Multi-scale spatio-temporal analysis of human mobility

Alessandretti, Laura; Sapiezynski, Piotr; Jørgensen, Sune Lehmann; Baronchelli, Andrea

Published in:
P L o S One

Link to article, DOI:
10.1371/journal.pone.0171686

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Multi-scale spatio-temporal analysis of human mobility

Laura Alessandretti¹, Piotr Sapiezynski², Sune Lehmann²,³, Andrea Baronchelli¹*

¹ City, University of London, London EC1V 0HB, United Kingdom, ² Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark, ³ Niels Bohr Institute, University of Copenhagen, DK-2100 København Ø, Denmark

* Andrea.Baronchelli.1@city.ac.uk

Abstract

The recent availability of digital traces generated by phone calls and online logins has significantly increased the scientific understanding of human mobility. Until now, however, limited data resolution and coverage have hindered a coherent description of human displacements across different spatial and temporal scales. Here, we characterise mobility behaviour across several orders of magnitude by analysing 850 individuals’ digital traces sampled every 16 seconds for 25 months with 10 meters spatial resolution. We show that the distributions of distances and waiting times between consecutive locations are best described by log-normal and gamma distributions, respectively, and that natural time-scales emerge from the regularity of human mobility. We point out that log-normal distributions also characterise the patterns of discovery of new places, implying that they are not a simple consequence of the routine of modern life.

Introduction

Characterising the statistical properties of individual trajectories is necessary to understand the underlying dynamics of human mobility and design reliable predictive models. A trajectory consists of displacements between locations and pauses at locations, where the individual stops and spends time (Fig 1). Thus, the distribution of waiting times (or pause durations), Δt, between movements and the distribution of distances, Δr, travelled between pauses are often used to quantitatively assess the dynamics of human mobility. For example, specific probability distributions of distances and waiting times characterise different types of diffusion processes. Thanks to the recent availability of data used as proxy for human trajectories including mobile phone call records (CDR), location based social networks (LBSN) data, and GPS trajectories of vehicles, the characteristic distributions of distances and waiting times between consecutive locations have been widely investigated. There is no agreement, however, on which distribution best describes these empirical datasets.

Pioneer studies, based on CDR [1, 2] and banknote records [3], found that the distribution of displacement ∆r is well approximated by a power-law, \( P(\Delta r) \sim \Delta r^{-\beta} \), (or ‘Lévy distribution’ [4], as typically 1 < \( \beta \) < 3), and that an exponential cut-off in the distribution may control boundary effects [2]. These findings were confirmed by studies based on GPS trajectories of
individuals [5–7] and vehicles [8, 9], as well as online social networks data [10–12]. It has been noted, however, that power-law behaviour may fail to describe intra-urban displacements [13]. Other analyses, based on online social network data [14–16] and GPS trajectories [17–20] showed that the distribution of displacements is well fitted by an exponential curve, 

\[ P(\Delta r) \sim e^{-\lambda \Delta r}, \]

in particular at short distances. Finally, analyses based on GPS on Taxis [21, 22] suggested that displacements may also obey log-normal distributions,

\[ P(\Delta r) \sim \left[ \frac{1}{\Delta r} \right] e^{-\left( \log \Delta r - \mu \right)^2/2\sigma^2}. \]

In Ref. [6], the authors found that this is the case also for single-transportation trips.

Fewer studies have explored the distribution of waiting times between displacements, \( \Delta t \), as trajectory sampling is often uneven (e.g., in CDR data location is recorded only when the phone user makes a call or texts, and LBSN data include the positions of individuals who actively “check-in” at specific places). Analyses based on evenly sampled trajectories from mobile phone call records [1, 23], and individuals GPS trajectories [5, 7] found that the distribution of waiting times can be also approximated by a power-law. A recent study based on GPS trajectories of vehicles, however, suggests that for waiting times larger than 4 hours, this distribution is best approximated by a log-normal function [24]. Several studies have highlighted the presence of natural temporal scales in individual routines: distributions of waiting times display peaks that corresponds to the typical times spent home on a typical day (~ 14 hours) and at work (~ 3 – 4 hours for a part-time job and ~ 8 – 9 hours for a full-time job) [23, 25, 26].

Fig 2 and Table 1 compare distributions obtained using different data sources. The spectrum of results reflects the heterogeneity of the considered datasets (see Fig 2). It is known in fact that data spatio-temporal resolution and coverage has an important influence on the results of the analyses performed [27–29].
First, the datasets considered have different spatial resolution and coverage, and few studies have so far considered the whole range of displacements occurring between $\sim 10$ and $10^7$ m (10000 km) (Fig 2). Our analysis suggests that constraining the analysis to a specific distance range may result in different interpretations of the distributions. Another difference concerns the temporal sampling in the datasets analysed so far. Uneven sampling typical of CDR and LBSN data (i) does not allow to distinguish phases of displacement and pause, since individuals could be active also while transiting between locations, and (ii) may fail to capture patterns other than regular ones [31, 32], because individuals’ voice-call/SMS/data activity may be higher in certain preferred locations. Finally, studies focusing on displacements effectuated using one or several specific transportation modality (private car [24, 33], taxi [20], public transportation [34], or walk [7]) capture only a specific aspect of human mobility behaviour.

In this paper, we analyse mobility patterns of $\sim 850$ individuals involved in the Copenhagen Network Study experiment for over 2 years [35]. Individual trajectories are determined combining GPS and Wi-Fi scans data resulting in a spatial resolution of $\sim 10$ m and even sampling every $\sim 16$ s. Trajectories span more than $\sim 10^7$ m. Previous studies with comparable spatial coverage (Fig 2) relied on single-transportation modality data [8], unevenly sampled data [16], or small samples (32 individuals in Ref. [5]). To our knowledge, the Copenhagen Network Study data has the best combination of spatio-temporal resolution and sample size among the datasets analysed in the literature to date (see Methods).

Results

We consider an individual to be pausing when he/she spends at least 10 consecutive minutes in the same location, and moving in the complementary case. In the following, we refer to locations as places where individuals pause. The distribution of displacements is robust with respect to variations of the pausing parameter (see S1 File for the results obtained with 15 and 20 minutes pausing).

We start by considering the three distributions most frequently reported in the literature (Table 1), namely

- The log-normal distribution of a random variable $x$, with parameters $\sigma$ and $\mu$, defined for $\sigma > 0$ and $x > 0$, with probability density function:
  \[
  P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\log x - \mu)^2}{2\sigma^2}} \tag{1}
  \]

- The Pareto distribution (i.e. power-law) of a random variable $x$, with parameter $\beta$, defined for $x \geq 1$, and $\beta > 1$, with probability density function:
  \[
  P(x) = (\beta - 1)(x)^{-\beta} \tag{2}
  \]
Table 1. Distribution of waiting times and displacements: A comparison of over 30 datasets on human mobility. The table reports for each dataset: the reference to the journal article/book where the study was published, the type of data (LBSN stands for Location Based Social Networks, CDR for Call Detail Record), the number of individuals (or vehicles in the case of car/taxi data) involved in the data collection, the duration of the data collection (M → months, Y → years, D → days, W → weeks), the minimum and maximum length of spatial displacements, the shape of the probability distribution of displacements with the corresponding parameters, the temporal sampling, the shape of the distribution of waiting times with the corresponding parameters. Power-law (T), indicates a truncated power-law. The table can also be found at http://lauralessandretti.weebly.com/plosmobilityreview.html.

<table>
<thead>
<tr>
<th>Data type</th>
<th>N</th>
<th>Dur.</th>
<th>Range</th>
<th>$P(\Delta x)$</th>
<th>Sampling $\delta t$</th>
<th>$P(\Delta t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1] (D1)</td>
<td>CDR</td>
<td>3.0 - $10^6$</td>
<td>1 Y</td>
<td>1 km 100 km</td>
<td>$\text{power-law (T)}$ $\beta = 1.55$</td>
<td>uneven</td>
</tr>
<tr>
<td>[1] (D2)</td>
<td>CDR</td>
<td>$10^3$</td>
<td>2 W</td>
<td>1 km 100 km</td>
<td>$\text{power-law (T)}$ $\beta = 1.75$</td>
<td>1 h $\text{power-law (T)}$ $\beta = 1.80$</td>
</tr>
<tr>
<td>[2] (D1)</td>
<td>CDR</td>
<td>$10^5$</td>
<td>6 M</td>
<td>1 km 1000 km</td>
<td>$\text{power-law (T)}$ $\beta = 1.75$</td>
<td>uneven</td>
</tr>
<tr>
<td>[2] (D2)</td>
<td>CDR</td>
<td>206</td>
<td>1 W</td>
<td>1 km 500 km</td>
<td>$\text{power-law (T)}$ $\beta = 1.75$</td>
<td>2 h</td>
</tr>
<tr>
<td>[3]</td>
<td>Bills records</td>
<td>4.6 - $10^5$</td>
<td>1.39 Y</td>
<td>100 m 3200 km</td>
<td>$10 \leq \Delta x &lt; 3200 \text{ km}$ $\text{power-law}$ $\beta = 1.59$</td>
<td>uneven</td>
</tr>
<tr>
<td>[5] (Geolife)</td>
<td>GPS</td>
<td>32</td>
<td>3.42 Y</td>
<td>10 m 10000 km</td>
<td>$0.01 \leq \Delta x &lt; 10 \text{ km}$ $\text{power-law}$ $\beta_0 = 1.25$</td>
<td>$2 \text{ min}$ $\text{power-law}$ $\beta = 1.98$</td>
</tr>
<tr>
<td>[6] (Nokia)</td>
<td>GPS</td>
<td>200</td>
<td>1.50 Y</td>
<td>100 m 10 km</td>
<td>$\text{power-law (T)}$ $\beta = 1.39$</td>
<td>10 sec</td>
</tr>
<tr>
<td>[6] (Geolife)</td>
<td>GPS</td>
<td>182</td>
<td>5.00 Y</td>
<td>100 m 10 km</td>
<td>$\text{power-law (T)}$ $\beta = 1.57$</td>
<td>$1 - 5 \text{ sec}$</td>
</tr>
<tr>
<td>[7] (5 datasets)</td>
<td>GPS</td>
<td>101</td>
<td>5 M</td>
<td>10 m 10 km</td>
<td>$\text{power-law (T)}$ $\beta = [1.35-1.82]$</td>
<td>10 sec $\text{power-law}$ $\beta = [1.45-2.68]$</td>
</tr>
<tr>
<td>[8]</td>
<td>Taxi (GPS)</td>
<td>50</td>
<td>6 M</td>
<td>1 km 100 km</td>
<td>$3 \leq \Delta x &lt; 23 \text{ km}$ $\text{power-law}$ $\beta_0 = 2.50$</td>
<td>10 sec</td>
</tr>
<tr>
<td>[9]</td>
<td>Taxi (GPS)</td>
<td>6.6 - $10^3$</td>
<td>1 Y</td>
<td>1 km 100 km</td>
<td>$\text{power-law (T)}$ $\beta = 1.20$</td>
<td>10 sec</td>
</tr>
<tr>
<td>[10]</td>
<td>Flickr</td>
<td>4.0 - $10^4$</td>
<td>1 km</td>
<td>100 km</td>
<td>$\text{power-law (T)}$ $\beta = 1.88$</td>
<td>uneven</td>
</tr>
<tr>
<td>[11]</td>
<td>LBSN</td>
<td>2.2 - $10^5$</td>
<td>4 M</td>
<td>1 km 500 km</td>
<td>$\text{power-law}$ $\beta = 1.88$</td>
<td>uneven</td>
</tr>
<tr>
<td>[12]</td>
<td>Twitter</td>
<td>1.3 - $10^7$</td>
<td>1 Y</td>
<td>1 km 100 km</td>
<td>$\text{power-law}$ $\beta = 1.62$</td>
<td>uneven</td>
</tr>
<tr>
<td>[13]</td>
<td>LBSN</td>
<td>9.2 - $10^5$</td>
<td>6 M</td>
<td>1 km 20000 km</td>
<td>$\text{power-law}$ $\beta = 1.50$</td>
<td>uneven</td>
</tr>
<tr>
<td>[13] (intracity)</td>
<td>LBSN</td>
<td>9.2 - $10^5$</td>
<td>6 M</td>
<td>10 m 100 km</td>
<td>$\text{power-law}$ $\beta = 4.67$</td>
<td>uneven</td>
</tr>
<tr>
<td>[14]</td>
<td>LBSN</td>
<td>2.6 - $10^5$</td>
<td>1 Y</td>
<td>10 m 50 km</td>
<td>$\text{exponential}$ $\lambda = 0.179$</td>
<td>uneven</td>
</tr>
<tr>
<td>[15]</td>
<td>LBSN</td>
<td>5.2 - $10^5$</td>
<td>1 Y</td>
<td>1 km 4000 km</td>
<td>$\text{exponential}$ $\lambda = 0.003$</td>
<td>uneven</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Data type</th>
<th>N</th>
<th>Dur.</th>
<th>Range (\Delta x)</th>
<th>(P(\Delta x))</th>
<th>Sampling (\delta t)</th>
<th>(P(\Delta t))</th>
</tr>
</thead>
</table>
| [16] Twitter           | 1.6 \(\times\) 10^5 | 8 M  | 10 m 4000 km        | 0.01 \(\leq \Delta x \leq 0.1\) km  
\(\lambda = 0.073\)  
0.1 \(\leq \Delta x \leq 100\) km  
Stretched power-law  
\(\beta_1 = 0.45\)  
100 \(\leq \Delta x \leq 4000\) km  
power-law  
\(\beta_2 = 1.32\) | uneven               |                  |
| [17] Taxi (GPS)        | 803        | 1.25 Y | 1 km 100 km         | \(\Delta x \leq 15\) km  
exponential  
\(\lambda = 0.36\)  
15 \(\leq \Delta x \leq 100\) km  
power-law  
\(\beta = 3.66\) | 30 sec               |                  |
| [18] (D1) Taxi (GPS)   | \(10^4\)  | 3 M  | 1 km 100 km         | 1 \(\leq \Delta x \leq 20\) km  
exponential  
\(\lambda_0 = 0.23\)  
20 \(\leq \Delta x \leq 100\) km  
exponential  
\(\lambda_1 = 0.17\) | 1 min                |                  |
| [18] (D2) Taxi (GPS)   | \(10^4\)  | 2 M  | 1 km 100 km         | 1 \(\leq \Delta x \leq 20\) km  
exponential  
\(\lambda_0 = 0.24\)  
20 \(\leq \Delta x \leq 100\) km  
exponential  
\(\lambda_1 = 0.18\) | 1 min                |                  |
| [19] Taxi (GPS)        | 6.6 \(\times\) 10^3 | 1 W  | 2 km 20 km          | exponential  
\(\lambda = [0.072-0.252]\) | 10 sec               |                  |
| [20] (3 datasets) Taxi (GPS) | \(10^4\)  | 1 M  | 600 m 10 km        | exponential | [9 – 177] s | power-law |
| [21] (6 datasets) Taxi (GPS) | 3.0 \(\times\) 10^4 | [1 M-2 Y] | 1 km 100 km    | log-normal  
\(\mu = [0.77-1.32]\),  
\(\sigma = [0.67-0.87]\) | [24 – 116] s |                  |
| [22] Taxi (GPS)        | 1.1 \(\times\) 10^3 | 6 M  | 100 m 30 km        | log-normal  
\(\mu = 0.38\),  
\(\sigma = 0.48\) | 30 sec               |                  |
| [23] Surveys           | \(10^6\)  | 1 Y  | self-reported      | power-law (T)  
\(\beta = 0.49\) |                  |                  |
| [24] Private Cars (GPS) | 7.8 \(\times\) 10^5 | 1 M  | 1 km 500 km        | superimposition Poisson  
\(\Delta t \leq 4h\)  
power-law  
\(\beta = 1.03\)  
1 \(\leq \Delta t \leq 200\) h  
log-normal  
\(\mu = 1.60\),  
\(\sigma = 1.60\) | 10 sec               |                  |
| [26] Private Cars (GPS) | 3.5 \(\times\) 10^4 | 1 M  | 300 m 100 km       | polynomial | 10 sec | power-law  
\(\beta = 0.97\) |                  |
The exponential distribution of a random variable $x$, with parameter $\lambda$, where $x \geq 0$, and $\lambda > 0$, with probability density function:

$$P(x) = \lambda e^{-\lambda x}$$  \hspace{1cm} (3)

In Eq (2) the probability density can be shifted by $x_0$ and/or scaled by $s$, as $P(x)$ is identically equivalent to $P(y)/s$, with $y = (x - x_0)$. In Eqs (1) and (3), $P(x)$ is identically equivalent to $P(y)$, with $y = (x - x_0)$. In this work, the shift ($x_0$) and scale ($s$) parameters are considered as additional parameters to take into account the data resolution. With few exceptions, the results presented below hold also imposing no shift, $x_0 = 0$ (see S1 File). Note also that Pareto distributions with exponential cut-off (or truncated Pareto) are considered below (see also Table 1).

### Distribution of displacements

We start our analysis by investigating the distribution of displacements between consecutive stop-locations $P(\Delta r)$. First, we consider the overall distribution of the displacements $\Delta r$ using all available data (851 individuals over 25 months). We find that $P(\Delta r)$ is best described by a log-normal distribution (Eq (1)) with parameters $\mu = 6.78 \pm 0.07$ and $\sigma = 2.45 \pm 0.04$, which maximises Akaike Information Criterion (see Methods)—among the three models considered—with Akaike weight $\sim 1$ (Fig 3, see also S1 File).

Second, we investigate if this results holds also for sub-samples of the entire dataset. We bootstrap data 1000 times for samples of 200 and 100 individuals, and we verify that the best distribution is log-normal for all samples, and the average parameters inferred through the bootstrapping procedure are consistent with the parameters found for the entire dataset (see S1 File). In fact, the errors on the value of the parameters reported above are computed by

---

<table>
<thead>
<tr>
<th>Data type</th>
<th>N</th>
<th>Dur.</th>
<th>Range $\Delta x$</th>
<th>$P(\Delta x)$</th>
<th>Sampling $\delta t$</th>
<th>$P(\Delta t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[30] (D1) CDR</td>
<td>$1.3 - 10^6$</td>
<td>1 M</td>
<td>1 km 200 km</td>
<td>power-law $\beta = 2.02$</td>
<td>uneven</td>
<td></td>
</tr>
<tr>
<td>[30] (D2) CDR</td>
<td>$6 - 10^5$</td>
<td>1 Y</td>
<td>1 km 500 km</td>
<td>power-law $\beta = 1.75$</td>
<td>uneven</td>
<td></td>
</tr>
<tr>
<td>[30] (D3) CDR</td>
<td>$4 Y$</td>
<td>1 km 100 km</td>
<td>power-law $\beta = 1.80$</td>
<td>uneven</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[34] Travel cards</td>
<td>$2.0 - 10^6$</td>
<td>1 W</td>
<td>100 m 50 km</td>
<td>negative binomial $\mu = 9.28, \sigma = 5.83$</td>
<td>uneven</td>
<td></td>
</tr>
<tr>
<td>[42] Travel Diaries</td>
<td>230</td>
<td>1.5 M</td>
<td>1 km 400 km</td>
<td>power-law (T) $\beta = 1.05$</td>
<td>self-reported</td>
<td></td>
</tr>
<tr>
<td>[56] Private Cars (GPS)</td>
<td>$7.5 - 10^4$</td>
<td>1 M</td>
<td>10 m 500 km</td>
<td>0.01 $\leq \Delta x \leq 20$ km exponential $20 &lt; \Delta x \leq 150$ km power-law $\beta = 3.30$</td>
<td>30 sec $\Delta t \leq 3h$ exponential $\lambda = 1.02$</td>
<td></td>
</tr>
<tr>
<td>[57] Taxi (GPS)</td>
<td>1 D</td>
<td>200 m 1000 km</td>
<td>power-law $\beta = 2.70$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. (Continued)

doi:10.1371/journal.pone.0171686.t001
bootsraping data for samples of 100 randomly selected individuals. This analysis ensures homogeneity within the population considered, and takes into account also that often smaller sample sizes were analysed in previous literature.

Third, we zoom in to the individual level. We find that the individual distribution of displacements is best described by a log-normal function for 96.2% of individuals. The best distribution is the Pareto distribution for 1.4%, and exponential for the remaining 2.4%. However, the number of data points per individual tend to be significantly lower in group of individuals exhibiting Pareto or exponential distributions, so that one should be cautious in interpreting the observed deviations from a log-normal distribution. Fig 4 reports the histogram of the individual $\mu$ parameters for the 96.2% of the population that is best described by a log-normal distribution, along with three examples of individual distributions.

Finally, we look at large $\Delta r$ in order to compare our results with precedent studies relying on data with larger spatial resolution. We find that limiting the analysis to large values of $\Delta r$ results in the selection of a Pareto distribution (Eq (2)). We identify the threshold $\Delta r^* = 7420$ m as the minimal resolution for which the best fit in $\Delta r < 10^7$ m is Pareto with coefficient $\beta = 1.81 \pm 0.03$ and not log-normal. By bootstrapping 1000 times over samples of 100 individuals we find that $\Delta r^* = 7488.3 \pm 328.2$ m. Thus, power-law distributions describe mobility behaviour only for large enough distances, while mobility patterns including distances smaller than 7420 m are better described by log-normal distributions.
Distribution of waiting times

We now analyse the distribution of waiting times between displacements. The best model describing the distribution of waiting times over all individuals is the log-normal distribution (Eq (1), Fig 5, see also S1 File), with parameters $\mu = -0.42 \pm 0.04$, $\sigma = 2.14 \pm 0.02$. As above, errors are found by bootstrapping over samples of 100 individuals. Also, by bootstrapping we find that the log-normal distribution is the best descriptor for samples of 200 and 100 randomly selected individuals (see S1 File). As in the case of displacements, we find that restricting the analysis to large values of our observable $\Delta t$, and specifically considering only $\Delta t > \Delta t^* = 13$ h, results in the selection of the Pareto distribution (Eq (2), see Fig 5), with coefficient $\beta = 1.44 \pm 0.01$. We find by averaging over 100 samples of 200 individuals that $\Delta t^* = 13.01 \pm 0.12$. Note that the log-normal distribution is selected as the best model also when the analysis is restricted to $\Delta t < \Delta t^*$.
The distribution of waiting times shows also the existence of “natural time-scales” of human mobility. We detect local maxima of the distribution at 14.0, 39.3, 64.8, and 89.9 hours. Hence, 14 hours is the typical amount of time that students in the experiment spent home every day, in agreement with previous analyses on human mobility [23, 25, 26]. Other peaks appear for intervals $\Delta t = 14 + n \cdot 24$, with $n = \{2, 3 \ldots\}$, suggesting individuals spend several days at home. Notice also that the distribution we consider is limited to $\Delta t < 5$ days, an interval much shorter than the observation time-window (about 2 years), a fact that guarantees the absence of possible spurious effects [29]. This limit is imposed to control the cases in which students leave their phones home. The upper bound is arbitrarily set to 5 days; however, we have verified that results are consistent with respect to variations of this choice.

Distribution of displacements between discoveries

Log-normal features also characterise patterns of exploration. We consider the temporal sequence of stop-locations that individuals visit for the first time—in our observational window—and characterise the distributions of displacements between these ‘discoveries’. We find that the distribution of distances between consecutive discoveries $P(\Delta r)$ is best described as a log-normal distribution with parameters $\mu = 6.59 \pm 0.02$, $\sigma = 1.99 \pm 0.01$, (Fig 6, see also S1 File). For $\Delta r > 2800$ m, the best model fitting the distribution of displacements is the Pareto distribution with coefficient $\beta = 2.07 \pm 0.02$. This results are verified by bootstrapping (see S1 File).
Correlations between pauses and displacements

We further investigate the properties of individual trajectories by analysing the correlations between the distance $\Delta r$ and the duration $\Delta t_{\text{disp}}$ characterising a displacement and the time $\Delta t$ spent at destination. **Fig 7A** shows a positive correlation between $\Delta r$ and $\Delta t_{\text{disp}}$ for $\Delta r \gtrsim 300$ m ($p < 0.01$). As $\Delta r$ is the distance between the displacement origin and destination, the absence of correlation at short distances could be due to individuals not taking the fastest route. A positive correlation characterises also the distance $\Delta r$ covered between origin and destination and the waiting time at destination for distances $30 \text{ m} \lesssim \Delta r \lesssim 10^4$ m ($p < 0.01$). Instead, the correlation is negative for distances larger than $5 \times 10^4$ m (**Fig 7B**). This could suggest that individuals break long trips with short pauses. We have verified that these results hold also when individuals’ most important locations (typically including university and home) are removed from the trajectory, implying that these correlations are not dominated by daily commuting.

**Further analysis: Selection of the best model among 68 distributions**

In the previous sections we have restricted the analysis of the distributions of displacements and waiting times to the three functional forms that are most frequently found in the literature. We now repeat the selection procedure considering a list of 68 models (see **S1 File** for the list of distributions) in order to confirm the results described above.

The distributions of displacements and displacements between discoveries are best described by log-normal distributions also when the choice is extended to 68 models, and tails
respectively for $\Delta r > \Delta r^* = 7420$ m and $\Delta r > \Delta r^* = 2800$ m) are better modelled as generalised Pareto distribution, with form:

$$P(x) = (1 + \xi x)^{\frac{1}{\xi}}$$

where $\xi$ is the parameters of the model, such that $x \geq 0$ if $\xi \geq 0$, and $0 \leq x \leq -\frac{1}{\xi}$ if $\xi < 0$.

The best model selected for the whole distribution of waiting time among the 68 models considered is a gamma distribution, defined for $x \in (0, \infty), k > 0$ and $\theta > 0$ as:

$$P(x) = \frac{1}{\Gamma(k) \theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

where $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$. Although the gamma distribution is the best model for the distribution of waiting times (see S1 File for the result of the fit), the presence of natural scales could indicate that the whole distribution may be better described as the composition of several models.
Discussion

Using high resolution data we have characterised human mobility patterns across a wide range of scales. We have shown that both the distribution of displacements and waiting times between displacements are best described by a log-normal distribution. We found, however, that power-law distributions are selected as the best model when only large spatial or temporal scales are considered, thus explaining (at least partially) the disagreement between previous studies. We also showed that log-normal distributions characterise the distribution of displacements between discoveries, implying that this property is not a simple consequence of the stability of human mobility but a characteristic feature of human behaviour. Finally, we have shown that there exist correlations between displacements’ length and the waiting time at destination.

The heavy tailed nature of human mobility has been attributed to various factors, including differences between individual trajectories [36], search optimisation [37–40], the hierarchical organisation of the streets network [41] and of the transportation system [6, 24, 42]. On the other hand log-normal distributions can result from multiplicative [43] and additive [44] processes and describe the inter-event time of different human activities such as writing emails, commenting/voting on online content [45] and creating friendship relations on online social networks [46]. Instead, the distribution of inter-event time in mobile-phone call communication activity can be described as the composition of power-laws [47–49], a feature attributed to the existence of characteristic scales in communication activity such as the time needed to answer a call, as well as the existence of circadian, weakly and monthly patterns. We also find clear signatures of circadian patterns, which could indicate that the whole distribution may be better described as the composition of several models. However, in our case the best description for times including $\Delta \tau < \Delta \tau^*$ is the gamma distribution, which thus is selected both when the whole range of scales is considered and when the analysis is restricted to short times.

Our results come from the analysis of a sample of ~850 University students, which of course represent a very specific sample of the whole population. Nevertheless, it is worth noting that many statistical properties of CNS students mobility patterns are consistent with previous results, such as the distribution of the radius of gyration, the Zipf-like behaviour of individual locations frequency-rank plot, and the power-law tail of the distribution of displacements ($\beta = 1.81 \pm 0.03$ vs. $\beta = 1.75 \pm 0.15$ of [2]). Details are reported in Supplementary Information of [50].

While identifying the mechanism responsible for the observed mobility patterns is beyond the scope of the present article, we anticipate that a more complete spatio-temporal description of human mobility will help us develop better models of human mobility behaviour [24, 51]. Our findings can also help the understanding of phenomena such as the spreading of epidemics at different spatial resolutions, since the nature of heterogeneous waiting times between displacements have a major impact on the spreading of diseases [52].

Methods

Data description and pre-processing

The Copenhagen Network Study data collection took place between September 2013 and February 2016 and involved 851 students of Technical University of Denmark (DTU) in Copenhagen. Data collection was approved by the Danish Data Protection Agency. All participants provided informed consent by filling an on-line consent form and all methods were performed in accordance with the relevant guidelines and regulations. Individual trajectories were inferred combining WiFi scans data and GPS scans data recorded on smartphones handed out to all participants. An anthropological field study included in the 2013 deployment of the experiment reported that participants did not alter their habits due to participation in the CNS experiment.
The WiFi scans data provides a time-series of wireless network scans performed by participants’ mobile devices. Each record \((i, t, SSID, BSSID, RSSI)\) indicates:

- the participant identifier, \(i\)
- the timestamp in seconds, \(t\)
- the name of the wireless network scanned, \(SSID\)
- the unique identifier of the access point (AP) providing access to the wireless network, \(BSSID\)
- the signal strength in dBm, \(RSSI\).

APs do not have geographical coordinates attached, but their position tend to be fixed. The geographical position of APs is estimated the procedure described in S1 File, which used participants’ sequences of GPS scans to obtain APs locations and remove mobile APs. Then, we clustered geo-localised APs to “locations” using a graph-based approach. With our definition, a “location” is a connected component in the graph \(G_d\), where a link exists between two APs if their distance is smaller than a threshold \(d\) (see [50], SI for more details). Here, we present results obtained for \(d = 2\) m. However, results are robust with respect to the choice of the threshold (see also [50]).

Throughout the experiment, participants’ devices scanned for WiFi every \(\Delta t\) seconds. The median time between scans is between \(\Delta t_M = 16\) s and \(\Delta t_M < 60\) s for 90% of the population (see also [50], SI). Data was temporally aggregated in bins of length \(\Delta t = 60\) s, since we focus here on the pauses between moves. If a participant visits more than one location within a time-bin, we assign the location in which they spent the most time to that bin. Given our definition of location and the given time-binning, the median daily time coverage (the fraction of minutes/day that an individual’s position is known, where the median is taken across all days) is included between 0.6 and 0.98 for 90% of the population.

**Model selection**

The best model is selected using Akaike weights [53]. First, we determine the best fit parameters for each of the models via Nelder-Mead numerical Likelihood maximisation [54] (maximisation is considered to fail if convergence with tolerance \(t = 0.0001\) is not reached after \(200 \cdot N\) iterations, where \(N\) is the length of the data). For each model \(m\), we compute the Akaike Information Criterion:

\[
AIC_m = -2 \log L_m + 2V_m + \frac{2V_m(V_m + 1)}{n - V_m - 1}
\]

where \(L_m\) is the maximum likelihood for the candidate model \(m\), \(V_m\) is the number of free parameters in the model, and \(n\) is the sample size. The \(AIC\) reaches its minimum value \(AIC_{\text{min}}\) for the model that minimises the expected information loss. Thus, \(AIC\) rewards descriptive accuracy via the maximum likelihood and penalises models with large number of parameters.

The Akaike \(w_m(AIC)\) weight of a model \(m\) corresponds to its relative likelihood with respect to a set of possible models. Measuring the Akaike weights allows us to compare the descriptive power of several models.

\[
w_m(AIC) = \frac{e^{-\frac{1}{2}(AIC_m - AIC_{\text{min}})}}{\sum_{k=1}^{K} e^{-\frac{1}{2}(AIC_k - AIC_{\text{min}})}}
\]
For all distributions considered in this paper, we found one model $m^*$ such that $w_{m^*} \sim 1$ (which implies all the other models have Akaike weight very close to 0).

Figures
All figures were generated using Matplotlib [55] package (version 1.5.3) for Python.

Related work
We present here more detailed analysis of the literature discussed in the paper.

Supporting information
S1 File. Supporting figures and tables. (PDF)

Author Contributions
Conceptualization: LA SL AB.
Data curation: PS.
Formal analysis: LA SL AB.
Investigation: LA SL AB.
Writing – original draft: LA AB SL.
Writing – review & editing: LA PS SL AB.

References


