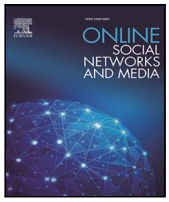
Online Social Networks and Media 28 (2022) 100196

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem



Contact duration: Intricacies of human mobility

Leonardo Tonetto [a,*], Malintha Adikari [a], Nitinder Mohan [a], Aaron Yi Ding [b], Jörg Ott [a]

[a] Technical University of Munich, Boltzmannstrasse 3, Garching, 85748, BY, Germany
[b] TU Delft, Jaffalaan 5, Delft, 2628, BX, Netherlands

ARTICLE INFO

Keywords:
COVID-19
Human mobility
Bluetooth sensing
Opportunistic forwarding

ABSTRACT

Human mobility shapes our daily lives, our urban environment and even the trajectory of a global pandemic. While various aspects of human mobility and inter-personal contact duration have already been studied separately, little is known about how these two key aspects of our daily lives are fundamentally connected. Better understanding of such interconnected human behaviors is crucial for studying infectious diseases, as well as opportunistic content forwarding. To address these deficiencies, we conducted a study on a mobile social network of human mobility and contact duration, using data from 71 persons based on GPS and Bluetooth logs for 2 months in 2018. We augment these data with location APIs, enabling a finer granular characterization of the users' mobility in addition to contact patterns. We model stops durations to reveal how time-unbounded-stops (*e.g.*, bars or restaurants) follow a log-normal distribution while time-bounded-stops (*e.g.*, offices, hotels) follow a power-law distribution. Furthermore, our analysis reveals contact duration adheres to a log-normal distribution, which we use to model the duration of contacts as a function of the duration of stays. We further extend our understanding of contact duration during trips by modeling these times as a Weibull distribution whose parameters are a function of trip length. These results could better inform models for information or epidemic spreading, helping guide the future design of network protocols as well as policy decisions.

1. Introduction

The SARS-COV-2 outbreak in 2020 showed us, once again, the importance of understanding human mobility, also reflected in the vast literature that exists and continues to increase (*e.g.*, [1–5]).

SARS-CoV2's spread is hard to control, as asymptomatic patients contribute to transmission. Most current epidemiological models are limited in how they assume uniformity in contacts between individuals [6,7], thereby overestimating the efficacy of lockdown measures [3,5,8]. *It still remains a challenge, however, to refine these models with more accurate information on individuals contact with one another in various locations as well as while on the move, which we address in this paper.*

To help curb the spread of the virus, various forms of contact tracing have been implemented, with varying degrees of success. Contact tracing efforts have been carried out in various countries in either manual (with the use of contact tracers which do not scale [9]) or automated ways (which only work if the majority of the population adopts and have a series of issues with privacy and trust [10]). From various automated contact tracing approaches, Bluetooth-based are the most popular [9]. Among others, the digital tracing based on Bluetooth sensing has been widely adopted by multiple countries, especially given the pervasiveness of this technology in today's smart-devices (e.g., phones, watches, tablets) and its shown efficacy in aggregating users in close proximity [11].

In this work, we capture and analyze data from a mobile social network of individuals, including multiple sensors from their mobile phones. This approach allows us to accurately sense physical encounters between persons through the ephemeral virtual network formed by their devices in close proximity [12]. We study the daily mobility from location traces of 71 subjects, containing GPS and Bluetooth data, for 2 months in 2018. Furthermore, we quantify different properties of *contacts* between these subjects as well as with nearby individuals through Bluetooth encounters.

As a result of our analysis, we show how overall stays are well modeled by a power-law. However, when breaking down the stops into *time-unbounded-stops* (typically do not follow a schedule, e.g., bars, restaurants, etc.) follow a log-normal distribution, while *time-bounded-stops* (*i.e.*, typically follow a schedule, such as office) follow a power-law. Previous studies report similar observations in web-content viewing time [13], where users spend time differently according to the content being viewed. Power-law distributions describe the duration of interactions with time-free content (*e.g.*, text, photos) while log-normal distribution best describe interactions with time-correlated content

* Corresponding author.
E-mail address: tonetto@in.tum.de (L. Tonetto).

https://doi.org/10.1016/j.osnem.2021.100196
Received 30 June 2021; Received in revised form 10 November 2021; Accepted 19 December 2021

(*e.g.*, videos). Human brain perception of *information* was used to explain these differences [14].

Inter-personal *contact duration*, however, shows a log-normal distribution. With this observation, we propose a model to estimate such values from the overall duration of stays (power-law). When characterizing trips, we observe trip length as well as trip time duration follow a log-normal, while contact duration during trips follows a Weibull distribution, in which its parameters are best described as a function of the distance traveled. Taken together, these results suggest how contacts happen in various modes of transportation, and could be used to guide planning of future urban environments and in coping with pandemic outbreaks.

2. Related work

The growing pervasiveness of smartphones and their sensors enabled researchers to study various aspects of *human mobility* in recent years. Random models for movements were replaced by Lévy-flight (power-law based) models [15,16]. Using data sets with higher resolution, these observations have been more recently revisited, and the distribution of displacements has been shown to follow a log-normal distribution [17,18] in urban scenarios while exponential in intra-urban trips [19].

Human mobility has also been modeled around social interactions [20,21], natural disasters [22], and income [23].

Another fundamental aspect of mobility that has been largely studied is information dissemination, either for opportunistic data forwarding [24] or contagious disease spread [5,8]. The seminal work by Hui et al. [25] revealed long-tailed distributions in *inter-contact time* (time interval between consecutive contacts of any pair of devices) instead of exponential distribution and its implications on opportunistic forwarding systems using a data set collected during a scientific conference. Furthermore, the complementary study by Chaintreau et al. [26] includes 8 different data sets, however all do not include either accurate measurements for location or contact duration and often include measurements done in a limited set of locations (*e.g.*, conference venues and university). Other similar studies include fine grained measurements also limited to certain locations, such as schools [6], conferences and museums [27]. The work by Sun et al. [12] studies contacts using a metropolitan scale data, but limited to public transport. In our study, we analyze mobility and contacts data by observing their daily lives.

While short *inter-contact times* are associated with lower latency in opportunistic networks, large *contact duration* can be seen as high throughput [28]. Regardless of their importance, most recent studies have focused on the former, mainly as recent advancements in wireless network technologies brought a nearly infinite bandwidth to mobile devices, even though data exchange capacity grows as contact duration gets longer. When modeling the spread of infectious diseases, however, *contact duration* is a key aspect [6,29].

Contact duration allows the study of how epidemics spread through a temporal network, in which edges between nodes evolve over time [30]. While such studies often better describe the dynamics of diseases outbreaks and their prevention, little is still known about how mobility and contacts are related. Therefore, to help bridge this gap, our study characterizes inter-personal contacts through a series of analysis of GPS and Bluetooth data. Our results while elementary also reveal intricate relationships between contacts and human mobility.

It is assumed in this study, that the well documented short range of Bluetooth is a good proxy for human contacts, and therefore a proxy for the possible transmissibility of an infectious disease, such as SARS-CoV2 [9]. In other words, our observations are shaped by the technology used in our measurements.

3. Background

In this section we define the notion of *contact*, *stop* and *trip* used in this paper, and describe the distribution functions observed, as well as the method for estimating their parameters.

Basic definitions

Contacts: We model a *contact* between two individuals through measurements of Bluetooth signals. Given the short range of this radio technology it can emulate well interactions between persons, especially in the context of airborne infectious diseases [9,31].

Stop: We define a stop (or a stay) as a prolonged visit to a well defined point of interest, *e.g.*, home, a shop or a transit station, but *not* a short break at a traffic light (Section 5).

Trip: Given the detection of *stops*, a *trip* is defined as the sequence of geographical coordinates between two identified locations where a subject spent enough time. We also define the total length of a trip as the sum of all distances between all consecutive points of a that trajectory, that is $\ell = \sum_t \|\mathbf{x_t} - \mathbf{x_{t-1}}\|$, where $\mathbf{x_t}$ is the location at time t (Section 6).

Empirical distribution functions

While limited when compared to highly parameterized models (*e.g.*, neural-networks), well-known distributions are highly *interpretable* (*i.e.*, changes in the distribution can often be explained by variations in parameters), *comparable* (*i.e.*, different parameter values or different distributions have intrinsic properties that can be contrasted), and *portable* while preserving the **privacy** of the subjects involved in the study (*i.e.*, models or data sets can be compared without any personal identifiable information being shared).

In this work, we observe three long-tailed distributions for stops (Section 5) and trips (Section 6), which we describe next, along with the implication of observing each one of them.

Log-normal: The probability density function (PDF) of this function, for a given random variable X for all $x > 0$, is defined by Eq. (1), with parameters μ (mean or *location*) and σ (standard deviation or *shape*). Intuitively, this distribution describes a Normal distribution for the logarithm of a random variable. This distribution has been used to describe trip length from GPS data [17,18,32] and for stop duration [18], for describing the length of textual internet content [13], and time users spend on individual internet content without a time component [14] (*e.g.*, images, text).

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

Weibull: The PDF of this distribution function, for a given random variable X for all $x > 0$, is defined by Eq. (2), with parameters λ (*scale*) and β (*shape*). While λ describes how spread-out the distribution is, β defines whether the tail of the distribution will be exponential (when $\beta > 1$) or long-tailed (when $\beta < 1$). This distribution has been used to describe trip length from Twitter data [33] and from taxi data [34], as well as users behavior on online social networks [35].

$$p(x) = \frac{\beta}{\lambda}\left(\frac{x}{\lambda}\right)^{\beta-1} e^{-\left(\frac{x}{\lambda}\right)^{\beta}} \tag{2}$$

Power-Law: The PDF of this distribution function, for a given random variable X, is defined by Eq. (3), with parameters α (*scale*) and $x_{\min}$ where $\alpha > 0$ and $x_{\min} > 0$. This distribution has been extensively used to model various naturally occurring phenomena [36] and is often explained by *preferential-attachment* in a time-evolving network [37]. Power-law models have been extensively used to describe trip length [16,38], friendship on online social networks [21], and the organization of the Web [39].

$$p(x) = \frac{\alpha - 1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha} \tag{3}$$

Parameters Estimation and Distribution Comparisons: To fit the parameters of these distributions we use the maximum-likelihood method

proposed by Clauset et al., which provably gives accurate parameter estimates in the limit of large sample sizes [40]. Once the best parameters are found for a distribution, a likelihood value is derived, which in turn, is used to compare the log-likelihood of which distribution best describes the data. Finally, following the method by Clauset et al. [40] we produce a *p-value* which allows us to infer the significance of this comparison (*i.e.*, that it was *not* due to chance). For this work, we adopt the common convention that a p-value < 0.05 is significant. That is, when comparing how well two distributions describe a set of data, a p-value < 0.05 indicates that there is a probability lower than 5% that the best distribution was chosen due to randomness. Therefore, whenever reporting a distribution fit, we provide the *p-value* to the comparison between the two best options.

4. Data collection

Our data collection had 71 registered subjects who agreed to participate in our study. Sensors data were collected using the Aware App [41], for a total of 2 months starting in April/2018. Subjects were mostly between 20 and 30 years old, living in Munich, Germany. Location data as well as Bluetooth scans had a median sampling rate of 3 minutes^{-1} (95th-% 9 minutes^{-1}), ensuring a high density and reliable source of data for our analysis.

Cohort Biases: The cohort of this study consists of young adults, living in a large metropolitan city, in Europe. Therefore, all of our observations do not represent how the totality of the human population behaves. For example, elderly age groups are likely to have their overall mobility far more constrained [42]. Populations in other regions of the world, where public transport is less developed will also show different patterns [23]. To support how well our data set captures different aspects of human mobility, we present some observations using both our data as well as the Geolife set (Section 4.3).

4.1. Location data

These were provided by the GPS sensor as well as the operating system as geographical coordinates together with an estimated accuracy [41]. The 85th-percentile of this uncertainty was 10 m, allowing us to accurately extract the stops (Section 5) and trips (Section 6) of each subject. These data include urban mobility as well as long distance commutes and international travels to 17 countries.

4.2. Bluetooth contact data

We study real-world contacts through the ephemeral social network built from the proximity between mobile devices. For that, we classified observed nearby devices into *human-held* and *static*, modeling human contacts using Bluetooth readings as our microscope. This classification was done in two phases, that we explain below.

In *phase 1*, we use the name broadcast by nearby devices, commonly used for discoverability. To these names, we cleared and tokenized their strings in order to filter out non-English/German words. Finally, we manually classify them in either human-held or stationary. These steps ensured any personally identifiable information was removed, while maximizing the coverage of possible human-held devices. Examples of this group include *battery_pack*, *camera*, *smart_watch* and *cigarette*, while examples of stationary devices include *light*, *home-theater*, and *printer*. In this step, we were able to classify nearly 6000 unique devices, which correspond to 5% of the total MAC addresses recorded.

In *phase 2*, we used a method by Alipour et al. [43] to classify Wi-Fi devices based on their MAC address. More specifically, it assumes vendors assign similar prefixes of the MAC address for similar devices. With this approach, we could classify an extra 16920 devices (15.5% of the total). A random 1% sample from this phase revealed names which attribute the type of devices as human-held, such as *cameras* and *portables speakers* (e.g., Canon, Bose), validating this classification.

Table 1
Summary of the data set used.

Users	Stops	Encounters	Trips
71	19317	12432	18438

Note that we could only classify 20% of the recorded nearby Bluetooth devices. All unclassified devices were discarded to eliminate possible biases and uncertainties. After these preparation steps, we identified a total of over 6500 human-held devices. We then assumed each of these devices to represent the person they belong to. Although this strong assumption held inexorable biases, the similarity with previous studies on the distribution of contact duration (discussed next) suggests that distortions do not invalidate our results.

In this study, we consider contacts which happened in either a *stop* or a *trip*, and not encounters which last for multiple events. Out of a total of 12,423 contacts studied from our collected data, 389 lasted for consecutive *stops* or *trips*. This was done in order to distinctively classify each encounter into a mobility modality as well as discard multiple devices a single subject could be carrying.

The distribution of all contacts duration, regardless of while moving or static, was best described by a log-normal distribution, with parameters $\mu = 6.67$ and $\sigma = 1.65$ (p-value $= 0.002$ to a power-law). As expected, compared to contacts during stops (Section 5), the biggest difference is observed in a significantly larger *shape* parameter (σ), supporting previous observations of short-tailed distributions for contacts [44]. A summary of the main features of our data set are summarized in Table 1. To extend our understanding on contact duration, we will focus on a clear separation between *stops* and *trips*, as will be presented in the next sections.

4.3. Supporting set - Geolife

We validate some of our observations with the Geolife data set [45]. It contains GPS trajectories from 182 subjects for 4.5 years, and sampling rates of 5 seconds^{-1} or 10 meters^{-1}, which we process using the same methods used in our data.

5. Stops

In this section, we characterize our *stops* (or stays) as well as construct a model of contacts observed at these locations.

5.1. Detection of stops

To ensure a robust and reproducible detection of stops, we apply the extensively used stop detection method for GPS traces proposed by Zheng et al. [46]. It defines two main parameters: `max_dist`, as the maximum distance allowed between any two geo-location points within an area, or location cluster; `min_stop_time`, as the minimum duration spent within a location cluster for it to be considered a *stop*.

To detect stops, we first cluster consecutive location records using `max_dist`, and continue adding new points to the cluster as long as its distance δ to *any* other point in the cluster is smaller than the threshold (*i.e.*, $\delta <$ `max_dist`). Once a new candidate point no longer fulfills this criterion the cluster is evaluated as a *stop*. This evaluation is done by comparing the total time spent at the cluster (τ) with `min_stop_time`, *i.e.*, if $\tau >$ `min_stop_time` then the cluster is a *stop*, otherwise it is discarded. Once a *stop* is identified, its location is saved as the centroid of the cluster.

Given the high accuracy of the location points in our collected data (Section 4), we chose `max_dist` $= 10$ meters. Furthermore, we evaluated possible values for `min_stop_time` between 5 min (location sampling rate, Section 4) and 50 min, at intervals of 1 min. The graph `min_stop_time` vs. total number of stops showed an inflection point between 10 and 15 min, leading us to select `min_stop_time` =

15 minutes for a more conservative choice, also inline with previous research. A *stop* of at least 15 min would also allow us to identify potential *close contacts* in the context of COVID-19, as defined by the CDC [29].

5.2. Stops enrichment

In order to characterize sojourn times in the various places visited, we further classified the observed *stops* in our collected data. First, the "home" locations of the subjects were identified, then all remaining *stops* were classified with a combination of multiple publicly available location API.

The detection of "home" is of key importance given its central role in a person's mobility [16,47]. Therefore, as a first step in classifying *stops*, we assign "home" to the *stop* location a subject had the highest frequency of visits between 7pm and 7am [47]. These places are then removed from all subsequent analyses as we are interested in how contacts happen outside people's homes, where they might have little control over whom they might encounter.

For the remaining *stops*, we searched 4 different location API: Google Places,[1] Tomtom Places,[2] Foursquare Places,[3] Here Geocoding and Search.[4] In all cases, these services provide a list of points of interest (POI) that are nearest to a given geographical coordinate. From this list of possible POI, we pick the one closest to a requested stop, within a maximum distance of 10 m. This variety of services ensured maximal coverage of the places visited by our subjects, allowing us to identify 57% of all *stops*.

The categories of POI identified were: apartment/residence, bank, bar, company/office, entertainment (*e.g.*, museum, art gallery), gas station, gym/sports facility, health facility (*e.g.*, hospital, clinic), hotel, library, religious center, restaurant, salon, shop, supermarket, theater (including cinemas), transport station (*e.g.*, train, bus), and university. When studying sojourn times, we use these categories to examine how the distributions of such times varies across different places.

5.3. Stops duration

Here we present the observations we have for stop (or stay) duration, often referred to as *sojourn time*. When taken without discrimination by category, the distribution of stops duration is well described by a power-law ($\alpha = 2.13$, p-value < 0.001 to a log-normal), which has a probability density function defined by Eq. (3) (Section 3), in which $x_{\min}$ is the minimal value chosen when fitting the parameter α of the distribution. For our analysis, as explained in Section 5.1, the minimum time we use was 15 min (*i.e.*, $x_{\min} = 900$). Fig. 1 depicts this distribution for our collected data, in accordance with the same analysis using the Geolife data set ($\alpha = 1.98$, p-value < 0.001 to a log-normal). Further supporting these observations, from a much larger data set based on call detail records, Song et al. also fitted a power-law with similar parameters values ($\alpha = 1.8$) to the distribution of stops duration [38]. This long-tailed distribution is often explained by preferential attachment, in which a person will tend to have few preferred locations to visit. In this way, various places will be visited rarely and for a shorter duration while few places are likely to see much longer stays.

Interestingly, when looking at these distributions based on the category of place visited (Section 5.1), some categories present a power-law distribution in their stops, while others present a log-normal distribution. The probability density function of a log-normal is defined by Eq. (1), in which μ defines the center and σ the scale (or log-variance) of the distribution. Unlike a power-law, a log-normal distribution has an exponential tail. This indicates that the underlying process described by this distribution is bounded by something, like resources. Furthermore, existing work by Kai et al. on human mobility has shown how the combination of log-normal processes can lead to a power-law distribution [17].

One common characteristic of stops described by a log-normal is that the distribution emerges in places where the user has no time constraints in either starting or ending a visit (time-unbounded-stop), such as bars, restaurants and gyms (which accounted for 55% of the total identified stops). On the contrary, stops where a user would typically follow a schedule to either start or stop a visit (time-bounded-stop) are better described by a power-law distribution, places such as offices, hotels, and transport stations (accounting for the remaining 45% of the total identified stops).

In the work by Gros et al. [14], the authors made a similar observation to file sizes from internet content. In their results, they observe power-law distributions to files without a time component (*e.g.*, text), and log-normal for objects for which the time is defining qualia (*e.g.*, videos). Finally, these findings were explained by maximum information entropy [48], in which the time component, when present, worked as an additional constraints to file sizes in the form of an exponential tail. For stop duration, we conjecture that a similar phenomenon appears whether or not the visit follows a schedule. Therefore, the end of the pre-allocated time for a visit would work as an added constraint to the total time spent at a place, yielding an exponential tail, characteristic of a log-normal distribution. Complementary, time-unbounded-stops not having this temporal constraint, yield a power-law distribution for visits, in line with our results.

These results highlight the importance of studying mobility with higher resolution sensors data, such as the one used presently, which allows us to further classify *stops*, revealing intrinsic properties of these stay durations which would not emerge in coarser measurements. Furthermore, for a given random variable T of stay durations, with a defined mean μ and standard deviation σ, a log-normal distribution produces the largest possible entropy, supporting the characterization of time-unbounded-stops as least predictable [49].

5.4. Contacts characterization at stops

The distribution of contacts is well described by a log-normal distribution (Eq. (1)). The data collected as well as the distribution fit to these data are presented in Fig. 2. Interestingly, the distribution of contacts remained constant (*i.e.*, with similar parameters) at different distances from each user's home. We grouped stops: (i) up to 1 km from home, (ii) between 1 km and 100 km, and (iii) above 100 km from home, and found similar parameters describing their contacts distribution.

Using the *stops* characterization discussed previously (see Section 5.3), we observe a similar distribution for contacts as for stop duration. In time-bounded-stops, contact duration was better described by a power-law ($\alpha = 2.21$, p-value $= 0.03$ to a log-normal), while in time-unbounded-stops, contacts were best described by a log-normal distribution ($\mu = 7.6, \sigma = 0.99$, p-value $= 0.04$ to a power-law).

As all individuals would tend to stay fixed amounts of time to fulfill a schedule at time-bounded-stops, they are more likely to produce long-tailed contacts when compared to time-unbounded-stops. As in the latter visits might be driven by a goal (*e.g.*, eat something at a restaurant), contacts show an exponential decay with a small *shape* parameter (σ).

[1] https://developers.google.com/places
[2] https://developer.tomtom.com/products/places-api
[3] https://developer.foursquare.com/docs/places-api/
[4] https://developer.here.com/documentation/geocoding-search-api

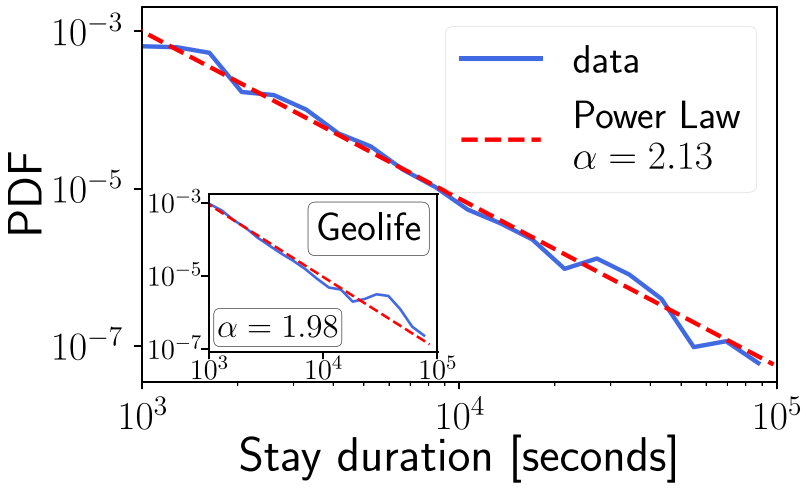


Fig. 1. Overall stop duration follows a **power-law** distribution.

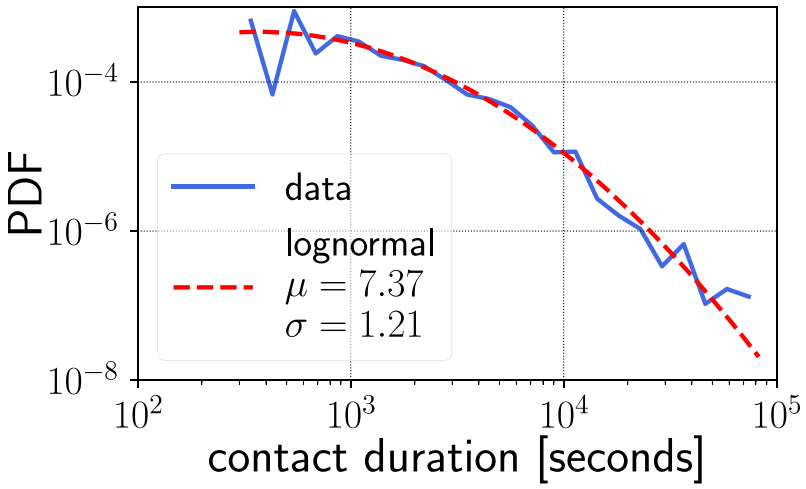


Fig. 2. Contact duration at stops follows a **log-normal** distribution.

5.5. Model of contacts during stops

Given the scarcity of data on inter-personal contact, we now propose a model capable of inferring a distribution of contact duration from a distribution of stay duration. Given the more common availability of location traces from which stop duration can be inferred, this model enables a simplified estimation of how long contacts will last. In turn, this can be used to better model the spread of information opportunistically as well as the spread of infectious diseases.

We first define the probability of a stay duration y as a power-law of the form $Pr(y) = Cy^{-\alpha}$, where $C = (\alpha - 1)x_{\min}^{(\alpha-1)}$, and α is the defining coefficient of the distribution. Then, as previously discussed, we know contact duration x follows a log-normal distribution, therefore we can write $e^x \propto N(\mu, \sigma)$, where $N(\mu, \sigma)$ is a normal distribution defined by μ and σ. To avoid the non-trivial estimation of $N(\mu, \sigma)$ we can approximate its probability density function with a uniform distribution.

This non-parametric estimation produces a constant loss-function in the interval of a stay duration (*i.e.*, from 0 to y). This observation emerges from the KL-Divergence between any target distribution P being approximated by a Uniform distribution U, in the interval $((a, b) = n)$ as in Eq. (4), where the final divergence is defined only by the desired interval n and the entropy of the target function $H(P)$.

$$
\begin{aligned}
D(P \parallel U) &= \sum_{i}^{n} P(X_i) \log_2 \left(\frac{P(X_i)}{U} \right) \\
&= \sum_{i}^{n} p_i \log_2 \left(\frac{p_i}{1/n} \right) \\
&= \log_2(n) + \sum_{i} p_i \log_2(p_i) \\
&= \log_2(n) - H(P) \qquad (4)
\end{aligned}
$$

We therefore can re-write the definition of x as $e^x \propto 1/y$. To find a relationship between x and y we can write $Pr(x)\, dx = Pr(y)\, dy$, as well as $dx \propto e^x\, dy$. Substituting, we get $Pr(x) \propto C\, e^{\alpha-1}$.

By definition, a given random variable Z, it is said to be described by a log-normal if it has the form $Z \sim e^{\mu+\sigma x}$ and if x is normally distributed. By comparing this equation with the inferred $Pr(x)$, we can compute $\mu \approx \ln(\alpha - 1)x_{\min}^{\alpha-1}$ and $\sigma \approx \alpha - 1$.

From our data, using $\alpha = 2.13$ (Section 5.3), we estimate $\hat{\mu} = 7.80$ and $\hat{\sigma} = 1.13$, which are close to the actual values (Section 5.4) $\mu = 7.37$ and $\sigma = 1.21$.

With this model, we vary α and plot the resulting distributions in Fig. 3. Interestingly, this model shows how overall shorter stays actually leads to a decrease in the probability of seeing users of shorter stay while increasing the probability of longer contacts. Numerically, an increase in α as a result of shorter stays increases both $\hat{\mu}$ (*i.e.*, the distribution shifts to the right) and $\hat{\sigma}$ (*i.e.*, the standard deviation, or spread, of the distribution increases).

Note that this does not mean that the frequency of longer contacts is going to be higher, but rather among the remaining contacts, those of longer duration will have a higher likelihood of being encountered.

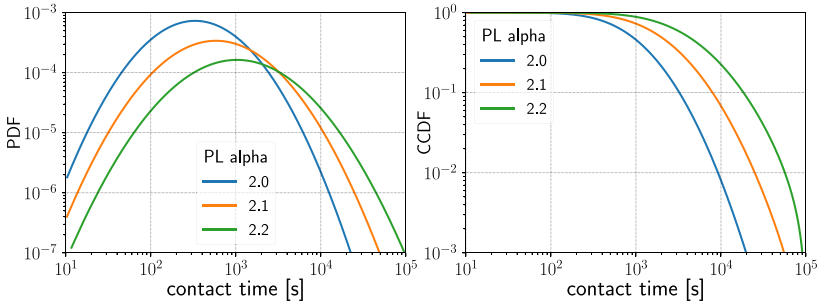


Fig. 3. Distribution of modeled contact duration for different values of the stay duration parameter (*alpha*). Larger values of *alpha* for stay duration indicate *higher* probability for shorter stays, leading to an increase in the probability of long-term contacts as short-term meets become less often.

Takeaway: As individuals in a population follow a similar mobility model in their visits, a pattern for contacts emerges. A shortening in stay duration leads to fewer contacts, where the remaining ones are inevitably longer. However, these changes in stay duration may not be possible across all locations as people tend to follow a schedule in some of them (time-bounded-stops).

6. Trips

In this section, we present the results for trips. To complement the characterization of contact duration at stops, we show how contact duration during trips actually follows a Weibull distribution. We discuss the implications of such distribution as well as its parameterization being a function of the distance traveled, along with its interpretation.

6.1. Detection of trips

To ensure the quality and validity of the inferred trips, we validated these in three steps. First, we only consider trajectories that start and end at an identified *stop*. This ensures the integrity of trips. Second, we impose a *temporal* constraint by eliminating any trajectory that contains a pair of coordinates recorded within a time interval greater than 1 h. This was done in order to avoid large fractions of trips to go untraced while allowing some discontinuity that could be caused by poor GPS reception indoors [50] or when a subject might have switched off their phone. Third and lastly, we impose a *spatial* constraint by eliminating any trajectory containing a distance between any pair of consecutive points ($d = \|\mathbf{x_t} - \mathbf{x_{t-1}}\|$) which was greater than 50% of the total trip length (ℓ). That is, for any d between two points in a trajectory, if $\ell/2 < d$ that trajectory is discarded from further analysis. This ensures the continuity of the traces as well as the reliability when characterizing contacts during trips.

After the aforementioned steps, we identified a total of 2512 trips which will be further analyzed next.

6.2. Trip duration and total length

In contrast to *stop* duration (Section 5.3), the time spent traveling, in our collected data, was best modeled by a log-normal distribution (p-value = 0.02), depicted in the left panel of Fig. 4. A similar observation was made in the Geolife set (p-value < 0.001), presented in the inset of that same panel. As the majority of trips in our collected data were in urban environments (Section 4), the exponentiation instead of a long-tail could be explained by a decrease in average population density in urban areas along lengthy trips [19]. Taken together, these observations reinforce the validity or our data collection as well as methods for *stop* and *trip* detection, while providing insights into how contacts happen during trips (Section 6.3).

In agreement with previous work by Alessandretti et al. [18] (N=850, GPS points at high temporal granularity) and our observations in the Geolife data set, trip length in our data is best modeled by a log-normal distribution, depicted in the right panel of Fig. 4. A fit with a power-law yielded $\alpha = 1.22$ (shown in dotted gray), however with a much lower log-likelihood than the log-normal (p-value = 0.009), in contrast to part of the previous literature [16,38]. The differences found in our work could be explained by measurements done with much finer granularity in all aforementioned data sets (*i.e.*, fine grained GPS vs. course grained cell tower records).

6.3. Model of contacts during trips

When taken as a whole, contact duration during trips did not have a good fit with either of the distribution functions discussed in Section 3 (*i.e.*, p-value > 0.05 when comparing some of these alternatives). The best fit was revealed when segmenting the trips based on distance traveled. This analysis revealed an intricate relationship between distance traveled and the characteristics of the contact duration, which are shown in Fig. 5.

Interpreting the changes of these parameters as a function of distance, reveals a set of interesting characteristics. As the trip distance increases, λ displays a bi-modal behavior. This parameter, often referred to as the *scale* of the distribution is directly proportional to the average (and median) of the Weibull (Section 3). Its bi-modal shape is likely capturing the tendency for people to take (crowded) public transport (bus,train,airplanes) with similar probability as a function of the distance traveled. Alternatively, people would either walk or drive and have less contact (or shorter contacts) with other people.

The parameter β, often referred to as the *shape* of the Weibull, decreases from ~1.2 to ~0.5 as the trip distance increases. When $\beta > 1$, it decays faster than an exponential, or in other words, the longer a person is seen nearby, the shorter they are likely to remain close by. When $\beta < 1$, it demonstrates a long-tail behavior, or in other words, the longer a person is nearby, the longer they are likely to stay. Once again, this is likely caused by the choice of means of transport, where walking is likely predominant for shorter distances, and vehicles for longer ones. Furthermore, in the latter such behavior is (probably) explained by people typically traveling together, and again, the longer someone stays next to you in the metro/train/bus, the longer they are to continue with you (*e.g.*, a commuter train has only spaced-out limited stops).

Different from *stop* duration, simulations of changes to trips requires a broader understanding of the intrinsic purposes people travel [47]. For example, trips are often taken from home to work (which cannot be easily changed), or to a shop or restaurant given someone's intentions, which are not captured in our data. Therefore, in order to fully grasp the changes introduced in trips distribution and consequent contact duration by lockdown measures, a recent data set with similar high density is required.

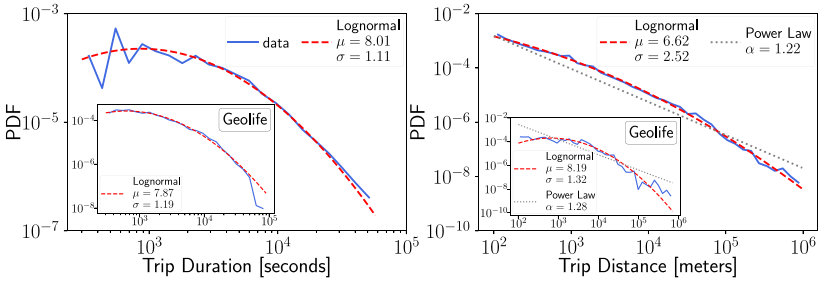


Fig. 4. Both trip time duration and length are best modeled by a log-normal.

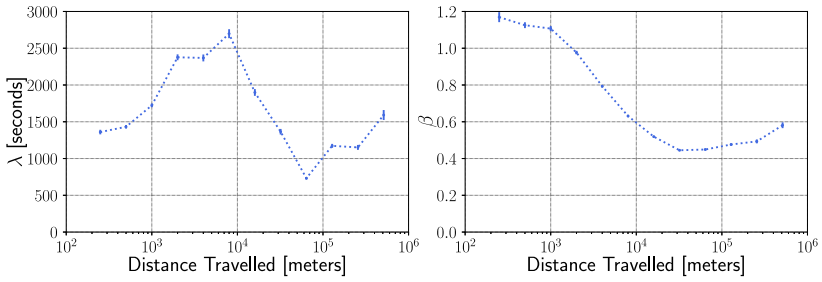


Fig. 5. Variation of the Weibull parameters as a function of distance traveled.

Takeaway: Restricting trips to only short distances may not necessarily lead to less (or shorter) contacts. As distances increase, people may choose other modes of transport (*e.g.*, driving) which do not expose them as much.

7. SIR model and implications

A commonly used model for information spreading, such as infectious diseases, is the SIR model. From a set of assumptions, the *contact ratio* (q) is a key parameter defining how fast an epidemic spreads, the maximum number of infective individuals, and the total number of individuals that will ever get infected (see [51]). Essentially, the *contact ratio* is defined by how often individuals from a population are in close contact for a sufficient amount of time, which is, in turn, defined by the information being transmitted (*e.g.*, a virus or a computer file).

Implications: With enough data about how the aforementioned aspects of human mobility changed in the COVID-19 pandemic, a more accurate modeling of the epidemic would be possible. The results of such models could better inform policy makers about new restrictive measures on movements, which, in turn, could include visits of limited duration. As shown in Section 5, shorter stops will lead to less contacts, while shorter trips might not necessarily lead to the same outcome (Section 6). Recent studies have demonstrated strong associations between *non-pharmaceutical interventions* and changes in the spread of SARS-COV-2, even though it remains challenging to study measures in isolation [3,7,8,52]. A large study including 11 European countries revealed that total lockdown measures were responsible for the reduction in 81% in the reproduction rate (R) in those countries [8], while in the US, for similar measures, the observed reduction in contacts was over 90% (10.86 → 0.89 interactions per day). Furthermore, contact restriction measures in China were associated with a 2.6 fold reduction in infections [52] when compared with an unrestricted scenario. Using a large data set, with over 98 million individuals for 6 months, Chang et al. [3] demonstrated, with temporal networks and a modified SIR model, that a reduction in the capacity of visits of places to 20% could lead to a reduction in the number of infections of up to 80%. Such cut in capacity could, for example, be achieved through a reduction in the overall stay duration, as shown in this work. Taken together, these results demonstrate the importance of a combined, timely and well informed set of changes to curb the spread of an infectious disease.

8. Limitations and discussions

Mode of Transport: Previous research on mobility [17] has shown the importance of studying distances traveled for each transportation mode. However, we did not perform any mode of transport inference, which might have limited our study. This was, in part, due to a lack of ground-truth to validate any model we may ever want to use. Future iterations of such study should include a reliable inference.

Models generalization: The unique combination of mobility and contact information in our traces, allowed us to apply well-established statistical models (see Section 2) to better understand how these two properties are related. Our robust results, supported by statistically significant measures, reveal the solid numerical relationship between mobility and contact duration. Furthermore, our approach could be applied to similar sets, yielding interpretable and comparable results.

Contact Opportunities: Fig. 6 shows how longer stays expectedly bring users in contact with more people. However, the link between these two is only a weak one, with a Spearman correlation of 0.3. This way, curfew measures, which would bring down contact duration (*e.g.*, take-away instead of dine-in), could bring down the overall number of contacts a person will have. In an opportunistic communication scenario, current curfew measures would significantly impact the performance of such systems. Furthermore on epidemics, as most remaining contacts will be long ones, such as with workers in a restaurant or a shop, these individuals should be isolated as much as possible from the general public, in order to reduce the probability of transmission.

Current Data Sets: Even though for the current COVID-19 pandemic there exist aggregate data on mobility changes from Google and Apple, those refer only to the number of visits to places but not to how long people stayed (*i.e.*, were potentially in contact) [53]. New measurements or another source of data would be needed to allow us to assess the impact of those changes in contact duration. Other non-pharmaceutical interventions with significant effect on the spread of COVID-19 as well as in the characterization of human contacts, that

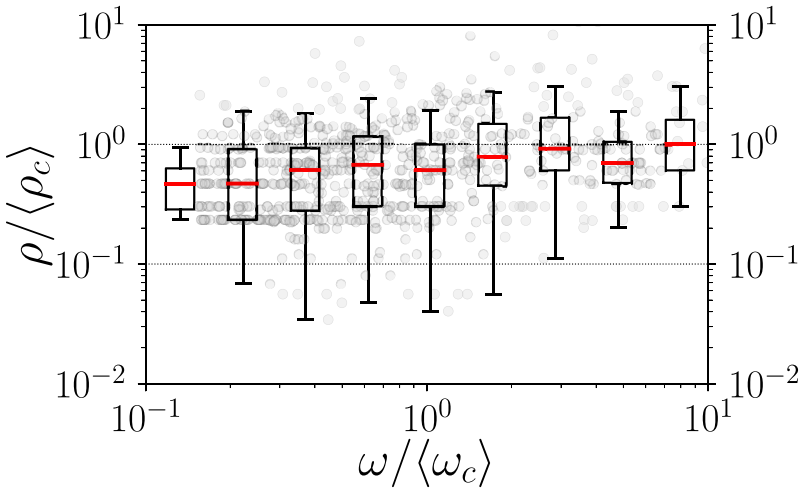


Fig. 6. Number of contacts increases with longer stays (Spearman correlation = 0.3, p-value <0.01). (Red) Lines inside boxes represent median values.

were not part of our study, include face-masks, hand-washing and prohibition of large events [8,52].

Other Applications: Other areas which could benefit from these results include smart urban planning as well as the design of mobile network protocols. The planning of public spaces as well as smart transportation could benefit from our insights in how changes in their utilization would lead to modified contacts. Mobile network protocols, especially in opportunistic scenarios, could utilize our models to better understand how information dissemination would occur under different circumstances or types of location, such as in disaster-stuck regions.

9. Conclusion

In this work we analyzed high resolution data from a mobile social network, including mobility and contacts, from a series of mobile phone users. We reveal a strong relationship between the distribution of stop duration and location types, where *time-bounded-stops* (*i.e.*, where there is a typical schedule) follow a power-law and time-unbounded-stops follow a log-normal. We further model the relationship between stop duration and contact duration, which could be further used in studies where contacts are not available. Furthermore, our analysis of trips reveals an intricate relationship between the distribution of contact duration and trip lengths, where the distribution of the former is best described by a stretched exponential for which both parameters are a function of the latter. These findings can be further used by researchers to develop more accurate models to better understand and deal with the current (and future) pandemic, as well as support the creation of better mobile network protocols.

10. Ethical considerations

For this study, all participating subjects voluntarily agreed to be tracked and have their data used for this study under a privacy agreement. The pre-processing steps described in Section 4 were designed and executed to ensure no personal identifiable information was ever disclosed, be it from the participant's device or those devices sensed nearby. No individual subject was ever studied in isolation, but only aggregates.

CRediT authorship contribution statement

Leonardo Tonetto: Conceptualization, Data collection, Methodology, Analysis, Writing. **Malintha Adikari:** Data curation, Writing. **Nitinder Mohan:** Writing. **Aaron Yi Ding:** Writing. **Jörg Ott:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank all volunteers for their data collection. This work was partially supported by TUM IGSSE MO3 project, Germany.

References

[1] D.S. Candido, et al., Evolution and epidemic spread of SARS-CoV-2 in Brazil, Science 369 (6508) (2020).
[2] S.M. Kissler, et al., Reductions in commuting mobility correlate with geographic differences in SARS-CoV-2 prevalence in new york city, Nature Commun. 11 (1) (2020).
[3] S. Chang, et al., Mobility network models of COVID-19 explain inequities and inform reopening, Nature (2020).
[4] K. Soltesz, et al., The effect of interventions on COVID-19, Nature 588 (7839) (2020).
[5] M.U. Kraemer, et al., The effect of human mobility and control measures on the COVID-19 epidemic in China, Science 368 (6490) (2020).
[6] M. Salathé, et al., A high-resolution human contact network for infectious disease transmission, Proc. Natl. Acad. Sci. USA 107 (51) (2010).
[7] A. Aleta, et al., Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19, Nat. Hum. Behav. 4 (9) (2020).
[8] S. Flaxman, et al., Estimating the effects of non-pharmaceutical interventions on COVID-19 in europe, Nature 584 (7820) (2020).
[9] L. Ferretti, et al., Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing, Science 368 (6491) (2020).
[10] I. Braithwaite, et al., Automated and partly automated contact tracing: a systematic review to inform the control of COVID-19, Lancet Digital Health (2020).
[11] A. Montanari, et al., A study of bluetooth low energy performance for human proximity detection in the workplace, in: 2017 IEEE PerCom, IEEE, 2017, pp. 90–99.
[12] L. Sun, et al., Understanding metropolitan patterns of daily encounters, Proc. Natl. Acad. Sci. USA 110 (34) (2013) 13774–13779.
[13] P. Sobkowicz, et al., Lognormal distributions of user post lengths in internet discussions-a consequence of the Weber-fechner law? EPJ Data Sci. 2 (1) (2013).
[14] C. Gros, et al., Neuropsychological constraints to human data production on a global scale, Eur. Phys. J. B 85 (2012).
[15] D. Brockmann, et al., The scaling laws of human travel, Nature 439 (7075) (2006).
[16] M.C. Gonzalez, et al., Understanding individual human mobility patterns, Nature 453 (7196) (2008).
[17] K. Zhao, et al., Explaining the power-law distribution of human mobility through transportationmodality decomposition, Sci. Rep. 5 (1) (2015).
[18] L. Alessandretti, et al., Multi-scale spatio-temporal analysis of human mobility, PLoS One 12 (2) (2017).

[19] X. Liang, et al., Unraveling the origin of exponential law in intra-urban human mobility, Sci. Rep. 3 (1) (2013).
[20] N. Eagle, et al., Reality mining: sensing complex social systems, Pers. Ubiquitous Comput. 10 (4) (2006).
[21] E. Cho, et al., Friendship and mobility: user movement in location-based social networks, in: Proceedings of ACM SIGKDD, 2011.
[22] X. Lu, et al., Predictability of population displacement after the 2010 haiti earthquake, Proc. Natl. Acad. Sci. USA 109 (29) (2012).
[23] M.U. Kraemer, et al., Mapping global variation in human mobility, Nat. Hum. Behav. 4 (8) (2020).
[24] P. Hui, et al., Bubble rap: Social-based forwarding in delay-tolerant networks, IEEE TMC 10 (11) (2010).
[25] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, C. Diot, Pocket switched networks and human mobility in conference environments, in: Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-Tolerant Networking, 2005, pp. 244–251.
[26] A. Chaintreau, et al., Impact of human mobility on opportunistic forwarding algorithms, IEEE TMC 6 (6) (2007).
[27] L. Isella, et al., What's in a crowd? Analysis of face-to-face behavioral networks, J. Theoret. Biol. 271 (1) (2011) 166–180.
[28] T. Hossmann, et al., Putting contacts into context: Mobility modeling beyond inter-contact times, in: Proceedings of the ACM MobiHoc, 2011.
[29] K.A. Fisher, et al., Community and close contact exposures associated with COVID-19 among symptomatic adults ≥ 18 years in 11 outpatient health care facilities—United States, july 2020, Morb. Mortal. Wkly. Rep. 69 (36) (2020).
[30] N. Masuda, et al., Predicting and controlling infectious disease epidemics using temporal networks, F1000prime Rep. 5 (2013).
[31] C. Cattuto, et al., Dynamics of person-to-person interactions from distributed RFID sensor networks, PLoS One 5 (7) (2010).
[32] W. Wang, et al., A comparative analysis of intra-city human mobility by taxi, Physica A 420 (2015).
[33] R. Jurdak, et al., Understanding human mobility from Twitter, PLoS One 10 (7) (2015).
[34] R. Gallotti, et al., A stochastic model of randomly accelerated walkers for human mobility, Nature Commun. 7 (1) (2016).
[35] L. Gyarmati, et al., Measuring user behavior in online social networks, IEEE Netw. 24 (5) (2010).
[36] N. Eikmeier, et al., Revisiting power-law distributions in spectra of real world networks, in: Proceedings of ACM SIGKDD, 2017.
[37] M.E. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2) (2001).
[38] C. Song, et al., Modelling the scaling properties of human mobility, Nat. Phys. 6 (10) (2010).
[39] L.A. Adamic, et al., Power-law distribution of the world wide web, Science 287 (5461) (2000).
[40] A. Clauset, et al., Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009).
[41] D. Ferreira, et al., AWARE: mobile context instrumentation framework, Front. ICT 2 (2015).
[42] A.L. Goldberger, et al., Fractal dynamics in physiology: alterations with disease and aging, Proc. Natl. Acad. Sci. USA 99 (suppl 1) (2002).
[43] B. Alipour, et al., Flutes vs. cellos: Analyzing mobility-traffic correlations in large wlan traces, in: IEEE Infocom, 2018.
[44] M. Musolesi, et al., A community based mobility model for ad hoc network research, in: Proceedings of the ACM REALMAN, 2006.
[45] Y. Zheng, et al., Geolife: A collaborative social networking service among user, location and trajectory., IEEE Data Eng. Bull. 33 (2) (2010).
[46] Y. Zheng, L. Zhang, et al., Mining interesting locations and travel sequences from GPS trajectories, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 791–800.
[47] P. Widhalm, et al., Discovering urban activity patterns in cell phone data, Transportation 42 (4) (2015).
[48] C. Gros, Complex and Adaptive Dynamical Systems, Springer, 2010.
[49] C. Song, et al., Limits of predictability in human mobility, Science 327 (5968) (2010).
[50] M.B. Kjærgaard, et al., Indoor positioning using GPS revisited, in: IEEE PerCom, Springer, 2010.
[51] R. Pastor-Satorras, et al., Epidemic processes in complex networks, Rev. Modern Phys. 87 (3) (2015).
[52] S. Lai, et al., Effect of non-pharmaceutical interventions to contain COVID-19 in China, Nature 585 (7825) (2020).
[53] P. Nouvellet, et al., Reduction in mobility and COVID-19 transmission, Nature Commun. 12 (1) (2021) 1–9.