A comparative analysis of intra-city human mobility by taxi

Wenjun Wang\textsuperscript{a,b}, Lin Pan\textsuperscript{a,b,*}, Ning Yuan\textsuperscript{a,b,c}, Sen Zhang\textsuperscript{d}, Dong Liu\textsuperscript{a,b,e}

\textsuperscript{a} School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
\textsuperscript{b} Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300072, China
\textsuperscript{c} Department of Basic Courses, Academy of Military Transportation, PLA, Tianjin 300161, China
\textsuperscript{d} Department of Math, Computer Science and Statistics, State University of New York College at Oneonta, Oneonta, NY 13820, USA
\textsuperscript{e} School of Computer and Information Technology, Henan Normal University, Xinxiang 453007, China

\section*{Highlights}
- We observe common patterns of human mobility by taxi in several cities.
- The displacement distributions of taxi trips tend to follow exponential laws.
- The trip durations follow log-normal distributions.
- The interevent time distributions have log-normal bodies followed by power law tails.
- Airports attract large amounts of taxi traffic at a certain distance level.

\section*{Article Info}

\textbf{Article history:} received 8 April 2014, received in revised form 26 October 2014, available online 6 November 2014

\textbf{Keywords:} Human mobility, Urban traffic, Taxi trajectory, Exponential distribution, Log-normal distribution

\section*{Abstract}
Quantitative understanding of human movement behaviors would provide helpful insights into the mechanisms of many socioeconomic phenomena. In this paper, we investigate human mobility patterns through analyzing taxi-trace datasets collected from five metropolitan cities in two countries. We focus on three statistics for each dataset: the displacement of each occupied trip, the duration of each occupied trip, and the time interval between successive occupied trips by the same taxi (interevent time). The results indicate that the displacement distributions of human travel by taxi tend to follow exponential laws in two displacement ranges rather than power laws; the trip duration distributions can be approximated by log-normal distributions; the interevent time distributions can be well characterized by log-normal bodies followed by power law tails. For each considered measure, the rescaled distributions of all cities collapsed into a master curve. These results provide empirical evidence supporting the common regularity of intra-city human mobility. Moreover, we show that airport locations could play a role in explaining the spikes of displacement distributions of taxi trips in certain cities.

\section*{1. Introduction}

Understanding, modeling and predicting human spatio-temporal movements is an active research subject relevant to urban studies [1], traffic engineering [2], urban planning [3], smart cities [4], epidemics controlling [5,6], emergency management [7,8], location-based services [9], and so on. Traditionally, researchers relied on using and analyzing the human movements data collected from performing travel surveys or observations [10,11]. More recently, researchers have been able...
to have access to massive individual movement data, including GPS traces of vehicles [12–15] and human beings [16,17], mobile phone records [18–20] and check-ins of online social network accounts [9,21]. These data have helped researchers to uncover the patterns of human mobility from more perspectives and gain deeper understanding of their underlying mechanisms than before.

A number of recent studies have found that statistical patterns of human mobility exhibit characteristics similar to Lévy flight [16–18,22], a type of random walk with the lengths of the steps following a certain probability distribution and the directions of the steps being isotropic and random. To be more specific, power law distribution was found in the dispersal of bank notes [22]. A power law with an exponential cutoff can be used to approximate the displacement distribution of human trajectories obtained from mobile phone datasets [18,20], GPS traces [16,17,23] and online location-based social networks [9]. Furthermore, Song et al. found a 93% potential predictability in individual’s daily mobility [24]. In order to explain the observed scaling laws, González et al. suggested a convolution of the population-based heterogeneity and individual Lévy flights [18]. Song et al. developed an individual mobility model which incorporates exploration and preferential return principles [25]. Han et al. demonstrated that the power-law-like displacement distributions may originate from the hierarchical organization of transportation systems [26]. Yan et al. explained the aggregated scaling law under the principle of maximum entropy [27].

The studies mentioned above have provided valuable insights into the patterns and mechanisms of human mobility. However, these work studied human movements at various scales. Since most daily travel activities are bounded by geographic borders [28], routine movements of individuals within cities have also been of a particular interest to the research community. Through the analysis of GPS traces collected from 50 taxicabs, Jiang et al. found scaling properties in trip lengths [12]. However, power law tails were not suggested in many cities by analyzing the check-ins of Foursquare users [21]. Daily mobility lengths by private vehicles [13], taxis [14,15] and mobile phones [19] were reported to follow exponential distributions rather than power laws. Using a taxi dataset in Lisbon, Veloso et al. showed that trip distances can be fitted with a Gamma distribution, while the decreasing interval of trip distances can be fitted with an exponential distribution [29]. A study using a mobile phone dataset in Portugal showed that the commuting distances fit the log-normal distribution [30]. In short, according to various datasets compiled from different cities, many empirical studies have demonstrated that the displacements occurred in the cities deviate from power law distributions.

In this paper, we investigate statistical laws for human mobility by exploring large amounts of GPS traces collected from the taxi transportation system in five metropolitan cities. We focus on discovering common patterns of collective human mobility by taxi at the large city scale. More precisely, we analyze the statistical distributions of trip displacements (i.e., the displacement between pick-up and drop-off locations of an occupied trip), trip durations (i.e., the travel time of an occupied trip) and interevent time (i.e., the time spent between a pick-up and the previous drop-off). The main contributions of this paper can be summarized as follows:

- We comparatively study human mobility patterns in five metropolitan cities by analyzing massive trajectories of GPS equipped taxis.
- We observe that the statistical properties of taxi trips in all the five studied cities display similar behaviors. For each property, the distributions of all datasets collapse after being rescaled, suggesting common patterns of human mobility by taxi in cities.
- We find that the displacements of human movements by taxi approximately follow exponential distributions at the small distance scale (less than the diameter of the main urban area), and the displacements at the large distance scale tend to follow the exponential distributions rather than the power laws.
- We note that travel costs and airport locations should have important effects on human activities of traveling by taxi.
- We show that the trip durations could be well described by a log-normal distribution.
- We observe similar traffic impact in different cities by exploring the correlations between trip displacements and trip durations.
- We also observe log-normal bodies in the distributions of the interevent time between two consecutive taxi services. The tails of the interevent time distributions display similar scaling properties with exponents ranging from 1.52 to 1.60.

2. Datasets

This research uses six taxi datasets containing a total of more than 18 billion GPS records, which were generated by about a total number of 30,000 taxis from four major cities (Beijing, Tianjin, Shanghai and Nanjing) in China and the San Francisco Bay Area in the United States.

Dataset $D_1$ contained the GPS traces of approximately 15% of taxis in Beijing during March 2009. Both $D_2$ and $D_3$ were generated by over 4000 taxis in Tianjin, which are about 12% of the taxis in Tianjin, but respectively, from August 1, 2011 to December 31, 2011 ($D_2$) and from January 1, 2012 to December 31, 2012 ($D_3$). $D_4$ [31,32] was collected from over 4000 taxis in Shanghai, where there were nearly 50,000 taxis. $D_5$ consisted of the GPS traces recorded over a two-year period for almost all the taxis in Nanjing. $D_6$ [33,34] was collected through the cabspotting project [35], and it contained taxicab mobility traces that were captured in the San Francisco Bay Area from May 17 to June 10, 2008. Some basic statistics of the six datasets are summarized in Table 1.
The diameters of the studied cities are all less than 100 km, we simply discarded trips with displacement $\Delta$ between and where $p$ non-smooth trips. Assume a taxi trip is recorded as the datasets do not directly show which trip is valid. Therefore, a filter with a sliding window is applied to detect and remove 3.3.1. Non-smooth trip filter

In consideration of the traffic conditions within urban areas, we also excluded trips with average velocity $\langle v \rangle > 5$ km/h or $\langle v \rangle > 120$ km/h.

### Table 1
Statistics of datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
<th>$D_4$</th>
<th>$D_5$</th>
<th>$D_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>Beijing</td>
<td>Tianjin</td>
<td>Tianjin</td>
<td>Shanghai</td>
<td>Nanjing</td>
<td>San Francisco</td>
</tr>
<tr>
<td>#Effective days</td>
<td>23</td>
<td>144</td>
<td>366</td>
<td>30</td>
<td>619</td>
<td>25</td>
</tr>
<tr>
<td>#Taxis</td>
<td>10 939</td>
<td>4328</td>
<td>4252</td>
<td>4671</td>
<td>8816</td>
<td>536</td>
</tr>
<tr>
<td>#Samples</td>
<td>3.65E+8</td>
<td>1.95E+9</td>
<td>4.54E+9</td>
<td>1.86E+8</td>
<td>1.13E+10</td>
<td>1.12E+7</td>
</tr>
<tr>
<td>Avg. sampling interval (s)</td>
<td>116</td>
<td>26</td>
<td>24</td>
<td>60</td>
<td>32</td>
<td>62</td>
</tr>
<tr>
<td>Avg. sampling distance (m)</td>
<td>218</td>
<td>66</td>
<td>62</td>
<td>133</td>
<td>106</td>
<td>388</td>
</tr>
<tr>
<td>#Occupied trips</td>
<td>3.90E+6</td>
<td>1.15E+7</td>
<td>2.50E+7</td>
<td>2.22E+6</td>
<td>1.20E+8</td>
<td>4.46E+5</td>
</tr>
<tr>
<td>Avg. trip duration (min)</td>
<td>17.23</td>
<td>13.65</td>
<td>13.62</td>
<td>13.92</td>
<td>15.96</td>
<td>11.03</td>
</tr>
<tr>
<td>Avg. trip displacement (km)</td>
<td>5.48</td>
<td>4.18</td>
<td>4.19</td>
<td>4.52</td>
<td>4.18</td>
<td>4.20</td>
</tr>
<tr>
<td>Trips within 20 km</td>
<td>97.7%</td>
<td>99.4%</td>
<td>99.4%</td>
<td>98.6%</td>
<td>99.1%</td>
<td>99.7%</td>
</tr>
</tbody>
</table>

Each GPS record consists of six attributes: taxi id, timestamp, latitude, longitude, instantaneous velocity, and taximeter state (vacant or occupied). The sampling rate of the studied datasets varies according to different GPS tracking systems in these cities as well as the nature of the trips. The taximeter state changes when the taxi picks up and drops off passengers.

### 3. Methods

3.1. Definitions of taxi trip

A taxi trajectory is a sequence of time-stamped points recorded for a given taxi, where each point $p$ contains timestamp $t$, latitude $lat$, longitude $lon$, instantaneous velocity $v$, and taximeter state $s$. A taxi is considered as vacant when there is no passenger in a taxi; or occupied otherwise. A taxi trip is a sub-trajectory with a single state, which could be one of two states: occupied or vacant. Given a taxi trip $p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$, we have $p_i.s = p_j.s$ for all $1 \leq i, j \leq n$. Let $p_i.s = 1$ be the occupied state, and a taxi trip is an occupied trip if for all $i, p_i.s = 1$. We define $p_0 = p_1$ as the pick-up point and $p_D = p_n$ as the drop-off point.

3.2. Measures of taxi trip

The trip displacement $\Delta r$ between $p_0$ and $p_D$ can be measured using the great-circle distance, which gives the shortest distance between the two points on the surface of the planet. The trip duration $\Delta t$ can be calculated as $\Delta t = |t_D - t_0|$. The interevent time $\tau$ represents the time interval between two consecutive occupied trips by the same taxi. It can be estimated via $\tau = |t_D - t'_D|$, where $t'_D$ is the drop-off time of the last occupied trip. In other words, $\tau$ indicates the duration that the taxi is vacant. In this work, $\Delta r$, $\Delta t$ and $\tau$ are measured in kilometers, minutes and minutes, respectively.

3.3. Data cleaning

We extracted raw taxi trips from the GPS samples by detecting the changes of taximeter state, which indicate pick-up/drop-down behaviors. To identify the actual vacant and occupied trips, we first excluded abnormal trajectories which contain a dramatically large number of trips for a time period.

Then we applied a sliding window filter to detect and remove non-smooth trips. Additionally, we discarded trips with duration $\Delta t < 1 \text{ min}$ or $\Delta t > 3 \text{ h}$ because passengers seldom travel by taxi for a few seconds or a great deal of time. As the diameters of the studied cities are all less than 100 km, we simply discarded trips with displacement $\Delta r > 100$ km. In consideration of the traffic conditions within urban areas, we also excluded trips with average velocity $\langle v \rangle < 5 \text{ km/h}$ or $\langle v \rangle > 120$ km/h.

3.3.1. Non-smooth trip filter

While we have access to the pick-up and the drop-off locations, the timestamps and several other crucial information, the datasets do not directly show which trip is valid. Therefore, a filter with a sliding window is applied to detect and remove non-smooth trips. Assume a taxi trip is recorded as $p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$. An indicator function $f(p_i \rightarrow p_j)$ is defined as

$$f(p_i \rightarrow p_j) = \begin{cases} 1, & v(p_i \rightarrow p_j) \leq \theta_1 \\ -1, & v(p_i \rightarrow p_j) > \theta_1 \end{cases},$$

where

$$v(p_i \rightarrow p_j) = \begin{cases} \Delta r/\Delta t, & i \neq j \\ v_i, & i = j \end{cases},$$

and $\theta_1$ a threshold which is predetermined according to the traffic regulation in the city. Here we use the displacement $\Delta r$ between $p_i$ and $p_j$ to approximate the path length. We further calculate the velocity indicator $g_i$ for point $p_i$ over a sliding
window of size $\theta_2$ as follows:

$$g_i = \sum_{j=1}^{\theta_2} f(p_{i-j} \rightarrow p_i) + \sum_{j=1}^{\theta_2} f(p_i \rightarrow p_{i+j}) + f(p_i, p_i).$$

(3)

If $g_i \leq 0$, point $p_i$ is considered to be incorrect. The trip should be excluded as it is invalid if either the total number of incorrect points exceeds a predetermined value $\theta_3$ or there is more than a certain proportion $\theta_4$ of incorrect points. The parameters $\theta_j$ $(j = 2, 3, 4)$ are chosen heuristically.

3.4. Logarithmic binning

Logarithmic binning has been suggested as an appropriate technique to reduce the statistical errors in the tails of empirical power-law-like distributions [36,37]. It is especially recommended for analyzing Lévy flight motions [38]. Logarithmic binning groups the samples into bins of exponentially growing widths. The number of observations which fall in a bin is normalized by the width of the bin.

3.5. Fitness metrics

We fit several commonly used distributions to our datasets, including power law, power law with an exponential cutoff, log-normal, Weibull, and exponential, which are defined in Eqs. (4)–(8), respectively. In practice, the examined heavy-tailed distributions (power law, log-normal and Weibull) are subexponential [39].

A distribution follows a power law with exponent $\beta$ if

$$P(x) \sim x^{-\beta}.$$  

The power law with an exponential cutoff is defined as

$$P(x) \sim x^{-\beta} \exp\left(-\frac{x}{\kappa}\right),$$

where $\beta$ is the exponent and $\kappa$ the cutoff value. The probability density function of a log-normal distribution is

$$P(x) = \frac{1}{\kappa \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right),$$

where $\mu$ and $\sigma$ denote, respectively, the mean and the standard deviation of the natural logarithm of the variable. The Weibull distribution is defined by the equation:

$$P(x) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^k\right),$$

where $k > 0$ is the shape parameter, $\lambda > 0$ the scale parameter. The exponential distribution can be expressed as

$$P(x) \sim \exp(-\lambda x),$$

where $\lambda$ is the rate parameter.

The fitting parameters are estimated by the maximum likelihood method (MLE) [40,41]. The Akaike information criterion (AIC) [40,42] is employed to select the best fitting distributions. The Akaike weights $w_i$ represent the relative likelihoods of models [42]. It is defined as the normalization of model likelihoods:

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{j=1}^{m} \exp(-\Delta_j/2)},$$

where $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$ and $m$ the number of models. The AIC score is calculated by

$$\text{AIC}_i = -2 \log \left( L_i(\hat{\theta}|\text{data}) \right) + 2K_i,$$

where $K_i$ is the number of parameters used in model $i$, $L_i(\cdot)$ the likelihood function, and $\hat{\theta}$ the model parameter estimated by MLE that maximizes the likelihood function.

4. Results

4.1. Trip displacement distribution

The traveling distance is an important measure to describe the travel behavior. Trip displacement is an alternative approach to estimate the distance between the pick-up and drop-off locations. The displacement, which is not influenced by details of the movement route, is widely adopted in mobility studies.
Fig. 1. Distributions of trip displacements in log–log scale. Logarithmic binning with normalization was applied to the data. The red dash–dot, blue solid, green dash–dot, magenta dashed curves denote the tested distribution models (power-law with exponential cutoff, log-normal, Weibull and exponential respectively). Common trends are observed in the six datasets. For each dataset, the full distribution can be divided into two parts by the line at 20 km (10 km for San Francisco). The first part increases at the beginning followed by a period of steadily decreasing, but the second part varies from city to city. The cyan and black dashed lines show the power law fit and the exponential fit for the second part, respectively. We also note that there are interesting spikes in the curves of Shanghai, Nanjing, and San Francisco. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 1 shows the distributions of occupied trip displacements computed from the datasets. From the figure, we can see that the distributions in all datasets display similar trends. It is obvious that power laws cannot accurately capture these distributions. Therefore, we compared four commonly used distributions, which are power law with an exponential cutoff, log-normal, Weibull, and exponential. According to the Akaike test results shown in Table 2, the trip displacement distributions fit best to log-normal and exponential distributions. The corresponding results for the best fittings are reported in Table 3. As can be seen from Table 3, the parameters for Tianjin, Shanghai, and Nanjing are almost identical, and they are not far from the ones for Beijing and San Francisco. However, no model appears to appropriately capture the distribution tails. For each dataset, the exponential law approximates the tail part better than the power law. The measured exponents are reported in Table 4.

In Fig. 1, another interesting observation is that each of the full distributions consists of two different regimes separated by the displacement of about 20 km (10 km for San Francisco). Similar observations were reported by Liang et al. [14] and Noulas et al. [21]. We note that more than 97% of occupied trips are within 20 km in all the studied datasets. There are approximately 90% of occupied trips with displacements less than 10 km in San Francisco. This is fairly consistent with our daily experience. Mainly attributing to the travel cost, individuals seldom take taxis for long distance trips. People typically use subway and other public transportation systems more than taxis for longer distances and more routine intra-city travel purposes. However, long distance taxi trips do happen for many reasons. For example, clients who plan to catch the flights...
or trains for inter-city travels or return from usually exhaustive inter-city travels via airports and train stations, especially when carrying big and/or heavy luggage, tend to use taxis, which quite often result in long distance intra-city trips.

Fig. 2 illustrates the distributions of displacements less than 20 km (10 km for San Francisco). For each city, the probability density function \( P(\Delta r) \) increases rapidly to a considerably high level and then gradually falls down as the trip displacement continues to increase. The peaks of \( P(\Delta r) \) are reached at \( \Delta r = r_m \) where \( r_m \) ranges from 1.3 km to 2.3 km. The rise of \( P(\Delta r) \) is reasonable because we rarely use taxis for very short trips. Besides, \( r_m \) confirms that we will take the travel cost into consideration in daily life. As we know, taxi fares are generally calculated as follows: the minimum fare applied to the taximeter when a passenger gets on (flag fall), plus extra charges for additional distance and waiting time, plus applicable surcharges. The flag fall in the four Chinese cities covers the first 3 km. It is interesting that we can obtain a value around the flag fall distance by simply multiplying \( r_m \) and the detour factor, which is about 1.5 according to Chalasani et al. [43]. San Francisco does not follow this calculation since the flag fall only charges the first 0.2 mile (about 0.322 km).

As plotted in Fig. 2, the best fit for the displacements from \( r_m \) to 20 km (10 km for San Francisco) is exponential. For each dataset, this part of displacements accounts for more than 67% of all the trips. Although the rate parameter \( \lambda \) may vary from city to city, they are quite close, especially among the ones of the four cities in China. Intuitively, this variation could be attributed to many potential variations across cities, such as the geographical structures of urban areas, the spatial distributions of locations of activities, and even the effects of social, economical and cultural factors.

As a final remark, our findings are consistent with the empirical observations reported in Refs. [14, 19, 29]. Veloso et al. [29] explored the taxi GPS traces in Lisbon. They considered the trips within 25 km and fitted the trip displacements with a Gamma distribution. After removing the first interval, they obtained an exponential distribution with \( \lambda = 0.26 \). Liang et al. [14] analyzed two taxi-trace datasets, which were generated by taxis in Beijing during different periods. They illustrated the full distribution of displacements with two ranges of exponential fits. For displacements less than 20 km, they obtained \( \lambda = 0.2329 \) and \( \lambda = 0.2403 \) from their two datasets, respectively. Moreover, the values of rate parameter \( \lambda \) provided by Veloso et al. [29] and Liang et al. [14] are comparable to ours. Using mobile phone datasets, which were collected in eight cities in China, Kang et al. also reported exponential distributions for intra-urban human movements with displacements less than 35 km [19]. To summarize, our study shows that the taxi trip displacements in the five cities tend to follow exponential distributions.

### Table 2

Akaikeweights for the tested distributions of displacements.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Power-law-with-exp-cutoff</th>
<th>Log-normal</th>
<th>Weibull</th>
<th>Exponential</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_1 )</td>
<td>0.0000</td>
<td>0.4778</td>
<td>0.0000</td>
<td>0.5222</td>
</tr>
<tr>
<td>( D_2 )</td>
<td>0.0000</td>
<td>0.9786</td>
<td>0.0000</td>
<td>0.0214</td>
</tr>
<tr>
<td>( D_3 )</td>
<td>0.0000</td>
<td>0.9983</td>
<td>0.0000</td>
<td>0.0017</td>
</tr>
<tr>
<td>( D_4 )</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>( D_5 )</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>( D_6 )</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

### Table 3

Results for the log-normal and exponential fittings of displacements.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Parameters (with 95% confidence bounds)</th>
<th>( \mu ) (LN)</th>
<th>( \sigma ) (LN)</th>
<th>( \lambda ) (EXP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_1 )</td>
<td>1.321 (1.307, 1.334) 0.8683 (0.8522, 0.8845) 0.2146 (0.2109, 0.2184)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_2 )</td>
<td>1.178 (1.171, 1.186) 0.6716 (0.6635, 0.6796) 0.3073 (0.2950, 0.3197)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_3 )</td>
<td>1.174 (1.166, 1.181) 0.6783 (0.6702, 0.6863) 0.3073 (0.2956, 0.3190)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_4 )</td>
<td>1.083 (1.069, 1.097) 0.7949 (0.7791, 0.8105) 0.2942 (0.2890, 0.3250)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_5 )</td>
<td>1.161 (1.149, 1.174) 0.6926 (0.6791, 0.7060) 0.3074 (0.2899, 0.3250)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_6 )</td>
<td>0.769 (0.753, 0.785) 0.6669 (0.6526, 0.6813) 0.4698 (0.4376, 0.5019)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4

Results for the exponential fittings of displacements in different ranges.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>First part</th>
<th>Tail part</th>
<th>( \lambda )</th>
<th>( R^2 )</th>
<th>( \lambda )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_1 )</td>
<td>0.2193 (0.2181, 0.2205) 0.9993</td>
<td>0.1628 (0.1473, 0.1783) 0.9928</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_2 )</td>
<td>0.3253 (0.3204, 0.3303) 0.9961</td>
<td>0.1066 (0.0799, 0.1332) 0.9453</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_3 )</td>
<td>0.3245 (0.3198, 0.3292) 0.9965</td>
<td>0.1103 (0.0841, 0.1366) 0.9502</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_4 )</td>
<td>0.2893 (0.2872, 0.2914) 0.9990</td>
<td>0.0741 (0.0560, 0.0921) 0.9472</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_5 )</td>
<td>0.3402 (0.3357, 0.3448) 0.9971</td>
<td>0.0838 (0.0114, 0.1562) 0.4956</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( D_6 )</td>
<td>0.4953 (0.4776, 0.5131) 0.9882</td>
<td>0.0458 (0.0373, 0.1288) 0.8889</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 2. Distributions of trip displacements in semi-log scale. The vertical lines show the peaks of $P(\Delta r)$. After the peak, these displacements decay following an exponential law for $\Delta r < 20$ km ($\Delta r < 10$ km for San Francisco).

4.2. Airport impacts on displacement distribution

We note that the interesting spikes, which are observed in Fig. 1, occur around the displacements which happen to be close to the displacements from the airports to the downtowns in Shanghai, Nanjing and San Francisco. To our experience, some taxi drivers would like to pick up passengers at airports because there are usually long distance travels along with considerable income. Hence we further examine how taxi airport trips impact the overall displacement distribution.

A taxi airport trip is an occupied trip with the pick-up location or the drop-off location in the airport areas. In this work, we acquired the airport locations through the online map services provided by Google maps.

Fig. 3 compares the displacement distributions for all the taxi trips with the corresponding non-airport parts. As can be seen from Fig. 3, there are great shifts in the spike regimes of displacement distributions before and after the removal of taxi airport trips. This reasonably suggests that the appearance of spikes in Shanghai, Nanjing and San Francisco is due to the relatively high portion of taxi airport trips in those cities. This result also provides empirical evidence that transfer stations of public transportation, such as airports, could play an important role in human mobility.

In addition, there are no spikes and no distinct shifts in Beijing and Tianjin. Intuitively, this variation across cities might be largely due to the urban morphology and the locations of airports. Thus, we further examine the crow-fly distances from the major airports to the centers of the cities. We find that the taxi airport trips are indistinguishable in displacements from daily traveling trips in Beijing and Tianjin. In contrast, the taxi airport trips are significant in Nanjing because the airport is located far from the heart of the city. Shanghai Pudong International Airport, which is further from the city center than the Hongqiao International Airport, contributed most of the taxi airport trips with displacement over 30 km. The displacement between
downtown San Francisco and the major airports, either San Francisco International Airport or Oakland International Airport, is around 20 km where the peak occurs. These results suggest that while people tend not to take taxis for long distances possibly in consideration of the total travel cost, they may still choose taxis for airport trips because of the convenience. Further studies are needed to reveal the relationship between airport locations and intra-city human mobility.

4.3. Trip duration distribution

Trip duration $\Delta t$ is another fundamental variable that characterizes human travel behaviors. From the six taxi-trace datasets, we computed the time durations of each occupied trip. Fig. 4 reports the distributions of trip durations. Again, we observed the same trend in all cities. Similar to the displacement distribution, $P(\Delta t)$ increases at the beginning and then decreases. The Akaike test is adopted to select the good fits as well. Table 5 shows the results of the model selection. For each dataset, the trip durations could be well described by a log-normal distribution. Table 6 presents the fitting results. Notice that the fitting parameters are nearly identical for all the five cities, suggesting a common pattern in human travel time. Our results are consistent with the regularities in travel-time budgets [44] and travel-energy budgets [45].

The observed deviations in the tails may be caused by the heterogeneity of long-duration trips. As we know, trip duration can be greatly increased by traffic jams that happen from time to time almost in all large cities. Apart from traffic jams, detours and bad weathers can also increase the trip duration. Nevertheless, long-distance travels, most of which happen between two centers of the city or between urban areas and suburbs, should lead to long-duration trips.
Fig. 4. Distributions of trip durations in log–log scale. Logarithmic binning with normalization was applied to the data. For each dataset, the blue solid line represents log-normal fitting to the overall distribution.

**Table 5**
Akaikeweights for the tested distributions of trip durations.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Akaikeweights</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power-law-with-exp-cutoff</td>
</tr>
<tr>
<td>$D_1$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_2$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_3$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_4$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_5$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_6$</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

**Table 6**
Results for the log-normal fittings of trip durations.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Parameters (with 95% confidence bounds)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>2.656 (2.635, 2.677)</td>
<td>0.6702 (0.6527, 0.6877)</td>
</tr>
<tr>
<td>$D_2$</td>
<td>2.440 (2.432, 2.448)</td>
<td>0.6070 (0.6002, 0.6137)</td>
</tr>
<tr>
<td>$D_3$</td>
<td>2.443 (2.436, 2.451)</td>
<td>0.6089 (0.6026, 0.6152)</td>
</tr>
<tr>
<td>$D_4$</td>
<td>2.447 (2.433, 2.461)</td>
<td>0.6566 (0.6451, 0.6680)</td>
</tr>
<tr>
<td>$D_5$</td>
<td>2.598 (2.593, 2.604)</td>
<td>0.6148 (0.6102, 0.6195)</td>
</tr>
<tr>
<td>$D_6$</td>
<td>2.278 (2.266, 2.289)</td>
<td>0.5149 (0.5054, 0.5244)</td>
</tr>
</tbody>
</table>
4.4. Correlations between trip displacement and trip duration

For each taxi trip, the displacement $\Delta r$ and the duration $\Delta t$ have been measured. Our study shows that, except for the tail parts, both $\Delta r$ and $\Delta t$ can be well approximated by log-normal distributions. In this sense, we examine the relations between them. In the simplest case, the equation $\Delta r = \langle v \rangle \Delta t$ holds for straight line motion with constant speed. However, a number of factors, such as road networks, traffic jams, incidents, and weather conditions, should affect taxi trajectories. By comparing the trip displacement distribution and the trip duration distribution, we found that taxi data, in some sense, could reflect the traffic conditions in different cities.

As shown in Fig. 5(a), the average displacement $\langle \Delta r \rangle$, which corresponds to taxi trips with the same duration $\Delta t$, increases rapidly as $\Delta t$ reaches to a considerably high level and then varies in different datasets. When $\Delta t$ goes to extremely large, $\langle \Delta r \rangle$ changes much slower, but with no particular trend observed. It increases, decreases and sometimes fluctuates with up and down jumps. The interesting jumps, especially for dataset $D_6$, which has the smallest sample sizes, may be caused by the relatively small number of long-duration trips in our datasets. These results can be partly explained by traffic jams that result in long-duration trips with small average speed. Another possible reason is that sometimes we have to take a detour due to the traffic constraints, for example, bridge closures or road closures. Except for dataset $D_6$, the rapidly increasing interval can be well approximated by $\langle \Delta r \rangle = k \Delta t$ with $k$ range from 0.22 to 0.32. Linear correlation generally holds for the four Chinese cities, but not San Francisco. This can be partially explained by that less highways are available or less highway advantage can be taken in large Chinese cities than San Francisco. Due to historical reasons, passing-through city highways are limited in these large Chinese cities, despite inter-city highways system has been rapidly deployed. This means that in these large Chinese cities, a longer travel distance generally means longer local rides and thus leads to longer trip time. As for San Francisco, a longer travel distance may lead the driver to take highway with faster speed thus resulting in relatively shorter time.

Fig. 5(b) shows that the average duration $\langle \Delta t \rangle$, which corresponds to taxi trips with the same displacement $\Delta r$, increases with $\Delta r$. The growth rate of $\langle \Delta t \rangle$ is a measure of the average speed. A slower growth rate implies a higher average speed. As shown in Fig. 5(b), it tends to grow slower especially when $\Delta r > 20$ km ($\Delta r > 10$ km for dataset $D_6$). A possible explanation for this phenomenon could be that large-displacement trips may be less affected by traffic jams in urban areas. To our experience, taxi trips with large displacements usually happen between urban areas and suburbs, or between two centers of the city (e.g., the main urban area of Tianjin and the Binhai New Area). The taxis can take the freeway for the main part of this kind of trips. We also notice that the traffic condition, which is indicated by the time spent on certain displacement, in the San Francisco Bay Area is better than those in the four cities in China according to the studied datasets.

4.5. Interevent time distribution

The previous analysis focused on the occupied trips, which are directly related to the travel demands of passengers. Here, we characterize human travels with a third measure, the interevent time $\tau$, which can reflect travel demands indirectly. After dropping off the passengers, taxis begin to seeking for new passengers. The interevent time, which measures the durations of taxis without passengers, reasonably shows how active human mobility by taxi is. Intuitively, the interevent time is the reciprocal of the capacity dedicated to the taxi services. In a statistical sense, shorter interevent time means more taxi demands.

Fig. 6 presents the distributions of interevent time. Table 7 summarizes the results of the model selection using AIC method. When the interevent time is less than 200 min, the log-normal model performs best in all the six datasets. Table 8 shows the parameters of the best fittings. The heavy-tailed distribution of the interevent time suggests the high heterogeneity of taxi waiting time for new passengers. The results demonstrate that the processes of serving passengers
consist of bursts of intensive activity alternating with long periods of inactivity. In Fig. 6, we also notice that the tails decay slower in all cities, and even display an up-tail phenomenon after about ten hours in three cities: Beijing, Tianjin, and Shanghai. These interesting tails might be attributed to several reasons. First, the drivers need to take a rest after a period of driving. According to everyday experience, ten hours is a typical length for sleeping and relaxing. Second, compared with daytime, fewer travels take place in cities at late night. Therefore, long interevent time is often observed during nighttime. Besides, after drop-offs in the suburbs, the taxi drivers tend to move to locations that have more chances of picking up new customers, such as airport, railway station and downtown, even if it means longer vacancy time.

### Table 7

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Akaikeweights</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power-law-with-exp-cutoff</td>
</tr>
<tr>
<td>$D_1$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_2$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_3$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_4$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_5$</td>
<td>0.0000</td>
</tr>
<tr>
<td>$D_6$</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Fig. 6. Distributions of interevent time in log–log scale. Logarithmic binning with normalization was applied to the data. The red dash–dot, blue solid, green dash–dot, magenta dashed lines indicate power-law with exponential cutoff, log-normal, Weibull and exponential models respectively, which apply to $\tau < 200$ min. The black dashed lines show power law fitting for $\tau > 30$ min. In (a)–(f), the scaling exponents are 1.52, 1.59, 1.60, 1.59, 1.60 and 1.57, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Fig. 7. Distributions and rescaled ones of all the six datasets. (a) and (d) plot the displacement distributions of non-airport trips, while (b) and (e) use all the occupied trips. In (d), (e), (f), each distribution is rescaled with the average trip displacement $\langle \Delta r \rangle$, the average trip duration $\langle \Delta t \rangle$, the average interevent time $\langle \tau \rangle$ of the respective dataset. After being rescaled, the curves of different datasets collapse.

4.6. Rescaling

In this section, we compare the studied measures of human mobility among different datasets. The rescaling method [18] was used to eliminate the spatial and temporal heterogeneity. We considered the relative indicators $\Delta r / \langle \Delta r \rangle$, $\Delta t / \langle \Delta t \rangle$ and $\tau / \langle \tau \rangle$, where $\langle \Delta r \rangle$, $\langle \Delta t \rangle$ and $\langle \tau \rangle$ are, respectively, the average trip displacement, the average trip duration and the average interevent time of the corresponding dataset. Here we did not use logarithmic binning because model fitting was not necessary.

As shown in Fig. 7, the rescaled distributions of trip displacement, trip duration and interevent time all collapsed into their respective master curves. It is interesting that the interevent time distributions already collapsed well before being rescaled, suggesting that a single interevent time distribution characterizes all datasets, independent of where they were generated. The results suggest that human mobility by taxis in all the studied cities could be characterized by the same model. Recent works have proposed several universal commuting models, such as the radiation model [46], the rank-distance model [21], and the modified doubly-constrained gravity model [47]. In the future, we plan to apply these models to check if they hold at the large city scale.

5. Conclusions

In this paper, we have empirically analyzed intra-city human mobility patterns by exploring six taxi datasets, which were collected from four Chinese cities (Beijing, Tianjin, Shanghai and Nanjing) and one American city (San Francisco). We
observed similar statistical regularities in the studied datasets. As we have shown, the distribution of the trip displacements tends to have an exponential body, and the distribution tail is better approximated by the exponential law than the power law. The trip durations follow a log-normal distribution, probably caused by the average travel time budgets. The distribution of the interevent time can be well approximated by a log-normal body along with a power law tail. The heavy-tailed distribution of the interevent time suggests the heterogeneity of time spent waiting for new passengers, which can be applied to validate the bursty behavior in human daily activities. The collapse phenomena observed in rescaled distributions confirm that there are common patterns of human movements within cities. We also noted that travel costs and airport locations could play important roles in shaping the displacement distribution of taxi trips.

We expect that the empirical results we have obtained from this research will help enrich our knowledge of human mobility for universal models of human mobility by taxi and further enhance our understanding to human mobility by taxi, at the large city scale. We are aware that taxi is only one mode of transport, while individuals travel in a variety of ways in practice. In the future, we plan to conduct more studies to examine alternative modes of transport. We would also like to further investigate human movement patterns at the large city scale by integrating taxi traces and other datasets, for example, transactions of smart cards, location data from mobile phones and check-ins of online social networks. Further research into the modeling and prediction of intra-city human mobility should be pursued. It would be interesting to explain why certain statistical distributions emerge at the large city scale from a modeling perspective.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions. We also thank the Wireless and Sensor networks Lab (WnSN) at Shanghai Jiao Tong University for providing the mobility traces of taxis in Shanghai. This work was partially supported by the National Natural Science Foundation of China under Grant No. 91224009, the National Key Technology R&D Program of the Ministry of Science and Technology of China under Grant Nos. 2012BAK03B06 and 2013BAK02B06, the Key Technology R&D Program of Tianjin under Grant Nos. 12CZDSF07200 and 13CZGDGX01099.

References


Table 8
Results for the log-normal fittings of interevent time.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Parameters (with 95% confidence bounds)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>$\mu = 2.156$, $\sigma = 1.578$</td>
<td>0.9836</td>
</tr>
<tr>
<td>$D_2$</td>
<td>$\mu = 1.956$, $\sigma = 1.469$</td>
<td>0.9944</td>
</tr>
<tr>
<td>$D_3$</td>
<td>$\mu = 1.913$, $\sigma = 1.475$</td>
<td>0.9945</td>
</tr>
<tr>
<td>$D_4$</td>
<td>$\mu = 1.876$, $\sigma = 1.377$</td>
<td>0.9956</td>
</tr>
<tr>
<td>$D_5$</td>
<td>$\mu = 1.882$, $\sigma = 1.551$</td>
<td>0.9923</td>
</tr>
<tr>
<td>$D_6$</td>
<td>$\mu = 2.200$, $\sigma = 1.387$</td>
<td>0.9942</td>
</tr>
</tbody>
</table>