ViFi-MobiScanner: Observe Human Mobility via Vehicular Internet Service

Lai Tu, Shuai Wang, Desheng Zhang, *Member, IEEE*, Fan Zhang, and Tian He, *Fellow, IEEE*

***Abstract*— Exploring human mobility is essential for urban applications. To observe human mobility, various data-driven techniques based on different data sources, such as cell phone and transportation data, have been proposed. This paper investigates human mobility through the emerging vehicular Internet service on public bus system. The key idea is that if a passenger is using WiFi on the bus or his/her WiFi device is activated in the background, we know that the passenger is traveling on the bus. By fusing the network events generated by WiFi devices with the data from the automatic fare collection (AFC) system, and the bus GPS information, we exploit not only the origin but also the destination of a passenger. Based on this idea, we develop a novel system called ViFi-MobiScanner which consists of about 4, 800 mobile routers distributed in a city with 1, 992 KM^2 urban area. We develop an ID matching algorithm that matches part of the users' network identities and their smartcard identities anonymously. As a result, we have built a set of labeled samples with the reference of observation from smartcard data and use them to train a classifier to infer users mobility from their network activities. We evaluate ViFi-MobiScanner with both field tests and collected datasets associated with 168 million network events, 3.6 million trips, and 1.4 million users. The evaluation results show that ViFi-MobiScanner increases the observability on the passengers and trips by about 53.9% and 48.1% over the smartcard observations. ViFi-MobiScanner also helps to estimate the passengers' destination that cannot be observed by current smartcard systems and the estimation can be accomplished in minutes. Thus it expands the observability of mobility in object, temporal and spatial dimensions and provides unique insights on human mobility at metropolitan scales.**

***Index Terms*— Human mobility observation, vehicular internet service, multi-source data mining.**

Manuscript received January 9, 2018; revised August 19, 2018, August 13, 2019, and October 28, 2019; accepted November 18, 2019. This work was supported in part by the China National Key Research and Development Program under Grant 2018YFB2100300, and in part by the National Natural Science Foundation of China under Grant 6167219 and Grant BK20190336. The Associate Editor for this article was C. F. Mecklenbräuker. *(Corresponding author: Lai Tu.)*

L. Tu is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: tulai@hust.edu.cn).

S. Wang is with the Key Laboratory of Computer Network and Information Integration, School of Computer Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: shuaiwang@seu.edu.cn).

D. Zhang is with the Department of Computer Science, Rutgers University, Piscataway, NJ 08854-8019 USA (e-mail: dz220@cs.rutgers.edu).

F. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: zhangfan@siat.ac.cn).

T. He is with the School of Computer Science and Engineering, Southeast University, Nanjing 210096, China, and also with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: tianhe@seu.edu.cn).

Digital Object Identifier 10.1109/TITS.2019.2956744

I. Introduction

UNDERSTANDING human mobility has great significance in a wide range of scientific studies such as anthropology, civil engineering, as well as urban applications in people's daily life, from predicting the spread of human viruses to urban traffic engineering and mobile network optimization. Mobility studies usually begin from observing human's moving behavior. Based on the observed moving behavior, mobility models are proposed by statistic analysis, feature extraction, and model abstraction. However, the legacy research on human mobility is based on hypotheses, e.g., Gravity and Radiation Models [1], or limited sampling data collected through surveys and censuses, which are often dated and incomplete.

Thanks to the era of big data and Internet of Things (IoT), there are more and more methods of observing the mobility. As people usually leave footprints when they use these infrastructures, by obtaining and processing the data generated by IoT, the mobility is observed. Since the locations of the IoT infrastructures are usually known, the mobility is then represented as a sequence of the location of the infrastructures with which the human associates.

In this paper, we study a new case of observing human mobility, which uses a new type of IoT infrastructure, i.e., the emerging vehicular Internet on buses. In the city of Shenzhen, which this study focuses on, about 4,800 buses are equipped with a mobile router (so called *ViFi*) that provides WiFi coverage and a 4G WAN gateway to the passengers.[1] By analyzing the network logs of these ViFi users, we now have a new opportunity to observe the human mobility. Based on this idea, we develop a system called ViFi-MobiScanner that utilizes the ViFi infrastructures and process the Network Event (i.e., the users' network activities logs) for mobility observation. The goal of the system is to increase the observability in object, temporal and spatial dimensions.

The typical scenario of ViFi-MobiScanner is illustrated in Figure 1, which shows the GPS location of the buses equipped with ViFi at 8:00AM in Shenzhen. Passengers who use ViFi leave footprints in term of Network Event (NE) logs on the ViFi device. The NE logs record the activity of ViFi users. It could be logs of users' ViFi connection or logs of the amount of networking traffic. The detailed types of NE logs are presented in Section IV. By studying the NE logs

[1] The system is called a ViFi system and the mobile router is named as ViFi router. Here ViFi stands for Vehicular WiFi and is same with WiFi in protocol and technical aspect, except that ViFi emphases on deployment on the mobile vehicle.

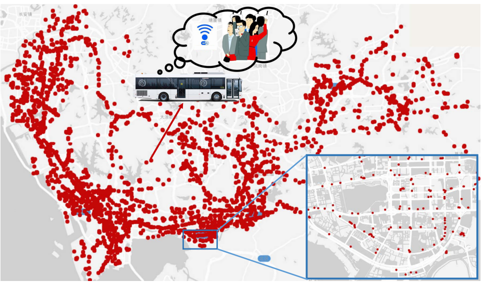


Fig. 1. An illustration of the on-bus WiFi coverage at 8:00AM in Shenzhen.

and fusing with the bus's GPS trace, we are able to extract the passengers mobility.

The ubiquitous mobile ViFi devices provides unique benefits to observe human mobility. Firstly, the ViFi system provides us a supplementary means to observe more human mobility. Moreover, it provides the passenger's fine-granularity trajectory information. Finally, since the ViFi system is connected with Internet, the mobility observation through the system will be convenient to link to a wide range of Internet applications and service such as context-aware mobile advertisement, real-time bus load notification [2] and mobility assisted handover [3].

To exploit the city wide human mobility through the ViFi system, we need to address the following challenges.

- First, although the ViFi system records users' network activities, it is possible that a user may have few ViFi activities. For example, users may not use the ViFi service during their ride and the ViFi system thus only records some sensible events, e.g., receiving *Probe Request* from a phone. Given the sparse network event, we need to develop techniques to discover the user's presence.
- Second, due to the feature of an open wireless, a mobile user who is not on a bus may occasionally connect to the nearby bus's ViFi router. Meanwhile, some on bus passengers have few interactions with the ViFi router for their inactive network usage. This results that discrimination base on naive handcrafted rules of hearing certain packets may misjudge a user's mobility status. Therefore, how to build a smart system that judges whether a mobile user is on the bus based on his/her network activity is a challenging task.

To address this problem, we design and deploy ViFi-MobiScanner system at both bus and cloud sides. We validate our system with both field test and consistency analysis with one-week smartcard data. Specifically, our contributions are as follows:

- We design a system called ViFi-MobiScanner to explore urban-scale human mobility with large-scale ViFi networks. In particular, ViFi-MobiScanner is a two-end system: on the front bus side, ViFi-Mobiscanner is implemented based on a mobile router, which is capable to appropriately record ViFi users' network actions and upload them to a central server; on the back cloud side, ViFi-Mobiscanner fuses different data sources to infer users' mobility status, i.e., whether the user is traveling on the bus.
- We develop an ID matching algorithm that matches part of the users' network identities and their smartcard identities anonymously. As a result, we have built a set of labeled samples with the reference observation from smartcard data and use them to train a classifier to infer users' mobility from their network activities.
- We deploy ViFi-MobiScanner on about 4,800 buses in Shenzhen, which is one of the most crowded city in China (17,150 people per KM^2). To the best of our knowledge, the ViFi-MobiScanner system is a combination of one of the largest urban vehicular networks and one of the largest mobile WiFi networks in public bus systems.
- We evaluate ViFi-MobiScanner based on the datasets associated with 168 million network event logs, 3.6 million trips and 1.4 million users. The results show that the observation of ViFi-MobiScanner covers users, time and locations that cannot be observed by other sensing systems, thus providing unique insights on human mobility at metropolitan scales.

In the rest of the paper, we first briefly review the existing research on mobility analytics from different data sources in Section II. Then the overview of the system is presented in Section III. Section IV and Section V introduce the design of the mobility sensing module and the mobility inference module. We evaluate the system in Section VI with both field tests and data driven simulations.

II. Related Work

A number of human mobility studies have been proposed in literatures, from analyzing the urban wide mobility using call detailed records (CDR) [4], [5], to tracking individuals by wireless fingerprints and phone sensing data [6], [7], from the buses and metros that carry tens and hundreds passengers [8], [9], to taxis and even rental bikes that only carry one or two [10]–[12]. These mobility study differs in observation objects, the data source, the granularity and scale.

Among various sources of mobility data, smartcard data are one of the most commonly used sources for passenger flow estimation in transportation systems [2], [13], [14]. Wireless signals including WiFi [15] and Bluetooth [16] are used to detect human mobility as well. However, inferring mobility from single source data like smartcard data can underestimate ridership [17], [18]. A mobility analytics platform, mPat is presented in [19], which explores human mobility using multi-source data from vehicles GPS trace, smartcards transaction records and cellphones CDR. A space alignment framework is proposed reconstruct individual mobility history by delicately aligning the monetary space and geo-spatial space with the temporal space [8].

Some other works on using wireless communications to detect the transportation mode of travellers are presented in [16], [20], [21]. These works are either based in small scale experiments, (e.g., one bus for 70Min [20] and 4 routes in one day [16]) or deployed in large scale but qualitatively studied on mobility observation (e.g., demonstration and visualization of passenger flows [21]).

Inspired by previous study, we use ViFi to observe passengers' mobility in urban scale and quantitatively investigate how and how much ViFi can expand the observability over other data source. We fuse ViFi data with bus systems' GPS data to obtain the passenger's fine-granularity trajectory information.

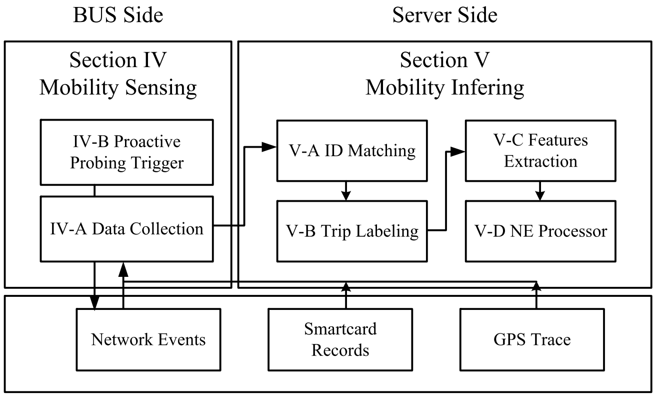


Fig. 2. Overview of the structure of ViFi-MobiScanner.

With the proposed ID matching algorithm, a part of the users' network identities and their smartcard identities are associated anonymously. This provides a new way that multiple data source are combined to complement each other for mobility observation. By exploiting these unique benefits, we conduct the first work to quantitatively explore urban scale human mobility with large scale vehicular Internet service data.

III. System Overview

We develop ViFi-MobiScanner to address the two challenges mentioned in Section I. Figure 2 shows the system architecture. The basic idea of ViFi-MobiScanner is to (i) trigger and collect multi-source data including ViFi data, smart card records and GPS data, and (ii) apply an automatic labeling technique to build a set of labeled samples that covers different users. Based on the labeled samples, ViFi-Mobiscanner trains a classifier as a mobility discriminator to infer a ViFi user's "*Mobility Status*", i.e., whether the user is traveling on the bus.

To achieve that, two core components, i.e., the *Mobility Sensing Module* and the *Mobility Inference Module*, run at the router and the center server respectively.

A. Mobility Sensing Module

Besides serving as the Internet portal, the ViFi router is modified to run the mobility sensing module that collects essential data for mobility inference and transmits them via any available wireless network to the cloud server. The mobility sensing module has two functions.

- **Data Collection:** The data collection submodule coordinates with multiple hardware devices on the bus to collect data. The collected data include (i)"*Network Events, NEs*" data from the ViFi router, (ii) smartcards' transaction data from the automatic fare collection (AFC) machine, and (iii) GPS data from the on-bus GPS devices.
- **Proactive Probing:** The proactive probing submodule is triggered when a user's sparse network events are detected. For example, few network events are recorded when phones are silent but have their WiFi turned on in the background. As a result, their status are not up to date. To increase the observability, this submodule sends some inquiry packets to trigger the phone to respond. If the phones respond, we label these events as "*Proactive Probing*" events.

We will discuss the data collection and proactive probing in details procedures in Section IV.

B. Mobility Inference Module

The collected data are compressed and sent to the cloud. The reason of processing the data remotely lies in both the limit of the computation capability of the router and more importantly, the need for global information for mobility inference.

The collected data are processed by the inference module which has four submodules. Their connections are as follows.

- **ID Matching:** The *ID Matching* submodule first exploits historical data to match the Network ID of a user's phone to the user's smartcard ID. With *ID Matching* function, the intersection of ViFi observation and smartcard observation is used to create a sub-dataset in which whether a ViFi user is on a bus can be revealed by its corresponding smartcard's usage record.
- **Trip Labeling:** Then in the *Trip Labeling* submodule, we label an *NE* sequence of a phone with tags of whether the *NE* sequence is associated with a bus trip. The smartcard usage data serve as the ground truth to label whether a user with a Network ID is on a bus. The first two steps accomplish *Automatic Labeling*. We thus build a training set of labeled *NE* sequence samples.
- **Feature Extractions:** With the training set, the *Classifier Learner* extracts features and trains a discriminator that takes *NE* sequence of a user as input and outputs his/her mobility status.
- **NE Processor:** Finally the discriminator works as *NE Processor* to process online *NE* stream data to output mobility observations.

Details on the inference module will be discussed in Section V.

IV. Mobility Sensing Module

In this section, we introduce the details of the mobility sensing module on the bus side. We first briefly introduce datasets that are collected for mobility inference. Then we present the way of triggering *NE* through proactive probing.

A. Data Collection

Table I lists the main attributes of the data used in the mobility inference. Three devices are involved, the on board GPS tracker, the Automatic Fare Collection (AFC) machine and the ViFi router. Each type of the device has a unique ID and can be associated with the bus plate number. The time stamps of the records in the three datasets are also synchronized. Therefore, by fusing the three datasets, it is possible to tell the time and location of a smartcard tapping the AFC machine on the bus and a user being sensed by the ViFi router.

Among the three datasets, the Network Events set is a set of logs that records the activities with respect to the people's using the ViFi service. Figure 3 illustrates the *NE* processing in the router side. When a phone interacts with the

TABLE I
ViFi-MobiScanner Dataset Descriptions

Dataset	Field	Remarks
GPS	GPS Tracker ID	Unique ID of the GPS Tracker
	GPS Coordinates	GPS Coordinates at the sample rate of twice per minute
	Time Stamp	Time of the GPS record
Smartcard	AFC ID	Unique ID of the Automatic Fare Collection machine
	User's SID	Smartcard ID of a Tapping record
	Time Stamp	Time of the Smartcard Tapping record
Network Events (NE)	User's NID	User's Network ID of an NE
	ViFi Device ID	Unique ID of the ViFi router
	NE Type and Value	See Table II
	Time Stamp	Time of the NE's occurrence

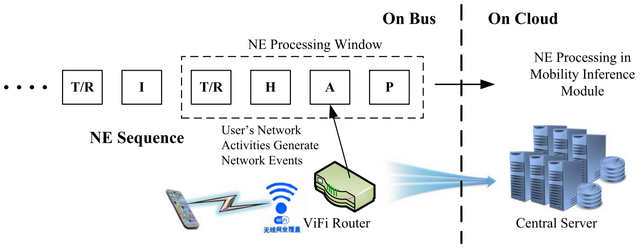

On Bus
On Cloud
NE Processing Window
. . . .
T/R
I
T/R
H
A
P
NE Processing in Mobility Inference Module
NE Sequence
User's Network Activities Generate Network Events
ViFi Router
Central Server

Fig. 3. An illustration of the Network Events generation and transmission on bus side.

ViFi router, the ViFi router generates a stream of *NEs*. Some *NEs* record the type of a WiFi frame, such as **A**ssociation and **D**issociation. Some *NEs* have specific value, such as the amount of **T**x/**R**x Traffic of a session. The *NEs* stream is then cut into sequences with a preset sliding window. The users can enjoy ViFi service either by using standard HTTP Internet service or by a portal mobile App of ViFi that uses a private protocol. If the user's usage generates an http request, an **H** event will pop up. If the user has some action on the ViFi App, it produces an **I** event.

The sequences are sent to the cloud server for data processing. All the *NEs* captured by the router are labeled with the unique *Network ID (NID)* of the terminal and the time stamp when the event occurs. The *NID* is implemented as the ciphered MAC address of a terminal. It is used to identify a terminal for our mobility sensing purpose while preserving privacy. The terminal's physical address can be interpreted from the *NID* by the router for network control while they are kept hidden to the system operators in the central server. The *NEs* used of mobility inference are presented in Table II.

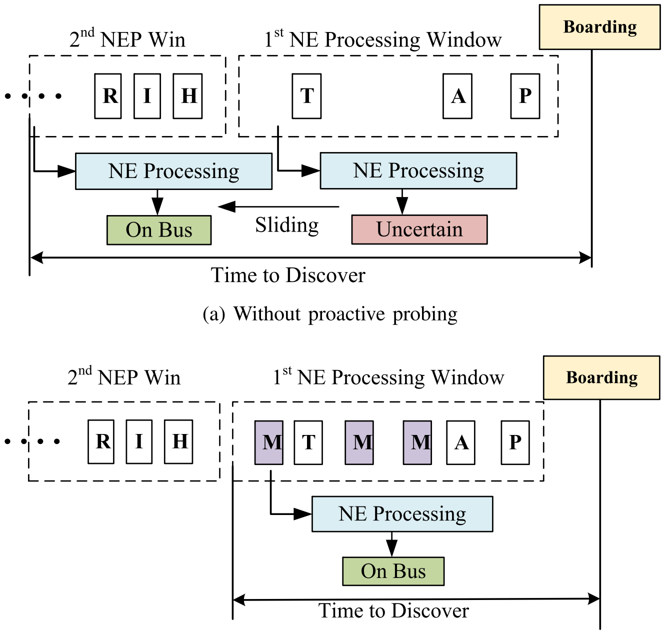
B. Proactive Probing

In addition to correctly determining a user's mobility status, we are expected to make a judgment as quickly as possible. When using the *NE* sequence as the input for mobility observation, the *NEs* come as real-time event streams. If we use statistic features of *NE* sequences for a judgment, it may take a long time to get enough *NEs* that can represent the distinguishing statistic features, as illustrated in Figure 4a.

Therefore, we generate extra *NEs* for mobility observation. Figure 4b illustrates the basic idea of proactive probing. When the native *NEs* of an associated terminal are sparse, the router proactively sends a small packet to the terminal and requires the terminal's response. There are many types of packets that can be used as the probe packet. If the probing target has associated with the ViFi router, we use a small MAC layer data frame as a probe packet. Upon receiving the frame, the targeted terminal shall respond the data frame with an ACK. On the other hand, if the probing target is in idle state, the best choice is using RTS control frames as probe packets, a.k.a. the "RTS Control Frame Attack" [22]. If the corresponding ACK or CTS is received by the router, the router inserts a manufactured *NE* with the targeted terminal

TABLE II
Network Events Descriptions

Abbr.	Description
A	**A**ssociation: The event represents that the terminal has successfully associated with the ViFi AP.
D	**D**issociation: The event represents that the terminal has disassociated from the ViFi AP. This may happen due to the terminal's initiated disconnection or its inactivity for a certain period of time (5 minutes in current router setting).
T/R	**T**x/**R**x Traffic: The event logs the amount of uplink traffic transmitted by phone to the AP and downlink traffic received by phone from AP.
P	**P**robe Request: The event occurs when a probe request package broadcast by a terminal is heard by the AP. A phone with WiFi turned on that is not associated with any AP may periodically broadcast probe request every 5 ∼ 10 minutes.
H	**H**TTP: The event logs the user's every http request when he uses different types of http based Internet services.
I	**I**nteraction: The event logs the user's interaction of his using mobile App.
M	**M**anufactured Event: The event logs the response of the terminal when it is proactive probed by the router.



(a) Without proactive probing

(b) With proactive probing

Fig. 4. An illustration of how the proactive probing helps speed up mobility status determination.

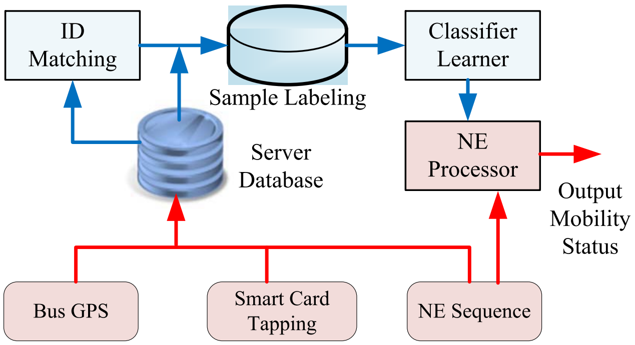


Fig. 5. Data flow and function components in mobility inference module.

NID in the *NE* stream. In this way, we fill the observing hole for positive samples and differentiate the positive samples with negative samples which may not respond the probing. The time to determine the mobility is shortened.

Currently, we set the proactive probing interval as 10 seconds per probing, which means about 100 times of the default BSSID beacon interval (100ms per beacon). Even in a bus full of passengers, the probing frequency won't produce traffic more than the BSSID broadcasting traffic with the highest beacon frequency (20ms per beacon). Meanwhile, the 10 seconds per probing will be enough for the requirement that estimating user's mobility in minutes.

V. The Mobility Inference Module

This section presents the design detail of the mobility inference module. We first introduce the workflow and the main function components in this module. Then we illustrate the algorithms in the key components.

A. The Workflow

The core function of the mobility inference module is to take multiple data as input to estimate the users' mobility status. The operations on the data are divided into two types, i.e., offline operation and online operation. The offline operation working on the historical data prepares strategies and parameter settings for the online operation. The online operations process real-time data streams and output estimations.

The workflow of the mobility inference module is illustrated as Figure 5. The red line and blue line indicate the data flow of online operation and offline operation respectively. The data streams collected from the routers are preprocessed to filter outliers by a set of rules (such as GPS drifting to irrational location, invalid timestamp) before using. Invalid values are dropped out to ease the process. The filter data are stored in the server database for future usage. The ID Matching component uses the archived historical data in the database to map a terminal's *Network ID (NID)* to its user's *Smartcard ID (SID)*. Then we can label whether a *NID*'s *NEs* sequence is presented in the corresponding *SID*'s trip. By doing so, we build a training set of *NEs* sequence samples that are labeled as "*Positive*" or "*Negative*" which represents whether they are presented in bus trips. With the training set, we train a classifier as the *NEs* processor that takes real-time data stream as input to output mobility status estimations.

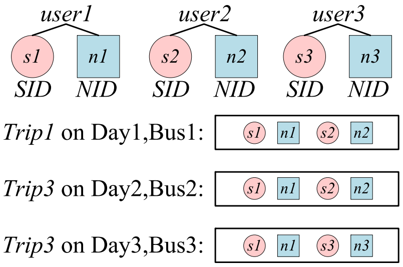


Fig. 6. An illustration of the possibility of matching ID.

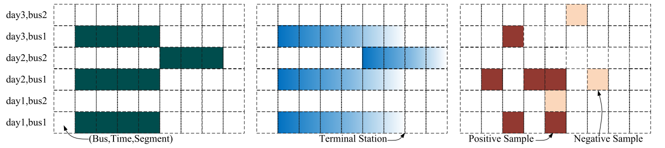
B. ID Matching and Trip Labeling

The automatic data labeling is one key component in mobility inference module, which operates based on the following assumptions:

- The purpose of matching the network ID with smartcard ID is to build a training set. This process does not require to match all ID pairs. We focus on matching IDs for passengers who have enough number of trips in a period of time. And we consider that these passengers' *NEs* sequence data are sufficient to cover different *NEs* sequences patterns for the training purpose.
- Secondly, we assume that the matching result is stable for a long period of time since most passengers do not change their smartcards or phones frequently. We thus run the ID matching algorithm on historical data and build a semi-static table of matched IDs. Periodically updates may run upon fresh data that load from new records. Here we only focus on the offline algorithm that takes static input from historical data.
- Finally, the basic idea of matching two elements in the two ID sets is based on the assumption that any two passengers are very unlikely to have exactly same trip patterns for many times in a long term period.

As a simplified example illustrated in Figure 6, two passengers *user1* and *user2* once get on a same bus at a same station, e.g., a trip *Trip1*. They will have much less probability to get on another same bus at the same time for another trip *Trip2*. Even this happens, it's almost impossible that *user1* meets *user2* in *Trip3*. Therefore given enough records of passengers' trips, the best matching *NID* for an *SID* is the *NID* that appears most frequently in the trips that the *SID* appears. So in Figure 6's example, $n1$ is matched with $s1$. It also noted that both $n1$ and $n2$ appear twice in $s2$'s trips. So we use a bipartite match algorithm to ensure that each *SID* is matched with at most only ONE *NID*. As $n1$ is more matched with $s1$ (based on the *Matching Score Calculation* method we present later in the section), $s1$ is excluded from the $n2$'s matching candidates.

In some minor probability cases when multiple users ALWAYS travel together, the ID matching procedure may not be able to distinguish one ID from another. However as their trips are exactly the same in such cases, matching the *NID* with



(a) Estimated OD Trip bitmap representation from SID Data (b) Fuzzy OD Trip bitmap representation from SID Data (c) Discrete Trip bitmap representation from NID Data

Fig. 7. Trip representation.

any *SID* of the co-travelers makes no difference for mobility observation.

Based on the above three assumptions, we match an ID in the SID set with an ID in the NID set and label a *NEs* sequence by the following four steps:

1) Trip Representation: As the ID match solution is based on the commonality of the trips extracted from SID and NID data, we first need to find a proper representation of the trips of a passenger in these two datasets. Physically, a passenger's trip on a bus can be represented as his/her association with the bus from the trip's *Origin (O)* to its *Destination (D)*. If we split a bus's route into segments between adjacent stops, we can denote a trip of a passenger as a bitmap (as shown in Figures 7) where a pixel stands for whether the passenger presents on a bus at a certain road segment at a time.

Accordingly, a trip of a smartcard begins from a user's boarding at O and ends at the alighting at D. However in most AFC systems, there are no records of alighting behavior in the smartcard dataset, i.e., the D is missing. Therefore, we present two alternative solutions:

- Trip with Estimated OD: In this solution, we pick up two kinds of special trips, i.e., (i) the trips that are followed by a transfer trip, and (ii) the trips that have regular round trip pattern. As shown in Figure 7a, the D of a trip can be estimated by the O of the next transfer trip or the return trip.
- Trip with Fuzzy OD: In case that the D cannot be estimated, as shown in Figure 7b, we use the terminal station to approximate the D. Considering that the terminal station is the farthest station that a passenger travels on a bus in a single trip, using the terminal station as the D leads to a relax criteria in connecting trips in two ID sets.
- Trip of NID: For a trip of an NID, as shown in Figure 7c, we examine the discrete segments where the NID is presented at least once. Then we use the trip of an SID as reference. If the presented discrete segments have overlap with the trip of an SID, we count that there is one common trip between the pair of SID and NID.

The approximation of the terminal to infer the D is only used for ID matching. As such approximation only increases the probability of ID collision linearly while the probability of two non-corresponding IDs (two users) to collide multiple times decrease exponential with the collision times. As we require high matching score (i.e., high collision number) to match two IDs, the error brought by this approximation is trivial.

2) Matching Score Calculation: Based on the above representation, we calculate the relation between the trips of an NID and an SID. We define a metric called "*Matching Score*" to measure how likely an SID matches to an NID. The *Matching Score M* of an SID to an NID depends on two factors: (i) the number of trips of an SID that have been found in NID's trips (denoted as N) and (ii) the ratio of trips of an SID that have been found in NID's trips (denoted as ρ). The *Matching Score* – M is thus given by $M = N\rho$.

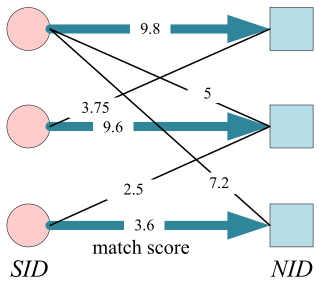


Fig. 8. Match IDs in the two sets by bipartite match.

Take the trip representations in Figure 7 as an example. There are 4 trips observed from SID data, i.e., row (*day1,bus1*), (*day2,bus1*), (*day2,bus2*), (*day3,bus1*), both with estimated OD and Fuzzy OD representations (shown in Figure 7a and Figure 7b). Among the 4 trips, 3 of them have observations in NID trip bitmap in corresponding pixels, i.e. row (*day1,bus1*), (*day2,bus1*), (*day3,bus1*), as shown in Figure 7c. So the number of the trips of the SID found in the NID's is 3 (i.e., $N = 3$) and the ratio of the ones found in the NID's is 3/4 (i.e., $\rho = 3/4$). The factor N and ρ reflect the commonality of two IDs in absolute aspect and relative aspect. By involving them together, we avoid bias caused by the difference in the travel frequencies of different users.

3) Solve Bipartite Match: After calculating the matching score of each pair of SID and NID, we map an SID to an NID by solving a weighted bipartite match problem. As shown in Figure 8, the nodes are the IDs in the two ID sets and the weight of a edge is the corresponding *Matching Score*. The weighted bipartite match problem can be solved in polynomial time. In our design, we use Hungarian algorithm to solve this match problem. In the matching results, we select the part of NIDs with high matching scores as training samples to ensure the matching confidence. We set the two thresholds for the factor N and ρ as the lower bounds for picking out a matching pair, which are $\rho \geq 50\%$ and $N \geq 10$. This means the NID and SID collide at least 10 times week (i.e., averagely 2 trips every business day) and a *Matching Score* of at least 2.5.

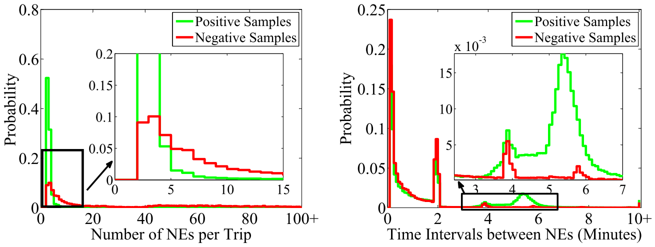
4) Label Trip Samples: After obtaining matching results between SID and NID, we label a *NE* sequence of an NID as "*Positive*" or "*Negative*" based on whether the corresponding trip is observed by the matched SID. Taking the example in Figure 7c, the *NE* sequences represented by dark red pixels in the NID trip bitmap are considered to be *Positive* samples. And the sequences represented by the orange pixels are considered to be *Negative* samples.

C. Features Extraction

As we plan to use the *NEs* sequence as input to determine a user's mobility status, we first study the statistic features

TABLE III
FEATURES USED FOR DISCRIMINATING THE MOBILITY STATUS

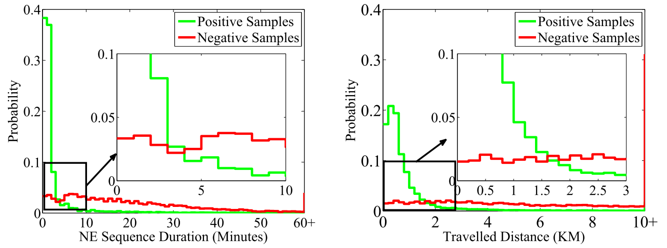
Category	No.	Features
Frequency features	0	The number of *NEs*
	1-4	The mean, minimum, maximum and middle value of the intervals of consecutive *NEs*
Temp-Spatial features	5	The duration of *NE* sequence
	6	The traveled distance of the *NE* sequence
	7	The straight line distance between the first and last *NEs*
NE type features	8	The number of user interactions
Traffic features	9	The amount of uplink traffic flow
	10	The amount of downlink traffic flow
	11	The asymmetry of up/down link traffic, i.e., the value of Feature 9 divided by the value of Feature 10



(a) The number of consecutive *NEs* associated with the same ViFi router in positive and negative samples.

(b) The time intervals of two consecutive *NEs* associated with the same ViFi router in positive and negative samples.

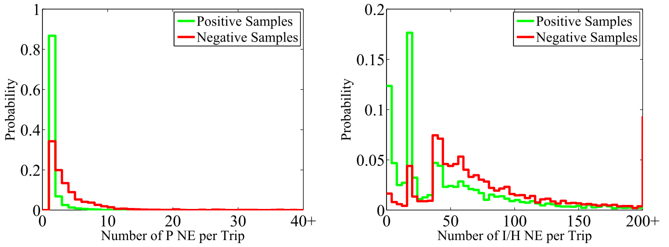
Fig. 9. Frequency features of the *NEs* sequence in positive and negative samples.



(a) Time duration distribution.

(b) Traveled distance distribution.

Fig. 10. Temporal-spatial features of the *NEs* sequence in positive and negative samples.



(a) The number of **P** *NEs* associated with the same ViFi router of on-bus and off-bus users.

(b) The number of **I/H** *NEs* associated with the same ViFi router of on-bus and off-bus users.

Fig. 11. Distributions of **P** and **I/H** *NEs*.

of the *NE* sequence. The features that contain distinguishable difference between positive and negative samples can be chosen as the features for the classifier in mobility inference. Based on the automatic labeled data, we build a set of $20,000$ samples with half positive samples and half negative samples. We study the features in Table III with regard to frequency features, temporal spatial features, user activity features and traffic features.

1) Frequency Features: Figure 9a shows the distribution of the number of consecutive NEs in positive and negative samples. Figure 9b shows the time intervals between consecutive *NEs* in positive and negative samples. The results show that statistically the negative samples have much larger probability to have less consecutive *NEs*. So numerous of consecutive *NEs* can be a sign of true bus trip. We also find that there is a peak at about 5 minutes in the distribution of *NEs* intervals. This is expected as 5 minutes happens to be the inactivity timeout threshold to disassociation. We can also utilize this feature as a criterion to determine whether an *NEs* sequence is responding to a true bus trip. If the *NEs* stream interrupt for more than 5 minutes, the user is more likely not traveling on the bus or s/he might have switched off the WiFi connection.

Note that there still exist some tiny peaks for large time intervals for negative samples in Fig. 9b. This is due to in some infrequent cases the buses stop at some place for a while. This sometimes happens when the buses experience traffic jam, or wait for traffic light or they are at their terminals. The nearby mobile users may intentionally or accidentally connect the ViFi routers for a while. That's also the reason why we involve spatial features like "The traveled distance of the NE sequence" for the classifier.

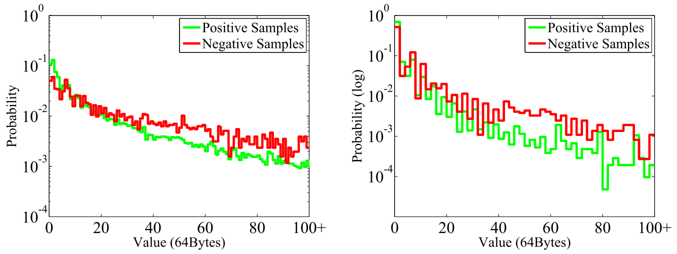
2) Temp-Spatial Features: Figure 10 depicts the distributions of the duration and traveled distance of the trips from the positive and negative *NEs* sequence samples. Same as the frequency and interval distributions, the negative samples show interrupts in both duration and traveled distance distributions. They are highly unlikely to have larger values as more than 5 minutes or more than 2 KM.

3) Features of Different Types of NE: We further separate the features to control related *NEs* and traffic related *NEs*. Figure 11a shows the number of **P** *NEs* in positive samples and negative samples. Although there is a little difference in the distribution of the number of **P** *NEs* in negative samples, their trends are similar. As a result, they contribute little in determining the users' mobility status. Another problem of probe requests is that some phones[2] may use randomized local MAC address when broadcasting a probe request [23], which makes the *NEs* hard to identify. This weakens its feature importance, and we exclude it from the features for classifier.

On the other hand, the number of user activities, including http requests and mobile App interactions, shows differences in distributions, as illustrated in Figure 11b. Statistically, there are more activities sensed in positive samples than the negative ones. We thus use this feature to determine the users' mobility status.

4) Traffic Features: Furthermore, we look at the traffic features. Figure 12 shows the amount of traffic in *NEs* in positive samples and negative samples. They both approximately obey

[2] E.g., iPhone 5s or newer using iOS newer than iOS 8



(a) Uplink flow (b) Downlink flow

Fig. 12. Traffic features of the positive and negative samples.

TABLE IV
CLASSIFIER PERFORMANCE COMPARISON

Classifier	5-fold cross validation score			
	Accuracy	Precision	Recall	F1-Measeure
Decision Tree	0.8741	0.8812	0.8653	0.8721
Extra Trees	0.9014	0.9047	0.8977	0.9004
Random Forrest	0.9135	0.9154	0.9108	0.9129
Gradient Tree Boosting	0.9150	0.9267	0.8989	0.9130
AdaBoost	0.9151	0.9291	0.9013	0.9132

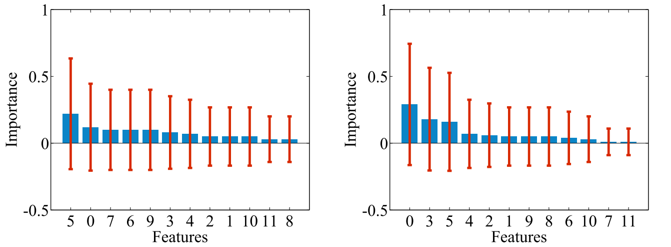
the power laws where the positive samples are slightly more possible to have larger amount value. We also involve them as one of the features in mobility status decision.

We also study other features such as the velocity of the bus at the time of the *NE*, the distance to the nearest stop of the bus at the time of the *NE*. Due to the limit of GPS reporting rate (30sec per report) in current ViFi devices, they provide few useful clues in discriminating the samples. Therefore, based on the statistical analysis, we preliminarily choose the above features for estimating a user's mobility status and use the data-driven approach to let the data trains the classifier to get the thresholds to partition the space.

D. Network Event Processor

Based on the preliminary analysis, we choose the twelve features listed in Table III and use them in the *NEs* processor. We evaluate several classifiers on the set of 20, 000 samples with half positive samples and half negative samples. We use 5-fold cross validation method to evaluate the classifiers, i.e., the samples are randomly and evenly divided into 5 parts, and we use 80% of the samples as the training set and 20% as the test set. The mean score of the accuracy, precision, recall and F1-measure metrics in the 5-fold cross validation are shown in Table IV. Results show that the Adaboost classifier performs best.

Figure 13a and Figure 13b give the results of "gini feature importance" [24] of the Adaboost classifier when proactive probing is and is not enabled respectively. Based on the training results, the duration of the sequence, the number of the *NEs* in the sequence, maximum interval between consecutive *NEs*, the traveled distance and the number of user actions, are the five most important features in discriminating the mobility status when proactive probing is not used. In case that proactive probing is enabled, the number of the *NEs* in the sequence and the maximum interval between consecutive *NEs* becomes more important. And the duration becomes less important. This makes the classifier rely less on the duration, which also helps *NEs* processor to be able to discriminate the mobility status quick in processing the real time *NEs* stream.



(a) Feature importance analysis w/o proactive probing (b) Feature importance analysis w/ proactive probing

Fig. 13. Feature importance analysis.

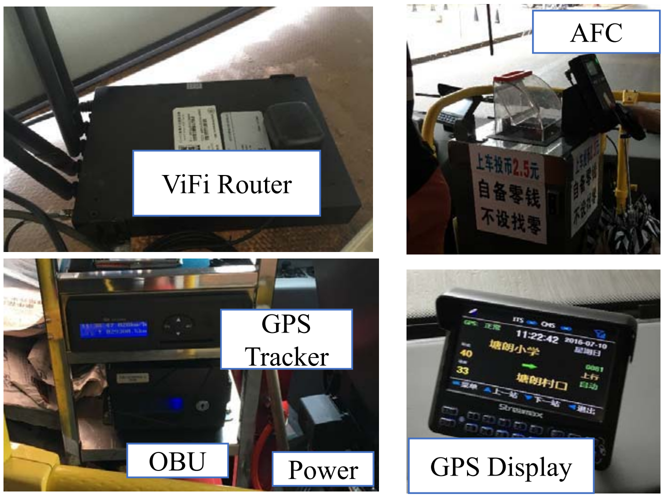


Fig. 14. Devices deployed on the buses.

VI. EVALUATIONS

The mobility study in this paper is based on the vehicular Internet service system that we deployed in the city of Shenzhen. In this section, we briefly introduce the design and deployment of system, and then present the performance evaluations of the city's mobility.

A. Implementation and Deployment

The vehicular Internet service in this work is based on deployment of 3G/4G mobile router on the buses. We develop a custom model specifically for vehicular Internet service. The router is developed based on linux system running on an Atheros 535MHz processor with 2G DDR and a 64 SSD or an external SD Card.

Figure 14 shows the devices deployed on the bus. The system also cooperates with the GPS tracker and the Automatic Fare Collection (AFC) devices on the bus. The GPS coordinates of the bus are periodically reported to the cloud twice a minute. And the AFC devices collect the smartcard tapping records of the passengers getting on the bus. Currently, an additional On Board Unit (OBU) is equipped on the bus to transform GPS tracker data and AFC data to specific format and upload to the central server. The router and OBU are installed behind the driver's seat and the AFC machine is installed at the entrance gate.

To attract users, currently the ViFi Internet service is free to all users. Any user within the WiFi coverage can connect to the on-board WiFi. To provide better user experience, we also provide mobile App for registered users at the same time. The registration is also free of charge but requests the users to accept several agreements of collecting data to improve the App. Despite some unpopular routes that have too few logs and the routers in repair and maintenance, the collected data from about 4,800 routers are involved in this study.

We use one week's data for our study. The collected data includes about 168M network event logs, 6M smartcard tapping records and 33M GPS records. The static data such as the bus route map, stations' locations and the device-to-vehicle correspondence information is also provided.

B. Validation

We first validate the correctness of our ViFi-MobiScanner system. As it is hard to obtain the precise mobility status of large amount of passengers', we begin with validating the result of conducted field experiments of 20 trips. Then we compare the statistical features of the mobility observations from ViFi data with the observation from smartcard data. Thirdly, we analyze the consistency of the result of ID matching over time.

1) Field Tests: To validate the mobility discriminator, we invited several participants to travel on different ViFi deployed buses and mimic different ViFi using behaviors. The participants were requested to turn on the WiFi of their phones. For each trip, two participants were traveling together. And one of them was requested to use the ViFi service and the other is requested not to use it. The ViFi user participants mimic to use different Internet services, such as browsing, instant messaging, streaming. There is no strict requirement how the they use the ViFi service. They may use for a while or for the whole trip. We conducted the field experiments of 20 trips. The NE sequences of their trips are collected. The time of the participants boarding and alighting are also recorded. We input the collected *NE* sequence into the NE processor described in last section and exam the output mobility discrimination and the "Time to Discover". The results show that all the trips of the participants are discriminated correct. Figure 15 shows 4 trips as examples of the total 20 trips. The two figures in the first row in Figure 15 are the *NE* sequences of participants who use the ViFi. The two in the second row are from participants who did not use ViFi on corresponding same trip. Compared to the actual time of their boarding, there exist delays of about 3 to 15 minutes to discriminate their mobility status. The delays are time to get enough *NEs* that can represent the distinguishing features.

2) Passenger Flow Analysis: Besides validation with small number of participants' experiments data, we also compare some statistical features of the observation of smartcard users and ViFi users with large number of users. The common trends between the results of the two observation methods can also partially validate the correctness of ViFi-MobiScanner.

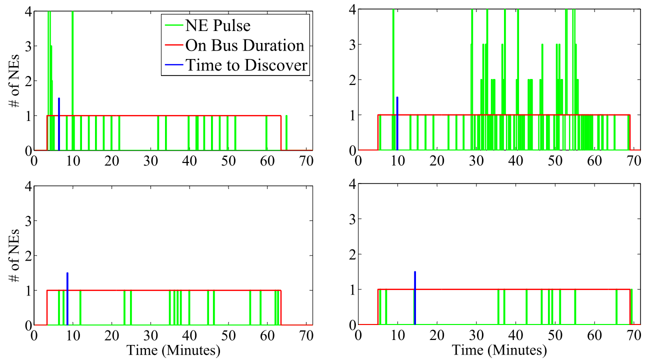


Fig. 15. Conducted field experiment.

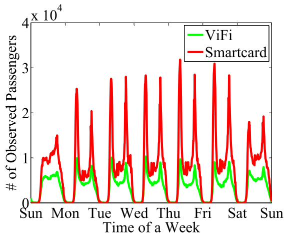


Fig. 16. The urban wide travel volume of observations of smartcard users and ViFi users.

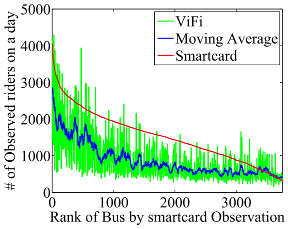


Fig. 17. The number of smartcard users and ViFi users carried by each bus on one day.

Figure 16 compares the urban wide travel volume from the observation of smartcard users and ViFi users. We can find that although not same or linearly proportional, they both show some common patterns such as two peaks in rush hours every day and increasing/decreasing trends.

We further study the number of smartcard users and ViFi users carried by each bus on one day. We rank the buses according to the number of smartcard users they carry on a day and exam the monotonicity of the number of observed ViFi users. The result in Figure 17 show that generally the buses carrying more smartcard users will observe more ViFi users. More smartcard users on bus is only one of the factors that lead to more ViFi users. But the relation is not simply linearly proportional nor monotonic.

3) Matching Consistency Analysis: Thirdly, we analyze consistency of the results of ID matching over four weeks. The ID matching operation runs on the data of each

TABLE V
THE CROSS-VERIFY RESULTS AMONG WEEKS WITH THE HIGH AND LOW ρ AND N PARAMETERS

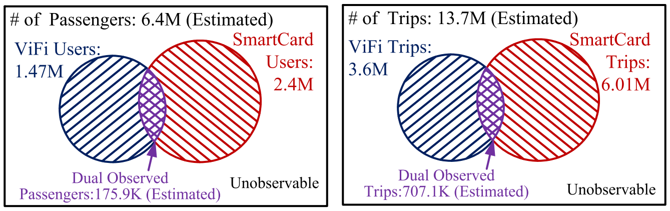
	X with Week1		X with Week2		X with Week3		X with Week4		Weekly matched pairs
$\rho = 50\%,\ N = 10$, (Absolute; Relative)									
	Verifiable	Conflicted	Verifiable	Conflicted	Verifiable	Conflicted	Verifiable	Conflicted	# of matched pairs
Week1			293; 30%	17; 1.8%	173; 18%	9; 0.9%	213; 22%	25; 2.6%	961
Week2	293; 14%	17; 0.7%			370; 17%	22; 1.0%	439; 20%	40; 1.8%	2142
Week3	173; 13%	9; 0.7%	370; 28%	22; 1.7%			383; 29%	23; 1.7%	1317
Week4	213; 7.9%	25; 0.9%	439; 16%	40; 1.5%	383; 14%	23; 0.9%			2696
$\rho = 30\%,\ N = 5$, (Absolute; Relative)									
	Verifiable	Conflicted	Verifiable	Conflicted	Verifiable	Conflicted	Verifiable	Conflicted	# of matched pairs
Week1			667; 32%	59; 2.8%	485; 23%	42; 2.0%	503; 24%	63; 3.0%	2109
Week2	667; 12%	59; 1.0%			1229; 22%	101; 1.8%	1447; 26%	149; 2.6%	5670
Week3	485; 10%	42; 0.9%	1229; 26%	101; 2.2%			1466; 31%	102; 2.2%	4692
Week4	503; 5.6%	64; 0.7%	1447; 16%	149; 1.7%	1466; 16%	102; 1.1%			8894

week separately. Totally we match 5566 NIDs to SIDs, among which only about 1% of match results have conflicts between two weeks. About 20% of match results can be cross validated by the results of different weeks. The other 79% match results are only found once. Note that we'd rather match an NID to no one than match it to a wrong SID that the NID happens to collide a few times. So we set a relative high matching score threshold ($\rho \geq 50\%$ and $N \geq 10$). However, this results in a relative low portion of matchable pairs. Only the commuters that often use ViFi on bus in the week will be the ID-matchable passengers. As some commuters may use ViFi intensively in a week while seldom use ViFi in another period, their matching results in the intensive using week cannot be cross validated by the other period. Nevertheless, the seldom using week won't produce a wrong matching result. That's why we have a very low inconsistent rate (around 1%) although the cross verifiable rate is not high. We also tested the cross verifiable and conflicted rate among weeks with $\rho = 30\%$ and $N = 5$. Matching results show that these lower threshold have provided much more valid matching pairs, which is 16511 matched ID pairs. The absolute numbers of both the cross-verifiable pairs and the conflicted pairs also increase. However the relative ratios of them over the whole number don't change too much. Table V gives the cross-verify results among weeks with the high and low parameters. Note that some matched ID pairs appear in three or four weeks. So the numbers of total matched ID pairs and cross-verifiable pairs are not the sum of the numbers of each week. We believe with data of longer period being applied in ID matching, we can increase the confidence of the match results and filter out the outliers and conflicted matches.

C. Observation Results

1) Observability in Object Domain: It is expected that using WiFi on the buses can observe a lot of network devices. Despite the randomized local NIDs, the collected NEs data have revealed 78M mobile devices with 168M NEs. The hundreds of millions NEs from the mobile devices provide a large amount of temporal spatial samples of mobility of hundreds of millions of human. To compare fairly, we care more about the NEs that are related to one's trips on buses.

Figure 18 illustrates the overall observations from smartcard and network events. Based on the result of NE processor running on the 168M NEs, there are 3.6M trips of 1.47M users are observed in the week. We compare them with the mobility observations from smartcard data. The smartcard data have recorded 6.01M trips of 2.4M smartcard users.


# of Passengers: 6.4M (Estimated)
ViFi Users: 1.47M
SmartCard Users: 2.4M
Dual Observed Passengers:175.9K (Estimated)
Unobservable
# of Trips: 13.7M (Estimated)
ViFi Trips: 3.6M
SmartCard Trips: 6.01M
Dual Observed Trips:707.1K (Estimated)
Unobservable

(a) The number of observed Passengers

(b) The number of observed Trips

Fig. 18. The numbers of users and trips observed from smartcard and network events.

Note that there must exist some unobservable passengers and trips, such as cash payers, fare evaders, free pass holders (for senior citizens and students) and period pass holders. So the actual sizes of the complete sets of passengers and trips can only be precisely measured by manual check, which is impractical for large scale measurement. We interviewed several bus drivers and were told that fare evaders and free pass or period pass holders are quite small portions compared to paid passengers. So the estimation based on toll data can be considered as approach between the ground truth and observable passengers and near the ground truth. Therefore, we use a downscale method to roughly estimate the total passengers and trips from the monthly toll data as reference. As the bus company does not audit every bus every day, we use the average monthly toll data on ViFi-equipped buslines to make the rough estimation. The estimated number of passengers and number of trips on the 4,800 ViFi-equipped buses in one week are 6.4M and 13.7M. At the same time, although we have matched 5566 NIDs to SIDs, the actual size of intersection of ViFi users and smartcard users is far more than 5566. This is because our ID Matching algorithm only picks out the matching results of the frequent travelers with high matching scores. The true numbers of the dual observable passengers and trips are immeasurable either. After reviewing relation between the number of ID matched travelers and the number of trips per user, as shown in Figure 19, we use a log-linear regression model to estimate the dual

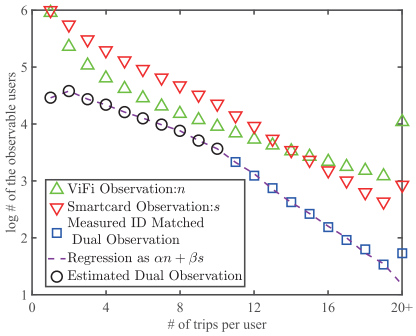
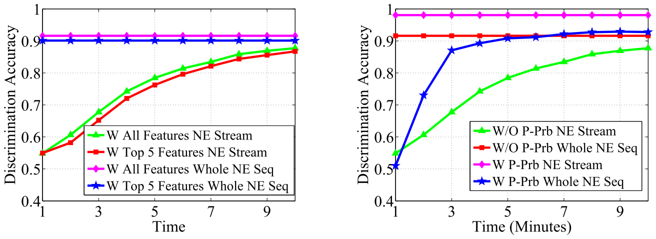


Fig. 19. Estimation of the dual observable passengers.



(a) Discrimination Accuracy without Proactive Probing (P-Prb)

(b) Discrimination Accuracy with Proactive Probing (P-Prb)

Fig. 20. An Illustration of that the proactive probing helps speed up mobility status determination.

observable passengers and trips. The estimated values are 175.9K and 707.1K. This implies that ViFi helps to reveal about $\frac{1.47}{6.4} = 23.0\%$ passengers and $\frac{3.6}{13.7} = 26.3\%$ trips of the total numbers, and adding ViFi data helps to increase the observability on the passengers and trips by $\frac{1.47-0.1759}{2.4} = 53.9\%$ and $\frac{3.6-0.7071}{6.01} = 48.1\%$ over the conventional single data source observation from smartcard.

2) Observability in Time Domain: Secondly, we take a look at the observability in time domain. We first split the samples of NE sequences into training set and test set. We load the whole NE sequence of every sample in the training set to the classifier learner and output the NE processor. We then get a reference score which is the classification accuracy of the NE processor that runs on the whole length sequence of sample data in the test set.

To mimic the system's processing the real time event streaming, we split the NE sequence samples in the test set into sub-sequences with a one-minute window. We let the NE processor load the sub-sequence one by one to see how long it takes to get close to the reference score.

Figure 20 presents the accuracies of the NE processor that runs on the sub-sequences over time. Figure 20a gives the result based on the solution without proactive probing. The reference score in accuracy is about 90% and it almost takes 8 minutes for the NE processor running on sub-sequence to get an accuracy of 80%.

Currently proactive probing is not implemented in the deployed ViFi router. We use data driven simulation to evaluate the effect of proactive probing. We insert **M** *NE* every minute to the positive *NE* sequence samples. For a negative *NE* sequence sample, the user is supposed not on the bus. If the bus moves far away from where it encounters the user, the user cannot response the probe. So **M** *NE* is only inserted if the bus position is still in a range of 500 meters to its last non-**M** *NE*. By doing so, we have a new set of positive and negative samples.

We use same method to evaluate the performance with the new set. Figure 20b show that with proactive probing, the reference score in accuracy is originally higher, i.e., about 98%. It only takes 4 $\sim$ 5 minutes and the accuracy reaches 90%. This suggests that the mobility status can be determined quicker and more accurate with proactive probing, which implies an improvement of the observability in time domain.

3) Observability in Space Domain: Finally, in the space domain, the main advantage of the observation from network events is that it outputs both the origin and an estimation of the destination of the trips. We compare the mobility observations from the smartcard data and the NE data. Among the observed trips 6.01M trips of 2.4M smartcard users, the destination of only about 35% trips can be estimated, which are 2.09M OD pairs. While the destination of the trip of a ViFi user is estimated by the bus stop next the user's last network events. Among the 3.6M observed trips from NE data, we can infer both O and D for all of them, which implies an improvement of the observability in space domain.

VII. Conclusions and Discussions

This paper proposed a novel human mobility observation system named ViFi-MobiScanner. ViFi-MobiScanner infers human mobility through fusing the information from the network event logs and buses' GPS traces. We develop an ID matching algorithm that matches part of the users' network identities and their smartcard identities anonymously, so that we have built a set of labeled samples to train a classifier to infer users' mobility from their network activities. Through the large-scale deployment on about 4,800 buses, we investigate the observability of ViFi-MobiScanner with both field tests and the collected datasets associated with 168 million network events, 3.6 million trips and 1.4 million users. The results show that ViFi-MobiScanner increases the observability on the passengers and trips by about 53.9% and 48.1% over the smartcard observations. ViFi-MobiScanner also helps to estimate the passengers' destination that cannot be observed by current smartcard systems.

Due to some restrictions of the hardware capability and the data source, our work still has some limitations: (i) First of all, our system relies on the availability of AFC data in the training phase. (ii) Proactive Probing is only simulated due to restriction of the commodity Wi-Fi chip set in the router. (iii) Due to the limit of the duration of data and its anonymity, there is no direct means to verify the ID matching results and mobility status discrimination in such large scale. (iv) Due to some values are not measurable such as the total number of passengers, we have to use some estimations in evaluation. (v) There are exceptional cases such as frequent co-riders, unexpected WiFi on/off due to human intervention or battery status change. These limitations may affect the accuracy and verifiability of our observations.

Nevertheless, we believe that in an urban wide scale, the benefit from the significant promotion in observability is more important than the above limitations. Although this paper only discusses the case of fusing ViFi data and AFC data, the methodology can be extended to some other similar applications. For instance, the ID matching method based on co-occurrence can also be used to connect face recognition result and wireless signal in search guilty suspect. For another example, we can apply the classifier trained with both ViFi and AFC data to other cases that AFC data is not available.

As future work, we are also working on involving more data in validation and evaluation and improving the accuracy and verifiability of ViFi-MobiScanner. We also plan to cooperate with the ViFi operator to design rich applications and provide more content to stimulate more passengers to use ViFi. Meanwhile the proposed method is also applicable to combine other multisource data to complement each other. Therefore we conclude that ViFi-MobiScanner provides some unique insights and a new methodology on human mobility observations.

References

[1] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008.

[2] J. Zhang *et al.*, "A real-time passenger flow estimation and prediction method for urban bus transit systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3168–3178, Nov. 2017.

[3] S. Sen, J. Yoon, J. Hare, J. Ormont, and S. Banerjee, "Can they hear me now?: A case for a client-assisted approach to monitoring wide-area wireless networks," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 99–116.

[4] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[5] P. Deville *et al.*, "Dynamic population mapping using mobile phone data," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 45, pp. 15888–15893, Nov. 2014.

[6] A. Thiagarajan, J. Biagioni, T. Gerlich, and J. Eriksson, "Cooperative transit tracking using smart-phones," in *Proc. 8th ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, New York, NY, USA, 2010, pp. 85–98.

[7] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden, and L. Girod, "Accurate, low-energy trajectory mapping for mobile devices," in *Proc. 8th USENIX Conf. Netw. Syst. Design Implement. (NSDI)*, Berkeley, CA, USA, 2011, pp. 267–280.

[8] N. J. Yuan, Y. Wang, F. Zhang, X. Xie, and G. Sun, "Reconstructing individual mobility from smart card transactions: A space alignment approach," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dallas, TX, USA, Dec. 2013, pp. 877–886.

[9] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. Part C, Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, Aug. 2011.

[10] D. Zhang, Y. Li, F. Zhang, M. Lu, Y. Liu, and T. He, "coRide: Carpool service with a win-win fare model for large-scale taxicab networks," in *Proc. 11th ACM Conf. Embedded Netw. Sensor Syst. (SenSys)*, New York, NY, USA, 2013, pp. 9:1–9:14.

[11] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proc. 13th ACM Int. Conf. Ubiquitous Comput. (UbiComp)*, New York, NY, USA, 2011, pp. 89–98.

[12] Z. Yang, J. Hu, Y. Shu, P. Cheng, J. Chen, and T. Moscibroda, "Mobility modeling and prediction in bike-sharing systems," in *Proc. 14th Annu. Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, New York, NY, USA, 2016, pp. 165–178.

[13] J. Zhao *et al.*, "Estimation of passenger route choice pattern using smart card data for complex metro systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 790–801, Apr. 2017.

[14] M. Dou, T. He, H. Yin, X. Zhou, Z. Chen, and B. Luo, *Predicting Passengers in Public Transportation Using Smart Card Data*. Cham, Switzerland: Springer, 2015, pp. 28–40.

[15] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann, "Tracking human mobility using WiFi signals," *PLOS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130824.

[16] V. Kostakos, T. Camacho, and C. Mantero, "Wireless detection of end-to-end passenger trips on public transport buses," in *Proc. 13th Int. IEEE Conf. Intell. Transp. Syst.*, Sep. 2010, pp. 1795–1800.

[17] A. V. Reddy, J. Kuhls, and A. Lu, "Measuring and controlling subway fare evasion: Improving safety and security at new york city transit authority," *Transp. Res. Rec.*, vol. 2216, no. 1, pp. 85–99, 2011, doi: 10.3141/2216-10.

[18] B. Barabino, S. Salis, and B. Useli, "Fare evasion in *proof-of-payment* transit systems: Deriving the optimum inspection level," *Transp. Res. B, Methodol.*, vol. 70, pp. 1–17, Dec. 2014.

[19] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, New York, NY, USA, 2014, pp. 201–212.

[20] T. Oransirikul, R. Nishide, I. Piumarta, and H. Takada, "Measuring bus passenger load by monitoring Wi-Fi transmissions from mobile devices," *Procedia Technol.*, vol. 18, pp. 120–125, Sep. 2014.

[21] P. M. Santos *et al.*, "PortoLivingLab: An IoT-based sensing platform for smart cities," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 523–532, Apr. 2018.

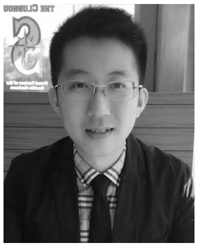
[22] J. Martin *et al.*, "A study of MAC address randomization in mobile devices and when it fails," Mar. 2017, *arXiv:1703.02874*. [Online]. Available: https://arxiv.org/abs/1703.02874

[23] C. J. Bernardos, J. C. Zúniga, and P. O'Hanlon, "Wi-Fi Internet connectivity and privacy: Hiding your tracks on the wireless Internet," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Tokyo, Japan, Oct. 2015, pp. 193–198.

[24] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, 1st ed. Boca Raton, FL, USA: CRC Press, 1984.

Lai Tu received the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology in 2007. From 2007 to 2010, he was a Post-Doctoral Fellow with the Department of EIE, Huazhong University of Science and Technology, China, and with the Department of CSIE, Nation Cheng Kung University, Taiwan. He was a Visiting Scholar with the University of Minnesota from 2015 to 2016. He is currently an Associate Professor with the School of Electronic and Information and Communications, Huazhong University of Science and Technology. His research areas include urban computing, human behavior study, mobile computing, and networking.

Shuai Wang received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, China, and the Ph.D. degree from the Department of Computer Science and Engineering, University of Minnesota, in 2017. He is currently a Professor with the School of Computer Science and Engineering, Southeast University. His research interests include the Internet of Things, cyber-physical systems, data science, and wireless networks and sensors.



Desheng Zhang (M'11) is an Assistant Professor with the Department of Computer Science, Rutgers University. He is broadly concentrated on bridging cyber-physical systems (also known as the Internet of Things under some contexts) and big urban data by technical integration of communication, computation, and control in data-intensive urban systems. He is focused on the life cycle of big data-driven urban systems, from multisource data collection to streaming-data processing, heterogeneous-data management, model abstraction, visualization, privacy, service design, and deployment in complex urban setting. He is currently interested in real-time interactions among heterogeneous urban systems including cellphone, smartcard, taxi, bus, truck, subway, bike, personal vehicle, electric vehicle, and road networks.

Fan Zhang received the Ph.D. degree in communication and information system from the Huazhong University of Science and Technology in 2007. He was a Post-Doctoral Fellow with the University of New Mexico and with the University of Nebraska-Lincoln from 2009 to 2011. He is currently a Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. He is also the Director of the Shenzhen Institute of Beidou Applied Technology (SIBAT). His research topics include intelligent transportation systems, urban computing, and big data and AI technology.

Tian He (M'03–SM'12–F'18) is currently a Professor with the School of Computer Science and Engineering, Southeast University. He is the author or coauthor of over 300 articles in premier network journals and conferences with over 23,000 citations (H-Index 70). His research interests include wireless networks, networked sensing systems, cyber-physical systems, real-time embedded systems, and distributed systems. He is a fellow of ACM. He was a recipient of the NSF CAREER Award in 2009, the McKnight Land-Grant Chaired Professorship in 2011, the George W. Taylor Distinguished Research Award in 2015, the China NSF Outstanding Overseas Young Researcher I and II in 2012 and 2016, and the eight best paper awards in international conferences, including MobiCom, SenSys, and ICDCS. He has served as the few general/program chair positions for international conferences and on many program committees and also has been an editorial board member for six international journals, including *ACM Transactions on Sensor Networks*, the IEEE TRANSACTIONS ON COMPUTERS, and *IEEE/ACM Transactions on Networking*.