IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 13 November 2013; revised 28 May 2014; accepted 13 August 2014.
Date of publication 8 September 2014; date of current version 30 October 2014.

Digital Object Identifier 10.1109/TETC.2014.2356493

A Carpooling Recommendation System for Taxicab Services

DESHENG ZHANG[1], (Student Member, IEEE), TIAN HE[1], (Senior Member, IEEE), YUNHUAI LIU[2], (Member, IEEE), SHAN LIN[3], (Member, IEEE), AND JOHN A. STANKOVIC[4], (Fellow, IEEE)

[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 USA
[2]Third Research Institute of Ministry of Public Security, Shanghai 201204, China
[3]Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11790 USA
[4]Department of Computer Science, University of Minnesota, Minneapolis, MN 55455 USA

CORRESPONDING AUTHOR: D. ZHANG (zhang@cs.umn.edu)

This work was supported in part by the National Science Foundation under Grant CNS-0845994, Grant CNS-0917097, Grant CNS-1239226, Grant CNS-1239108, Grant IIS-1231680, Grant CNS-1218718, and Grant CNS-1239483, and in part by UMN Thesis Travel Grant.

ABSTRACT Carpooling taxicab services hold the promise of providing additional transportation supply, especially in the extreme weather or rush hour when regular taxicab services are insufficient. Although many recommendation systems about regular taxicab services have been proposed recently, little research, if any, has been done to assist passengers to find a successful taxicab ride with carpooling. In this paper, we present the first systematic work to design a unified recommendation system for both the regular and carpooling services, called CallCab, based on a data-driven approach. In response to a passenger's real-time request, CallCab aims to recommend either: 1) a vacant taxicab for a regular service with no detour or 2) an occupied taxicab heading to the similar direction for a carpooling service with the minimum detour, yet without assuming any knowledge of destinations of passengers already in taxicabs. To analyze these unknown destinations of occupied taxicabs, CallCab generates and refines taxicab trip distributions based on GPS data sets and context information collected in the existing taxicab infrastructure. To improve CallCab's efficiency to process such a big data set, we augment the efficient MapReduce model with a Measure phase tailored for our CallCab. Finally, we design a reciprocal price mechanism to facilitate the taxicab carpooling implementation in the real world. We evaluate CallCab with a real-world data set of 14 000 taxicabs, and results show that compared with the ground truth, CallCab reduces 60% of the total mileage to deliver all passengers and 41% of passenger's waiting time. Our price mechanism reduces 23% of passengers' fares and increases 28% of drivers' profits simultaneously.

INDEX TERMS Taxicab network, recommendation system, carpooling.

I. INTRODUCTION

Among all transportation modes, taxicabs play a prominent role in residents' commutes in metropolitan areas, *e.g.*, New York City, over 100 companies operate 13, 000 taxicabs with daily demand of 660, 000 passengers [1]. In taxicab services, *availability* and *affordability* are two important criteria: the top comments from passengers about taxicabs are that taxicabs are not available when needed and fares are higher than expected. According to a survey [2], the average waiting time for a taxicab in the rush hour in big cities, *e.g.*, New York City, is more than 13 minutes, and the average taxicab fare is more than 6 times of a public transit fare, *e.g.*, a bus.

To improve both the availability and the affordability, a *carpooling service* is proposed in dense urban areas. In the carpooling service, a passenger can hail an occupied taxicab on streets or wait at a taxicab stand to carpool with the existing passengers. For the availability, a well-designed carpooling schedule groups related passengers into a single taxicab trip with the minimum detour mileage, thus delivering the same number of passengers with fewer taxicabs and lower mileage; for the affordability, a practical carpooling price mechanism reduces the passengers' fares, since the total fares and tolls are shared by all passengers on the same taxicab.

Different from regular taxicab services where any vacant taxicab can take a passenger in any direction, in a carpooling service, however, a new passenger has to find a *carpoolable* taxicab, which refers to an occupied taxicab with the existing passengers heading to the similar direction (no need to be the same destination) with this new passenger. But finding such a carpoolable taxicab is challenging because in the existing taxicab network infrastructure, even with real-time taxicab GPS tracking, a dispatching center cannot know future directions of the taxicabs that pick up passengers along streets, since the destinations of these passengers are normally unknown to the dispatching center. Unfortunately, almost all the existing taxicab recommendation systems [3]–[10] are focused on vacant taxicabs. Little work, if any, is focused on *how to find a carpoolable taxicab* for a passenger. Thus, we face a challenge to assist passengers to find carpoolable taxicabs in the existing infrastructures.

In this paper, we argue that a *data driven approach* is a promising solution to address such an issue. In the existing taxicab infrastructure of big cities, taxicabs' locations and status are uploaded to a dispatching center periodically in real time, forming a large GPS dataset. This dataset has a large volume (several TBs) and grows fast (1TB per year), and it can be used to draw *taxicab trip distributions* to analyze passengers' destinations (thus the future directions of occupied taxicabs) based on context information, *e.g.*, the route this taxicab has already passed. Thus, we employ these distributions to assist passengers to find carpoolable taxicabs.

In this work, we conduct the first effort to design a unified recommendation system called *CallCab* for both CArpooLing and reguLar taxiCAB services in dense urban areas such as New York City, Beijing, or Shenzhen, based on both GPS datasets and contexts collected in the existing infrastructure. Specifically, the key contributions of this paper are as follows:

- To the best of our knowledge, we conduct the first work to recommend either a vacant or a carpoolable taxicab for a passenger with a unified method, and provide a comprehensive study of how to analyze occupied taxicabs' routes without destinations of passengers.
- We design *CallCab*, which mines trip distributions from GPS datasets collected in the existing infrastructure. Then, according to these distributions conditioning on collected contexts for a particular new passenger, *CallCab* recommends either a vacant taxicab for a direct route (no detour distance), or a carpoolable taxicab for a carpooling route (small detour distance) in real time based on the similarities between directions of this new passenger and potential taxicabs.
- To quantify the similarity between directions, we design a novel metric called *Detour Ratio*, a ratio between a particular passenger's detour distance and the distance of the direct route. This detour ratio unifies recommendations for both regular services (with detour ratios equal to 0) and carpooling services (with detour ratios larger than 0). Thus, *CallCab* recommends the taxicab (either vacant or occupied) with the minimum detour ratio for a new passenger.
- To efficiently process GPS datasets for the detour ratio calculation, we present a generic *Map Reduce Measure* model by adding a new *Measure* operation to the *MapReduce*. This model provides three kinds of abstractions to hide details of data processing, and can be used for various applications.
- To facilitate the taxicab carpooling implementation in the real world, we present a simple yet effective reciprocal price mechanism to lower passengers' fares and simultaneously to improve drivers' profits, thus providing the economic incentives for carpooling.

We test *CallCab* on a real-world dataset consisting of GPS records from more than 14, 000 taxicabs in Shenzhen, the most crowded city in China. The results show that compared with the ground truth, *CallCab* reduces 60% of the total mileage and reduces 41% of the waiting time. Our price mechanism reduces 23% of the passengers' fares, and increases 28% of the drivers' profits, simultaneously.

The rest of the paper is organized as follows. Section II presents motivations. Section III shows our main idea. Section IV depicts the computing model. Sections V and VI describe the design and the price mechanism. Section VII evaluates our system. Section VIII introduces the related work, followed by the conclusion in Section IX.

II. MOTIVATION

In this section, based on a dataset collected in the infrastructure of Shenzhen, we present our motivation to show two inefficiencies of taxicab services, and to provide evidence for carpooling services to address these inefficiencies.

A. INFRASTRUCTURE DESCRIPTION

In the existing taxicab networks of large cities, *e.g.*, New York City, Beijing, and Shenzhen, taxicabs are equipped with GPS and communication devices, in addition to fare meters. To monitor the global status of all taxicabs, dispatching centers with cloud servers are also established in the most taxicab networks. Thus, as shown in Figure 1, the existing taxicab infrastructure typically consists of two parts: taxicabs in the frontend; dispatching centers in the backend.

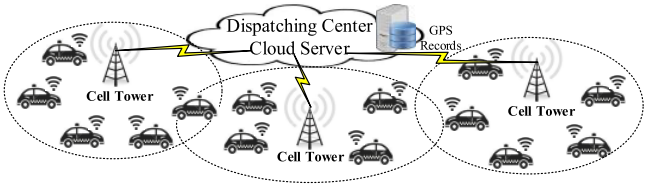


FIGURE 1. **Existing infrastructure.**

In such an infrastructure, (i) taxicabs record their physical status, *e.g.*, the current location, with GPS devices; (ii) taxicabs record their logical status with fare meters, *i.e.*, with passengers or not; (iii) physical and logical status

TABLE 1. Dataset summary.

Collection Period	6 Months
Collection Date	01/01-06/30
Number of Taxicabs	14,453
Number of Trips	98,472,628
Data Size	450 GB
Number of GPS records	3.9 Billion

is uploaded periodically to the dispatching centers via cell towers, by GPS records, which mainly consist of the following parameters: Plate Number; Date and Time; GPS Coordinates; Status Bit: with passengers or not when this record is uploaded. Table 1 gives statistics about such a GPS dataset of Shenzhen.

As in Table 1, this half-year dataset contains almost four billion GPS records. The partial aggregated data used in this work have been made for public access in the website of Transport Committee of Shenzhen Municipality [11] on a monthly basis. Such a large raw dataset has a very high resolution, which can be used to locate a particular taxicab at a fine granularity in terms of both time and space. But such a detailed GPS dataset has many records of no interest. Thus, we mine some semantics from this large fine-granular raw dataset to produce logical concepts, *i.e.*, trips, for our system design in Section IV.

Specifically, based on GPS records, we separate individual trips from the entire dataset by continuously observing the change of the status bit on the GPS records of the same taxicab. If a status bit turns to 1 from 0 in two consecutive records of a taxicab, then it indicates that this taxicab just *picked up* a passenger in the location indicated by the GPS coordinates, which is considered as an ***origin*** or a ***pickup*** location of a trip; if a status bit turns to 0 from 1, then it indicates that this taxicab just *droped off* a passenger at the location considered as a ***destination*** or a ***dropoff*** location of a trip. A GPS record set consisting of ***visited locations*** between an origin and its corresponding destination is considered as a ***trip***, which is the key for our design.

Figure 2 gives an example of a trip by mapping several GPS records on a map. A taxicab starts with no passengers at location L_1, and picks up a passenger between L_2 and L_3, and drops off this passenger between L_4 and L_5, and picks up a new passenger between L_6 and L_7, and finally leaves the map at L_8. Thus, a complete trip is given from L_3 to L_5.

B. INEFFICIENCIES OF TAXICAB SERVICES

We use the empirical data in Shenzhen to show two key inefficiencies, *i.e.*, low affordability and low availability, for current regular taxicab services.

1) LOW AFFORDABILITY OF TAXICAB SERVICES

Table 2 shows the taxicab fares (including surcharges) in USD for a 3KM trip in eight large cities in the world. We found that the taxicab fares are typically higher in developed countries than those in developing countries. In addition, according to

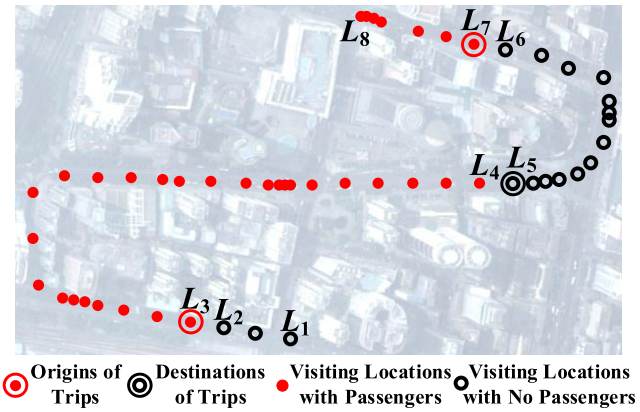


FIGURE 2. Taxicab trip.

TABLE 2. Taxicab fares.

Mexico City	$2.14
Beijing	$3.06
Shenzhen	$3.50
Hong Kong	$5.16
New York City	$12.00
Paris	$14.49
Tokyo	$15.66
Zurich	$26.97

a survey in New York City [2], the average taxicab fare is 5.8 times of public transit fare, *e.g.*, a bus. Further, according to the average paid taxicab fare in an hour basis from our Shenzhen dataset, the average fare of 22.9 CNY ($3.5) for Shenzhen taxicabs is 11 times of a bus fare, and is 11% of the average daily income [12].

2) LOW AVAILABILITY OF TAXICAB SERVICES

We investigate the availability of taxicab services in Figure 3, which shows taxicab occupancy ratios, and a high ratio indicates fewer empty taxicabs on streets, *i.e.*, low availability. It indicates more than 80% of taxicabs is occupied on average during the rush hour. Figure 4 plots time intervals between taxicab trips. A small interval indicates that a taxicab picks up a new passenger right after it drops off an old passenger, *i.e.*, low availability. It shows the average time interval in the rush hour is less than 3 mins.

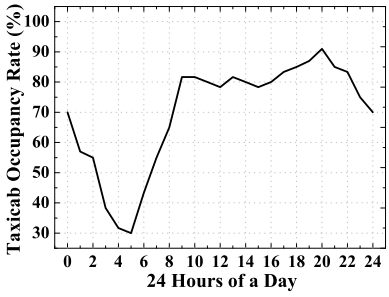


FIGURE 3. Occupancy rate.

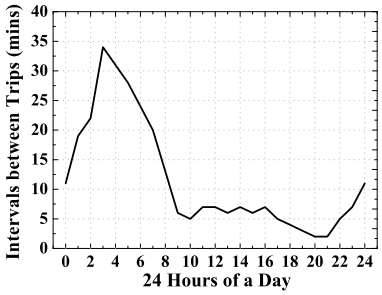


FIGURE 4. **Trip interval.**

C. OPPORTUNITIES FOR TAXICAB CARPOOLING

Taxicab carpooling services employ fewer taxicabs to deliver the same number of passengers, and lower the individual passenger fare by letting more passengers share the same taxicab. Thus, carpooling services are promising endeavors to improve both the affordability and availability of taxicab services.

In this subsection, to enable a practical service, we discuss three factors to show how likely carpooling services can be achieved in reality: (i) the distance between passengers' origins as well as the distance between passengers' destinations; (ii) the travel distances of shared routes between passengers; (iii) the passenger preference to the carpooling services. The benefit of carpooling services can be further unleashed, if we have more passengers who (i) start from close origins or end at close destinations, and (ii) share the long-distance common routes, and (iii) are willing to accept carpooling services.

1) CLOSE ORIGINS AND CLOSE DESTINATIONS

Based on the dataset, we show 200 consecutive trips to an airport in Figure 5 where most passengers came to the airport from the downtown and several hot spots.

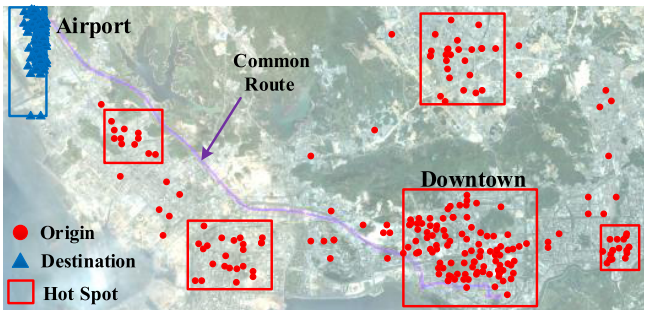


FIGURE 5. **Trips to an airport.**

In Figure 6, we show the cumulative distribution function (hereafter CDF) of distances between the origins of 1, 000 trips to the airport. Similarly, almost 50% of the trips have an origin closer than 1 KM to another origin, and almost 90% of trips have an origin closer than 5 KM to another origin. In Figure 7, we show the CDF of distances between destinations of 1, 000 trips from the airport. Almost 60% of trips has a destination closer than 1 KM to another destination,

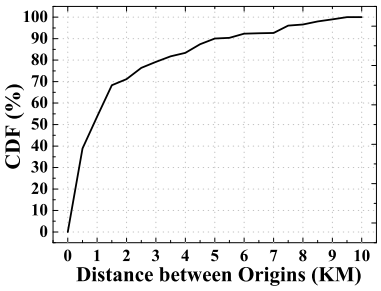


FIGURE 6. **Close origins.**

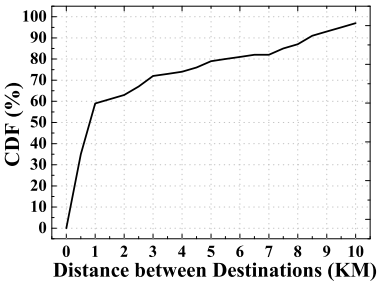


FIGURE 7. **Close destinations.**

and almost 80% of trips has a destination closer than 5 KM to another destination.

2) SHARED ROUTES

Based on the dataset, Figure 8 shows the CDF of distances of shared routes of 1, 000 trips to the airport. More than 90% of trips share at least 7.5 KM with another trip, and more than 50% of trips share at least 20 KM with another trip. Figure 9 shows the CDF of distances of shared routes of 1, 000 trips from the airport. Similarly, more than 90% of trips share at least 5 KM with another trip, and more than 50% of trips share at least 20 KM with another trip.

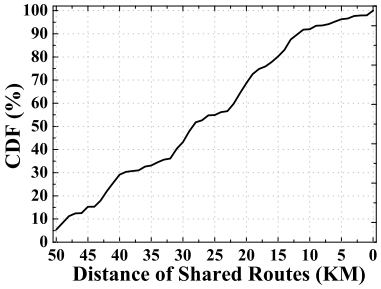


FIGURE 8. **Shared dist. to AP.**

3) PASSENGER PREFERENCES

Based on a taxicab service survey held at Beijing [13], we found that most passengers are willing to accept carpooling services. According to this survey, 75% of interviewees accept carpooling services; 73% of interviewees accept a

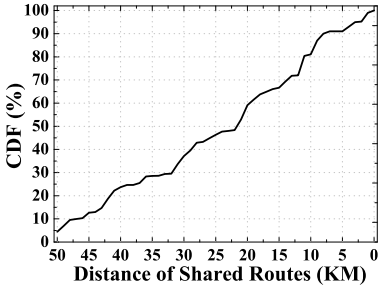


FIGURE 9. Shared dist. from AP.

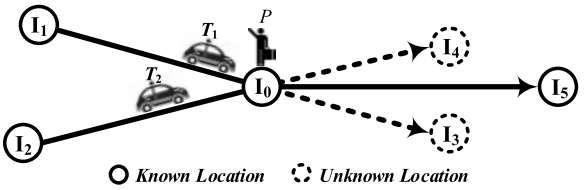


FIGURE 10. Taxicab operating scenario.

simple carpooling price mechanism where every passenger pays 60% of the regular service fare for the shared distance, leading to extra profits for drivers.

D. POSITIONS AND CHALLENGES OF CARPOOLING

Instead of completely replacing the traditional taxicab services, carpooling taxicab services aim to serve as a key supplement for the situations where traditional taxicab services are insufficient in the rush hour or the extreme weather, or where some passengers would like to take transportation cheaper than regular taxicabs yet more convenient than buses.

Given the mined semantics, the existing recommendation systems can easily locate and recommend a vacant taxicab to a new passenger based on their locations. But if no nearby vacant taxicab is available, they cannot recommend an occupied taxicab for a carpooling service, since destinations of the most existing passengers in taxicabs are unknown. But the large GPS dataset and contexts provide us an opportunity to predict the future directions of occupied taxicabs, and thus to locate carpoolable taxicabs. Note that though using the historical taxicab data presents a constraint, it offers valuable insights for future services since the taxicab trips are highly patterned due to regular commutes [14].

III. METHODOLOGY

Our *CallCab* aims for both the regular and carpooling services. Since regular services are commonly understood, we give an example of carpooling services, and then present the main idea of *CallCab*.

A. TAXICAB CARPOOLING SCENARIO

Figure 10 gives a passenger P waiting at origin I_0 and heading to destination I_5. Under the existing infrastructure, P provides a request with origin I_0 and destination I_5 to a recommendation system for a taxicab. Based on real-time GPS records, a recommendation system locates two nearby occupied taxicabs T_1 and T_2 that will pass P's origin I_0 soon.

To recommend T_1 or T_2 to P, a recommendation system has to analyze the actual traveling distance for P to be carpooled into T_1 or T_2. For example, if carpooled into T_1 at origin I_0, P first has to be "involuntarily" taken to a location I_3 (which is unknown destination of existing passengers on T_1) before to P's destination I_5, by the "First Come, First Served" policy. Thus, the actual traveling distance for P to be carpooled into T_1 is a ***carpool distance*** ($|\cdot|$) of a carpool route, *i.e.*, $|I_0 \Rightarrow I_3| + |I_3 \Rightarrow I_5|$, instead of a direct route with a ***direct distance*** of $|I_0 \Rightarrow I_5|$. The difference between the carpool distance and the direct distance leads to a ***detour distance*** of $(|I_0 \Rightarrow I_3| + |I_3 \Rightarrow I_5|) - |I_0 \Rightarrow I_5|$. With both the detour distance and the direct distance, we have a ***Detour Ratio*** $\rho_{T_1}^{P} = \frac{\text{detour distance}}{\text{direct distance}}$ to show the utility of P being carpooled into T_1.

Note that though the "First Come, First Served" policy is mostly adopted, it may not be the best choice. For example, if P's destination I_5 is on the path from I_0 to I_3, we can ask T_1 to pick up P at I_0 and then to drop off P at I_5 during the process of T_1 delivering the existing onboard passenger from I_0 to I_3, *i.e.*, serving P first. There is no additional detour for the existing passenger of T_1, since both I_0 and I_5 are on the route from I_1 to I_3. As a result, when calculating the carpool distance, if the destination of a carpooling passenger P is on the path of the existing taxicab service, then the carpool distance is equal to the direct distance, since there is no detour to deliver the carpooling passenger.

Different occupied taxicabs passing I_0 have different destinations, leading to different detour ratios for P to carpool. The optimal strategy is usually to select the taxicab with the minimum detour ratio.

However, only the origins of passengers on T_1 or T_2 (*i.e.*, I_1 or I_2) are known for the recommendation system, and their destinations (*i.e.*, I_3 or I_4) are mostly unknown in the existing infrastructure. Thus, the existing recommendation system cannot calculate detour ratios, thus failing to recommend a taxicab with a smaller detour ratio to P.

But in the existing infrastructure, although destinations are unknown during trips, the destinations are stored in terms of GPS records, after passengers are dropped off. These historical destinations and the collected real-time contexts are used to analyze unknown destinations of existing passengers in taxicabs, and thus to analyze detour ratios for new passengers to carpool with the existing passengers.

B. RECOMMENDATION PROCEDURES

The overview of recommendation procedures in *CallCab* is shown in Figure 11. First, we continuously maintain the trip distributions offline. Second, when a carpooling passenger request arrives in the real time, we calculate the detour ratio for every nearby taxicab based on the passenger request and the real-time contexts. Third, we recommend the taxicab with the minimum detour ratio.

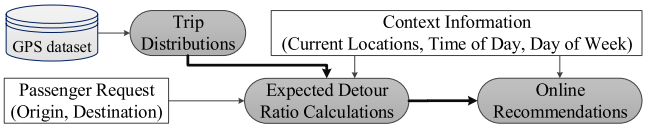


FIGURE 11. Overview of recommendation.

1) TRIP DISTRIBUTIONS

In *CallCab*, GPS records for all taxicabs are stored as a big dataset. By GPS records, destinations and corresponding origins comprise numerous trips, used to construct *trip distributions*. Such distributions generate destinations of trips that start at a particular origin and pass a particular location.

2) DETOUR RATIO CALCULATIONS

Upon receiving a *request* from a passenger P, *CallCab* uses trip distributions to calculate an *Expected Detour Ratio* $\rho^P_{T_1}$ for P to carpool with a nearby taxicab T_1, by a basic and an advanced design. In both the basic and advanced design, *CallCab* (i) calculates a potential destination set DS_{T_1} for T_1 (DS_{T_1} includes all destinations previously associated to the origin of existing passengers on T_1; this origin is obtained by T_1's last pickup locations), and then (ii) reduces the size of DS_{T_1} by contexts, and (iii) assigns probabilities for all destinations in reduced DS_{T_1} to calculate a weighted average $\rho^P_{T_1}$. The key differences between the basic and advanced designs are (i) how to reduce the size of DS_{T_1}, and (ii) how to assign the probabilities for the destinations in DS_{T_1}, as shown by Figure 12.

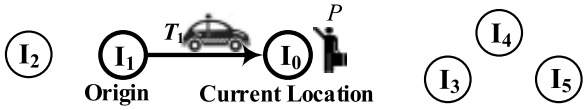


FIGURE 12. Detour ratio calculations.

a: BASIC DESIGN

(i) Based on trip distributions and T_1's last pickup location I_1, *CallCab* calculates $DS_{T_1} = \{I_2, I_3, I_4, I_5\}$. (ii) Assuming drivers use the shortest trips to deliver all passengers, we eliminate some destinations in DS_{T_1}, according to T_1's current location I_0. *CallCab* first obtains shortest paths between I_1 to all destinations in DS_{T_1} from the dataset. Then, a destination I_i is eliminated from DS_{T_1}, if the shortest path from I_1 to I_i does not include I_0. For example, *CallCab* eliminates I_2 from DS_{T_1}, since the shortest path from I_1 to I_2 does not include I_0 in the normal situation; *i.e.*, a normal trip starting at I_1 and passing I_0 is not the shortest trip from I_1 to I_2, so I_2 is not a potential destination for a trip starting at I_1 and passing I_0. (iii) By assigning equal probabilities (*i.e.*, 33%) for the remaining I_3, I_4 and I_5 in DS_{T_1}, *CallCab* calculates a weighted average $\rho^P_{T_1}$ by their locations.

b: ADVANCED DESIGN

The advanced design is built upon the basic design. But in the advanced design, (i) based on richer contexts, *CallCab* further reduces the size of DS_{T_1} obtained in the basic design, *e.g.*, *CallCab* can eliminate I_5 from DS_{T_1}, if I_5 has never been a destination for a trip at the current time of day and day of week. (ii) Instead of assigning equal probabilities for the remaining I_3 and I_4 as in the basic design, *CallCab* assigns probabilities to I_3 and I_4 based on their frequencies in the distributions to more accurately calculate $\rho^P_{T_1}$ in the advanced design, *e.g.*, if among six trips starting from I_1 in the distribution, four of them have I_3 as their destinations, while others have I_4 as their destinations, then *CallCab* assigns $\Pr(I_3) = \frac{4}{6}$ and $\Pr(I_4) = \frac{2}{6}$ to calculate a weighted average $\rho^P_{T_1}$. Note that if T_1's destination is known as one of the richer contexts (*e.g.*, reservation based pickups) to the dispatching center, then we reduce the destination set DS_{T_1} to only one destination with 100% probability.

To summarize, the basic design conditions trip distributions on only limited *contexts*, *e.g.*, origin and current locations of taxicabs, while the advanced design further considers the richer *contexts*, *e.g.*, popularity of destinations and time of day. Thus, the basic design is suitable for the scenario with the limited contexts, while the advanced design is suitable for the scenario with the richer contexts.

3) ONLINE RECOMMENDATION

With the detour ratio for every taxicab within a given recommendation radius, *CallCab* recommends the taxicab with the minimum expected detour ratio (either a vacant taxicab with no detour or an occupied taxicab with a small detour) for this passenger. With updated contexts, *e.g.*, taxicab locations, this recommendation is constantly updated.

C. OPPORTUNITY FOR MAPRDUCE IN CALLCAB DESIGN

The key step in *CallCab* is how to obtain trip distributions based on a raw GPS dataset. However, the raw GPS dataset shows physical aspects of taxicabs, while our design is focused on logical concepts, *e.g.*, trips, not directly given in the raw datasets. Further, the raw GPS dataset typically has a large volume and interconnects multi-dimensional GPS records with high resolutions, so though detailed enough, much of the raw dataset is of no interest in our design. Thus, to tackle this big dataset regarding these features [15], we need to map this raw physical GPS dataset to a filtered and compressed logical dataset (*i.e.*, trips) for analyses. In this work, we are inspired by *MapReduce* model proposed to deal with such big datasets [16], and augment it by an additional *Measure* phase to present a generic model called *MapReduceMeasure*, which can be used independently from our design. In Sections IV and V, employing the recommendation system as a showcase, we show how to use our model to tackle a big dataset that is not in a format ready for analyses.

TABLE 3. Model operations.

Name	Input	Output	Note
$MapByIS$	Trip Dataset	Set of $[IS, TRIP]$ pairs	$TRIP$ is a trip including intersection IS
$MapByTD$	Trip Dataset	Set of $[TD, TRIP]$ pairs	$TRIP$ is a trip starting at Time of Day TD
$MapByDW$	Trip Dataset	Set of $[DW, TRIP]$ pairs	$TRIP$ is a trip starting at Day of Week DW
$ReduceByIS1$	I_1, Set of $[TRIP]$	Set of $[TRIP_1]$	$TRIP_1$ is one of trips with I_1 as first intersection
$ReduceByIS2$	I_2, Set of $[TRIP]$	Set of $[TRIP_2]$	$TRIP_2$ is one of trips with I_2 as middle intersection
$ReduceByTD$	TD, Set of $[TRIP]$	Set of $[TRIP_3]$	$TRIP_3$ is one of trips starting at time of day as TD
$ReduceByDW$	DW, Set of $[TRIP]$	Set of $[TRIP_4]$	$TRIP_4$ is one of trips starting at day of week as DW
$MeasureB$	I_O^P, I_D^P, Set of $[TRIP]_{T_i}$	Detour Ratio $\rho_{T_i}^P$	Basic design for average expected detour ratio
$MeasureA$	I_O^P, I_D^P, Set of $[TRIP]_{T_i}$	Detour Ratio $\rho_{T_i}^P$	Advanced design for average expected detour ratio

IV. MAPREDUCEMEASURE MODEL

In this section, we first introduce the basic yet generic *MapReduceMeasure* model, and then present preliminaries, and define *Map*, *Reduce*, and *Measure* operations tailored for our application.

A. MAPREDUCEMEASURE INTRODUCTION

Our *MapReduceMeasure* model is mainly based on *MapReduce*, which is designed as a generic design and programming model for processing and generating large datasets. *MapReduce* has two key operations: *Map* and *Reduce*. A dataset user specifies a *Map* operation that takes *key*/*value* pairs as input to generate a set of intermediate *key*/*value* pairs, and a *Reduce* operation that takes all intermediate values associated with the same intermediate keys as inputs to generate a set of output values.

Even though sufficiently generic to perform many real world tasks, the two-phase *MapReduce* model is best at generating a set of values based on the same key. The impact of one key on the values generated by another key is difficult to evaluate in the current model. In this work, we present a third phase *Measure*, and it measures the impact of one key on the values generated by another key, and outputs a new value as a metric to show the impact. The generic types of our model are given as in Eq.(1).

$$\begin{aligned} \text{Map} &: (key_1, value_1) \rightarrow \mathbf{Set}[key_2, value_2]; \\ \text{Reduce} &: (key_2, \mathbf{Set}[value_2]) \rightarrow \mathbf{Set}[value_2]; \\ \text{Measure} &: (key_3, \mathbf{Set}[value_2]) \rightarrow value_3. \end{aligned} \tag{1}$$

Note that in this work we design the *Measure* operation as a separate operation in order to increase the parallelism on the operation levels, even though the *Measure* operation can be merged into the *Reduced* operation. The *Measure* operation is the designated point for multi-input operations. It obtains its input from several different MapReduce programs and makes use of the fact that the output in the *Reduced* operation is typically partitioned and sorted. Thus, every parallel instance of the *Measure* operation may select any data subset form all its input partitions to enable the flexible combination of data for parallel processing. In short, our three-operation design is flexibly performed in parallel to enable a fast processing of the taxicab GPS traces to meet the real-time requirement of carpooling applications.

B. PRELIMINARIES

To convert the raw GPS dataset into a format ready for our model, we present a mathematical concept, ***Carpool Graph***, and convert a set of raw GPS records into a logical trip record based on the carpool graph.

The basic unit for a passenger to carpool with others is a road segment. Thus, we define a carpool graph as a simple graph where vertices represent intersections and edges represent road segments between adjacent intersections. Figure 13 shows a carpool graph created by a given road map.

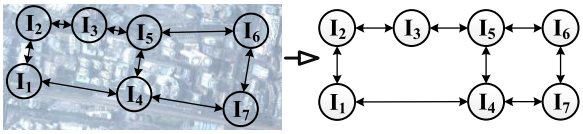


FIGURE 13. Carpool graph.

A set of raw GPS records belonging to a single logical trip is identified as shown in Section II-A by several key GPS records, indicating the origin, the visited locations, and the destination. Based on the set of GPS records belonging to a single logical trip, we create a trip record to capture the key information about this trip, *e.g.*, the origin, the destination, the intersections passed, the time and the date.

C. MODEL OPERATIONS

Via the trip record dataset obtained in the last subsection, we present three model operations.

1) MAP OPERATIONS

The *Map* operation is to reorganize the trip record dataset by generating pairs containing new keys (*e.g.*, a specific intersection) and the values associated to these new keys (*e.g.*, a trip includes this specific intersection). Three *Map* operations are presented in Table 3, *e.g.*, *MapByIS* generates a set of [*key*=intersection, *value*=trip] pairs, *e.g.*, [I_1, Trip#1] where Trip#1 includes intersection I_1.

2) REDUCE OPERATIONS

The *Reduce* operation is to reduce the size of sets of values associated to the same key. We present four *Reduce* operations as in Table 3, *e.g.*, *ReduceByIS*1 takes an intersection I_1 and

all trips associated to I_1 as input, and generates a smaller set of trips, including I_1 as their first intersection.

3) MEASURE OPERATION

We present *MeasureB* and *MeasureA* for the Basic and Advanced design as in Section III-B2, which both take the following as input: (i) a new passenger P's Origin I_O^P; (ii) P's Destination I_D^P; (iii) a trip set $[TRIP]_{T_i}$, indicating a particular trip distribution about a taxicab T_i. Both operations output a detour ratio $\rho_{T_i}^P$ for P to be carpooled into T_i as in Eq.(2):

$$\sum_{I_{D_i}^{T_i} \in DS_{T_i}} (\Pr(I_{D_i}^{T_i}) \cdot \frac{(|I_O^P \Rightarrow I_{D_i}^{T_i}| + |I_{D_i}^{T_i} \Rightarrow I_D^P|) - |I_O^P \Rightarrow I_D^P|}{|I_O^P \Rightarrow I_D^P|}) \tag{2}$$

where DS_{T_i} is the destination set of $[TRIP]_{T_i}$, and includes all distinct destinations of trips in $[TRIP]_{T_i}$. In *MeasureB* for the basic design, assuming every destination has an equal probability, $\Pr(I_{D_i}^{T_i}) = \frac{1}{|DS_{T_i}|}$ where $|DS_{T_i}|$ is the size of DS_{T_i}; whereas in *MeasureA* for the advanced design, assuming every destination has a different probability according to the times it appears in the trip set $[TRIP]_{T_i}$ (*i.e.*, frequency), $\Pr(I_{D_i}^{T_i}) = \frac{|I_{D_i}^{T_i}|}{|[TRIP]_{T_i}|}$ where $|I_{D_i}^{T_i}|$ is the number of $I_{D_i}^{T_i}$ appearing in $[TRIP]_{T_i}$ as a destination. Note that if T_i is a vacant taxicab, both operations return 0 as the ratio, since no detour is needed for a vacant taxicab. Further, if $I_{D_i}^{T_i}$ is on the route from I_O^P to I_D^P, both operations also return 0 as the ratio, because there is no detour to drop off the carpooling passenger first.

V. CALLCAB DESIGN

With the model presented in the last section, we present our design for a unified recommendation for both vacant and occupied taxicabs.

A. TRIP DISTRIBUTIONS

We envision a scenario where in the existing infrastructure, *CallCab* maintains trip distributions based on GPS records received by a dispatching center. By our *Map* operations, we generate different trip distributions for a particular intersection, the time of day, or the day of week. For example, a trip distribution for a particular intersection indicates how many taxicab trips pass such an intersection among the total taxicab trips.

B. EXPECTED DETOUR RATIO CALCULATIONS

When a passenger P wants to find a taxicab, P makes a request with the Origin I_O^P and the Destination I_D^P to *CallCab*. Based on I_O^P and real-time GPS records, *CallCab* collects the following contexts. **Time of Day** *TD* **and Day of Week** *DW***:** We consider both Time of Day (in terms of hourly windows) and Day of Week (in terms of SUN, MON, TUS, ..., and SAT). **Nearby Taxicab Set** T**:** As potential candidates, T is a set of taxicabs (either vacant or occupied) close and heading to P's origin I_O^P, within a recommendation radius R^T to I_O^P (*e.g.*, 100M). For every taxicab $T_i \in T$, based on real-time GPS records, *CallCab* further obtains (i) Last Pickup Location $I_O^{T_i}$ (*i.e.*, the Origin of existing passengers on T_i), and (ii) Current Location of T_i, which equals to I_O^P, since T_i is heading to P. Based on the above contexts, *CallCab* generates several particular distributions by model operations, which are used to calculate an expected detour ratio for P to be carpooled into a taxicab $T_i \in T$.

1) BASIC DESIGN

In the basic design, for a particular taxicab $T_i \in T$, *CallCab* generates two distributions and combines them together: (i) the trip distribution on intersection $I_O^{T_i}$ (the last pickup location of T_i); (ii) the trip distribution on intersection I_O^P (P's origin, *i.e.*, T_i's current location), by the following operations in Eq.(3).

$$\begin{aligned} \text{TripSet}(I_O^{T_i}) &= \mathit{ReduceByIS1}(I_O^{T_i}, \mathit{MapByIS}); \\ \text{TripSet}(I_O^P) &= \mathit{ReduceByIS2}(I_O^P, \mathit{MapByIS}); \\ \text{TripSet(B)} &= \text{TripSet}(I_O^{T_i}) \cap \text{TripSet}(I_O^P). \end{aligned} \tag{3}$$

According to the above TripSet(B), *CallCab* obtains the expected detour ratio $\rho_{T_i}^P$ as in Eq.(4).

$$\rho_{T_i}^P = \mathit{MeasureB}(I_O^P, I_D^P, \text{TripSet(B)}) \tag{4}$$

2) ADVANCED DESIGN

CallCab generates two more trip distributions and combines them with the TripSet(B) as in Eq.(5).

$$\begin{aligned} \text{TripSet}(TD) &= \mathit{ReduceByTD}(TD, \mathit{MapByTD}); \\ \text{TripSet}(DW) &= \mathit{ReduceByDW}(DW, \mathit{MapByDW}); \\ \text{TripSet(A)} &= \text{TripSet}(TD) \cap \text{TripSet}(DW) \cap \text{TripSet(B)}. \end{aligned} \tag{5}$$

According to TripSet(A), *CallCab* obtains the expected detour ratio $\rho_{T_i}^P$ as in Eq.(6).

$$\rho_{T_i}^P = \mathit{MeasureA}(I_O^P, I_D^P, \text{TripSet(A)}) \tag{6}$$

C. ONLINE RECOMMENDATION

Among all $\rho_{T_i}^P$ where $T_i \in T$, the taxicab T_{MIN} associated with the minimum ρ is the taxicab *CallCab* recommended to the passenger P. *CallCab* sorts all nearby taxicabs according to ρ, and if two or more taxicabs have the same ρ, the tie is broken by the distances to the passenger P. Further, *CallCab* marks all nearby taxicabs with ρ and plate numbers on a carpool graph sent back to the passenger's mobile device. We envision that a passenger follows this carpool graph to hail the recommended taxicab. During this process, some context information, *e.g.*, the passenger's location or the nearby taxicabs' current locations, will be changed, which may change detour ratios of the recommended taxicabs. Thus, *CallCab* updates this carpool graph, until the passenger is moving together with a taxicab, indicating this passenger has already found a ride.

VI. RECIPROCAL PRICE MECHANISM

The objective of our price mechanism on carpooling services is to employ a simple formula to lower the passenger fare and to improve the driver profit together, compared to non-carpooling situations. For a carpooling trip, the whole distance is dividend into the shared distances and the non-shared distances. For the non-shared distances, the passenger has to pay the fare according to the existing price mechanism, which is highly diverse in different cities. In this paper, we focus how to calculate fare for the shared distances. Figure 14 gives an example about sharing distance among multiple passengers.

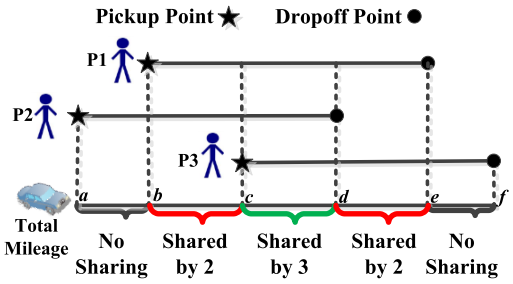


FIGURE 14. **Shared distances.**

Figure 14 shows 3 carpooled passengers who are picked up by the same taxicab at locations b, a, c, and dropped off at e, d, f, respectively. In this concurrent carpool trip, the mileage between a and b as well as between e and f is not shared by more than one passenger and is considered as regular taxicab services, and should be charged by the existing price mechanism. The rest of mileage is shared by either two or three passengers. For the shared mileage, every passenger should pay the carpooling fare, instead of the regular fare. For a passenger P_i sharing with other k passengers, the carpool fare CF is calculated based on the regular fare RF and a variable fare sharing ratio, *i.e*, $CF = RF \times r$ where $\frac{1}{1+k} \leq r \leq 1$. With $r \geq \frac{1}{1+k}$, we ensure that the total fare paid by all passengers together is at least equal to the regular fare for the benefit of drivers; whereas with $r \leq 1$, we ensure that every passenger pays less than the regular fare RF for the benefit of passengers. Thus, by this price mechanism, every passenger at most pays for the regular fare, and the driver at least collects the regular fare. A r bigger than $\frac{1}{1+k}$ and smaller than 1 leads a reciprocal situation where every passenger pays a carpool fare less than the regular fare, and the driver collects the aggregated carpool fare more than the regular fare.

Figure 15 gives the relationship between the total increased profit for the driver and fare sharing ratio r for passengers when 1 to 4 passengers are carpooled. In all carpool scenarios, the relationship between the total fare increased and r is linear. To encourage the carpooling service for both drivers and passengers, our objective is: "the more the carpooled passengers, the less fare every passenger pays, the more profit the driver has". Thus, we suggest three models for the price mechanism where the maximum passenger number is 4 as

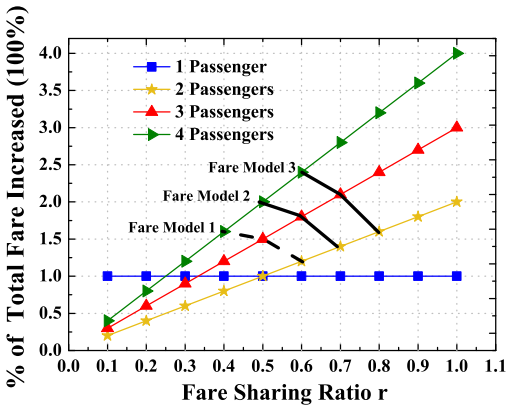


FIGURE 15. **Sharing ratio.**

TABLE 4. **Price mechanism.**

# of Passengers	Reduced Fare (%)	Increased Profit (%)
2	40%	$2 \times (1 - 40\%) - 1$=20%
3	50%	$3 \times (1 - 50\%) - 1$=50%
4	60%	$4 \times (1 - 60\%) - 1$=60%

in Figure 15. The mechanism 1 is given in Table 4, *e.g.*, if four passengers are carpooled, every passenger pays 40% of the regular fare, leading to 60% more fare as the profit for the driver. The mechanism 2 and 3 are similar to model 1, except increasing the starting fare sharing ratio, leading more benefits for the driver. In the more advanced design, this ratio r could change from time to time based on various factors, *i.e.*, the supply and demand relationship in taxicab networks.

VII. CALLCAB EVALUATION

We draw a sample with one week of GPS records from the dataset in Section II-A to test *CallCab*.

A. EVALUATION OVERVIEW

We compare two versions of *CallCab*, **Basic** and **Advanced**, against a **Heuristic** recommendation. Based on GPS datasets, we also obtain trip records which show the real passenger requests. Then, we use the requests that happened in the dataset of one day as the future requests to test *CallCab*. Based on a trip record such as [pickup time, origin, dropoff time, destination] in the dataset, we generate a passenger request [request time=pickup time, origin, destination]. According to a request, all systems first locate a nearby taxicab set T where taxicabs are within R^T radius to the origin, based on traces of taxicabs in the dataset for a particular day. If there are vacant taxicabs in T, all schemes recommend the closest vacant taxicab to passengers. Otherwise, (i) Heuristic recommends the closest taxicab in T to the passenger; (ii) Basic calculates the expected detour ratio for every taxicab in T based on the basic design in Section V-B1, and then recommends the taxicab with the minimum ratio; (iii) Advanced works similarly, except that it calculates the detour ratio based on the advanced design in Section V-B1.

We use **Actual Detour Ratio** as a key metric to show the efficiency, which is obtained by $\frac{\text{actual travel distance}-\text{direct distance}}{\text{direct distance}}$, and given a specific recommended taxicab, this metric is calculated by the method given in Section III-A.

We investigate **Percentage of Reduced Mileage**, which is used to evaluate how much the total mileage can be reduced by an efficient system recommending more suitable occupied taxicabs. It equals to $\frac{\text{M}-\text{m}}{\text{M}}$ where M is the total mileage used to deliver all passengers separately (*i.e.*, only regular services with vacant taxicabs), and m is the total mileage used to deliver all passengers with either vacant or occupied taxicabs recommended.

We justify carpooling services by showing **Percentage of Reduced Waiting Time** due to carpooling. With carpooling, a passenger can significantly reduce the waiting time to take a carpooled taxicab, instead of waiting for a vacant taxicab. But in the current dataset, the actual waiting time for a passenger is not given. However, the upper bound of the waiting time is determined by the time that two taxicabs pass the same pickup location. For example, if a GPS dataset shows that (i) when a vacant taxicab T_1 passes a location L at time τ, T_1 does not pick up any passenger, and (ii) when another vacant taxicab T_2 passes the same location L later at time $\tau + \Delta\tau$, T_2 picks up a passenger, then the upper bound of waiting time for this passenger is $\Delta\tau$. Assuming the actual waiting time is equally distributed from 0 to $\Delta\tau$, and then we obtain an expected waiting time $\Delta\tau_r$ for a regular service in the dataset. The waiting time $\Delta\tau_c$ for carpooling is decided by the time when the passenger starts to wait (obtained by $\tau+\Delta\tau-\Delta\tau_r$) and the time when the recommended occupied taxicab passes the passenger's location (obtained from the dataset). Based on the waiting time and expected trip time, we also investigate the **Total Travel Time** for passengers from the time when they start to wait to the time when they are dropped off. The expected trip time is obtained by average travel time for all historical trips traveling the same route.

Finally, we investigate **Reduced Fares** for passengers and **Increased Profits** for drivers by our reciprocal price mechanism.

We evaluate the performance for different hourly windows for weekdays and weekends, and at different radii R^T, which determine the size of the nearby taxicab set T. The default setting of R^T is 250M. For both weekdays and weekends, we use requests from an one-day dataset and test all systems with traces of taxicabs on other days. The average results are reported.

We maintain the trip distributions offline and update them on a daily basis. Thus, in the real-time mode, the running time for the recommendation is mostly depended on the number of nearby taxicabs. We process the data with a Hadoop cluster of 10 nodes (8 processors in each node), and recommend a taxicab among a fair number of nearby taxicabs under reasonable response time, e.g., we recommend a taxicab among 10 taxicabs within 3 seconds. We omit the related figures due to space limitations.

B. IMPACT OF DAY OF WEEK AND TIME OF DAY

Figure 16 plots the average actual detour ratio on five weekdays. During the rush hour of a weekday, *e.g.*, 7-10, the average actual detour ratios for all four schemes are higher than those of the non-rush hour, *e.g.*, 1-7. This is because there are many vacant taxicabs during the non-rush hour, whereas in the rush hour passengers have to use carpooling services, which leads to high actual detour ratios. But the Basic and Advanced solutions outperform Heuristic, which have a high average actual detour ratios during the rush hour, *i.e.*, 60% and 55%. Advanced outperforms Basic by 25% in the rush hour.

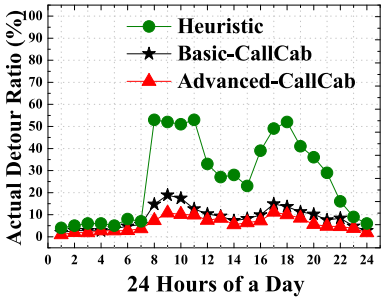


FIGURE 16. Detour in weekday.

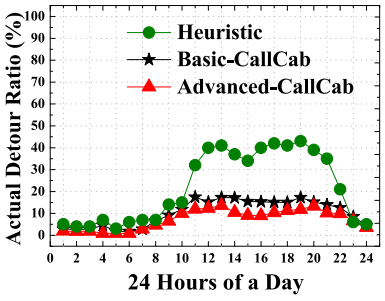


FIGURE 17. Detour in weekend.

Figure 17 gives the average actual detour ratios for two weekends. During the rush hour of a weekend, *e.g.*, 10-21, the average actual detour ratios for Basic and Advanced are much lower than others. This is because during the rush hour, Heuristic recommends more occupied taxicabs with long detours to passengers. But both versions of *CallCab* utilize the trip distribution to recommend occupied taxicabs with less expected ratios, so it leads to a lower ratio.

Figure 18 shows the percentage of reduced mileage for five weekdays. During the rush hour of a weekday, *e.g.*, 7-10, the percentage of reduced mileage is higher than that of the non-rush hour for all four schemes. This is because during the rush hour, there are more carpooling services than regular services, which leads to the reduction of the total mileage to deliver the same number of passengers. But Basic and Advanced outperform Heuristic during both the rush and non-rush hour, which shows the effectiveness of *CallCab*.

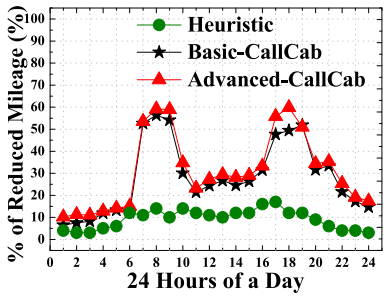


FIGURE 18. **Mileage in weekday.**

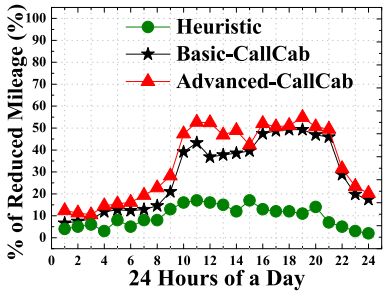


FIGURE 19. **Mileage in weekend.**

Figure 19 shows the percentage of reduced mileage for two weekends. Different from weekdays, for the weekend, the high percentages of reduced mileage are between 10-21 for both versions of *CallCab*. The performance on weekends is different than that on weekdays, since people take taxicabs at different times on weekdays and weekends. There is no significant high percentage of reduced mileage in certain time windows among 10-21 than others. The relative performance is similar as in Figure 18.

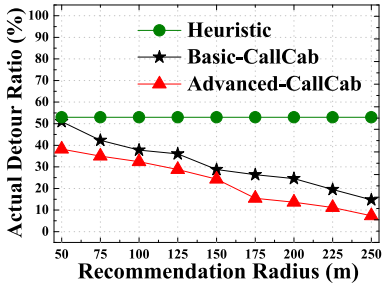


FIGURE 20. **Detour vs. radius.**

C. IMPACT OF RECOMMENDATION RADIUS

Figure 20 shows the effects of recommendation radii on the average actual detour ratio from 8-9 of a weekday. We increase the recommendation radius from 50 meters to 250 meters, which increases the size of potential taxicabs that can be recommended. Heuristic is not affected by such an increase, since it only recommends the closest taxicab. But with the increase of the radius, both Advanced and Basic

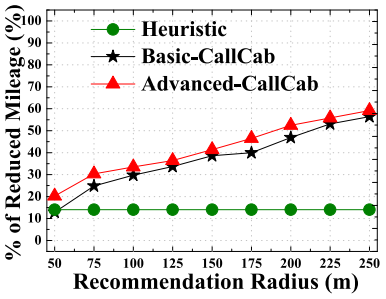


FIGURE 21. **Mileage vs. radius.**

have lower actual detour ratios, because a large recommendation radius gives them more taxicabs to select for a better recommendation.

Figure 21 shows effects of recommendation radii on percentage of reduced mileage from 8-9 of a weekday. With the increase of the radius from 50 to 250 meters, the performance of both versions of *CallCab* increases, while the others stay the same. But as the radius is close to 250M, the increase for *CallCab* slows down, since the radius is large enough to have the sufficient taxicabs for recommendations, and a larger radius does not help.

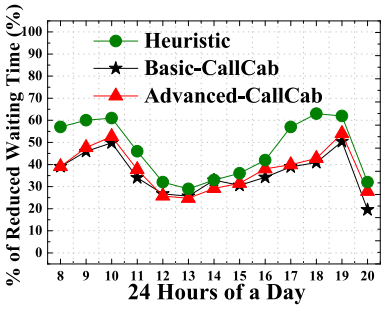


FIGURE 22. **Waiting time in WD.**

D. WAITING TIME AND TOTAL TRAVEL TIME

In this subsection, we show the percentage of reduced waiting time due to carpooling in the weekday (WD) in Figure 22. Because the method we use to calculate waiting time is based on taxicabs passing locations of pickup events, we present the percentage of reduced waiting time from 8 to 20 of a weekday, due to the high densities of taxicabs and pickup events. During the rush hour, *e.g.*, from 8 to 10 A.M., all systems with carpooling services reduce the waiting time by as much as 41% on average. Heuristic outperforms the rest because it recommends the closest occupied taxicab for carpooling services, and other systems perform similarly to each other. We also show the total travel time by adding the expected trip time to the waiting time as in Figure 23. During the rush hour, all systems with carpooling services reduce the total travel time by as much as 28%, due to the prolonged waiting time in regular services. But in the non rush hour, because of the decreased waiting time, carpooling services

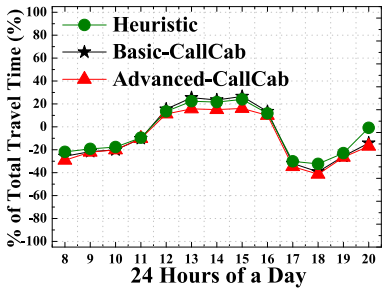


FIGURE 23. Total time in WD.

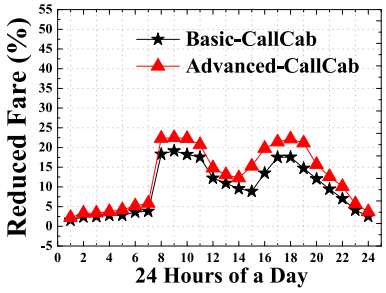


FIGURE 24. Reduced fares.

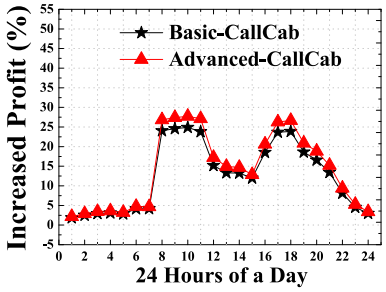


FIGURE 25. Increased profits.

increase the passenger total travel time by as much as 29% by taking detour to deliver other passengers.

E. REDUCED FARES AND INCREASED PROFITS

In this subsection, we evaluate the performance of *CallCab*'s reciprocal price mechanism. Based on the datasets, we have the ground truth for regular fares of individual passengers, and based on the recommended taxicabs, we have the fares for carpooling services. We use our price mechanism 1 as in Section VI. In Figures 24 and 25, we can see that in the rush hour of a day, the fare for every passenger decreases significantly by as much as 23% on average, and the profit for the drivers increases by as much as 28% on average. In the non-rush hour of a day, we still achieve both a 10% passenger fare reduction and a 9% driver profit improvement, which indicates the advantage of our reciprocal price mechanism.

VIII. RELATED WORK

Due to the increasing availability of GPS devices in taxicabs, taxicab GPS records have been employed by several systems to improve the efficiency of regular taxicab services, *e.g.*, discovering temporal and spatial causal interactions to provide timely and efficient services in certain urban areas [7]; detecting anomalous taxicab trips to discover driver fraud or road network changes [17]; allowing taxicab passengers to query the expected duration and fare of a planed trip based on previous trips [14]; querying real-time taxicab availability to make informed transportation choices [3]; recommending optimal pickup locations or routes [4]. Moveover, taxicab GPS records can help beyond the taxicab business: assisting other drivers to improve their driving performance with GPS records from experienced taxicab drivers [18]; navigating newer drivers to smart routes based on those of the experienced taxicab drivers [19]; better understanding traffic conditions of cities [20]. Yet existing research on taxicab systems are focused on vacant taxicabs, assuming that one taxicab can accommodate only a single delivery request at a time. In contrast, our recommendation system aims for both vacant and occupied taxicabs.

Currently, carpooling taxicab services exist in big cities in an *ad hoc* fashion. For example, in New York City, up to four passengers can carpool together in a single taxicab ride during 6 AM to 10 AM on a weekday, along three preset routes in Manhattan at a flat fare of $3 or $4 per passenger, significantly less than the regular metered rates [21]. However, no systematic method under the existing infrastructure is provided to improve the efficiency of carpooling.

IX. CONCLUSION

In this work, we analyze, design, and evaluate a recommendation system *CallCab* for both carpooling and regular taxicab services in taxicab networks. *CallCab* mines taxicab trip distributions from historical GPS datasets collected in an existing infrastructure, and recommends either a vacant taxicab with no detour distance or a carpool route with a small detour distance. We verify *CallCab* with a real world dataset of 14, 000 taxicabs, and results show that compared to ground truth, *CallCab* decreases 60% of the total mileage, 41% of the passenger's waiting time and 28% of the total travel time.

REFERENCES

[1] S. Consulting. *The New York City Taxicab Fact Book*. [Online]. Available: http://www.schallerconsult.com/taxi/taxifb.pdf, accessed 2006.

[2] *Taxi of Tomorrow Survey*, New York City Taxi & Limousine Commission, New York, NY, USA, 2011.

[3] W. Wu, W. S. Ng, S. Krishnaswamy, and A. Sinha, "To taxi or not to taxi?—Enabling personalised and real-time transportation decisions for mobile users," in *Proc. IEEE 13th Int. Conf. Mobile Data Manage. (MDM)*, Jul. 2012, pp. 320–323.

[4] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2010, pp. 899–908.

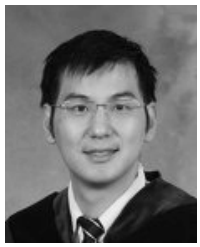
[5] H. Yang, C. S. Fung, K. I. Wong, and S. C. Wong, "Nonlinear pricing of taxi services," *Transp. Res. A, Policy Pract.*, vol. 44, no. 5, pp. 337–348, 2010.

[6] K. Yamamoto, K. Uesugi, and T. Watanabe, "Adaptive routing of cruising taxis by mutual exchange of pathways," in *Knowledge-Based Intelligent Information and Engineering Systems*. Berlin, Germany: Springer-Verlag, 2010.

[7] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 1010–1018.
[8] H.-W. Chang, Y.-C. Tai, and J. Y.-J. Hsu, "Context-aware taxi demand hotspots prediction," *Int. J. Bus. Intell. Data Mining*, vol. 5, no. 1, pp. 3–18, 2010.
[9] H. Gonzalez, J. Han, X. Li, M. Myslinska, and J. P. Sondag, "Adaptive fastest path computation on a road network: A traffic mining approach," in *Proc. 33rd Int. Conf. Very Large Data Bases (VLDB)*, 2007, pp. 794–805.
[10] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell, "Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior," in *Proc. 10th Int. Conf. Ubiquitous Comput. (UbiComp)*, 2008, pp. 322–331.
[11] Transport Committee of Shenzhen. (2014). *Statistical Data for Transportation in Shenzhen*. [Online]. Available: http://www.sz.gov.cn/jtj/tjsj/zxtjxx/
[12] *Average Salary in Shenzhen*. [Online]. Available: http://bsy.sz.bendibao.com/bsyDetail/4826.html, accessed 2013.
[13] Data100 Company. (2012). *Taixcab Carpooling Survey*. [Online]. Available: http://wenku.baidu.com/view/2f0fea1f964bcf84b9d57b4c.html
[14] R. K. Balan, K. X. Nguyen, and L. Jiang, "Real-time trip information service for a large taxi fleet," in *Proc. 9th Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2011, pp. 99–122.
[15] D. Agrawal *et al.* (2012). *Challenges and Opportunities With Big Data*. [Online]. Available: http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf
[16] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in *Proc. 6th Conf. Symp. Opeart. Syst. Design Implement. (OSDI)*, vol. 6. 2004, pp. 1–13.
[17] D. Zhang, N. Li, Z.-H. Zhou, C. Chen, L. Sun, and S. Li, "iBAT: Detecting anomalous taxi trajectories from GPS traces," in *Proc. 13th Int. Conf. Ubiquitous Comput. (UbiComp)*, 2011, pp. 99–108.
[18] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. 17th Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 316–324.
[19] L.-Y. Wei, Y. Zheng, and W.-C. Peng, "Constructing popular routes from uncertain trajectories," in *Proc. 18th Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 195–203.
[20] W. Zhang, S. Li, and G. Pan, "Mining the semantics of origin-destination flows using taxi traces," in *Proc. ACM Conf. Ubiquitous Comput. (UbiComp)*, 2012, pp. 943–949.
[21] New York Times. *Limited Share-a-Cab Test to Begin Soon*. [Online]. Available: http://www.nytimes.com/2010/02/22/nyregion/22ataxis.html, accessed 2010.
[22] D. Zhang, T. He, Y. Liu, and J. A. Stankovic, "CallCab: A unified recommendation system for carpooling and regular taxicab services," in *Proc. IEEE Int. Conf. Big Data (BIGDATA)*, Oct. 2013, pp. 439–447.

DESHENG ZHANG (M'10) is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA. He was a recipient of the Chinese Government Award for Outstanding Students Abroad, and was one of the four computer science student recipients selected from over 2 50 000 Chinese students studying in the U.S. in 2013. His research includes big data analytics and intelligent transportation systems.

TIAN HE (M'03–SM'12) is currently an Associate Professor with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA. As a recipient of the U.S. National Science Foundation CAREER Award'09, he served a few Program Chair position in international conferences, and also serves as an Editorial Member for several journals, including the *ACM Transactions on Sensor Networks*. His research includes wireless sensor networks and distributed systems.



YUNHUAI LIU (M'03) received the Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2008. He is currently the Deputy Director of the Research and Development Center of Internet of Things with the Third Research Institute of Ministry of Public Security, Shanghai, China. His research interests include wireless sensor networks, cognitive radio networks, and extreme-scale datacenter and data networks.

SHAN LIN (M'03) is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY, USA. He received the Ph.D. degree in computer science from the University of Virginia, Charlottesville, VA, USA. His research is in the area of networked systems, with an emphasis on feedback control-based design in cyber physical systems and sensor systems. He works on wireless network protocols, medical devices, and smart transportation systems.

JOHN A. STANKOVIC (F'94) is currently the BP America Professor with the Department of Computer Science, University of Virginia, Charlottesville, VA, USA. He served as the Chair of the Department for eight years. He was a recipient of the IEEE Real-Time Systems Technical Committee Award for Outstanding Technical Contributions and Leadership, the IEEE Technical Committee on Distributed Processing Distinguished Achievement Award (inaugural winner), and six best paper awards, including the 2006 ACM Conference on Embedded Networked Sensor Systems. He was the Editor-in-Chief of the *IEEE Transactions on Distributed and Parallel Systems*. His research interests are in cyber physical systems, distributed computing, real-time systems, and wireless sensor networks.