

Chapter 2

Dealing with Vehicular Traces

2.1 Introduction to the Shanghai Grid Project

Intelligent transportation systems (ITSs) [72–74] have been evolving rapidly in the past two decades, leveraging advanced computing and communication technologies. ITSs help coordinate traffic condition, improve safety, reduce environmental impact, and make efficient use of available resources. Shanghai, the largest metropolis in China, covers an area of 5,800 km² and has a large population of 18.7 million. The economy of Shanghai is soaring today and the growing traffic has become a serious challenge. In response to the challenge and the needs of the public, the Shanghai government has established the SG project cooperated with SJTU since 2005, with the ambitious goal of building a metropolitan-scale traffic information system. The goals of the project are twofold. First, it tries to make the available transportation infrastructure to be used more efficiently. Second, it aims to provide the public with a wide spectrum of ITS applications, ranging from real-time traffic information, trip planning and optimal route selection, to congestion avoidance and bus arrival prediction.

In this chapter, we first introduce three vehicular trace data sets involving tens of thousands of public vehicles collected from the SG project and from Shenzhen, another metropolis in south China. The reason that we collect these data is mainly to better understand vehicular mobility and to conduct informed design of message forwarding algorithms between vehicles. Then, we present the main challenges encountered to process those data for future VANET studies.

2.2 Collecting Vehicular GPS Traces

In the SG project, each experimental vehicle is deployed with a GPS unit and a GPRS wireless communication module. As such a vehicle runs along the roads in the city, it periodically sends a GPS report back to a data center via a GPRS channel. Due to the GPRS communication cost for data transmission, reports are



Fig. 2.1 A taxi with a commercial GPS device installed (*highlighted in the inset*), the location and operational information thus can be periodically sent back via GPRS wireless channels

usually sent at rather large intervals, typically once per minute. We have collected three datasets consisting of GPS traces of buses and taxis from two cities in China:

Shanghai Taxis: We collected the GPS trace of taxis in Shanghai collected between Feb 1 and Mar 3, 2007. We chose 2,109 taxis in the datasets which have consecutive GPS reports on each day during the 31 days. The specific information contained in such a report includes: ID, the longitude and latitude coordinates of the current location, timestamp, moving speed, and heading direction. In addition, the information contained in a taxi GPS report also reports whether passengers are onboard. The granularity of reports is 1 min for taxis with passengers and about 15 s for vacant ones. Figure 2.1 illustrates an experimental taxi in Shanghai. In Fig. 2.2, the geometry of destinations of all taxi deliveries on Shanghai map during Feb of 2007 is shown, where every colored dot presents the average number of destinations per taxi per day located in the corresponding $300\text{ m} \times 300\text{ m}$ square area on the map.

Shanghai Buses: The trace consists of GPS reports sent from 2,501 buses which serve on 199 routes and cover the main downtown area (within Neihuan Viaducts about 120 km^2) between Feb 19 and Mar 5, 2007. Figure 2.3 shows the coverage of all experimental bus routes. A commuting bus periodically sends GPS reports back to a backend data center via GPRS channel. The information contained in a report is similar to that of taxis except that there are more fields contained, such as whether the bus is arriving at a stop or a terminal is also sent. Due to the GPRS communication cost for data transmission, reports are sent at a granularity of around 1 min.

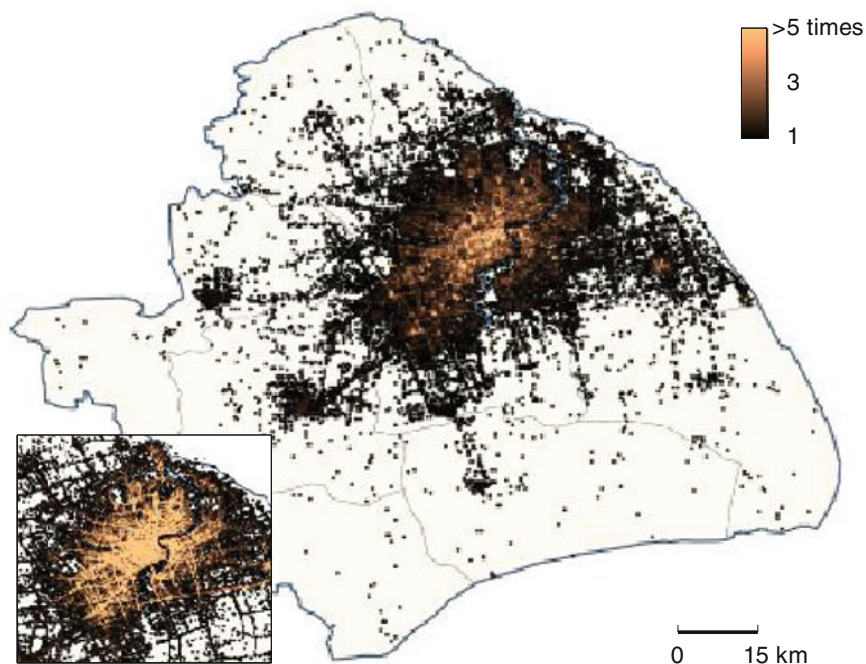


Fig. 2.2 The geometry of destinations of all taxi deliveries on Shanghai map during Feb 2007. Every *colored dot* presents the average number of destinations per taxi per day located in the corresponding $300\text{ m} \times 300\text{ m}$ square area on the map

Fig. 2.3 The distribution of bus lines within the downtown area of Shanghai city, with 199 bus lines denoted by *red lines*



Shenzhen Taxis: We also collected the GPS trace of taxis in Shenzhen in October, 2009. The data format is similar to that of Shanghai taxi trace. We chose 8,291 taxis which continuously send GPS reports during the whole period. Taxis in Shenzhen always send GPS reports on every 1 min. Figure 2.4 demonstrates the

Fig. 2.4 The geographical distribution of GPS reports from all experimental taxis in Shenzhen on Oct 1, 2009

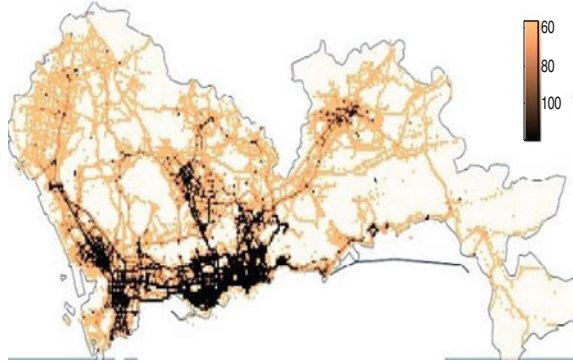


Table 2.1 Comparison of three data sets

Data set	Shanghai bus	Shanghai taxi	Shenzhen taxi
Number of vehicles	2,501	2,109	8,291
From date	Feb 19, 2007	Feb 1, 2007	Oct 1, 2009
Duration (day)	15	31	31
Granularity (second)	60	15 [*] , 60 ^{**}	60
Number of contacts	1,229,380	22,053,178	23,968,860
Mean ICT (minute)	31.8	47.6	30.5

^{*} Vacant

^{**} Passengers onboard

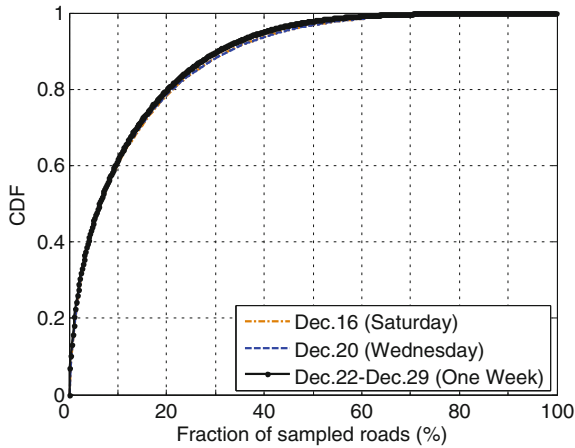
geographical distribution of GPS reports from all experimental taxis on October 1, 2009, where every colored dot presents the average number of reports per taxi per day located in the corresponding $300\text{ m} \times 300\text{ m}$ square area on the map.

We choose taxis and buses to study for two reasons. First, taxis and buses shows two distinct mobility patterns, namely, rather random and well scheduled, respectively. Second, the privacy problem is less concerned since we use public vehicles. As privacy preservation schemes progress and more mobility data of private vehicles available, it is invaluable to study private vehicles in the future. Key statistics of the traces are listed in Table 2.1.

2.3 Challenges and Issues in Data Analysis

To study VANETs in urban scenarios, it is ideal to collect GPS reports for a sufficient long period of time from various types of vehicles 24 h a day with a granularity measured in seconds. In practice, due to the deployment and communication costs and privacy issues, we only collect GPS reports from public vehicles, i.e., taxis and buses with a granularity measured in about 1 min. As a result, the data sets are very sparse in terms of temporal and spatial distributions.

Fig. 2.5 CDFs of GPS sample density at each road



We first examine the geographic distribution of GPS data. For example, from Fig. 2.2, it can be seen that most of the GPS samples are scattered in the downtown area where taxis congregate more densely than in suburbs. The cumulative distribution functions (CDF) of sample density on each road are shown in Fig. 2.5. The data are taken on a weekend, on a workday and for a whole week, respectively. We observe an obvious Pareto distribution in which the “80-20 rule” [27] stands (i.e., 20 % of the road segments owns 80 % of the data).

We then examine the distribution of taxi GPS data in time dimension. We are interested in the probability distribution of the inter-report times, which refers to the time intervals between any two consecutive reports received from a location over time. Figure 2.6 shows the complementary cumulative distribution function (CCDF) of inter-report times. It can be seen that the middle part of the CCDF is almost linear in log–log scale, which indicates a power law. This means a location may frequently has no sensory data for a long time. Figure 2.7 shows the CCDFs

Fig. 2.6 CCDF of inter-report times

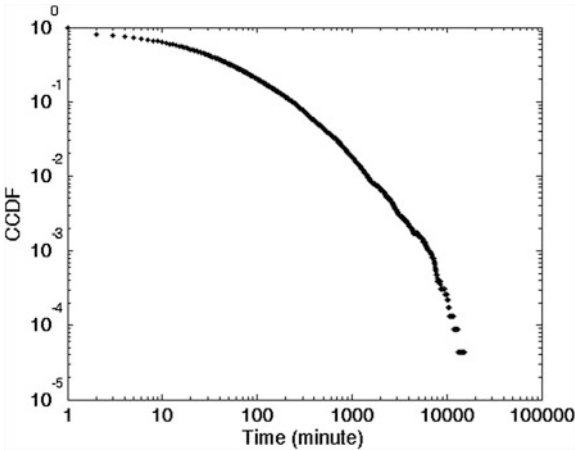
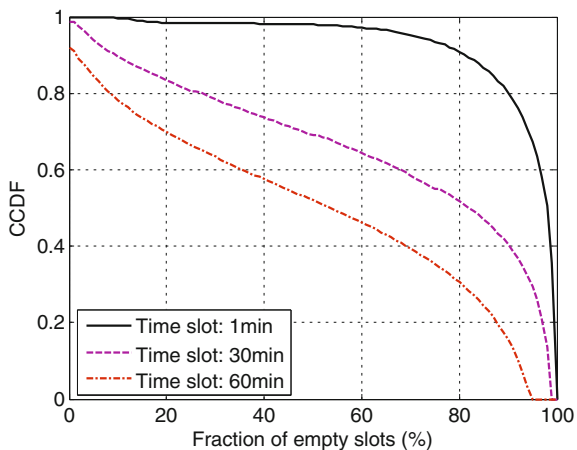


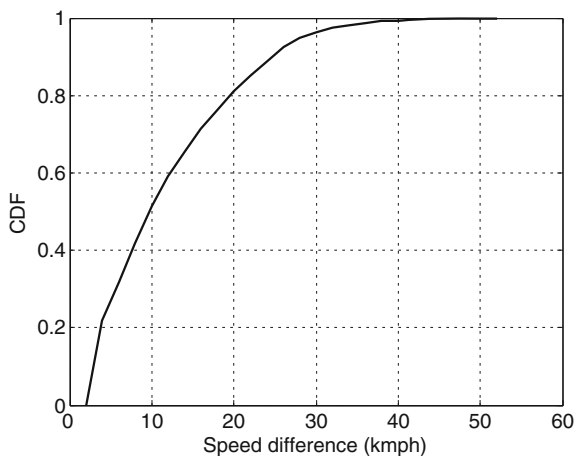
Fig. 2.7 CCDFs of the proportion of time with no sensory data



of the proportion of time with no sensory data in a day in different observation granularities. The time windows used to collect sensory reports are 1, 30 and 60 min, respectively. It shows that about 90 % of roads have no samples in 80 % of the 1,440 min in a day. The fraction is about 50 % when counting the number of road segments that are short of samples for 12 h in a day.

Besides the sparseness, the collected GPS trace data are also erroneous with noise. In the city setting with dense high buildings and viaducts, the error of GPS reports from taxis can be as large as 100 m. To tell which road a taxi is actually monitoring, we need to recover each sample back on track. We deal with this problem using our map-matching algorithm. More specifically, the algorithm prefers those roads with minimum projection distance from the report and minimum angle deviation between the heading direction of the taxi and the road. This simple yet effective strategy works well in most situations and can gain very high accuracy compared with real itineraries. In more complicated cases where the

Fig. 2.8 CDF of speed difference at the same location at the same time



geographical distance between these two consecutive records could exceed thousands of meters, we need to consider the mobile context of the taxi. The algorithm examines several previous and successive reports to determine the most possible road segment where the report issued. Our on-road experiment results show that our map-matching algorithm can reach about 98 % accuracy with the left regarded as an inevitable source of noise.

In addition, we also find that individual reports vary significantly even they are collected from the same location at the same time. Figure 2.8 shows the CDF of speed difference derived from reports at the same location at the same time. It can be seen that the CDF increases slowly with a relatively long tail, which implies the individual reports can vary largely. The derivation of this variance may be ascribed to individual driving behavior. For example, a taxi may stop arbitrarily to pick up or drop passengers. In other words, each sensory data report is associated with a certain degree of noise. Despite these inaccuracies, the GPS trace data are very valuable to study VANETs since they cover thousands of vehicles and last for 1 month.



<http://www.springer.com/978-1-4614-8047-1>

Studies on Urban Vehicular Ad-hoc Networks

Zhu, H.; Li, M.

2013, IX, 124 p. 78 illus., Softcover

ISBN: 978-1-4614-8047-1