [7] claims that Uber can reduce the waiting time significantly but may not always give the lowest price; [8], [9], [10] study and discuss the change of market share of taxi and public transportation services before and after Uber's entrance; [11] performs a spatio-temporal head-to-head comparison between these two services; [12], [13] focus on the market effects of Uber's entrance, such as the relationship with public transit, the changes to drivers' behavior, etc.

As mentioned, dynamic pricing is one of the key features of RoD service, and is also studied from different perspectives. For examples, [14], [15], [16] concentrate on dynamic pricing's effects in balancing and redistributing the supply and demand, increasing driver revenue and reducing passenger waiting time; [17] is one of the early work that tries to mine data based on simulated users and evaluate Uber's surge pricing mechanism treating it as a black-box; [6], [18], [19], instead, study and analyze the demand, the effect of dynamic pricing, passengers' reaction to prices, and dynamic price prediction, based on real data from typical RoD services. Besides studying RoD service based on data and computation methodologies, some studies analyze dynamic pricing [14] and its effects on supply elasticity [20] and consumer surplus [21] from economics perspective.

Seeking route recommendation. Seeking strategies analysis and route recommendation are two steps in enabling drivers to earn more, and both of them have been studied extensively in taxi service. Seeking strategies analysis can be regarded as macro-level studies. In taxi service, for examples, seeking strategies are studied by mining GPS trajectories [22], [23], to identify the most profitable strategies under different circumstances. In RoD service, studies are rare and few of them consider dynamic pricing. As an example, [1] studies seeking strategies by mining from multi-source urban data including ROD data, taxi data, public transportation service data and POI data.

For seeking route recommendation, considerable efforts have been done in taxi service. For instances, [24] recommends routes to drivers to minimize the distance between the taxi and an anticipated customer request; [3], [4] build a Markov Decision Process model to help drivers to earn more; [25] uses reinforcement learning to solve the same problem; [26] also uses Markov Decision Process model, but to recommend routes for electric taxis, by incorporating the charging process and battery constraint; [5] applies the force-directed approach and recommends routes to taxi drivers; etc. There are also studies aiming to improve driver-passenger matching probability from other perspectives – e.g., [27] introduces the concept of dynamic waiting to enable one driver to be matched to more than one passengers.

In RoD service, however, seeking route recommendation has not received extensive attention. For example, [28] attempts to optimize earning in on-demand ride-hailing based on theoretical modelling of drivers, cities and the service itself. Most existing studies do not take into account new features in RoD service such as dynamic pricing.

Force-directed approach. It models a problem as a system of particles with forces acting between them, and the system would then go into an equilibrium or behave in some particular way with the interaction forces. This approach has been widely used in design automation [29], graph visualization [30], map-matching [31], urban computing [5], and etc. Among them, [5] tackles a problem similar to ours – seeking route recommendation in taxi service.

Different from the above works, our study on seeking route recommendation emphasizes the differences between RoD and taxi service. We summarize that "*mobile-app-based*" and "*dynamic pricing*" are two key features in RoD service, and claim that we should take into account *matching-in-advance*, *dynamic pricing*, and *status of other transportation services* when recommending routes to drivers. The joint effort of any one or more of these three factors on seeking route recommendation in RoD service has never been studied before. While adopting and extending the force-directed approach, we also show new patterns in RoD service, study the way of incorporating features from multi-source urban data that describe dynamic prices and status of other transportation services, evaluate the effectiveness of such incorporation, and etc. Besides, our study is based on city-scale real multi-source urban data, making our results more tenable.

3 Multi-source Urban Data

We extract features from multi-source urban data to describe the status of RoD service, dynamic pricing, and status of taxi, bus and metro service. In this section, we explain the RoD service data (including the order data, GPS trajectories data, and the event-log data), taxi GPS trajectories data and bus & metro distribution data. Tab. 1 shows examples of data entries in these datasets, for illustration purpose only.

3.1 RoD Service Data

RoD service data is certainly the fundamental data we use in this study. Studies on taxi services mostly rely on GPS trajectories data that can describe how taxis move during a certain period. In RoD service, however, more datasets are available due to the mobile-app-based feature. All communication messages between passengers, drivers and the service provider are carried out through mobile apps. Hence, besides GPS trajectories data of RoD cars, there are also order data and event-log data available.

TABLE 1
Examples of data entries in multi-source urban data.

Dataset	Example data entry
RoD order data	*boarding_time*: "2015-12-21 13:55", *boarding_loc*: "116.478202, 39.910898", *arriving_time*: "2015-12-21 14:13", *arriving_loc*: "116.460157, 39.926769", *user_ID*: "2145622446", *driver_ID*: "21881", *car_ID*: "4569", *type*: "business".
RoD trajectories data	*upload_time*: "2015-11-01 05:19:13", *loc*: "116.442497, 39.852982", *car_ID*: "4731", *speed*: "0", *direction*: "180".
RoD event-log data	*event_time*: "2015-11-01 11:53:07", *event_loc*: "116.449501, 39.931717", *estimated_trip_fare*: "49.6", *price_mul*: "1.6", *user_ID*: "2145622446".
Taxi trajectories data	similar to the example entry in RoD trajectories data.
Bus & Metro distribution data	*loc*:"116.298, 39.878", *bus_station_count*: "5", *metro_station_count*: "1", *bus_line_count*: "26", *metro_line_count*: "4".

Our data is from Shenzhou UCar (https://bit.ly/2MG47xz), a major RoD service provider in China. As mentioned, in RoD service trips are created on mobile apps. Fig. 1 shows the user interface of its *passenger app*. Basically, after the user has filled in addresses and chosen "when to ride" or coupons, the app sends the information to the service provider and obtains (a) the estimated trip fare and (b) the current dynamic price multiplier. The user then chooses to accept the price (i.e., "Ride a Car!") or to give up the current request if s/he considers the price multiplier too high.

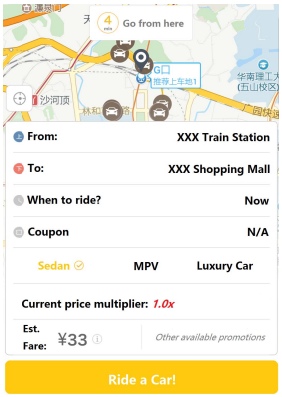

Fig. 1. The user interface of a typical RoD service.

We obtain the following three datasets:

The Order Data. This dataset describe each order's boarding/arriving time and location, the unique ID of the user/driver/car/order, the type of order, etc. Our data is from Beijing, as it is one of the most representative metropolis as well as the biggest market of the service provider. All user and driver IDs are anonymized so that one cannot relate an ID to a real person or car. The dataset is from Nov. 2015 to Mar. 2016, and contains about 2.7 million orders. This time range of almost 5 months is a relatively long period, making the datasets representative enough. Moreover, there is only one major holiday during this time range - the Spring Festival (on Feb. 2016) - and this reduces the impacts on traffic and trip patterns from holiday seasons (the Christmas Day is not a holiday in China, and on the New Year Day people only get one day off).

GPS Trajectories. This dataset is similar to the commonly used GPS trajectories dataset in taxi service, containing the GPS records of every single car. Fields include the location of the car (i.e., longitude and latitude), data upload time, the unique ID of the car, etc. For each car, the time interval between two consecutive records is two minutes. The time range of the dataset is the same to the order data, and on each day there are roughly 3,500 cars on the road working for the service provider.

The Event-log Data. The event-log data is new in RoD service. By saying an "event", we mean the *EstimateFee* event generated when the *passenger app* sends back the information to the service provider. This dataset contains the record of *EstimateFee* event in the same time range, describing the event time, event location, estimated trip fare, price multiplier, the unique user ID, etc. In total there are 14,832,418 entries.

In our study, the last dataset is related to the "dynamic pricing" feature. Firstly, our order data does not give information such as the trip fare or the price multiplier of each order, possibly due to privacy concerns. Secondly, it covers more information than orders, as those fare estimations that do not lead to order creations are also recorded. Later in Section 5, we will give more details about how to quantitatively describe dynamic pricing based on all these datasets.

3.2 Taxi Service GPS Trajectories Data

Though our study is based on RoD service, we also use taxi GPS trajectories data for two reasons. Firstly, [1] points out that as RoD service is similar to taxi service on many aspects, they have influences on each other. [1] concludes that these two services are complementary rather than competitive to each other, e.g., a region with more taxis is also profitable for RoD drivers to seek for passengers. In other words, the status of taxi service is an indication of a region's popularity. Secondly, the taxi service data helps to characterize the general traffic condition of different regions or locations, e.g., the number of taxis, the average speed of taxis, etc.

Similar to the GPS trajectories in Section 3.1, this dataset covers about 30,000 taxis in Beijing from Nov. 2015 to Mar. 2016, but the upload time interval is 30 seconds. For each day, the volume of dataset ranges from 45 to 50 million entries.

3.3 Bus & Metro Distribution Data

Compared to taxi service, public transportation services such as bus and metro are less similar to RoD service, but they are also influential. [1] reaches a conclusion from

real data that bus and metro are also complementary to RoD service. Besides indicating a region's popularity, the presence of bus and metro stations also makes it possible for RoD drivers to provide connecting services – picking up a passenger who just alighted from a bus or train, or delivering a passenger to a bus or metro station.

Different from the RoD or taxi service data that provides exact information of each driver, passenger or trip, our bus & metro distribution data only counts the number of bus and metro stations and lines within a 500-meter radius of a given location. This choice is not as accurate as the number of buses or metro trains around, but as bus and metro have relatively fixed time tables, most people decide whether to take public transportation only based on the availability of bus & metro lines or stations nearby. The dataset is crawled from AMap service [32] (one of the largest digital map service providers in China), and for the whole city, there are more than 7,700 bus stations and 380 metro stations.

4 Patterns in RoD Service

In this section, we show, based on real service data, some patterns and observations related to drivers in RoD service. The motivations of studying these patterns in RoD service are three-fold:

- Improving the understanding about RoD service. RoD service is a new transportation service: its patterns are either different from that of taxi, or unavailable in taxi service. These statistics and patterns help to understand the motivation and methodologies.
- Inspiring the modelling in our approach. The ideas or assumptions in our approach and its modelling (see Section 5) are not out of imagination; they are based on the observations of patterns from real data.
- Providing baselines for our evaluation. These patterns also provide metrics, as baselines, for our model evaluation (see Section 5 and 6), so that we could compare the performance of our seeking route recommendation with that of ground truth.

Patterns presented here belong to four categories: *utilization rate*, *passenger density*, *driving pattern* and *revenue efficiency*. They are about drivers' revenue-making capability.

4.1 Utilization Rate

We calculate the distance and time utilization rate of each driver. The utilization rates are calculated based on a single driver's driving history during one single day. To calculate these utilization rates, four quantities are defined: T as the total driving time during the day, t as the total driving time with a passenger on board, D as the total driving distance during the day, and d as the total driving distance with a passenger on board. Then the time utilization rate τ_t and distance utilization rate τ_d are defined as:

$$\tau_t = \frac{t}{T}, \tau_d = \frac{d}{D}. \quad (1)$$

For each driver, the time and distance utilization rate characterize the driver's efficiency in finding passengers. Following (1), we can calculate the utilization rates for every driver on every day across our RoD dataset. Specifically, the driving time and distance can both be calculated based on the GPS trajectory dataset of RoD service. To get an intuitive understanding of utilization rates, we first choose three timeslots – [7am, 9am] as the morning rush hours, [5pm, 7pm] as the evening rush hours, and [10am, 1pm] as typical non-rush hours around noon. These three timeslots are typical rush and non-rush hours during a day, and the representativeness of these timeslots has already been verified in [6], [19]. We then plot the distribution of distance and time utilization rate on weekdays and weekends in Fig. 2 to Fig. 5. In each figure, we plot the distribution across the whole day and during the three timeslots.

We have the following observations:

- For the distance utilization rate on weekdays, it is clear that the rates are higher during morning and evening rush hours: during these time periods, drivers take shorter trips to seek for passengers due to more passenger requests. By comparison, the distance utilization rates are lower during non-rush hours, for the possible reason that the number of passenger requests is reduced.
- Comparing the distance utilization rates on weekdays and weekends, we first notice that across the whole day, the rates are a little bit lower on weekends (e.g., the most frequently seen utilization rate is about 15% smaller on weekends than on weekdays). Also, the distance utilization rates during the three typical rush and non-rush hours are fairly close, indicating that there is few or no fluctuation of passengers' requests during the day on weekends, which has been verified in [1], [18].
- For the time utilization rate, we also have similar observations. Besides, Fig. 4 and Fig. 5 also show that for the time utilization rates on both weekdays and weekends, the differences between three typical rush and non-rush hours are much more obvious, compared to the differences for the distance utilization rates. This is due to the impact of driving speed. [1] observe that the average driving speed is faster in non-rush hours than in rush hours, so the bigger differences in time utilization rates are compensated by the difference of driving speed, leading to smaller differences in distance utilization rates.

4.2 Passenger Density

Passenger density is one of the most important features to take into account in recommending seeking routes. For example, a taxi driver chooses a region to seek for passengers based on personal experience that there are many potential passenger requests in this region during a particular time-of-day. Studies on seeking route recommendation in taxi service are also intended to recommend "hot spots" (i.e., locations with a high demand), either locally or globally, to drivers, or to rank a number of locations according to their popularity before choosing one or more candidates based on some certain criteria.

It is thus necessary to inspect the patterns of passenger density in RoD service. By "passenger", in this paper we only count the met demand – i.e., passenger requests that

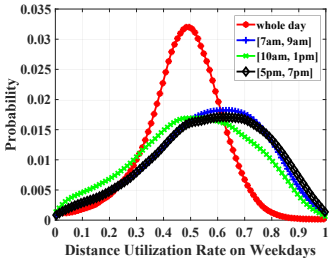

Fig. 2. The distribution of distance utilization rate on weekdays.

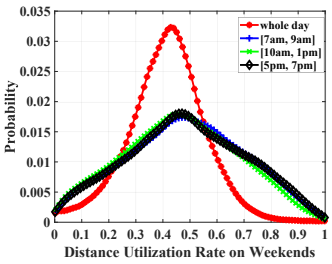

Fig. 3. The distribution of distance utilization rate on weekends.

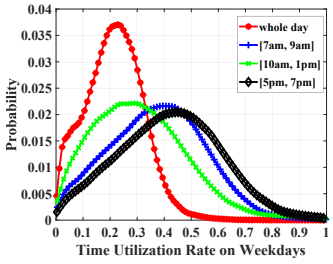

Fig. 4. The distribution of time utilization rate on weekdays.

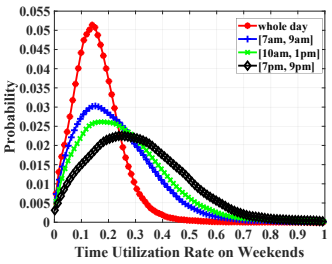

Fig. 5. The distribution of time utilization rate on weekends.

are later fulfilled, and one fulfilled order means one passenger. We first divide the map of Beijing into rectangular cells. The map of Beijing chosen in our paper is a rectangle, ranging from 116.1 to 116.8 (east) in longitude, and from 39.7 to 40.2 (north) in latitude, as regions outside this area see much fewer trips. Each cell is 0.02 longitude by 0.02 latitude, and in total there are 875 cells across our area. For a given timeslot (e.g., the morning rush hours [7am, 9am]) on a particular day-of-week (e.g., on Mondays), we define:

- *passenger density*: the total number of passengers starting their orders in one cell, during the given timeslot, on the particular day-of-week;
- *average passenger density*: the average of passenger density across all cells, during the given timeslot, on the particular day-of-week;
- *relative passenger density*: the passenger density of this cell divided by the average passenger density, for those cells with non-zero passenger density.

Hence, *passenger density* and *relative passenger density* are defined on each cell, whereas *average passenger density* is defined across the city. As an example, Fig. 6 shows the histogram of relative passenger density of cells with non-zero passenger density, on morning rush hours on Mondays. To make it clear for relative passenger density greater than 1, we also zoom in part of Fig. 6 inside it.

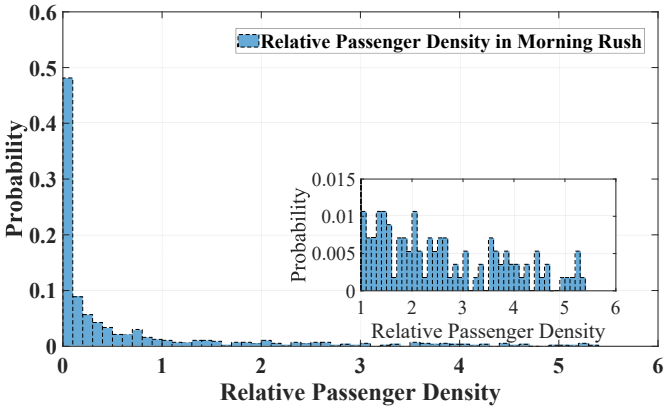

Fig. 6. The relative passenger density on morning rush hours on Mondays.

It is obvious in Fig. 6 that the distribution of relative passenger density has a long tail:

- Most cells have a relative passenger density smaller than 0.1 – numerically, during morning rush hours on Mondays the exact percentage is about 48%. One should note that we already exclude cells with zero passenger density, e.g., those cells with parks, rivers, mountains or other forms of inaccessible terrain.
- In the meantime, there are still a non-negligible number of cells with higher relative density. In Fig. 6, about 23% cells have a relative passenger density greater than 1, and some of them even have a relative density greater than 5.

In other words, the passenger density has an unbalanced distribution. This requires drivers to carefully choose seeking locations to avoid cells without enough potential passengers. Similarly, it is also necessary for seeking route recommendation to consider passenger density.

4.3 Driving Pattern

Our discussions on driving pattern contain two parts. The first part is on the number of drivers' visits to different cells, and the second is on driving traces characterization.

The number of drivers' visits to city cells reflects not only the distribution of the supply of RoD cars, but also drivers' preferences to choose different seeking locations or regions. Similar to 4.2, we divide the city map into cells, and count the number of visits to each cell. Fig. 7 shows the number of visits to city cells in Beijing, and the darker the cell, the more visits there are.

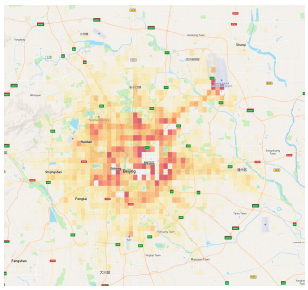

Fig. 7. The number of drivers' visits to city cells in Beijing.

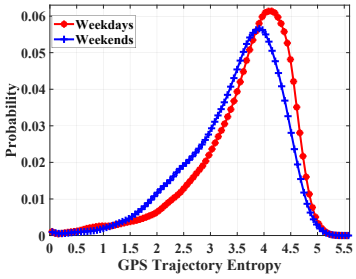

Fig. 8. The distribution of driver entropies.

It is intuitive to see from Fig. 7 that the supply of cars, or drivers' preferences to seek in different locations, is also highly unbalanced. Previous studies on seeking route recommendation in taxi service reveal the fact that only a small fraction of drivers can effectively plan their driving routes in order to earn more, and that's where seeking route

recommendation can help. Actually, compared to the city map and city planning, we observe that drivers tend to seek in crowded regions or major functional areas such as business (e.g., CBD), residential (e.g., some big living communities) or transportation areas (e.g., airports and railway stations). This observation may not be clearly emphasized in Fig. 7 due to the limited space.

The inspiration from Fig. 7 is that drivers do not take different locations or regions as equals. It is thus necessary to weigh them differently. In our study, we do not go into city-specific details such as the distribution and characteristics of city functional areas, planned or spontaneous events, the distribution of traffic in rush or non-rush hours, etc; instead, we try to characterize a location based on features relevant to RoD service, taxi service and public transportation services. For example, the supply of RoD cars or taxis is an indication of a location's popularity. Feature extraction and explanation will be covered later in Section 5.

The second part of our discussions on driving pattern is on driving traces characterization. Drivers' driving traces may be influenced by their various driving habits – some like to focus on smaller regions they are familiar with, while some tend to wander around a much larger region looking for passengers in a more random fashion. Another source of influences is temporal features. For example, during weekdays, the large number of orders brings drivers to more locations, increasing their driving traces' diversity.

To characterize a driver's driving traces during one day, we adopt the definition of a 2-dimensional entropy of the driver's GPS trajectories from [33]. For the city cell with horizontal index i and vertical index j, we use p_{ij} to denote the empirical probability of the driver passing this cell during the day. Based on p_{ij}, some entropy measures can be defined:

$$H(E) = -\Sigma_{i=1}(\Sigma_{j=1}p_{ij})ln(\Sigma_{j=1}p_{ij}), \quad (2)$$

$$H(P/E) = -\Sigma_{i=1}[\Sigma_{j=1}p_{ij}ln(p_{ij}/p_i)]. \quad (3)$$

In (3), p_i is the sum of p_{ij} over all js, and $H(P/E)$ represents the weighted average of entropy of GPS traces for rows. In (2), $H(E)$ represents the entropy of the sums of columns. Then, the 2-dimensional entropy $H(E \cdot P)$ is defined as:

$$H(E \cdot P) = H(E) + H(P/E). \quad (4)$$

Hence, the 2-dimensional entropy $H(E \cdot P)$ describes the degree of disorder of a driver's GPS trajectories over one day, and thus characterizes his/her driving pattern.

We calculate the 2-dimensional entropy for each driver on each day across our RoD GPS trajectories dataset, and in Fig. 8 we show the distribution of drivers' entropy on weekdays and weekends. We have the following observations:

- Different drivers have various driving patterns. The drivers' entropies are widely distributed from 0 to more than 5 either on weekdays or weekends, indicating that drivers always have their own perceptions of "how to seek to earn more". Some of these perceptions may not be good enough, and that is where seeking route recommendation works.
- Driving patterns are also influenced by temporal features. Fig. 8 justifies that drivers' entropies are higher on weekdays, and this agrees to our earlier conjecture in this section. In fact, drivers' entropies are different in smaller timeslots such as rush or non-rush hours, and these are now shown here due to the limited space.

4.4 Revenue Efficiency

We calculate the revenue efficiency in RoD service, including the efficiency in distance and in time. For each driver on a working day, the revenue efficiency in distance is the total revenue of the day divided by the total driving distance of that day (including seeking for and delivering passengers), and similarly, the revenue efficiency in time is the total revenue of the day divided by the total driving time.

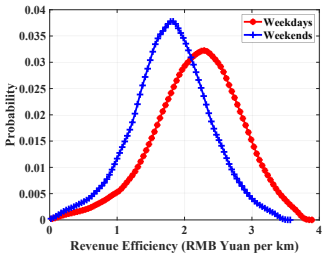

Fig. 9. The distribution of revenue efficiency in distance.

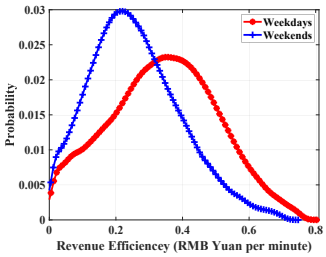

Fig. 10. The distribution of revenue efficiency in time.

Fig. 9 and Fig. 10 show the distribution of revenue efficiency in distance and in time, respectively. One thing to note is that as our data date back to early 2016, when RoD services have just come into business practices and the number of users was still climbing, the demand is not as high as we regularly see now, and so is the driver revenue. It is thus reasonable to see driver revenue efficiency not as high as was shown in previous taxi studies. The evaluation of our approach is based on the current data, and we will update results when we obtain newer datasets.

We observe from Fig. 9 and 10 that the mean revenue efficiency in distance is 2.179 and 1.796 RMB (Yuan) per km, and the mean revenue efficiency in time is 0.338 and 0.259 RMB (Yuan) per minute. Moreover, it is also clear that revenue efficiency, either in time or in distance, on weekdays is significantly higher than on weekends. Lastly, revenue efficiency varies greatly between drivers, indicating plenty of opportunities in seeking route recommendation.

5 The Force-directed Approach

We discuss the force-directed approach in this section, including problem formulation, the basic idea of the approach, determination of traffic charge, and route recommendation. Besides the approach itself, the emphasis is put on the above-mentioned three new factors that need to be addressed in RoD service – namely, matching-in-advance, dynamic pricing, and status of other transportation services.

5.1 Problem Formulation

As was typically done in previous taxi studies, in our study the road network is extracted from OpenStreetMap, and

consists of two different elements – intersections and directed road segments. We thus use $G = (N, E)$ to denote the road network, where $N = \{N_1, N_2, \cdots, N_r\}$ is a finite set of r intersections (or "nodes") and $E = \{E_1, E_2, \cdots, E_m\}$ is a finite set of m directed road segments (or directed "edges"). Each road segment has a starting and ending intersection.

Seeking route recommendation then tries to recommend a road segment to a vacant RoD car driver as soon as s/he arrives at an intersection, such that the driver may earn a higher profit.

Similar to previous studies in taxi's seeking route recommendation, our study is based on the **assumption** that "*in modelling and evaluating the approach, a relatively small number of drivers are assumed to adopt the recommended seeking routes, and their behaviours do not have a visible impact on the whole service.*" This assumption is for the convenience of evaluation, and should not be relevant to research significance. In order words, we extract features, calculate parameters and design the approach based on the whole datasets, and recommend seeking routes for a subset of drivers, and we assume that the behaviour of these drivers would not, in turn, influence the dynamic prices, the status of other transportation services, features extracted from multi-source urban datasets, etc. Hence it is not necessary to consider problems such as "whether we should re-predict the change of dynamic prices due to drivers' adoption of our recommendation?". If, without such assumption, then we may need to consider drivers' response, adoption rate of our recommendation, the prediction of dynamic prices at the city scale, etc., and these are left for future work.

Based on the discussions above, we have the following definition of seeking route recommendation problem:

Definition 5.1 (Seeking Route Recommendation)**.** Given the road network $G = (N, E)$ with intersections and road segments, the multi-source urban datasets, and a subset of RoD cars X, try to find the optimal seeking route for each car in X to increase earnings. Specifically, as soon as a vacant car driver reaches an intersection, recommend the next road segment for him/her to follow, until s/he picks up a passenger.

5.2 Basic Idea of the Force-directed Approach

The force-directed approach is built on the analogy between our topic to study and a certain scenario in physics. Specifically, the problem is modelled as a system of particles with forces acting between them, and that's also why such approach is called as "force-directed". The force-directed approach has been used in graph visualization, GPS trajectory map-matching, seeking route recommendation, etc.

In our study, the force-directed approach models the relationship between vacant cars and passengers as the relationship between positive and negative charges. In electrostatic field, Coulomb's law explains the electrostatic force between two point charges at certain distance in free space. Coulomb's law states that the magnitude of the electrostatic force between two charges is proportional to the amount of electrostatic charge on each of them, and is inversely proportional to the square of their distance:

$$\overrightarrow{F_{12}} = \frac{k \cdot Q_1 \cdot Q_2 \cdot (\overrightarrow{r_2} - \overrightarrow{r_1})}{|\overrightarrow{r_2} - \overrightarrow{r_1}|^3} \tag{5}$$

In (5), $\overrightarrow{F_{12}}$ is the electrostatic force acting on the charge Q_2 due to the charge Q_1, $\overrightarrow{r_2} - \overrightarrow{r_1}$ is the vector pointing from Q_1 to Q_2, k is a constant, and Q_1 and Q_2 are the amount of charge (both can be positive or negative). (5) also reveals another characteristic of electrostatic force – *likes repel and opposites attract* – considering Q_1 and Q_2 being positive and (or) negative charges.

These characteristics are very similar to the characteristics of the relationship between vacant cars and passengers:

- Vacant cars and passengers can both be regarded as points in free space.
- Potential passengers attract vacant cars, while vacant cars repel each other.
- For a vacant car, the influence from nearby vacant cars or passengers is attenuated quickly if they move further.

Such similarities inspire us to view vacant cars as positive point charges, and passengers as negative point charges. For a particular vacant car at an intersection, based on the distribution of vacant cars and passengers nearby, we can calculate the aggregated force acting on this particular vacant car due to the nearby cars and passengers. The aggregated force will then "drag" the vacant cars towards the right road segment. In other words, the right road segment is the outbound road segment closest to the direction of the aggregated force. For two objects q_1 and q_2, we define the force between them as:

$$\overrightarrow{F_{q_1 q_2}} = \frac{k' \cdot C_{q_1 q_2} \cdot \overrightarrow{e_{q_1,q_2}}}{|\overrightarrow{r_2} - \overrightarrow{r_1}|^2} \tag{6}$$

Similar to (5), in (6) $\overrightarrow{e_{q_1,q_2}}$ is the unit vector pointing from q_1 to q_2, $\overrightarrow{r_2} - \overrightarrow{r_1}$ is the vector pointing from q_1 to q_2, k' is a constant. Actually, the power to $|\overrightarrow{r_2} - \overrightarrow{r_1}|$ can be any positive value, but previous work [5], [31] indicate that 2 is already a good enough choice. $C_{q_1 q_2}$ is the counterpart of $Q_1 \cdot Q_2$ in (5), and we can call it as *traffic charge*. Traffic charge represents not only the existence of vacant cars or passengers, but also dynamic prices and the status of other transportation services. We will discuss traffic charge in more details in Section 5.3 and 5.4.

Considering the case of calculating the right road segment for a particular vacant car, we can thus always regard this vacant car in question as object q_1, and hence it is enough to only consider q_2 in (6). q_2 can be vacant cars or passengers very close to q_1, or a little bit further but still in the vicinity of q_1, or much further away. Similar to previous sections, we divide the city map into rectangular cells of 0.01 longitude by 0.01 latitude. In (6) the force decreases quickly when the distance increases, so it is safe to ignore those objects that are much further away from q_1. We then adopt the concept of extended region from [5], as shown in Fig. 11. In Fig. 11, the vacant car in question (i.e., q_1) is in cell R_0, and we call the 8 cells around R_0, denoted as $R_1, R_2, \cdots, R_8$, as the extended region. Then, we only consider the forces from all vacant cars and passengers in R_0 and the extended region – objects outside this area is omitted, as the corresponding forces become small enough.

The size of the extended region – with 3*3 cells as mentioned above – is chosen for the following reasons. Firstly, the interaction force is inversely proportional to the square

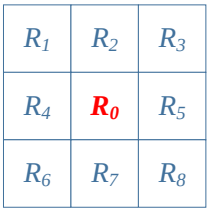

Fig. 11. The concept of extended region.

of distance between two objects, and hence the magnitude of the force decreases quickly when the distance increases. Even though objects in one cell far away from R_0 may change the magnitude and direction of the aggregated force, the road segment closest to the direction of aggregated force may remain unchanged. Choosing this size of the extended region is also supported by our daily experience – 2 to 3 km is already a large enough distance when one considers vacant cars and potential passenger demand nearby. Secondly, the amount of computation is reduced when the extended region is small: for example, if the extended region contain 4*4 cells, the amount of computation is almost doubled.

The aggregated force $\overrightarrow{F_{q_1}}$ is the vector sum of the forces from all vacant cars and passengers in R_0 and the extended region, on the vacant car q_1:

$$\overrightarrow{F_{q_1}} = \Sigma_{r=R_1}^{R_8} \frac{k' C_r \overrightarrow{e_{r,R_0}}}{d_{r,R_0}^2} + \Sigma_{q \in R_0} \frac{k' C_{R_0} \overrightarrow{e_{q,q_1}}}{d_{q,q_1}^2} \quad (7)$$

Note that the definition of traffic charge is on a cell instead of on a single car or passenger. In (7), the first term represents the aggregated force from the extended region. For each cell r in the extended region, C_r is the traffic charge of this cell, $\overrightarrow{e_{r,R_0}}$ is the unit vector pointing from the center of r to that of R_0, and d_{r,R_0} is distance between r and R_0. The second term is the aggregated force from objects (denoted by q) within R_0 – $\overrightarrow{e_{q,q_1}}$ and d_{q,q_1} are the unit vector and distance from any object q to q_1. The distance between two cells (e.g., d_{r,R_0}) is chosen, as verified by [5], as the shortest road distance between the two center intersections of these cells.

5.3 Regular and Recent Traffic Charge

It is common to consider both the regular traffic pattern and burst events together in seeking route recommendation. In our study, we calculate both the regular traffic charge and recent traffic charge of a cell. Before discussing them, we emphasize that the traffic charge is defined on each cell, and in each hour, based on the features extracted from multi-source urban data. For a particular cell, we calculate the traffic charge every hour. The choice of features and the relationship between the traffic charge and features will be discussed in Section 5.4.

The traffic charge is only related to the cell and the time, but when we want to calculate the aggregated force (as shown in (7)), we need to consider regular and recent traffic charge. "Regular" and "recent" are relative to the time referred to by the aggregated force. In the following discussion, we assume that our target is the aggregated force on a vacant car during hour $[t, t+1]$, on one day of day-of-week Y. We then define:

- Regular traffic charge is the characterization of regular traffic pattern in a cell. Specifically, regular traffic charge of a cell r, denoted by $C_{r,regular}$, refers to the average of traffic charges of r during hour $[t, t+1]$ on all days of day-of-week Y.
- Recent traffic charge is the characterization of burst events in a cell. Specifically, recent traffic charge of a cell r, $C_{r,recent}$, refers to the traffic charge C_r during hour $[t-1, t]$ on the very day of day-of-week Y.

For the first term of (7) (i.e., the forces from extended region, including cells R_1 to R_8), we calculate both the regular and recent traffic charge, and use a weighted sum as the final traffic charge C_r for cell r:

$$C_r = (1-\omega) C_{r,regular} + \omega C_{r,recent} \quad (8)$$

In (8), ω is the weight between regular and recent traffic charge. For the second term of (7) (i.e., the forces from objects within R_0), we include only the recent traffic charge:

$$C_{R_0} = C_{R_0,recent} \quad (9)$$

This is under the consideration that for the second term, we involve the locations of individual cars or passengers instead of the whole cell's collective properties. Hence recent traffic charge is much more important than the regular traffic charge.

5.4 Determination of Traffic Charge

In this section we discuss, in details, the calculation of traffic charge of a cell r, during hour $[t, t+1]$, on one particular day of day-of-week Y.

As mentioned previously, RoD service is a special service, but is still similar to traditional taxi service. Hence, to perform seeking route recommendation in RoD service, we should consider not only the data and features of RoD service, but also the status of other transportation services such as taxi, bus and metro. In addition, dynamic pricing, as one of the core features of RoD service, should also be presented in calculating the recommendation results. All these requirements are fulfilled in the design and determination of traffic charge.

5.4.1 *Features from RoD Service Data*

We extract the following features from our RoD service datasets, including the order data, GPS trajectories, and the event-log data. All these features can be calculated offline.

Density of passengers $PA_{t,r}$: the total number of passengers appearing in cell r during hour $[t, t+1]$ on this day.

Average density of passengers PA_t: among those cells across the city that have passengers appearing during hour $[t, t+1]$ on this day, PA_t is the average number of passengers.

Density of vacant cars $VC_{t,r}$: the total number of vacant cars appearing in cell r during hour $[t, t+1]$ on this day.

Density of all cars $AC_{t,r}$: the total number of cars appearing in this cell r during hour $[t, t+1]$ on this day.

Average dynamic price multiplier $DP_{t,r}$: based on the event-log data, we can calculate the average dynamic price multiplier from all *EstimateFee* events taking place in this cell r during hour $[t, t+1]$ on this day.

Maximum average dynamic price multiplier DP_t: among those cells with non-zero average dynamic price multiplier, DP_t is the maximum of $DP_{t,r}$ among all cells.

Average order revenue $OR_{t,r}$: based on all orders starting from cell r during hour $[t, t+1]$, the average order revenue is $OR_{t,r}$. Different from $DP_{t,r}$, $OR_{t,r}$ reflects not only the dynamic price multiplier, but also orders' distance and time.

Maximum average order revenue OR_t: among those cells with non-zero average order revenue, OR_t is the maximum of $OR_{t,r}$ among all cells.

Average order speed $OS_{t,r}$: similar to $OR_{t,r}$, but $OS_{t,r}$ represents the average of order speed among all orders.

Maximum average order speed OS_t: among those cells with non-zero average order speed, OS_t is the maximum of $OS_{t,r}$ among all cells.

Among these features, densities of passengers or cars are descriptions of supply and demand in RoD service; features of dynamic price multiplier are the direct representations of dynamic prices; features of order revenues reflect dynamic prices and order distances; and features of order speed describe, from another perspective, the ability of revenue-making in cell r, as previous work [1] suggests the relation between order speed and revenue efficiency.

5.4.2 Features from Taxi Service Data

We extract the following features from our taxi service GPS trajectories data. Similarly, all these features can be calculated offline. The goal of including these features is to describe the cell's popularity as well as traffic condition.

Taxi up count $UC_{t,r}$: the total number of taxi trips starting from cell r during hour $[t, t+1]$ on this day.

Maximum taxi up count UC_t: among those cells across the city that have non-zero taxi up counts, UC_t is the maximum of $UC_{t,r}$ among all cells.

Taxi down count $DC_{t,r}$: the total number of taxi trips ending in cell r during hour $[t, t+1]$ on this day.

Maximum taxi down count DC_t: among those cells across the city with non-zero taxi down counts, DC_t is the maximum of $DC_{t,r}$ among all cells.

Average speed of full taxi $FS_{t,r}$: the average speed of taxis with passengers on board that pass by cell r during hour $[t, t+1]$ on this day.

Maximum average speed of full taxi FS_t: among those cells with non-zero $FS_{t,r}$, FS_t is the maximum of $FS_{t,r}$ among all cells.

Density of vacant taxis $VT_{t,r}$: the total number of vacant taxis appearing in cell r during hour $[t, t+1]$ on this day.

Density of all taxis $AT_{t,r}$: the total number of taxis appearing in this cell r during hour $[t, t+1]$ on this day.

We choose the above features because:

- Some features describe a cell's popularity. For example, *taxi up count* and *taxi down count* reflect the number of people taking trips away and coming to the cell, respectively. Also, as verified in [1], the popularity of taxis in a cell does not mean competition to RoD service; instead, they are complementary to each other – the more popular a cell is to taxis, the more profitable it is for RoD drivers to seek in. Hence, we also include *density of vacant/all taxis*.
- Some features characterize a cell's traffic condition. For example, *average speed of full taxi* is a representation of general cars' speed in the cell, and it is a more accurate representation, as the number of taxis is larger than the number of RoD cars.

5.4.3 Features from Bus & Metro Data

We extract the following features to describe the status public transportation services in a cell. Note that these features are not time-dependent – they are only relevant to the cell r. These features are more about the existence of bus or metro stations and lines, than about the real-time operation status of such services.

Number of bus stations BS_r: the total number of bus stations in cell r.

Maximum number of bus stations BS_0: among those cells with non-zero BS_r, BS_0 is the maximum.

Number of bus lines BL_r: the total number of bus lines stopping by any bus station in cell r.

Maximum number of bus lines BL_0: among those cells with non-zero BL_r, BL_0 is the maximum.

Number of metro stations MS_r: the total number of metro stations in cell r.

Maximum number of metro stations MS_0: among those cells with non-zero MS_r, MS_0 is the maximum.

Number of metro lines ML_r: the total number of metro lines stopping by any metro station in cell r.

Maximum number of metro lines ML_0: among those cells with non-zero ML_r, ML_0 is the maximum.

As mentioned in Section 3.3 and verified in [1], these features extracted from public transportation services not only describe the popularity of a cell, but also characterize the possibilities of RoD cars to provide connecting services to passengers – this should attract drivers to seek in corresponding cells.

5.4.4 Integrating Features into Traffic Charge

Based on features extracted from multi-source urban datasets, we can combine them into traffic charge, and define the traffic charge, $C_{t,r}$ of cell r during hour $[t, t+1]$ on one day of day-of-week Y.

We denote the terms related to RoD service data, taxi service data, and bus & metro data by C_{RoD}, C_{taxi} and C_{public}, respectively. The traffic charge is then defined in a multiplicative form:

$$C_{t,r} = C_{RoD}^{1-2\alpha} \cdot C_{taxi}^{\alpha} \cdot C_{public}^{\alpha}, \tag{10}$$

In (10), we let the weights of these three terms add up to 1, and the taxi term and bus & metro term have equal weights. Additionally, as the traffic charge is calculated for RoD service, the RoD term should at least have a larger weight than other terms together, i.e., $1-2\alpha \geq 2\alpha$ or, rather, $\alpha \leq 1/4$. The idea that the RoD service data have much larger impacts on drivers' seeking for passenger has been verified in [1].

For the RoD term, we also define it in a multiplicative form,

$$\begin{aligned} C_{RoD} = &\frac{PA_{t,r}}{PA_t} \cdot (2 - \frac{VC_{t,r}}{AC_{t,r}}) \cdot (1 + \frac{DP_{t,r}}{DP_t}) \\ &\cdot (1 + \frac{OR_{t,r}}{OR_t}) \cdot (1 + \frac{OS_{t,r}}{OS_t}) \end{aligned} \tag{11}$$

In (11), we have the following concerns:

- The impact of passenger density on the traffic charge should be the strongest. Hence, we use the ratio between the passenger density of cell r and the average passenger density across the city to represent such impact. The ratio $\frac{PA_{t,r}}{PA_t}$ can be in the range $[0, +\infty)$.
- The ratio of vacant cars in a cell has a negative impact on attracting drivers to come. For such negative impact, we use the "2-minus" to quantify, and it is in the range $[1, 2]$.
- For other three multiplicative items, we first use the ratio between the value of cell r and the maximum value across the city to represent the relative value, and then use the "1-plus" form to quantify the positive impact on attracting drivers to come. The "1-plus" form also has a range $[1, 2]$.
- It is intuitive that the dynamic price multiplier, order revenue and order speed all have a positive impact on attracting drivers to seek in a cell. This is also verified in previous observations [1].

For the taxi term, similarly we have,

$$\begin{aligned} C_{taxi} =&(1 + \frac{UC_{t,r}}{UC_t}) \cdot (1 + \frac{DC_{t,r}}{DC_t}) \\ &\cdot (1 + \frac{FS_{t,r}}{FS_t}) \cdot (2 - \frac{VT_{t,r}}{AT_{t,r}}) \end{aligned} \tag{12}$$

And for the bus & metro term,

$$\begin{aligned} C_{public} =&(1 + \frac{BS_r}{BS_0}) \cdot (1 + \frac{BL_r}{BL_0}) \\ &\cdot (1 + \frac{MS_r}{MS_0}) \cdot (1 + \frac{ML_r}{ML_0}) \end{aligned} \tag{13}$$

The concerns are similar to those for (11), and the following concerns help us determine whether one feature has a positive or negative impact on a cell's attractiveness to drivers:

- As mentioned previously, RoD and taxi service are more like complementary than competitive to each other. Hence, the more attractive a cell to taxis, the more it is to RoD cars either.
- Similarly, the availability of public transportation services (bus & metro) makes it possible for drivers to provide connecting services to people.

Finally, combining (10) to (13), we can calculate the traffic charge $C_{t,r}$.

5.5 Miscellaneous Calculations

Besides the determination of traffic charge, as discussed in details in Section 5.4, some miscellaneous calculations should also be done before we are able to recommend the right road segments to drivers. We list them below.

The locations of vacant cars and passengers. In the second term of (7), to calculate the forces from objects within cell R_0, it is necessary to have the distance d_{q,q_1} between the vacant car in question and any other vacant car or passenger within cell R_0. To do this,

- Step 1: for each hour $[t, t+1]$ on one day of day-of-week Y, we first further divide this hour into six 10-minute timeslots.
- Step 2: then, for each 10-minute timeslot, we use the average longitude and latitude as the location of a vacant car in a timeslot. For a passenger's location, as passengers are assumed to be not moving during a request, it is not necessary to take any average.

The above two steps are about listing the locations of vacant cars and passengers, and can be done offline in advance. The calculation of the aggregated force, on the other hand, can only be performed online, with the movement of the target vacant taxi.

The average speed on each road segment. In simulating a driver's movement in seeking, it is necessary to estimate the time it takes from one intersection to the next intersection. To do that, we need an estimate of the speed on the corresponding road segment, during hour $[t, t+1]$ on one day of day-of-week Y. We gather all the GPS trajectories that pass by this road segment during hour $[t, t+1]$ on all days of day-of-week Y, and use the average speed of these trajectories as the estimate. The reason of gathering trajectories on *all* days of day-of-week Y is to have enough data to avoid any possible inaccuracies.

If, for some target road segment, no trajectories pass it by, we then calculate the average speeds of all road segments that lie in the same cell with the target road segment, and use the average among these average speeds as the estimate. This is very rare in our data – about 0.02% cases (road segments during different time periods) need to be processed in this way.

The key information of RoD orders. Based on the order dataset, we obtain the following key information of RoD orders: the starting and ending intersections and time, the order's total distance, and the estimated order's fare. As soon as a seeking driver picks up a passenger, s/he jumps from the starting intersection to the order's ending intersection, with a driving distance and making a revenue equal to the order's fare. The timeline also jumps from the order's starting time to the ending time.

The average speed on each road segment, as well as the order information, can all be processed offline.

5.6 Seeking Route Recommendation

This section wraps up all the previous sections, and provides a detailed step-by-step explanation for seeking route recommendation. We divide our seeking route recommendation into two parts – offline data preparation and online recommendation.

5.6.1 Offline Data Preparation

Step 1: List the locations of vacant cars and passengers: for every day across our RoD datasets, list the locations of vacant cars and passengers in every timeslot (i.e., in 10-minute long), as explained in Section 5.5.

Step 2: Calculate the average speed on each road segment: for days of each day-of-week, calculate the average speed on each road segment in every hour, as in Section 5.5.

Step 3: Extract key information of RoD orders: for every order in RoD order dataset, obtain the starting and ending intersections and time, the order's total distance, and the estimated order's fare, as in Section 5.5.

Step 4: Extract features from multi-source urban datasets: extract features from RoD service data, taxi service GPS trajectories data, and bus & metro distribution data, as in Section 5.4.1, 5.4.2 and 5.4.3, respectively.

Step 5: Integrating the above features into traffic charge: first, calculate the corresponding traffic charge term C_{RoD}, C_{taxi} and C_{public} based on (11), (12) and (13), respectively. Then, the traffic charge $C_{t,r}$ is calculated based on (10). Note that we calculate the traffic charge for every city cell, on every hour, and on every day.

Step 6: Calculating the distances between any two cells: according to (7), the distance between two cells will be used to determine the forces from objects within the same cell of the vacant car in question. As mentioned in Section 5.2, we use the shortest road distance between the two center intersections of these cells.

5.6.2 Online Recommendation

The online recommendation involves, for each driver of the subset of RoD cars X, recommending the next road segment the driver should take when his/her car is vacant and s/he arrives at an intersection. This recommendation is performed until the driver picks up one passenger.

As pointed out in Section 5.1, our recommendation is for a subset X of RoD cars, and in the meantime we assume other RoD cars and passengers will be the exactly same as they are in our dataset. In other words, other RoD cars keep the same trajectories as shown in RoD GPS trajectories data; all passengers request for rides from and to the exactly same locations, and at the exactly same time, as in RoD order data. Also, because the number of cars in X is relatively small, we can safely assume that the changes of trajectories and orders of these cars would not influence other RoD cars, passengers, taxi service, public transportation services, etc.

Matching-in-advance, as stated earlier in this paper, is another key difference between seeking route recommendation for RoD service and that for taxi service. When a vacant car and a passenger are not close enough to be within each other's sight, the service provider will attempt to match them if the passenger is the closest to the vacant car and the distance between them is less than a threshold (called as *matching distance*). This can increase the possibility of matching – even though the seeking RoD car does not pass directly along a passenger, they can be matched. In our study, we set the *matching distance* to be 1 km. On one side, 1 km is a setting corresponding to our daily experience using the service; on the other side, related works (e.g., [4]) try to identify a value between 0.5 to 1.5 km as the matching distance.

In fact, the matching distance may not be a fixed value. It may be a value dependent on some spatio-temporal features, set by the service provider. For example, in a distant region, when there are few vacant cars, it is beneficial to set a larger matching distance, so that a vacant car driver can be matched to a passenger earlier; similarly, in busy business region, the matching distance could be smaller. Matching distance could also be in a probabilistic form – instead of being set by the service provider, the matching distance could be mined from real data to answer "by what empirical probability a vacant car and a passenger could be matched when they are a particular distance apart". Such cases, however, are difficult to identify at this stage, as we don't have enough data to find out the exact location where a vacant car actually accepts an order. Hence, we use a fixed matching distance (e.g., 1 km) in our study, and more flexible settings are left for future work.

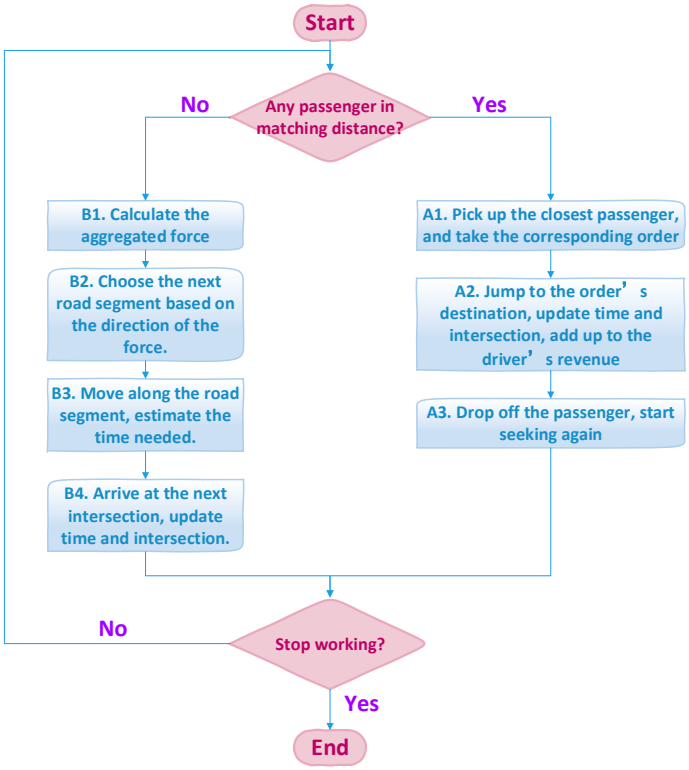

Fig. 12. The flowchart of online recommendation.

Fig. 12 illustrates the flowchart of our online recommendation, in the following we explain each of these steps with more details. The following explanations are based on a single driver (and the car) in the subset X.

Start. At the very beginning of the recommendation, the driver starts working at an initial intersection e_0 and time t_0. At this stage, there is not any passenger on board, so the driver also starts seeking.

"Any passenger in matching distance?" At every intersection, when the car is vacant, there are two possible cases the driver is faced with: a) if there is at least one passenger within the matching distance, then the closest passenger is matched with the driver; and b) otherwise, the driver needs a road segment recommendation so that s/he could go to the next intersection. For case a), then we go to step A1 to A3; otherwise, we go to step B1 to B4.

Step A1. If there is more than one passengers within matching-distance, then the closest passenger will be matched to the driver. The corresponding order is taken from our RoD order dataset, and obtain the order information.

Step A2. In this study, we care only the seeking process, and do not pay attention to the passenger delivery process. Hence, we directly jump to the drop-off location of the order:

- The driver is then at the ending intersection, with the status changed from "delivering" to "seeking".

- The timeline is moved to the ending time of the order.
- The driver's revenue is updated, adding up this order's fare.

Step A3. At this stage, the driver starts seeking again.

Step B1. We assume that the seeking driver is currently at intersection e_i at time t_i, and so the cell e_i is in is R_0, with the neighboring cells (the extended region) as R_1 to R_8. At this step, the aggregated force on the driver is calculated following (7).

Step B2. The absolute magnitude of the aggregated force on the driver is not important, but the direction is. We choose the next road segment for the driver as the outbound road segment closest to the direction of the aggregated force. If the aggregated force is zero, then we choose one random outbound road segment as the next road segment.

Step B3. The driver then move along the selected road segment. The time to move along the road segment, Δt, is estimated as the length of the segment divided by the average speed on the segment (calculated in Step 2 in Section 5.6.1).

Step B4. The driver arrives at the next intersection e_{i+1} through the selected road segment. The timeline moves from t_i to $t_{i+1} = t_i + \Delta t$.

"Stop working?" Either after step A3 or B4, the driver has already arrived at a new intersection with no passenger on board, and hence s/he is in the seeking process. This judgement tries to identify if it is time to stop working. For each driver on a particular day, we can obtain, from RoD GPS trajectories data, the time the driver stops working (maybe another driver takes over the shift, or maybe it is already late night, etc.), denoted by t_{stop}. If the current time is later then t_{stop}, then the driver should stop working, and the online recommendation for this driver stops; otherwise, the flow goes back to "*Any passenger in matching distance*" judgement.

6 Evaluations

We simulate our force-directed approach based on our multi-source urban datasets to verify its effectiveness. As mentioned previously in Section 3, our RoD and taxi datasets cover a time range from Nov. 2015 to March. 2016. We choose a Monday in Nov. 2015 to simulate our approach, and note that even though our simulation is based on the one chosen Monday, some features are calculated based on all Mondays (e.g., the average speeds on road segments), and also some features are independent of the day-of-week (e.g., the distribution of bus & metro services).

We choose 50 drivers who work on this Monday. Basically, the chosen drivers should be the most active ones:

- they work for longer time in the chosen Monday, and their GPS trajectories have few errors;
- in the RoD order dataset, each driver has the number of orders close to the average, and the orders are effective (i.e., not with a close-to-zero trip time or trip distance);
- they also work for most of other days.

The reason for choosing active drivers to simulate our approach is to avoid any possible inaccuracies due to problematic GPS trajectories, driver behavior, or orders.

TABLE 2
The comparison of average revenue efficiency of different approaches.

Approach	Rev. effi. in distance (Yuan per km)	Rev. effi. in time (Yuan per minute)
ground truth	2.188	0.343
force-directed approach	2.980	0.541
random	1.932	0.307
local hotspot	2.243	0.372

We also choose $\omega = 0.2$: in the calculation of the force from any cell in extended region, the weights for regular and recent traffic charge are 0.8 and 0.2, respectively. Note that based on the explanation of (9), in the calculation of the force within R_0, we only consider recent traffic charge.

In the most of the following evaluation, we also choose $\alpha = 0.2$: α is the power to the taxi and public transportation terms in (10). Hence, the powers to the RoD, taxi and public transportation term are 0.6, 0.2, 0.2, respectively. We also evaluate the effects of different αs below.

The simulation is performed in a step-by-step way using simple Python codes. For each chosen driver, the initial state is the first intersection and the corresponding time, obtained from ground truth data, at which the driver starts his one day's business. For passengers, we assume that every passenger appears in the intersection as in ground truth, and goes to the same destination using the same amount of time. Then, based on the flowchart of on-line recommendation (i.e., Fig. 12), the driver jumps between intersections, with driver revenue and timeline updated accordingly, as presented in step A1 to A3 and step B1 to B4 in Section 5.6.2. We then simulate this step-by-step process for each driver for 10 times.

6.1 Baselines

To evaluate the effectiveness of our approach, we would compare the average revenue efficiency of the chosen drivers to that of ground truth and two other baselines:

- **Random**: for a vacant car at an intersection, let the car randomly choose a connected road segment as the next road segment. This corresponds to the case that drivers have no recommendation or data support, and that they have no reliable experience.
- **Local hotspot**: for a vacant car at an intersection, choose the "local hotspot" cell, and follow the shortest route to the center intersection of the cell. The "local hotspot" cell is one of the 8 cells around the cell the car is currently in (i.e., the extended region) that has the highest passenger density. This corresponds to the case that drivers blindly chase for passengers.

6.2 Basic Results

We first compare the average revenue efficiency of the force-directed approach, the ground truth, and the two baselines, among the 50 chosen drivers. Tab. 2 shows the results.

We have the following observations on these results:

- The force-directed approach indeed has the best performance. Both the revenue efficiency in distance and in time are much higher than those of ground truth. Actually, the revenue efficiency in distance and in time are about the 89- and 90-percentile of revenue efficiencies in ground truth, respectively.
- The "random" baseline has the worst performance – even worse than the ground truth. This means that even in the ground truth, seeking based on drivers' personal experience is better than seeking randomly. Hence, the "random" baseline can be a borderline.
- The "local hotspot" baseline has an average revenue efficiency between ground truth and our approach. Basically, it tries to identify cells with higher passenger density as ideal seeking locations for drivers. Our approach, on the other hand, also takes into account many other factors such as dynamic prices, status of other transportation services, etc, and this proves to be helpful in improving drivers' revenue efficiency.

Additionally, we also calculate the 2-dimensional GPS trajectory entropy (see Section 4.3) of these drivers. It is shown that the average entropy of these drivers is 17% higher. This indicates, to some extent, that our approach also tries to distribute drivers into more cells, which, in turn, may help to increase service responsiveness and passengers' quality of experience.

6.3 Effects of Multi-source Urban Data

We evaluate the effects of introducing multi-source urban data into our seeking route recommendation in the following directions:

- the effects of α – this is an indication of the weights of different datasets;
- the effects of different datasets – we evaluate the revenue efficiency when only some of our datasets are involved;
- the effects of dynamic prices – we evaluate the revenue efficiency with and without features related to dynamic prices;

6.3.1 Effects of α

To evaluate the effects of α, we try three different αs in our approach. α is the power to the taxi term and the public transportation term in (10), and correspondingly, the power to the RoD term is $1-2\alpha$. We have mentioned in Section 5.4.4 that the RoD term should have a weight at least larger than other terms together, and that hence $\alpha \leq 1/4$. We choose three different αs: 0.2, $1/4$ and 0.3 – corresponding to the cases that the RoD term have a weight larger, equal, and less than other two terms together, respectively.

Tab. 3 shows the comparison of average revenue efficiency with different αs. It is clear that $\alpha = 0.2$ gives the highest revenue efficiencies. Considering the revenue efficiency in distance, $\alpha = 0.25$ gives a 4.9% smaller revenue efficiency compared to $\alpha = 0.2$ does, and $\alpha = 0.3$ gives a 9.4% smaller revenue efficiency compared to $\alpha = 0.25$ does. This verifies our earlier claim that the RoD term should have a weight at least larger than other terms together. Specifically, the average revenue efficiency drops faster when α grows larger than $1/4$. Earlier observations in [1] prove that including multi-source urban dataset indeed improves the accuracy in predicting drivers' revenue-making capability, but still the RoD data itself has the highest impact. Our observations regarding α agree to this.

TABLE 3
The comparison of average revenue efficiency with different αs.

α	Rev. effi. in distance (Yuan per km)	Rev. effi. in time (Yuan per minute)
0.2	2.980	0.541
0.25	2.834	0.508
0.3	2.569	0.465

TABLE 4
The comparison of average revenue efficiency using different datasets.

Datasets	Rev. effi. in distance (Yuan per km)	Rev. effi. in time (Yuan per minute)
all	2.980	0.541
RoD	2.636	0.481
RoD+taxi	2.732	0.495
RoD+public	2.658	0.486

It may be true that $\alpha = 0.2$ is not the optimal choice of α, but finding the exactly optimal α requires brute-force enumeration, meaning a lot of computation and trial-and-error. The above comparison already justifies the effects of α, and we thus consider it enough.

6.3.2 Effects of Different Datasets

In this section we try to answer if is necessary to introduce multi-source urban data into our approach. To do that, we vary the traffic charge calculation from (10) to only using RoD datasets,

$$C_{t,r} = C_{RoD}, \tag{14}$$

and, only using RoD and taxi datasets,

$$C_{t,r} = C_{RoD}^{0.6} \cdot C_{taxi}^{0.4}, \tag{15}$$

and, only using RoD and public transportation datasets,

$$C_{t,r} = C_{RoD}^{0.6} \cdot C_{public}^{0.4}. \tag{16}$$

In (15) and (16), the power to the RoD term is set to 0.6 so that it is comparable to our basic results in which the power to the RoD term is also $1 - 2\alpha = 0.6$.

Tab. 4 shows the comparison of average revenue efficiency using different datasets. In this table, "all", "RoD", "Rod+taxi" and "RoD+public" refer to using (10), (14), (15) and (16) to calculate traffic charge, respectively. We have the following observations,

- Using multi-source datasets indeed improves the average revenue efficiencies significantly. Comparing between "all" and "RoD", the improvement of the revenue efficiency in distance (and in time) is 13% (and 12.5%).
- The importance of taxi data is greater than that of public transportation data, shown by the higher revenue efficiencies of "RoD+taxi". Taxi service, compared to bus or metro, is more similar and related to

RoD service, and thus taxi data provide more clues for RoD drivers. This also corresponds to previous observations in [1] that features extracted from taxi data have larger weights than those extracted from public transportation data in determining drivers' revenue-making capabilities.

6.3.3 Effects of Dynamic Prices

Dynamic pricing, as mentioned previously, is one of the core new features of RoD service. There are four features extracted from RoD data (see Section 5.4.1) related to the dynamic prices – $DP_{t,r}$, DP_t, $OR_{t,r}$ and OR_t. Among them, $DP_{t,r}$ and DP_t are directly related to dynamic prices, whereas $OR_{t,r}$ and OR_t are indirectly related to dynamic prices.

For $OR_{t,r}$ and OR_t, the revenue of a single order is calculated as:

$$f = p * (15 + 2.8 * d) \tag{17}$$

In (17), the service provider sets the flag-fall to be 15 RMB Yuan ($\approx$ 2.18 USD), and each additional kilometre costs 2.8 Yuan ($\approx$ 0.41 USD). d is the order distance, and p is the corresponding dynamic price multiplier of this order.

To compare the revenue efficiencies with and without dynamic prices, we remove the influence of dynamic prices from the above four features. For $DP_{t,r}$ and DP_t, we set them to be 1, i.e., assuming every city cell has a price multiplier of 1. For $OR_{t,r}$ and OR_t, in calculating every order's revenue, we set p (i.e., the order's dynamic price multiplier) to 1.

Our simulation results show that without the influence of dynamic prices, the revenue efficiency in distance (and in time) is 2.675 (and 0.491). This indicates that:

- including features of dynamic prices in seeking route recommendation is of significant importance. The revenue efficiency in distance (and in time) is reduced by 10.2% (and 9.3%) without considering dynamic prices.
- combining results from Section 6.3.2 and 6.3.3, it is shown that both dynamic prices and multi-source urban data should be considered in recommending seeking route to drivers in RoD service. This is reasonable as they are the new features that make RoD service distinct from taxi service.

6.4 Results on Different Days

In Section 6.1 through 6.3, we present evaluation results on a chosen Monday in Nov. 2015 – actually it is the first Monday in this month. To justify the representativeness of our results, in this section we show the revenue efficiency in distance on other Mondays. We still choose Mondays as the representative day-of-week, and due to the limited space, we do not show results on other days-of-week here. Also, we only show the revenue efficiency in distance; the revenue efficiency in time shows similar patterns, and is thus omitted due to the limited space.

From Nov. 1, 2015 to Feb. 29, 2016, there are altogether 18 Mondays. During this time range, there is one major holiday season – the Spring Festival – this holiday is the most important holiday in China, and people get 7 days off (from Feb. 7 to Feb. 13, 2016). Before, during and after this holiday season, there are less people, less traffic and different trip and traffic patterns. In Fig. 13 we show the revenue efficiency in distance, from both the ground truth and the force-directed approach, among these 18 Mondays. The first Monday is Nov. 2, 2015, and the last is Feb. 29, 2016.

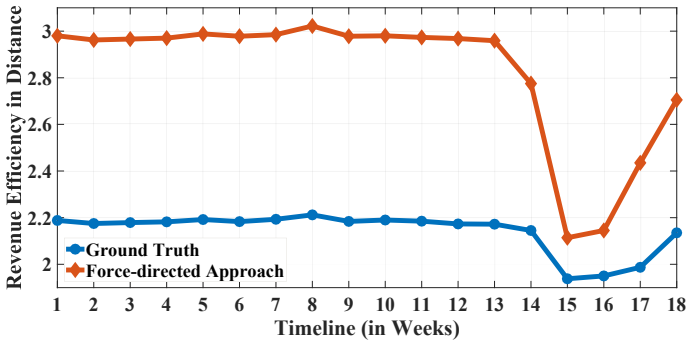

Fig. 13. The revenue efficiency in distance among different Mondays.

We have the following observations:

- The revenue efficiencies, from the ground truth and our force-directed approach, are relatively stable from the 1st to 13th Monday (Jan. 25, 2016). Firstly, the ground truth shows that on normal Mondays, drivers have stable driving habits and revenue efficiencies. Secondly, our force-directed approach has a stable performance throughout different days.
- The revenue efficiencies drop dramatically on the 15th and 16th Monday due to the holiday season. It is also clear that the revenue efficiencies of our approach drop more significantly than that of ground truth. The performance drop of our approach may be due to the fact that there is less passenger demand on the road, so there is little room for improvement.

7 Discussions

We present some discussions on relevant questions in this section.

Recommending the same route to different drivers. A typical question in route recommending studies is "*is the recommending scheme giving the same recommendation to different drivers*?". For example, a "local or global hotspot" scheme tries to recommend drivers to the cell with the highest passenger density; but if all nearby drivers follow such suggestion, this high-passenger-density cell would soon have a supply greater, or even much greater, than the passenger demand, making it impossible to satisfy these drivers. There are a number of heuristics to avoid recommending the same route to different drivers, e.g., generating several recommendations and giving a random one of them to each driver, setting a timer to measure the usability of a recommendation, distinguishing drivers by finer-granularity features, etc.

Our force-directed approach avoid giving the same recommendation to drivers by calculating the aggregated force from nearby vacant cars and drivers at finer-granularity. As

in (7), the aggregated force consists of two parts – the forces from extended region, and the forces from objects within the same cell. Hence,

- For two vacant car drivers in different cells, they receive different forces (both in magnitude and direction) from extended region, and thus the aggregated forces on two drivers in different cells are different;
- For two vacant car drivers in the same cell, the forces from objects (i.e., other vacant cars and passengers) within the cell are different for these two drivers, as the distances between them and these objects are not the same. Hence the aggregated forces on two drivers in the same cell are also different.

As a result, the directions of the aggregated forces on nearby drivers are different, and they thus receive different road segment recommendations.

Choosing a small number of drivers to simulate. This is another predicament in similar studies. If a majority of drivers follow the recommended road segment and change their seeking patterns, this would in turn reshape the service, change the distribution of supply and demand, vary the distribution of dynamic prices, and even influence the operation of taxi, bus or metro services. This effect may weaken the applicability of the original recommendation.

As a result, if a study aims to fully take into account the above impacts, it is necessary to:

- measure the adoption rate among drivers – how many drivers follow the recommendation?
- study and predict the supply and demand when these drivers adopt the recommendation;
- understand and predict the spatio-temporal changes in dynamic prices;
- study the interaction between multiple transportation services.

On one hand, such studies require the collection and analysis of business or sensitive data, and some even require certain forms of real experiment on passengers or drivers (e.g., AB test), which are, at this stage, difficult to accomplish; on the other hand, each of these studies involves a lot of effort and can be an independent research topic rather than a sub-problem in our study.

Instead, choosing a small number of drivers to simulate, as in our study, avoids all the above impacts on transportation services and enables us to focus on the core problem of seeking route recommendation. We thus study how to apply the force-directed approach into seeking route recommendation in RoD service, and how to incorporate the influence of dynamic prices and other transportation services. The evaluation of the approach under a large number of drivers is left for future work when we have the necessary datasets.

Computational efficiency of the approach. As the calculation for the next road segment happens at every intersection for every driver, it is necessary to analyze the computational efficiency of our force-directed approach. Our calculations could be divided into off-line and on-line calculations, and they are discussed separately.

We have pointed out in Section 5.6.1 that many preparation work could be done off-line, e.g., calculating the average speed, extracting RoD order information, extracting features from multi-source urban data, calculating traffic charges of all city cells, calculating the distances between any two cells, etc. Such off-line calculations could be performed periodically (e.g., daily, hourly, etc.) or do not need to be updated at all. We do not need to worry about the computational efficiency of these off-line work.

On-line calculations are discussed in Section 5.6.2, about the aggregated force. In calculating the aggregated force, the required traffic charges, as well as distances, are already available after off-line calculations. Only the following two tasks need to be calculated on-line:

- Performing the division between the traffic charge and the square of the distance, for the force from each cell in the extended region, or objects within cell R_0.
- Performing the summation among all forces.

Considering the fact that there are only eight neighboring cells in the extended region, and that the cell itself is only about 1 km^2 and could not accommodate many cars and passengers, the division and summation operations above would not take a long time. In our simulation, it takes less than 10 ms to calculate the aggregated force for each driver. Furthermore, on-line calculations could be carried out in a parallel fashion, and hence the computational efficiency could be guaranteed when calculating the force for a large number of drivers.

Directions for future work. Based on our datasets, approach, and evaluation, we have a number of possible directions for future work. They are not included in this study due to the limited space, the lack of data, the overwhelming computation, or because they are peripheral topics of less importance. Some of them are listed below:

- Study drivers' responses to the recommendation, and the corresponding adoption rate;
- Use numerical method to characterize the relationship between revenue efficiencies and the value of α, and find out the optimal α that leads to the highest revenue efficiencies;
- In the calculation of the RoD, taxi and bus & metro term (i.e., (11), (12) and (13)), experiment with the possibilities that each feature could have different importance. For example, (11) could be generalized to:

$$C_{RoD} = (\frac{PA_{t,r}}{PA_t})^{\gamma_1} \cdot (2 - \frac{VC_{t,r}}{AC_{t,r}})^{\gamma_2} \cdot (1 + \frac{DP_{t,r}}{DP_t})^{\gamma_3} \cdot (1 + \frac{OR_{t,r}}{OR_t})^{\gamma_4} \cdot (1 + \frac{OS_{t,r}}{OS_t})^{\gamma_5} \tag{18}$$

- With the generalization of the RoD, taxi and bus & metro term such as (18), it is also possible to study the force-directed approach with different parameters in various spatio-temporal settings, e.g., during morning rush hours, during non-rush hours, on weekdays, on weekends, around central business area, etc. Finding out the relationship between parameters and spatio-temporal settings may help to improve the performance and applicability of the approach.

8 Conclusion

In this paper, we study the seeking route recommendation problem in RoD service: recommending the right road segments to RoD drivers at every intersection. RoD service is distinct with taxi service by two key features – *mobile-app-based* and *dynamic pricing* – and they require the consideration of three more factors on the problem than in taxi service: matching-in-advance, dynamic pricing, and status of other transportation services.

We adopt the force-directed approach to tackle the problem. The force-directed approach models the relationship between vacant cars and passengers as that between positive and negative charges, and hence the aggregated force on a vacant car "drags" it into the right road segment in seeking for passengers. The modelling of the relationship is based on features extracted from multi-source urban data, including the RoD data, taxi data and bus & metro data.

Our evaluation results show that the approach not only outperforms two baselines on the revenue efficiency, but also tries to distribute drivers to more random cells. We also evaluate the effects of multi-source urban data and dynamic prices, and it is found that both of them help improve the recommendation significantly, and that it is necessary to weight carefully between features from different datasets.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (62002135, 61872050, 61602067, 61572059, 61825204), the Fundamental Research Funds for the Central Universities (2018cdqyjsj0024, 11619310), the Chongqing Basic and Frontier Research Program (cstc2018jcyjAX0551), the Science and Technology Project of Beijing (Z181100003518001), the Open Foundation of TUCSU (TUCSU-K-17002-01), and Beijing Outstanding Young Scientist Project (BJJWZYJH01201910003011). Chao Chen is the corresponding author.

References

[1] S. Guo, C. Chen, J. Wang, Y. Liu, K. Xu, Z. Yu, D. Zhang, and D. M. Chiu, "ROD-Revenue: Seeking strategies analysis and revenue prediction in ride-on-demand service using multi-source urban data," *IEEE Transactions on Mobile Computing*, vol. PP, no. 99, pp. 1–18, 2019.

[2] H.-w. Chang, Y.-c. Tai, and Y.-j. Hsu, "Context-aware taxi demand hotspots prediction," *International Journal of Business Intelligence and Data Mining*, vol. 5, no. 1, pp. 3–18, 2009.

[3] H. Rong, X. Zhou, C. Yang, Z. Shafiq, and A. Liu, "The rich and the poor: A markov decision process approach to optimizing taxi driver revenue efficiency," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '16. ACM, 2016, pp. 2329–2334.

[4] X. Yu, S. Gao, X. Hu, and H. Park, "A markov decision process approach to vacant taxi routing with e-hailing," *Transportation Research Part B: Methodological*, vol. 121, pp. 114–134, 2019.

[5] Y. Lai, Z. Lv, K.-C. Li, and M. Liao, "Urban traffic coulombs law: A new approach for taxi route recommendation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 3024–3037, 2018.

[6] S. Guo, C. Chen, J. Wang, Y. Liu, K. Xu, D. Zhang, and D. M. Chiu, "A simple but quantifiable approach to dynamic price prediction in ride-on-demand services leveraging multi-source urban data," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 112:1–112:24, 2018.

[7] A. Picchi, "Uber vs. Taxi: Which Is Cheaper?" 2016. [Online]. Available: http://bit.ly/2DMgrMc

[8] Y. M. Nie, "How can the taxi industry survive the tide of ridesourcing? evidence from shenzhen, china," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 242–256, 2017.

[9] S. Jiang, L. Chen, A. Mislove, and C. Wilson, "On ridesharing competition and accessibility: Evidence from uber, lyft, and taxi," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, ser. WWW'18, 2018, pp. 863–872.

[10] L. Rayle, D. Dai, N. Chan, R. Cervero, and S. Shaheen, "Just a better taxi? a survey-based comparison of taxis, transit, and ridesourcing services in san francisco," *Transport Policy*, vol. 45, pp. 168–178, 2016.

[11] V. Salnikov, R. Lambiotte, A. Noulas, and C. Mascolo, "Openstreetcab: exploiting taxi mobility patterns in new york city to reduce commuter costs," *arXiv preprint arXiv:1503.03021*, 2015.

[12] J. D. Hall, C. Palsson, and J. Price, "Is Uber a substitute or complement for public transit?" 2017. [Online]. Available: https://bit.ly/2K6Vs7L

[13] T. Berger, C. Chen, and C. B. Frey, "Drivers of disruption? estimating the uber effect," *European Economic Review*, vol. 110, pp. 197–210, 2018.

[14] J. Hall, C. Kendrick, and C. Nosko, "The effects of Uber's surge pricing: a case study," Oct. 2015. [Online]. Available: http://bit.ly/2kayk9O

[15] J. Gan, B. An, H. Wang, X. Sun, and Z. Shi, "Optimal pricing for improving efficiency of taxi systems." in *Proceedings of the 22th International Joint Conferences on Artificial Intelligence*, ser. IJCAI '13. AAAI, 2013, pp. 2811–2818.

[16] L. Rayle, S. Shaheen, N. Chan, D. Dai, and R. Cervero, "App-based, on-demand ride services: Comparing taxi and ridesourcing trips and user characteristics in San Francisco," 2014. [Online]. Available: http://bit.ly/2kVkahg

[17] L. Chen, A. Mislove, and C. Wilson, "Peeking beneath the hood of Uber," in *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, ser. IMC '15. New York, NY, USA: ACM, 2015, pp. 495–508.

[18] S. Guo, Y. Liu, K. Xu, and D. M. Chiu, "Understanding ride-on-demand service: Demand and dynamic pricing," in *Pervasive Computing and Communication Workshops (PerCom Workshops), 2017 IEEE International Conference on*. IEEE, 2017, pp. 509–514.

[19] S. Guo, C. Chen, Y. Liu, K. Xu, and D. M. Chiu, "Modelling passengers' reaction to dynamic prices in ride-on-demand services: A search for the best fare," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 136:1–136:23, 2018.

[20] M. K. Chen, "Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform," in *Proceedings of the 2016 ACM Conference on Economics and Computation*, ser. EC '16. New York, NY, USA: ACM, 2016, pp. 455–455.

[21] P. Cohen, R. Hahn, J. Hall, S. Levitt, and R. Metcalfe, "Using big data to estimate consumer surplus: The case of uber," 2016. [Online]. Available: http://bit.ly/2pqXiWo

[22] B. Li, D. Zhang, L. Sun, C. Chen, S. Li, G. Qi, and Q. Yang, "Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset," in *Pervasive Computing and Communication Workshops (PerCom Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 63–68.

[23] D. Zhang, L. Sun, B. Li, C. Chen, G. Pan, S. Li, and Z. Wu, "Understanding taxi service strategies from taxi gps traces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 123–135, 2015.

[24] N. Garg and S. Ranu, "Route recommendations for idle taxi drivers: Find me the shortest route to a customer!" in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 1425–1434.

[25] Y. Gao, D. Jiang, and Y. Xu, "Optimize taxi driving strategies based on reinforcement learning," *International Journal of Geographical Information Science*, vol. 32, no. 8, pp. 1677–1696, 2018.

[26] C.-M. Tseng, S. C.-K. Chau, and X. Liu, "Improving viability of electric taxis by taxi service strategy optimization: A big data study of new york city," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1–13, 2018.

[27] C. Yan, H. Zhu, N. Korolko, and D. Woodard, "Dynamic pricing and matching in ride-hailing platforms," *Naval Research Logistics (NRL)*, pp. 1–20, 2019.

[28] H. A. Chaudhari, J. W. Byers, and E. Terzi, "Putting data in the driver's seat: Optimizing earnings for on-demand ride-hailing," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 90–98.

[29] P. G. Paulin and J. P. Knight, "Force-directed scheduling in automatic data path synthesis," in *Proceedings of the 24th ACM/IEEE Design Automation Conference*, ser. DAC '87. New York, NY, USA: ACM, 1987, pp. 195–202.
[30] Y. Hu, "Efficient, high-quality force-directed graph drawing," *The Mathematica Journal*, vol. 10, no. 1, pp. 37–71, 2006.
[31] E. Rappos, S. Robert, and P. Cudré-Mauroux, "A force-directed approach for offline gps trajectory map matching," in *Proceedings of the 2018 ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '18. New York, NY, USA: ACM, 2018, pp. 319–328.
[32] AMap, "API of AMap Service," 2017. [Online]. Available: http://bit.ly/2n8YRbZ
[33] Q. Ma, H. Yang, H. Zhang, K. Xie, and Z. Wang, "Modeling and analysis of daily driving patterns of taxis in reshuffled ride-hailing service market," *Journal of Transportation Engineering, Part A: Systems*, vol. 145, no. 10, pp. 1–35, 2019.

Suiming Guo received his Ph.D. from the Chinese University of Hong Kong. He is currently an associate professor in College of Information Science and Technology, Jinan University, Guangzhou, China. His research interests include data mining, urban computing, pervasive computing and smart cities studies.

Chao Chen received his Ph.D. from UMPC (Paris 6) and Telecom SudParis. He is currently a full professor of computer science at Chongqing University, China. His research interests include pervasive computing, social network analysis, and mobile crowdsensing.

Jingyuan Wang received his Ph.D. from Tsinghua University. He is currently an associate professor at Beihang University. His general area of research is data mining and machine learning, with special interests in smart cities.

Yan Ding is a Ph.D. student of computer science in State University of New York, Binghamton, US. He obtained his B. Sc and M. Sc from Chongqing University, China. His research interests include urban driving, urban computing, autonomous intelligent robotics and artificial intelligence.

Yaxiao Liu received his Ph.D. from Tsinghua University. He is currently a senior manager in AWS China. His research interests are in AI based cloud architecture, spatio-temporal big data, stream computing and smart cities.

Ke Xu received his Ph.D. from Tsinghua University. He is currently a full professor in the Department of Computer Science and Technology, Tsinghua University. His research interests include next generation Internet, P2P systems, Internet of Things(IoT), network virtualization and optimization.

Zhiwen Yu received his Ph.D. from Northwestern Polytechnical University. He is a full professor with Northwestern Polytechnical University. His research interests include pervasive computing, context-aware systems, human-computer interaction, mobile social networks and personalization.

Daqing Zhang received his Ph.D. from University of Rome "La Sapienza" and University of L'Aquila. He is a full professor in Institut Mines-Telecom/Telecom SudPais. His research interests include large-scale data mining, urban computing, context-aware computing, and ambient assistive living.