$CallCab$: A Unified Recommendation System for Carpooling and Regular Taxicab Services

Desheng Zhang and Tian He
Department of Computer Science and Engineering
University of Minnesota, USA
{zhang,tianhe}@cs.umn.edu

Yunhuai Liu
Third Research Institute
Ministry of Public Security, China
yunhuai@trimps.ac.cn

John A. Stankovic
Department of Computer Science
University of Virginia, USA
stankovic@cs.virginia.edu

***Abstract*—Carpooling taxicab services hold the promise of providing additional transportation supply, especially in 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 regular and carpooling services, called $CallCab$, based on a data driven approach. In response to a passenger's request, $CallCab$ aims to recommend either (i) a vacant taxicab for a regular service with no detour, or (ii) an occupied taxicab heading to the similar direction for a carpooling service with less detour, yet without assuming any knowledge of destinations of passengers already on occupied taxicabs. To analyze these unknown destinations of occupied taxicabs, $CallCab$ generates and refines taxicab trip distributions based on GPS datasets and context information collected in the existing taxicab infrastructure. To improve $CallCab$'s efficiency to process such a big dataset, we augment the efficient $MapReduce$ model with a $Measure$ phase tailored for our application. We evaluate $CallCab$ with a real world dataset of $14,000$ taxicabs, and results show that compared to ground truth, $CallCab$ can reduce 64% of the total mileage to deliver all passengers and 63% of passenger's waiting time.**

I. Introduction

Among all transportation modes, taxicabs play a prominent role in residents' daily commutes in metropolitan areas [1], *e.g.*, in New York City [2], over 100 companies operate more than $13,000$ taxicabs with a daily demand of $660,000$ passengers. But regular taxicab services are very insufficient during extreme weather or rush hour in big cities, *e.g.*, New York City and Beijing, where the average waiting time for a taxicab in rush hour is more than 30 minutes [1]. To address such an issue, carpooling taxicab services are proposed in some taxicab networks to provide additional transportation supply, yet with the same number of taxicabs. In a carpooling service, a passenger can hail an occupied taxicab on streets or wait at a taxicab stand to carpool.

However, different from well known regular taxicab services where any vacant taxicab can take a passenger to any reasonable direction, in a carpooling taxicab service, 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. For example, a passenger heading to a direction to the east may not consider a taxicab with existing passengers heading to the west as a carpoolable taxicab, since according to "First Come, First Served" policy, there will be a long distance detour for this new passenger. So when vacant taxicabs are not available, a key question for a passenger is how to find a carpoolable taxicab?

Unfortunately, almost all taxicab recommendation systems [3] [4] [5] [6] [7] are focused on *how to find a vacant taxicab*. Little work, if any, is focused on *how to find a carpoolable taxicab* for a passenger. More importantly, how to find a carpoolable taxicab cannot be addressed by configuring the existing solutions for finding a vacant taxicab. This is because for a particular new passenger, a recommendation system can recommend *any vacant* taxicab in a regular service; but in a carpooling service, it has to recommend a *particular carpoolable* taxicab heading to the similar direction with this new passenger. This is challenging because in the existing taxicab infrastructure, even with the real-time taxicab GPS tracking, a recommendation system cannot know future directions of occupied taxicabs, since the destinations of passengers on these occupied taxicabs are unknown, until the passengers are dropped off. A straightforward yet trivial solution is to let drivers log passengers' destinations right after passengers enter taxicabs. But such a system is not feasible in the real world, since it requires an infrastructure upgrade with hardware for drivers to manually input destinations, which can be potential hindrances in terms of both cost and efficiency. So it is challenging to assist a specific passenger to find a carpoolable taxicab in the existing infrastructure, when no vacant taxicabs are available.

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, taxicab's location and status are uploaded to a dispatching center periodically (*e.g.*, 2 records/minute), 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 recreate daily operations of thousands taxicabs in the networks. More importantly, from this dataset, we can draw *taxicab trip distributions* to analyze an occupied taxicab's destination (thus future directions) based on context information, *e.g.*, the route this taxicab has already passed, time of day, or day of week, *etc*. These distributions enable a recommendation system to assist passengers to find carpoolable taxicabs.

Admittedly, dispatching taxicabs or informing passengers based on GPS datasets is not a new method. But existing dispatching or recommendation systems assume that only vacant taxicabs can pick up passengers, and little research, if any, has been done on carpooling taxicab services. In this

Prof. Tian He is the corresponding author of this paper.

work, we conduct the first effort to propose a unified recommendation system for both CArpooLing and reguLar taxiCAB services, called $CallCab$, based on both GPS datasets and context information collected in the existing infrastructure. But dealing with such a big dataset (in terms of high volume and velocity, yet with raw unstructured format) requires an efficient design and processing model. In this paper, we are inspired by the efficient $MapReduce$ model [8] proposed by Google to handle large datasets, and augment it with a new $Measure$ phase for our application. Specifically, the key contributions of this paper are as follows:

- To the best of our knowledge, we conduct the first work that recommends either a vacant or a carpoolable taxicab (an occupied taxicab heading the similar direction) in responds to a passenger's request with a unified method, and provide a comprehensive study of how to analyze occupied taxicabs' routes without destinations of their passengers for large-scale taxicab networks.
- To achieve our goal, we propose $CallCab$, which mines taxicab trip distributions from historical and real-time GPS datasets collected in the existing taxicab infrastructure without extra costs. Then, according to these trip distributions conditioning on collected context information (*e.g.*, the last pickup locations and current locations of nearby taxicabs, time of day, day of week, *etc*) for a particular new passenger, $CallCab$ recommends either a vacant taxicab for a direct route (no detour distance), or a carpoolable taxicab for a carpool route (small detour distance) based on the similarities between directions of this new passenger and potential taxicabs.
- To quantify the similarity between directions, we propose a novel metric called *Detour Ratio*, which is shown as 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 a taxicab (either vacant or occupied) with the minimum detour ratio for a particular new passenger.
- To efficiently process GPS datasets for detour ratio calculation, we propose a generic $MapReduceMeasure$ model, inspired by $MapReduce$. This model provides 3 kinds of abstractions to hide details of data processing, and can be used for various applications.

Our evaluation effort is comprehensive. We test $CallCab$ on a real world dataset consisting of GPS records from more than $14,000$ taxicabs in a big metropolitan area with a population of more than 10 million. The results show that compared with ground truth, $CallCab$ can reduce 64% of the total mileage and reduce 63% of passenger waiting time, simultaneously.

The rest of the paper is organized as follows. Section II introduces the related work. Section III presents the existing infrastructure. Section IV proposes our main idea. Section V depicts our $MapReduceMeasure$ model. Section VI describes our detailed design. Section VII validates our design with datasets, followed by the conclusion in Section VIII.

II. Related Work

The concept of carpooling taxicab services is not brand new, but in the real world it is normally negotiated privately by drivers and passengers in an *ad hoc* manner, when vacant taxicabs are not available. We lack a systematic design for a unified recommendation for both regular and carpooling taxicab services. Two types of previous work related to our work are introduced as follows.

A. Regular Taxicab Services

Due to the increasing availability of GPS devices on taxicabs, taxicab GPS records have been employed by several systems to improve the efficiency of regular taxicab services. For example, taxicab GPS records are able to help taxicab operators better oversee taxicabs and provide timely services to passengers, *e.g.*, discovering temporal and spatial causal interactions to provide timely and efficient services in certain areas with disequilibrium [6] [7], and detecting anomalous taxicab trips to discover driver fraud or road network changes [9]. In addition to taxicab operators, several systems are proposed for the benefit of passengers or drivers, *e.g.*, allowing taxicab passengers to query the expected duration and fare of a planed trip based on previous trips [3] and query real-time taxicab availability to make informed transportation choices [4], and recommending optimal pickup locations or routes [5]. Moveover, taxicab GPS records can help beyond the taxicab business: (i) traces consisting of GPS records from experienced taxicab drivers can assist other drivers improve their driving performance [10]; (ii) GPS records can be used for navigating newer drivers to smart routes based on those of experienced taxicab drivers [11]; (iii) large scale taxicab GPS traces enable us to better understand traffic conditions of cities [12]. 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. Technically, we focus on recommending a taxicab ride (vacant or occupied) to passengers with the minimum detour distance, which has not been investigated before.

B. Ad Hoc Carpooling Taxicab Services

Currently, carpooling taxicab services exist in both developed and developing countries in an *ad hoc* fashion. For example, in Beijing, *ad hoc* taxicab carpooling is allowed with the consent of both passengers and drivers, and every passenger pays 60% of the regular fare. 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 [13]. However, in the aforementioned services, both time and locations are preset in a small-scale *ad hoc* manner, and no systematic method under the existing infrastructure is provided to improve the efficiency of carpooling.

III. Existing Taxicab Infrastructure

In this section, we introduce the existing taxicab infrastructure and discuss semantics data mined from a big dataset collected in this infrastructure.

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 global status of all taxicabs, dispatching centers with cloud servers are also established in 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.

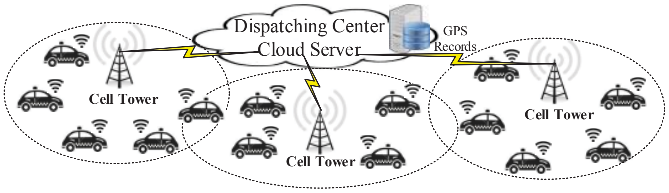


Fig 1: Existing Infrastructure

In such an infrastructure, (i) with GPS devices, taxicabs record their physical status, *e.g.*, current location, direction, *etc*; (ii) with fare meters, taxicabs record their logical status at any time, *i.e.*, with passengers or not; (iii) with communication devices, both physical and logical status are uploaded periodically to dispatching centers via cell towers, in terms of a GPS record, which mainly consists of the following parameters: Plate Number; Date and Time; GPS Coordinates; Status Bit: with passengers or not when this record is uploaded. Thus, a large GPS dataset is stored in cloud servers of dispatching centers for analysis. Figure 2 gives statistics about such a GPS dataset of a Chinese city Shenzhen with 10 million population.

GPS Dataset Summary	
Collection Period	6 Months
Collection Date	01/01-06/30
# of Taxicabs	14,453
# of Pickup Events	98,472,628
# of GPS records	3.9 Billion
Uploading Speed	2 Records/mins

Fig 2: Dataset Summary

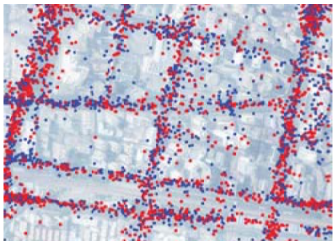


Fig 3: Origins & Destinations

As shown in Figure 2, a half-year dataset contains almost four billion GPS records. Such a raw large dataset has a very high resolution, which can be used to locate a particular taxicab at fine-granularity in terms of both time and space. But such a fine-granular large GPS dataset has many records of no interest, and such a raw GPS dataset is not in a format ready for analysis. In the next subsection, we data mine some semantics from this large fine-granular raw dataset, which is used to reduce the resolution (smaller yet compact size) of the raw dataset and to produce logical concepts, *i.e.*, trips, for our system design in Section V.

B. *Semantics in Existing Infrastructure*

Based on the historical and real-time GPS records, we separate individual trips from the entire dataset by continuously observing the change of Status Bit on 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. Figure 3 gives examples of origins and destinations. A GPS record set consisting of ***visited locations*** between an origin and its corresponding destination is considered as a ***trip***, which is the key unit of our design.

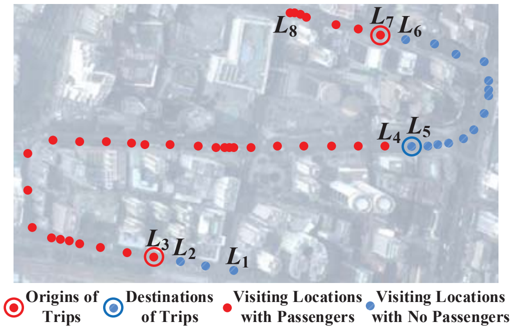


Fig 4: Taxicab Trips

Figure 4 gives examples of a trip, based on a recreated scenario in an aerial map according to GPS records. A taxicab starts with no passengers at location L_1, and picks up a passenger between L_2 and L_3 (the GPS record uploaded in L_2 indicates no passengers in this taxicab, while the GPS record uploaded in L_3 indicates a passenger in this taxicab), 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.

Given the semantics data mined from the real-time GPS dataset, the existing recommendation systems can easily locate and recommend a vacant taxicab to new passengers based on their locations. But if no nearby vacant taxicab is available, the existing recommendation systems cannot recommend an occupied taxicab for a carpooling service, since destinations of existing passengers on occupied taxicabs are unknown, and thus they fail to recommend a carpoolable taxicab heading in a similar direction as the new passenger.

But the large GPS dataset and context information provide is an opportunity to analyze the future directions of occupied taxicabs without destinations of passengers already in these taxicabs, and thus to locate carpoolable taxicabs, which is introduced in the next section.

IV. Main Idea

Our $CallCab$ aims for both regular and carpooling taxicab services. Since regular services are commonly understood, we give an example of a scenario where carpooling services are applied, and then present the main idea of $CallCab$.

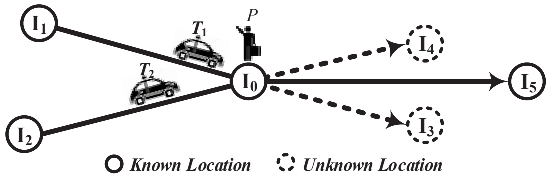


Fig 5: Taxicab Operating Scenario

A. Taxicab Carpooling Scenario

Figure 5 gives a scenario where a passenger P is 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 cannot find an available vacant taxicab, but can locate two nearby occupied taxicabs T_1 and T_2 that will pass P's origin I_0 soon, as potential carpoolable taxicabs.

To recommend T_1 or T_2 to P, a recommendation system has to analyze 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 being dropped off at P's own destination I_5, according to a "First Come, First Served" policy. Thus, the actual traveling distance for P to be carpooled into T_1 is the 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 route and the direct route leads to a *detour distance* of $(|I_0 \Rightarrow I_3| + |I_3 \Rightarrow I_5|) - |I_0 \Rightarrow I_5|$. With both a detour distance and a direct route's distance, we can have a ***Detour Ratio*** $\rho^P_{T_1} = \frac{\text{detour distance}}{\text{direct distance}}$ to show the utility of a passenger P being carpooled into a taxicab T_1. Different occupied taxicabs passing I_0 have different destinations, leading to different detour ratios for P to carpool. One of the optimal carpooling strategies for P is to select a taxicab with the smallest detour ratio as the carpoolable taxicab.

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 *completely unknown* in the existing infrastructure. Therefore, the 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 the trip, destinations are stored in terms of GPS records, after passengers are dropped off. These historical destinations and collected real-time context information is used to analyze unknown destinations of existing passengers in taxicabs, and thus to analyze detour ratios for a particular new passenger to carpool with these existing passengers in taxicabs. The details are given next.

B. Main Idea

The main idea of $CallCab$ is shown in Figure 6.

1) ***Trip Distributions*** *:* In $CallCab$, GPS records for all taxicabs are stored as a big dataset. Thus, inferred from GPS records, destinations and corresponding origins comprise numerous trips, which are used to construct *trip distributions*. Such distributions can generate destinations of trips that start at a particular origin and pass a particular location.

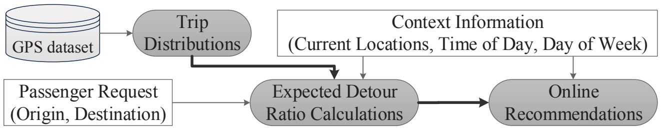

GPS dataset
Trip Distributions
Context Information (Current Locations, Time of Day, Day of Week)
Passenger Request (Origin, Destination)
Expected Detour Ratio Calculations
Online Recommendations

Fig 6: Main Idea

2) ***Detour Ratio Calculations*** *:* Upon receiving a *request* from a passenger P, $CallCab$ will employ trip distributions to calculate an *Expected Detour Ratio* $\rho^P_{T_1}$ for P to carpool with a particular nearby taxicab T_1, according to a basic and an advanced design. In both basic and advanced design, based on trip distributions, $CallCab$ (i) calculates a potential destination set DS_{T_1} for T_1 (DS_{T_1} includes all destinations 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 context information, and finally (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 destinations in DS_{T_1}, as shown by Figure 7.

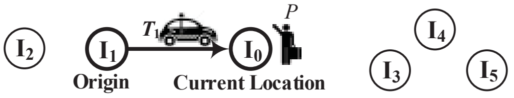


Fig 7: Detour Ratio Calculations

Basic Design: (i) Based on trip distributions and T_1's last pickup location I_1 (the origin of existing passengers on T_1), $CallCab$ calculates $DS_{T_1} = \{I_2, I_3, I_4, I_5\}$. (ii) If we assume that drivers will use shortest trips to deliver all passengers, then we can 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) 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.

Advanced Design: The advanced design is built upon the basic design. However, in the advanced design, (i) based on richer context information, $CallCab$ 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 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 destination, while others have I_4 as their destination, then $CallCab$ assigns $\Pr(I_3) = \frac{4}{6}$ and $\Pr(I_4) = \frac{2}{6}$ to calculate a weighted average $\rho^P_{T_1}$.

To summarize, the basic design conditions trip distributions on only limited *context information*, *e.g.*, origin and current locations of taxicabs, while the advanced design further considers richer *context information*, *e.g.*, popularity of destinations, time of day, day of week, *etc*.

3) ***Online Recommendation :*** By analyzing the detour ratio for every nearby taxicab, $CallCab$ recommends a taxicab with the minimum expected detour ratio (either a vacant taxicab with 0 detour ratio or an occupied taxicab with a small detour ratio) for this passenger. Based on updated *context information*, *e.g.*, taxicab current location, this recommendation can be constantly updated.

C. Opportunity for $MapRduce$ in $CallCab$ Design

The key step of $CallCab$ is how to obtain trip distributions based on raw GPS datasets. However, the raw GPS dataset shows physical aspects of taxicabs, while our design is focused on logical concepts, *e.g.*, trips, which are not directly given in raw datasets. Further, the raw GPS dataset typically has a large volume and interconnects multi-dimensional GPS records with high resolution, so though detailed enough, much of the raw dataset is of no interest in our design.

The above features fit the common understanding of big data [14]. To tackle this big dataset regarding these features, we need to map this raw physical GPS dataset to a filtered and compressed logical dataset (*e.g.*, trips) for analysis. In addition, we have to intelligently process this raw physical GPS dataset to a compacted size with only interesting data (*e.g.*, trips including a particular intersection). In this work, we are inspired by $MapReduce$ model proposed to deal with such big datasets [8], and augment it by an additional $Measure$ phase to present a generic $MapReduceMeasure$ model, which can be used independently from our design. In Section V and VI, employing the recommendation system as a showcase, we demonstrate how to use our model to tackle a big dataset that is not in a format ready for analysis.

V. $MapReduceMeasure$ Model

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

A. $MapReduceMeasure$ Introduction

Our $MapReduceMeasure$ model is mainly based on $MapReduce$, which is proposed 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.

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 propose 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 $MapReduceMeasure$ model is given as follows.

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

B. Preliminaries

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

1) Carpool Graph: The basic unit for a passenger to carpool with others is a road segment between intersections. Therefore, we define a carpool graph as a simple graph where vertices represent intersections and edges represent road segments between adjacent intersections. Figure 8 shows a carpool graph created by a given road map.

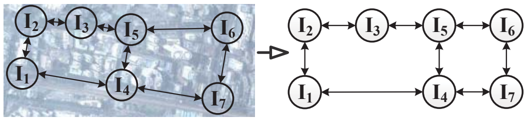
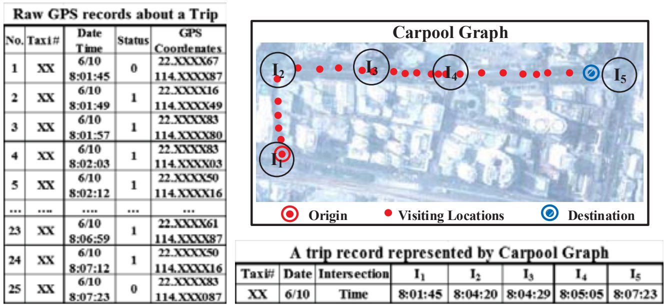


Fig 8: Carpool Graph

2) GPS Record Conversion: A set of raw GPS records belonging to a single logical trip is identified by several key GPS records, indicating the origin, the visited locations, and the destination. How to identify them is mentioned in Section III. 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 date.



Raw GPS records about a Trip

No.	Taxi#	Date Time	Status	GPS Coordenates
1	XX	6/10 8:01:45	0	22.XXXX67 114.XXXX87
2	XX	6/10 8:01:49	1	22.XXXX16 114.XXXX49
3	XX	6/10 8:01:57	1	22.XXXX83 114.XXXX80
4	XX	6/10 8:02:03	1	22.XXXX83 114.XXXX03
5	XX	6/10 8:02:12	1	22.XXXX50 114.XXXX16
...	....	....	...	...
23	XX	6/10 8:06:59	1	22.XXXX61 114.XXXX87
24	XX	6/10 8:07:12	1	22.XXXX50 114.XXXX16
25	XX	6/10 8:07:23	0	22.XXXX83 114.XXX087

A trip record represented by Carpool Graph

Taxi#	Date	Intersection	I_1	I_2	I_3	I_4	I_5
XX	6/10	Time	8:01:45	8:04:20	8:04:29	8:05:05	8:07:23

Fig 9: a Trip Record Based on a Carpool Graph

Based on a carpool graph, Figure 9 shows how to convert 25 raw GPS records to a trip record. Masks on Taxicab Number and coordinates are for privacy concerns. 25 GPS records describe the detailed trajectory of a trip, which can be mapped on a given road map, corresponding to an unique carpool graph. Thus, a trajectory is represented by a series of carpool graph's vertices, indicating intersections this trip went through, *e.g.*, intersection I_2, I_3, and I_4. We also add two intersections I_1 and I_5 near to the origin and the destination to complete the representation as in the figure. By the similar preprocessing, we convert the entire GPS dataset to a new trip record dataset as inputs for our model.

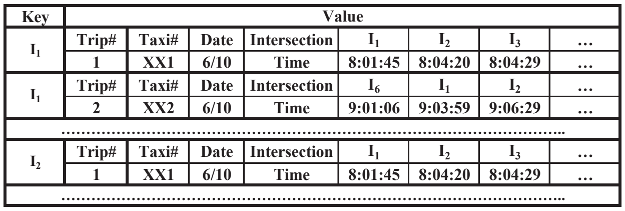
TABLE I. Model Operations

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 including I_1 as their first intersection
$ReduceByIS2$	I_2, Set of $[TRIP]$	Set of $[TRIP_2]$	$TRIP_2$ is one of trips including I_2 as their intermediate 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 to obtain the weighted average expected detour ratio
$MeasureA$	I_O^P, I_D^P, Set of $[TRIP]_{T_i}$	Detour Ratio $\rho_{T_i}^P$	Advanced design to obtain the weighted average expected detour ratio

C. Model Operations

Via the trip record dataset obtained in the last subsection, we propose Map, $Reduce$, and $Measure$ 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, time of day, or day of week) and the values associated to these new keys (*e.g.*, a trip includes this specific intersection, or a trip starts at a specific time of day or day of week). Three Map operations are proposed in Table I, *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. Note that multiple keys are paired with the same value, *e.g.*, I_2 is also paired with Trip#1. Fig 10 gives examples of $MapByIS$ on all intersections (*e.g.*, I_1, I_2, *etc*).



Key	Value							
I_1	Trip#	Taxi#	Date	Intersection	I_1	I_2	I_3	...
	1	XX1	6/10	Time	8:01:45	8:04:20	8:04:29	...
I_1	Trip#	Taxi#	Date	Intersection	I_6	I_1	I_2	...
	2	XX2	6/10	Time	9:01:06	9:03:59	9:06:29	...
	....							
I_2	Trip#	Taxi#	Date	Intersection	I_1	I_2	I_3	...
	1	XX1	6/10	Time	8:01:45	8:04:20	8:04:29	...
	....							

Fig 10: Map Operation

2) $Reduce$ Operations: Based on Map operations, $Reduce$ operation is to reduce the size of sets of values associated to the same key. We propose four $Reduce$ operations as in Table I, *e.g.*, $ReduceByIS1$ takes an intersection I_1 and all trips associated to I_1 as an input, and generates a smaller set of trips that include I_1 as their first intersection.

3) $Measure$ Operation: We propose $MeasureB$ and $MeasureA$ for the Basic and Advanced design as in Section IV-B2, respectively, which both take the following as inputs: (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 follows.

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

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 detour ratio, since no detour is needed for a vacant taxicab.

VI. $CallCab$ Design

In this section, based on the model we proposed in the last section, we present the detailed $CallCab$ 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 that a dispatching center received. By our Map operations in the model we generate different trip distributions for a particular intersection, time of day, or 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. Figure 10 gives the partial outputs of $MapByIS$ operations on all intersections.

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 context information.

- **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) in our context information. It has be shown in previous work that taxicab trips are highly patterned in terms of Time of Day and Day of Week [3]. For example, a trip distribution for $8AM$ on one Monday is very similar to a trip distribution for $8AM$ on another Monday, but it may be significantly different from a trip distribution for $11PM$ on a Sunday.
- **Nearby Taxicab Set T:** As potential taxicab candidates, T is a set of taxicabs (either vacant or occupied) heading to P's origin I_O^P, within a recommendation radius R^T to I_O^P (*e.g.*, 100 meters). For every taxicab $T_i \in T$, based on real-time GPS records, $CallCab$ can further obtain (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 context information, $CallCab$ can generate several particular distributions by model operations. These distributions are used to calculate an expected detour ratio for P to be carpooled into any taxicab $T_i \in T$.

*1) **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 model operations.

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

According to the above TripSet(Basic), $CallCab$ can obtain the expected detour ratio $\rho_{T_i}^P$ as follows, under the assumption that every destination has the same probability.

$$\rho_{T_i}^P = MeasureB(I_O^P, I_D^P, \text{TripSet(Basic)})$$

*2) **Advanced Design**:* $CallCab$ further generates two more trip distributions and combines them with TripSet(Basic) obtained in the basic design.

$$\begin{aligned}\text{TripSet}(TD) &= ReduceByTD(TD, MapByTD);\\ \text{TripSet}(DW) &= ReduceByDW(DW, MapByDW);\\ \text{TripSet(Advanced)} &= \text{TripSet}(TD) \cap \text{TripSet}(DW) \cap \text{TripSet(Basic)}.\end{aligned}$$

According to TripSet(Advanced), $CallCab$ obtains the expected detour ratio $\rho_{T_i}^P$ as follows, under the assumption that every destination has a different probability according to its frequencies in TripSet(Advanced).

$$\rho_{T_i}^P = MeasureA(I_O^P, I_D^P, \text{TripSet(Advanced)})$$

C. Online Recommendation

In our online recommendation, among all $\rho_{T_i}^P$ where $T_i \in T$, the taxicab T_{MIN} associated with the minimum ρ is the one $CallCab$ 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 on the 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 recommended taxicabs. Therefore, $CallCab$ updates this carpool graph based on the updated information, until the passenger is moving together with a taxicab, which indicates this passenger has already found a ride.

VII. $CallCab$ EVALUATION

We draw a sample dataset from the entire dataset introduced in Section III to test $CallCab$. This sample dataset includes one week of GPS records of more than $14,000$ taxicabs. Due to the large size of the datasets, we mainly found two kinds of errors. (i) Location Error: GPS coordinates indicate that a taxicab is off the road. (ii) Missing Records: a fair amount of GPS records are missing. The errors may result from different causes, *e.g.*, GPS device malfunctions, software issues, *etc*. We perform a simple preprocessing to clean the datasets to rule out taxicabs with more than 10% of missing or errant records.

A. Evaluation Overview

We compare two versions of $CallCab$, **Basic** and **Advanced**, against a **Random** and 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 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 the 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) Random recommends one of taxicabs in T at random; (ii) Heuristic recommends the closest taxicab in T to the passenger; (iii) Basic calculates the expected detour ratio for every taxicab in T based on the basic design in Section VI-B1, and then recommends the taxicab with the minimum expected detour ratio; (iv) Advanced works similarly, except that it calculates the detour ratio based on the advanced design in Section VI-B2.

Several metrics are proposed to show system performance.

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 can be calculated by the same method as in Section IV-A.

Further, we investigate **Percentage of Reduced Mileage**. Different from the first metric focusing the benefit of an individual passenger, this metric is used to evaluate how much the total mileage can be reduced (leading to less gas consumption and less traffic congestion) by an efficient system recommending more suitable occupied taxicabs for passengers. Assuming 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 by a specific system, then the percentage of reduced mileage equals $\frac{\text{M}-\text{m}}{\text{M}}$.

More importantly, we justify carpooling services by showing **Percentage of Reduced Waiting Time** due to carpooling, which has not been investigated before. 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).

Finally, we investigate the impact of **Insufficient Datasets** on the performance of $CallCab$, *e.g.*, only one day dataset is available for weekday trip distribution analysis.

We evaluate the performance according to the above metrics 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 250 meters. For both weekdays and weekends, we use requests from a one day dataset and test all systems with traces of taxicabs on other days. The average results are reported.

B. Actual Detour Ratio

In this subsection, we evaluate $CallCab$'s performance in terms of the average actual detour ratio.

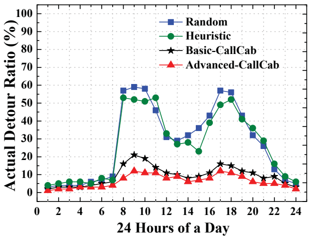


Fig 11: Detour Ratio in Weekday

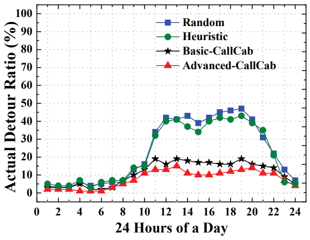


Fig 12: Detour Ratio in Weekend

1) *Weekday Detour Ratio:* Figure 11 plots the average actual detour ratio of all passengers in different one hour time windows of five weekdays. During the rush hours of a weekday, *e.g.*, 7-10, the average actual detour ratios for all four schemes are higher than those of non-rush hours, *e.g.*, 1-7. This is because there are many vacant taxicabs during non-rush hours, whereas in rush hours passengers have to use carpooling services, which leads to high actual detour ratios. But the Basic and Advanced solutions outperform Random and Heuristic, which have a high average actual detour ratios during rush hours, *i.e.*, 60% and 55%. Advanced outperforms Basic by 25% on average during rush hours, indicating that the probability considering frequency in Advanced leads to a better performance.

2) *Weekend Detour Ratio:* Figure 12 gives the average actual detour ratios in different one hour time windows for two weekends. During rush hours of a weekend, *e.g.*, 10-21, the average actual detour ratios for Basic and Advanced are much lower than those of Random and Heuristic. This is because during rush hours, Random and Heuristic recommend more occupied taxicabs with long detours to passengers according to their recommendation methods. But both versions of $CallCab$ utilize the trip distribution to recommend occupied taxicabs with less expected detour ratios, so it leads to a lower actual detour ratio. During rush hours, Advanced outperforms Basic by 34%, indicating that with more carpool opportunities during rush hours, Advanced more accurately assigns probabilities for every potential destination for recommendations of corresponding taxicabs.

C. Percentage of Reduced Mileage

In this subsection, we investigate $CallCab$'s performance via the percentage of the reduced total mileage.

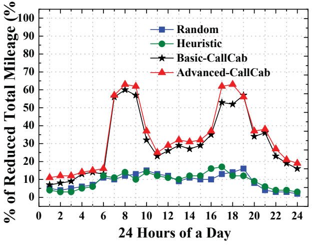


Fig 13: Mileage in Weekday

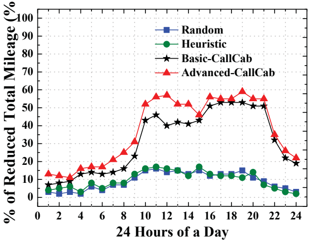


Fig 14: Mileage in Weekend

1) *Weekday Reduced Mileage:* Figure 13 shows the percentage of reduced mileage in different one hour time slots for five weekdays. During the rush hours of a weekday, *e.g.*, 7-10, the percentage of reduced mileage is higher than that of non-rush hours for all four schemes. This is because during rush hours, 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 Random and Heuristic during both rush hours and non-rush hours, which shows the effectiveness of $CallCab$. In addition, Advanced outperforms Basic by 18% on average during rush hours, indicating the superiority of Advanced.

2) *Weekend Reduced Mileage:* Figure 14 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 those 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 performances are similar as in Figure 13.

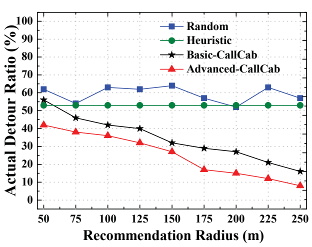


Fig 15: Detour Ratio VS. Radius

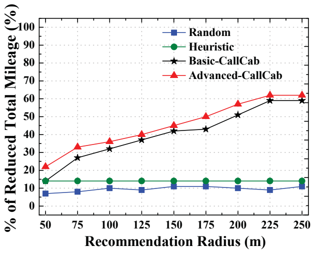


Fig 16: Mileage VS. Radius

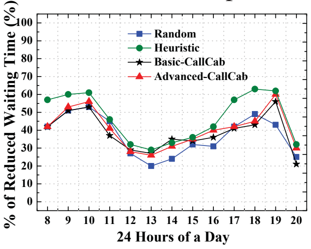
D. Impact of Recommendation Radius

In this subsection, we study the impact of recommendation radius on $CallCab$'s performance.

1) *Actual Detour Ratio with Different Radius:* Figure 15 shows the effects of different recommendation radii on the performance of the four schemes in terms of the average actual detour ratio from 8AM to 9AM 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. Random is not significantly affected, since it recommends a taxicab based a random selection. But with the increase of radius, both Advanced and Basic have better actual detour ratios, because a large recommendation radius gives them more taxicabs to select for a better recommendation. Also, with the increase of radius, Advanced always outperforms Basic, which confirms our observations in the previous subsections.

2) Reduced Mileage with Different Radii: Figure 16 shows the effects of different recommendation radii on percentage of reduced total mileage from 8-9 of a weekday. With the increase of the radius from 50 meters to 250 meters, the performance of both versions of $CallCab$ increases, while those of Random and Heuristic stay the same. But when the radius is close to 250M, the increase for both versions of $CallCab$ slows down, which is because the radius is large enough to have a sufficient number of taxicabs for recommendations, and an even larger radius would not help.



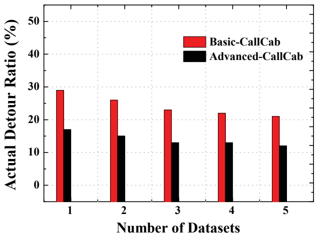


Fig 17: Waiting Time in Weekday Fig 18: Detour Ratio VS. Dataset

E. Percentage of Reduced Waiting Time

In this subsection, we show the percentage of reduced waiting time due to carpooling in Figure 17. 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 hours 8 to 20 of a weekday, due to the high densities of taxicabs and pickup events. During rush hours, *e.g.*, from 8 to 10 A.M., all systems with carpool services reduce the waiting time by as much as 63% on average. Heuristic outperforms the rest because it recommends the closest occupied taxicab for carpooling service, and other systems perform similarly to each other. In short, carpooling services can significantly reduce passenger's waiting time, when regular taxicab services are not sufficient.

F. Impact of Insufficient Datasets

In this subsection, we evaluate $CallCab$ when there is no sufficient dataset for analysis. Figure 18 shows the performance of $CallCab$ in terms of detour ratio in one rush hour 8AM to 9AM of a weekday with the maximum recommendation radius, when one to five days of dataset are used for analysis. We observe that even though with only a dataset of one day, $CallCab$ still achieves satisfactory performance, *e.g.*, Advanced and Basic achieves a 16% and 28% detour ratio, respectively. When more datasets are available, the performance of both versions of $CallCab$ becomes better.

VIII. Conclusion

In this work, we analyze, design, and evaluate a recommendation system $CallCab$ for both carpooling and regular taxicab services in large-scale taxicab networks. $CallCab$ data 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 employ a generic $MapReduceMeasure$ model to efficiently tackle the raw GPS dataset to obtain recommended taxicabs. We verify $CallCab$ with a real world dataset of $14,000$ taxicabs, and results show that compared to ground truth, $CallCab$ can save the passenger's initial wait time and fare, by decreasing both 64% of the total mileage and 63% of the passenger's waiting time, leading to a faster and affordable service.

IX. Acknowledgements

This work was supported in part by US NSF Grants CNS-0845994, CNS-0917097, CNS-1239226, CNS-1239483, and by the China NSFC Grants 61170247 and 61322205, as well as China National Major Science and Technology Project 2013ZX01033002-003.

References

[1] S. Consulting, "The new york city taxicab fact book," *http://www.schallerconsult.com/taxi/taxifb.pdf.*

[2] N. Taxi and L. Commission, "Taxi of tomorrow survey," 2011.

[3] R. K. Balan, K. X. Nguyen, and L. Jiang, "Real-time trip information service for a large taxi fleet," in *Proceedings of the international conference on Mobile systems, applications, and services*, ser. MobiSys '11.

[4] 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 *Proceedings of the 2012 IEEE 13th International Conference on Mobile Data Management*, ser. MDM '12.

[5] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '10, 2010.

[6] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '11, 2011.

[7] Y. Huang and J. W. Powell, "Detecting regions of disequilibrium in taxi services under uncertainty," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '12, 2012.

[8] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," in *Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, ser. OSDI'04.

[9] D. Zhang, N. Li, Z.-H. Zhou, C. Chen, L. Sun, and S. Li, "ibat: detecting anomalous taxi trajectories from gps traces," in *Proceedings of the 13th international conference on Ubiquitous computing*, ser. UbiComp '11, 2011.

[10] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the international conference on Knowledge discovery and data mining*, ser. KDD '11.

[11] L.-Y. Wei, Y. Zheng, and W.-C. Peng, "Constructing popular routes from uncertain trajectories," in *Proceedings of the international conference on Knowledge discovery and data mining*, ser. KDD '12.

[12] W. Zhang, S. Li, and G. Pan, "Mining the semantics of origin-destination flows using taxi traces," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12, 2012.

[13] N. Y. Times, "Limited share-a-cab test to begin soon," *http://www.nytimes.com/2010/02/22/nyregion/22ataxis.html.*

[14] D. Agrawal, P. Bernstein, E. Bertino, and etc, "Challenges and opportunities with big data," in *http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf*, ser. 2012.