Individual mobility in context: from high resolution trajectories to social behaviour

Laura Alessandretti

A doctoral dissertation completed for the degree of Doctor of Philosophy

Department of Mathematics
School of Mathematics, Computer Science and Engineering
City, University of London

London – January 2018

Laura Alessandretti :
Individual mobility in context:
From high resolution trajectories to social behaviour, © January 2018

SUPERVISOR:
Andrea Baronchelli
SECOND SUPERVISOR:
Mark Broom

EXAMINERS:
Filippo Simini
Elsa Arcaute

ABSTRACT

Understanding human mobility can help creating solutions to society-wide issues, from urban planning and traffic forecasting, to the modelling of epidemics. Existing studies have shown that knowledge on how single individuals take spatial decisions is fundamental for modelling collective mobility patterns. However, individual mobility remains poorly understood, also due to the lack of suitable data. In this thesis, we use novel datasets to characterize and model mobility in relation to other individual aspects: social behaviour, personality, and demographic attributes. Our study focuses on mobility across unprecedented spatial ranges, from $\sim$ 10 m to $\sim$ 10000 Km, and temporal scales, from seconds to years.

PUBLICATIONS

This thesis is based on the following papers:

[I] Laura Alessandretti, Piotr Sapiezynski, Sune Lehmann, and Andrea Baronchelli. "Multi-scale spatio-temporal analysis of human mobility." In: *PloS one* 12.2 (2017), e0171686.

[II] Laura Alessandretti, Piotr Sapiezynski, Sune Lehmann, and Andrea Baronchelli. "Evidence for a conserved quantity in human mobility." In: *arXiv preprint 1609.03526 (under review at Nature Human Behaviour)* (submitted in 2017).

[III] Laura Alessandretti, Sune Lehmann, and Andrea Baronchelli. "Individual mobility and social behaviour: Two sides of the same coin." In: *arXiv preprint 1801.03962 (submitted to EPJ Data Science)* (2018).

[IV] Laura Alessandretti, Kaiyuan Sun, Andrea Baronchelli, and Nicola Perra. "Random walks on activity-driven networks with attractiveness." In: *Physical Review E* 95.5 (2017), p. 052318.

[V] Laura Alessandretti, Márton Karsai, and Laetitia Gauvin. "User-based representation of time-resolved multimodal public transportation networks." In: *Open Science* 3.7 (2016), p. 160156.

Other publications:

[VI] Abeer ElBahrawy, Laura Alessandretti, Anne Kandler, Romualdo Pastor-Satorras, and Andrea Baronchelli. "Evolutionary dynamics of the cryptocurrency market." In: *Open Science* 4.11 (2017), p. 170623.

ACKNOWLEDGMENTS

I have worked on this thesis surrounded by a vibrant and warm research community. I am very thankful:

To Andrea Baronchelli who has guided me through my PhD with care and trust, challenging me to work to my full potential. I feel privileged for the time he spent with me discussing research and career. Andrea has shown me how to work with passion, dedication and ambition, and I will try to follow his example.

To Sune Lehmann, who contributed substantially to the ideas presented in this thesis and who has given me the unparalleled opportunity to play with rich data and amazing hardware.

To the staff within the Mathematics Department at City, who have welcomed me in a friendly research environment. Among them, a special thanks goes to Anne Kandler, Alessandro De Martino, Mark Broom, Andreas Fring, and Yang Hui He.

To my office mates, who have brought light to a dark room. Thanks to Lleonard Rubio, Adam Varga, Davide Bianchini, Veronika Witzke, Malte Probst, Niamh Farrell, Hamish Forbes, Johan Bauer, Patrick Serwene, Abrar Ali, Roberta Amato and Antoine Pierson. A special thanks to Abeer ElBahrawy, for sharing the joy and sorrow of our common fate.

To Piotr Sapiezynski, Radu Gatej, and Vedran Sekara who have worked closely with me and efficiently helped access and process data.

To my friends from the Complex Networks group at Queen Mary University, who have shared with me the good and the bad of living as a PhD student in London. Thanks to Federico Battiston, Valerio Ciotti, Moreno Bonaventura, Jacopo Jacovacci, Iacopo Iacopini, Andrea Santoro and Piero Mazzarisi.

To Marco De Nadai and Simone Centellegher, for carefully reading my work, and for helping with feedback and ideas.

To my Master's thesis advisors Laetitia Gauvin and Marton Karsai, who showed me the way to start a career in research. To their research groups at the ISI foundation and the IXXI Insitute of Com-

plex Systems, among which I felt and still feel welcome.

To Luca Aiello, Rossano Schifanella, Nicola Perra and Miriam Redi for insightful discussions on life in research, and great time in London. An additional thanks to Nicola, for organizing an unforgettable summer school.

To the Machine Intelligence and Data Science group at Information Sciences Institute, University of Southern California, who hosted me for a three months internship. A special thanks goes to Emilio Ferrara and Aram Galstyam who warmly welcomed me and worked with me during my stay.

To the friends from yrCSS, with whom I had fun organising activities for young researchers. Thanks to Elisa Omodei, Federico Botta, Alice Patania, Giovanni Petri, Alberto Antonioni, Aleksandra Aloric, Massimo Stella and Francesca Lipari. Thanks to the Complex Systems Society, in the persons of Alain Barrat and Yamir Moreno, for their support.

To the Pint of Science central team in London, who has given me the chance to join the organization of a unique event.

To Elsa Arcaute and Filippo Simini, who have kindly accepted to act as examiners for this thesis and my defence.

Finally, I wish to thank my family and friends for bearing with me through all the stages that led me to write this thesis.

DECLARATION

I, Laura Alessandretti, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

London , January 2018

Laura Alessandretti

CONTENTS

1

INTRODUCTION

Mobility, or 'the tendency to move between places' [1] is a key aspect of human life, to which we allocate resources, including time [2], money [3] and energy [4]. Existing studies have emphasized that knowledge on how single individuals take spatial decisions is fundamental for modelling collective mobility patterns [5]. This understanding can help creating new solutions to society-wide issues, from urban planning [6] and traffic forecasting [7], to the modelling and simulation of epidemics [8]. Despite its importance, individual spatial behaviour remains poorly understood, partly due to the lack of high-resolution and comprehensive datasets. In this thesis, we use novel multi-dimensional data to characterize and model mobility in relation to other individual aspects: social behaviour, personality, and demographic attributes.

Our work builds upon over a century of academic activity, spanning several disciplines and encompassing a variety of topics, from the movements of large groups of people to the spatial choices of single individuals. The study of mobility patterns at the collective level was pioneered in Geography: In the late 19th century, researchers exploited census data to quantify migrations within the UK [9]. Their research laid the groundwork for later studies on collective movements carried out in Sociology [10], Economics [11], Statistics [12], and, some decades later, Transportation planning and Urban Engineeing [13]. In the same period, researchers in Sociology [14–16] and Economics [17] conducted the first studies on time-allocation using self-administered travel diary. Their work aimed at understanding how individuals allocate time to different activities and places, but neglected the role played by the physical distance between locations.

In 1969, Torsten Hägerstraand's paper 'What about people in regional science?' [5] marked a turning point in the history of human mobility studies. In the article, the author suggested that, when it comes to the interaction between humans and space, '*there are fundamental direct links [...] between the micro-situation of the individuals and the large scale aggregated outcome*'. His research framework, later referred to as 'time-space geography', linked previous studies on time-allocation with those on collective mobility patterns and space utilization. It is worth noticing that, concurrently, the economist Thomas Schelling developed one of the earliest agent-based models [18], to assess the same problem: How the interactions between autonomous

agents determine the emergence of collective phenomena. Until then, this question had been addressed mostly in statistical physics to understand natural phenomena. Today, it underlies researches across various fields of humanities and the natural sciences [19]. In particular, it is the main focus of the Complex Systems Science [20, 21], whose methods and intuitions will be widely adopted along this thesis.

Hägerstraand microscopic theory fostered the development of new research fields that are still lively today. Transportation researchers proposed new ways to model travel demand such as disaggregate and activity based models [22–25]; Behavioural geographers focused on the cognitive processes underlying spatial decision making [26–28]. Until recently, however, theoretical advancements were hindered by the lack of resoluted empirical data on human displacements. The main data sources remained census, surveys and self-reported travel diaries, which suffered from limitations including limited resolution, small sample sizes and reporting biases [29, 30].

In the early 21st century, the introduction and subsequent diffusion of mobile phone devices and other positioning techonologies, marked another decisive step for mobility studies. As we will see along this thesis, today human trajectories can be inferred by various sources including Wireless RFID sensors data, bluetooth records [31], mobile phone call logs [32], location based social networks information [33], and data collected from GPS devices [34].
Concurrently, the Complex Systems Science field has experienced a rapid growth and brought together tools and intuitions from Complex Networks Theory [35], Statistical Physics [36], and Machine Learning [37] to study non-linear, and interconnected systems previously considered intractable. This framework is ideal to investigate real-world systems where, just as in Hägerstraand's representation of human mobility, interacting components generate aggregated outcomes.

These recent developments have renewed the interest of the scientific community for human mobility [38], leading to remarkable advancements that we will describe along this thesis. On the one hand, we have better understanding of collective movements such as migration [39–42] and urban flows [6, 43–45]. On the other hand, we have a clearer picture of how single individuals take spatial decisions [7, 46]. Despite these progresses, many questions have remained opened. The lack of evenly sampled data, necessary to access short temporal scales (seconds to minutes), has hindered a fine-grained description of individual behaviour [47, 48]. The evolutions of mobility across months and years have been poorly investigated, due to the lack of longitudinal datasets. Finally, the unavailability of multi-channel datasets has

partially impeded the investigations of the connections between mobility and other aspects of behaviour.

In this thesis, we address these challenges relying on novel datasets and tools from Complex Systems Science including statistical physics, probability theory, linear algebra, machine learning, agent-based modelling, complex networks theory, numerical simulations, random walks, and stochastic models. Our contribution is twofold: First, we focus on a wide range of spatial and temporal scales, from meters to thousands of kilometres, from seconds to years. As it will be clear in the following chapters, this approach allows to broaden the existing knowledge and improve current models. Secondly, we study mobility in relation to other individual traits, such as personality and social behaviour. To this end, our research framework includes theoretical concepts from human geography, the social sciences, and personality psychology. Our results come from the analysis of novel high-resolution longitudinal datasets consisting of WiFi, GPS, and call detail records collected from mobile phones. High-resolution data allows to describe individual mobility behaviour in an entirely new regime, including displacements between neighbouring buildings, and visits lasting as little as few minutes. In chapter 2, we describe these datasets and the novel techniques developed to pre-process them.

Our research builds upon previous knowledge on three issues that have raised growing interest within the scientific community [38]. Below, we present how the thesis is structured around these three questions. Our main findings, presented in chapters 3 to 7 are supported by thorough analyses. To help the reader following the narrative of the text, results of robustness and sensitivity tests are reported at the end of the thesis, in appendices A to D.

What are the statistical properties of human trajectories?

Trajectories of individuals living in western society seem to display homogeneous statistical properties [32, 49], similar to those characterizing foraging patterns of animals and hunter-gatherers [50]. This evidence has been attributed to various factors including the structure of the physical space [51–55] and the ancestral necessity to optimize search for food and resources [56–59]. This topic is the focus of chapter 3. In the chapter, we contribute to the field with a review of over 40 studies on the statistics of human displacements. We show that they disagree, to a large extent, due to the heterogeneity of the considered data. We resolve this controversy by analysing the properties of trajectories with high spatio-temporal resolution, long duration, and large sample size. We show that the distribution of displacements

and waiting times between displacements are best described by log-normal distributions, but power-law distributions are selected when only large spatial or temporal scales are selected. The chapter is based on research published in [I] and the literature review included in it is regularly updated online [1].

How do humans allocate time among different locations?

Under the time-space geography framework, individuals describe a path in time and space, but their possibilities are limited by constraints [5]. 'Capability constraints' are due to biological limitations, 'Coupling constraints' relate to the necessity of joining other individuals and access resources, 'Authority constraints' are exerted by external authorities including national laws and social norms. The study of visitation patterns has shed light how humans balance the trade-off between exploring novel opportunities and exploiting known options while subject to these constraints [45, 48, 60, 61], allowing to develop predictive model of mobility [7, 46]. The existing literature, however, has focused on patterns recurring within time periods typically shorter than six months. In chapter 4, we study visitation patterns in an entirely new regime, characterized by the slowly-evolving dynamics taking place across months and years. We contribute to the field with the discovery that routines are unstable in the long term because of the continual exploration of new locations, but the number of locations an individual visits regularly is conserved over time. The work is based on research presented in [II] and currently under peer-review.

What are the connections between individuals' mobility and social behaviour?

Multi-channel datasets allow to investigate how the need to join other individuals affects our decisions on where and when to move in space. Existing researches, however, have focused on the coupling between mobility and social proximity only at the level of pairs of individuals [37, 62–66]. In chapter 5, we study for the first time the connection between how single individuals take decisions in the spatial and social realms. We show that there is a connection between the way in which individuals explore new resources and exploit known assets in the social and spatial spheres. This connection can be partly explained in terms of personality traits. The work is based on research presented in [III] and currently under peer-review.

1 http://lauraalessandretti.weebly.com/plosmobilityreview.html

In chapter 6, we dig more into the study of social behaviour and we develop an agent-based model of social interacftions. We adopt a time-varying network model perspective[67–69], where nodes are individuals and links with limited duration are the interactions between them. We present a new dynamic model where tie formation is driven by two distinctive and persistent properties of individuals: their *activity*, or propensity to engage in social interactions, and their *attractiveness*, or propensity to attract connections. We show that these stable dispositions have a major impact on diffusion processes spreading on the network. Future developments of the presented model will include a spatial component to account for the correlations between social and spatial individual dispositions. The chapter is based on work published in [IV].

Information on the displacements of single individuals can be aggregated to study flows of people travelling between different areas. In the last part of this thesis, we shift the attention to the study of mobility patterns at the urban scale. In chapter 7, we develop a framework to compare the transportation network and the commuters flows within a given urban agglomeration. In line with the rest of the thesis, we propose a user-based representation of transportation systems, that accounts for individuals' need to limit the total travel time and the number of line changes. We will use this method allow to assess the efficiency of public transportation systems of several French cities. The chapter is based on research published in [V].

2

MOBILITY DATA DESCRIPTION AND PRE-PROCESSING

Our research is, to a large extent, based on the analyses of trajectories from two novel datasets with fixed temporal sampling and long-term duration: the displacements of $\sim$ 37000 users of the SONY LifeLog mobile application, followed for 19 months, and those of the 850 participants in the Copenhagen Networks Study [70] longitudinal experiment, sampled every $\sim 16s$ for 24 months. Fixed sampling enables to capture patterns beyond regular ones such as home-work commuting, avoiding the sampling biases of mobile phone calls and location-based social networks data [47, 48]. Long duration enables to capture mobility behaviour in an entirely new regime, characterized by a slowly-evolving dynamics. Most results were corroborated with data from two other experiments: the Lausanne Data Collection Campaign[71] and the Reality Mining dataset [72]. Some of these sources include information on individuals' interactions across multiple channels (phone calls, sms, Facebook), and their personality extracted from questionnaires (for more info see chapter 2). In table 2.1, we present important charachteristics of the datasets considered: the number of individuals N, the duration of data collection T, the spatial (δx) and temporal (δt) resolution and the time coverage. The time coverage is used to quantify the fraction of time a user's position is known. Individuals' location is unknown, for example, when users' turn off their phone (all datasets), when they disable the sensor from which the trace is sensed, WiFI (CNS) or GPS (Lifelog), or, in the case of the MDC and RM datasets, when they can not connect to an antenna. For the CNS dataset, moreover, we consider WiFi traces, and in some cases, the set of scanned WiFi access points can not be geolocalized. My ongoing work include aggregating the CNS WiFi and GPS traces to obtain higher time-coverage.

This chapter is organized as follows: In sections 2.1 to 2.4, we provide a description of the 4 datasets considered (see table 2.1) and the data pre-processing applied to extract stop-locations from GPS and WiFi sensors data. Data pre-processing relies both on previously developed and novel methodologies. In section 2.5, we compare some properties of these datasets with the state-of-the-art literature.

	N	δt	T	δx	TC
Lifelog	36898	change in motion	19 months	10 m	0.57*
Lifelog (selected users)	2272	change in motion	19 months	10 m	0.66*
CNS	850	16 s	24 months	10 m	0.84
MDC	185	60 s	19 months	100-200m	0.73
RM	95	16 s	10 months	100-200m	0.93

*computed from data including only stop-locations, after pre-processing internal at SONY Mobile

Table 2.1: **Characteristics of the mobility datasets considered.** N is the number of individuals, δt the temporal resolution (for the Lifelog dataset, location is recorded at every change in motion), T the duration of data collection, δx the spatial resolution, TC the median weekly time coverage, defined as the fraction of time an individual's location is known. Note that TC for Lifelog trajectories is computed from data including only stop-locations, where users stop for more than 10 minutes. In the other datasets, stop-locations account on average for 80/90% of the total TC. We also validated results considering a subset of Lifelog users with high time coverage (second row). See also fig. 2.1, fig. 2.2, fig. 2.3.

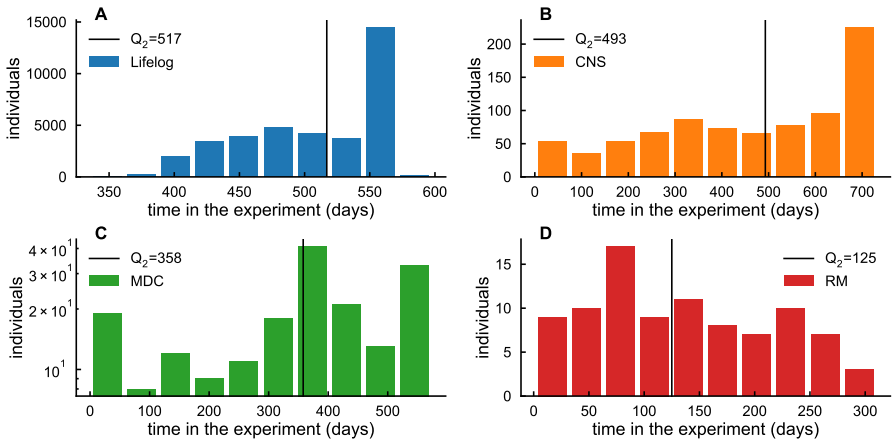

Figure 2.1: **Long duration of the datasets considered** Frequency histogram of individuals' based on collection duration for Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)**. The black line is the median.

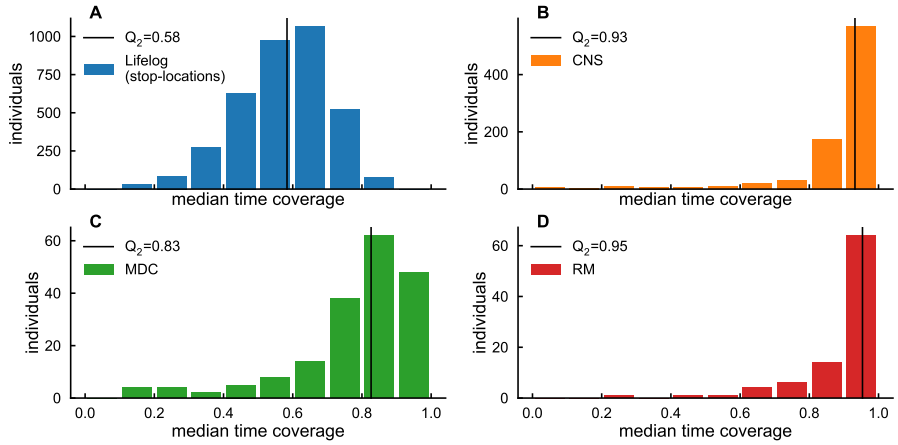

Figure 2.2: **High individual temporal resolution** Frequency histogram of individuals' median weekly time coverage for Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)**. The black line is the median.

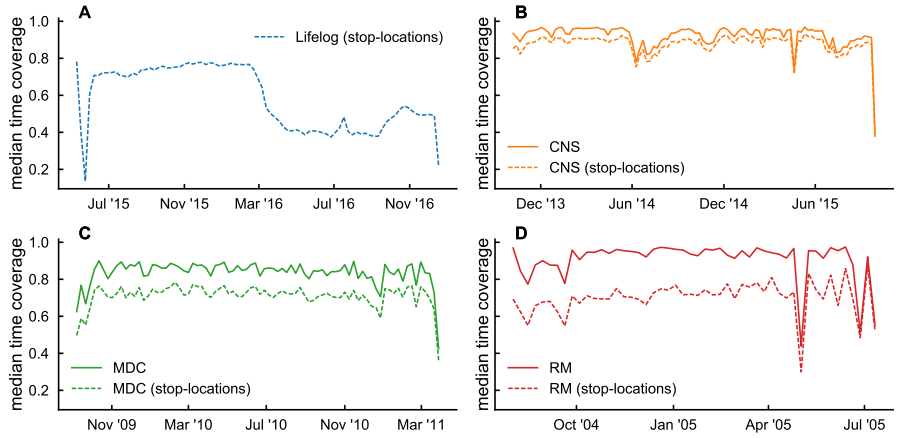

Figure 2.3: **High temporal resolution** Median weekly time coverage as a function of time for the Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)** datasets. Filled lines are computed considering all locations, dashed lines are computed considering only stop-locations.

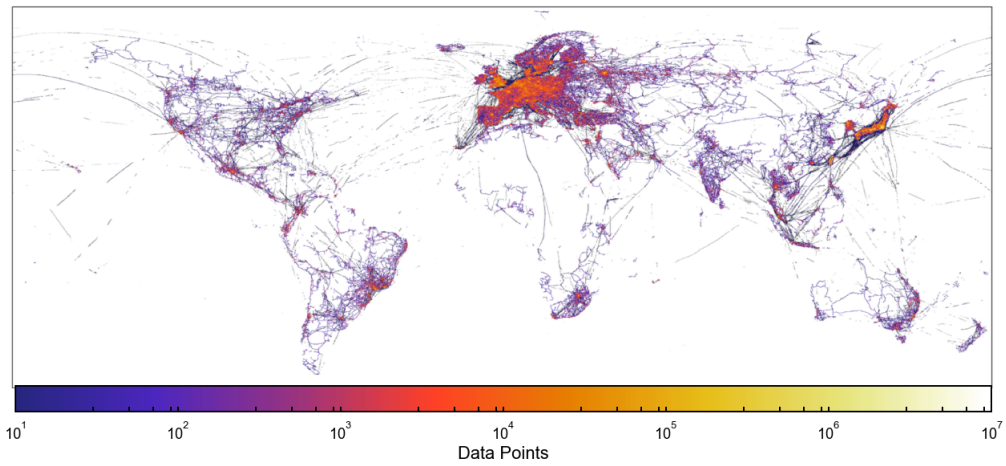

Figure 2.4: **Broad spatial coverage of the Lifelog dataset.** Heatmap showing the spatial distribution of data points in the Lifelog dataset.

2.1 SONY LIFELOG DATASET

2.1.1 *Description*

The dataset consists of anonymized GPS location data for $\sim$ 37000 users of Lifelog, an activity tracker app for Android phones, collected between 2015 and 2016. Lifelog users are geo-localised across the world (see fig. 2.4), and are aged between 18 and 65 years old, with average at 36 years old. About 1/3 of users are female. Data is not collected with a fixed time interval. Instead, the app gets updates when there is a change is the motion-state of the device (if the accelerometer registers a change). Location estimation error is below 100 meters for 93% of data points. To preserve privacy, GPS traces were pre-processed (internally at SONY Mobile) to infer stop-locations using the method described below. Data collection has been approved by the Sony Mobile Logging Board and informed consent has been obtained for all study participants according to the Sony Mobile Application Terms of Service and the Sony Mobile Privacy Policy.

2.1.2 *Pre-processing*

SELECTION OF USERS We have selected users who have data for at least 365 days ($\sim$ 36.000*users*).

DEFINITION OF LOCATIONS GPS data is pre-processed (internally at Sony Mobile) to infer stop-locations using the distance grouping method described in [73]. The method is built on the idea that a stop corresponds to a temporal sequence of locations within a maximal

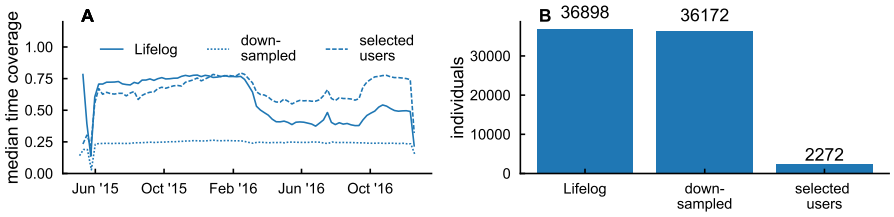

Figure 2.5: **Lifelog dataset: pre-processing (A)** Median weekly time coverage across the population as a function of time for the raw Lifelog dataset (filled line), after downsampling (dotted line) and after user selection (dashed line). **(B)** Number of individuals in the dataset (filled bar), after downsampling (dotted bar), and after user selection (dashed bar).

distance d_{max} from each other. In this work, we present results obtained for $d_{max} = 50m$, $d_{max} = 30m$, and $d_{max} = 40m$ (see chapter 4).

DATA CLEANING During the data collection period, the app settings changed causing a considerable change in time coverage for a subset of users (see fig. 2.3). We propose two methods to solve this issue (see fig. 2.5):

(a) Users selection: We consider only the subset of users for which there is no change in time-coverage over time ($\sim$ 6% of all users)

(b) Temporal down-sampling: We down-sample data to achieve constant time-coverage across time. The method used relies on:

- Find for each user i the week w_m with lowest weekly time-coverage $tC(w_m)$.
- Down-sample weeks with weekly time-coverage higher than w_m by selecting a random sample of total duration $tC(w_m) *$ 60 minutes.

Results presented in chapter 4 are produced with method (a) and (b).

2.2 COPENHAGEN NETWORKS STUDY

2.2.1 *Description*

The Copenhagen Networks Study experiment took place between September 2013 and September 2015 [70] and involved 851 Technical University of Denmark students ($\sim$ 22% female, $\sim$ 78% male) typically aged between 19 and 21 years old. Participants' position over time was estimated combining their smart-phones WiFi and GPS data using the method described in [74] (see below). The location estimation error is below 50 meters in 95% of the cases. Participants calls and sms activity was also collected as part of the experiment. Individuals' background information were obtained through a 310 questions

survey including the Big Five Inventory [75], measuring five broad domains of human personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism). Data collection was approved by the Danish Data Protection Agency. All participants provided informed consent.

2.2.2 *Pre-processing*

Location data is obtained combining Wi-Fi data (sampled every $\sim 15s$) with GPS data (high spatial resolution). The following methodology was implemented to estimate the sequences of individuals stop-locations:

ESTIMATION OF WI-FI ACCESS POINTS (AP) POSITION Access Points (AP) positions were estimated using participants' sequences of GPS scans. We discarded *mobile APs*, that are located on buses or trains, and *moved APs* that were displaced during the experiment (for example by residents of Copenhagen changing apartment, taking their APs with them). Then, we considered all WiFi scans happening within the same second as a GPS scan to estimate APs location. The APs location estimation error is below 50 meters in 95% cases. Most of the APs are located in the Copenhagen area (see [74] for a detailed description of the methodology).

DEFINITION OF LOCATIONS We find locations by clustering APs based on the distance between them. First, we built the indirect graph of APs simultaneous detection $G = (V, E)$. V is the set of geo-localised APs, links $e(j,k)$ exist between pairs of access points that have ever been scanned in the same 1 min bin by at least one user. Then, we compute the physical distances $dist(j,k)$ for all pairs of $(j,k) \in E$. and we consider the set of links $E_D \subset E$ such that $dist(j,k) < d$, where d is a threshold value, to define a new graph $G_d = (V, E_d)$. Finally, we define a *location* as a connected component in the graph G_d, with coordinates equal to the median latitude and longitude of nodes in the component. For $d = 5m$ the maximal distance between two APs in the same location is smaller than $10m$ for most locations and at most $\sim 200m$ (see fig. 2.6-A). The number of APs in the same location is lower that 10 for most locations, but reaches ~ 1000 for dense areas such as the University Campuses (see fig. 2.6-B).For this reason, we chose a threshold of 5 m. The AP localization error is below $5m$ for 50% of APs in the dataset. However, it is important to notice that our work focuses on a specific subset of the APs: Those where individuals stop for more than 10 consecutive minutes, and that, additionally, are seen multiple times. The localization accuracy is higher for these APs. In fact, the more the visits, the largest the number of GPS-APs simultaneous observations (see [74], Figure 2). For example, for 30

simultaneous observations, the localization error is below 5m for 75% of APs. Note that GPS scans on average every 5 minutes. An example of APs clustering for $d = 5m$ and $d = 10m$ is shown in fig. 2.6-C and fig. 2.6-D. We show below that our findings do not depend on the choice of the threshold (see chapter 4). Compared to methods defining stops based on the GPS trajectory [73], this method allows to assign locations a unique id for all users. In the case considered, it works well in all areas, including those that are highly visited and where access points are densely distributed (such as the University Campus, see fig. 2.6) . However, the method may not be scalable if much larger populations are considered.

TEMPORAL AGGREGATION Data was aggregated in bins of length 1 min, where for each bin we selected the most likely location.

2.3 LAUSANNE MOBILE DATA CHALLENGE

Data was collected by the Lausanne data Collection Campaign between October 2009 and March 2011. The campaign involved an heterogeneous sample of $\sim$ 185 volunteers with mixed backgrounds from the Lake Geneva region (Switzerland), who were allocated smartphones [76]. In this work we used GSM data, that has the highest temporal sampling ($\sim$ 60 seconds, see fig. 2.2). Following Nokia's privacy policy, individuals participating in the study provided informed consent [76]. The Lausanne Mobile Data Challenge experiment involves 62% male and 38% female participants, where the age range 22-33 year-old accounts for roughly 2/3 of the population [77].

2.4 REALITY MINING

The Reality Mining project was conducted from 2004-2005 at the MIT Media Laboratory. It measured 94 subjects using mobile phones over the course of nine months. Of these 94 subjects, 68 were colleagues working in the same building on campus (90% graduate students, 10% staff) while the remaining 26 subjects were incoming students at the university's business school [72]. An application installed on users' phones continuously logs location data from cell towers ids at fixed rate sampling (see fig. 2.2). The study was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). Subjects were provided with detailed information about the type of information captured and provided informed consent [78].

2.5 COMPARISON WITH PREVIOUS RESEARCH

Our datasets displays statistical properties consistent with previously analysed data on human mobility.

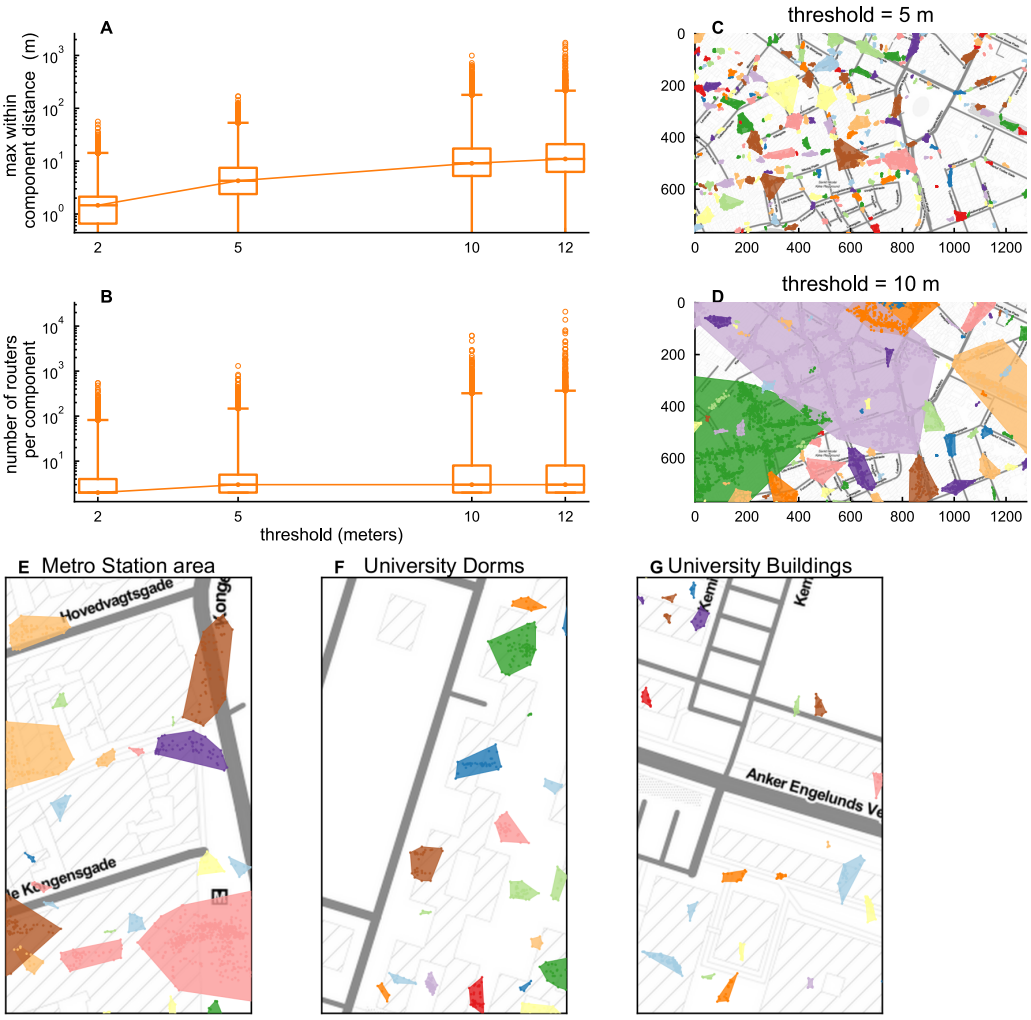

Figure 2.6: **CNS dataset: Different definitions of locations (A)** The boxplots of the maximal distance between pairs of geo-localized APs forming a location, as a function of the threshold d used to merge APs. Boxes are set at the 1st and 3rd quantile, while whiskers at 2.5% and 97.5%. **(B)** The boxplots of the locations size (number of APs) as a function of the threshold d. **C-D**) An example of the clustering of APs located within Copenhagen city for thresholds $d = 5m$ (C) and $d = 10m$ (D). Dots corresponds to geo-localized APs, colored according to the location they belong to. Note that APs are typically geo-localized outdoor due to poor GPS signal inside buildings [74]. Coloured regions are the convex hulls of the set of APs in a same location. Grey lines are streets. **E-G**) Three examples of APs clustering for thresholds $d = 5m$.

- **Rank-frequency distribution of locations:** The visitation frequency of a location, defined as the fraction of visits to that location, goes with the location rank r as $r^{-\zeta}$, with $\zeta \sim 1$.(fig. 2.7-A). Our result is consistent with [32], where the authors found $f(r) \propto 1/r$, and [46], where it was found $f(r) \propto r^{-1.2}$.

- **Distribution of displacements:** The distribution between consecutive jumps $P(\Delta r)$ has a power law tail (fig. 2.7-B), with exponent $\beta = 1.81$. Gonzalez et. al [32] found $\beta = 1.77$ for the truncated power-law distribution, Song et. al [46] found a power-law tail with exponent $\beta = 1.55$.

- **Growth of the radius of gyration:** Individuals' total radius of gyration (see [32], SI for definition) growth across time is consistent with the logarithmic growth described in [46](fig. 2.7-C).

- **Distribution of the radius of gyration:** Individuals are distributed heterogeneously with respect to their total radius of gyration measured at the end of the experiment, with the probability distribution $P(r_g)$ (fig. 2.7-D) decaying as a power-law with coefficient $\beta = -1.47$. This is comparable with the results found in [32], $\beta = -1.65$ and [46] $\beta = -1.55$, where both studies relied on CDRs.

- **Returners and explorers:** In accordance with Pappalardo et al. [79], the distribution of r_g^2/r_g is bimodal (fig. 2.8A and B), where r_g is the radius of gyration computed across a window of 20 weeks and r_g^2 is the radius of gyration computed within the same window including only the top 2 locations (see [79] for the definition). Hence, within each window, an individual can be categorized as either a *returner* (if $r_g^2/r_g < 0.5$) or as an *explorer* (if $r_g^2/r_g > 0.5$). We find that these categorization is stable in time for $\sim 50\%$ of individuals (fig. 2.8D).

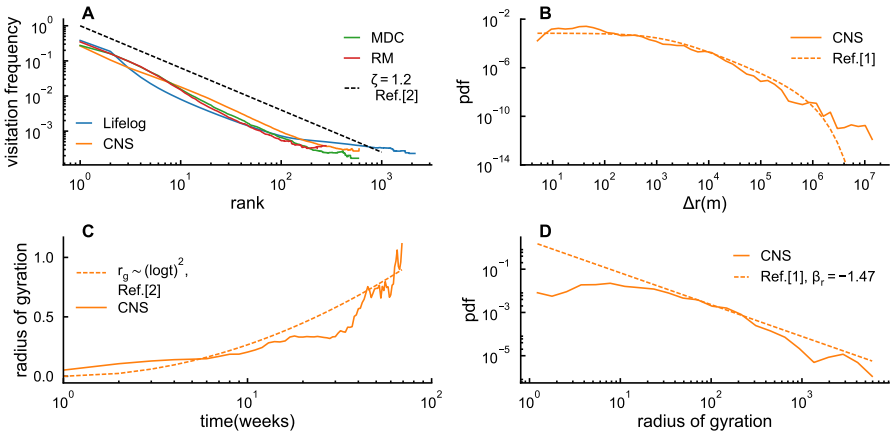

Figure 2.7: **Agreement with previous research. A**) The average visitation frequency f_k as a function of a location rank for the four datasets (filled lines) and the power law fit $f_k \propto k^{\zeta}$, with $\zeta = 1.2$ found in [46] (dashed line). **(B)** CNS dataset: The probability density distribution of jump lengths (in m) between consecutive stop-locations (filled line), and the truncated power-law found as the best fit in [32]. **(C)** CNS dataset: Evolution of the average radius of gyration r_g as a function of time (filled line) and a logarithmic curve $r_g \sim (logt)^2$ found in [46] as the best fit. **(D)** CNS dataset: The probability density function of individuals final radius of gyration (filled line) and the power-law fit (dashed line) found in [32].

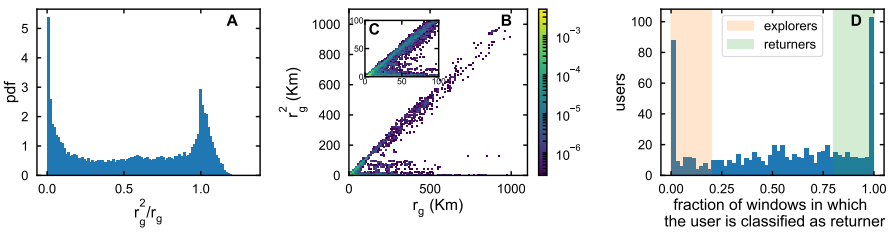

Figure 2.8: **Returners and Explorers Dichotomy in the CNS dataset. (A)** Distribution of r_g^2/r_g, where r_g is the total radius of gyration and r_g^2 is the radius of gyration computed considering only the top 2 locations. These quantities are computed across windows of length 20 weeks for all individuals. **(B)** Heatmap displaying the joint probability density $p(r_g, r_g^2)$. **(C)** The joint probability density $p(r_g, r_g^2)$, for r_g and $r_g^2 < 100Km$. **(D)** Based on the definition in [79], *returners* have $r_g^2/r_g > 0.5$, while *explorers* have $r_g^2/r_g < 0.5$. The figure is the histogram of individuals based on the fraction of times they are assigned to the *returner* category in a window of 20 weeks (blue bars). For our study (see chapter 4), we consider as *returners* (green shaded area) and *explorers* (red shaded area) only individuals falling in the same category in at least 75% of time-windows (about 50% of all CNS participants).

Recently, the writer listened to a conversation between two educators who were talking about a survey made on students' reasons for choosing a certain small college. One educator asked the other to guess the most important reasons. A half dozen were suggested. "You have missed the most important," was the reply. "It is simply proximity."

— Samuel A Stouffer [10]

3 STATISTICS OF DISPLACEMENTS

What are the characteristic properties of human trajectories? Since decades, researchers have tried to provide a statistical description of human motion in order to understand its dynamics and design reliable predictive models.
In this chapter, we review over 40 studies on the statistics of human trajectories (see section 3.1). We show that they disagree, to a large extent, due to the heterogeneity of the considered datasets. In section 3.2, we present the analysis of the trajectories collected by the Copenhagen Networks Study (see section 2.2), that have the best combination of spatio-temporal resolution and sample size among the datasets analysed in the literature to date. Subjects in the experiment are homogeneous with respect to socio-demographic indicators affecting mobility behavior [80], and their displacements are constrained by a similar academic schedule. In the following, we rely on the hypothesis that many statistical properties of mobility are distinctive of human behaviour, and therefore consistent across samples. In fact, several characteristics of CNS students mobility patterns are consistent with previous results. This chapter is based on research published in [I].

3.1 STATE OF THE ART

Trajectories can be understood as series of *displacements* between locations and *pauses* at locations, where an individual stops and spends time (fig. 3.1). Thus, the distribution of waiting times (or pause durations), Δt, between movements and the distribution of distances, Δr, travelled between pauses are among the most studied statistical properties. In fact, 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, these distributions have been widely investigated. There is no agree-

Figure 3.1: **Example of an individual trajectory**. An individual trajectory is composed of pauses (red dots) and displacements (dashed black line). The trajectory shows the positions of one individual across 26 hours. Location is estimated from individual's WiFi scans as detailed in the text and the data is sampled in 1 min bins. Red dots correspond to locations where the individual spent more than 10 consecutive minutes. The coordinates of these locations have been slightly altered to protect the subject privacy. The map was generated with the Matplotlib Basemap toolkit for Python (https://pypi.python.org/pypi/basemap). Map data © OpenStreetMap contributors (License: http://http://www.openstreetmap.org/copyright). Map tiles by Stamen Design, under CC BY 3.0.

ment, however, on which distribution best describes these empirical datasets.

Pioneer studies, based on CDR [32, 46] and banknote records [81], found that the distribution of displacement Δr is well approximated by a power-law, $P(\Delta r) \sim \Delta r^{-\beta}$, (or 'Lévy distribution'[50], as typically $1 < \beta < 3$), and that an exponential cut-off in the distribution may control boundary effects [32]. These findings were confirmed by studies based on GPS trajectories of individuals [49, 51, 82] and vehicles [83, 84], as well as online social networks data [85–87]. It has been noted, however, that power-law behaviour may fail to describe intra-urban displacements [88]. Other analyses, based on online social network data [89–91] and GPS trajectories [92–95] 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 [96, 97] suggested that displacements may also obey log-normal distributions, $P(\Delta r) \sim (1/\Delta r) * e^{-(\log \Delta r - \mu)^2/2\sigma^2}$. In Ref. [51], the authors found that this is the case also for single-transportation trips.

Fewer studies have explored the distribution of waiting times between displacements, Δt, 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 [46, 98], and individuals GPS trajectories [49, 82] 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 [52]. Several studies have highlighted the presence of natural temporal scales in individual routines: distributions of waiting times display peaks in that corresponds to the typical times spent home on a typical day ($\sim$ 14 hours) and at work ($\sim 3-4$ hours for a part-time job and $\sim 8-9$ hours for a full-time job)[43, 98, 99].

fig. 3.2 and table 3.1 compare distributions obtained using different data sources. The spectrum of results reflects the heterogeneity of the considered datasets (see fig. 3.2). It is known in fact that data spatio-temporal resolution and coverage has an important influence on the results of the analyses performed [100–102].

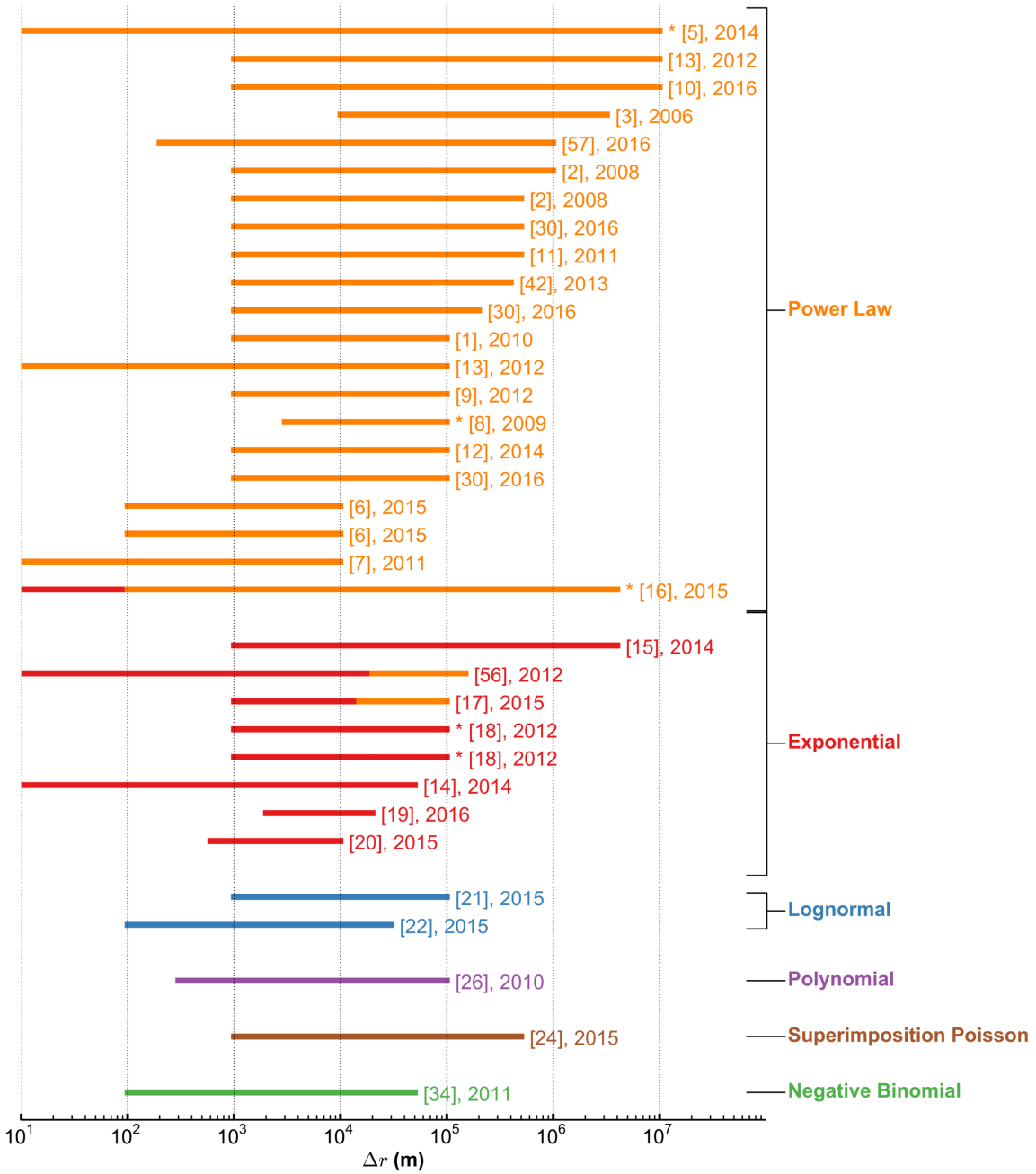

Figure 3.2: **The distribution of displacements $P(\Delta r)$: heterogeneity of results found in the literature.** Each horizontal line corresponds to a different dataset. Lines extend from the minimum Δr (*i.e.* the spatial resolution of the data or the minimum value considered for the fit of the distribution), to the maximal length of displacement considered (both in meters). Colours correspond to the model fitting $P(\Delta r)$ according to the study reported at the end of each line. If the distribution is not unique, but varies for different ranges of Δr, the line is divided in segments. Lines are marked with '*' if the corresponding data is modelled as a sequence of two distributions of the same type with different parameters, for different ranges Δr. Refs [32, 51, 93, 103] analyse more than one dataset. In [88] the authors analyse the same dataset for different ranges Δr. A more detailed table is presented in section "Related Work".

Table 3.1: **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 N of individuals (or vehicles in the case of car/taxi data) involved in the data collection, the duration of the data collection (M $\rightarrow$ months, Y $\rightarrow$ years, D $\rightarrow$ days, W $\rightarrow$ weeks), the range of spatial displacements considered $x_{min} - x_{max}$, the shape of the probability distribution of displacements $P(\Delta x)$ with the corresponding parameters, the temporal sampling δt (u stands for uneven), the shape of the distribution of waiting times $P(\Delta t)$ with the corresponding parameters

	Data type	**N**	**Dur.**	x_{min} x_{max}	$P(\Delta x)$	δt	$P(\Delta t)$
[46] (D1)	CDR	$3.0 \cdot 10^6$	1 Y	1 km 10^2 km	**Power Law** β=1.55 Δr_0=0.00 $k = 100.00$	u	
[46] (D2)	CDR	10^3	2 W	1 km 10^2 km		1 h	**Power Law** β=1.80 Δt_0=0.00 k=17.00
[32] (D1)	CDR	10^2	6 M	1 km 10^3 km	**Power Law** β=1.75 Δr_0=1.50 $k = 400.00$	u	
[32] (D2)	CDR	206	1 W	1 km 500 km	**Power Law** β=1.75 Δr_0=1.50 $k = 80.00$	2 h	
[81]	Bills	$4.6 \cdot 10^5$ bills	1.39 Y	10^2 m 10^2 km	**Power Law** β=1.59 Δr_0=0.00 $k = \infty$	u	
[82] (Geolife)	GPS	165	3.42 Y	10 m 10^4 km	0.01- 10 km **Power Law** β_0=1.25 Δr_0=0.00 $k_0 = \infty$ 10 - 10^4 km **Power Law** β_1=1.90 Δr_1=0.00 $k_1 = \infty$	2 min	**Power Law** β=1.98 Δt_0=0.00 k=∞
[51] (MDC)	GPS	200	1.50 Y	10^2 m 10^4 km	**Power Law** β=1.39 Δr_0=0.00 $k = 6250.00$	10 sec	

Continued on next page

Table 3.1 Continued from previous page

	Data type	N	Dur.	x_{min} x_{max}	$P(\Delta x)$	δt	$P(\Delta t)$
[51] (Geolife)	GPS	182	5.00 Y	10^2 m 10^4 km	**Power Law** β=1.57 Δr_0=0.00 $k = 3891.00$	5 sec	
[49]	GPS	44	5 M	10 m 10 km	**Power Law** β=[1.35-1.82] Δr_0=0.00 k=inf	10 sec	**Power Law** β=[1.45-2.68] Δt_0=0.00 k=∞
[83]	Taxi	50	6 M	10 m 10^4 km	3 - 23 km **Power Law** β_1=4.60 Δr_1=0.00 $k_1 = \infty$ 23 - 10^2 km **Power Law** β_0=2.50 Δr_0=0.00 $k_0 = \infty$	10 sec	
[84]	Taxi	$6.7 \cdot 10^3$	1 W	1 km 10^2 km	**Power Law** β=1.20 Δr_0=0.30 $k = 10.00$	10 sec	
[85]	Flickr	$4.0 \cdot 10^4$		1 km 10^4 km	**Power Law**	u	
[86]	LBSN	$2.2 \cdot 10^5$	1.25 Y	1 km 500 km	**Power Law** β=1.88 Δr_0=0.00 $k = \infty$	u	
[87]	Twitter	$1.3 \cdot 10^7$	1 Y	1 km 10^2 km	**Power Law** β=1.62 Δr_0=0.00 $k = 0.00$	u	
[88]	LBSN	$9.2 \cdot 10^5$	6 M	1 km 3200 km	**Power Law** β=1.50 Δr_0=2.87 $k = \infty$	u	
[88] (intracity)	LBSN	$9.2 \cdot 10^5$	6 M	10 m 10^2 km	**Power Law ("poor")** β=4.67 Δr_0=18.42 $k = \infty$	u	
[89]	LBSN	$2.6 \cdot 10^5$	1 M	10 m 50 km	**Exponential** k=5.58	u	
[90]	LBSN	$5.0 \cdot 10^5$	1 M	1 km 4000 km	**Exponential** k=333.00	u	

Continued on next page

Table 3.1 Continued from previous page

	Data type	N	Dur.	x_{min} x_{max}	$P(\Delta x)$	δt	$P(\Delta t)$
[91]	Twitter	$1.6 \cdot 10^5$	1.58 Y	10 m 10^4 km	0.01 - 0.1 km **Exponential** k_0=13.69 0.1 - 10^2 km **Stretched Power Law** β_1=0.45 Δr_1=0.00 $k_1 = 90.90$ 10^2 - 10^4 km **Power Law** β_2=1.32 Δr_2=0.00 $k_2 = \infty$	u	
[92]	Taxi	803	1.25 Y	1 km 10^2 km	0 - 15 km **Exponential** k_0=2.80 15 - 10^2 km **Power Law** β_1=3.66 Δr_1=0.00 $k_1 = \infty$	30 sec	
[93] (D1)	Taxi	10^4	3 M	1 km 10^2 km	1 - 20 km **Exponential** k_0=4.29 20 - 10^2 km **Exponential** k_1=5.89	1 min	
[93] (D2)	Taxi	10^4	2 M	1 km 10^2 km	1 - 20 km **Exponential** k_0=4.16 20 - 10^2 km **Exponential** k_1=5.65	1 min	
[94]	Taxi	$6.6 \cdot 10^3$	1 W	2 km 20 km	**Exponential** k=[3.96-13.89]	10 sec	
[95]	Taxi	$1.0 \cdot 10^4$	1 M	600 m 10 km	**Exponential**	29 min	**Power Law**
[104]	Travel cards	$2.0 \cdot 10^6$	1 Y	10^2 m 50 km	**Lognormal** μ=9.28 σ=5.83	u	
[96]	Taxi	$3.0 \cdot 10^4$	1.69 Y	1 km 10^2 km	**Lognormal** μ=[0.70-1.30] σ=[0.67-0.86]	1 min	
[97]	Taxi	$1.1 \cdot 10^3$	6 M	10^2 m 30 km	**Lognormal** μ=0.38 σ=0.48	30 sec	

Continued on next page

Table 3.1 Continued from previous page

	Data type	**N**	**Dur.**	x_{min} x_{max}	$P(\Delta x)$	δt	$P(\Delta t)$
[98]	Surveys	10^4	1 Y			self reported	**Power Law** β=0.49 Δt_0=0.00 k=1.45
[34]	Private cars	$7.8 \cdot 10^5$	1 M	1 km 500 km	**Superimposition Poisson**	10 sec	0 - 4h **Power Law** β_0=1.03 Δt_0=0.00 $k_0 = \infty$ 4- 200 h **Lognormal** μ_1=1.60 σ_1=1.60
[43]	Travel cards	626	3 M			u	
[99]	Private cars	$3.5 \cdot 10^4$	1 M	300 m 10^2 km	**Polynomial**	10 sec	**Power Law** β=0.97 Δt_0=0.00 k=∞
[36]	Private cars	$7.5 \cdot 10^4$	1 M	1 km 500 km	0.3 - 20 km **Exponential** 20 - 150 km **Power Law** β_1=3.30 Δr_1=0.00 $k_1 = \infty$	30 sec	**Exponential** k=0.98
[53]	Travel Diaries	230	1 M	1 km 400 km	**Power Law** β=1.05 Δr_0=0.00 $k = 50.00$	self reported	
[105]	Taxi		1 D	200 m 10^3 km	**Power Law** β=2.70 Δr_0=0.00 $k = \infty$		
[103] (D1)	CDR	$1.3 \cdot 10^6$	1 M	1 km 200 km	**Power Law** $\beta = 2.02$	u	
[103] (D2)	CDR	$6 \cdot 10^6$	1 Y	1 km 500 km	**Power Law** $\beta = 1.75$	u	
[103] (D3)	CDR		4 Y	1 km 10^2 km	**Power Law** $\beta = 1.80$	u	

Concluded

First, the datasets considered have different *spatial resolution and coverage*, and few studies have so far considered the whole range of displacements occurring between ~ 10 and 10^7 m (10000 km) (fig. 3.2). The analysis presented below suggests that constraining the study 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 [47, 48], 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 [52, 106], taxi [95], public transportation [104], or walk [49]) capture only a specific aspect of human mobility behaviour.

3.2 A MULTI-SCALE ANALYSIS

In this section, we present the analysis of the mobility trajectories collected by the Copenhagen Networks Study (see section 2.2). Individual trajectories have spatial resolution of ~ 10 m, even sampling every ~ 16 s, and they span more than $\sim 10^7$ m. Previous studies with comparable spatial coverage (fig. 3.2) relied on single-transportation modality data [83], unevenly sampled data [91], or small samples (32 individuals in Ref. [82]). To our knowledge, the Copenhagen Networks Study data has the best combination of spatio-temporal resolution and sample size among the datasets analysed in the literature to date.

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 appendix A.1 for the results obtained with 15 and 20 minutes pausing).

We start by considering the three distributions most frequently reported in the literature (table 3.1), namely

- *The log-normal distribution* of a random variable x, with parameters σ and μ, defined for $\sigma > 0$ and $x > 0$, with probability density function:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}x} e^{-\frac{1}{2}\frac{(\log x - \mu)^2}{\sigma^2}} \tag{3.1}$$

- *The Pareto distribution* (*i.e.* power-law) of a random variable x, with parameter β, defined for $x \geq 1$, and $\beta > 1$, with probability density function:

$$P(x) = (\beta - 1)\,(x)^{-\beta} \tag{3.2}$$

- *The exponential distribution* of a random variable x, with parameter λ, where $x \geq 0$, and $\lambda > 0$, with probability density function:

$$P(x) = \lambda e^{-\lambda x} \tag{3.3}$$

In eq. (3.2) 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 $y = \dfrac{(x - x_0)}{s}$. In eq. (3.1), and eq. (3.3), $P(x)$ 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, $x_0 = 0$ (seeappendix A.1). Note also that Pareto distributions with exponential cut-off (or truncated Pareto) are considered below (see also table 3.1).

3.2.1 *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 Δr using all available data (851 individuals over 25 months). We find that $P(\Delta r)$ is best described by a log-normal distribution (eq. (3.1)) with parameters $\mu = 6.78 \pm 0.07$ and $\sigma = 2.45 \pm 0.04$, which maximises Akaike Information Criterion [107] — among the three models considered — with Akaike weight ~ 1 (fig. 3.3, see also appendix A.1).

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 appendix A.1). In fact, the errors on the value of the parameters reported above are computed by bootstrapping 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

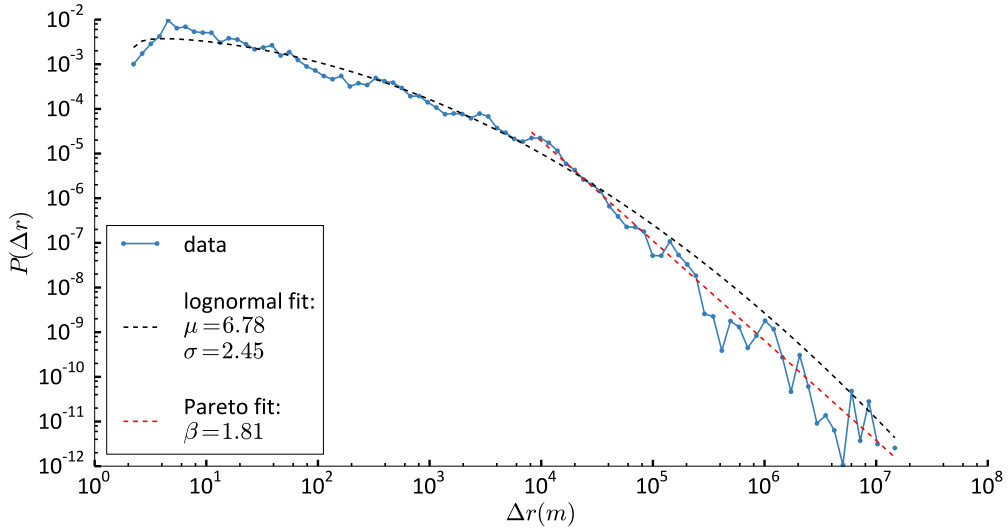

Figure 3.3: **Distribution of displacements**. Blue dotted line: data. Black dashed line: log-normal fit with characteristic parameter μ and σ. Red dashed line: Pareto fit with characteristic parameter β for $\Delta r > 7420$ m.

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. 3.4 reports the histogram of the individual μ 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 Δr 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 Δr results in the selection of a Pareto distribution (eq. (3.2)). We identify the threshold $\Delta r* = 7420$ m as the minimal resolution for which the best fit in $\Delta r* < \Delta r < 10^7$ 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 $\Delta \hat{r}* = 7488.3 \pm 328.2$ 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.

3.2.2 *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. (3.1), fig. 3.5,

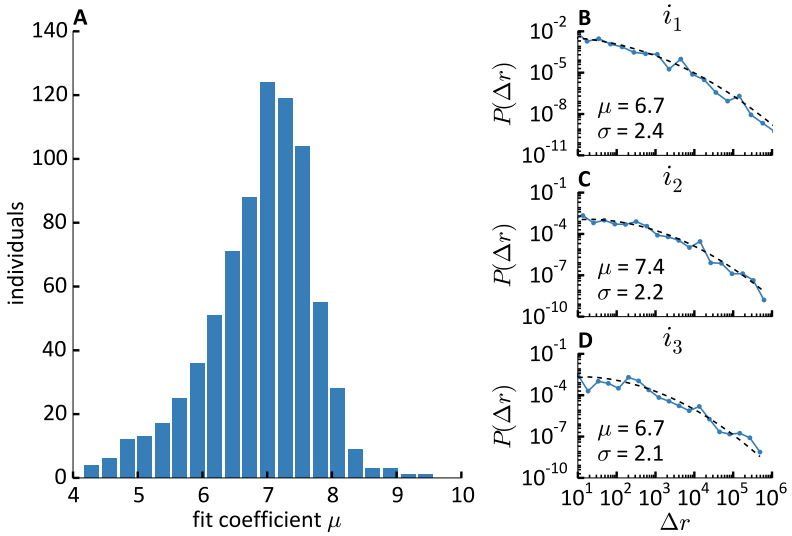

Figure 3.4: **Distribution of individual displacements**. **A**) Frequency histogram of 96.2% of individuals for which the individual distribution of displacement is log-normal, according to the value of the log-normal fit coefficient μ. **B-C-D**) Examples of the distribution of displacements $P(\Delta r)$ of three individuals i_1 (B), i_2 (C), i_3 (D) (dotted line), with the corresponding log-normal fit (dashed line). The value of the fit coefficients μ and σ are reported in each subfigure.

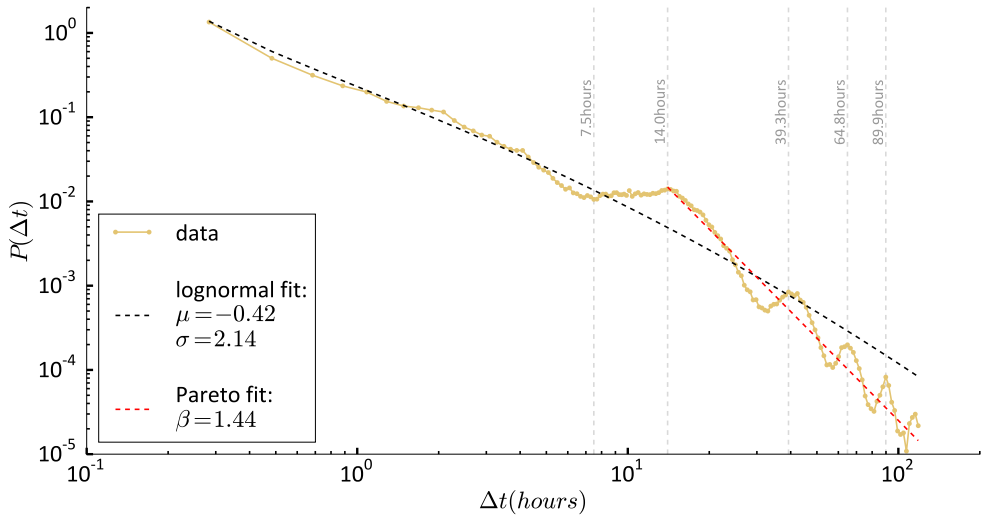

Figure 3.5: **Distribution of waiting times between displacements.** Yellow dotted line: data. Black dashed line: Log-normal fit with characteristic parameter μ and σ. Red dashed line: Pareto fit with characteristic parameter β for $\Delta t > 13$ h.

see also appendix A.1), 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 appendix A.1). As in the case of displacements, we find that restricting the analysis to large values of our observable Δt, and specifically considering only $\Delta t > \Delta t* = 13$ h, results in the selection of the Pareto distribution (eq. (3.2), see fig. 3.5), with coefficient $\beta = 1.44 \pm 0.01$. We find by averaging over 100 samples of 200 individuals that $\Delta\hat{t}* = 13.01 \pm 0.12$. 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 [43, 98, 99]. Other peaks appear for intervals $\Delta t \approx 14 + n \cdot 24$, with $n = \{2, 3...\}$, 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[102]. 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

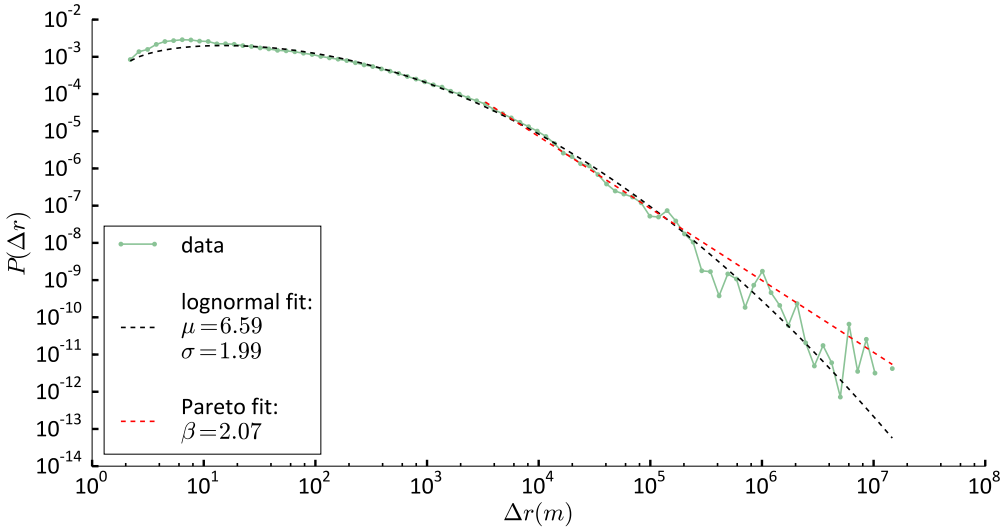

Figure 3.6: **Distribution of displacements between discoveries**. Green dotted line: data. Black dashed line: Log-normal fit with characteristic parameter μ and σ. Red dashed line: Pareto fit with characteristic parameter β for $\Delta r > 2800$ m.

variations of this choice, including considering periods shorter than 100 hours.

3.2.3 *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. 3.6, see also appendix A.1). 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 appendix A.1).

3.2.4 *Correlations between pauses and displacements*

We further investigate the properties of individual trajectories by analysing the correlations between the distance Δr and the duration Δt_{disp} characterising a displacement and the time Δt spent at destination. fig. 6.8A shows a positive correlation between Δr and Δt_{disp} for $\Delta r \gtrsim 300m$ ($p < 0.01$). As Δr 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 char-

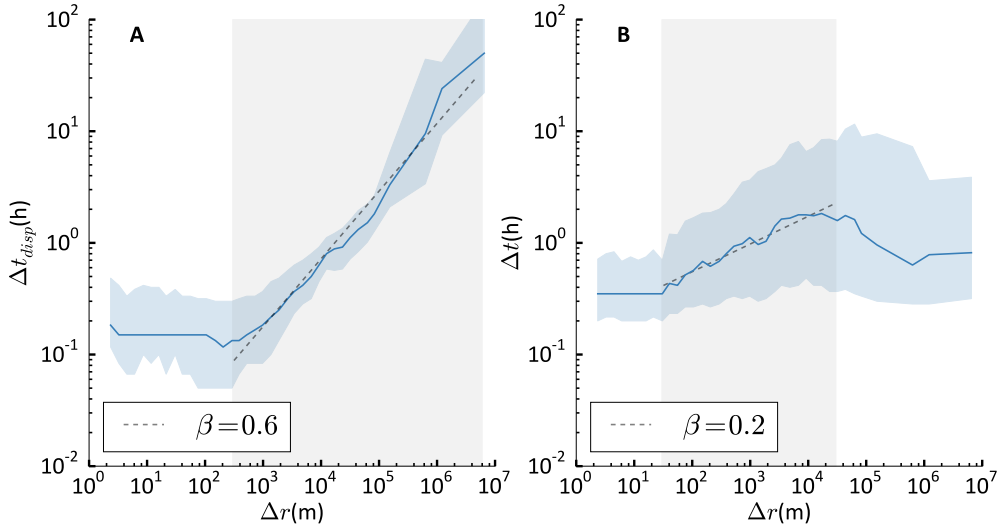

Figure 3.7: **Correlations between displacements and pauses**. **A)** The duration Δt_{disp} of a displacement vs the distance Δr between origin and destination. The blue line is the median value of Δr and Δt_{disp} computed within log-spaced 2-dimensional bins. The filled blue area corresponds to the 25-75 percentile range. The value of the Pearson correlation coefficient within the shaded grey area indicates a positive correlation, with $p - value < 0.01$. The dashed line is a power-law function with coefficient β, as a guide for the eye. **B)** The waiting time Δt at destination vs the distance Δr between origin and destination. The blue line is the median value of Δr and Δt computed within log-spaced 2-dimensional bins. The filled blue area corresponds to the 25-75 percentile range. The value of the Pearson correlation coefficient within the shaded grey area indicates a positive correlation, with $p - value < 0.01$. The dashed line is a power-law function with coefficient β, as a guide for the eye.

acterises also the distance Δr covered between origin and destination and the waiting time at destination for distances $30m \lesssim \Delta r \lesssim 10^4 m$ ($p < 0.01$). Instead, the correlation is negative for distances larger than $5 \times 10^4 m$ (fig. 6.8B). 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.

3.2.5 *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 appendix A.1 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:

$$P(x) = (1 + \xi x)^{-\frac{\xi+1}{\xi}} \tag{3.4}$$

where ξ 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 appendix A.1 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.

3.3 SUMMARY

In this chapter, we have presented a comprehensive review of over 40 studies analysing the distributions of human trajectories, showing there is large disagreement among them, due to the difference in data collection and resolution. Then, using high resolution data collected by the Copenhagen Networks Study (see section 2.2), 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 [108], search optimisation [56–59], the hierarchical organisation

of the streets network [109] and of the transportation system [51–53]. Given that individuals involved in the CNS experiment mostly live in cities, all these factors could contribute to explain why the distribution of displacements of the CNS dataset is heavy-tailed. On the other hand log-normal distributions can result from multiplicative [110] and additive [111] processes and describe the inter-event time of different human activities such as writing emails, commenting/voting on online content [112] and creating friendship relations on online social networks [113]. Instead, the distribution of inter-event time in mobile-phone call communication activity can be described as the composition of power-laws [114–116], 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 $\Delta t < \Delta t^*$ 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 $\sim$ 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 [32]) (Details are reported in chapter 2).

While identifying the mechanism responsible for the observed mobility patterns is beyond the scope of the present work, we anticipate that a more complete spatio-temporal description of human mobility will help us develop better models of human mobility behaviour [52, 117]. 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 [118].

Life becomes an astronomically large series of small events,
most of which are routine and some of which represent very critical gates.

— Torsten Hägerstraand [5]

4

LONG-TERM VISITATION PATTERNS

How do humans allocate time among different locations? Since the beginning of the 20th century, this question has intrigued sociologists, anthropologists, psychologists, economists and geographers. Only recently, the availability of passively collected trajectories has allowed to characterize and model human visitation patterns at high resolution. In this chapter, we present a brief overview of these recent experimental studies. We show that the existing literature had focused on patterns recurring within short time periods, typically less than six months. Then, we present our research (see also [II]), where we analyse visitation patterns of $\sim$ 40000 individuals to study changes in human mobility in an entirely new regime, characterized by the slowly-evolving dynamics taking place on longer time-scales.

4.1 STATE OF THE ART: THE DAILY AND WEEKLY TIME-SCALES

The theoretical foundations of research on time-space allocation lie primarily in time geography [5]. From the 1970s, time-geographers recognized the role of cultural, social and legal constraints on the space-time fixity of daily activities [5, 119, 120]. Their researches were based on experimental studies relying mostly on self-reported travel diaries. Recent studies based on digital traces including mobile phone records [45, 60], online location-based social networks [35, 86, 88, 121, 122], and GPS location data of vehicles[36, 83, 93, 99, 123, 124] have confirmed that individuals universally exhibit a markedly regular pattern characterized by few locations where they return regularly [79, 125] and predictably [46]. However, the observed regularity mainly concerns human activities taking place at the daily [98, 126] or weekly [45, 60, 86] time-scales, such as commuting between home and office [45, 48, 60, 61], pursuing habitual leisure activities, and socializing with established friends and acquaintances [35]. Thus, while the role played by slowly occurring changes on the evolution of individuals' social relationships has been widely investigated [127–134], their effects on human mobility behavior are not well understood and not included in most available models [7, 46, 109, 135–139].

4.2 LONG-TERM VISITATION PATTERNS

In this chapter, we present a study of individuals' routines across months and years. Our research is based on the analysis of $\sim 40\,000$ high resolution mobility trajectories of two samples of individuals measured for at least 12 months (see table 2.1): the users of the Lifelog mobile application, traced over 19 months, and the participants in the Copenhagen Networks Study (CNS) [70], spanning 24 months. Results were corroborated with data from two other experiments with fixed rate temporal sampling, but lower spatial resolution and sample size (table 2.1): the Lausanne Data Collection Campaign (MDC), lasted for 19 months [71, 76] and the Reality Mining dataset (RM) [72, 140], spanning 10 months.

Our datasets rely on different types of location data and collection methods (see chapter 2), but share the high spatial resolution and temporal sampling necessary to capture mobility patterns beyond highly regular ones such as home-work commuting [47].

All the datasets considered display statistical properties consistent with those reported in previous studies focusing on larger samples but shorter timescales [32, 46] (see also section 2.5), and their temporal resolution and duration make them ideal for investigating the evolution of individual geo-spatial behaviours on longer timescales. Moreover, three of the datasets considered (CNS, MDC, RM) include also information on individuals' interactions across multiple social channels (phone calls, sms, Facebook, see also chapter 2), allowing us to connect individuals' spatial and social behaviours across long timescales. Two of the datasets (CNS and RM) consist of the trajectories of university students (CNS, RM) and faculty members (RM). These subjects are homogeneous with respect to socio-demographic indicators affecting mobility behaviour [80], and their displacements are constrained by a similar academic schedule.Notwithstanding this possible source of bias, all results presented below hold for the four considered datasets.

4.2.1 *Individuals' set of visited locations grows with characteristic sublinear exponent.*

When initiating a transition from a place to another, individuals may either choose to return to a previously visited place, or explore a new location. To characterize this exploration-exploitation trade-off, we represent individual geo-spatial trajectories as sequences of locations, where 'locations' are defined as places where participants in the study stopped for more than 10 minutes (fig. 4.1A, see also chapter 2). CNS locations' typical extent after pre-processing matches that of places like commercial activities, metro stations, classrooms and other areas within the University campus (see fig. 2.6). Despite the

differences in data spatial resolution, the number of unique locations visited weekly is comparable among all 4 datasets.

A first question concerning the long term *exploration behaviour* of the individuals is whether an individual's set of known locations continuously expands, or saturates over time. We find that the total number of unique locations $L_i(t)$ an individual i has discovered up to time t grows as $L_i \propto t^{\alpha_i}$ (fig. 4.1B), and that individuals' exploration is homogeneous across the populations studied, with α_i peaked around $\overline{\alpha}$ (Lifelog: $\overline{\alpha} = 0.73$, CNS: $\overline{\alpha} = 0.61$, MDC: $\overline{\alpha} = 0.69$, RM: $\overline{\alpha} = 0.76$) (fig. 4.1C). This sub-linear growth occurs regardless of how locations are defined or when in time the measurement starts (seefig. B.8). This behavior is a characteristic signature of Heaps' law [141], and consistent with findings from previous studies focusing on shorter timescales [46].

4.2.2 *In spite of exploration the number of familiar locations is constant*

. We also find that, while continually exploring new places, individuals allocate most of their time among a small subset of all visited locations (see appendix B.1), in agreement with previous research on human mobility behavior [46, 79, 125] and time-geography [5, 142–145]. Hence, at any point in time, each individual is characterized by a set of familiar locations within which she visits as a result of her daily activities [143, 146]. Operationally, we define it as the set $S_i(t) = \{\ell_1, \ell_2, ..., \ell_k, ...\ell_C\}$ of locations ℓ_k that individual i visited at least twice and where she spent on average more than 10 minutes/week during a time-window of 20 consecutive weeks preceding time t. Typical locations visited repeatedly but for short periods of time correspond to shops/commercial activities or to transportation hubs. The results presented below are robust with respect to variations of this definition, such as changes of the time-window size or the definition of a location (see appendix B.1).

Thus, individuals continually explore new places yet they are loyal to a limited number of familiar ones. But how does discovery of new places affect an individual's set of familiar locations? We find that the average probability $\overline{P}$ that a newly discovered location will enter in the set stabilizes at $\overline{P}$ (CNS: $\overline{P} = 15\%$, Lifelog: $\overline{P} = 7\%$, MDC: $\overline{P} = 15\%$, RM: $\overline{P} = 20\%$) over the long term, indicating that individuals' sets of familiar locations are inherently unstable and new locations are continually added. However, over time individuals may also cease to visit certain familiar locations. The balance between newly added and dismissed locations is captured by the temporal evolution of the set, which we characterize by the *location capacity* and *net gain*. We define *location capacity* C_i as the number of an individual's familiar locations, at any given moment. Operationally, it is computed as the size of the activity set. The *net gain* G_i is defined as the difference

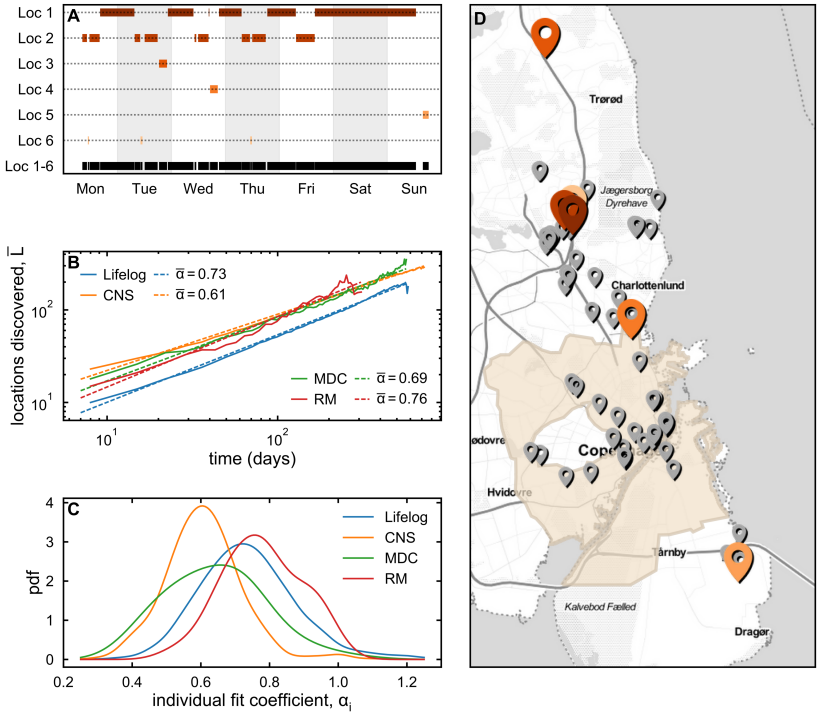

Figure 4.1: **Familiar locations and exploration.** (**A**) An example of an individual's mobility trace. The visiting temporal pattern of the six most visited locations are shown (Loc1, ..., Loc6) along with the black trace including all visits to these 6 locations (Loc1-6). (**B**) Total number of discovered locations in time. The figure shows the average across users for each dataset (coloured filled lines), and a power-law fitting function (dashed lines) with exponent α. (**C**) The probability density functions of individuals' power-law fit coefficients for different datasets (coloured filled lines) are peaked around their average value. (**D**) Example of an individual's set of familiar locations. Locations are represented as pins on a map. The six most visited locations are displayed as larger pins using the same color scheme of panel **A**. The light orange area shows the city of Copenhagen.

between the number of locations that are respectively added (A_i) and removed (D_i) from the activity set at a specific time, hence $G_i = A_i - D_i$. Note that a location is removed from the activity set at time t if it is not visited multiple times and for more than 10 minutes/week in the 20 weeks preceding t. fig. 4.2A shows the evolution of the average capacity $\overline{C}$ for the populations considered, normalized to account for the effects due to different data collection methods (see chapter 2).

We find that $\overline{C}$ is constant in time, with a linear fit of the form $\overline{C} = a + b \cdot t$ yielding b not significantly different than 0 (Lifelog: $b = 0.0013 \pm 0.0040$, CNS: $b = -0.0024 \pm 0.0025$, MDC: $b = 0.0003 \pm 0.0032$, RM: $b = 0.0044 \pm 0.0189$). Analogously, a power-law fit of the form $\overline{C(t)} \propto t^{\beta}$ yields β consistent with 0 (Lifelog: $5 \cdot 10^{-4} \pm 3 \cdot 10^{-2}$, CNS: $-2 \cdot 10^{-3} \pm 4 \cdot 10^{-2}$, MDC: $-2 \cdot 10^{-4} \pm 3 \cdot 10^{-3}$, RM: $4 \cdot 10^{-3} \pm 2 \cdot 10^{-2}$), suggesting a linear relationship between t and $\overline{C}$. As a further control, we performed a multiple hypothesis test with false discovery rate correction to compare the averages of the capacity distribution at different times (see appendix B.1). We find no evidence for rejecting the hypothesis that the average capacity does not change in time. Additionally, we find that the spatial extent of the set of familiar locations, measured by its radius of gyration [32], is on average constant in time (see appendix B) under the two tests above. Thus, despite individual set of familiar locations evolving over time, the average location capacity is a conserved quantity.

4.2.3 *Conservation of location capacity holds for individuals.*

The conservation of the average location capacity may result from either (i) each individual maintaining a stable number of familiar locations over time or (ii) a substantial heterogeneity of the populations considered, with certain individuals shrinking their set of familiar locations and other expanding theirs. We test the two hypotheses by measuring the individual average net gain across time $\langle G_i \rangle$ and its standard deviation $\sigma_{G,i}$. If a participant's average gain is closer than one standard deviation from 0, hence $|\langle G_i \rangle|/\sigma_{G,i} < 1$, then the net gain is consistent with $\langle G_i \rangle = 0$. If this is true for the majority of individuals, the location capacity is conserved at the individual level and hypothesis (i) holds. If, on the other hand, $|\langle G_i \rangle|/\sigma_{G,i} \geq 1$, the individual capacity must either increase or decrease in time, supporting hypothesis (ii). We find that hypothesis (i) holds for most individuals (Lifelog: 99.39%, CNS: 97.44%, MDC: 95.42%, RM: 87.80%) (fig. 4.2C-F, see also appendix B.1). For the large majority of each population, the average net gain of familiar locations added or removed at any instant of time is not significantly different from 0, hence their individual capacity is conserved. Also, we find that the individual capacity has low variability with the ratio between the average individual capacity and its standard deviation typically limited below 30% (Lifelog: 30%,

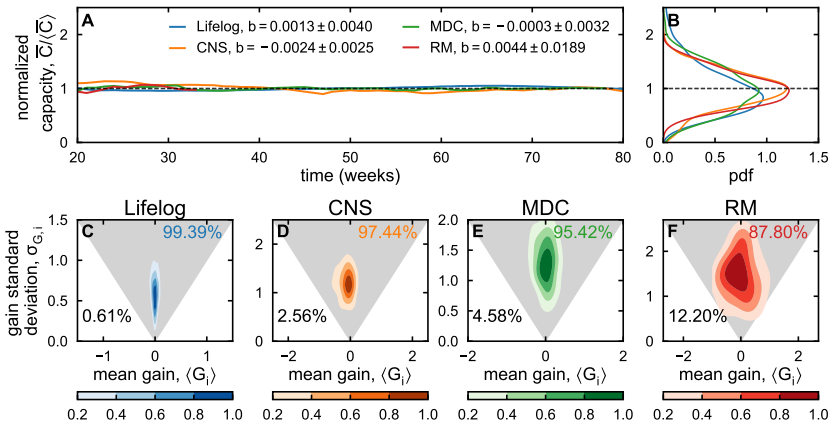

Figure 4.2: **Conserved size of evolving sets of familiar locations. (A)** Evolution of the average normalized capacity for the 4 datasets considered. The dashed black line corresponds to constant capacity. The error on the angular coefficient b of a linear fit, reported in the legend, shows that the fit is compatible with a constant line. **(B)** Probability density function of individuals' average capacity for the 4 datasets considered. **(C, D, E, F)** Gain standard deviation $\sigma_{G,i}$ vs the average gain $\langle G_i \rangle$ for the Lifelog **(E)**, CNS **(F)**, MDC **(G)** and RM **(H)** datasets. Lines representing cumulative probabilities are obtained through a kernel density estimation from the data, the grey area corresponds to individuals for which $|\langle G_i \rangle| < \sigma_{G,i}$, i.e. whose average gain is compatible with zero. It contains 99.39% (Lifelog), 97.44% (CNS), 95.42% (MDC) and 87.80% (RM) of the population.

CNS: 28%, MDC: 27%, RM: 14%), demonstrating that fluctuations of the capacity are relatively small.

Fixed size capacity is not a consequence of time constraints. These results indicate that each individual is characterized by a fixed-size but evolving set of familiar locations. We find that the typical size of the set saturates at $\sim$ 25 for increasingly larger values of the time-window defining this set (see appendix B.1). This value is consistent across all 4 samples, prior rescaling to account for the differences in time coverage. Individuals' values are homogeneously distributed around the sample mean (fig. 4.2B, see also appendix B.1), and it is worth noting that the so-called *explorers* [79] have significantly higher capacity than the so-called *returners* [79] (see appendix B).

To interpret the information contained in the measured value of the location capacity, we randomize the temporal sequences of locations in two ways, preserving routines of individuals only up to the daily level. After breaking individual time series into modules of 1 day length, (a) we randomize individual timeseries preserving the module/day units (local randomizations) or (b) we create new sequences by assembling together modules extracted randomly by the whole set of individual traces (global randomization) (see appendix B.1). Due to the absence of temporal correlations, the capacity is constant

in time also for the randomized datasets. However, the capacity of the random sets is significantly higher than in the real time series for both randomizations under the Kolmogorov-Smirnov test (see appendix B.1), implying that the observed value in real data is not a simple consequence of time constraints. In all cases, the similarity is higher for the local randomization than for the data. In one case, this is true also for the global randomization. This result has to do with the ratio between the location in the AS (the spatial capacity) and the total number of locations that ever enter in the AS. This ratio is typically larger for the local randomization (and in one case also for the global randomization) than for the data, implying that the Jaccard similarity is on average higher in the randomized case. In fact, for two subsets of k objects picked from a list of n, the expected value of the intersection is k^2/n and the union is $k(2-k/n)$, hence the Jaccard Similarity increases with the ratio $r = k/n$ as $r(1/(2-r))$, where $r <= 1$.

Instead, the fixed capacity is an inherent property of human behaviour.

Time-evolution shows that familiar locations change gradually. The time evolution of the set of familiar locations supports this finding. We measure the turnover of familiar locations using the Jaccard similarity $J_i(t, \gamma)$ between the weekly set at t and at $t + \gamma$ (see fig. 4.3). Despite seasonality effects which imply fluctuations around a typical behaviour, J_i does not depend on the initial point but only on the waiting time γ, and we can consider $J_i(\gamma)$ independently of t (see appendix B.1). We find that the average similarity decreases as a power law $\bar{J} \propto \gamma^{\lambda}$ with coefficient significantly different than 0 (Lifelog: $\lambda = -0.16$, CNS: $\lambda = -0.31$, MDC: $\lambda = -0.5$, RM: $\lambda = -3.00$) . On the other hand, for the randomized sequences, the Jaccard similarity is constant in time as familiar locations are never abandoned ($\bar{J} \propto \gamma^0$). This confirms that individual sets of familiar locations change continually and individual routines evolve gradually in time.

In order to characterize the structure of the set of familiar locations, we investigate how individuals allocate time among different classes of locations defined on the basis of their average visit duration. We consider intervals ΔT, with ΔT ranging from 10 to 30 minutes per week (the time it takes to visit a bus stop or grocery shop) up to 48 to 168 hours per week (such as for home locations). For each of these locations classes, we compute the evolution of the *capacity* $c_i^{\Delta T}$ and the *gain* $G_i^{\Delta T}$, and test the hypothesis $G_i^{\Delta T} = 0$, as above. We find that, although these subsets are continuously evolving, $c_i^{\Delta T}$ is conserved for each ΔT (fig. 4.4, see also appendix B.1), indicating that the number of places where individuals spend a range of time ΔT does not change over time. This result holds independently of the choice of specific ΔT and implies that the individual capacity $C_i =$

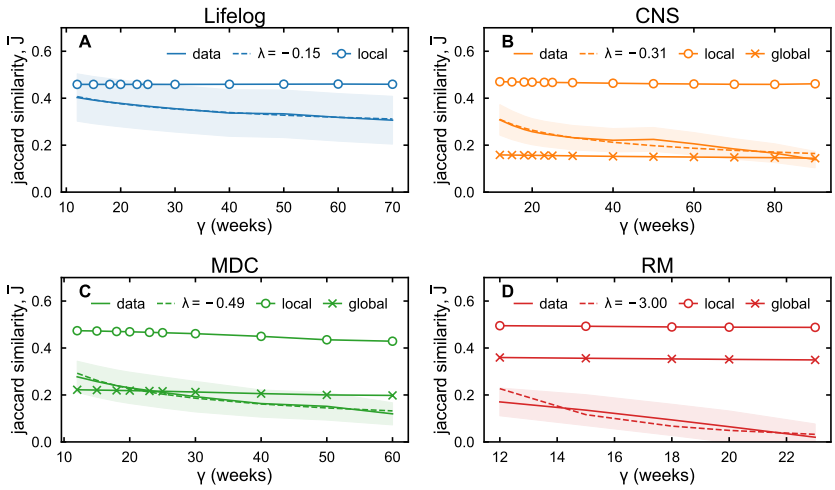

Figure 4.3: **Evolution of the set of familiar locations** The average Jaccard similarity $\overline{J}$ between the set measured at t and $t + \gamma$ as a function of γ for data (filled lines), the globally randomized series (lines with crosses) and the locally randomized series (lines with dots). Filled areas correspond to the 50% interquartile range. Dashed lines correspond to power-law fits $\overline{J} \sim \gamma^{\lambda}$. Results are shown for the Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)** datasets, with $w = 10$ weeks. The anonymization procedure applied by SONY Mobile before supplying the data makes impossible to perform the global randomization on the Lifelog trajectories.

$\sum c_i^{\Delta T}$, where both C_i and each $c_i^{\Delta T}$ are conserved across time. Thus, both location capacity and time allocation are conserved quantities.

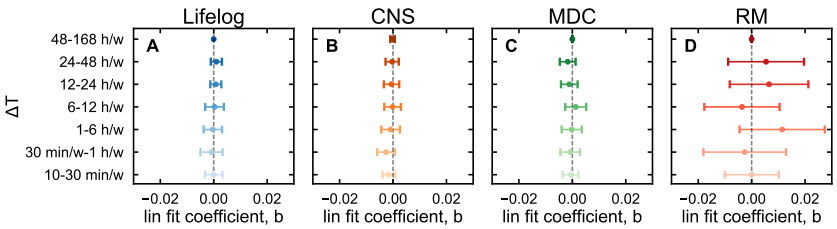

Figure 4.4: **Conservation of time allocation.** Linear fit coefficients of the average capacity vs time for several categories of locations ΔT are consistent with 0 within errors. Results are shown for the Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)** datasets.

4.2.4 *Including long-term evolution of familiar locations improves modelling*

Our results have consequences for the modeling of human mobility. The renown exploration and preferential return model [7, 46] describe agents that, when not exploring a new location, return to a previously visited place selected with a probability proportional to the number of former visits. According to the EPR model, at a given transition n,

an individual explores a new location with probability $P_{new} = \rho S - \gamma$, or returns to a previously visited location with probability $1 - P_{new}$, with S the number of previously visited locations, and ρ and γ parameters of the model. If the individual returns to a previously visited location, she chooses location i with probability $\Pi_i = m_i / \sum_i m_i(n)$ where m_i is the total number of visits to location i occurring before transition n. In the EPR model, time scales with the number of transitions as $\sim n^{1/\beta}$, with β a parameter of the model. This models reproduce some of the empirical observations described above, including the conservation of the *location capacity* (Fig. 4.5). In fact, in the EPR model, the number of visits to location i after n steps is $m_i(n) = n/n_i$, where n_i is the step at which location i was first seen. The number of visits after $n + w$ steps (where w is the size of the rolling window) is $m_i(n + w) = (n + w)/n_i$. Hence, during w, location i is visited w/n_i times. This value is independent of n. The condition $w/n_i > 2$, holds for the limited set of locations such that $n_i < w/2$. A similar argument can be used to justify that the number of places where users spend at least a given amount of time during the window w is fixed. The model, however, fail to describe the time evolution of the activity set (Fig. 4.5).

To overcome this limitation, we start from the observation that the exploitation probability for a location is time-dependent [147, 148] and endow the agents with a finite memory M so that the probability of returning to a location is based on the number of visits occurred in the last M days. Hence, the return probability to a given location i is $\Pi_i = m_i / \sum_i m_i(n)$, where m_i is the total number of visits to location i occurring at most M time units before transition n. The model including this simple modification qualitatively reproduces all the observations, including the long-term evolution of the activity set (see Fig.4.5). Note that with this choice of parameters, the capacity averages at $\sim$ 42 locations. In my ongoing work, I am studying the relation between the model parameters and the value of the capacity.

Exploration rates in spatial and social domain are connected. Finally, we analyse the connection between the social and spatial domain. Empirical observations suggest that there are upper limits to the size of an individual's social circle, the so-called Dunbar number [131, 132, 149, 150], due to cognitive constraints [149], and it has been hypothesized that the geography of one's familiar locations is proportional to one's social network geography [151]. Motivated by these observations, we test the hypothesis of a correlation between individuals' *location capacity* and the size of their social circle, as measured by the people contacted by either phone call or sms over a period of 20 weeks. We find that a significant positive correlation exists (see fig. 4.6). Furthermore, for the CNS dataset, we are able to show that both quantities correlate with the individual personality trait of extraversion [152], which tend to be manifested in outgoing,

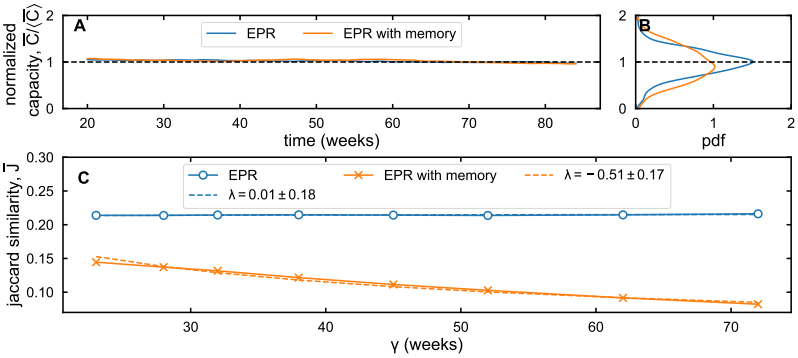

Figure 4.5: **Including finite memory improves modeling. (A)** Average normalized capacity for the EPR model [46] (blue line), and the EPR model with finite memory (orange line). **(B)** Probability density of the average normalized capacity across the population for the three models. **(C)** The average Jaccard similarity $\overline{J}$ between the set measured at t and $t + \gamma$ as a function of γ for the three models. Dashed lines correspond to power-law fits $\overline{J} \sim \gamma^{\lambda}$. Simulations are ran for 10^3 individuals. Parameters are taken from [46] and $\rho = 0.6$, $\gamma = 0.2$ and $\beta = 0.8$ (for the two models), $M = 200$ days (for the EPR model with memory). We consider that 1 time unit in the simulation (the shortest duration extracted from the distribution of waiting times) corresponds to 1 minute (the time unit considered to analyse our data). All measures are computed after waiting for a period corresponding to 7 months.

talkative and energetic behavior [153] (Pearson correlation $\rho = 0.21$, 2-tailed $p < 10^{-7}$ for location capacity vs extraversion; $\rho = 0.42$, 2-tailed $p < 10^{-27}$ for size of social network vs extraversion[154], see appendix B). We consider that these observations call for further analyses on the connections between human social and spatial behaviour.

4.3 SUMMARY

In summary, we have shown that the number of locations an individual visits regularly is conserved over time, even while individual routines are unstable in the long term because of the continual exploration of new locations. This individual *location capacity* is peaked around a typical value of ~ 25 locations across the population, and significantly (typically, at least 30%) smaller than what would be expected if only time-constraints were at play (see appendix B.1).

The *location capacity* is hierarchically structured, indicating that individual time allocation for categories of places is also conserved. These results have allowed us to improve existing models of human mobility which are unable to fully account for long-term instabilities and fixed-capacity effects.

Taken together, these findings shed new light on the underlying dynamics shaping human mobility, with potential impact for a better un-

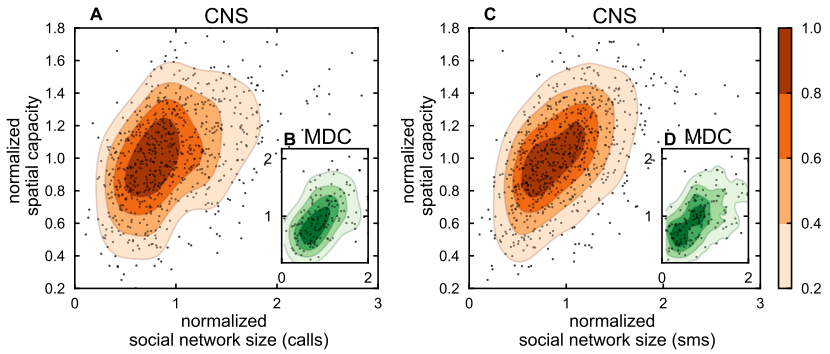

Figure 4.6: **Correlation between location capacity and social network size.** Values of individuals' average normalized location capacity vs their normalized social network size computed from phone calls interactions **(A, C)** and sms interactions **(B, D)** (black dots). Coloured filled areas correspond to cumulative probabilities estimated via Gaussian Kernel Density estimations. Results are shown for the CNS **(A, B)** and MDC **(C, D)** datasets. The values of the Pearson correlation coefficient are 0.31 **(A)**, 0.48 **(B)**, 0.52 **(C)**, 0.54 **(D)** (2-tailed $p < 10^{-13}$ in all cases). Social network size is normalized to the population average value.

derstanding of phenomena such as urban development and epidemic spreading. Extending our scope beyond mobility, we have shown that individuals' *location capacity* is correlated with the size of their social circles. In this respect, it is interesting to note that fixed-size effects in the social domain [131, 132, 149, 150] have been put in direct relation with human cognitive abilities [149]. We anticipate that our results will stimulate new research exploring this connection.

The path inside the daily prism is to a pronounced degree ruled by 'coupling constraints'. These define where, when, and for how long, the individual has to join other individuals.

— Torsten Hägerstraand [5]

5 THE CONNECTION BETWEEN SOCIAL AND SPATIAL BEHAVIOUR

In the 1970s, time geographers theorized that individual spatial choices are affected by the necessity to interact and coordinate with others. In recent years, the availability of multi-channel data has allowed researchers to explore the connections between individuals' social and spatial behaviours. However, these recent analyses have focused on understanding this relation at the level of pairs of individuals (see section 5.1). In this chapter (see section 5.2), we study the the connection between social, spatial behaviour and personality at the level of the single individual. Our analysis is based on the analysis of trajectories and interactions of $\sim$ 1000 individuals. The chapter is based on the work described in [III].

5.1 STATE OF THE ART

Individual-level variability in social and spatial behaviour has mostly been investigated in isolation so far, with few notable efforts to reconcile the two. Here, we briefly review the empirical findings in the two domains.

5.1.1 *The social domain*

Individuals deal with limited time and cognitive capacity resulting in finite social networks [149, 150] by distributing time unevenly across their social circle [132, 155–159]. While this is a shared strategy, there is clear evidence for individual-level variation. First, social circles vary in terms of diversity: they differ in size [160] - within a maximum upper-bound of $\sim$ 150 individuals [149] - and in structure [132, 161]. Second, individuals display different attitudes towards exploration of social opportunities as they are more or less keen on creating new connections [162–165]. Finally, individuals manage social interactions over time in different ways. Some are characterised by high level of stability as they maintain a very stable social circle, while others renew their social ties at high pace [131].

These heterogeneities can be partially explained by factors including gender [166, 167], age [168–170], socio-economic status [171, 172] and physical attractiveness [173]. Moreover, as conjectured by personality psychologists [174, 175], differences in personalities partially explain the variability in social circle composition [153, 154, 160, 176–180], and the different attitudes towards forming [162, 181], developing [182, 183] and replacing [161] social connections. It is worth noticing that many of these findings are recent, resulting from the analysis of digital communication traces.

5.1.2 *The spatial domain*

Constraints including physical capabilities, the distribution of resources, and the need to coordinate with others limit our possibilities to move in space [5]. Individuals cope with these limitations by allocating their time within an activity space of repeatedly visited locations [144], whose size is conserved over several years according to a recent study based on high-resolution trajectories [II], and previous ones based on unevenly sampled and low spatial resolution data [184, 185]. The activity space varies across individuals in terms of size [II] and shape [79]: it was shown that two distinct classes of individuals can be identified based on the spatial distribution of their locations, similarly to the social domain [131]. Heterogeneities in spatial behaviour can be explained in terms of gender [186], age [187, 188], socio-economic [169, 189] and ethnic [190] differences. There has only been sporadic efforts to include personality measures in geographic research, despite the strong connections between the two [191]. Recent works [180, 192] suggest that spatial behaviour can be partially explained from personality traits. However, in [192], this understanding is based on biased data collected from location-based social networks. In [180], the connection between spatial behaviour and personality is not investigated extensively, as it is not the main focus of the study.

5.1.3 *Social and spatial connection*

Recently, connections between the social and spatial behaviour of pairs [37, 62–66] and groups [193] of individuals have been demonstrated, and used to design predictive models of mobility [35, 62, 194] or social ties [63, 195–197]. Shifting the attention to the individual level, recent works based on online social network data [86, 198], mobile phone calls data [66] and evenly sampled high resolution mobility trajectories [II] have shown correlations between the activity space size and the ego network structure, calling for further research to more closely examine the connections between social and spatial behaviour at the individual level.

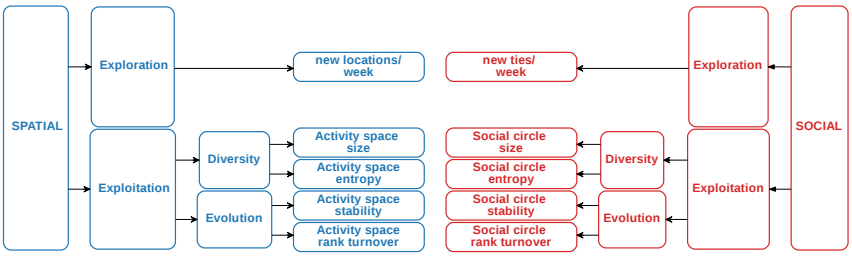

Figure 5.1: **Schematic description of our framework.**

5.2 SOCIAL, SPATIAL BEHAVIOUR AND PERSONALITY: AN EMPIRICAL STUDY

In this section, we present an empirical study on the connection between social, spatial behaviour and personality. The analysis is based on two of the datasets described in chapter 2: the CNS dataset, collecting high resolution trajectories and call records of 850 individuals, and the MDC dataset, considering 185 individuals.
We analyse the relationship between social and spatial strategies by adopting the exploration-exploitation perspective [199, 200]. In fact, both in their spatial and social behaviour, individuals are constantly balancing a trade-off between the exploitation of familiar options (such as returning to a favourite restaurant or spending time with an old friend) and the exploration of new opportunities (such as visiting a new bar or going on a first date) [199]. We quantify the propensity for exploration and exploitation within each individual, i, using the metrics reported in table 5.1, fig. 5.1 and described in section 5.2.1. We focus on two aspects of exploitation, (i) *diversity*, characterising how diverse individuals' routine are, and (ii) *evolution*, characterising the tendency to change exploited locations and friends over time. Finally, we explain part of the variability in the data by considering individuals' personality traits, that had been previously collected as part of the CNS dataset. CNS participants' background information were obtained through a 310 questions survey including the Big Five Inventory [75], which measures how individuals score on five broad domains of human personality traits: openness, conscientiousness, extraversion, agreeableness, neuroticism. The personality questionnaire used in the study is a version of the Big Five Inventory [75], translated from English into Danish. It contains 44 individual items and each trait is computed as the average of 7-10 items.

Methods are described in section 5.2.1 and the analysis is organized as follows. First, we verify that individuals' strategies are persistent in time (section 5.2.2). Then, we test the hypothesis that the strategies individuals adopt in order to choose where to go and with whom to interact are similar (section 5.2.3). Then, we identify and characterise the prevailing socio-spatial profiles appearing in the datasets(section 5.2.5).

	Exploration	Exploitation: Diversity	Exploitation: Evolution
Spatial	New loc./week, n_{loc}	Activity space size, C	Activity space stability, J_{AS}
		Activity space entropy, H_{AS}	Activity space rank turnover, R_{AS}
Social	New ties/week, n_{tie}	Social circle size, k	Social circle stability, J_{SC}
		Social circle entropy, H_{SC}	Social circle rank turnover, R_{SC}

Table 5.1: **Metrics characterising social and spatial behaviour.** The metrics are defined in section 5.2.1.

Finally, we show that socio-spatial profiles can be partially explained by the widely adopted big-five personality trait model, often used to describe aspects of the social and emotional life [160, 176, 178, 183, 201–204] (section 5.2.5).

5.2.1 *Metrics*

In this section, we define the concepts and metrics used to quantify the social and spatial behaviour of an individual i.
Exploration behaviour is characterised by the following quantities:

NUMBER OF NEW LOCATIONS/WEEK: $n_{loc}(i,t)$ is the number of locations discovered by i in the week preceding t. We discard data collected in the first 20 weeks.

NUMBER OF NEW TIES/WEEK: $n_{tie}(i,t)$ is the number of individuals who had contact with i (by sms or call) for the first time in the week preceding t.

Exploitation behaviour can be quantified by considering:

ACTIVITY SPACE: The set $AS(i,t) = \{\ell_1, \ell_2, ..., \ell_j, ...\ell_C\}$ of locations ℓ_j that individual i visited at least twice and where she spent a time τ_j larger than $200min$ during a time-window of $T = 20$ weeks preceding time t (see appendix C for the analysis with $T = 30$ weeks). Among the locations in the activity space, i visited ℓ_j with probability $p(\ell_j) = \tau_j / \sum \tau_j$. (It is worth noting that this time-based definition of activity space includes all significant locations independently of their spatial position and it is only loosely connected with space-oriented definitions widespread in the geography literature such as the “standard deviational ellipse" and the “road network buffer" [146]).

	Activity space	**Social circle**
1)	$C(i,t) = \lvert AS_i(t)\rvert$	$k(i,t) = \lvert SC(i,t)\rvert$
2)	$H_{AS}(i,t) = -\sum_{j=1}^{C(i,t)} p(j)\log p(j)$	$H_{SC}(i,t) = -\sum_{j=1}^{k(i,t)} p(j)\log p(j)$
3)	$J_{SC}(i,t) = \frac{\lvert SC(i,t)\cap SC(i,t-T)\rvert}{\lvert SC(i,t)\cup SC(i,t-T)\rvert}^{*}$	$J_{AS}(i,t) = \frac{\lvert AS(i,t)\cap AS(i,t-T)\rvert}{\lvert AS(i,t)\cup AS(i,t-T)\rvert}^{*}$
4)	$R_{AS}(i,t) = \sum_{j=1}^{N} \frac{\lvert r(j,t)-r(j,t-T)\rvert^{**}}{N}$	$R_{SC}(i,t) = \sum_{j=1}^{N} \frac{\lvert r(j,t)-r(j,t-T)\rvert^{**}}{N}$

* Here $T = 20$ weeks, see appendix C for the analysis with $T = 30$ weeks

** $r(\ell_k,t)$ and $r(u_k,t)$ denote the rank of a location ℓ_k and individual u_k at t, respectively. Locations that en

Table 5.2: **Definition of the metrics characterising the activity space and the social circle.** 1) The *size* of a set is the number of elements in the set 2) We compute the *entropy* of a set considering the probability $p(j)$ associated to each element j of the set. 3) We measure the *stability* J_{AS} by computing the Jaccard similarity between the activity space at t and at $t-T$, with $T = 20$ weeks. J_{SC} is computed in the same way for the social circle. 4) We compute the *rank turnover* of a set by measuring for each of its elements j the absolute change in rank between two consecutive time windows of length $T = 20$ weeks. The rank is attributed based on the probability $p(j)$. The average absolute change in rank across all elements corresponds to the rank turnover.

.

SOCIAL CIRCLE: The set $SC(i,t) = \{u_1, u_2, ..., u_j, ...u_k\}$ of individuals u_j with whom individual i had a number of contacts $n_j > 5$ by sms or call during a time-window of $T = 20$ consecutive weeks preceding time t (see appendix C for the analysis with $T = 30$ weeks). The probability that i has contact with a given member u_j of her social circle is $p(u_j) = n_j / \sum n_j$.

For these two sets $AS(i,t)$ and $SC(i,t)$, we consider their sizes $C(i,t)$ and $k(i,t)$, quantifying the number of favoured locations and social ties, respectively; their entropies $H_{AS}(i,t)$ and $H_{SC}(i,t)$, measuring how time is allocated among locations and ties; their stabilities $J_{AS}(i,t)$ and $J_{SC}(i,t)$, quantifying the fraction of conserved locations and ties, respectively, across consecutive non-overlapping windows of $T = 20$ weeks (seeappendix C for $T = 30$); their rank turnovers $R_{AS}(i,t)$ and $R_{SC}(i,t)$ measuring the average absolute change in rank of an element in the set between consecutive windows. The mathematical definition of these quantities is provided in table 5.2.

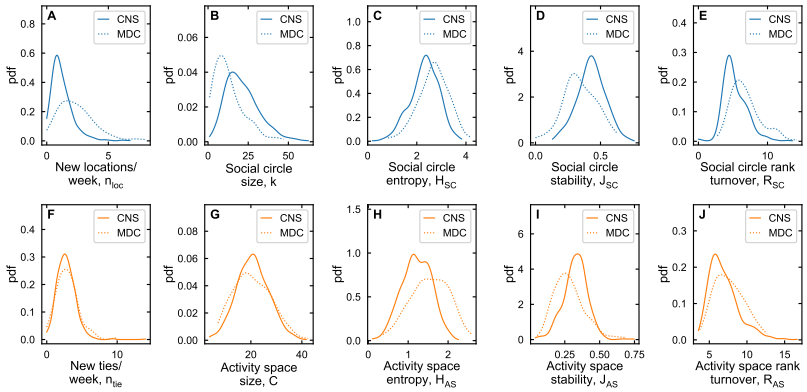

Figure 5.2: **Distribution of social (above line) and spatial (bottom line) metrics for the CNS and MDC datasets.**

5.2.2 *Exploration and exploitation are persistent in time.*

First, we verify that individual behaviour is persistent in time. For all the aforementioned measures, we compare the individual self-variation across time $d_{self}(i)$ with a reference difference $d_{ref}(i,j)$ between individuals i and j. In the case of the activity space size, for example, self-variation is measured as $d_{self} = \langle |C(i,t) - C(i,t-T)| \rangle$, where $\langle \cdot \rangle$ is the average across time and $T = 20$ weeks (see appendix C for $T = 30$) ; the reference difference is computed as $d_{ref}(i,j) = |\langle C(i,t) \rangle - \langle C(j,t) \rangle|$. If $d_{self}(i) < d_{ref}(i,j)$ for most j, we can conclude that for individual i, fluctuations of the activity space size are negligible compared to the difference with other individuals. The same procedure is followed for all metrics with an adjustment in the case of entropies: The persistence of the entropy H_{AS} is verified by comparing the Janson-Shannon divergences $d_{self} = JSD(AS(i,t), AS(i,t-T))$ and $d_{ref} = JSD(AS(i,t), AS(j,t))$, where $JSD(P_1,P_2) = H(\frac{1}{2}P_1 + \frac{1}{2}P_2) - \frac{1}{2}(H(P_1) + H(P_2))$. The same method was used for H_{SC} (see Methods and [132]).

Results from the CNS dataset reported in table 5.3 show that for all metrics $d_{self}(i) < d_{ref}(i,j)$ holds in more than 99% of cases on average (MDC: 97%, see appendix C). Moreover, the average self-variation across the population $\overline{d_{self}}$ is consistent with $\overline{d_{self}} = 0$ within errors, and $\overline{d_{self}}$ significantly smaller than the average reference difference $\overline{d_{ref}}$ (see table 5.3 andappendix C).

These results extend previous findings [II, 132] and suggest that each individual is characterised by a distinctive socio-spatial behaviour captured by the ensemble of these metrics averaged across time. In fact, these averages are heterogeneously distributed across the samples considered (see fig. 5.2).

	$\overline{d_{self}}$	$\overline{d_{ref}}$	$\overline{d_{self}(i) < d_{ref}(i,j)}$
Social circle size, k	0.04 ± 0.09	12 ± 5	99%
Activity space size, C	0.04 ± 0.07	7 ± 3	99%
New locations/week, n_{loc}	0.05 ± 0.10	0.9 ± 0.5	96%
New ties/week, n_{tie}	0.10 ± 0.17	1 ± 1	95%
Social circle entropy, H_{SC}	0.002 ± 0.007	0.7 ± 0.2	99%
Activity space entropy, H_{AS}	0.002 ± 0.005	0.4 ± 0.1	99%
Social circle stability, J_{SC}	$(9 \pm 22) \cdot 10^{-4}$	0.13 ± 0.05	100%
Activity space stability, J_{AS}	$(9 \pm 26) \cdot 10^{-4}$	0.10 ± 0.04	99%
Social circle rank turnover, R_{SC}	0.05 ± 0.39	2 ± 1	99%
Activity space rank turnover, R_{AS}	0.04 ± 0.10	2 ± 1	99%

Table 5.3: **CNS dataset: Persistence of social and spatial behaviour.** For each of the social and spatial metrics, $\overline{d_{self}}$ is the average self-distance and $\overline{d_{ref}}$ is the reference distance between an individual and all others, averaged across individuals. The third column reports the fraction of cases where $\overline{d_{self}(i) < d_{ref}(i,j)}$, averaged across the population.

5.2.3 *Exploration and exploitation are correlated in the social and spatial domain.*

A natural way to test the interdependency between social and spatial behaviours is measuring the correlation between a given social metric and a corresponding spatial one. We find positive and significant correlations for all metrics and datasets (see fig. 5.3 and appendix C).

We find that individuals with high propensity to explore new locations are also more keen on exploring social opportunities (see fig. 5.3A). Those with diverse mobility routine are also likely to have a correspondingly large social circle (see fig. 5.3B), and those that often replace social ties, have also an unstable set of favourite locations (see fig. 5.3C and D).

We verify that the observed correlations are not spurious by performing multiple regression analyses that control for other possible sources of variation: gender, age, and time coverage (the average time an individual position is known). We implement five multiple linear regression models M1, M2, M3, M4 and M5. Each regression model predicts a given spatial metric (the activity space size C,the activity space entropy H_{AS}, the number of new locations/week n_{loc}, the activity space stability J_{AS} and the rank turnover R_{AS}) using the corresponding social metric and the control variables (age, gender and time coverage) as regressors. The relative importance of each regressor is assessed using the LMG [205] method, which estimate the proportion of the R^2 contributed by each individual regressor. Given a sequence of regressors, the contribution of each of them is computed as the increase in R^2 obtained by adding the regressor to the model. The order of elements in the sequence may influence the results if the regressors are correlated. Hence, the LMG method estimates the contribution of each regressor as the average contribution over possible orderings of the sequence of regressors.

Results obtained via weighted least square regression (see section 5.2.3 and appendix C) reveal that the social metrics are significant predictors for spatial metrics (p value$>$ 0.01 in all cases except for M4 in the MDC dataset), and they typically have more importance than factors such as gender, time, coverage and age group (see fig. 5.4).

Among the control variables, gender is a significant predictor of spatial behaviour in the CNS dataset: Females display higher level of routine diversity and propensity towards exploration, in accordance with [206]. Time coverage, measuring the fraction of time an individual position is known, plays a significant role in explaining spatial entropy and activity space stability, since individuals who spend long time in the same place (or leave their phone in the same place) are more easily geo-localised. Age differences are not present within the sample of students participating in the CNS study, and they are not

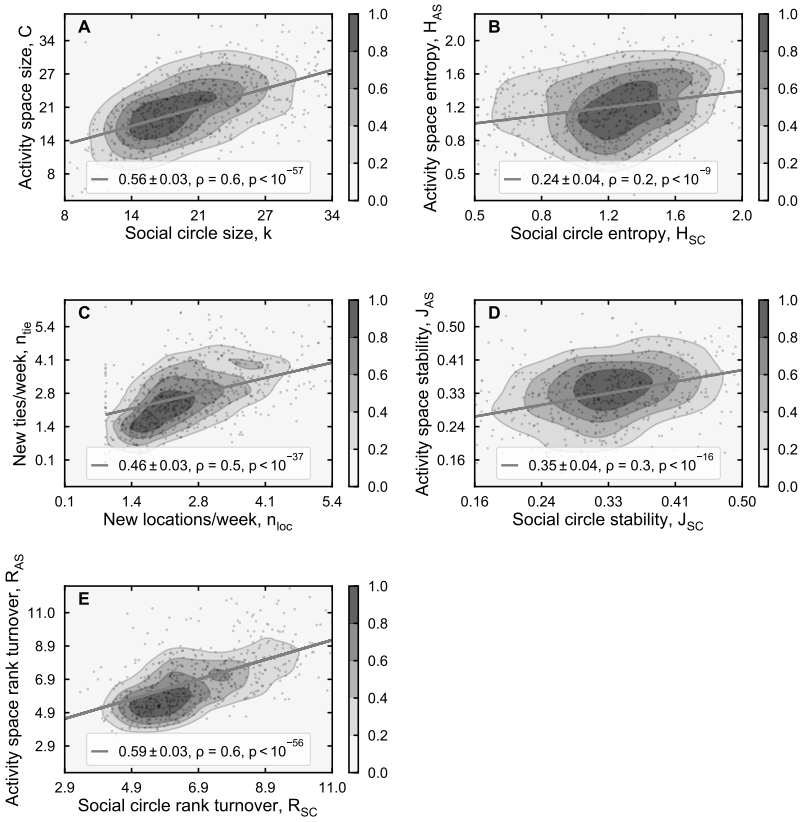

Figure 5.3: **CNS dataset: correlation between the four dimensions of social and spatial behaviour.** (A) Activity space vs social circle size. (B) Activity space vs social circle composition measured as their entropy. (C) Average number of new locations vs new ties per week. (D) Stability of the activity space vs the stability of the social circle measured as the Jaccard similarity between their composition in consecutive time-windows. (E) Rank turnover of the activity space vs the rank turnover of the social circle. Coloured filled areas correspond to cumulative probabilities estimated via Gaussian Kernel Density estimations. Grey lines correspond to linear fit with angular coefficient b reported in the legend. The Pearson correlation coefficient, with corresponding p-value, is reported in the legend.

Model M1: Activity space size, $C = coeff_1 \cdot k + coeff_2 \cdot (gender) + coeff_3 \cdot (timecoverage) + coeff_4$	coeff	p val	LMG
Social circle size, k	4 ± 0	$< 10^{-50}$	0.94
gender	-0.4 ± 0.2	0.05	0.05
time coverage	0.4 ± 0.2	0.06	0.01
[$R^2 = 0.32$, $F = 100.44$, $p_F = 0.0$]			
Model M2: Activity space entropy, $H_{AS} = coeff_1 \cdot H_{SC} + coeff_2 \cdot (gender) + coeff_3 \cdot (timecoverage) + coeff_4$			
Social circle entropy, H_{SC}	0.07 ± 0.01	$< 10^{-6}$	0.42
gender	-0.06 ± 0.01	$< 10^{-4}$	0.22
time coverage	-0.07 ± 0.01	$< 10^{-5}$	0.36
[$R^2 = 0.11$, $F = 27.30$, $p_F = 0.0$]			
Model M3: New locations/week, $n_{tie} = coeff_1 \cdot n_{loc} + coeff_2 \cdot (gender) + coeff_3 \cdot (timecoverage) + coeff_4$			
New ties/week, n_{loc}	0.60 ± 0.05	$< 10^{-32}$	0.9
gender	-0.16 ± 0.05	$< 10^{-3}$	0.08
time coverage	0.001 ± 0.047	1.0	0.01
[$R^2 = 0.22$, $F = 61.99$, $p_F = 0.0$]			
Model M4: Activity space stability, $J_{AS} = coeff_1 \cdot J_{SC} + coeff_2 \cdot (gender) + coeff_3 \cdot (timecoverage) + coeff_4$			
Social circle stability, J_{SC}	0.024 ± 0.004	$< 10^{-10}$	0.6
gender	0.007 ± 0.003	0.05	0.04
time coverage	0.017 ± 0.004	$< 10^{-5}$	0.36
[$R^2 = 0.16$, $F = 33.36$, $p_F = 0.0$]			
Model M5: Activity space rank turnover, $R_{AS} = coeff_1 \cdot R_{SC} + coeff_2 \cdot (gender) + coeff_3 \cdot (timecoverage) + coeff_4$			
Social circle rank turnover, R_{SC}	1 ± 0	$< 10^{-56}$	0.98
gender	0.12 ± 0.07	0.06	0.01
time coverage	-0.12 ± 0.07	0.07	0.01
[$R^2 = 0.36$, $F = 108.31$, $p_F = 0.0$]			

Table 5.4: **Linear regression models for the CNS dataset.**

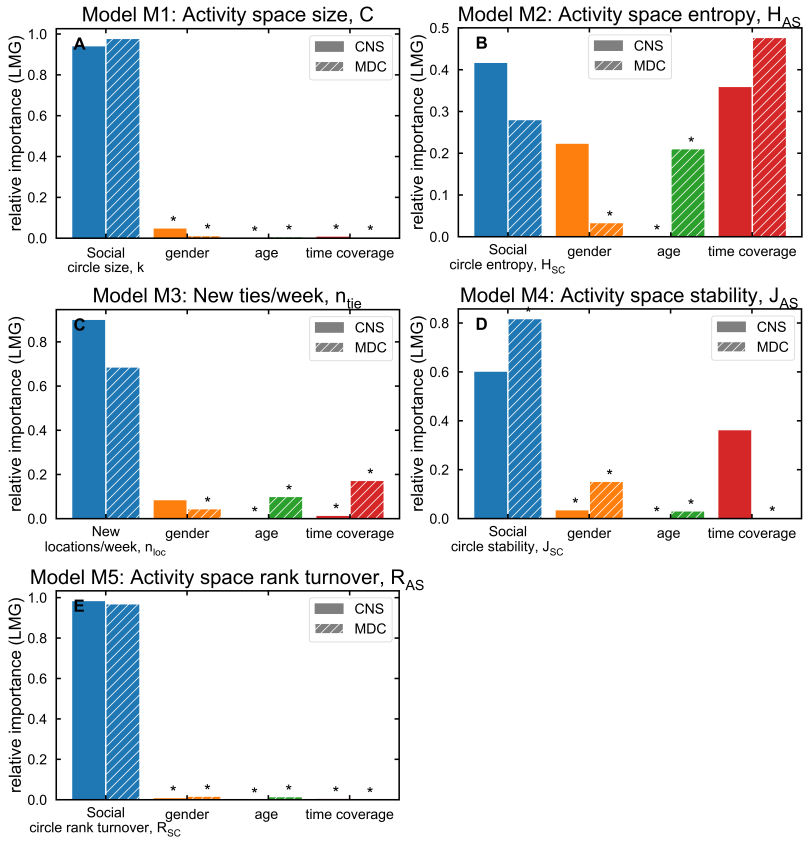

Figure 5.4: **Relative importance of regressors** LMG of each regressor computed using the Lindeman, Merenda and Gold method [205] for models M1 (A), M2 (B), M3 (C), M4 (D) and M5 (D). Plain bars show results for the CNS dataset, dashed bars for the MDC dataset. Variables that are not significant in the regression model are marked with *.

	PC 0	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
CNS	0.39	0.17	0.12	0.08	0.07	0.06	0.04	0.03	0.03	0.01
MDC	0.43	0.14	0.13	0.08	0.07	0.06	0.04	0.03	0.02	0.01

Table 5.5: **Variance explained by principal components.** The fraction of variance explained by each principal component for the CNS and MDC dataset.

estimated to be relevant with respect to spatial behaviour in the MDC study.

5.2.4 *We do not identify distinct classes of individuals.*

A natural question is whether or not, in the samples considered, there is evidence for distinct classes of individuals based on their socio-spatial behaviour [79, 131]. We approach this problem by reducing the set of metrics to a smaller number of uncorrelated variables by applying Principal Component Analysis [207, 208]. To find principal components, each variable was priorly rescaled, by subtracting the mean and dividing by the standard deviation.

In both datasets, we find that a single predominant component explains $\sim$ 40% of the differences between individuals (see table 5.5). This dimension is dominated by the metrics quantifying exploration and routine diversity (see table 5.6). This suggests that individuals with higher exploration propensity tend to have larger social circle and more diverse spatial routine, while those who explore less also have less diverse routines.

The second principal component, which accounts for $\sim$ 15% of the total variation, is dominated by the effects of evolving routines over long time scales (see table 5.6). We consider the first two principal components, PC 0 and PC 1, to reduce the effects of noise and we test the hypothesis that there exists different classes of individuals applying the gap statistic method [209]. We apply it by looking at the gap $G(K)$ between the within-cluster dispersion expected under a reference uniform distribution and the dispersion obtained after applying K-means to the data. The number of clusters $\hat{K}$ is chosen to be the smallest such that $G(K) \geq G(K+1) - s_{K+1}$, where s_{k+1} is the standard error obtained by sampling 1000 times from the uniform distribution. We find that $\hat{K} = 1$.

	CNS		MDC	
	PC 0	**PC 1**	**PC 0**	**PC 1**
Social circle size, k	0.41	0.16	0.37	-0.15
Activity space size, C	0.42	-0.24	0.42	-0.08
New locations/week, n_{loc}	0.33	0.28	0.27	0.33
New ties/week, n_{tie}	0.38	-0.05	0.37	0.19
Social circle entropy, H_{SC}	0.31	0.30	0.34	0.09
Activity space entropy, H_{AS}	0.38	-0.16	0.30	-0.07
Social circle stability, J_{SC}	-0.16	-0.46	0.07	-0.72
Activity space stability, J_{AS}	-0.10	-0.49	-0.12	-0.51
Social circle rank turnover, R_{SC}	-0.20	0.28	-0.33	0.10
Activity space rank turnover, R_{AS}	-0.30	0.44	-0.38	0.17

Table 5.6: **Principal Components.** The weight of each metric in the first two principal components, for both datasets.

Trait	Related Adjectives
Extraversion	Active, Assertive, Energetic, Enthusiastic, Outgoing, Talkative
Agreeableness	Appreciative, Forgiving, Generous, Kind, Sympathetic
Conscientiousness	Efficient, Organised, Planful, Reliable, Responsible, Thorough
Neuroticism	Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying
Openness to Experience	Artistic, Curious, Imaginative, Insightful, Original, Wide Interests

Table 5.7: **The Big-Five traits and examples of adjectives describing them [210]**

5.2.5 *The big-five personality traits partly explain spatial and social behaviour.*

We verify if the differences between individuals can be explained by the Big five personality traits model [75], typically used to describe social and emotional life (see table 5.7). We build two multiple linear regression models that use the Big five personality traits as regressors and one of the principal components describing socio-spatial behaviour as target. Results, shown in table 5.8, show that three personality traits, neuroticism, openness and extraversion, are relevant predictors for socio-spatial behaviour. In particular, extraversion is the most important predictor of the first principal component: it characterises the tendency to diversify routine and to explore opportunities. Neuroticism and openness explain instead the second principal component, which characterises the tendency to change routine over time (see also fig. 5.5). We verify that the same analysis performed considering only the spatial metrics leads to similar results (see table 5.9, table 5.10, table 5.11 and fig. 5.6). All the result presented above hold when choosing a time-window with length $T = 30$ weeks (see appendix C). As an additional test, we verify that the Pearson correlations between extraversion and the first principal component, between openness and the second principal component, between neuroticism and the second principal component are significant ($p < 0.01$). The same test performed after shuffling the scores on each trait across users yields no significant correlation.

	PC 0 $R^2 = 0.17$, $F = 21.40$, $p_F = 0.0$			**PC 1** $R^2 = 0.03$, $F = 3.64$, $p_F = 0.0$		
	coeff	**p val**	**LMG**	**coeff**	**p val**	**LMG**
E	0.85 ± 0.09	$< 10^{-19}$	0.85	0.12 ± 0.06	0.05	0.14
O	-0.17 ± 0.08	0.03	0.02	0.13 ± 0.06	0.02	0.33
N	0.25 ± 0.09	0.004	0.04	0.15 ± 0.06	0.02	0.3
A	0.11 ± 0.08	0.2	0.04	-0.07 ± 0.06	0.2	0.12
C	0.06 ± 0.08	0.4	0.04	-0.07 ± 0.06	0.2	0.11

Table 5.8: **Extraversion, openness, and neuroticism explain socio-spatial behaviour.** The result of a multiple linear regression explaining principal components of socio-spatial data (see table 5.6) from personality traits (E: extraversion, O: openness, N: neuroticism, A: agreableness, C: conscientiousness). The value of each coefficient (coeff) is reported together with the probability (p val) that the coefficient is not relevant for the model. The relative importance of each coefficient (LMG) is computed using the LMG method [205].

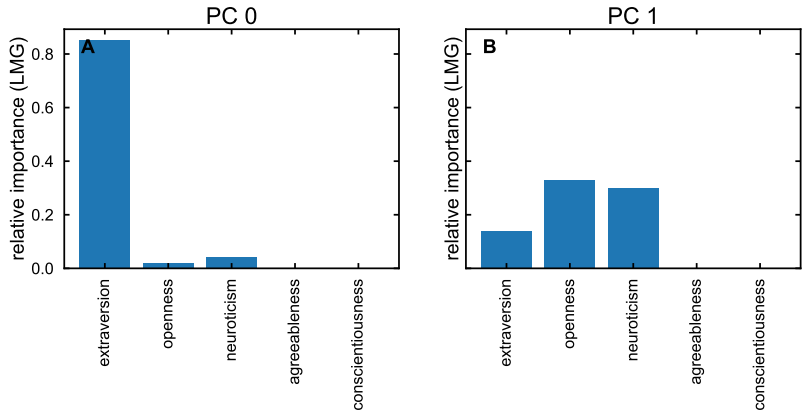

Figure 5.5: **Relative importance of personality traits for socio-spatial behaviour** Relative importance of each personality trait (computed using the Lindeman, Merenda and Gold method [205]) in explaining the first (A) and the second (B) principal components of socio-spatial behaviour (see also table 5.8).

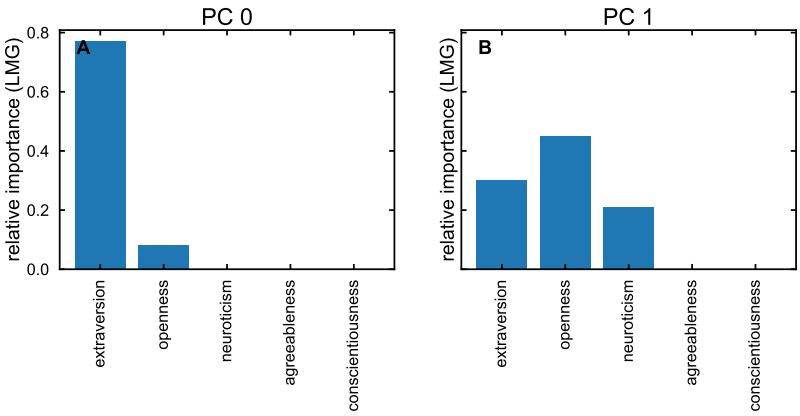

Figure 5.6: **Relative importance of personalitry traits for spatial behaviour** Relative importance of each personality trait (computed using the Lindeman, Merenda and Gold method [205]) in explaining the first (A) and the second (B) principal components of spatial behaviour (see also table 5.11).

	PC 0	**PC 1**	**PC 2**	**PC 3**	**PC 4**
CNS	0.53	0.21	0.13	0.10	0.04
MDC	0.56	0.19	0.13	0.07	0.04

Table 5.9: **Variance explained by principal components (only spatial data).** The fraction of variance explained by each principal component for the CNS and MDC dataset.

	CNS		**MDC**	
	PC 0	**PC 1**	**PC 0**	**PC 1**
Activity space size, C	-0.58	0.02	0.55	0.12
New ties/week, n_{tie}	-0.48	-0.19	0.51	-0.09
Activity space entropy, H_{AS}	-0.50	-0.08	0.43	0.20
Activity space stability, J_{AS}	-0.02	0.94	-0.19	0.95
Activity space rank turnover, R_{AS}	0.43	-0.25	-0.47	-0.16

Table 5.10: **Principal Components (only spatial data).** The weight of each metric in the first two principal components, for both datasets.

	PC 0 $R^2 = 0.10$, $F = 12.83$, $p_F = 0.0$			**PC 1** $R^2 = 0.03$, $F = 3.50$, $p_F = 0.0$		
	coeff	**p val**	**LMG**	**coeff**	**p val**	**LMG**
E	-0.50 ± 0.07	$< 10^{-10}$	0.77	-0.11 ± 0.05	0.02	0.3
O	0.19 ± 0.07	0.004	0.08	-0.11 ± 0.04	0.009	0.45
N	-0.07 ± 0.07	0.4	0.03	-0.10 ± 0.05	0.03	0.21
A	-0.10 ± 0.07	0.2	0.07	0.01 ± 0.05	0.8	0.01
C	-0.05 ± 0.07	0.5	0.05	0.03 ± 0.04	0.4	0.03

Table 5.11: **Extraversion, openness, and neuroticism explain spatial behaviour.** The result of a multiple linear regression explaining principal components of spatial data from personality traits (E: extraversion, O: openness, N: neuroticism, A: agreableness, C: conscientiousness) (see table 5.6). The value of each coefficient (coeff) is reported together with the probability (p val) that the coefficient is not relevant for the model. The relative importance of each coefficient (LMG) is computed using the LMG method [205].

5.3 SUMMARY

Using high resolution data from two large scale studies, we have investigated the connection between social and spatial behaviour for the first time. We have shown that, in both domains, individuals balance the trade-off between exploring new opportunities and exploiting known options in a distinctive and persistent manner. We have found that, to a significant extent, individuals adopt a similar strategy in the social and spatial sphere. These strategies are heterogeneous across the two samples considered, and there is no evidence suggesting that there exist distinct classes of individuals. Finally, we have shown that the big five personality traits explain related aspects of both social and spatial behaviour. In particular, we have found that extraverted individuals are more explorative and have diverse routines in both the social and the spatial sphere while neuroticism and openness associate with high level of routine instability in the social and spatial domain.

Our findings confirm the usefulness of mobile phone data to study the connections between behaviour and personality [161, 177, 180, 211–213]. The results are in line with previous findings on the relation between personality and social behaviour: extraversion correlates with social network size [154, 179, 201], openness to experience to social network turnover [161] and neuroticism does not correlate with

social network size [176]. Finally, our findings establish a relation between personality and spatial behaviour, validating the theories suggesting that spatial choices are partially dictated by personality dispositions [214] and that a single set of personality traits underlies all aspect of a person's behaviour [175, 215]. The individual characterisation of spatial behaviour is also fundamental to develop conceptual [191] and predictive [7] models of travel behaviour accounting for individual-level differences.

6

MODELLING THE DYNAMICS OF SOCIAL INTERACTIONS

As we have seen in the previous chapters, there are deep connections between an individual's social and spatial behaviours: For example, sociable individuals are likely to be more explorative in the spatial domain, and vice versa. While such individual characteristics tend to be distinctive and persistent in time, a dynamic description of social and spatial phenomena is necessary to capture short time scale events as well as long-term developments.

In this chapter, we shift the attention towards the modelling of social interactions at short temporal scales. We adopt a time-varying network model perspective, where nodes are individuals and links with limited duration are the interactions between them. We present a dynamic model where tie formation is driven by two distinctive and persistent properties of individuals: their *activity*, or propensity to engage in social interactions, and their *attractiveness*, or propensity to attract connections. Then, we study the role played by these individual properties on diffusion processes spreading on the network, both analytically and numerically. In future studies we aim at including a spatial component in the model, accounting for the observed correlations between social and spatial attitudes.

The chapter is based on work published in [IV] and it is organized as follows. In section 6.1 we review the relevant literature; in section 6.2 we introduce the network model; in section 6.3, we study the interplay between activity and attractiveness in real networks. In section 6.4 we study the stationary state of the random walks diffusing on the model. In section 6.4.1 we study the mean first passage time.

6.1 STATE OF THE ART

Small-world phenomena along with heterogeneity in the number and frequency of contacts are among the most well known properties of social networks [216–218]. They are often referred to as late or time-integrated properties [67, 68] because they emerge integrating interactions over long time-scales. Traditionally, the modelling efforts put forward to characterise social systems and dynamical processes unfolding on their fabrics focused mainly on these features [216, 219], neglecting the dynamics acting at much shorter time-scales. This was

due to the challenges of introducing the temporal dimension in any mathematical construct and to the lack of real time-resolved datasets. While the former obstacle remains largely unsolved, significant progresses have been made to tackle the later [67–69]. Indeed, the digital revolution has enabled scientists to access a wealth of offline and online data describing social interactions in time. The access to the temporal dimension allows to observe properties of social behaviour that are invisible in time-integrated datasets, and can help characterise microscopic mechanisms driving the dynamics of social acts at all time-scales [132, 220–229]. As a result, an intense research effort has been recently devoted to modeling the temporal dynamics characterising the emergence and evolution of networks. Furthermore, much attention has been directed to understand the effects of these dynamics on processes unfolding on the network such as the spreading of infectious diseases, idea, rumours, or memes [218, 220, 230–254].

Observations in a range of real social networks show that the propensity of individuals to engage in social acts is highly heterogenous [220–222, 229, 232]. Also, it was found that the establishment of connections is highly correlated in time [114, 221–223, 255, 256]. Several studies have focused on understanding the effects of *local* memory in the creation of links. It was shown that different types of local reinforcement mechanisms, including proximity in the space of ideologies, are able to mimic characteristic aspects of social networks such as the emergence of strong and weak ties [221–223, 257–259].

However, in certain circumstances *local* mechanisms alone can not explain the creation of social ties. For example, in online social networks like Twitter individuals can interact with popular figures and access topical pieces of information. Arguably, the creation of these connections does not follow the same local rules driving the emergence of close social ties. Instead, at least to some extent, they may be driven by *global* effects such as interest towards celebrities or for the information provided by popular accounts. Despite the widespread diffusion of these platforms, the modelling of global mechanisms for link creation and the understanding of its effects on diffusion processes unfolding on the network remain largely unexplored. This is especially true when short-time scales and thus time-varying dynamics are considered.

In this chapter, we propose a temporal model of interactions driven by global popularity. In particular, we extend the activity-driven framework [220] in which nodes are assigned an activity defining their propensity to establish contacts per unit time. In its first formulation active nodes connect to others through a memoryless and random selection process [220]. More realistic mechanisms based on local reinforcement of ties have been then proposed [221, 222, 257]. Here, we present a new variation in which nodes are characterised by an attractiveness [260–262], or a popularity index, that might or might not be

correlated with activity and drive the contact selection process. In particular, we consider a classic linear preferential attachment [263]. We then study a random walk process unfolding at the same time-scale in which the connections are created. For sake of simplicity, we consider the fundamental random walk process, which has recently been investigated on different kinds of temporal networks [231, 232, 239, 241, 252, 264, 265]. We find analytical solutions for the stationary state of the process as well as its mean first passage time (MFPT) that match the results produced by numerical simulations. The solutions are general and allow to analytically characterise the interplay between activity and attractiveness considering also their correlations. We ground our results with empirical observations by measuring such correlations on different real datasets and we discuss their repercussions on the random walks.

6.2 TIME-VARYING NETWORK MODEL

In the activity-driven network framework [220] the N nodes of the network are assigned an activity rate a describing their propensity to engage in social acts [220–222, 229, 232]. Here, we consider nodes characterised also by another quantity, namely their *attractiveness* b, describing their popularity in the system [260, 261, 266]. In general, these two quantities are correlated and extracted from a joint distribution $H(a, b)$. At each time step, a node i is activated with probability $a_i \Delta t$ and connects to m others. The generic node j is selected with probability $b_j / \langle b \rangle N$. Each link has a duration of Δt. In fig. 6.1, we show the statistical features of the emerging network considering $N = 10^5, m = 6$ for an uncorrelated system where $H(a, b) = F(a)G(b) \sim a^{-2} b^{-2.5}$, integrating over time τ. Here, we chose $F(a)$ and $G(b)$ to be power-laws functions with exponents similar to those observed in real-world systems. τ is expressed in units of the average time between consecutive activations a_0^{-1}, where $a_0 = \sum_i a_i = \langle a \rangle N$[267]. As clear from the figure, the heterogeneity in activity and attractiveness induces heavy-tailed degree, strength, weight distributions. Note that the exponent characterizing the distribution of strength and degree of a node is the same, since the two are directly related. This is analogous to what is observed in the case of nodes with heterogeneous activity.

6.3 CORRELATION BETWEEN ACTIVITY AND ATTRACTIVENESS IN REAL NETWORKS

The activity measures the propensity of nodes to initiate a social interaction, while attractiveness quantifies the probability of being selected to participate to such interactions, i.e. popularity. These two quantities and their correlation can be studied in real networks, pro-

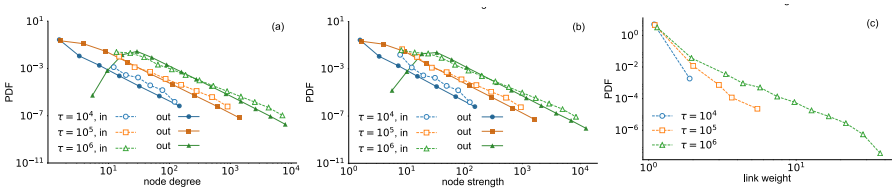

Figure 6.1: **Statistical properties of the time-aggregated network** Probability density function of nodes of given in and out-degree (a) and strength (b) for different values of time-window τ. Probability density function of links of given weight for different values of the time-window τ (c). Results are shown for $N = 10^5$, $m = 6$, $F(a) \sim a^{-2}$, $G(b) \sim b^{-2.5}$. For $\tau = 10^4$, the average in-degree is $\langle k_{in} \rangle = 0.6$, for $\tau = 10^5$, $\langle k_{in} \rangle = 5.7$, for $\tau = 10^6$, $\langle k_{in} \rangle = 57.9$. Note that the average out-degree equals the average in-degree.

vided that interactions are directed and allow to distinguish between the activation and selection process. Here, we consider two datasets. The first describes wall-posts interactions between 45,813 Facebook users over a timespan of 1,591 days [268, 269]. The second describes email replies among 26,885 users involved in the Linux kernel development over 2,921 days [270]. For the sake of this model, we consider the out-strength and in-strength of nodes as proxies for their activity and attractiveness respectively. Hence, activity and attractiveness of node i are computed as $a_i = s_{i,out} / \sum_j s_{j,out}$ and $b_i = s_{i,in} / \sum_j (s_{j,in})$, where $s_{i,in}$ and $s_{i,out}$ are the node in-strength and out-strength integrated across the entire time-span, respectively. Activity and attractiveness are computed aggregating across the whole period of data collection. In fact, observations in a range of real datasets such as co-authorship networks [220, 222], online social networks [220, 232] mobile phone networks [221], and networks created by R&D alliances between firms [229] show that the form of the activity distribution is independent of the aggregation window. In fig. 6.2 we show the distributions of activity and attractiveness in the two datasets. Not surprisingly, in the two datasets both activity and attractiveness follow heavy-tailed distributions spanning several order of magnitude [271]. In fig. 6.3 we plot the correlation between activity and attractiveness considering each node in the two datasets. A positive correlation is clear and in both cases the median follows a power-law with exponent very close to one, i.e. $a \sim b^{\beta}$, $\beta \sim 1$.

6.4 RANDOM WALK

We consider a Markovian and homogenous random walk [272] unfolding on networks generated with the model described above. We focus on the case in which the walker moves at the same time scale describing the evolution of links, moving from node to node when a link is present. The properties of the diffusion process thus are highly

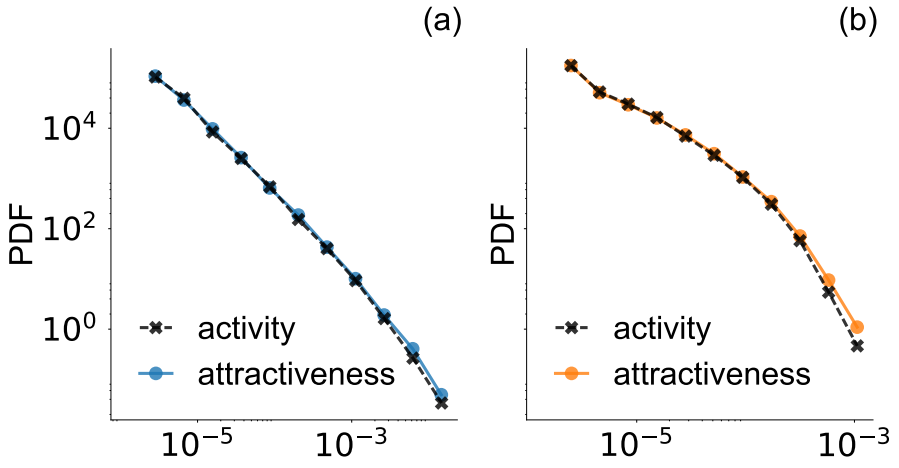

Figure 6.2: **Distribution of activity and attractiveness in real datasets.** Probability density function of activity (a) and attractiveness (b) for the Linux dataset (a) and the Facebook dataset (b). In the Linux network there are $N = 2.7 \cdot 10^4$ nodes, $E = 1.0 \cdot 10^6$ edges, and the period of measurement lasts $T = 2921$ days. For the Facebook network, $N = 4.6 \cdot 10^4$, $E = 8.6 \cdot 10^5$, $T = 1591$ days

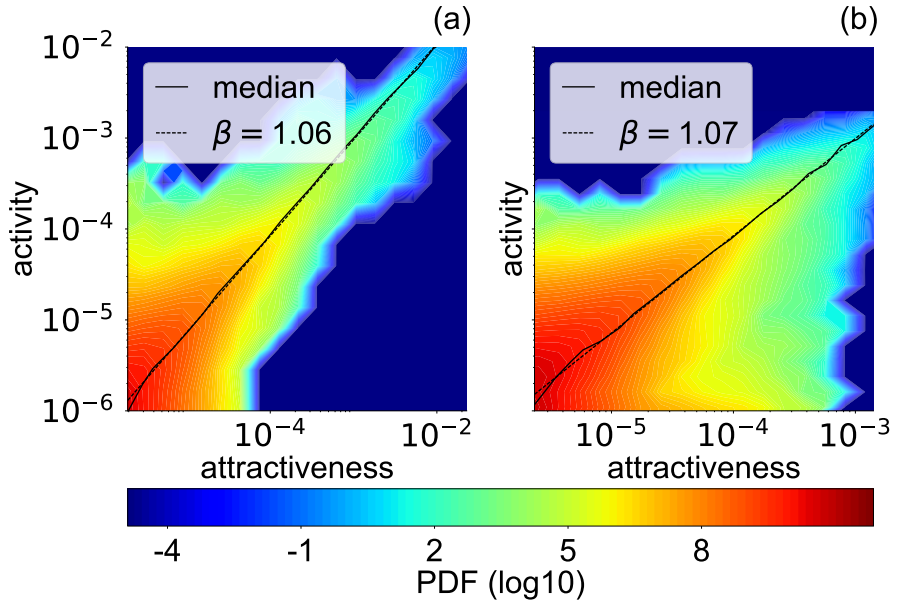

Figure 6.3: **Correlation between activity and attractiveness in real datasets.** Heat map showing the correlation between activity and attractiveness in two real datasets describing interactions between people involved in the development of Linux (a) and on Facebook (b). The continuous line describes the median correlation and the dashed line a power-law fit with exponent β.

affected by the dynamics driving the evolution of the connections.

Let us define $P_i(t)$ as the probability that the walker is in node i at time t. This quantity follows the following master equation:

$$P(i,t+\Delta t) = P(i,t)[1-\sum_{j\neq i}\Pi^{\Delta t}_{i\to j}] + \sum_{j\neq i} P(j,t)\Pi^{\Delta t}_{j\to i} \tag{6.1}$$

where $\Pi^{\Delta t}_{i\to j}$ is the propagator of the random walk that describes the probability that the walker moves from i to j in a time interval Δt. A link between i and j can be created as consequence of the activation of i or j. The probability that i is active and selects j is:

$$p(i\to j) = \frac{ma_i\Delta t b_j}{N\langle b\rangle}. \tag{6.2}$$

Note that this approximation is valid only if $bj/(N\langle b\rangle) << 1$, so the probability to connect to the same node twice is small. This is verified for the values $N = 10^5$ and $\gamma_2 = 2.5$, that we will use along the chapter (see fig. 6.4). In this case, the instantaneous degree of i is:

$$k_i = m + \frac{m\langle a\rangle\Delta t b_i}{N\langle b\rangle}. \tag{6.3}$$

Indeed, i will generate m links and will potentially receive links from other active nodes. The probability that j is active and selects i is instead:

$$p(j\to i) = \frac{ma_j\Delta t b_i}{N\langle b\rangle}. \tag{6.4}$$

The instantaneous degree of i will be:

$$k_i = 1 + \frac{m\langle a\rangle\Delta t b_i}{N\langle b\rangle}. \tag{6.5}$$

In the limit $\Delta t \to 0$, the events described by equations (6.2) and (6.4) do not happen simultaneously. Putting all together is easy to show that, for $\Delta t \to 0$:

$$\begin{aligned}\Pi^{\Delta t}_{i\to j} &= \frac{ma_i\Delta t b_j}{N\langle b\rangle}\frac{1}{m+\frac{m\langle a\rangle\Delta t b_i}{N\langle b\rangle}} + \frac{ma_j\Delta t b_i}{N\langle b\rangle}\frac{1}{1+\frac{m\langle a\rangle\Delta t b_i}{N\langle b\rangle}} \\ &\simeq \frac{\Delta t}{N\langle b\rangle}(a_i b_j + m a_j b_i).\end{aligned} \tag{6.6}$$

In the limit $\Delta t \to 0$ we can write the equation describing the evolution of $P_i(t)$ by substituting the expression of the propagator ineq. (6.1):

$$\begin{aligned}\dot{P}(i,t) = -\frac{P(i,t)}{N\langle b\rangle}\sum_{j\neq i}(a_i b_j + m a_j b_i) + \sum_{j\neq i}\frac{P(j,t)}{N\langle b\rangle}(a_j b_i + m a_i b_j) = \\ -\frac{P(i,t)}{\langle b\rangle}[a_i\langle b\rangle + m b_i\langle a\rangle] + \frac{b_i}{N\langle b\rangle}\sum_j P(j,t)a_j + \frac{ma_i}{N\langle b\rangle}\sum_j P(j,t)b_j.\end{aligned} \tag{6.7}$$

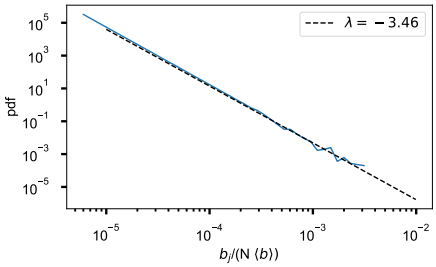

Figure 6.4: **Low probability of selecting the same node twice.** The distribution of $b_j/(N\langle b\rangle$. We considered $N = 10^5$ and $\gamma_2 = 2.5$. The black line shows a power law fit with exponent $\lambda = -3.46$.

We obtain a system level description of the process by grouping nodes in the same activity class a and attractiveness b, assuming that they are statistically equivalent [273]. Then, we define the walkers in a given node of class a and b at time t as $W_{ab}(t) = [NH(a,b)]^{-1} W \sum_{i\in a \& \in b} P_i(t)$, where, W is the total number of walkers in the system. By considering the continuous a and b limit,section 6.4 can be rewritten as:

$$\begin{aligned}
\dot{W}_{ab}(t) &= -\frac{W_{ab}(t)}{\langle b\rangle}[a\langle b\rangle + mb\langle a\rangle] \\
&+ \frac{b}{\langle b\rangle}\iint a' W_{a'b'}(t) H(a',b') da' db' \\
&+ \frac{ma}{\langle b\rangle}\iint b' W_{a'b'}(t) H(a',b') da' db' \qquad (6.8) \\
&= -\frac{W_{ab}(t)}{\langle b\rangle}[a\langle b\rangle + mb\langle a\rangle] + \frac{b}{\langle b\rangle}\phi_1 + \frac{ma}{\langle b\rangle}\phi_2, \qquad (6.9)
\end{aligned}$$

where $\phi_1 = \iint a' W_{a'b'}(t) H(a',b') da' db'$ and $\phi_2 = \iint b' W_{a'b'}(t) H(a',b') da' db'$. In the stationary state, the changes of $W_{ab}(t)$ are zero, thus we have:

$$W_{ab}(t) = \frac{b\phi_1 + ma\phi_2}{a\langle b\rangle + mb\langle a\rangle}. \qquad (6.10)$$

The stationary state features a and b in both numerator and denominator. Hence, the dynamical properties of the random walk are function of the interplay between the two quantities. It is important to notice that at the stationary state ϕ_1 and ϕ_2 are constant. Their value can be computed self-consistently by solving this system of integral equations:

$$\begin{aligned}
W &= N\iint H(a,b)\frac{b\phi_1 + ma\phi_2}{a\langle b\rangle + mb\langle a\rangle} da db \\
\phi_2 &= \iint bH(a,b)\frac{b\phi_1 + ma\phi_2}{a\langle b\rangle + mb\langle a\rangle} da db, \qquad (6.11)
\end{aligned}$$

where the first equation follows from the conservation of walkers in the system.
We test the analytical solutions against numerical simulations run following the Gillespie algorithm [267]. The algorithm works as follows:

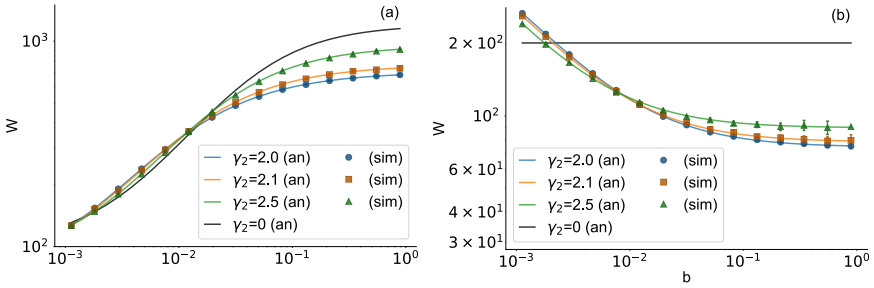

Figure 6.5: **Stationary state of the random walk process.** The average number of walkers per node of class a (a) and b (b) computed analytically (continuous lines) and through numerical simulations (dots, squares and triangles) for different values of exponent of the attractiveness distribution γ_2. Error-bars are standard deviations obtained by averaging across 10^3 different network configurations. In the above panel, they are not visible on the scale of the graph. We considered $N = 10^5$, $m = 6$, $W/N = 200$, and $\gamma_1 = 2$. The black line shows the case where $b_i = 1/N$ for all nodes.

First, one extracts the time elapsed between a node activation and the following from an exponential distribution, with mean equal to $1/\sum a_i$; Then, the activated node i is chosen with probability $a_i/\sum a_i$. The node connects instantaneously with m other nodes, where each connected node j is picked with probability $b_j/\sum b_j$. As a first step, let us consider the uncorrelated case in which both a and b are extracted from a power-law distribution: $H(a,b) = F(a)G(b)$ where $F(a) = Aa^{-\gamma_1}$ and $G(b) = Cb^{-\gamma_2}$. In both cases values are extracted in the range $x \in [10^{-3}, 1]$. In fig. 6.5 we plot the comparison between the average number of walkers per nodes of class a and b separately. In fig. 6.6 we plot instead $W_{a,b}(t)$ as a heat map. In both cases, the agreement between simulations and analytical predictions is clear.

Taken together, the two figures present a rich picture. First, they show that the larger the activity, the larger the capability of gathering walkers. The trend holds up to a saturation point after which an increase in activity does not translate to an increase of walkers, similarly to what is observed in Ref. [231] for the case of constant attractiveness, i.e. random tie selection process (see also fig. 6.5, bottom panel, black filled line). Second, they reveal an opposite trend for increasing values of b, as, before saturation, the larger the attractiveness the smaller the number of walkers in the stationary state. While this finding could seem counterintuitive, it can be understood considering the structure of the instantaneous network where walkers move. In the limit $\Delta t \to 0$, the degree of an active node i is $k_i \sim m$, while the degree of a node j connected by i is $k_j \sim 1$ as non-active nodes do not 'have time to' accumulate multiple connections. Thus, even extremely attractive nodes, that are involved in many connections across time, appear instantaneously as nodes with degree 1. Consequently, a node

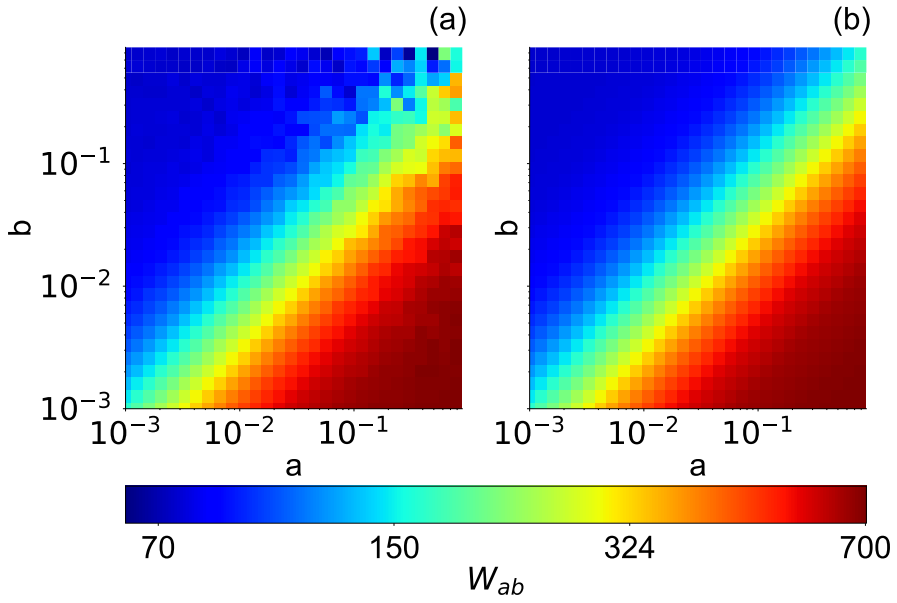

Figure 6.6: **Stationary state of the random walk process for nodes of class (a, b).** Heat map giving the average number of walkers W_{ab} per node of class (a, b) computed through numerical simulations (a) and analytically (b). Colours are attributed based on W_{ab} as shown in the colorbar on the bottom of the figure. We considered $N = 10^5$, $m = 6$, $W/N = 200$, $\gamma_1 = 2$, and $\gamma_2 = 2$.

selected by i receives on average a fraction $1/m$ of the walkers of i, but it sends all its walkers to i. This fact explains the decreasing trend of $W(b)$ and shows at a fundamental level the effects of temporal interactions on diffusion processes taking place on the same timescale. As a consequence, in the case of a random-tie selection process nodes with large activity are able to collect more walkers than in the case of heterogeneous b, due to the tendency to select nodes holding fewer walkers than average in the latter case.

To further understand these effects, we study the case of random walks unfolding on static networks obtained by integrating activity-driven networks with attractiveness over time windows of size τ. In doing so, we let nodes activate and connect to other nodes for a time τ. Then, we let the random walk unfolds on the union of such networks. Note that, in this case, interactions are not instantaneous. In fig. 6.7, we show the stationary state of the process as a function of the nodes activity and attractiveness, for different value of τ. In contrast to what observed when the diffusion process and the topology evolve at the same timescale, here the walkers concentrate also on highly attractive nodes. This result is expected. The stationary state of random walks unfolding on any static network is linearly proportional to the degree [216, 272]. In our case, nodes with large attractiveness are likely to be hubs: characterised by large degree values. These effects are more evident for increasing values of the time-aggregation window.

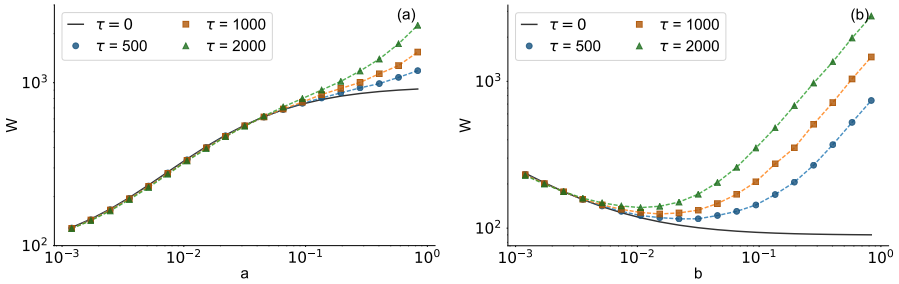

Figure 6.7: **Stationary state of the random walk in the aggregated case.** The average number of walkers per node of class a (a) and b (b) computed analytically for $\tau = 0$ (continuous line) and through numerical simulations for several values of τ (dots, squares and triangles). Dashed lines are shown as a guide for the eye. Errorbars obtained by averaging across 10^3 different network configurations are not visible on the scale of the graph. We considered $N = 10^5$, $m = 6$, $W/N = 200$, $\gamma_1 = 2$, and $\gamma_2 = 2.5$.

Indeed, the larger τ, the larger the degree of highly attractive nodes. For similar reasons, the same qualitative behaviour is observed also for nodes with high activity.

Considering the observations in real datasets, we turn now the attention to scenarios in which activity and attractiveness are correlated. In particular, we consider for each node a deterministic correlation of the form $b \sim a^\beta$, or more in general $b = J(a)$ where J is a generic function. The joint probability can then be written as $H(a,b) = F(a)\delta(b - J(a))$, where $\delta(x)$ is the Dirac delta. In fig. 6.8 we show the stationary state of the random walks for several values of β.

For $\beta < 1$ trends are not far from the uncorrelated case. For larger activity, nodes have higher capability of gathering walker and $W(a)$ saturates for large values of a, while the opposite trend holds for $W(b)$. Indeed, the negative correlation reinforces what is observed in the uncorrelated case since nodes with low-activity have also high attractiveness. Hence, we observe that the larger β, the smaller is the number of walkers collected by nodes with low activity and the faster $W(a)$ saturates.

Instead, for $\beta > 1$, the larger the activity, the lower the capability of gathering walkers. In this case, the set of nodes more frequently engaged in active interaction has also attractiveness much larger than the average node. These nodes tend not to hold walkers but to exchange them continuously. Instead, walkers are likely to be trapped in nodes that are unlikely to engage in interaction.

For $\beta = 1$, since the rate at which node is activated and the probability to be selected are exactly the same, $W(a)$ and $W(b)$ are constant.

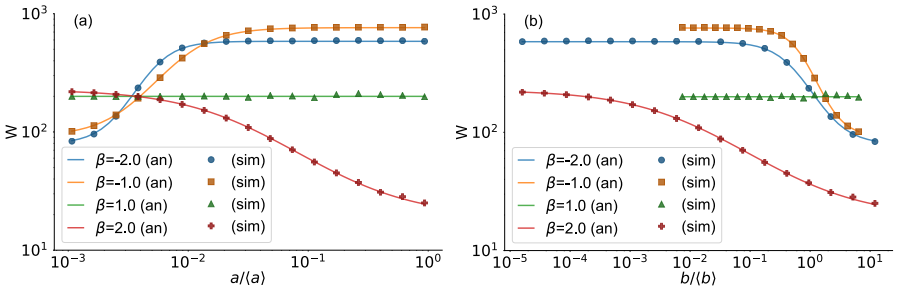

Figure 6.8: **Stationary state of the random walk in the correlated case.** The average number of walkers per node of class a (a) and b (b) computed analytically (continuous lines) and through numerical simulations (dots, squares, triangles and crosses) for different values of the correlation exponent β. Error-bars obtained by averaging across 10^3 different network configurations are not visible on the scale of the graph. We considered $N = 10^5$, $m = 6$, $W/N = 200$, and $\gamma_1 = 2$.

6.4.1 *Mean First Passage Time*

We now consider the mean first passage time (MFPT), defined as the average number of time steps needed for a walker to visit a node i starting from any other node in the system [272, 274, 275].

Let us consider $p(i,n)$ as the probability that the walker reaches i (the target) for the first time at time $t = n\Delta t$. Considering that each node could be connected directly to any other, we have:

$$p(i,n) = \xi_i(1-\xi_i)^{n-1}, \tag{6.12}$$

where ξ_i is the probability that the walker jumps in node i in a time interval Δt, that is:

$$\xi_i = \sum_j \frac{W(a_j,b_j)}{W}\Pi_{j\to i}^{\Delta t}. \tag{6.13}$$

Indeed, the propagator by definition encodes the probability that walkers moves from j to i, and $W(a_j,b_j)/W$ describes the probability that the walker is in j at time t (in the stationary state). Thus, we can estimate the MFPT as:

$$\begin{aligned} MFPT_i &= \sum_{n=0}^{\infty} n\Delta t \cdot p(i,n) = \frac{\Delta t}{\xi_i} \\ &= \frac{N\langle b\rangle w}{b_i\sum_j W(a_jb_j)a_j + ma_i\sum_j W(a_jb_j)b_j} \\ &= \frac{\langle b\rangle w}{b_i\phi_1 + ma_i\phi_2}. \end{aligned} \tag{6.14}$$

It is interesting to notice how in static and annealed networks (where the timescale of the random walk is either much faster or slower with

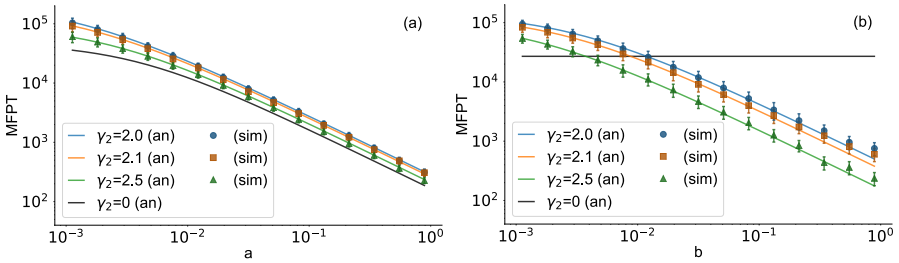

Figure 6.9: **Mean first passage time.** The average MFPT as function of a (a) and b (b) computed analytically (continuous lines) and through numerical simulations (dots, squares and triangles) for different values of exponent of the attractiveness distribution γ_2. Errorbars are standard deviations obtained by averaging across 10^3 simulations. We considered $N = 10^3$, $m = 6$, and $\gamma_1 = 2$.

respect to changes in the topology where it is unfolding) ξ_i is equivalent to the stationary state of the random walk, i.e. $\xi_i = W_i/W$. In time-varying networks instead this is not the case as the walker can be trapped in an inactive or unpopular node for several time steps [231]. Consequently, the expression of ξ considers explicitly the dynamical connectivity patterns to account for such delays.

In fig. 6.9 we test the validity of the analytical expression for the MFPT. We fixed γ_1 and considered different values of γ_2 assuming uncorrelated activities and attractiveness. In fig. 6.10 we show the comparison between the average values of MFPT for nodes of class a and b. In both cases we find very good agreement between theory and simulations. However, we observe that for the most attractive nodes, the $MFPT \sim 200$ time units. This is less than the time needed to reach the stationary state. In fact, given that the average time between consecutive activations is $1/\sum a_i$, there are only ~ 100 node activations occurring in the first 200 time units under our choices $N = 1000$ and $\gamma_1 = 2$. This explains why the analytic prediction slightly underestimates the MFPT for very attractive nodes, that are the more likely to be targeted in the initial steps. It is interesting to observe that the effect of heterogeneous attractiveness is to introduce delays in the transport dynamics since the MFPT is larger for all nodes with respect to the random-tie selection process case (fig. 6.9, bottom panel, black line).

6.5 SUMMARY

We presented a model of time-varying networks in which nodes are characterised by activity and attractiveness, regulating their propensity to initiate an interaction and their popularity, respectively. In particular, we extended the framework of activity-driven networks by introducing a tie selection mechanism based on a global linear pref-

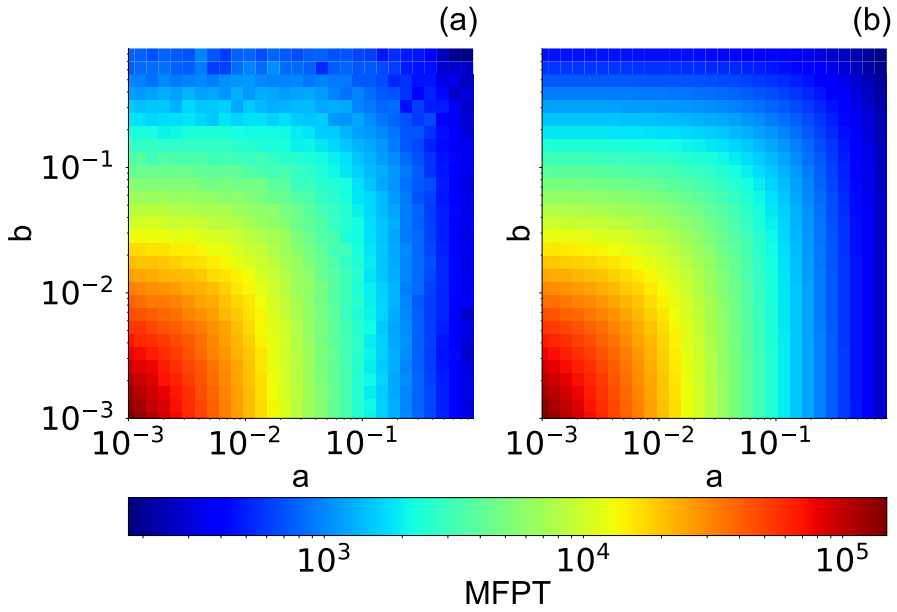

Figure 6.10: **Mean first passage time for nodes of class (a, b).** Heat map for the average values of MFPT per node of class (a, b) computed through numerical simulations (a) and analytically (b). Colours are attributed based on the value of MFPT as shown in the colorbar on the bottom of the figure. We considered $N = 10^3$, $m = 6$, $\gamma_1 = 2$, and $\gamma_2 = 2$.

erential attachment. We grounded our model with empirical observations by measuring activity and attractiveness from the out-strength and in-strength of nodes in two real time-varying networks describing interactions between i) users on Facebook and ii) people involved in the development of a software. Interestingly, we observed that both activity and attractiveness are heterogeneously distributed and correlated. In the two datasets the correlation is positive.

We then studied the interplay between activity and attractiveness and its effects on the prototypical random walk process. We derived analytical expressions for the stationary state and for the MFPT of the process unfolding on the time-varying network model. We thoroughly tested the analytical predictions via large-scale numerical simulations obtaining very good agreement between the two. Overall, the results shed light on how the presence of temporal connectivity patterns significantly alters the standard picture obtained in static and annealed networks. The presence of a global tie selection process and the possible correlation between activity and attractiveness introduce non-trivial effects. The stationary state and MFPT are significantly different from those obtained in activity-driven networks characterised by a random tie selection mechanism. In the uncorrelated case the effect of heterogeneous attractiveness is to limit the capability of very active nodes to gather walkers. In the case of positive correlations between activity and attractiveness, observed in real scenarios, the

stationary state of the process is substantially altered: The average number of walkers per node decreases as a function of the node activity for $\beta > 1$, it is constant for $\beta = 1$. Heterogeneous attractiveness furthermore slows down the transport dynamics, as we observe that in this case the MFPT is larger for all nodes.

The presented model can be further enriched in several ways. In particular, the activation dynamics it describes is Poissonian rather than bursty as typically observed in real systems [115, 257, 276–282]. The tie selection process is driven only by global popularity and neglects local tie reinforcement mechanisms responsible for high-order organisation of real networks. The framework of activity-driven networks has been extended in several instances to include such features [221, 222, 257]. However, while heterogeneous distributions of popularity characterize various real-world networks, from social-media to financial systems, the study of nodes' popularity and its effect on networks' dynamical properties was missing. Random walks are prototypical example of dynamical processes spreading upon networks. With opportune modifications, random walks on activity driven network with attractiveness could be a starting point to model the exchanges of assets within financial markets, or the diffusion of information over a social network. In the first case, modifications include accounting for the fact that nodes do not necessarily exchange their entire capital; while in the second case one should consider that nodes always keep a copy of the piece of information they exchange.

7

URBAN MOBILITY PATTERN AND MULTILAYER TRANSPORTATION SYSTEMS

Information on the displacements of single individuals can be aggregated to study flows of people travelling between different areas. This understanding is fundamental to create new solutions to society-wide technological problems and policy issues, from urban planning and traffic forecasting, to controlling international migrations.

In this chapter, we shift the attention from the individual level description of trajectories to the study of mobility patterns at the urban scale. In particular, we suggest a method to compare the transportation network of a city and the commuters flows within the same urban agglomeration. Our methodology accounts for the need to limit the total travel time and the number of line changes.

The chapter is based on research published in [V], and it is organized as follows: In section 7.1, we provide a brief summary on the study of public transportation systems under a network science perspective; In section 7.2 we present a novel representation of public transportation systems as multimodal networks; In section 7.3 we apply this framework to compare the transportation systems and the commuting flows of several French cities.

7.1 STATE OF THE ART

Urban transportation systems interweave our everyday life and although their construction is based on conscious design they appear with complex structural and dynamical features [283].

They build up from different transportation means, which connect places in a geographical space. Their most straightforward description is given by networks [284, 285] where stations are identified as nodes and links are the transportation connections between them. Based on this representation [286] considerable research efforts have been dedicated to address their sustainability [287], to optimise their efficiency [288, 289], reliability [290–292] or even to estimate risk they carry due to interdependency with other infrastructure networks in case of terrorist attacks [293].

All transportation networks share a few common features: (a) they are all embedded in space, setting constraints in their structural design, (b) networks of different transportation means may coexist in

the same space, and (c) they are all inherently temporally-resolved. Such details of several transportation networks became available lately [294] through the collection of large open datasets describing complete multimodal transportation systems in cities, regions, countries, and even internationally. These advancements were induced by novel data collection techniques including Smart Card Data [295], Automatic vehicle location (AVL) data [296], and mobile phone data from GSM providers [297]. On the top of these developments the advent of a new common non-proprietary transit data format, the General Transit Feed Specification (GTFS), further amplified actual trends in urban policy propagating smart city programs and real time online user services. As of February 2016, 325 public transportation companies around the world have released official GTFS feeds [298], which are regularly modified by online communities that are adding extensions and optional fields to adapt to different transit services [299]. In transportation, the confluence of open data, GTFS, ubiquitous mobile computing, sensing and communication technologies, has allowed to study the efficiency and performance of public transportation systems under different perspectives [300–304]. As a consequence GTFS data is now used for trip planning, ride-sharing, timetable creation, mobile data, visualisation, accessibility, and to provide real-time service informations.

These recent developments in data collection practices and in the corresponding fields of complex networks and human dynamics provided the opportunity to quantitively study transportation systems using a data-driven approach. These studies showed that geographical constraints largely determine the structure and scaling of transportation networks [305–307] but for their better understanding one needs to consider the actual urban environment and development level [289, 308, 309]. At the same time the emerging field of multilayer networks provided the methodology to consider their multimodal character [310, 311]. In this representation each layer corresponds to the network of a single transportation mean (bus, tram, train, etc.), which are defined on the same set of nodes (stations). This way they account for possible multiple links of different modes between the same stations [312]. This representation can be extended to capture the temporal nature of the system by using some aggregated information extracted from the transportation schedule [313] or, as a future challenge, by considering each time slot as a layer where journeys between stations are represented as temporal links [310, 314].

Here we build on these contemporary advancements and provide a novel representation, which combines multi-edge and P-space representations of transportation networks. The proposed scheme considers the system from the user's point of view by incorporating the minimisation of the total travel time, its variability across the schedule, and the number of transfers between lines. Our subsequent aim is to

adjust earlier defined characterisation techniques to the proposed representation in order help its analysis. We use the adjusted techniques (a) to identify patterns of privileged connections in the transportation network, which are not evidently present through their overall design; and (b) to quantify their overall efficiency as compared to the commuting flow. We carry out our analysis using openly shared GTFS datasets describing extensive transportation networks of French municipalities like larger Paris, Toulouse, Nantes, and Strasbourg.

As follows first we describe the actual time-resolved multilayer network representation and introduce our methodology incorporating travel routes and times to identify efficient transportation connections. Next we apply a matrix factorisation method to extract underlying connectivity patterns to analyse them from the commuter point of view, and quantify their overall efficiency. Finally we conclude our results and discuss possible applications and future directions of research. Note that the implementation of the proposed methodology is openly accessible online[1].

7.2 REPRESENTATION OF PUBLIC TRANSPORTATION NETWORKS

The proposed methodology integrates several sequential steps to detect origin-destination areas that are conveniently connected by public transportation with respect to user preferences. In the following description, first we define a user-based representation of a Public Transportation (PT) system, which limits the effect of its spatial embeddedness, but accounts for its multilayer structure, and its temporal dimension. Next, we calculate shortest time paths between stops by adapting a conventional algorithm [315] to the actual graph representation, and finally we select preferred connections, taking into account distance travelled over time.

7.2.1 *User-based multi-edge P-space representation*

Earlier studies revealed that the choice of users to select transportation means for commuting is mainly affected by the average travel time, and by the variability of the total travel time [316] [2], [317][3], in addition to the number of transfers they need to do. Our principal goal here is to introduce a novel representation of PT networks, which

1 https://github.com/lalessan/user_basedPT

2 This document provides a study on the factors influencing the choice of the transportation means. The study is based on a literature review and statistical analysis of surveys.

3 This document presents results of a survey about transport and mobility of households in France. The authors consider the distance travelled, if it is a long distance trip or short trip made on a daily scale. Moreover they look at the differences of behaviours in transportation use among different regions and individuals with different socio-demographic characteristics.

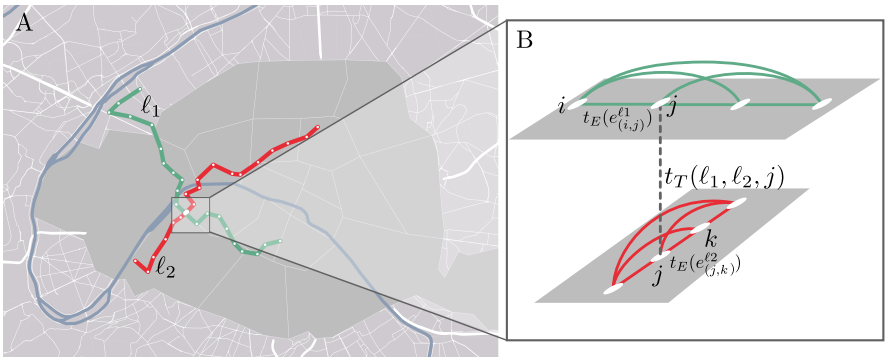

Figure 7.1: **Illustration of the user-based multi-edge P-space representation.** A) Two geo-localised crossing PT lines ℓ_1 and ℓ_2 are shown on the map of central Paris. B) Schematic illustration of the P-space multi-edge representation for a section of the network: all pairs of nodes corresponding to stops on the same line are connected by edges with the same label.

incorporates the aforementioned aspects decisive for users (while neglecting other less determinant factors such as travel cost or comfort), and which minimises the effects due to the spatial embeddedness of the system. In order to do so we combine a multi-edge [318] and a P-space representation of the transportation network [319–321] to describe the PT systems. The multi-edge representation accounts for the presence of several transportation lines in the same PT network by allowing the existence of multiple labelled edges within a single pair of nodes. On the other hand, the P-space representation takes into account that transfers between lines is time-consuming and may not be convenient for the user; also, it considers connections between stops located at large distance thus it reduces the effect of the geographical distances. The combination of these two representations constitutes an ideal framework to investigate complex features of PT systems from the user perspective. A schematic example of this representation is displayed in fig. 7.1: on the left, we show two crossing PT lines, and on the right we illustrate the corresponding P-space multi-edge representation. Two stops i (resp. k) and j on the same line ℓ_1 (resp. ℓ_2) are linked through the edge $e^{\ell_1}_{(i,j)}$ (resp. $e^{\ell_2}_{(j,k)}$), with weight $t_E(e^{\ell_1}_{(i,j)})$ (resp. $t_E(e^{\ell_2}_{(j,k)})$). At node j a transfer is possible between the two lines, which is represented by a link with weight $t_T(\ell_1, \ell_2, j)$ corresponding to the actual time of transfer.

Formally, the public transportation system is defined as a weighted, directed, edge labelled graph $G = (V, E, t_E, T, t_T)$ with vertex set V with cardinality N, corresponding to the public transportation stops, edge set E with weight function t_E, and set of transfers T with weight function t_T. If a line ℓ_k is defined as an ordered sequence of stops connected consecutively, in the corresponding P-space graph G there will be a direct labelled-edge $e^{\ell_k}_{ij} \in E$ connecting each pair of nodes (i, j)

on the given line, such that stop i precedes stop j in the sequence of line ℓ_k. This way each transportation line appears as a fully connected clique in the P-space representation. We define M as the total number of lines in the PT system. Further, a set $T \subset M \times M \times N$ of transfers identifies triplets of two lines and one node, $e^T_{\ell_1,\ell_2,j} = (\ell_1, \ell_2, j)$ assigning a possible transfer between lines ℓ_1 and ℓ_2 at station j. Each edge in E is weighted by the average travel time on the actual line. It is computed through a *time* function $t_E : E \to \mathbb{R}^+$, quantifying for each edge $e^{\ell_k}_{ij}$ the time needed to get from i to j along the line ℓ_k averaged on a selected time window $[h1, h2]$ over N_w weeks. The travel time assigned to an edge $e^{\ell_k}_{ij} \in E$ is then calculated as the sum of the average waiting time and the average time spent on the vehicle as

$$t_E(e^{\ell_k}_{ij}) = \frac{1}{2f_{\ell_k}} + \Delta t^{\ell_k}_{ij} \tag{7.1}$$

where f_{ℓ_k} is the average frequency of line ℓ_k and $\Delta t^{\ell_k}_{ij}$ is the average time one needs to spend on line ℓ_k to go from stop i to stop j. This formula is designed to consider the case where a user would go blindly to a stop (without looking at the schedule). An other approach would include that certain passengers attempt to reduce their waiting time by timing their arrival at transit stops to an optimal period before vehicle departure. Most studies report that passengers facing short headways or low reliability do not generally pursue these strategies [322–324]. Hence, we choose our approach to favour lines with high frequency and less variability due to unexpected perturbations while accounting for preference of users for low unexpected variability in the total travel time. Finally the transfer time function $t_T : T \to \mathbb{R}^+$ quantifies for each transfer $e^T_{\ell_1,\ell_2,j}$ the time needed to change between lines ℓ_1 and ℓ_2 at node j.

In such description the temporality of the system is included through the weights. The choice not to model the system as a temporal graph is motivated by the fact that in urban public transportation systems the total travel time is subject to variability and this factor matters considerably for the user when deciding to opt for public transportation service.

7.2.2 *Uncovering efficient transportation connections*

The previously defined public transportation graph $G = (V, E, t_E, T, t_T)$ is used to calculate shortest time paths between stops. In the multi-edge representation a path is defined as a sequence of edges [4] $P_E = \{e^{\ell_{i_1}}, e^{\ell_{i_2}}, ..., e^{\ell_{i_n}}\}_{o,d}$ connecting an origin node o to a destination node d through a sequence of consecutive trips made on n lines, $\ell_{i_1}, \ell_{i_2}, ..., \ell_{i_n}$. Considering also the sequence of corresponding transfers between

4 In the current paragraph, to simplify notations, we do not index edges by node names.

lines $P_T = \{e^T_{\ell_{i_1},\ell_{i_2}}, e^T_{\ell_{i_2},\ell_{i_3}} ..., e^T_{\ell_{i_{n-1}},\ell_{i_n}}\}_{o,d}$ the shortest time paths between origin and destination are taken as the smallest durations measured among the different alternative paths. Each time length is defined as

$$L_P = \sum_{j=1}^{n} t_E(e^{\ell_{i_j}}) + \sum_{j=1}^{n-1} t_T(e^T_{\ell_{i_j},\ell_{i_{j+1}}}) \tag{7.2}$$

i.e. the sum of the average time needed to wait, travel and transfer between lines.

We adapted the Dijkstra algorithm [315] to provide approximated shortest path lengths between any pair of stops in the user-based multi-layer representation, while keeping the interpretable description of the PT system and reduced computation time. The original version of the algorithm computes the minimal distance between any origin o and destination d nodes by considering the sum of link weights. Instead, the modified version accounts for the fact that not only the link weights have to be taken into consideration but also the transfer time, i.e. the cost to change between different layers (see appendix D.3). Also, to consider the preference of users to change lines in a limited number of times the algorithm allows at most two transfers in a single path, i.e., we limit $n \leq 3$. Due to these limitations the algorithm provides us an approximate solution, however which differs from the correct solution only in few cases. We find that more than 95% of the paths with at most 3 line changes computed with the unlimited (correct) algorithm and the limited (approximate) algorithm have the same temporal length in all cities. After computing the shortest paths between all nodes in the graph we characterise the distribution of shortest travelling times between all nodes whose physical distance falls within a specific range. Using these informations we identify privileged connections, i.e. fastest routes at a given distance.

7.2.3 *Implementation of the user-based representation*

The methodology presented above rely on informations, which are typically included in data given in GTFS format [5] such as trips, routes, travelling times, frequencies and transfer times recorded for each service line and station in the transportation system (for further details see appendix D.1). Using such data we build the P-space multi-edge representations of larger Paris, Strasbourg, Nantes ,and Toulouse. We decided to use a period of $N_w = 4$ weeks in each case, such that the total number of trips per day presents only weak fluctuations. We were interested in trips planned between $h1 = 7am$ and $h2 = 10am$ (though the choices of N_w, $h1$, and $h2$ are adjustable parameters). This choice of time window was made to focus on morning commuting patterns,

5 https://developers.google.com/transit/gtfs/reference

and because during this time interval the frequency of services is considerably higher with respect to the rest of the day. Typical line frequencies and trip durations are then defined as their averages over the selected time window over the four weeks. All PT systems considered rely substantially on three transportation modalities: metro (Paris, Toulouse) or tram (Nantes and Strasbourg), bus and rail. However, they differ considerably in terms of size (see table D.2), routes length, number of stops per route and route frequencies (see fig. D.1).

Finally, building on the multi-edge P-space representation and the estimation of the typical times and frequencies, we compute the typical shortest time paths between any pairs of origin and destination in the city. The implementation of this methodology is available online[6] and requires as input any dataset in GTFS fomat. The shortest paths are computed using a modified version of the Dijkstra algorithm, which allows to find the shortest path between a source node s and all other nodes (see fig. D.2). In the original Dijkstra, all nodes are originally assigned an infinite 'tentative distance' from the source node, while the source node is assigned a 'tentative distance' equal to zero. At each iteration, the node v with the shortest tentative distance is visited. The 'tentative distance' of each of its neighbours u is updated as the minimum between the current tentative distance of u and the distance from u to v plus the tentative distance from s to v. Then, node v is marked as visited (it will not be visited again) and its distance to the source will be equal to its current tentative distance. We introduce two substantial modifications to this algorithm. First, we include the time required to transfer between edges by adding its value to the tentative distance (which in our case is actually a temporal distance). Second, we update the tentative distance of a node only if the number of edges one should travel to get there is smaller than 3. These two approximations are introduced in order to reduce the computation time, reduce the complexity of the PT system representation, and take into account the fact that users are typically not willing to effectuate more than 3 line changes. The algorithm may overestimate the shortest path lengths in cases were locally optimal strategies does not provide a globally optimal solution. However, it provides in general a very good approximation as it is demonstrated in appendix D.3.

7.3 ILLUSTRATION: FINGERPRINTS OF PUBLIC TRANSPORTATION NETWORKS

We demonstrate one possible use of our framework through the examples of the PT systems of larger Paris, Strasbourg, Nantes, and Toulouse. After selecting privileged connections, we apply non-negative matrix factorisation to the graph of the privileged connections to iden-

6 https://github.com/lalessan/user_basedPT

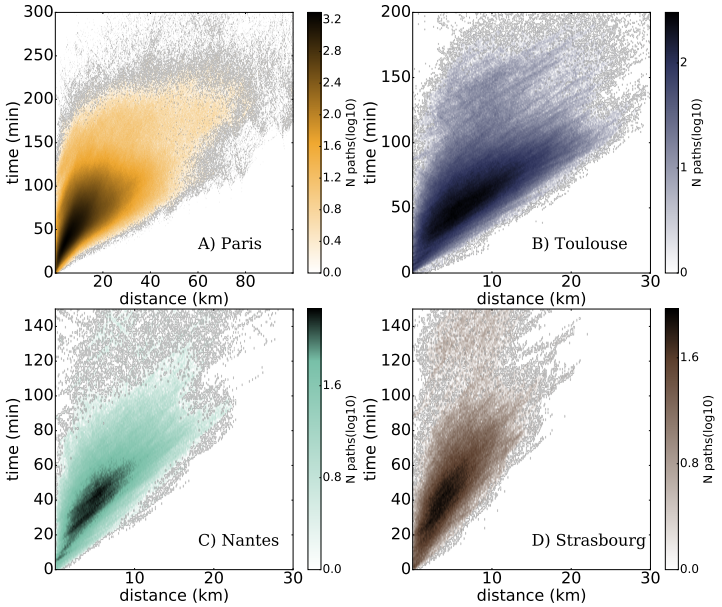

Figure 7.2: **Scatter plot of time versus physical distance associated to shortest time paths for each origin-destination pair.** The points are coloured according to the number of points in the area considered. Scatter plots are shown for the cities of Paris (A), Toulouse (B), Nantes (C), and Strasbourg (D). Colours indicate the logarithm of the number of origin-destination pairs in a given range time-distance bin.

tify underlying patterns, which may not be present due to overall design. Finally, we compare our findings with independent measures of commuting patterns, which allow us to give an estimation about the efficiency of the PT systems.

7.3.1 *Selection of efficient connections*

We used the method previously presented to compute the shortest time paths for each origin-destination pairs of the transportation systems of bus, train and metro. With the selection of a suitable time-window of size $N_w = 4$ weeks, we find that the average standard deviation of a route frequency across the 4 weeks, is about 0.05 transits/hour for all the city considered (focusing only on week-days). This result confirms that the system behaviour is subject to low variability during the period considered. Based on the shortest paths calculations, we built a time-distance map, which assigns the physical distance $d(o,d)$ and the shortest time path length $\Delta t(o,d)$ to each origin (o) - destination (d) pair. This time-distance map was drawn as a heat-map in fig. 7.2 for Paris and the other investigated cities, and can be used to identify patterns of privileged connections. We considered distance-bins with equal size 100 meters and time bins of size 1 minute.

In order to focus on the most efficient (privileged) connections with respect to the public transportation system of the city considered, we selected the trips responsible for the 1% lower part of the time distributions for each distance. To estimate, whether these connections are among the best at the urban agglomeration level as compared to travels by car for the same distances, we computed the travel time factor. More precisely, after building the histogram of shortest time paths for every distance bin, we compared the travel time of selected paths with the travel time needed to cover the same distance by car. Car commuting times were extracted from the French 2008 Enquête Nationale Transports et Déplacements 2007-2008 dataset [325] [7] describing the global mobility of people living in France. To collect this data individuals were asked how far (with resolution of 1 km), how long (with resolution of 1 minute), and by which transportation mean they travel every day. Based on this dataset we computed the median of the travel time distributions at each distance using the entire sample to measure the typical time needed to commute to a particular distance by car. Similarly, we calculated the medians of the best $1, 2, 5\%$ of the time distribution at each distance (i.e. shortest times for a given distance) travelled by public transportation. This enables to compute the travel time factor as displayed in fig. 7.3 for different selections of the best times taken by public transportation. By selecting the best connections responsible for the 1% lower part of the time distributions for each distance, in Paris agglomeration, we found that trips' durations are at most 1.71 times the time needed by car. This is in close agreement with the travel time factor tolerated by users [316], which was shown to be maximum 1.6 in [316]. For the other agglomerations studied, the travel time factor goes above this value for distances travelled greater than $5km$. We remark that while in Paris the travel time factor tends to saturate at large distance meaning that efficient connections exist also at the inter-city level, this is not conspicuous for the other cities (see fig. 7.3.b-d), where PT seems to provide an efficient alternative to car mainly for short trips.

Let us notice again that in the histograms and travel time factor calculations we do not use the best absolute time to travel at a given distance but we consider a waiting time assuming that a user arriving blindly at a stop, in order to take into account the preferences of users for path with small variability in time. In addition, the time travelled by car for each distance is taken from a data considering car trips in the whole country. These two points may lead to an overestimation of time travel factors, and this way the travel time factor cannot be used directly as a criterion to select the best connections but only gives a common metric to look at the different public transportation systems.

7 This reference links to a file about the home-work flow of individuals by transportation mean.

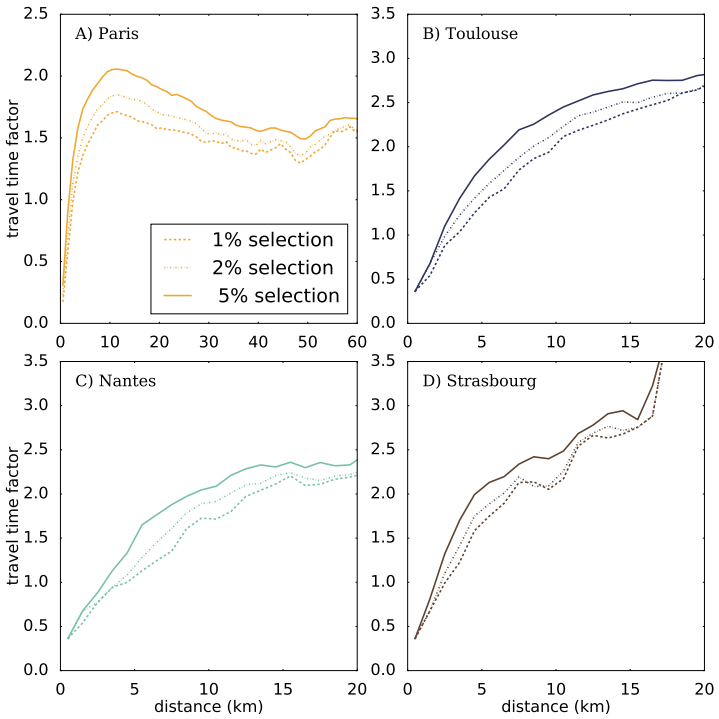

Figure 7.3: **Travel time factors with respect to distance travelled.** The factors have been computed using the 1%, 2% and 5% lower part of the time distribution for each distance travelled by public transportation for the following cities (including their surrounding areas) (A) Paris, (B) Toulouse, (C) Nantes, and (D) Strasbourg.

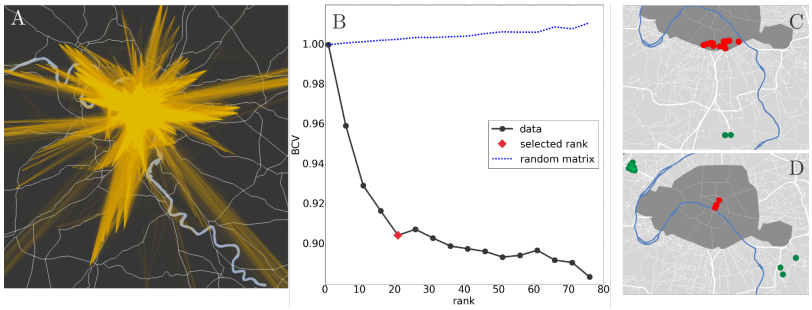

Figure 7.4: **Pattern detection using the multi-edge P-space representation.** (A) Geographic representation of graph G_{SP}, where links correspond to the 1% best shortest paths of the whole public transportation network. (B) The normalised BiCross validation error computed for the adjacency matrix $X_{SP}(10Km, 11Km)$ (black line with markers) of the same graph, for the associated random matrix $X_{SPrandom}(10Km, 11Km)$ (blue dashed line). The selected number of structures k_s is assigned by a red rhombus. (C,D) Two of the structures revealed in the PT system of Paris. Green dots are ingoing, while red dots are outgoing affiliated.

For all the cities considered, privileged connections include shortest paths with no line changes at very short distances, and increasing number of line changes as a function of the distance travelled. In the Paris area, only 2 transportation modalities are used in 80% of the paths up to 60Km distance and the most represented modalities include metro, bus, and rail, with metro dominating at short distances. In smallest cities, almost all paths at distances up to 20 Km involve less than 2 changes. The most represented modalities are bus and tram for Strasbourg and Nantes, bus and metro for Toulouse; instead, rail is present only when the path distance is larger than 15/20 Km.

In fig. 7.4A, we show the profile of the Paris urban agglomeration, where links correspond to the selected privileged connections. The city profile differs clearly from the profile obtained for single-modality single-layer representations since it accounts for the interconnectedness of several transportation modes (see appendix D.7). Note that, since we do not have access to the transfer times between lines in cities other than in Paris, the cost of transferring between layers was estimated for each city based on the data of Paris (see table D.3). This way, transfer times depends on the corresponding transportation modes, what a naive representation with all modes on a single layer would not be able to consider.

7.3.2 *Pattern extraction*

The question remains whether the identified set of privileged connections reveals any higher order meaningful patterns in the design of transportation systems. We expect that some stops, like stations

located in residential neighbourhoods, may have similar connectivity patterns to the rest of the network like to the city centre or to working areas. In order to identify such patterns, we first built an undirected, unweighted graph $G_{SP} = (V_{SP}, E_{SP})$, where $V_{SP} \subset V$ and E_{SP} is a set of edges linking origin-destination locations connected by privileged connections (for an example for Paris see fig. 7.4A). To compare commuters travelling at particular distances we analysed subgraphs $G_{SP}(d_1, d_2)$ (represented by an adjacency matrix $X_{SP}(d_1, d_2)$) of G_{SP}, where edges join stops at particular distances d ($d_1 < d \leq d_2$). For Paris we considered distances with resolution $d_2 - d_1 = 1$ kilometre, while for smaller cities we took the resolution $d_2 - d_1 = 5$ kilometres as the transportation networks were typically sparser there (see fig. D.1).

We expected to find both cohesive, and bipartite patterns in these subgraphs. The cohesive structures would correspond to sets of stations well connected between themselves while bipartite ones would single out two groups of stops with several connections between them. The connections may not be direct but should have durations comparable to the average time taken by car for the same distance.

To detect such patterns we considered the likelihood of having a connection between any two stations, which can be expressed in terms of possible connections of these stations to the same structures. Formally, it means we can express each term of the adjacency matrix representing G_{SP} as

$$X_{SP}(i, j) = \sum_k W_{ik} H_{kj}, \tag{7.3}$$

where W_{ik} quantifies the ingoing membership of node i to structure k and H_{kj} quantifies the outgoing affiliation of the node j to the structure. In order to find matrices $\mathbf{W}$ and $\mathbf{H}$, we performed matrix factorisation, thus minimising numerically the distance

$$\|\mathbf{X} - \mathbf{WH}\|_F^2, \tag{7.4}$$

where $\|\mathbf{X}\|_F$ is the Frobenius norm of matrix $\mathbf{X}$ (for further details see appendix D.1.4). Note that matrix factorisation was used earlier successfully to detect communities and higher order structures in graphs [326–332].

The number of structures to be detected was determined by the Bi-Cross validation (BiCv) approach proposed in [333] based on cross-validation, a common machine learning model validation technique. This consists of measuring an error, called BCV here, between an estimation of left out entries using a low rank approximation of the retained data and the actual left out entries. This error is decreasing with respect to the number of structures extracted toward a minimum that indicates how many structures are representative of the subgraphs, while, on the contrary, such behaviour is note visible

in case the network is close to random. To identify whether there are structures in subgraphs, we compared the BCV error with the one obtained for the corresponding null models (fig. 7.4B). Such null models were defined for each adjacency matrix $X_{SP}(d1, d2)$ as their corresponding random matrices $X_{SPrandom}(d1, d2)$, built by shuffling the values of the original matrix. An example of the behaviour of such a quantity for Paris public transportation network is displayed on fig. 7.4B (for other cities see appendix D.4). This quantity was computed for each subgraph and guided us on how many structures characterise each system at each range of distance. For some distance ranges and cities, the evolution of BCV is close to the random case assigning no strong attempt to link preferentially to some areas at the considered range of distance (see appendix D.4). However, in several cases we find bipartite structures, consisting of disjoint sets of ingoing and outgoing affiliated nodes. In fig. 7.4C and D we show two examples of bipartite structures detected in the Paris network. Given a structure k, green dots corresponds to nodes i that are ingoing affiliated to the structure (e.g. such that $W_{ik}! = 0$), while red dots corresponds to nodes j that are outgoing affiliated (e.g such that $H_{kj}! = 0$). The bipartite structures can be assimilated to strategical areas that are particularly well connected by PT. For example the structure shown in fig. 7.4C, connects stops located around Paris Orly airport to stops located at the border of Paris central area. In fig. 7.4D, the structure reveals the existence of privileged connections the Nanterre and Creteil areas in one side (both with high employment density) [8], with Paris centre on the other side. As these structures are latent patterns extracted from the networks of privileged connections, we consider them as the privileged origin-destination patterns representative of the transportation systems.

7.3.3 *Network efficiency: pattern analysis from the commuter point of view*

To estimate how well the different public transportation networks are devoted to answer the needs of commuters, we compared the identified privileged origin-destination patterns to the flows of commuters. We used the data of the 2010 French census [334] including origin-destination commuter flows per transportation mean at the level of the municipality for the larger areas of Strasbourg, Toulouse, and Nantes, and at the level of the municipal arrondissement (neighbourhood) for the Paris agglomeration. Using this dataset we compared the detected privileged origin-destination patterns to the commuting patterns by car and PT. We only considered inter-municipality trips for the comparison as the resolution provided for the commuter

8 http://insee.fr/fr/themes/document.asp?reg_id=20&ref_id=20718&page=alapage/alap417/alap417_carte.htm#carte1

dataset was given at the municipality level (for the number of intra-city trips see table D.5).

To draw a comparison, we first built the PT structural pattern network $G_C = (V_C, E_C)$ of each urban agglomeration as an unweighted, undirected graph. Here the set of nodes V_C is defined as municipalities and a link $(a, b) \in E_C$ between municipalities a and b exists if at least one stop located in a and one stop in located b appear in each side of a detected bipartite structure. In other words, the structural pattern networks are composed of links between municipalities presumably well connected by public transportation. At the same time, exploiting census data, we built a commuter flow network for each city and its surrounding area, as a weighted, directed graph $G^{TM}_{com} = (V^M_{com}, E^{TM}_{com}, W^M_{com})$. Here V^{TM}_{com} is the set of municipalities, and a link $(a, b) \in E^M_{com}$ with weight w_{ab} represents the flow of individuals commuting from a to b by mean TM (either PT or car). We compared the structural pattern graph with the commuter flow graphs both of the car and the PT of each urban agglomeration by computing a weighted Jaccard index s between the sets of links associated to each graph. This weighted index is defined as the sum of the flow graph weights of the links in common between the two graphs - structural and flow by the selected transportation mean - divided by the total flow for the transportation mean considered. More formally

$$s_{TM} = \frac{\sum_{(a,b)\in E_C \cap E^{TM}_{com}} w_{ab}}{\sum_{(a,b)\in E_C \cup E^{TM}_{com}} w_{ab}} \tag{7.5}$$

both for $TM = car$ and $TM = PT$. It represents the fraction of commuters using respectively car and PT, who have access to privileged PT connections (i.e. for which there exists a link corresponding to their commute in the PT structural pattern).

Bar charts of fig. 7.5 show the comparison between commuting flows and privileged connections for several urban agglomerations. Full bars refer to commuters choosing the car, while dashed bars refer to the choice of PT. The width of full bars are set to be equal for all the cities considered, and the width of the corresponding dashed bar is set proportionally. Hence, for each city the number of black (resp. white) stick men over the total number of stick men corresponds to the fraction of individuals choosing to commute by PT (resp. car). The total number of commuters (using either car or PT) in each city is indicated below each bar. For example, in the Paris central agglomeration (Paris PC, on the right of the figure) there are about 1.6 millions commuters using either car or PT. Among them, for every three commuters choosing the car, about seven choose PT.

The height of the full bars (resp. dashed filled) assigns the weighted Jaccard index s_{car} (resp. s_{PT}), indicating the fraction of individuals choosing the car (resp. PT) even with access to privileged PT connections. For example, in the Paris central agglomeration, among all

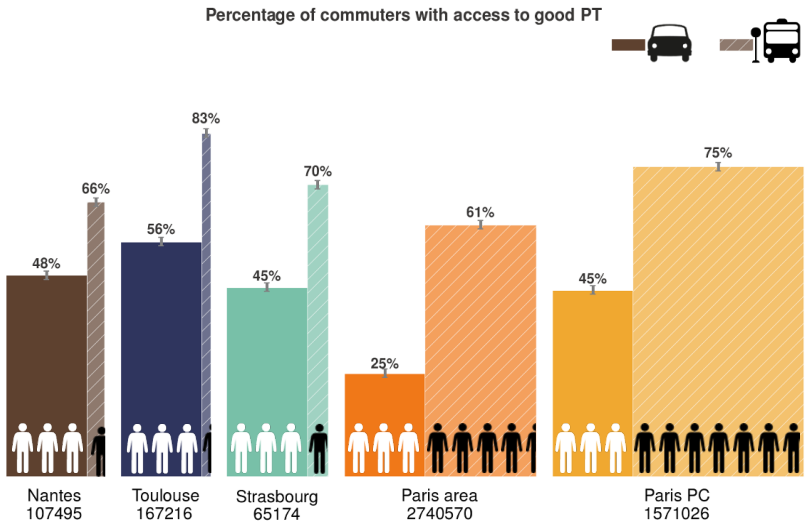

Figure 7.5: **Similarity between commuter flows and PT privileged connections in French municipal areas.** For each of the urban agglomerations considered (Toulouse, Nantes, Strasbourg, Paris area, and Paris Petite Couronne), the bar chart's height indicates the weighted Jaccard index s_{TM} between the commuter flow network G_{com}^{TM} and the PT structural pattern network G_C (for further explanation see text).

commuters choosing the car only 45% would have access to privileged connections, while among those who are choosing PT, 75% can rely on efficient transportation. Error bars on the top of the bars are obtained by repeating the methodology 100 times, with different random matrices initialising NMF. The small size of the error bars shows the robustness of the pattern detection.

A significant difference between the commuting practice in Paris agglomeration and other urban areas is evident. For Paris urban agglomeration, the flow of inter-municipality commuters choosing PT is larger than that of people commuting by car, in contrast to the other investigated cities. This may be partly explained by a travel time factor, which increases above the tolerated value for Toulouse, Nantes and Strasbourg (see fig. 7.3). Besides, fig. 7.5 indicates that among commuters choosing public transportation, a large fraction has access to privileged connections (Paris: 75%, Strasbourg: 70%, Toulouse: 83%, Nantes: 66%). Among commuters choosing the car, typically less than 50% have access to privileged connections (Paris: 45%, Strasbourg: 45%, Toulouse 56%, Nantes 48%). This comforts our definition of privileged connections based on commuting time with little variability and a limited transfer number. This corroborates the strong role of the latter factors in the decision making to use PT or car. Furthermore, we observe that in the larger Paris area only 25% of car commuters have access to privileged transportation connection. Instead, in other cities, although more than 48% of car drivers have

access to rapid connections, they still commute by car. In particular, in Toulouse a large percentage of commuters have access to good services according to the criteria introduced here, as there is large overlap between privileged connections and both PT (83%) and car (56%) commuting flows. However, there is still a non-negligible amount of people commuting by car. Based on this analysis we can distinguish between two main trends in commuting: (a) there are cities where a large part of the population tend to do inter-municipality trips by car disregarding the quality of PT services, examples are Nantes, Toulouse, and Strasbourg. (b) On the other hand, in Paris and its agglomeration, according to the metrics introduced, there is a good agreement between the needed and provided services of public transportations. This result is supported by a pairwise comparison between the car and the PT commuting flows for every pair of municipalities (see appendix D.5).

7.4 SUMMARY

Efficient analysis of public transportation networks is possible via abstract representations, which in turn help us to reveal hidden characteristics of such systems. As our main scientific contribution we provided a solution for this challenge by introducing a novel description, which combines multi-edge and P-space representations of multilayer transportation networks. We characterise these systems from the user's point of view through a description, which is detached from constraints imposed by their spatial embeddedness, but which incorporates their temporal variance. To further develop our framework we adjusted earlier defined methods and used them to identify effective routes and hidden transportation patterns, which were not evidently built due to overall design. We found cohesive and bipartite patterns of privileged connections induced by different ways of access of far-apart urban areas in French municipals such as larger Paris, Toulouse, Nantes, or Strasbourg. We further analysed the overall efficiency of the corresponding transportation systems as compared to the commuting flow. We found that while the transportation system of Paris is somewhat meeting overall demands and preferred to be used over the car alternative, in smaller cities the transportation systems may not meet the user expectations, leaving room for improvement, and even people have access to fast transportation options they prefer to use car instead.

We made some assumptions during our study, which set some limitations on the generalisation of our results. First of all we considered only a 3 hours time window to build our user-based representation. Extending this time window or considering different periods would potentially highlight further transportation patterns, assigning a direction to explore in the future. Further, we operated with average

frequencies of services neglecting the effect of any perturbation in the transportation system. This was a valid approach in our case as no major variance was observed during the analysed period. Nevertheless, to come around this limitation one can easily adjust or definition such that it considers dynamically unexpected perturbations on each line. Finally we assumed that passengers go blindly to a stop without considering that certain passengers attempt to reduce their waiting time by timing their arrival at transit stops to an optimal period before vehicle departure. On the other hand this can be easily considered in our representation by introducing arrival times of users e.g. depending on the frequency of the first line they take. In addition, note that aspects as adaptive travelling behaviour or the prediction of individual mobility patterns are out of the scope of the present methodology but they indicate possible future directions of research.

Several extensions of our methodology is possible. Parameters like the periods in focus, length of observations, number of transfers, etc. can be tailored for other systems, while a further refinement is possible by considering needs of various types of users. Our way of characterisation of privileged connections may be used to profile and compare different transportation systems to disclose generalities in their design. Use of this methodology in the future could help to enhance resilience of local transportation systems to provide better design policies for future developments.

8

CONCLUSIONS

This thesis was stimulated by the recent developments in the interdisciplinary area that studies human mobility [38]. Compared with previously existing approaches, our contribution has been twofold. First, we have studied individual behaviour across a broad range of temporal and spatial scales. Secondly, we have studied mobility in relation with other aspects of behaviour, establishing a connection between human geography, personality psychology, and the social sciences. Our advancements relied on the analysis of multi-channel data with unprecedented spatial, temporal resolution and duration. In chapter 2, we have proposed novel pre-processing techniques to extract trajectories from these comprehensive datasets which include WiFi, GPS and call detail records collected from mobile phones.

Our research has addressed three questions that had been raising growing interest within the scientific community [38].

What are the statistical properties of human mobility?

In chapter 3, we have tackled the unresolved controversy about the statistical distributions characterising human motion. We have studied the statistical properties of data with the best combination of resolution, duration, spatial range and sample size among those considered in the literature so far. We have shown that the distribution of displacements and waiting times between displacements are best described by log-normal distributions, but power-law distributions are selected when only large spatial or temporal scales are selected. While identifying the mechanism responsible for the observed mobility patterns was beyond the scope of the present work, we anticipate that a more complete spatio-temporal description of human mobility will help us develop better models of human mobility behaviour [52, 117].

How do humans allocate time among different locations?

In chapter 4, we have solved the tension between the state-of-the-art understanding of human mobility as highly predictable and stable over time, and the fact that individual lives are constantly evolving due to changing needs and circumstances. We have shown that this tension vanishes when the long-term evolution of human visitation patterns is considered. We have found that routines are unstable in

the long term because of the continual exploration of new locations, but the number of locations an individual visits regularly is conserved over time. We have shown that this individual 'location capacity' is peaked around a typical value of ~ 25 locations and correlates with individuals' social circles size. In this respect, it is interesting to note that fixed-size effects in the social domain [131, 132, 149, 150] have been put in direct relation with human cognitive abilities [149]. We anticipate that our results will stimulate new research exploring this connection.

What are the connections between individuals' mobility and social behaviour?

In chapter 5, we have explored, for the first time to our knowledge, the connection between individual spatial and social behaviour. We have shown that there is a connection between the way in which individuals explore new resources and exploit known assets in the social and spatial spheres. We have pointed out that different individuals balance the exploration-exploitation trade-off in different ways and we have explained part of the variability in the data by the big five personality traits. These findings establish a relation between personality and spatial behaviour, validating the theories suggesting that spatial choices are partially dictated by personality dispositions [214] and that a single set of personality traits underlies all aspect of a person's behaviour [175, 215].

In chapter 6, taking further the study of social behaviour, we have presented a model of interactions in which individuals are characterised by stable dispositions towards establishing and receiving connections. We have demonstrated that these individual characteristics have a major impact on diffusion processes spreading on the network. These results contribute towards the development of a comprehensive picture about how the dynamics *of* networks affect the dynamics unfolding *upon* networks.

In the second part of this thesis we have addressed the challenge of quantifying the efficiency of public transportation system. We have introduced a novel multilayer description of public transportation networks. We have adjusted earlier defined methods to identify effective routes and hidden transportation patterns, accounting for the fact that people prefer short trips with few line changes. We have found privileged connections in French municipals and analysed the efficiency of their transportation systems as compared to the commuting flows. We have found that, while the transportation system of a large city like Paris is meeting overall demands, and preferred to be used over the car, in smaller cities the transportation systems do not meet the user expectations, and even people have access to fast transportation, they prefer to use car. The use of the proposed methodology

could help providing solutions to urban and transportation design.

Present and future work will move in several directions. We list here briefly the topics we are currently addressing, or we plan to investigate in the near future.

- **Effects of location semantics**. In chapter 4 and chapter 5, we have characterized the evolution of individuals' sets of familiar locations (also referred to as 'Activity spaces'). We are now investigating the composition of individuals' activity spaces in term of different types of locations (e.g. commercial activities, leisure locations, homes, ...). This project relies on matching individuals' stop-locations with enriched semantic information collected from open-source data [1].

- **Co-evolution of individuals' visitation patterns.** In chapter 4, we have shown that individuals' visitation patterns evolve over long time scales. Our ongoing research focuses on studying the co-evolution of the visitation patterns of pairs individuals. Our hypothesis is that long-term developments in the social and spatial sphere are correlated. Furthermore, we are investigating the EPR model with memory, to understand the role played by the parameters on the specific size of the spatial capacity.

- **Spatial stochastic block model for communication data.** In chapter 7, we have shown how non-negative matrix factorization can be applied to extract meaningful structures from networks. One of my ongoing researches, in collaboration with Prof. Emilio Ferrara and Prof. Aram Galstyan focuses on a similar technique, the stochastic block modelling. Our work extends the current framework to account for the role played by the physical distance between the network nodes. We are applying this method to study telecommunication flows in urban areas.

- **Modelling the cryptocurrency market.** Alongside the interest for human mobility, we have developed a new research direction. Our work aims at characterizing and modelling the short and long-term dynamics of the cryptocurrency market. Our initial results are included in [VI].

1 https://www.openstreetmap.org/

Part I

APPENDIX

A

APPENDIX TO CHAPTER 1

In this appendix we present additional findings supporting the results presented in chapter 3. In appendix A.1, we show our findings are robust with respect to variation of the definition of stop-locations and across the sample considered. In appendix A.1, we present the list of distributions considered for our analysis.

A.1 ROBUSTNESS OF RESULTS

Results of the model selection

The selection of the log-normal distribution as the best model among the exponential, the log-normal and the Pareto distribution is made using the Akaike Information Criterion (AIC) weights. In tables A.1 to A.3 we report the AIC weights values for the four models considered as well as the Akaike information Criterion (AIC), the Bayesian Information Criterion (BIC) weights, the Residual Sum of Squares (RSS). These metrics provide additional information on the goodness-of-fit. In figs. A.1 to A.3, we show the results of the fit with the three distributions considered.

Bootstrapping

By bootstrapping data 1000 times for samples of 100 and 200 individuals, we find that for all groups the aggregated distributions of displacements and waiting times are best described by the same models

	AIC	AIC weights	BIC weights	RSS
expon	2.1e+07	0	0	3.1e-11
lognorm	1.9e+07	1	1	2.9e-11
pareto	2.0e+07	0	0	2.8e-11

Table A.1: **Distribution of displacements: model selection**. For the three distributions considered, the table reports the Akaike Information Criterion (AIC), the AIC weights (see Model selection section), the Bayesian Information Criterion (BIC) and the residual sum of squares (RSS).

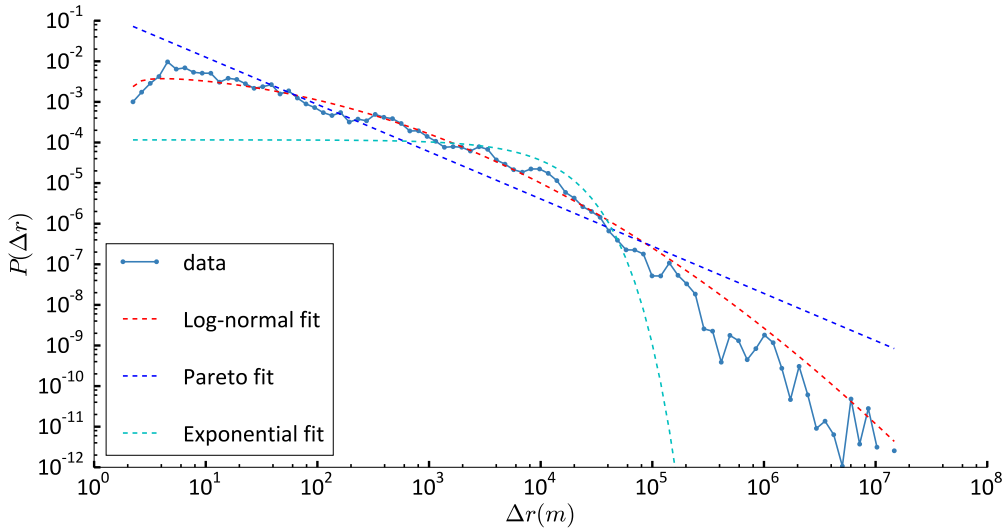

Figure A.1: **Distribution of displacements: comparison of three models**. Blue dotted line: data. Red dashed line: Maximum likelihood Log-normal fit. Blue dashed line: Maximum likelihood Pareto fit. Light blue dashed line: Maximum likelihood Exponential fit.

	AIC	AIC weights	BIC weights	RSS
expon	4.62e+06	0	0	0.061
lognorm	3.68e+06	1	1	0.026
pareto	3.79e+06	0	0	0.025

Table A.2: **Distribution of waiting times: model selection**. For the three distributions considered, the table reports the Akaike Information Criterion (AIC), the AIC weights (see Model selection section), the Bayesian Information Criterion (BIC) and the residual sum of squares (RSS).

	AIC	AIC weights	BIC weights	RSS
lognorm	2.7e+07	1	1	3.0e-11
pareto	2.9e+07	0	0	2.8e-11
expon	3.0e+07	0	0	3.1e-11

Table A.3: **Distribution of displacements between discoveries: model selection**. For the three distributions considered, the table reports the Akaike Information Criterion (AIC), the AIC weights (see Model selection section), the Bayesian Information Criterion (BIC) and the residual sum of squares (RSS).

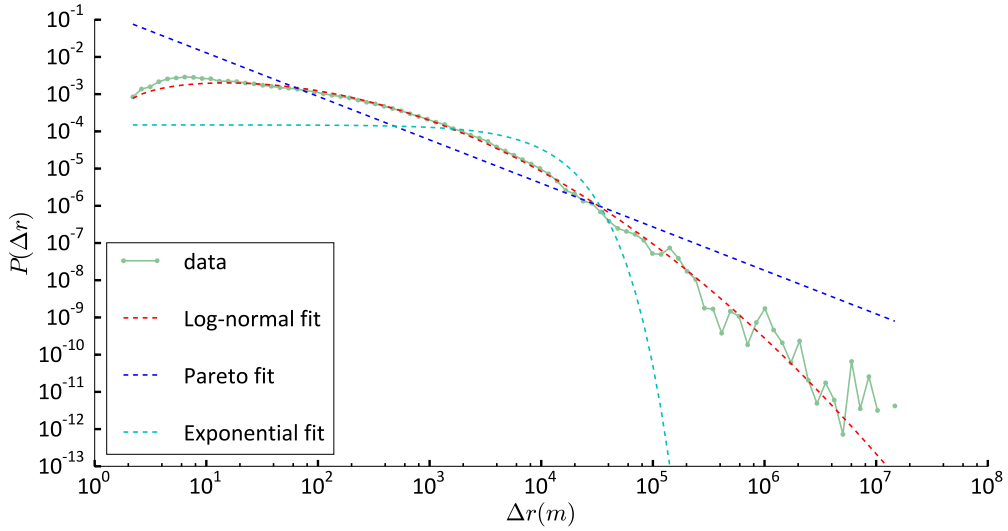

Figure A.2: **Distribution of waiting times: comparison of three models**. Yellow dotted line: data. Red dashed line: Maximum likelihood Log-normal fit. Blue dashed line: Maximum likelihood Pareto fit. Light blue dashed line: Maximum likelihood Exponential fit.

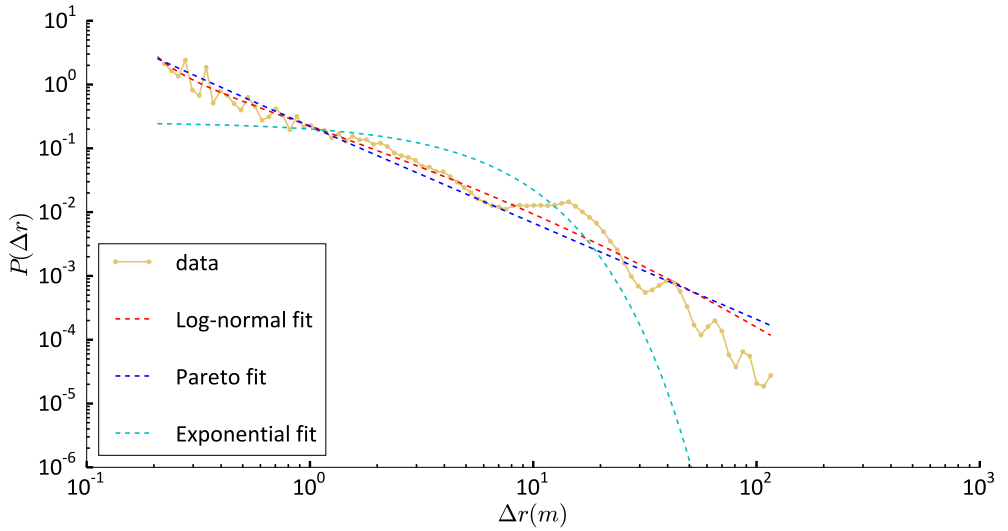

Figure A.3: **Distribution of displacements between discoveries: comparison of three models**. Green dotted line: data. Red dashed line: Maximum likelihood Log-normal fit. Blue dashed line: Maximum likelihood Pareto fit. Light blue dashed line: Maximum likelihood Exponential fit.

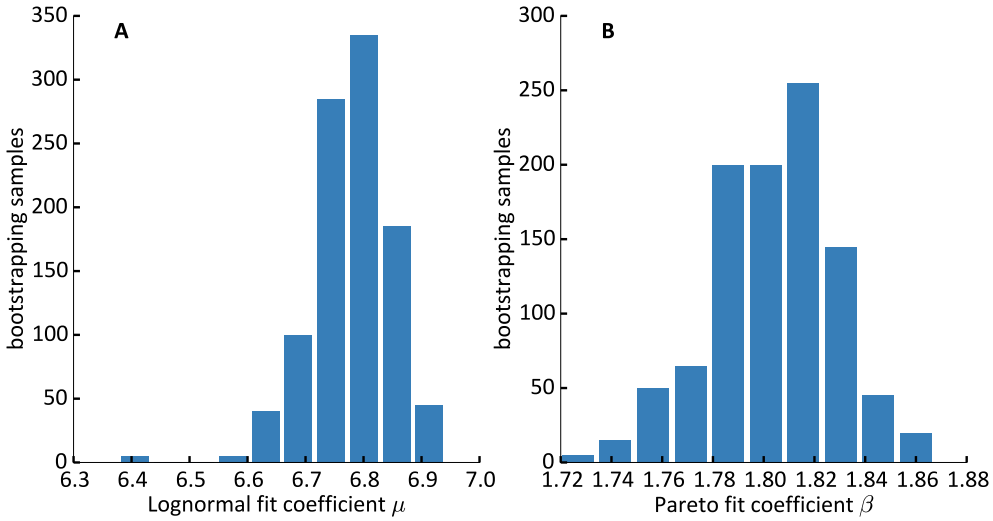

Figure A.4: **Displacements: distribution of parameters found by bootstrapping**. **A)**The distribution over 1000 bootstrapping samples of the log-normal fit coefficient μ, characterising the aggregated distribution of displacements. **B)**The distribution over 1000 bootstrapping samples of the Pareto fit coefficient β, characterising the tail of the aggregated distribution of displacements. Samples include 100 randomly selected individuals.

found for the entire dataset.
Here, we report the distribution of parameters found for the distribution of displacements (fig. A.4), waiting times (fig. A.5), and displacements between discoveries (fig. A.6), in the case of samples of 100 individuals.

Sensitivity to the definition of pausing

The distribution of displacements is robust with respect to the definition of *pausing*. The results reported in the main text refer to pauses longer than $P = 10$ minutes. Both for $P = 15$ minutes and $P = 20$ minutes, the distribution of displacements is best described by a log-normal model when the entire distribution is taken into account, and by a Pareto distribution, when only long distances are considered (see figs. A.7 and A.8). The same results hold for the distributions of waiting times (see figs. A.9 and A.10)

Interpretation of the shift and scale parameters

The shift and scale parameters are necessary to account for the fact that, in the cases considered, the lower bound of the distributions support is controlled by the data minimal resolution.

For example, the log-normal distribution of a random variable x is defined for $x \in (0, \infty)$. In our case the fit is performed for a shifted

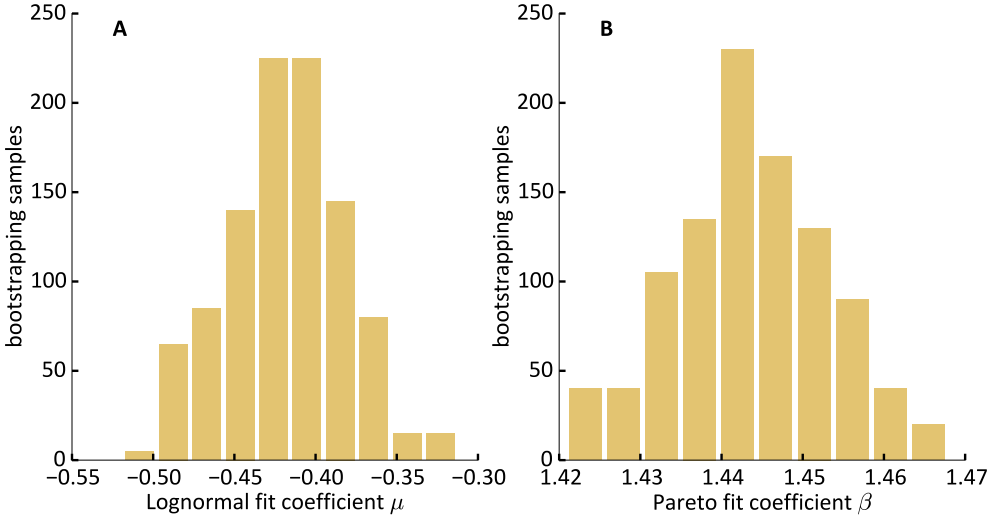

Figure A.5: **Waiting Times: distribution of parameters found by bootstrapping**. **A)**The distribution over 1000 bootstrapping samples of the log-normal fit coefficient μ, characterising the aggregated distribution of waiting times. **B)**The distribution over 1000 bootstrapping samples of the Pareto fit coefficient β, characterising the tail of the aggregated distribution of waiting times. Samples include 100 randomly selected individuals.

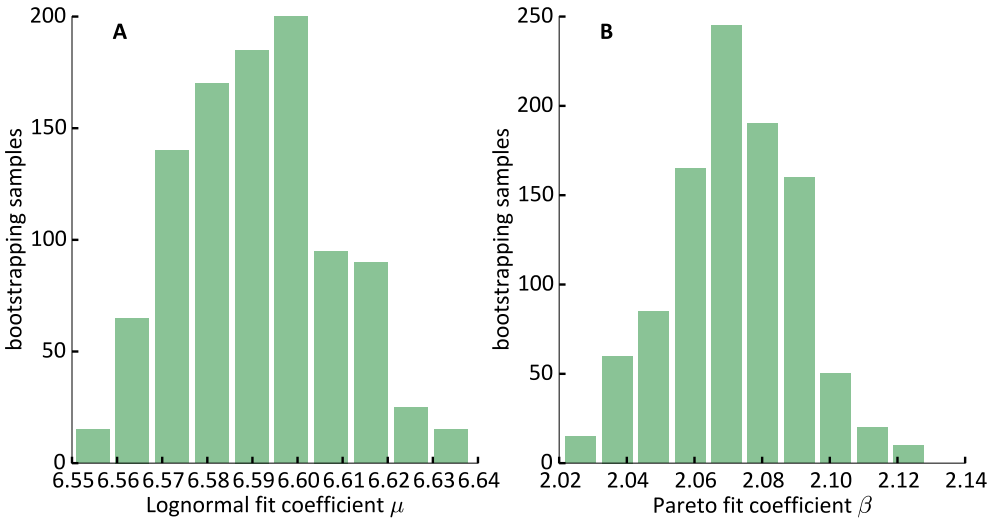

Figure A.6: **Displacements between discoveries: distribution of parameters found by bootstrapping**. **A)**The distribution over 1000 bootstrapping samples of the log-normal fit coefficient μ, characterising the aggregated distribution of displacements between discoveries. **B)**The distribution over 1000 bootstrapping samples of the Pareto fit coefficient β, characterising the tail of the aggregated distribution of displacements between discoveries. Samples include 100 randomly selected individuals.

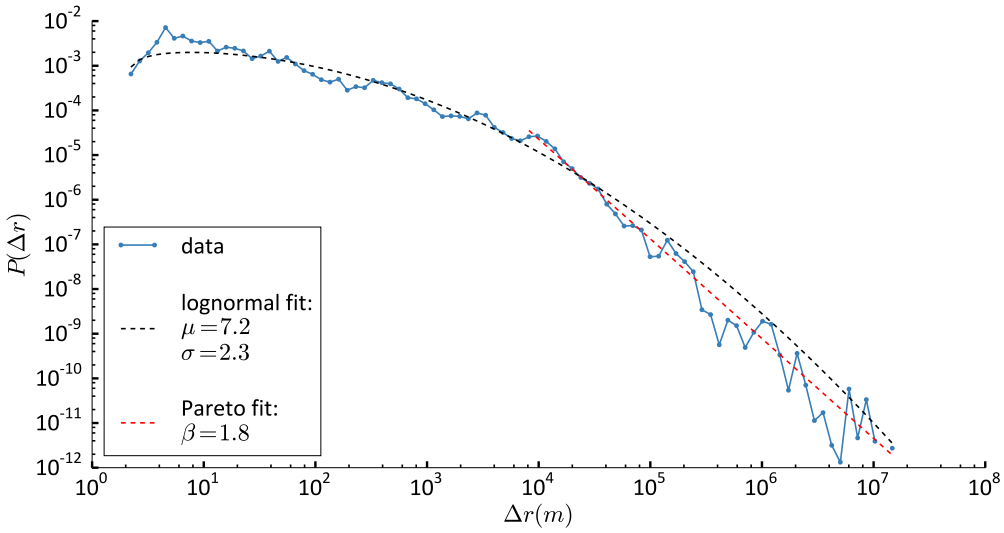

Figure A.7: **Distribution of displacements for pausing P=15 minutes**. Blue dotted line: data. Black dashed line: Log-normal fit with characteristic parameter μ and σ. Red dashed line: Pareto fit with characteristic parameter β for $\Delta r > 7420$ m.

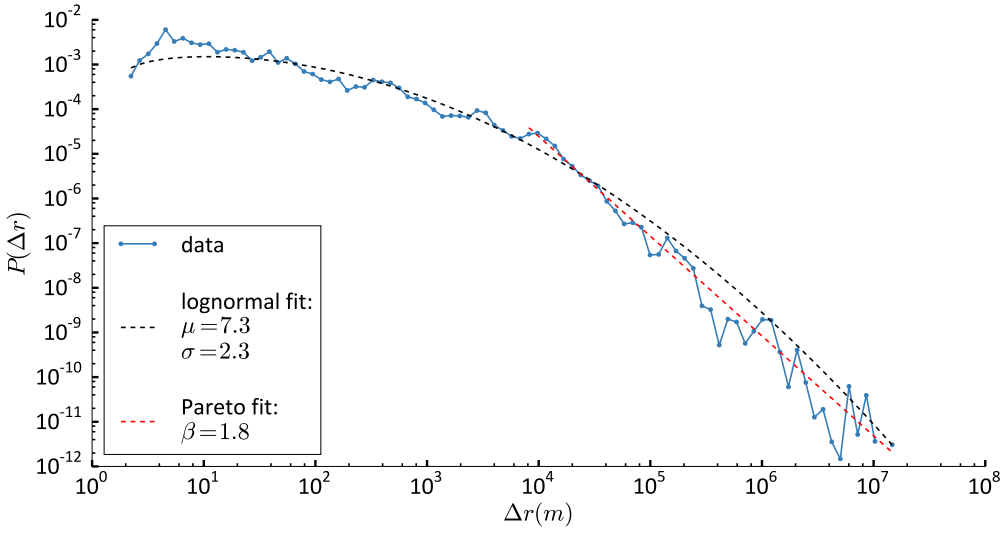

Figure A.8: **Distribution of displacements for pausing P=20 minutes**. Blue dotted line: data. Black dashed line: Log-normal fit with characteristic parameter μ and σ. Red dashed line: Pareto fit with characteristic parameter β for $\Delta r > 7420$ m.

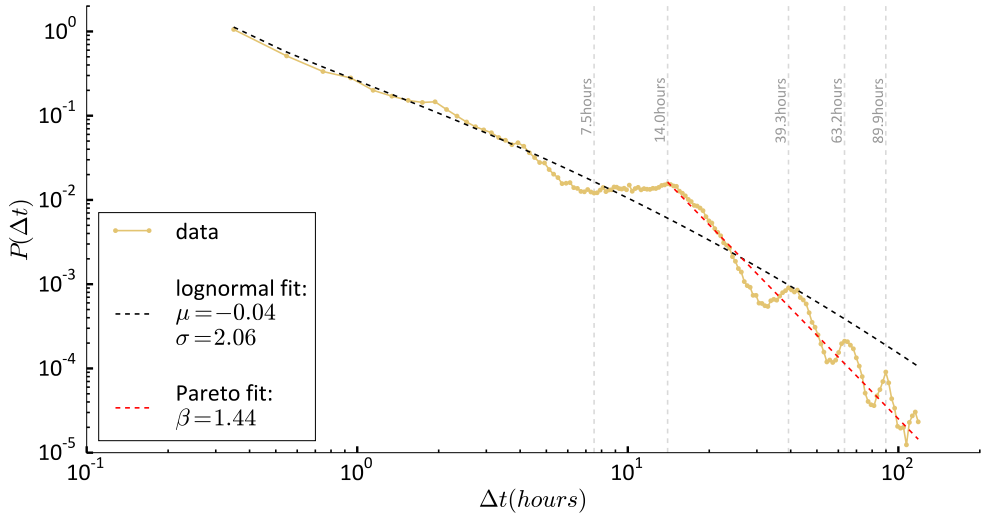

Figure A.9: **Distribution of waiting times for pausing P=15 minutes**. Yellow dotted line: data. Black dashed line: Log-normal fit with characteristic parameter μ and σ. Red dashed line: Pareto fit with characteristic parameter β for $\Delta t > 13h$.

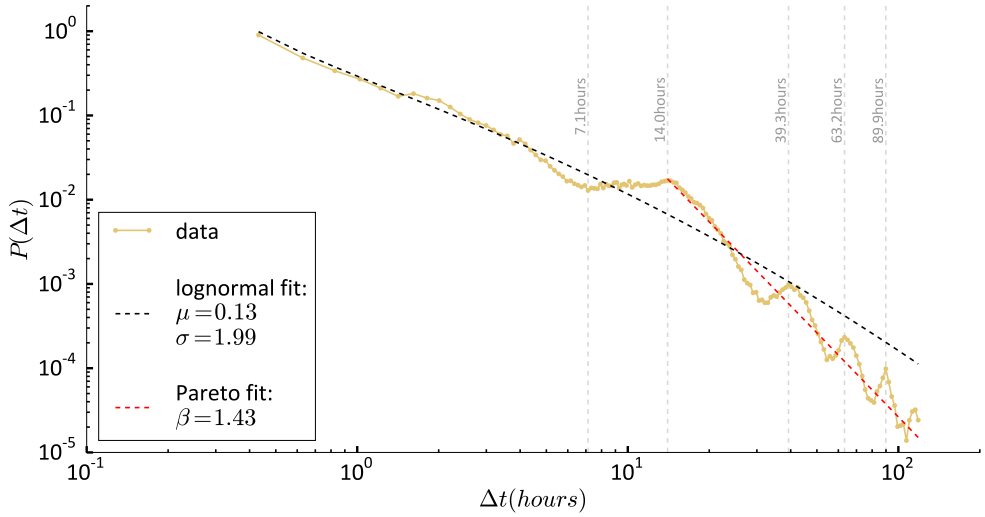

Figure A.10: **Distribution of displacements for pausing P=20 minutes**. Yellow dotted line: data. Black dashed line: Log-normal fit with characteristic parameter μ and σ. Red dashed line: Pareto Fit with characteristic parameter β for $\Delta t > 13$ h.

	Shift (Lognormal)	Shift (Pareto)	Scale (Pareto)
Displacements	2.02 m	-11.41 m	7431.83 m
Waiting times	0.18 h	-0.03 h	13.03 h
Discoveries	1.9 m	-1.34 m	2801.35 m

Table A.4: **The scale and shift parameters.** The values of the shift parameter of the Lognormal fit (first column), the shift and scale parameter of the Pareto fit of the distributions' tails (second and third columns).

distribution, with $x \in (x_0, \infty)$, where x_0 is the data minimal resolution. This reflects the fact that the reason why there are no data points for $x < x_0$ is not low probability but lack of information within this range (or in some cases it's due to the choice of fitting only the tail of the distribution).

Similarly, the Pareto distribution is defined for $x \in (1, \infty)$. The shift x_0 and the scale parameter s allow instead to consider $x \in (s + x_0, \infty)$, where $s + x_0$ is the minimum data point considered. Values of the shift and scale parameters could be set to fit the minimal resolution. However, in our case x_0 and s are additional parameters of the model. We have verified that the values recovered by the fitting algorithm are consistent with those expected.

We report in table A.4 the values of the shift s and scale parameters x_0. The results presented in the main text do not change when we set $x_0 = 0$ except for the distribution of waiting times, where we find Pareto as the best distribution if $x_0 = 0$.

Further analysis: Selection of the best model among 68 distributions

In the case of the distribution of waiting times, the best model among 68 distributions is the gamma distribution. Results of the gamma fit are shown in fig. A.11.

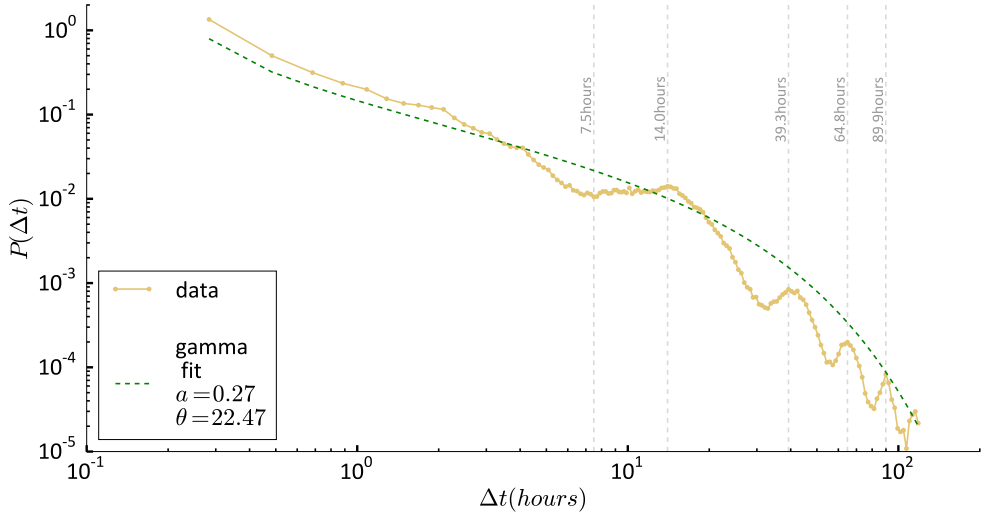

Figure A.11: **Distribution of waiting times: selection of the best model among 68 distributions**. Yellow dotted line: data. Green dashed line: Gamma Distribution fit with characteristic parameters $a = 0.27$ and $\theta = 22.47$

DISTRIBUTIONS CONSIDERED

The list of distributions is based on the scipy.stats Python [335] module which contains the implementation of over 80 probability distributions, including those reported in the literature on human mobility. We have excluded distributions with more than 3 parameters (including scale and shift), unless they were found in previous studies on human mobility. The distribution considered are the following:

Levy alpha-stable, Anglit, arcsine, Bradford, Cauchy, chi, chi-squared, cosine, double gamma, double Weibull, exponential, exponential power, fatigue-life, Fisk, folded Cauchy, folded normal, Frechet left, Frechet right, gamma, generalized extreme value, generalized Gamma, generalized half-logistic, generalized logistic, Generalized Pareto, Gilbrat, Gompertz, left-skewed Gumbel, right-skewed Gumbel, half-Cauchy, half-logistic, half-normal, hyperbolic secant, inverted gamma, inverse Gaussian, inverted Weibull, General Kolmogorov-Smirnov, Laplace, Levy, left-skewed Levy, log gamma, logistic, log-Laplace, lognormal, Lomax, Maxwell, Nakagami, normal, Pareto, Pearson type III, power-function, power log-normal , power normal, Rayleigh, R, Reciprocal inverse Gauss, Rice, semicircular, Student's T, triangular, truncated exponential, truncated normal, Tukey-Lambda, Truncated Pareto, Uniform, Von Mises, Wald, Weibull maximum, Weibull minimum, wrapped Cauchy

B

APPENDIX TO CHAPTER 2

This chapter is an appendix to chapter 4. In appendix B.1, we present tests assessing that the results presented in chapter 4 are robust with respect to choices such as the definition of familiar location. In appendix B.2, we discuss the evolution of spatial properties of the set of familiar locations.

B.1 ROBUSTNESS TESTS

The results presented in chapter 4 do not depend on how locations are defined, nor on the time-window used to investigate the long-term behaviour. In this section, we show how the results are derived and we demonstrate their statistical robustness. To avoid confusion, we will indicate with $\overline{x}$ the average value of a quantity x across the population, and $\langle x \rangle$ the average across time.

Conservation of the location capacity

The set of familiar locations is defined here as the set $S_i(t) = \{\ell_1, \ell_2, ..., \ell_k, ...\ell_C\}$ of locations ℓ_k that individual i visited at least twice and where she spent on average more than 10 minutes/week during a time-window of W consecutive weeks preceding time t. In fig. B.1, we show that for $W = 10$ weeks, the set contains on average a small fraction of all locations seen during the same 10 weeks. Yet, the time spent in these locations is on average close to the total time (fig. B.1). Given this definition, the number of locations an individual i visits regularly is equivalent to the set size $C_i(t) = |S_i(t)|$. We call this quantity *location capacity*.

Evidence 1 The average individual capacity $\overline{C}$ is constant in time regardless of the definition of location or the choice of the window size W (table B.1 and table B.2). This result is tested in several ways:

LINEAR FIT TEST We perform a linear fit of the form $\overline{C(t)} = a + b \cdot t$, computed with the least squares method. We test the hypothesis H_0 : $b = 0$, under independent 2-samples t-tests.

POWER LAW FIT TEST We perform a power-law fit of the form $\overline{C(t)} \propto t^{\beta}$, computed with the least squares method. We test the hypothesis $H_1 : \beta = 0$, under independent 2-samples t-tests.

MULTIPLE INTERVALS TEST We compare the value of $\overline{C}$ across different time-intervals δt_k. We divide the total time range into time-intervals δt_k spanning w weeks. We compute the average capacity

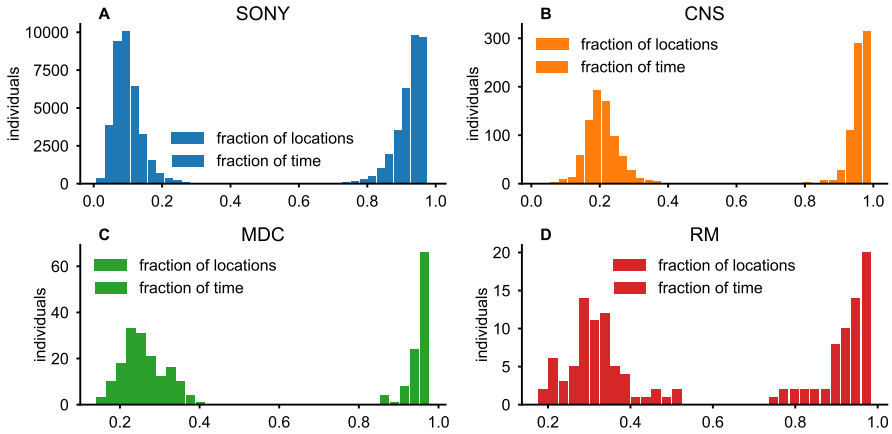

Figure B.1: **Establishment of the set of familiar locations.** Frequency histograms of individuals based on the fraction of all locations seen in a week that are part of the set of familiar locations (dashed bars), and on the fraction of time of the week spent in familiar locations (full bars). The set is computed for $W = 10$ weeks. Results are shown for the Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)**

$\overline{C(\delta t_k)}$ and its standard deviation $\sigma_C(\delta t_k)$ for each time-interval δt_k. We test the hypotheses $H_{j,k} : \overline{C(\delta t_j)} = \overline{C(\delta t_k)}$ for all pairs $\delta t_k, \delta t_j$.

For all the datasets considered, all choices of W, and definitions of locations the hypotheses H_0, H_1 and $H_{j,k}$ (for all intervals j and k) can not be rejected at $\alpha = 0.05$ with p-value$> \alpha$ under 2-tailed tests. Results are reported in table B.1.

Evidence 2 The individual weekly *net gain* of locations is equal to zero. The *net gain* defined as $G_i(t) = A_i(t) - D_i(t)$, where $A_i(t) = |S_i(t) \setminus S_i(t - dt)|$ is the number of location added and $D_i(t) = |S_i(t - dt) \setminus S_i(t)|$ (the difference between the sets) is the number of location removed from the set during dt, where $dt = 1$ week. This is verified by testing for all individuals i if the ratio $\sigma_{G,i}/\langle G_i \rangle > 1$, where $\sigma_{G,i}$ is the standard deviation of the average individual net gain across time (see chapter 4). We find that $\sigma_{G,i}/\langle G_i \rangle > 1$ hold for a large majority of individuals, under different definitions of locations and choices of W, for all datasets considered. Results are reported in table B.2 and fig. B.2.

Evidence 3 The average value of location capacity saturates for increasing values of the time-window W. We find that for all datasets the average time coverage $\overline{\langle C \rangle} \sim 25$. This result is obtained after accounting for the differences in data collection by considering the normalized location capacity C_i/TC_i, where TC_i is the weekly time coverage of individual i (see figs. B.3 and B.4). Individuals' capacity values are distributed homogeneously around the mean (fig. B.5).

data	d (m)	W	H_0		H_1		$H_{j,k}$
			b	p	β	p	(rej.)
Lifelog							
(sel.user)	50	10	$(-0.27 \pm 3.04) \cdot 10^{-3}$	0.94	$(-0.01 \pm 3.45) \cdot 10^{-2}$	1.00	0%
Lifelog	30	10	$(-1.47 \pm 3.24) \cdot 10^{-3}$	0.73	$(-0.07 \pm 2.25) \cdot 10^{-2}$	0.98	0%
Lifelog	40	10	$(-0.52 \pm 3.28) \cdot 10^{-3}$	0.90	$(-0.03 \pm 2.30) \cdot 10^{-2}$	0.99	0%
Lifelog	50	4	$(-0.67 \pm 2.66) \cdot 10^{-3}$	0.84	$(-0.03 \pm 1.68) \cdot 10^{-2}$	0.99	0%
Lifelog	50	6	$(-0.40 \pm 2.85) \cdot 10^{-3}$	0.91	$(-0.021 \pm 1.89) \cdot 10^{-2}$	0.99	0%
Lifelog	50	8	$(-0.18 \pm 3.02) \cdot 10^{-3}$	0.96	$(-0.01 \pm 2.08) \cdot 10^{-2}$	1.00	0%
Lifelog	50	10	$(-0.15 \pm 3.11) \cdot 10^{-3}$	0.97	$(-0.09 \pm 2.20) \cdot 10^{-2}$	1.00	0%
Lifelog	50	12	$0.26 \pm 3.33) \cdot 10^{-3}$	0.95	$(0.08 \pm 2.42) \cdot 10^{-2}$	1.00	0%
Lifelog	50	40	$(2.59 \pm 7.05) \cdot 10^{-3}$	0.78	$(0.11 \pm 5.77) \cdot 10^{-2}$	0.99	0%
Lifelog	50	20	$(1.27 \pm 4.00) \cdot 10^{-3}$	0.80	$(0.05 \pm 3.09) \cdot 10^{-2}$	0.99	0%
CNS	2	10	$(-3.74 \pm 3.42) \cdot 10^{-3}$	0.47	$(-0.15 \pm 4.21) \cdot 10^{-2}$	0.98	0%
CNS	5	4	$(-2.06 \pm 3.66) \cdot 10^{-3}$	0.67	$(-0.10 \pm 3.39) \cdot 10^{-2}$	0.98	0%
CNS	5	6	$(-1.81 \pm 3.57) \cdot 10^{-3}$	0.70	$(-0.08 \pm 3.71) \cdot 10^{-2}$	0.99	0%
CNS	5	8	$(-2.92 \pm 3.50) \cdot 10^{-3}$	0.56	$(-0.12 \pm 3.94) \cdot 10^{-2}$	0.98	0%
CNS	5	10	$(-3.84 \pm 3.43) \cdot 10^{-3}$	0.46	$(-0.15 \pm 4.10) \cdot 10^{-2}$	0.98	0%
CNS	5	12	$(-4.09 \pm 3.33) \cdot 10^{-3}$	0.43	$(-0.17 \pm 4.18) \cdot 10^{-2}$	0.97	0%
CNS	5	40	$(-1.77 \pm 8.92) \cdot 10^{-3}$	0.87	$(-0.01 \pm 1.41) \cdot 10^{-1}$	1.00	0%
CNS	5	50	$(-0.28 \pm 1.78) \cdot 10^{-2}$	0.90	$(-0.12 \pm 2.86) \cdot 10^{-1}$	1.00	0%
CNS	10	10	$(-3.39 \pm 3.39) \cdot 10^{-3}$	0.50	$(-0.14 \pm 4.04) \cdot 10^{-2}$	0.98	0%
MDC		4	$(-1.08 \pm 2.70) \cdot 10^{-3}$	0.76	$(-0.05 \pm 2.74) \cdot 10^{-2}$	0.99	0%
MDC		6	$(-0.95 \pm 2.75) \cdot 10^{-3}$	0.79	$(-0.05 \pm 3.11) \cdot 10^{-2}$	0.99	0%
MDC		8	$(-0.72 \pm 2.82) \cdot 10^{-3}$	0.84	$(-0.03 \pm 3.41) \cdot 10^{-2}$	0.99	0%
MDC		10	$(-0.59 \pm 2.88) \cdot 10^{-3}$	0.87	$-0.30 \pm 3.64) \cdot 10^{-2}$	0.99	0%
MDC		12	$(-0.45 \pm 2.95) \cdot 10^{-3}$	0.90	$(-0.02 \pm 3.83) \cdot 10^{-2}$	1.00	0%
MDC		40	$(1.74 \pm 5.13) \cdot 10^{-3}$	0.79	$(0.07 \pm 7.85) \cdot 10^{-2}$	0.99	0%
MDC		50	$(3.77 \pm 7.52) \cdot 10^{-3}$	0.70	$(0.02 \pm 1.19) \cdot 10^{-1}$	0.99	0%
MDC		20	$(-0.28 \pm 3.21) \cdot 10^{-3}$	0.94	$(-0.01 \pm 4.51) \cdot 10^{-2}$	1.00	0%
RM		4	$(4.73 \pm 7.05) \cdot 10^{-3}$	0.62	$(0.11 \pm 7.76) \cdot 10^{-2}$	0.99	0%
RM		6	$(3.77 \pm 8.47) \cdot 10^{-3}$	0.73	$(0.01 \pm 1.08) \cdot 10^{-1}$	0.99	0%
RM		8	$(4.31 \pm 8.87) \cdot 10^{-3}$	0.71	$(0.01 \pm 1.23) \cdot 10^{-1}$	1.00	0%
RM		10	$(2.16 \pm 9.46) \cdot 10^{-3}$	0.86	$(0.01 \pm 1.38) \cdot 10^{-1}$	1.00	0%
RM		12	$(-0.03 \pm 1.05) \cdot 10^{-2}$	0.98	$(-0.01 \pm 1.60) \cdot 10^{-1}$	1.00	0%
RM		20	$(0.44 \pm 1.89) \cdot 10^{-2}$	0.85	$(0.01 \pm 3.19) \cdot 10^{-1}$	1.00	0%

Table B.1: **Conservation of capacity: evidence 1**. The results of hypotheses testing H_0, H_1 and $H_{j,k}$ (see appendix B.1) for different values of the threshold used to define locations d, and sliding window size W. For H_0, we report the value of the linear fit coefficient b and the p-value. $H_0 : b = 0$ is rejected for $p < 0.05$. For H_1, we report the value of the power-law fit coefficient β and the corresponding p-value. $H_1 : \beta = 0$ is rejected for $p < 0.05$. For $H_{j,k}$, we report the percentage of rejected hypotheses $H_{j,k} : C_j = C_k$, with j and k two different time-intervals.

data	d (m)	W	$\lvert G_i \rvert < \sigma_{G_i}$	data	d (m)	W	$\lvert G_i \rvert < \sigma_{G_i}$
Lifelog	30	10	98%	CNS	5	50	94%
Lifelog	40	10	98%	CNS	10	10	98%
Lifelog	50	4	99%	MDC		4	98%
Lifelog	50	6	99%	MDC		6	95%
Lifelog	50	8	98%	MDC		8	97%
Lifelog	50	10	98%	MDC		10	99%
Lifelog	50	12	98%	MDC		12	99%
Lifelog	50	40	27%	MDC		40	94%
Lifelog	50	20	89%	MDC		50	83%
CNS	2	10	98%	MDC		20	95%
CNS	5	4	98%	RM		4	93%
CNS	5	6	98%	RM		6	90%
CNS	5	8	97%	RM		8	87%
CNS	5	10	98%	RM		10	84%
CNS	5	12	98%	RM		12	88%
CNS	5	40	95%	RM		20	88%

Table B.2: **Conservation of capacity: evidence 2.** For different values of the threshold used to define locations d, and sliding window size W, the percentage of individuals such that $\lvert G_i \rvert < \sigma_{G_i}$ (see appendix B.1).

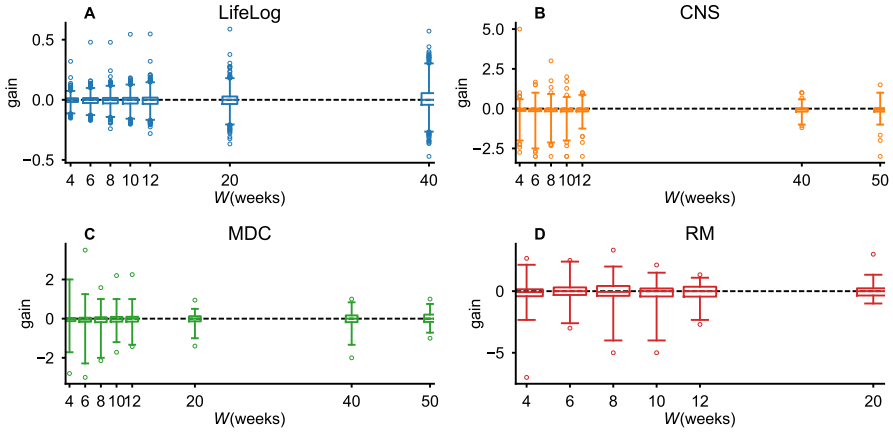

Figure B.2: **Gain: window size dependency** The boxplots of the individual average gain, as a function of the sliding window size for the Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)** datasets. Boxes contains the population interquartile (25 to 75 percentiles) and whiskers contain the 95% of the population (2.5 to 97.5 percentiles).

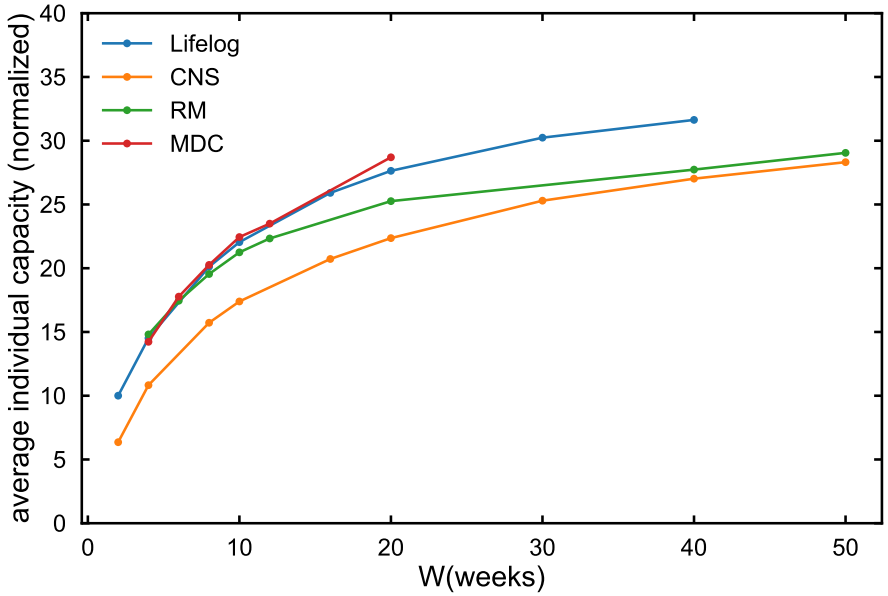

Figure B.3: **Saturation of the average normalized capacity.** The average value of the normalized capacity computed for increasing values of the time-window W. This result is obtained after accounting for the differences in data collection by computing the normalized location capacity C_i/TC_i, where TC_i is the weekly time coverage of individual i.

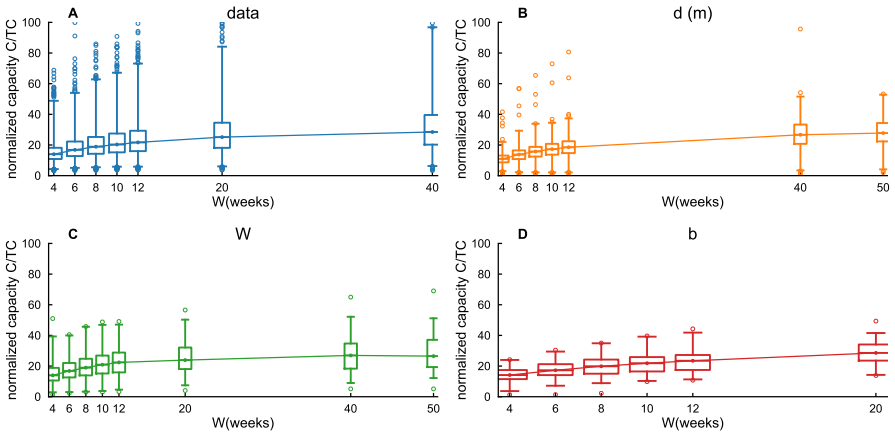

Figure B.4: **Capacity: window size dependency** The boxplots of the individual average capacity, as a function of the sliding window size for the Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)** datasets. Boxes contains the population interquartile (25 to 75 percentiles) and whiskers contain the 95% of the population (2.5 to 97.5 percentiles).

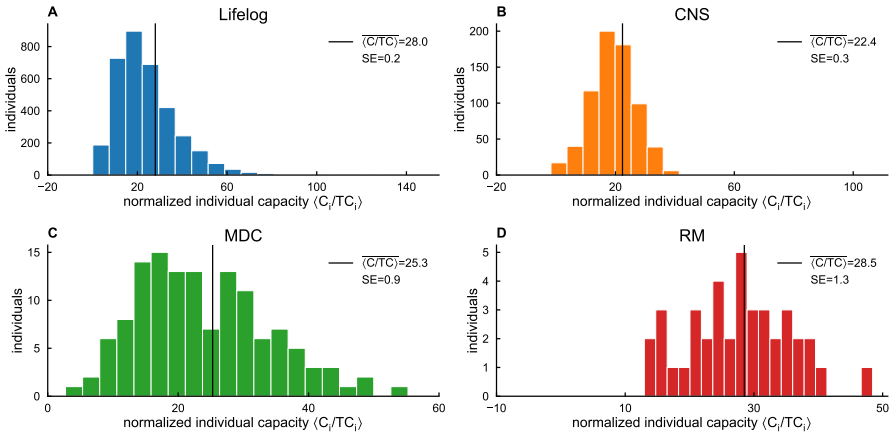

Figure B.5: **Individual capacity: population homogeneity** The frequency histogram of the normalized individual capacity $\langle C_i/TC_i \rangle$, where C_i and TC_i are respectively the location capacity and the time coverage of individual i. The average value $\overline{\langle C/TC \rangle}$ (black line) has standard error SE. Results are shown for the Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)**, computed with $W = 20$.

Evolution of the set of familiar locations: Invariance under time translation

We verified that the evolution of the set of familiar locations is not influenced by the particular time at which the data collection started or by the time elapsed from that moment. We borrow the concept of *aging* from the physics of glassy systems [336, 337]. A system is said to be in equilibrium when it shows invariance under time translations; if this holds, any observable comparing the system at time t with the system at time $t + \gamma$ is independent of the starting time t. In contrast, a system undergoing aging is not invariant under time translation. This property can be revealed by measuring correlations of the system at different times.

We measure the evolution of the set of familiar locations starting at different initial times t to verify if the system undergoes aging effects. The evolution is quantified measuring the Jaccard similarity $J_i(t,\gamma) = |S_i(t) \cap S_i(t+\gamma)|/|S_i(t) \cup S_i(t+\gamma)|$ (see MS). The average similarity $\overline{J(t,\gamma)}$ decreases in time: power-law fits of the form $\overline{J(t,\gamma)} =\sim \gamma^{\lambda(t)}$ yield $\lambda < 0$ for all t. The fit coefficient $\lambda(t)$ fluctuates around a typical value, because of seasonality effects, but does not changes substantially as a function of the starting time t (fig. B.6), hence $\overline{J(t,\gamma)} = \overline{J(\gamma)}$. This implies that the rate at which the set of familiar locations evolves does not substantially depends on when the measure is initiated. We conclude that our data reflect the 'equilibrium' behaviour of the monitored individuals. The fact that our dataset allow us to replicate measures performed on other datasets obtained with different methods (see above) further confirms this

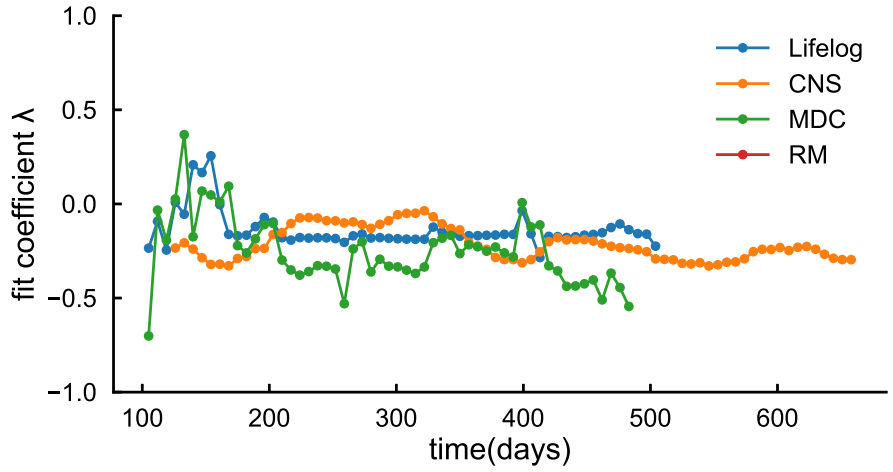

Figure B.6: **Evolution of set of familiar locations: invariance under time translation** The PL fit coefficients λ describing the evolution of the set of familiar locations as a function of the starting time of the measurement, for different datasets.

finding.

Sub-linear growth of number of locations

Individual exploration behaviour is quantified measuring the number of locations $L_i(t)$ discovered up to day t. In the MS, we show that $L_i(t)$ grows sub-linearly in time. Here, we show that this holds also changing the definition of locations (See fig. B.7). This property of exploration behaviour is not affected by the waiting time before starting the measure as we verify by repeating the same measures starting M months after the participant received the phone, for several values of M (See fig. B.8).

Discrepancy relative to the randomized cases

Individual capacity is lower than it could be if individuals were only subject to time constraints. We showed this by randomizing individual temporal sequences of stop-locations for 100 times, and then comparing the average randomized capacity $\langle C_{rand,i} \rangle$ with the real capacity $\langle C_i \rangle$. We perform two types of randomizations (see fig. B.9):

- (1) Local randomization: For each individual i, we split her digital traces in segments of length 1 day. We shuffle days of each individual.

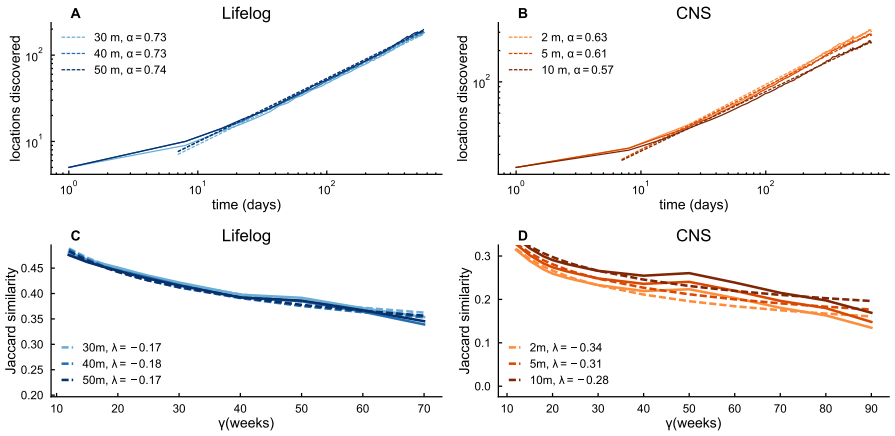

Figure B.7: **Effects of different definitions of locations** The average number of locations discovered up to a given day for different definitions of location, and the corresponding power-law fits (dashed line) with coefficient α, for the Lifelog **(A)** and CNS **(B)** datasets. **(C,D)** The average overlap (Jaccard similarity) between the set of familiar locations at week t and week $t+\gamma$ (full line), and the corresponding power law fit $\overline{J(\gamma)} \sim \gamma^{\lambda}$ (dashed lines) (dashed line) for different definitions of location. Results are shown for the Lifelog **(C)** and CNS **(D)** datasets.

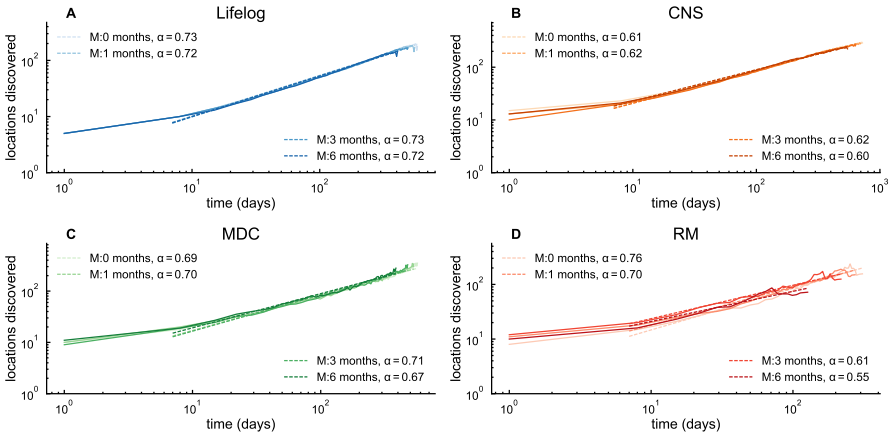

Figure B.8: **Exploration behaviour: invariance under time translation** The average number of locations individually discovered in time, measured after waiting M months, and the corresponding power-law function fit with coefficients α (dashed lines) for different values of M. Results are shown for the Lifelog **(A)**, CNS **(B)**, MDC **(C)** and RM **(D)** datasets.

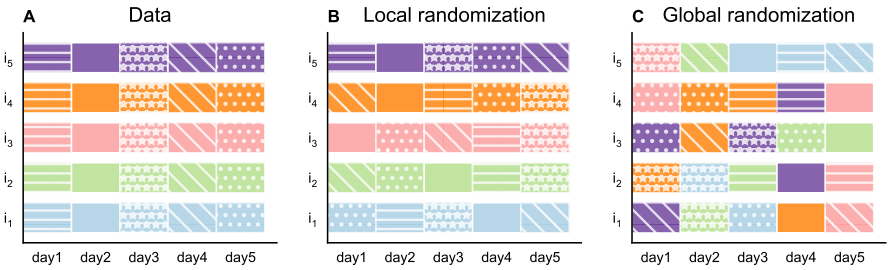

Figure B.9: **Data randomization schema.** A schematic representation of local and global randomization. **(A)** Individual time series for 5 individuals are divided into modules of 1 day length (each day has a specific color pattern). **(B)** In the *local randomization* individual timeseries are shuffled preserving the module units. **(C)** In the *global randomization* new sequences are created assembling together modules extracted randomly from the whole set of individual traces.

data	KS (local)	p (local)	KS (global)	p (global)
Lifelog	0.21	0		
CNS	0.29	0.0	0.94	0.0
MDC	0.36	0.0	0.99	0.0
RM	0.35	0.0	0.99	0.0

Table B.3: **Discrepancy with the randomized case.** The Kolomogorov-Smirnov (KS) test statistics measuring the discrepancy between the capacity in the real and randomized case, with the corresponding p-values. Since $p < 0.05$ we can reject the hypothesis that the distributions underlying the two samples are the same under a 2-tailed test. Results are shown for the local and global randomization, for different datasets.

- (2) Global randomization: For each individual i, we split her digital traces in segments of length 1 day. We shuffle days of different individuals.

The individual randomized capacity $\langle C_{rand,i} \rangle$ averaged across time, (see fig. B.10), is higher than in the real case both for the global and the local randomization cases. We compute the Kolmogorov–Smirnov test-statistics (table B.3) to compare the real sample with the randomized samples. We reject the hypothesis that the two samples are extracted from the same distribution since $p < \alpha$ with $\alpha = 0.05$.

Conservation of time allocation

Individuals allocate time heterogeneously among locations, due to their different functions (homes, work-places, shops, universities, leisure places...). We study time allocation between different classes of loca-

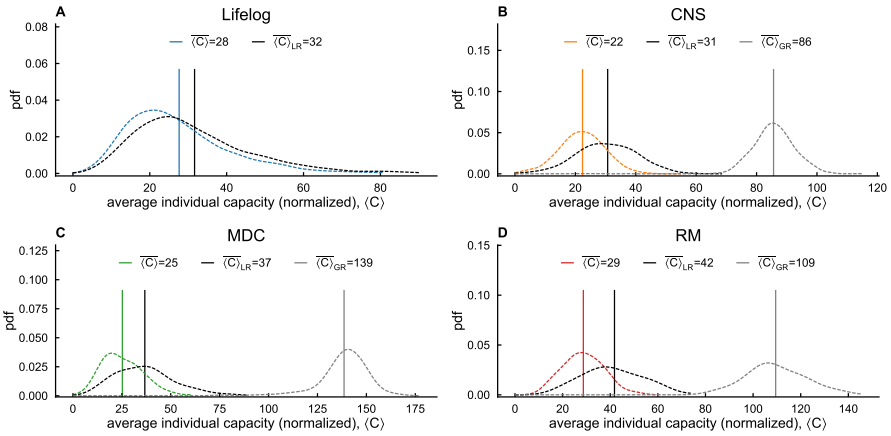

Figure B.10: **Discrepancy with the randomized cases.** The Kernel Density of the average individual capacity (normalized to account for the differences in time coverage) for data ($\langle C \rangle$), local ($\langle C \rangle_{LR}$) and global ($\langle C \rangle_{GR}$) randomizations (dashed lines), and the corresponding average values (full lines) computed across the population. The Kolmogorov–Smirnov test-statistics (table B.3) rejects the hypothesis that the three samples are extracted from the same distribution.

tions considering subsets of the set of familiar locations defined on the basis of the total visitation time. The subsets $S_i(t)^{\Delta T} \in S_i(t)$ include all locations seen in the W weeks preceding t at least twice and such that $W * \Delta t(0) < T_{i,\ell}(t) < W * \Delta t(1)$ where $T_{i,\ell}(t)$ is the time of observation of location ℓ during the W weeks preceding t.
We test several choices of intervals ΔT. We find that when ΔT increases, the subsets are empty for many individuals, since no locations satisfy the above-mentioned criteria. In figs. B.11 to B.14 we show the distribution of average individual sub-capacities $\langle C_i^{\Delta T} \rangle$. Only subsets with small enough ΔT are significant for more than 50% of the population, and typically each individual has 1 location where he/she spend more than 48 hours per week.
The average sub-capacities $\overline{C}^{\Delta T}(t)$ are constant in time for several choices of ΔT and different definitions of location. This is verified with the linear fit test as detailed in a previous section (see table B.4).

B.2 SPATIAL PROPERTIES OF THE SET OF FAMILIAR LOCATIONS.

We consider two spatial properties of the set of familiar locations S (see appendix B.1): its center of mass and its radius of gyration (see also [32, 79]). The center of mass is computed as:

$$\vec{r_{cm}} = \frac{1}{T} \sum_{j \in L} t_j \vec{r_j}$$

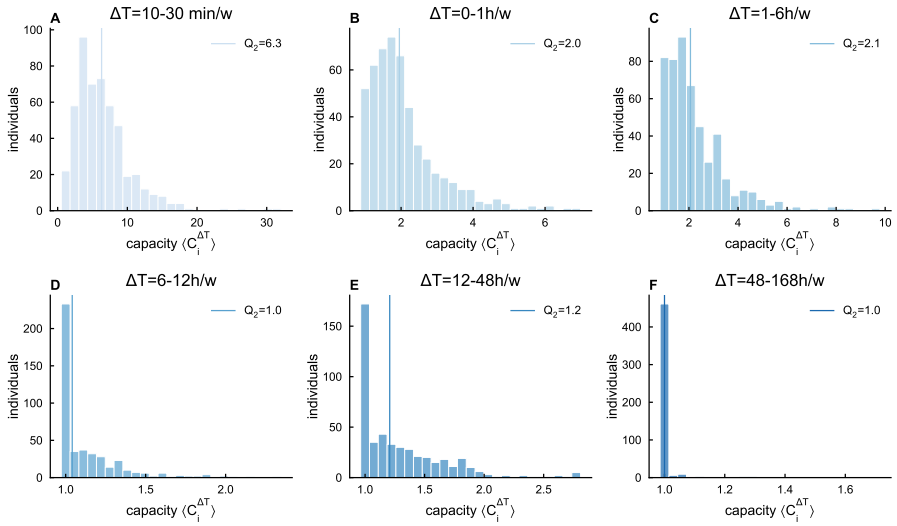

Figure B.11: **Lifelog dataset: Composition of the set of familiar locations. A-F**) The distribution of the average individual capacity $\langle C_i \rangle^{\Delta T}\rangle$, considering locations seen for a time included in ΔT.

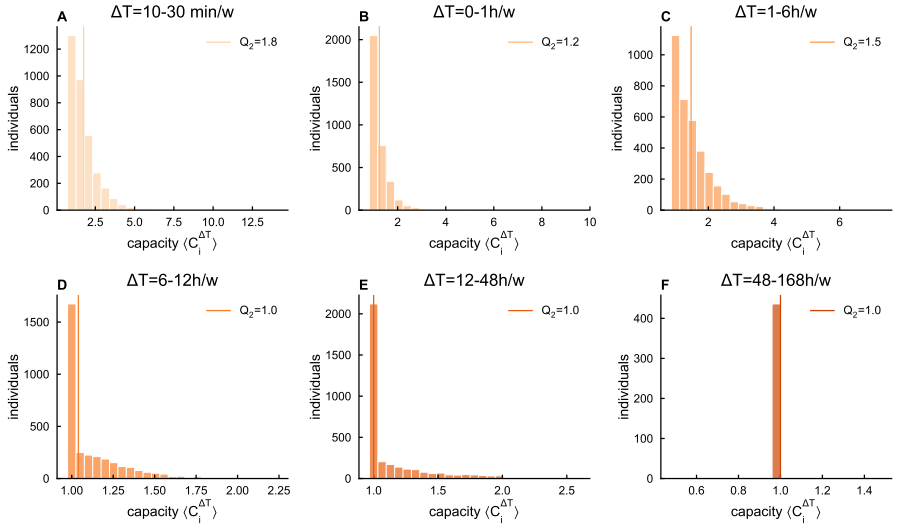

Figure B.12: **CNS dataset: Composition of the set of familiar locations A-F**) The distribution of the average individual capacity $\langle C_i \rangle^{\Delta T}\rangle$, considering locations seen for a time included in ΔT.

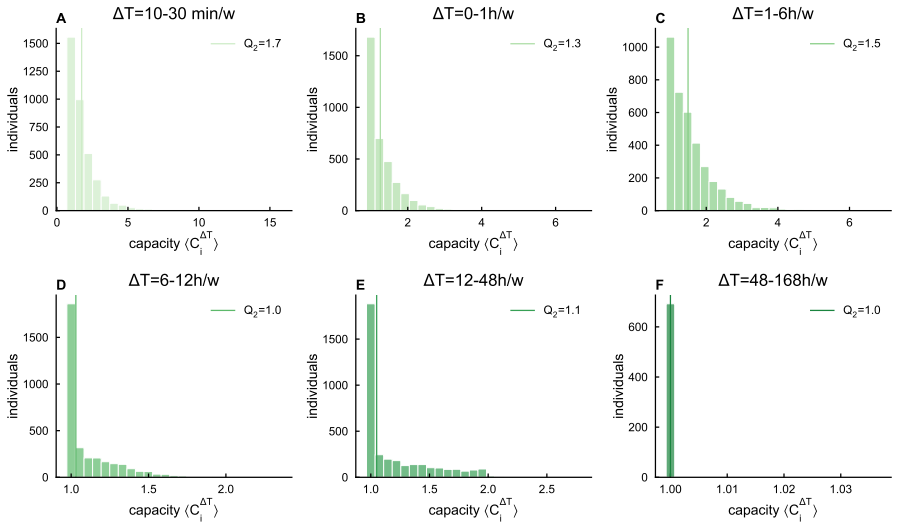

Figure B.13: **MDC dataset: Composition of the set of familiar locations. A-F)** The distribution of the average individual capacity $\langle C_i \rangle^{\Delta T}\rangle$, considering locations seen for a time included in ΔT.

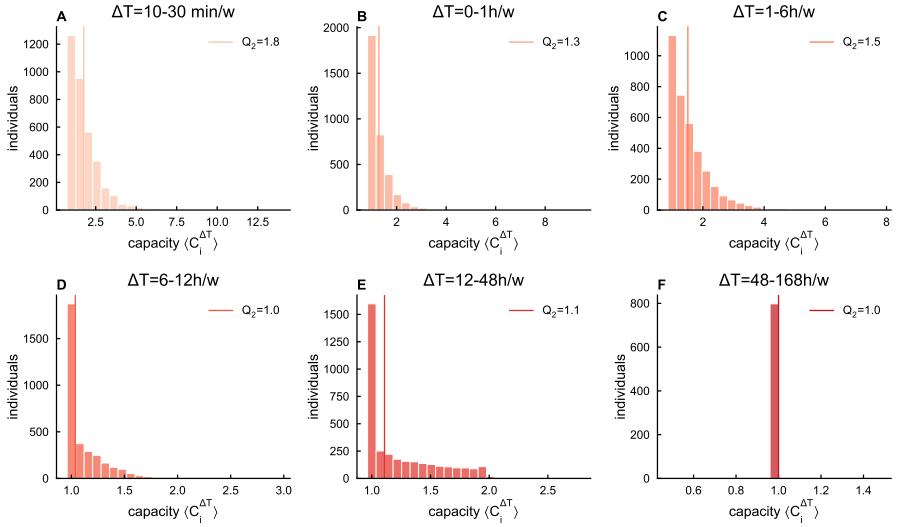

Figure B.14: **RM dataset: Composition of the set of familiar locations. A-F)** The distribution of the average individual capacity $\langle C_i \rangle^{\Delta T}\rangle$, considering locations seen for a time included in ΔT.

data	d (m)	W	10-30min	30-60 min	1-6 h	6-12h	12-48 h	48 h
			p	p	p	p	p	p
Lifelog	30	10	1.00	1.00	0.98	0.99	1.00	1.00
Lifelog	40	10	0.96	1.00	1.00	1.00	1.00	1.00
Lifelog	50	4	1.00	1.00	0.99	0.99	1.00	1.00
Lifelog	50	6	0.99	1.00	1.00	0.99	1.00	1.00
Lifelog	50	8	0.92	0.98	0.97	0.99	1.00	1.00
Lifelog	50	10	0.99	1.00	0.98	0.99	1.00	1.00
Lifelog	50	12	0.99	0.99	0.98	0.98	1.00	1.00
Lifelog	50	40	0.75	0.96	0.78	0.99	0.98	1.00
Lifelog	50	20	0.83	0.99	1.00	0.98	0.99	1.00
CNS	2	10	0.94	0.98	1.00	0.99	0.99	1.00
CNS	5	4	0.97	0.99	0.99	1.00	0.99	1.00
CNS	5	6	0.97	0.99	0.99	1.00	0.99	1.00
CNS	5	8	0.96	0.98	0.99	1.00	1.00	1.00
CNS	5	10	0.94	0.98	0.99	0.99	1.00	1.00
CNS	5	12	0.93	0.98	0.97	1.00	0.99	1.00
CNS	5	40	0.94	0.99	0.94	0.99	0.99	1.00
CNS	5	50	0.92	0.98	0.92	0.99	0.99	0.99
CNS	10	10	0.95	0.98	0.99	0.99	0.99	0.99
MDC	0	4	0.96	0.99	0.97	0.97	0.96	1.00
MDC	0	6	0.97	0.99	0.99	0.98	0.97	1.00
MDC	0	8	0.98	1.00	0.99	0.97	0.96	1.00
MDC	0	10	0.96	0.99	0.99	0.97	0.96	1.00
MDC	0	12	0.95	0.99	0.98	0.99	0.96	0.99
MDC	0	40	0.91	0.96	0.98	0.95	1.00	0.99
MDC	0	50	0.95	0.91	0.90	0.95	0.94	0.99
MDC	0	20	0.90	0.98	0.97	0.99	0.97	0.99
RM	0	4	0.97	0.95	0.93	0.91	0.96	0.99
RM	0	6	0.93	0.93	1.00	0.93	0.98	0.99
RM	0	8	0.97	0.84	0.99	0.97	0.98	0.99
RM	0	10	0.99	0.89	0.95	0.94	0.95	0.99
RM	0	12	0.94	0.87	0.92	0.93	0.93	0.99
RM	0	20	0.80	0.86	0.89	0.96	0.87	1.00

Table B.4: **Conservation of time allocation.** The results of hypotheses testing H_0 for different classes of locations ΔT. Results are shown for different values of the threshold used to define locations d, and sliding window size W. We report the p-value, testing the hypothesis $H_0 : b = 0$ is rejected for $p < 0.05$.

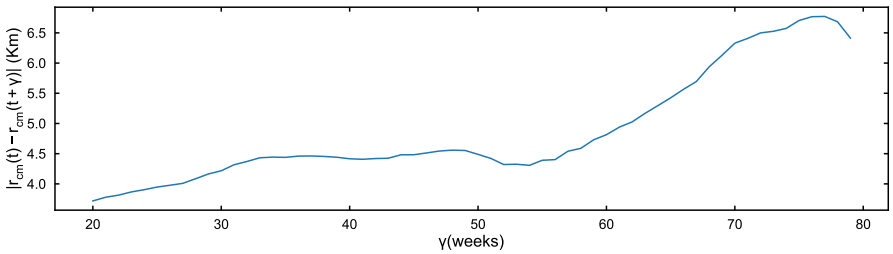

Figure B.15: **CNS dataset: Displacement of the set of familiar locations center of mass.** The average distance between the center of mass of the set of familiar locations $r_cm(t)$ computed at time t, and the same quantity computed at time $r_cm(t+\gamma)$. The distance is averaged across values of t and plotted as a function of the delay γ. Results are shown for sets computed using a sliding window of size $w = 20$ weeks.

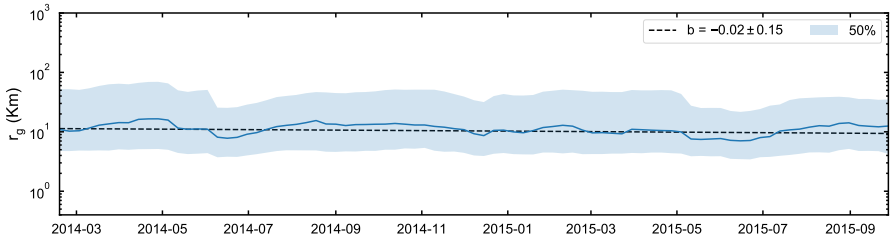

Figure B.16: **CNS dataset: Constant radius of gyration of the set of familiar locations.** Median value of the radius of gyration $r_g(t)$ of the set of familiar locations as a function of time (blue line). The light blue shaded area is the 50% of the sample around the median. The dashed line is a linear fit with coefficient $b = -0.02 \pm 0.15$.

where T is the total time spent in familiar locations, t_j is the time spent in location j and $\vec{r}_j$ is the spatial position of location j. The radius of gyration is computed as:

$$r_{g,i}(t) = \sqrt{\frac{1}{T}\sum_{j \in L} t_j (\vec{r}_j - \vec{r_{cm}})^2}$$

For the CNS data, we compute these quantities for all individuals and for all windows of length 20 weeks within the range of the experiment. We find that the center of mass of the set of familiar locations moves in time, on average (see fig. B.15): The average distance between $r_{cm}(t)$ and $r_{cm}(t+\gamma)$ increases as a function of γ but not constantly in time. The offset at $\sim 3Km$ reveals that individuals' sets may be highly unstable. This is partly explained by individuals' renewing favourite locations within the Copenhagen area and partly by their long-range travelling. Instead, the median radius of gyration is constant in time, with a linear fit $r_{cm} = a + b \cdot t$ yielding $a = 11 \pm 8$ Km and $b = -0.02 \pm 0.15$ $Km/week$ (see fig. B.16).

Finally, we find that the location capacity is significantly different between the so-called *'returners'* and *'explorers'* (see [79] and section

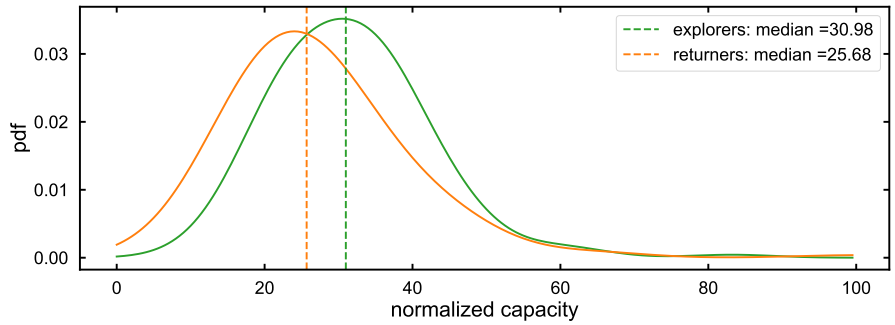

Figure B.17: **CNS dataset: Differences between returners and explorers.** Probability distribution of the average normalized location capacity for *returners* (orange line) and *explorers* (green line), according to the definition in [79].

'*Comparison with previous research*') . Individuals defined as explorers (see fig. 2.8) have higher capacity under the Kolmogorov–Smirnov test-statistics (see fig. B.17).

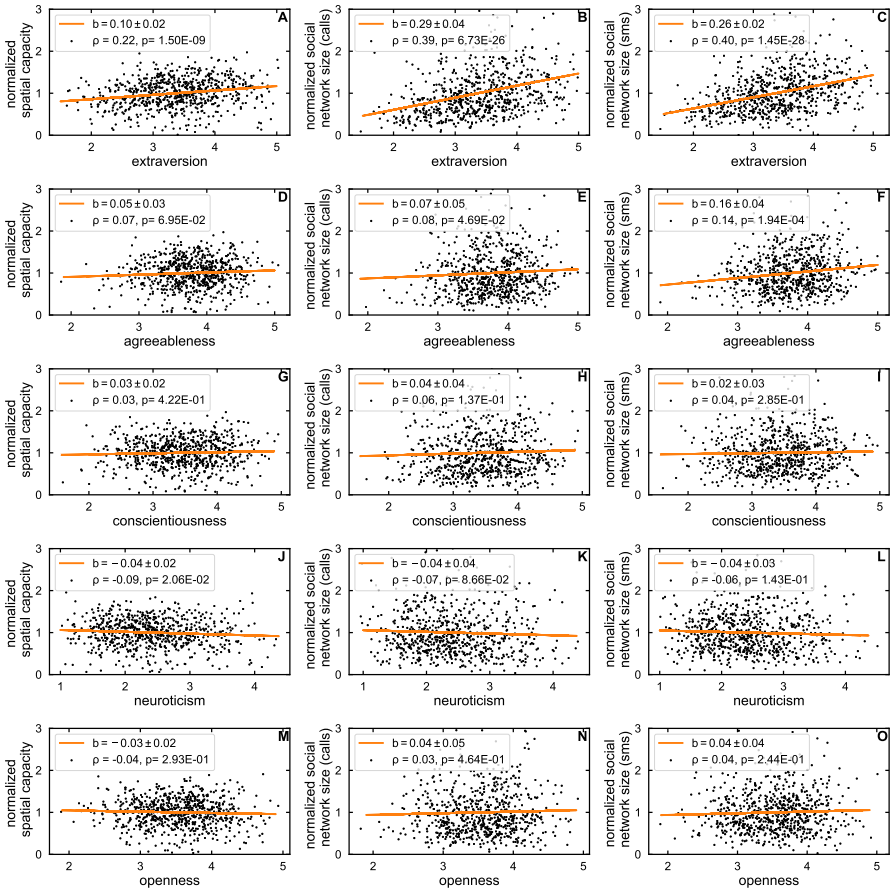

Figure B.18: **Social network size and location capacity correlate with extraversion.** The average normalized location capacity (left column), social network size computed from calls interactions(center column) and sms interactions (right column) as a function of each of the BigFive personality traits, measured on a scale from 0 to 5. The personality traits are: extraversion (first row), agreeableness (second row), conscientiousness (third row), neuroticism (fourth row) and openness (fifth row). The legend report the value of the slope b of a linear regression line, the Pearson correlation coefficient ρ, with associated p-value p. Results are shown for $W = 20$ weeks.

C

APPENDIX TO CHAPTER 3

In this appendix, we provide additional results to chapter 5. In appendix C.1, we present the same analyses ran in the main text, but for the MDC dataset. In appendix C.2, the results are shown for a sliding window of length 30 weeks.

C.1 RESULTS OBTAINED WITH THE MDC DATASET

tables C.1 and C.2 and fig. C.1 report the results of the persistence analysis, the multiple regression analysis, and the correlation analysis for the MDC dataset.

	$\overline{d_{self}}$	$\overline{d_{ref}}$	$\overline{d_{self}(i) < d_{ref}(i,j)}$
Social circle size, k	0.05 ± 0.13	10 ± 5	97%
Activity space size, C	0.07 ± 0.12	8 ± 3	97%
New locations/week, n_{loc}	0.2 ± 0.3	2 ± 1	91%
New ties/week, n_{tie}	0.2 ± 0.6	2 ± 1	90%
Social circle entropy, H_{SC}	0.006 ± 0.014	0.7 ± 0.3	97%
Activity space entropy, H_{AS}	0.004 ± 0.008	0.5 ± 0.2	97%
Social circle stability, J_{SC}	0.002 ± 0.005	0.15 ± 0.05	99%
Activity space stability, J_{AS}	0.002 ± 0.004	0.12 ± 0.05	99%
Social circle rank turnover, R_{SC}	0.07 ± 0.15	2 ± 1	98%
Activity space rank turnover, R_{AS}	0.2 ± 0.6	2 ± 1	97%

Table C.1: **MDC dataset: Persistence of social and spatial behaviour.** For each of the social and spatial metrics, $\overline{d_{self}}$ is the average self-distance and $\overline{d_{ref}}$ is the reference distance between an individual and all others, averaged across individuals. The third column reports the fraction of cases where $\overline{d_{self}(i) < d_{ref}(i,j)}$, averaged across the population.

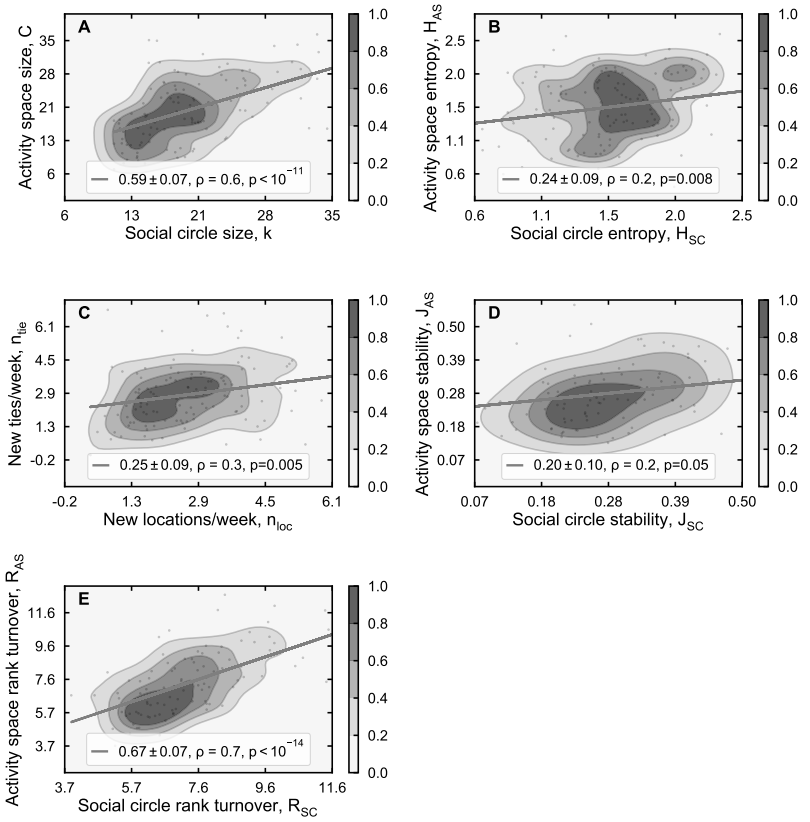

Figure C.1: **MDC dataset: correlation between the four dimensions of social and spatial behaviour.** (A) Activity space vs social circle size. (B) Activity space vs social circle composition measured as their entropy. (C) Average number of new locations vs new ties per week. (D) Stability of the activity space vs the stability of the social circle measured as the Jaccard similarity between their composition in consecutive time-windows. (E) Rank turnover of the activity space vs the rank turnover of the social circle. Coloured filled areas correspond to cumulative probabilities estimated via Gaussian Kernel Density estimations. Grey lines correspond to linear fit with angular coefficient b reported in the legend. The Pearson correlation coefficient, with corresponding p-value, is reported in the legend.

Model M1: Activity space size, C	coeff	p val	LMG
Social circle size, k	5 ± 1	$< 10^{-11}$	0.98
gender	0.1 ± 0.6	0.8	0.01
age group	0.6 ± 0.6	0.3	0.01
time coverage	-0.4 ± 0.6	0.4	0.0
$[R^2 = 0.40, F = 16.80, p_F = 0.0]$			
Model M2: Activity space entropy, H_{AS}			
Social circle entropy, H_{SC}	0.11 ± 0.04	0.009	0.28
gender	0.04 ± 0.04	0.3	0.03
age group	-0.08 ± 0.04	0.06	0.21
time coverage	-0.14 ± 0.04	0.002	0.48
$[R^2 = 0.20, F = 6.50, p_F = 0.0]$			
Model M3: New ties/week, n_{tie}			
New locations/week, n_{loc}	0.5 ± 0.1	0.002	0.69
gender	0.01 ± 0.15	0.9	0.04
age group	0.2 ± 0.1	0.2	0.1
time coverage	-0.3 ± 0.1	0.06	0.17
$[R^2 = 0.13, F = 3.78, p_F = 0.0]$			
Model M4: Activity space stability, J_{AS}			
Social circle stability, J_{SC}	0.02 ± 0.01	0.1	0.82
gender	-0.006 ± 0.012	0.6	0.15
age group	-0.003 ± 0.012	0.8	0.03
time coverage	$(-10 \pm 1213) \cdot 10^{-5}$	1.0	0.0
$[R^2 = 0.04, F = 0.80, p_F = 0.5]$			
Model M5: Activity space rank turnover, R_{AS}			
Social circle rank turnover, R_{SC}	1 ± 0	$< 10^{-15}$	0.97
gender	0.04 ± 0.15	0.8	0.02
age group	-0.2 ± 0.1	0.1	0.01
time coverage	-0.06 ± 0.15	0.7	0.0
$[R^2 = 0.55, F = 27.24, p_F = 0.0]$			

Table C.2: **Linear regression models for the MDC dataset.** For each model, we report the R^2 goodness of fit, the $F - test$ statistics with the corresponding p-value p_F. We show the coefficients (coeff) calculated by the regression model, the probability (p val) that the variable is not relevant, and the relative importance (LMG) of each regressor computed using the Lindeman, Merenda and Gold method. Gender is a binary variable taking value 1 for females and 2 for males.

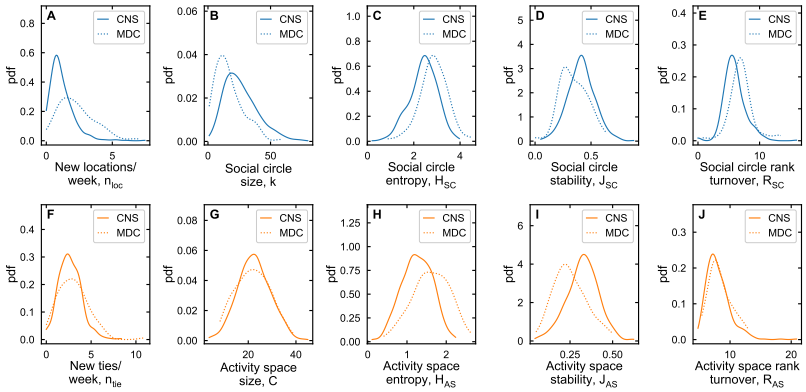

Figure C.2: **T=30, Distribution of social (above line) and spatial (bottom line) metrics for the CNS and MDC datasets.**

C.2 RESULTS OBTAINED WITH OTHER WINDOWS

figs. C.2 to C.6 and tables C.3 to C.10 report the results obtained choosing a time-window with length $T = 30$ weeks (see chapter 5).

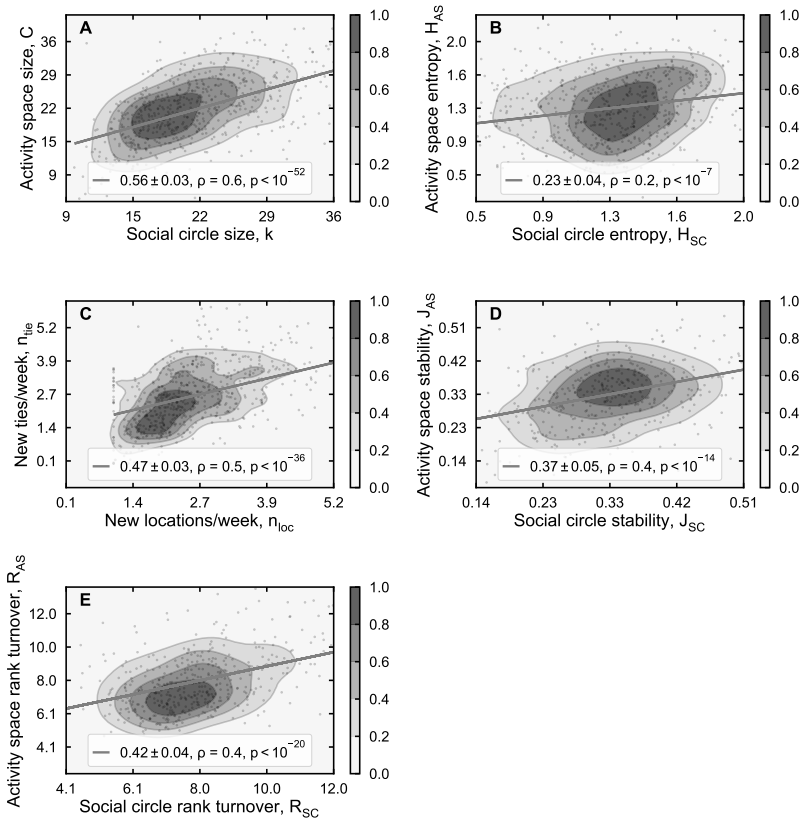

Figure C.3: **T=30, CNS dataset: correlation between the four dimensions of social and spatial behaviour.** (A) Activity space vs social circle size. (B) Activity space vs social circle composition measured as their entropy. (C) Average number of new locations vs new ties per week. (D) Stability of the activity space vs the stability of the social circle measured as the Jaccard similarity between their composition in consecutive time-windows. (E) Rank turnover of the activity space vs the rank turnover of the social circle. Coloured filled areas correspond to cumulative probabilities estimated via Gaussian Kernel Density estimations. Grey lines correspond to linear fit with angular coefficient b reported in the legend. The Pearson correlation coefficient, with corresponding p-value, is reported in the legend.

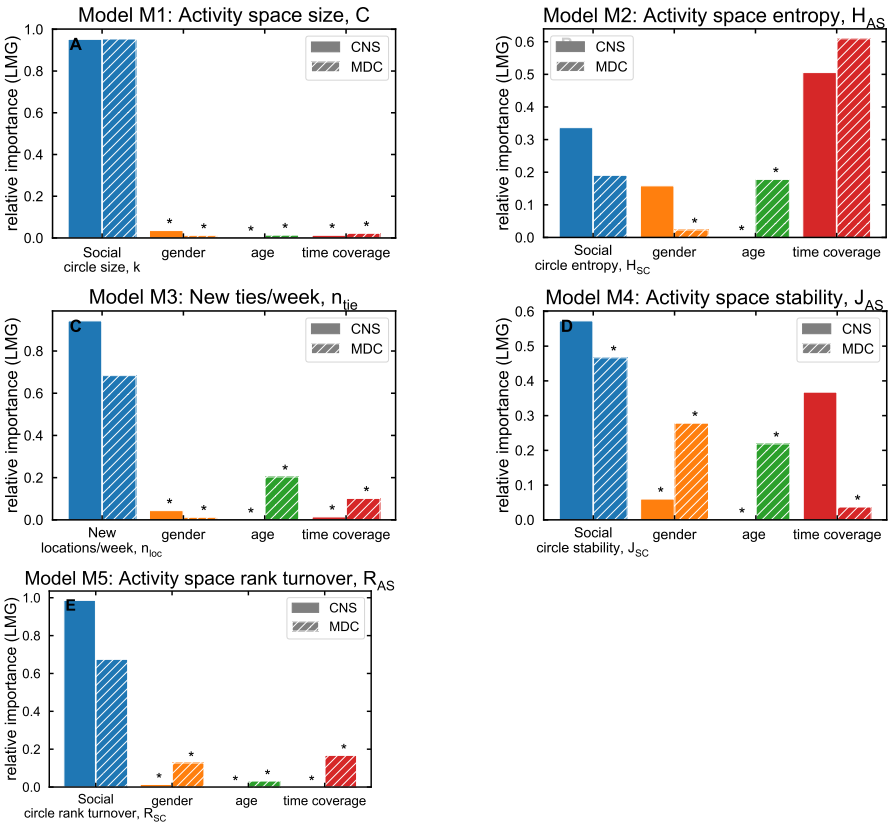

Figure C.4: **T=30, Relative importance of regressors** LMG of each regressor computed using the Lindeman, Merenda and Gold method for models M1 (A), M2 (B), M3 (C), M4 (D) and M5 (E). Plain bars show results for the CNS dataset, dashed bars for the MDC dataset. Variables that are not significant in the regression model are marked with *.

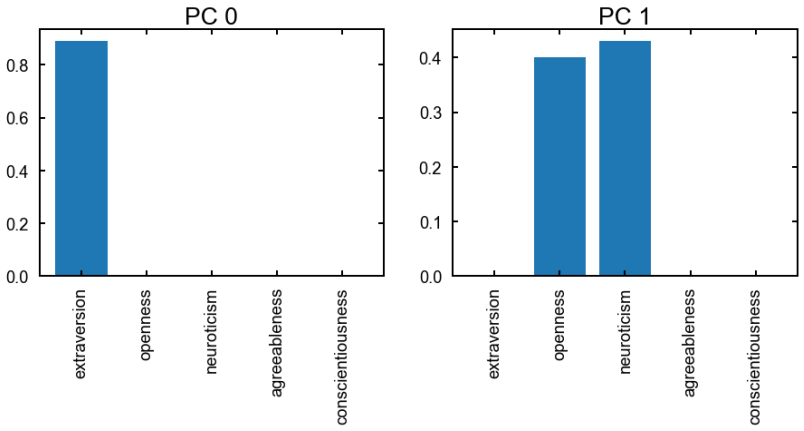

Figure C.5: **T=30, Relative importance of personality traits for socio-spatial behaviour** LMG of each regressor computed using the Lindeman, Merenda and Gold method for the multiple regression model of the principal components (table C.7).

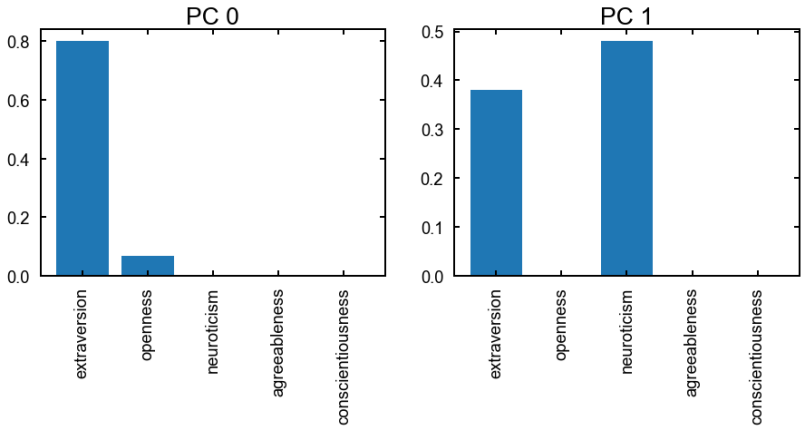

Figure C.6: **T=30, Relative importance of personality traits for spatial behaviour** LMG of each regressor computed using the Lindeman, Merenda and Gold method for the multiple regression model of the principal components (table C.10).

	$\overline{d_{self}}$	$\overline{d_{ref}}$	$\overline{d_{self}(i) < d_{ref}(i,j)}$
Social circle size, k	0.04 ± 0.13	15 ± 6	100%
Activity space size, C	0.04 ± 0.07	8 ± 3	99%
New locations/week, n_{loc}	0.06 ± 0.12	0.9 ± 0.5	96%
New ties/week, n_{tie}	0.1 ± 0.2	1 ± 1	95%
Social circle entropy, H_{SC}	0.002 ± 0.005	0.7 ± 0.3	99%
Activity space entropy, H_{AS}	0.002 ± 0.006	0.4 ± 0.1	99%
Social circle stability, J_{SC}	$(6 \pm 15) \cdot 10^{-4}$	0.14 ± 0.05	100%
Activity space stability, J_{AS}	$(6 \pm 11) \cdot 10^{-4}$	0.10 ± 0.04	100%
Social circle rank turnover, R_{SC}	0.04 ± 0.11	2 ± 1	99%
Activity space rank turnover, R_{AS}	0.04 ± 0.20	2 ± 1	99%

Table C.3: **T=30, CNS dataset: Persistence of social and spatial behaviour.** For each of the social and spatial metrics, $\overline{d_{self}}$ is the average self-distance and $\overline{d_{ref}}$ is the reference distance between an individual and all others, averaged across individuals. The third column reports the fraction of cases where $\overline{d_{self}(i) < d_{ref}(i,j)}$, averaged across the population.

Model M1: Activity space size, C	coeff	p val	LMG
Social circle size, k	4 ± 0	$< 10^{-46}$	0.95
gender	-0.3 ± 0.2	0.2	0.04
time coverage	0.5 ± 0.2	0.05	0.01
$[R^2 = 0.32, F = 91.23, p_F = 0.0]$			
Model M2: Activity space entropy, H_{AS}			
Social circle entropy, H_{SC}	0.07 ± 0.02	$< 10^{-4}$	0.34
gender	-0.05 ± 0.02	$< 10^{-3}$	0.16
time coverage	-0.09 ± 0.02	$< 10^{-8}$	0.51
$[R^2 = 0.12, F = 26.82, p_F = 0.0]$			
Model M3: New ties/week, n_{tie}			
New locations/week, n_{loc}	0.58 ± 0.05	$< 10^{-30}$	0.94
gender	-0.09 ± 0.05	0.04	0.04
time coverage	0.03 ± 0.05	0.5	0.01
$[R^2 = 0.22, F = 55.56, p_F = 0.0]$			
Model M4: Activity space stability, J_{AS}			
Social circle stability, J_{SC}	0.027 ± 0.004	$< 10^{-9}$	0.57
gender	0.009 ± 0.004	0.02	0.06
time coverage	0.020 ± 0.004	$< 10^{-5}$	0.37
$[R^2 = 0.18, F = 30.32, p_F = 0.0]$			
Model M5: Activity space rank turnover, R_{AS}			
Social circle rank turnover, R_{SC}	0.81 ± 0.08	$< 10^{-19}$	0.99
gender	0.09 ± 0.08	0.3	0.01
time coverage	-0.001 ± 0.084	1.0	0.0
$[R^2 = 0.18, F = 31.70, p_F = 0.0]$			

Table C.4: T=30, Linear regression models for the CNS dataset. For each model, we report the R^2 goodness of fit, the $F - test$ statistics with the corresponding p-value p_F. We show the coefficients (coeff) calculated by the regression model, the probability (p val) that the variable is not relevant, and the relative importance (LMG) of each regressor computed using the Lindeman, Merenda and Gold method. Gender is a binary variable taking value 1 for females and 2 for males. For this dataset, age is not relevant as all participants have similar age.

	PC 0	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
CNS	0.40	0.18	0.10	0.07	0.07	0.06	0.04	0.03	0.03	0.02
MDC	0.39	0.19	0.12	0.10	0.06	0.05	0.05	0.03	0.02	0.01

Table C.5: **T=30, Variance explained by principal components.** The fraction of variance explained by each principal component for the CNS and MDC dataset.

	CNS		MDC	
	PC 0	PC 1	PC 0	PC 1
Social circle size, k	0.41	0.18	-0.36	0.04
Activity space size, C	0.42	-0.23	-0.40	-0.05
New locations/week, n_{loc}	0.33	0.27	-0.24	-0.35
New ties/week, n_{tie}	0.39	-0.11	-0.37	-0.22
Social circle entropy, H_{SC}	0.29	0.33	-0.36	-0.23
Activity space entropy, H_{AS}	0.38	-0.10	-0.35	0.13
Social circle stability, J_{SC}	-0.12	-0.50	-0.11	0.56
Activity space stability, J_{AS}	-0.06	-0.50	-0.03	0.62
Social circle rank turnover, R_{SC}	-0.17	0.26	0.28	-0.19
Activity space rank turnover, R_{AS}	-0.35	0.37	0.42	-0.18

Table C.6: **T=30, Principal Components.** The weight of each metric in the first two principal components, for both datasets.

	PC 0 $R^2 = 0.17$, $F = 17.08$, $p_F = 0.0$			PC 1 $R^2 = 0.02$, $F = 2.05$, $p_F = 0.1$		
	coeff	p val	LMG	coeff	p val	LMG
E	0.9 ± 0.1	$< 10^{-15}$	0.89	0.06 ± 0.07	0.4	0.04
O	-0.18 ± 0.09	0.06	0.02	0.13 ± 0.07	0.05	0.4
N	0.1 ± 0.1	0.1	0.03	0.15 ± 0.07	0.04	0.43
A	0.05 ± 0.10	0.6	0.02	-0.04 ± 0.07	0.5	0.09
C	0.04 ± 0.10	0.7	0.04	-0.03 ± 0.07	0.7	0.04

Table C.7: **T=30, Extraversion, openness, and neuroticism explain socio-spatial behaviour.** The result of a multiple linear regression explaining principal components of socio-spatial data (see table 5.6) from personality traits (E: extraversion, O: openness, N: neuroticism, A: agreableness, C: conscientiousness). The value of each coefficient (coeff) is reported together with the probability (p val) that the coefficient is not relevant for the model. The relative importance of each coefficient (LMG) is computed using the LMG method [205].

	PC 0	PC 1	PC 2	PC 3	PC 4
CNS	0.57	0.21	0.10	0.08	0.04
MDC	0.55	0.24	0.12	0.06	0.03

Table C.8: **T=30, Variance explained by principal components (only spatial data).** The fraction of variance explained by each principal component for the CNS and MDC dataset.

	CNS		MDC	
	PC 0	PC 1	PC 0	PC 1
Activity space size, C	-0.55	0.01	-0.55	-0.10
New ties/week, n_{tie}	-0.48	-0.16	-0.48	-0.35
Activity space entropy, H_{AS}	-0.48	-0.14	-0.44	0.20
Activity space stability, J_{AS}	-0.06	0.96	-0.02	0.88
Activity space rank turnover, R_{AS}	0.48	-0.16	0.51	-0.22

Table C.9: **T=30, Principal Components (only spatial data).** The weight of each metric in the first two principal components, for both datasets.

	PC 0 $R^2 = 0.11$, $F = 11.55$, $p_F = 0.0$			PC 1 $R^2 = 0.02$, $F = 2.02$, $p_F = 0.1$		
	coeff	p val	LMG	coeff	p val	LMG
E	-0.56 ± 0.09	$< 10^{-9}$	0.8	-0.12 ± 0.05	0.03	0.38
O	0.20 ± 0.08	0.01	0.07	-0.04 ± 0.05	0.5	0.08
N	0.03 ± 0.08	0.7	0.07	-0.14 ± 0.05	0.01	0.48
A	-0.05 ± 0.08	0.5	0.03	-0.002 ± 0.052	1.0	0.01
C	-0.005 ± 0.081	1.0	0.03	-0.03 ± 0.05	0.6	0.04

Table C.10: **T=30, Extraversion, openness, and neuroticism explain spatial behaviour.** The result of a multiple linear regression explaining principal components of spatial data (see table 5.6) from personality traits (E: extraversion, O: openness, N: neuroticism, A: agreableness, C: conscientiousness). The value of each coefficient (coeff) is reported together with the probability (p val) that the coefficient is not relevant for the model. The relative importance of each coefficient (LMG) is computed using the LMG method [205].

D

APPENDIX TO CHAPTER 7

This is the appendix to chapter 7. In appendix D.1, we present the datasets. In appendix D.2, we present the Non negative Matrix factorization method. In appendix D.3, the modified Dijkstra algorithm. In appendix D.4, we present patterns detected for the city of Strasbourg, Nantes, and Toulouse. In appendix D.5 we compare the patterns detected and commuting flows. In appendix D.6, we show the characteristics of privileged connections. In appendix D.7, we compare our method with a single layer representation.

D.1 DATA DESCRIPTION

With the aim of catching a comprehensive picture of the public transportation (PT) networks in French municipal areas we made use of datasets provided by local public transportation companies. The characteristics of the datasets used for the different cities are listed in table D.1. Estimated timetable schedules for the public transport service are made publicly available online and frequently updated by the companies.

City	Area	Period	Companies
Paris	47.96N-49.45N 1.15W-3.51W	Sep-Oct 2013	RATP (Bus, Metro, Tram, RER) SNCF (RER,Train)
Toulouse	43.43N- 43.74N 1.17W- 1.69W	Sep-Oct 2014	Tisséo (Bus, Tram, Metro) SNCF (Train)
Nantes	47.12N- 47.32N 1.75W-1.34W	Jan 2015	Semitain (Bus, Tram, Ferry) SNCF (Train)
Strasbourg	48.46N- 48.68N 7.60W-7.83W	Jan 2015	CTS (Bus, Tram) SNCF (Train)

Table D.1: **Datasets** Table listing the main characteristics of the data used for each of the cities.

All datasets are provided in General Transit Feed Specification (GTFS) format [338]. GTFS is a common format for PT schedules and associated geographic information. It is composed of a series of text files: stops, routes, trips, and other schedule data. In particular, the follow-

ing objects and associated attributes are of relevance to the purpose of this study:

- **stop**: the physical location where a vehicle stops to pick up or drop off passengers. It is associated to a unique *stop_id* and it has attributes *stop_name*, *stop_lat*, *stop_lon*, respectively the name and the geographic coordinates. (Example: 4025460, "PONT NEUF - QUAI DU LOUVRE", 48.858588, 2.340932)
- **route**: a public transportation line (in the following we refer to "line" or "route" as interchangeable terms) identified by a unique *route_ id*. It has attributes *route_ type*, identifying the type of vehicle, and *route_ name*. (Example : 831555, metro, "14"). Note that the two directions of a same service are identified by two different routes, and that services with multiple termini are identified by several different routes.
- **trip**: a journey of a vehicle, identified by a unique *trip_id*. It refers to the unique route of the actual line, and also to a set of dates indicating in which days of the year that trip is running. It is also associated to an ordered sequence of stops of the vehicle, and with the list of arrival and departure time at each stop. Example:

trip_id	stop_id	arrival_time	departure_time
1013644000942075	4025388	16:10:00	16:10:00
	4025390	16:11:00	16:11:00
	4025392	16:12:00	16:12:00
	4025393	16:13:00	16:13:00
	...	...	...

D.1.1 *Coarse graining network stops*

To model the transportation network, it was necessary to coarse grain the data by grouping nearby stops together. table D.2 summarises the information contained in each of the datasets before and after coarse-graining.

D.1.1.1 *Paris*

The transportation system described in the RATP dataset contains 11850 stops. Some of these stops closely located to each other can be functionally replaced by a single station via a careful merging method. In order to merge stops, we used the information provided in the GTFS dataset. Data provides the list of stop pairs that are located at a short distance from each other, allowing people to transfer walking,

Area	Stops	Routes	Train stops	Train routes	Tot stops after merging
Paris	11850	1058	494	169	5690
Toulouse	1913	106	59	31	1920
Nantes	3412	61	27	18	1038
Strasbourg	1330	53	31	17	601

Table D.2: **Main characteristics of the PT systems datasets.** For each urban agglomeration (Area), the table indicates the number of Bus, Metro, Tram and RER stops and routes before coarse graining (Stops, Routes), the number of train stops (Train stops) and routes (Train routes), the total number of stops after coarse-graining and merging the two datasets (Tot stops after merging).

from one route to a different one in a given amount of time (that is also given in the dataset). It is for example the case of main railway stations or big squares, where many stops are concentrated in a relatively small area. We merged corresponding stops according to the information provided by the RATP company on possible transfers, as well as bus stops located in front of each other at the two opposite sides of the same road. After coarse graining, the total number of stops for the RATP dataset was reduced to 4596.

In the SNCF dataset, there is a total number of 494 suburban railway stations. It is necessary to identify stops/train stations present both in the SNCF and RATP datasets (i.e "Gare du Nord" is both a RER station and a metro stop). To do so we built a grid with a resolution of 0.25 Km and we identified for each of the train stations the cell it belongs to. A train station was then identified by the closest RATP stops present in the actual cell or in neighbouring cells otherwise. In the city centre, all the train stations were identified with RATP stops, while in the suburbs it was not always the case.

D.1.1.2 *Nantes*

The Semitain dataset contains 3412 stops. It indicates for each stop whether it is part of a larger station complex (stops that are located on the opposite side of a same road are considered part of a unique station). Using such information, it was straightforward to merge close-by stops. Since transfer time was not provided, we estimated the time to change line based on the data provided by RATP (average transfer time, see table D.3). After coarse graining, the network includes 1036 stops. The SNCF dataset was used to include the train stations which are located in the area served by the Semitain company. Using the

same method we used for Paris, we found their corresponding stops in the Semitain dataset.

D.1.1.3 *Toulouse*

The Tisséo dataset contains 5694 stops. As in the case of Nantes, the Tisséo dataset provides information on parent stations. We merged stops accordingly received 1913 stops in total. Since transfer time was not provided, we estimated the time to change line based on the data provided by RATP (average transfer time, see table D.3). From the SNCF dataset, we selected 59 stops that located in the same area served by the Tisséo company.

D.1.1.4 *Strasbourg*

The CTS dataset contains 1330 stops. Even if it does not provide information on parent station, we could merge stops based on their *stop_id*. Since transfer time was not provided, we estimated the time to change line based on the data provided by RATP (average transfer time, see table D.3). In fact, in this dataset all stops that are part of a larger station complex have the same name and in addition a unique number (Example: stops {DANTE_01, DANTE_02, DANTE_03} are part of a same large station complex). After coarse graining this way 595 stops were identified in the CTS dataset. From the SNCF dataset, we selected 31 stops that are located in the same area served by the CTS company.

D.1.2 *Choice of a representative day*

The datasets provide the schedule over several months in normal situations (which means no perturbation due to traffic jams or to system breakdowns) with a 1-minute resolution. We do not consider exact travel time at a given departure time but an estimation of the time taken in a "typical" day. The description of a typical day is given below.

In order to draw typical commuting times we first selected a window of $N_w = 4$ consecutive weeks. A week $w_i = \{d_1, d_2, d_3, d_4, d_5\}$ is defined as a set of five consecutive days, from Monday to Friday. The separation week-end/week days is necessary as the system behaviour is different in these two cases. For every span of consecutive weeks $W = \{w_1, w_2, w_3, w_4\}$, we calculated the average daily number of trips $\langle Nt_W \rangle = \sum_{d \in W} Nt_d / D$. Here D is the number of days ($D = 5 \times 4 = 20$), Nt_d is the number of trips during day $d \in W$. Then, by looking at fluctuations from the average $\sigma_W^2 = \sum_d (Nt_d - \langle Nt_W \rangle)^2 / D$, we selected the four weeks span W for which σ_W^2 is the smallest. For each city the selected period is indicated in table D.1.

Mode 1	Mode 2	Average transfertime (sec)
bus	bus	70
subway	rail	326
tram	rail	222
rail	rail	60
subway	bus	230
tram	bus	92
subway	subway	172
rail	bus	232
tram	subway	212
tram	tram	66

Table D.3: **Transfer time** Average transfer time (in seconds) between different transportation modalities.

The reason to select a span of time where the number of trips is not fluctuating is motivated by the need to work with meaningful averaged quantities. We are aware that the results of the illustration may not generalise well, as they are relative to a specific selected period of time. Future work could include a comparison to the system behaviour during weekends, and at different times of the year.

For the purpose of this work, as we aimed at comparing our results with the flux of commuters, we limited the analysis to the 7-10am time interval. Indeed, as a first step, we selected all trips occurring between $h1 = 7am$ and $h2 = 10am$ within the selected period. Further work could include the study of the system evolution at different times of the day.

As a second step, we calculated for each route ℓ_k and each day $d \in W$ the total number $Nt_{\ell_k,d}$ of trips tr occurring on day d between 7 and 10 am and computed its average over the four selected weeks $\langle Nt_{\ell_k} \rangle = \sum_{d \in W} Nt_{\ell_k,d}/D$. In this way, we received the average frequency $f_{\ell_k} = \langle Nt_{\ell_k} \rangle / 3h$ (3h is the length of the time interval) in the selected period for each metro, bus or train line. Also, we computed in equivalent way, the average duration of a trip between any two stops i and j along line ℓ_k: $\langle \Delta t_{ij}^{\ell_k} \rangle = \sum_{tr} (\Delta t_{ij}^{\ell_k}) / \sum_{d \in W} Nt_{\ell_k,d}$ considering all selected trips tr.

In fig. D.1, we show the characteristics of the PT datasets for the 4 cities considered. We observe that stops are highly heterogeneous with respect to the number of routes in all the cities considered,

with few highly connected stops and a considerable number of stops served by only one or two routes. The number of stops per route vary greatly for all transportation modalities in Paris; for other cities only buses routes have more than 40 stops, and rail routes are relatively short (up to 10/15 stops). Also service frequencies have large variations depending on the transportation modality, with metro lines running considerably more frequently than other services (up to $\sim$ 45 times per hour), and rail services running at most $\sim$ 10 times per hour.

D.1.3 *The INSEE datasets*

In order to analyse commuting patterns, we gathered two datasets of the French Institute of Statistics (INSEE): the *Enquête Nationale Transports et Déplacements 2007-2008* [339] used for computing the commuting travelling times, and the 2010 French census (*Recensement de la population 2010*) [334] to extract origin-destination commuting patterns.

We used the file "Q_ind_lieu_teg.csv" of the **first dataset** providing for each individual several informations about their daily journey to work/school. We estimated the average time needed to commute a specific distance by car by scanning over the following variables *V1_BTRAVDIST*, i.e. the distance covered daily (resolution 1Km), *V1_BTRAVTEMPSA* i.e. the time needed to cover such distance (5 minutes resolution), and *V1_BTRAVMOYEN1S*, i.e. the transportation mean used. The time computed for a given distance is the time average over the trips with the same distance and travelled by car.

The flow of commuters for each origin-destination trip was estimated using the file "FD_MOBPRO_2010.txt" of the **second dataset**, in which each line provides several variables related to an individual interviewed. In particular, the following variables were needed: *COMMUNE* and *ARM*, respectively indicating the INSEE code associated to the municipality and the arrondissement (available only for central Paris) where the individual interviewed lives, *DCLT* the INSEE code indicating the municipality and the neighbourhood (only for Paris) of work, and *TRANS* referring to the transportation mean used to commute (either by foot, two-wheeler, car/camion/van, PT). We also considered the variable *IPOND* to take into account that, because not every single citizen is interviewed for the census, each individual has a statistical weight to infer a representative behaviour. tables D.4 and D.5 provide an overview on the data for each of the urban agglomerations considered for this study.

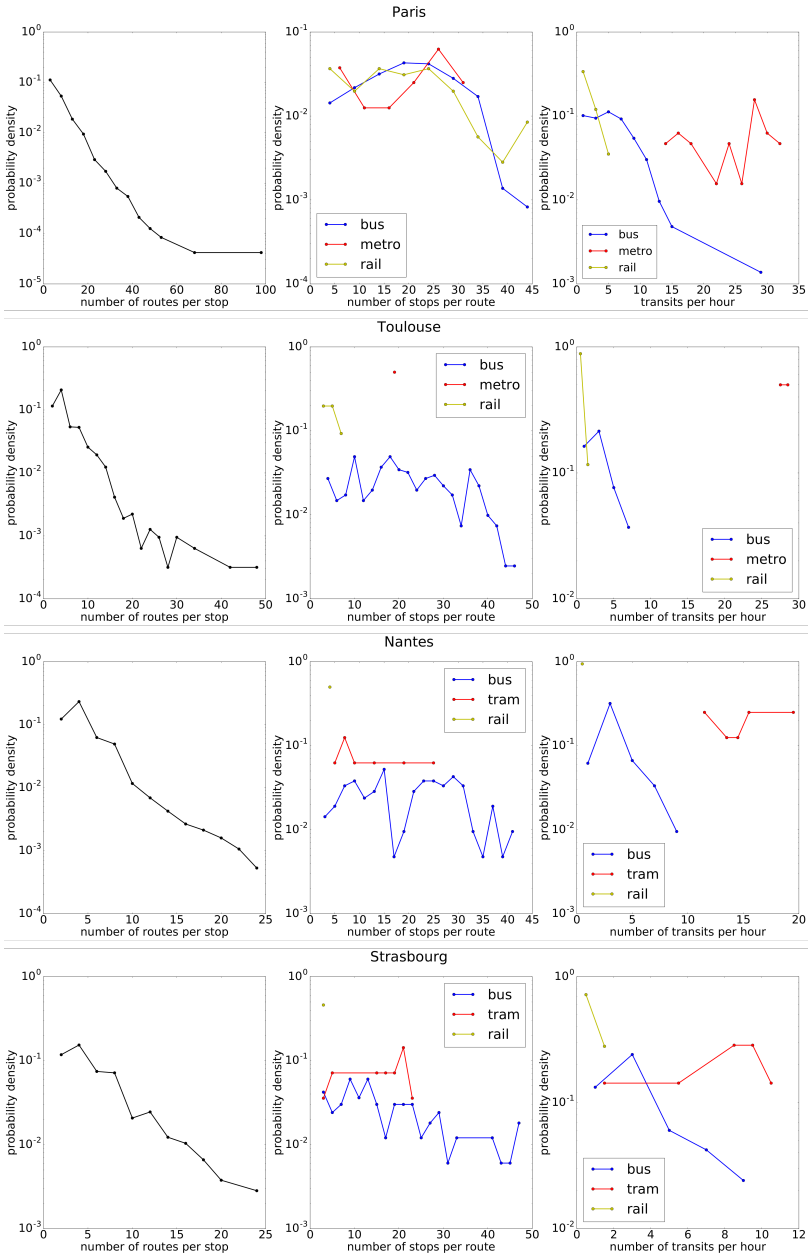

Figure D.1: **Characteristics of the PT datasets for the 4 cities considered.** For Paris, Toulouse, Nantes, and Strasbourg (top to bottom), we show the probability density distribution of the number of routes per stop after coarse-graining (left), the probability density distribution of the number of stops per line, for different transport modalities (center) and the probability density distribution of the number of transits per hour on each route, considering the period between 8 and 10 am (right).

Area	Mun	O-D pairs	Tot comm	Car comm	PT comm
Paris	460	61897	4321011	1542640	2017768
Toulouse	89	2319	363679	249642	57269
Strasbourg	57	618	170337	92275	36576
Nantes	26	524	225026	143441	45455

Table D.4: For each one of the urban areas considered (Area), the table provides with the number of municipalities considered (Mun), the number of origin-destination pairs travelled by commuters (O-D pairs), the total number of commuters (Tot comm), the number of commuters travelling by car (Car comm), the number of commuters travelling by PT (PT comm).

Area	IC comm	IC car comm	IC PT comm
Paris	1369535	390105	429735
Toulouse	187797	99142	40553
Strasbourg	99041	40579	23098
Nantes	111434	55444	25957

Table D.5: For each one of the urban areas considered (Area), the table provides with the total number of intra-city commuters (IC comm), the number of intra-city commuters using the car (IC car comm), the number of intra-city commuters using PT (IC PT comm).

D.1.4 *Matching the INSEE datasets and the PT datasets*

In order to establish a comparison between the commuting patterns and the efficient connections of the transportation systems, we matched the INSEE dataset with the PT data by associating to each of the stops in the PT data its corresponding municipality (or neighbourhood in the case of Paris). We made use of the Google Maps API [340] to assign to the latitude-longitude coordinates of each PT stop its corresponding address. Then, we matched the municipality to its corresponding INSEE code via the file *Base communale des aires urbaines 2010* provided by INSEE. [341]

D.2 STRUCTURE DETECTION WITH NON-NEGATIVE MATRIX FACTORISATION

In this section, we explain non-negative factorisation was achieved in order to extract structures from the transportation system dataset.

Algorithm

Aiming at minimising the Euclidean distance loss function between the original matrix and the factorized one, we implemented the standard multiplicative rule developed by Lee and Seung in [342]:

$$\mathbf{H}_{ci} \leftarrow \mathbf{H}_{ci} \frac{(\mathbf{W}^T\mathbf{V})_{ci}}{(\mathbf{W}^T\mathbf{W}\mathbf{H})_{ci}} \qquad \mathbf{W}_{ic} \leftarrow \mathbf{W}_{ic} \frac{(\mathbf{V}\mathbf{H}^T)_{ic}}{(\mathbf{W}\mathbf{H}\mathbf{H}^\mathbf{T})_{ci}}$$

Initialisation

The NMF algorithm may not converge to the same solution at each run, depending on the initial conditions. To address this problem we initialise the matrices $\mathbf{W}$ and $\mathbf{H}$ randomly and run the algorithm 500 times. At each iteration we compute the divergence $\|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2$ and we select the iteration for which the error was minimal.

In the present case, the algorithm turns out to be stable and the results are robust for large networks, future development of this work could however include the study of a consensus clustering procedure. Consensus clustering is the problem of reconciling clustering information about the same data set coming from different runs of the same algorithm. For NMF, some efforts have been done in this directions [343], however, as the result of the clustering is described through two different matrices and the partitioning is soft, the problem is not trivial to solve.

Soft/Hard partitioning

The results of NMF provide a soft clustering of the stops to the structures. Such information is included in matrices $\mathbf{W}$ and $\mathbf{H}$. For a given

node i and a given structure k, W_{ik} is the out-going affiliation of node i to structure k, while H_{ki} is the in-going affiliation. As the original matrix can be very sparse, and the NMF algorithm can hardly produce zero-values, many of the values in $\mathbf{W}$ and $\mathbf{H}$ are positive but very close to zero. In order to overcome this problem and to make sure we are capturing only the most relevant information, we applied a method to binaries the matrices $\mathbf{W}$ and $\mathbf{H}$ as follows: For each structure c, vectors $\mathbf{H_c}$ and $\mathbf{W}_c^T$ contain respectively the in-going and out-going affiliation of each node $i \in V$ to the structure c. With the goal of selecting only nodes that are strongly affiliated to c, we applied k-means clustering on these two vectors. k-means clustering partitions the $|V|$ affiliation values into k clusters. By choosing $k = 2$ for each of the structures c we distinguished a subset of not-affiliated nodes, whose affiliation value was very small, and a subset of affiliated nodes, whose affiliation value was significantly different from zero. Using this partitioning we defined a binary matrix $\mathbf{H}'$ such that $\mathbf{H'_{ci}} = 1$ if node i is in-going affiliated to community c and $\mathbf{H'_{ci}} = 0$ if it not. In the same way, we define $\mathbf{W}'$, for the out-going affiliation.

D.3 THE MODIFIED DIJKSTRA ALGORITHM

We devised a modified version of the Dijkstra algorithm allowing to compute approximated shortest paths in a weighted, labeled-edge graph. . The algorithm requires:

- A graph $G = (V, E, t_E, T, t_T)$ with vertex set V with cardinality N, edge set E with weight function t_E, and set of transfers T with weight function t_T
- A cut-off L_{max}(the maximal number of line changes allowed)

The algorithm returns:

- An array *dist* of length $N - 1$, where $dist[u]$ is the approximated shortest path length between nodes s and u
- an array Π_{node} of length $N - 1$, where $p = \Pi_{node}[u]$ is the *parent node* of node u, that precedes it in the shortest path between the source s and u itself
- the array of *parent edges* Π_{node}, of length $N - 1$, where $\Pi_{edge}[u]$ is the edge connecting u and its parent node p in the approximated shortest path connecting u and the source s

In the pseudo-code, the following notations are introduced: *lenPath* assigns to each vertex v the number of edges to reach source s, Q is a min-priority queue initialised with all nodes in V_G, where priority is given to nodes that are at shortest distance from the source s, $EXTRACT - MINQ$ is the operation of selecting and removing the

node with highest priority from Q, $e_{uv}^{\ell_k}$ is an edge in E connecting nodes u and v via line ℓ_k, and u is a neighbour of v if at least one of such connections exists and e_m is the edge connecting two nodes in the fastest way, also taking into account possible line transfers when coming from an other node, t_m is the associated time.

For each vertex $v \in V_G$
  $dist[v] = \infty$
  $\Pi_{node}[v] = NIL$
  $\Pi_{edge}[v] = NIL$
  $lenPath[v] = 0$
$dist[s] = 0$
$Q = V_G$
While $Q \neq \emptyset$
  $u = Extract - MinQ$
  For each v in neighbors u:
    If $\Pi_{edge}[v] == NIL$:
      $t_m, e_m = min, argmin(t_E(e_{uv}^{\ell_k}))$
    Else:
      $t_m, e_m = min, argmin(t_E(e_{uv}^{\ell_k}) + t_T(\Pi_{edge}(u), e_{uv}^{\ell_k}))$
    If $dist[v] > dist[u] + t_m$ AND $lenPath[u] + 1 <= L_{max}$
      $dist[v] = dist[u] + t_m$
      $\Pi_{node}[v] = u$
      $\Pi_{edge}[v] = e_m$
      $lenPath[v] = lenPath[u] + 1$
Return $dist, \Pi_{node}, \Pi_{edge}$

Figure D.2: Pseudo-code for the modified Dijkstra algorithm

Two main approximations are introduced in order to reduce the computation time, reduce the complexity of the PT system representation, and to account for individuals transportation strategy. We show that the approximations introduced do not affect the results on the overall efficiency of the PT systems.

- We limit the number of total line changes to L_{max} to account for individual choices of not changing line several times. This may lead to overestimate the shortest path lengths for paths with L_{max} changes. In this case, under a locally optimal strategy, one may change line before the best moment. As an example let's consider the following network:

  $(A, B, l1, w = 1)$
  $(B, C, l1, w = 2)$
  $(B, C, l2, w = 1)$
  $(C, D, l2, w = 3)$
  $(C, D, l3, w = 1)$
  The shortest path from A to D with $L_{max} = 1$ is $\{e_{A,B}^{l1}, e_{B,C}^{l1}, e_{C,D}^{l3}\}$

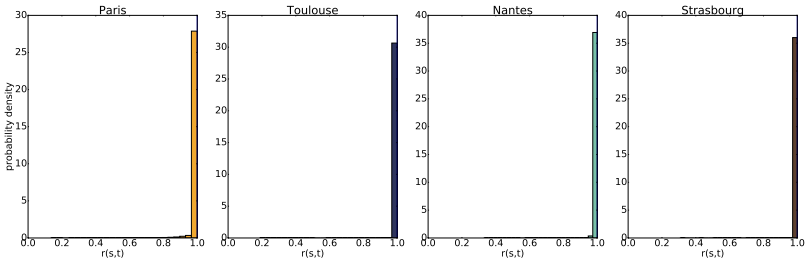

Figure D.3: **Comparison between the approximated and the traditional Dijkstra algorithm.** For each of the cities considered, we show the probability density of paths between nodes s and t with respect to the ratio $r(s,t)$. The 5% percentile is equal to one for all cities.

with total weight = 4. However, the algorithm will find the path $\{e^{l1}_{A,B}, e^{l2}_{B,C}, e^{l2}_{C,D}\}$ with total weight = 5.

- The representation includes labelled edges but not labelled nodes, to compromise between a complex description of the system and an efficient one. One way to resolve this issue would be to introduce transfers as links between labelled nodes, which would dramatically increase the network size. Instead, the algorithm includes the transferring time by adding its value to edge weights. The algorithm may overestimate the shortest path lengths in cases were the local optimal strategy of choosing the fastest transportation mean does not provide a globally optimal solution due to long transfer times. However, it provides a very good approximation when the transfer weights are small in comparison with edges weights (i.e. for long distances).

To quantify the impact of the approximations introduced, we calculate shortest paths for the 4 cities considered using both the approximated and the traditional version of the Dijkstra algorithm, where $L_{max} = \infty$, and both nodes and edges are labelled, resulting in a much larger network. For all pairs of nodes (s,t), such that a path with less than L_{max} changes exist, we compute the ratio $r(s,t) = dist_{correct}(s,t)/dist_{approx}(s,t)$, where $dist_{correct}(s,t)$ and $dist_{approx}(s,t)$ are the lengths of the shortest paths computed with the traditional and the approximated versions of the algorithm, respectively. We show the probability density distribution of $r(s,t)$ in fig. D.3. We find that in all cities the 95% of all paths have the same length in the two cases (see fig. D.3).

D.4 PATTERN DETECTION FOR STRASBOURG, NANTES, AND TOULOUSE

For the urban agglomerations of Strasbourg, Nantes, and Toulouse we detected structural patterns by considering intervals for distances with resolution of $d_2 - d_1 = 5$ kilometres. An example of structure

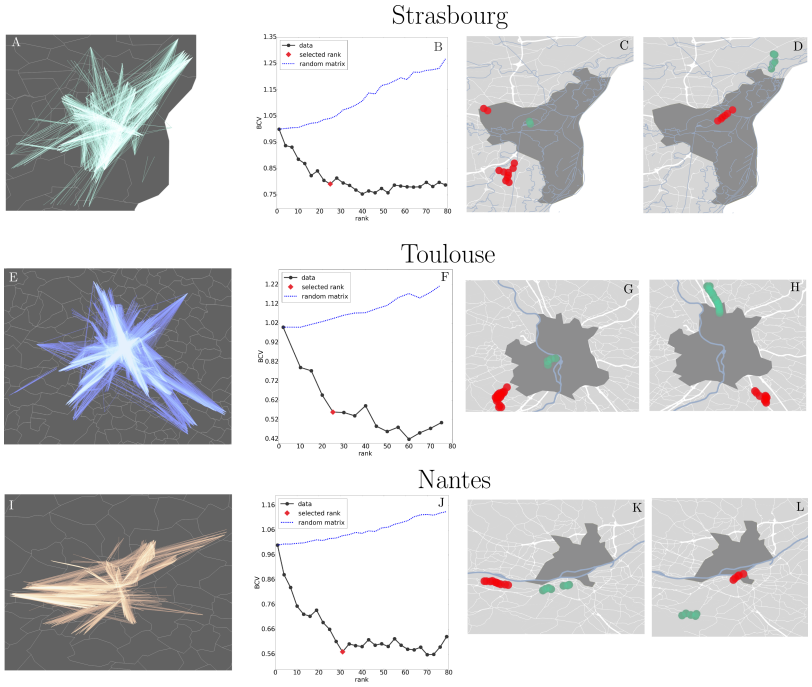

Figure D.4: **Pattern detection using the multi-edge P-space representation.** For Strasbourg, Toulouse, and Nantes, we show respectively in A,E and I the geographic representation of graph G_{SP}, where links correspond to the 1% best shortest paths of the public transportation network. In B, F and J, we show the normalised BiCross validation errors computed for the adjacency matrix $X_{SP}(0Km, 5Km)$ (grey full line) of the same graphs, for the associated random matrix $X_{SPrandom}(0Km, 5Km)$ (dashed line). The selected number of structures k_s is marked with a red rhombus. In C and D,G and H,K and L, two examples of structures revealed in the PT system are presented. Green dots are in-going, while red dots are out-going affiliated.

detected for each city is shown in fig. D.4. For an interval range $(d_1, d_2) = (5, 10)km$, both for Strasbourg and Nantes, we observed that the BiCross validation error computed for the adjacency matrix $X_{SP}(5Km, 10Km)$ is similar to the BiCross validation error of the associated random matrix $X_{SPrandom}(0Km, 5Km)$ (fig. D.7). This suggests that there is a lack of structure in the subgraph $G_{SP}(5Km, 10Km)$.

D.5 COMPARISON OF THE PATTERNS DETECTED AND THE COMMUTER FLOWS

We further investigate commuters behaviour, by identifying each pair of municipalities such that a flow of commuters exists between them and computing the corresponding PT-car flow ratio as the fraction of commuters using PT over the total people commuting between the two cities. We then compare the cases where the two municipalities are well (fig. D.8, A,C,E,G) or badly (fig. D.8, B,D,F,H) connected by

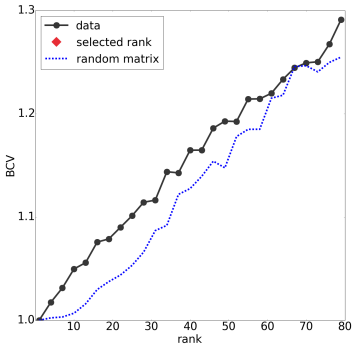

Figure D.5: Strasbourg,$d = 5 - 10Km$

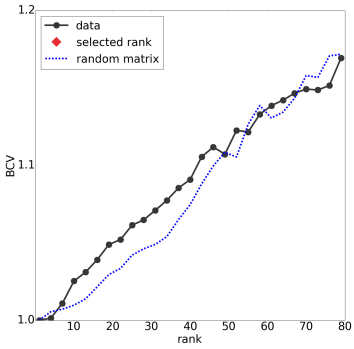

Figure D.6: Nantes,$d = 5 - 10Km$

Figure D.7: For the cities of Strasbourg (a), and Nantes (b), the normalised BiCross validation error computed for the adjacency matri $X_{SP}(5Km, 10Km)$ (grey full line) is similar to the BiCross validation error of the associated random matrix $X_{SPrandom}(0Km, 5Km)$ (dashed line). Therefore, no rank is selected and structures were not extracted.

PT according to our definition, considering the distribution of the PT-car flow ratio.
More precisely, we consider the PT structural pattern network $G_C = (V_C, E_C)$, and the commuter flow network $G^{TM}_{com} = (V^M_{com}, E^{TM}_{com}, W^M_{com})$, where $M = car$ or $M = PT$; first for each edge $(u, v) \in E_C$, we compute the fraction of commuters using PT, $f(u, v) = (W^{PT}_{com}(u, v) + W^{PT}_{com}(v, u))/(W^{PT}_{com}(u, v) + W^{PT}_{com}(v, u) + W^{car}_{com}(v, u) + W^{car}_{com}(v, u))$.Then, we compute the same quantity for all edges $(u, v) \in E_{com}$ that are not in E_C. For each city, we finally look at the distribution of $f(u, v)$ for both well and badly connected municipalities (fig. D.8).

In the case of Paris agglomeration, there is a significant difference between the case of privileged connections, where the distribution is left-side skewed (fig. D.8 A), and not privileged connections, where the distribution is more symmetrical (fig. D.8 A). This indicates that when the PT provides with good transportation according to our method, commuters prefer PT with respect to car. On the other hand, for Toulouse, Nantes and Strasbourg agglomerations, there is significantly less difference in the distribution of the PT-car flow ratios for well and badly connected pairs of cities. On the one hand, this may suggest commuters tend to use the car even where good connections are provided. On the other hand, we have to consider both that our selection was less strict for these cities, and that self loops (inter-city connections) may play an important contribution which could not be considered here due the resolution limit of the commuter dataset.

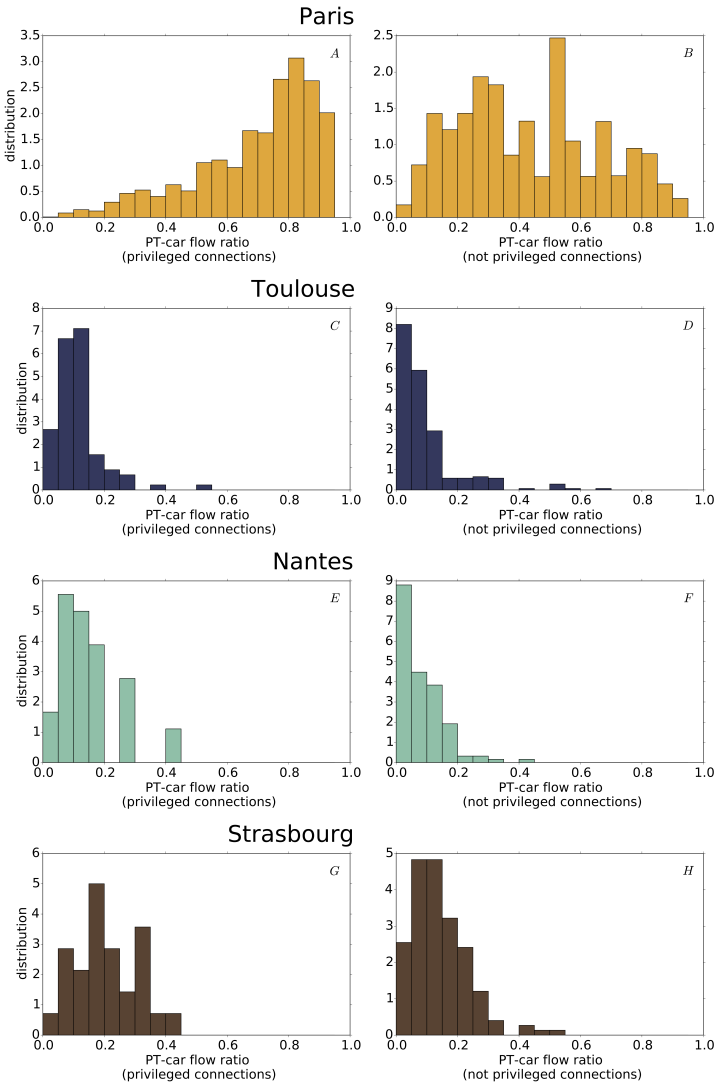

Figure D.8: For each city, we show the distribution of the PT-car flow ratio $f(u,v)$ when u and v are well (as defined in the main text) connected (A) or badly (the complementary connections) connected (B)

.

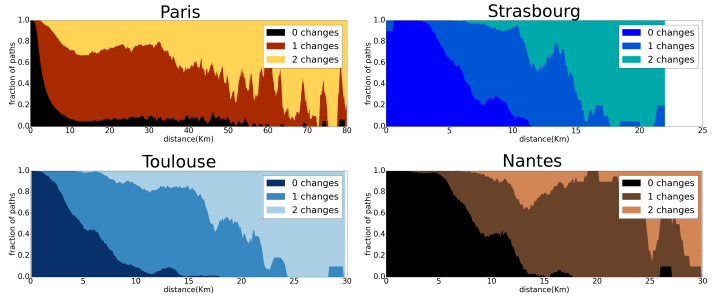

Figure D.9: We show for each city the fraction of privileged shortest paths with 0, 1, or 2 number of line changes in different colors, as a function of the shortest path distance.

.

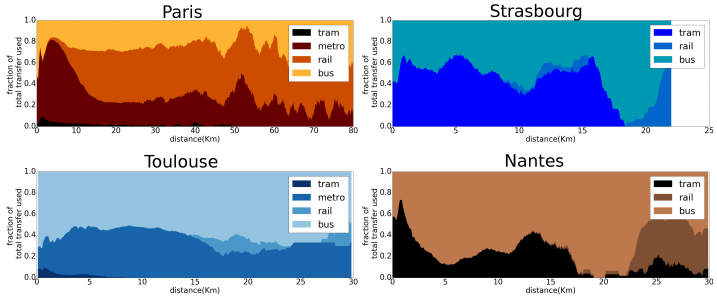

Figure D.10: For each city, we consider all edges occurring in privileged shortest paths. As a function of the shortest path distance, we show the fraction of edges according to their transportation modality.

.

D.6 CHARACTERISTICS OF PRIVILEGED CONNECTIONS

In this section we show some of the characteristics of the selected privileged connections. For each city, we show the distribution of the number of line changes (fig. D.9), of the number of different modalities in the same path (fig. D.10), and the occurrences of each possible transportation modality (fig. D.11), as a function of the shortest path distance.

D.7 COMPARISON WITH THE SINGLE-LAYER REPRESENTATIONS

The straightforward graph representation (fig. D.12), widely used for PT systems, where for each modality stops correspond to PT stops, and edges connect consecutive stops (connected by a vehicle without stopping between stops) does not allow to identify privileged connections and well-connected areas within the city.
We showed the city profile obtained considering privileged connections in the multilayer representation. Here, we compare the city profile with the one obtained considering the transportation modalities

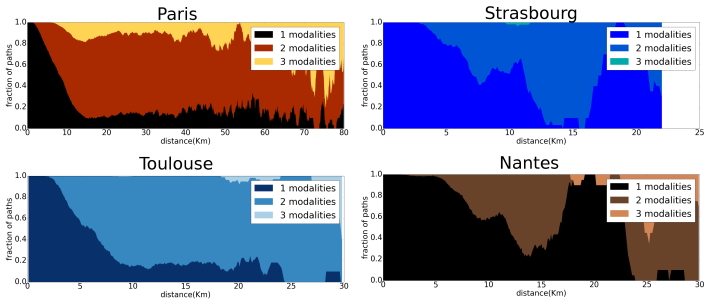

Figure D.11: We show for each city the fraction of privileged shortest paths including 1, 2, or 3 transportation modalities in different colors, as a function of the shortest path distance.

.

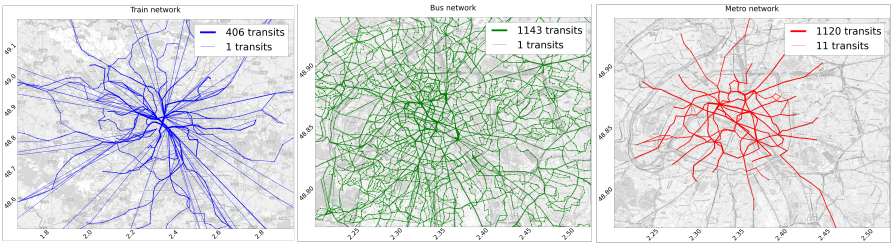

Figure D.12: Naive representation of the multi-layer PT system of Paris. For each modality (Tram, Bus, Metro, from left to right), dots correspond to nodes, and connecting edges have a thickness proportional to edge the number of transits per day

.

separately. We compute the shortest paths taking into account only buses, metro and rail connections, and we select privileged connections in the same way detailed for the entire multi-layer. We show on the same figure (fig. D.13, left) the profiles obtained for each transportation modality in Paris. The city single-layer profile differs from the one obtained considering all transportation modalities (fig. D.13, right) since the advantages due to the interconnectedness of several transportation modes are not accounted.

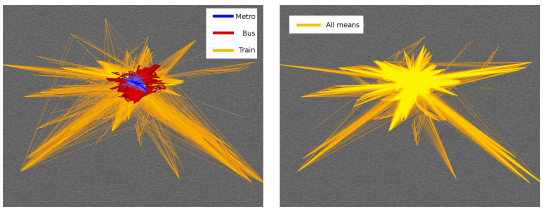

Figure D.13: Comparison between the city profile of Paris considering the single-modality single-layer representation (left), and the multi-layer representation (right).

.

BIBLIOGRAPHY

[1] Macmillan Dictionary. *Macmillan Dictionary*. 2015.

[2] Glenn Lyons and John Urry. "Travel time use in the information age." In: *Transportation Research Part A: Policy and Practice* 39.2-3 (2005), pp. 257–276.

[3] Patricia L Mokhtarian and Cynthia Chen. "TTB or not TTB, that is the question: a review and analysis of the empirical literature on travel time (and money) budgets." In: *Transportation Research Part A: Policy and Practice* 38.9-10 (2004), pp. 643–675.

[4] David L Greene. "Transportation and energy." In: *The geography of urban transportation* 3 (2004), pp. 274–293.

[5] Torsten Hägerstraand. "What about people in regional science?" In: *Papers in regional science* 24.1 (1970), pp. 7–24.

[6] Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. "A tale of one city: Using cellular network data for urban planning." In: *IEEE Pervasive Computing* 10.4 (2011), pp. 18–26.

[7] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. "The TimeGeo modeling framework for urban motility without travel surveys." In: *Proceedings of the National Academy of Sciences* (2016), p. 201524261.

[8] Michele Tizzoni, Paolo Bajardi, Adeline Decuyper, Guillaume Kon Kam King, Christian M Schneider, Vincent Blondel, Zbigniew Smoreda, Marta C González, and Vittoria Colizza. "On the use of human mobility proxies for modeling epidemics." In: *PLoS computational biology* 10.7 (2014), e1003716.

[9] Ernest George Ravenstein. "The laws of migration." In: *Journal of the statistical society of London* 48.2 (1885), pp. 167–235.

[10] Samuel A Stouffer. "Intervening opportunities: a theory relating mobility and distance." In: *American sociological review* 5.6 (1940), pp. 845–867.

[11] Helen Makower, Jacob Marschak, and Howard Waterhouse Robinson. "Studies in mobility of labour: a tentative statistical measure." In: *Oxford Economic Papers* 1 (1938), pp. 83–123.

[12] George Kingsley Zipf. "The P 1 P 2/D hypothesis: on the intercity movement of persons." In: *American sociological review* 11.6 (1946), pp. 677–686.

[13] Martin Beckmann, Charles B McGuire, and Christopher B Winsten. *Studies in the Economics of Transportation*. Tech. rep. 1956.

[14] George Esdras Bevans. *How workingmen spend their spare time*. Columbia University., 1913.

[15] George Andrew Lundberg, Mirra Komarovsky, and Mary Alice McInerny. *Leisure: A suburban study*. Columbia University Press, 1934.

[16] Pitirim Aleksandrovich Sorokin and Clarence Quinn Berger. *Time-budgets of human behavior*. Vol. 2. Not Avail, 1939.

[17] Gary S Becker. "A Theory of the Allocation of Time." In: *The economic journal* (1965), pp. 493–517.

[18] Thomas C Schelling. "Dynamic models of segregation." In: *Journal of mathematical sociology* 1.2 (1971), pp. 143–186.

[19] Henrik Jeldtoft Jensen. *Self-organized criticality: emergent complex behavior in physical and biological systems*. Vol. 10. Cambridge university press, 1998.

[20] Yaneer Bar-Yam. *Dynamics of complex systems*. Vol. 213. Addison-Wesley Reading, MA, 1997.

[21] Gregoire Nicolis, Ilya Prigogine, and G Nocolis. "Exploring complexity." In: (1989).

[22] Thomas F Golob, Abraham D Horowitz, and Martin Wachs. *Attitude-behavior relationships in travel demand modelling*. Tech. rep. 1977.

[23] David A. Hensher and Peter R. Stopher. *Behavioural Travel Modelling*. Croom Helm, 1979. ISBN: 9780856648199. URL: https://books.google.co.uk/books?id=jqEOAAAAQAAJ.

[24] Tijs Neutens, Tim Schwanen, and Frank Witlox. "The prism of everyday life: towards a new research agenda for time geography." In: *Transport reviews* 31.1 (2011), pp. 25–47.

[25] William Alonso. "A theory of movements." In: (1978).

[26] Pat Burnett. "The dimensions of alternatives in spatial choice processes." In: *Geographical Analysis* 5.3 (1973), pp. 181–204.

[27] Reginald G Golledge, Lawrence A Brown, and Frank Williamson. "Behavioural approaches in geography: an overview." In: *The Australian Geographer* 12.1 (1972), pp. 59–79.

[28] Dennis James Walmsley and Gareth J Lewis. *People and environment: Behavioural approaches in human geography*. Routledge, 2014.

[29] Paul Kelly, Aiden Doherty, Anja Mizdrak, S Marshall, J Kerr, A Legge, S Godbole, H Badland, Mark Oliver, and C Foster. "High group level validity but high random error of a self-report travel diary, as assessed by wearable cameras." In: *Journal of Transport & Health* 1.3 (2014), pp. 190–201.

[30] Jean Louise Wolf. "Using GPS data loggers to replace travel diaries in the collection of travel data." PhD thesis. School of Civil and Environmental Engineering, Georgia Institute of Technology, 2000.

[31] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. "Dynamics of person-to-person interactions from distributed RFID sensor networks." In: *PloS one* 5.7 (2010), e11596.

[32] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns." In: *Nature* 453.7196 (2008), pp. 779–782.

[33] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. "An Empirical Study of Geographic User Activity Patterns in Foursquare." In: *ICwSM* 11 (2011), pp. 70–573.

[34] Riccardo Gallotti, Armando Bazzani, Sandro Rambaldi, and Marc Barthelemy. "How transportation hierarchy shapes human mobility." In: *arXiv preprint arXiv:1509.03752* (2015).

[35] Eunjoon Cho, Seth A Myers, and Jure Leskovec. "Friendship and mobility: user movement in location-based social networks." In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, pp. 1082–1090.

[36] Riccardo Gallotti, Armando Bazzani, and Sandro Rambaldi. "Towards a statistical physics of human mobility." In: *International Journal of Modern Physics C* 23.09 (2012), p. 1250061.

[37] Lars Backstrom, Eric Sun, and Cameron Marlow. "Find me if you can: improving geographical prediction with social and spatial proximity." In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 61–70.

[38] Hugo Barbosa-Filho, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. "Human mobility: models and applications." In: *arXiv preprint arXiv:1710.00004* (2017).

[39] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. "A universal model for mobility and migration patterns." In: *Nature* 484.7392 (2012), pp. 96–100.

[40] Jaylson J Silveira, Aquino L Espíndola, and TJP Penna. "Agent-based model to rural–urban migration analysis." In: *Physica A: Statistical Mechanics and its Applications* 364 (2006), pp. 445–456.

[41] Hang-Hyun Jo, Jari Saramäki, Robin I. M. Dunbar, and Kimmo Kaski. *Dynamics of close relationships for the life-course migration*. July 2014. arXiv: 1407.4896. URL: http://arxiv.org/abs/1407.4896.

[42] Adrian M Tompkins and Nicky McCreesh. "Migration statistics relevant for malaria transmission in Senegal derived from mobile phone data and used in an agent-based migration model." In: *Geospatial health* 11.1s (2016).

[43] Samiul Hasan, Christian M Schneider, Satish V Ukkusuri, and Marta C González. "Spatiotemporal patterns of urban human mobility." In: *Journal of Statistical Physics* 151.1-2 (2013), pp. 304–318.

[44] Anastasios Noulas, Blake Shaw, Renaud Lambiotte, and Cecilia Mascolo. "Topological Properties and Temporal Dynamics of Place Networks in Urban Environments." In: *arXiv preprint arXiv:1502.07979* (2015).

[45] Andres Sevtsuk and Carlo Ratti. "Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks." In: *Journal of Urban Technology* 17.1 (2010), pp. 41–60.

[46] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. "Modelling the scaling properties of human mobility." In: *Nature Physics* 6.10 (2010), pp. 818–823.

[47] Serdar Çolak, Lauren P Alexander, Bernardo Guatimosim Alvim, Shomik R Mehndiretta, and Marta C González. "ANALYZING CELL PHONE LOCATION DATA FOR URBAN TRAVEL: CURRENT 2 METHODS, LIMITATIONS AND OPPORTUNITIES 3." In: *Transportation Research Board 94th Annual Meeting*. 15-5279. 2015.

[48] Gyan Ranjan, Hui Zang, Zhi-Li Zhang, and Jean Bolot. "Are call detail records biased for sampling human mobility?" In: *ACM SIGMOBILE Mobile Computing and Communications Review* 16.3 (2012), pp. 33–44.

[49] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. "On the levy-walk nature of human mobility." In: *IEEE/ACM transactions on networking (TON)* 19.3 (2011), pp. 630–643.

[50] Andrea Baronchelli and Filippo Radicchi. "Lévy flights in human behavior and cognition." In: *Chaos, Solitons & Fractals* 56 (2013), pp. 101–105.

[51] Kai Zhao, Mirco Musolesi, Pan Hui, Weixiong Rao, and Sasu Tarkoma. "Explaining the power-law distribution of human mobility through transportation modality decomposition." In: *Scientific reports* 5 (2015).

[52] Riccardo Gallotti, Armando Bazzani, Sandro Rambaldi, and Marc Barthelemy. "A stochastic model of randomly accelerated walkers for human mobility." In: *Nature Communications* 7 (2016), p. 12600.

[53] Xiao-Yong Yan, Xiao-Pu Han, Bing-Hong Wang, and Tao Zhou. "Diversity of individual mobility patterns and emergence of aggregated scaling laws." In: *Scientific reports* 3 (2013).

[54] Roberto Murcio, A. Paolo Masucci, Elsa Arcaute, and Michael Batty. "Multifractal to monofractal evolution of the London street network." In: *Phys. Rev. E* 92 (6 2015), p. 062130. DOI: 10.1103/PhysRevE.92.062130. URL: https://link.aps.org/doi/10.1103/PhysRevE.92.062130.

[55] Per Bak, Chao Tang, and Kurt Wiesenfeld. "Self-organized criticality: An explanation of the 1/f noise." In: *Physical review letters* 59.4 (1987), p. 381.

[56] Gandimohan M Viswanathan, Sergey V Buldyrev, Shlomo Havlin, Marcos GE Da Luz, EP Raposo, and H Eugene Stanley. "Optimizing the success of random searches." In: *Nature* 401.6756 (1999), pp. 911–914.

[57] Michael A Lomholt, Koren Tal, Ralf Metzler, and Klafter Joseph. "Lévy strategies in intermittent search processes are advantageous." In: *Proceedings of the National Academy of Sciences* 105.32 (2008), pp. 11055–11059.

[58] EP Raposo, SV Buldyrev, MGE Da Luz, GM Viswanathan, and HE Stanley. "Lévy flights and random searches." In: *Journal of Physics A: mathematical and theoretical* 42.43 (2009), p. 434003.

[59] MC Santos, D Boyer, O Miramontes, GM Viswanathan, EP Raposo, JL Mateos, and MGE Da Luz. "Origin of power-law distributions in deterministic walks: The influence of landscape geometry." In: *Physical Review E* 75.6 (2007), p. 061114.

[60] Balázs Cs Csáji, Arnaud Browet, Vincent A Traag, Jean-Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D Blondel. "Exploring the mobility of mobile phone users." In: *Physica A: Statistical Mechanics and its Applications* 392.6 (2013), pp. 1459–1473.

[61] Hui Zang and Jean Bolot. "Anonymization of location data does not work: A large-scale measurement study." In: *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM. 2011, pp. 145–156.

[62] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. "Location prediction in social media based on tie strength." In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM. 2013, pp. 459–468.

[63] Adam Sadilek, Henry Kautz, and Jeffrey P Bigham. "Finding your friends and following them to where you are." In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM. 2012, pp. 723–732.

[64] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. "Inferring social ties from geographic coincidences." In: *Proceedings of the National Academy of Sciences* 107.52 (2010), pp. 22436–22441.

[65] Przemyslaw A Grabowicz, Jose J Ramasco, Bruno Gonçalves, and Víctor M Eguíluz. "Entangling mobility and interactions in social media." In: *PloS one* 9.3 (2014), e92196.

[66] Jameson L Toole, Carlos Herrera-Yaqüe, Christian M Schneider, and Marta C González. "Coupling human mobility and social ties." In: *Journal of The Royal Society Interface* 12.105 (2015), p. 20141128.

[67] Petter. Holme and Jari Saramäki. "Temporal Networks." In: *Phys. Rep.* 519 (2012), p. 97.

[68] Petter Holme. "Modern temporal network theory: a colloquium." In: *The European Physical Journal B* 88.9 (2015), pp. 1–30.

[69] Naoki Masuda and Renaud Lambiotte. *A guide to temporal networks*. Vol. 4. World Scientific, 2016.

[70] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. "Measuring large-scale social networks with high resolution." In: *PloS one* 9.4 (2014), e95978.

[71] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. "Towards rich mobile phone datasets: Lausanne data collection campaign." In: *Proc. ICPS, Berlin* (2010).

[72] Nathan Eagle and Alex Sandy Pentland. "Reality mining: sensing complex social systems." In: *Personal and ubiquitous computing* 10.4 (2006), pp. 255–268.

[73] Andrea Cuttone, Sune Lehmann, and Jakob Eg Larsen. "Inferring human mobility from sparse low accuracy mobile sensing data." In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM. 2014, pp. 995–1004.

[74] Piotr Sapiezynski, Radu Gatej, Alan Mislove, and Sune Lehmann. "Opportunities and Challenges in Crowdsourced Wardriving." In: *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*. ACM. 2015, pp. 267–273.

[75] Oliver P John and Sanjay Srivastava. "The Big Five trait taxonomy: History, measurement, and theoretical perspectives." In: *Handbook of personality: Theory and research* 2.1999 (1999), pp. 102–138.

[76] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. "The mobile data challenge: Big data for mobile computing research." In: *Pervasive Computing*. EPFL-CONF-192489. 2012.

[77] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh Minh Tri Do, Olivier Dousse, Julien Eberle, and Markus Miettinen. "From big smartphone data to worldwide research: The mobile data challenge." In: *Pervasive and Mobile Computing* 9.6 (2013), pp. 752–771.

[78] Nathan Eagle. *The reality mining data*. 2010.

[79] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. "Returners and explorers dichotomy in human mobility." In: *Nature communications* 6 (2015).

[80] Maxime Lenormand, Thomas Louail, Oliva G Cantú-Ros, Miguel Picornell, Ricardo Herranz, Juan Murillo Arias, Marc Barthelemy, Maxi San Miguel, and José J Ramasco. "Influence of sociodemographic characteristics on human mobility." In: *arXiv preprint arXiv:1411.7895* (2014).

[81] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. "The scaling laws of human travel." In: *Nature* 439.7075 (2006), pp. 462–465.

[82] Xiang-Wen Wang, Xiao-Pu Han, and Bing-Hong Wang. "Correlations and scaling laws in human mobility." In: *PloS one* 9.1 (2014), e84954.

[83] Bin Jiang, Junjun Yin, and Sijian Zhao. "Characterizing the human mobility pattern in a large street network." In: *Physical Review E* 80.2 (2009), p. 021136.

[84] Yu Liu, Chaogui Kang, Song Gao, Yu Xiao, and Yuan Tian. "Understanding intra-urban trip patterns from taxi trajectory data." In: *Journal of geographical systems* 14.4 (2012), pp. 463–483.

[85] Mariano G Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. "Predicting human mobility through the assimilation of social media traces into mobility models." In: *arXiv preprint arXiv:1601.04560* (2016).

[86] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. "Exploring Millions of Footprints in Location Sharing Services." In: *ICWSM* 2011 (2011), pp. 81–88.

[87] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. "Geo-located Twitter as proxy for global mobility patterns." In: *Cartography and Geographic Information Science* 41.3 (2014), pp. 260–271.

[88] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. "A tale of many cities: universal patterns in human urban mobility." In: *PloS one* 7.5 (2012), e37027.

[89] Lun Wu, Ye Zhi, Zhengwei Sui, and Yu Liu. "Intra-urban human mobility and activity transition: evidence from social media check-in data." In: *PloS one* 9.5 (2014), e97010.

[90] Yu Liu, Zhengwei Sui, Chaogui Kang, and Yong Gao. "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data." In: *PloS one* 9.1 (2014), e86026.

[91] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. "Understanding human mobility from Twitter." In: *PloS one* 10.7 (2015), e0131469.

[92] Hsing Liu, Ying-Hsing Chen, and Jiann-Shing Lih. "Crossover from exponential to power-law scaling for human mobility pattern in urban, suburban and rural areas." In: *The European Physical Journal B* 88.5 (2015), pp. 1–7.

[93] Xiao Liang, Xudong Zheng, Weifeng Lv, Tongyu Zhu, and Ke Xu. "The scaling of human mobility by taxis is exponential." In: *Physica A: Statistical Mechanics and its Applications* 391.5 (2012), pp. 2135–2144.

[94] Li Gong, Xi Liu, Lun Wu, and Yu Liu. "Inferring trip purposes and uncovering travel patterns from taxi trajectory data." In: *Cartography and Geographic Information Science* 43.2 (2016), pp. 103–114.

[95] Kai Zhao, MP Chinnasamy, and Sasu Tarkoma. "Automatic City Region Analysis for Urban Routing." In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE. 2015, pp. 1136–1142.

[96] Wenjun Wang, Lin Pan, Ning Yuan, Sen Zhang, and Dong Liu. "A comparative analysis of intra-city human mobility by taxi." In: *Physica A: Statistical Mechanics and its Applications* 420 (2015), pp. 134–147.

[97] Jinjun Tang, Fang Liu, Yinhai Wang, and Hua Wang. "Uncovering urban human mobility from large scale taxi GPS data." In: *Physica A: Statistical Mechanics and its Applications* 438 (2015), pp. 140–153.

[98] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. "Unravelling daily human mobility motifs." In: *Journal of The Royal Society Interface* 10.84 (2013), p. 20130246.

[99] Armando Bazzani, Bruno Giorgini, Sandro Rambaldi, Riccardo Gallotti, and Luca Giovannini. "Statistical laws in urban mobility from microscopic GPS data in the area of Florence." In: *Journal of Statistical Mechanics: Theory and Experiment* 2010.05 (2010), P05001.

[100] Tuhin Paul, Kevin Stanley, Nathaniel Osgood, Scott Bell, and Nazeem Muhajarine. "Scaling Behavior of Human Mobility Distributions." In: *International Conference on Geographic Information Science*. Springer. 2016, pp. 145–159.

[101] Adeline Decuyper, Arnaud Browet, Vincent Traag, Vincent D Blondel, and Jean-Charles Delvenne. "Clean up or mess up: the effect of sampling biases on measurements of degree distributions in mobile phone datasets." In: *arXiv preprint arXiv:1609.09413* (2016).

[102] Mikko Kivelä and Mason A Porter. "Estimating interevent time distributions from finite observation periods in communication networks." In: *Physical Review E* 92.5 (2015), p. 052813.

[103] Pierre Deville, Chaoming Song, Nathan Eagle, Vincent D Blondel, Albert-László Barabási, and Dashun Wang. "Scaling identity connects human mobility and social interactions." In: *Proceedings of the National Academy of Sciences* 113.26 (2016), pp. 7047–7052.

[104] Camille Roth, Soong Moon Kang, Michael Batty, and Marc Barthélemy. "Structure of urban movements: polycentric activity and entangled hierarchical flows." In: *PloS one* 6.1 (2011), e15923.

[105] Can-Zhong Yao and Ji-Nan Lin. "A study of human mobility behavior dynamics: A perspective of a single vehicle with taxi." In: *Transportation Research Part A: Policy and Practice* 87 (2016), pp. 51–58.

[106] Riccardo Gallotti, Armando Bazzani, and Sandro Rambaldi. "Understanding the variability of daily travel-time expenditures using GPS trajectory data." In: *EPJ Data Science* 4.1 (2015), p. 1.

[107] Eric-Jan Wagenmakers and Simon Farrell. "AIC model selection using Akaike weights." In: *Psychonomic bulletin & review* 11.1 (2004), pp. 192–196.

[108] Sergei Petrovskii, Alla Mashanova, and Vincent AA Jansen. "Variation in individual walking behavior creates the impression of a Lévy flight." In: *Proceedings of the National Academy of Sciences* 108.21 (2011), pp. 8704–8707.

[109] Xiao-Pu Han, Qiang Hao, Bing-Hong Wang, and Tao Zhou. "Origin of the scaling law in human mobility: Hierarchy of traffic systems." In: *Physical Review E* 83.3 (2011), p. 036117.

[110] Michael Mitzenmacher. "A brief history of generative models for power law and lognormal distributions." In: *Internet mathematics* 1.2 (2004), pp. 226–251.

[111] Hideaki Mouri. "Log-normal distribution from a process that is not multiplicative but is additive." In: *Physical Review E* 88.4 (2013), p. 042124.

[112] P Van Mieghem, N Blenn, and C Doerr. "Lognormal distribution in the digg online social network." In: *The European Physical Journal B* 83.2 (2011), pp. 251–261.

[113] Norbert Blenn and Piet Van Mieghem. "Are human interactivity times lognormal?" In: *arXiv preprint arXiv:1607.02952* (2016).

[114] Márton Karsai, Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, A-L Barabási, and Jari Saramäki. "Small but slow world: How network topology and burstiness slow down spreading." In: *Physical Review E* 83.2 (2011), p. 025102.

[115] Hang-Hyun Jo, Márton Karsai, János Kertész, and Kimmo Kaski. "Circadian pattern and burstiness in mobile phone communication." In: *New Journal of Physics* 14.1 (2012), p. 013055.

[116] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D Blondel, and Jari Saramäki. "Effects of time window size and placement on the structure of an aggregated communication network." In: *EPJ Data Science* 1.1 (2012), p. 1.

[117] Mario Gutiérrez-Roig, Oleguer Sagarra, Aitana Oltra, Frederic Bartumeus, Albert Diaz-Guilera, and Josep Perelló. "Active and reactive behaviour in human mobility: the influence of attraction points on pedestrians." In: *arXiv preprint arXiv:1511.03604* (2015).

[118] Chiara Poletto, Michele Tizzoni, and Vittoria Colizza. "Human mobility and time spent at destination: impact on spatial epidemic spreading." In: *Journal of theoretical biology* 338 (2013), pp. 41–58.

[119] Lawrence D Burns. "Transportation, temporal, and spatial components of accessibility." In: (1980).

[120] Tim Schwanen, Mei-Po Kwan, and Fang Ren. "How fixed is fixed? Gendered rigidity of space–time constraints and geographies of everyday activities." In: *Geoforum* 39.6 (2008), pp. 2109–2121.

[121] Chloë Brown, Neal Lathia, Cecilia Mascolo, Anastasios Noulas, and Vincent Blondel. "Group colocation behavior in technological social networks." In: *PloS one* 9.8 (2014), e105816.

[122] Halgurt Bapierre, Chakajkla Jesdabodi, and Georg Groh. "Mobile Homophily and Social Location Prediction." In: *arXiv preprint arXiv:1506.07763* (2015).

[123] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. "Unveiling the complexity of human mobility by querying and mining massive trajectory data." In: *The VLDB Journal—The International Journal on Very Large Data Bases* 20.5 (2011), pp. 695–719.

[124] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. "Nextplace: a spatio-temporal prediction framework for pervasive systems." In: *Pervasive Computing*. Springer, 2011, pp. 152–169.

[125] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. "Identifying important places in people's lives from cellular network data." In: *Pervasive computing*. Springer, 2011, pp. 133–151.

[126] James P Bagrow and Yu-Ru Lin. "Mesoscopic structure and social aspects of human mobility." In: *PloS one* 7.5 (2012), e37676.

[127] Gueorgi Kossinets and Duncan J Watts. "Empirical analysis of an evolving social network." In: *science* 311.5757 (2006), pp. 88–90.

[128] Gueorgi Kossinets and Duncan J Watts. "Origins of homophily in an evolving social network 1." In: *American journal of sociology* 115.2 (2009), pp. 405–450.

[129] Daniel Mauricio Romero, Brendan Meeder, Vladimir Barash, and Jon Kleinberg. "Maintaining ties on social media sites: The competing effects of balance, exchange, and betweenness." In: *Fifth International AAAI Conference on Weblogs and Social Media*. 2011.

[130] John Levi Martin and King-To Yeung. "Persistence of close personal ties over a 12-year period." In: *Social Networks* 28.4 (2006), pp. 331–362.

[131] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. "Limited communication capacity unveils strategies for human interaction." In: *Scientific reports* 3 (2013).

[132] Jari Saramäki, EA Leicht, Eduardo López, Sam GB Roberts, Felix Reed-Tsochas, and Robin IM Dunbar. "Persistence of social signatures in human communication." In: *Proceedings of the National Academy of Sciences* 111.3 (2014), pp. 942–947.

[133] Ronald S Burt. "Decay functions." In: *Social networks* 22.1 (2000), pp. 1–28.

[134] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin Dunbar. "Dynamics of personal social relationships in online social networks: a study on twitter." In: *Proceedings of the first ACM conference on Online social networks*. ACM. 2013, pp. 15–26.

[135] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. "Human mobility modeling at metropolitan scales." In: *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM. 2012, pp. 239–252.

[136] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. "Slaw: A new mobility model for human walks." In: *INFOCOM 2009, IEEE*. IEEE. 2009, pp. 855–863.

[137] Minkyong Kim, David Kotz, and Songkuk Kim. "Extracting a Mobility Model from Real User Traces." In: *INFOCOM*. Vol. 6. 2006, pp. 1–13.

[138] Tao Jia, Bin Jiang, Kenneth Carling, Magnus Bolin, and Yifang Ban. "An empirical study on human mobility and its agent-based modeling." In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.11 (2012), P11024.

[139] Luca Pappalardo, Salvatore Rinzivillo, and Filippo Simini. "Human Mobility Modelling: Exploration and Preferential Return Meet the Gravity Model." In: *Procedia Computer Science* 83 (2016), pp. 934 –939. ISSN: 1877-0509. DOI: http://dx.doi.org/10.1016/j.procs.2016.04.188. URL: http://www.sciencedirect.com/science/article/pii/S1877050916302216.

[140] Nathan Eagle, Alex Sandy Pentland, and David Lazer. "Inferring friendship network structure by using mobile phone data." In: *Proceedings of the national academy of sciences* 106.36 (2009), pp. 15274–15278.

[141] Harold Stanley Heaps. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc., 1978.

[142] Frank E Horton and David R Reynolds. "Effects of urban spatial structure on individual behavior." In: *Economic Geography* 47.1 (1971), pp. 36–48.

[143] Mary Ellen Mazey. "The effect of a physio-political barrier upon urban activity space." In: (1981).

[144] Reginald G Golledge. *Spatial behavior: A geographic perspective*. Guilford Press, 1997.

[145] Yihong Yuan and Martin Raubal. "Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study." In: *International Journal of Geographical Information Science* 30.8 (2016), pp. 1594–1621.

[146] Jill E Sherman, John Spencer, John S Preisser, Wilbert M Gesler, and Thomas A Arcury. "A suite of methods for representing activity space in a healthcare accessibility study." In: *International journal of health geographics* 4.1 (2005), p. 24.

[147] Hugo Barbosa, Fernando B de Lima-Neto, Alexandre Evsukoff, and Ronaldo Menezes. "The effect of recency to human mobility." In: *EPJ Data Science* 4.1 (2015), p. 21.

[148] Michael Szell, Roberta Sinatra, Giovanni Petri, Stefan Thurner, and Vito Latora. "Understanding mobility in a social petri dish." In: *Scientific reports* 2 (2012).

[149] Robin IM Dunbar. "Coevolution of neocortical size, group size and language in humans." In: *Behavioral and brain sciences* 16.04 (1993), pp. 681–694.

[150] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. "Modeling users' activity on twitter networks: Validation of dunbar's number." In: *PloS one* 6.8 (2011), e22656.

[151] Kay W Axhausen. "Activity spaces, biographies, social networks and their welfare gains and externalities: some hypotheses and empirical results." In: *Mobilities* 2.1 (2007), pp. 15–36.

[152] Paul T Costa and Robert R McCrae. "Four ways five factors are basic." In: *Personality and individual differences* 13.6 (1992), pp. 653–665.

[153] Yuval Kalish and Garry Robins. "Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure." In: *Social Networks* 28.1 (2006), pp. 56–84.

[154] Thomas V Pollet, Sam GB Roberts, and Robin IM Dunbar. "Extraverts have larger social network layers." In: *Journal of Individual Differences* (2011).

[155] Alistair Sutcliffe, Robin Dunbar, Jens Binder, and Holly Arrow. "Relationships and the social brain: integrating psychological and evolutionary perspectives." In: *British journal of psychology* 103.2 (2012), pp. 149–168.

[156] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. "Analysis of ego network structure in online social networks." In: *Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (SocialCom)*. IEEE. 2012, pp. 31–40.

[157] Giovanna Miritello, Esteban Moro, Rubén Lara, Rocío Martínez-López, John Belchamber, Sam GB Roberts, and Robin IM Dunbar. "Time as a limited resource: Communication strategy in mobile phone networks." In: *Social Networks* 35.1 (2013), pp. 89–95.

[158] W-X Zhou, Didier Sornette, Russell A Hill, and Robin IM Dunbar. "Discrete hierarchical organization of social group sizes." In: *Proceedings of the Royal Society of London B: Biological Sciences* 272.1561 (2005), pp. 439–444.

[159] Cameron Marlow, L Byron, T Lento, and I Rosenn. "Maintained relationships on Facebook." In: *Retrieved February* 15.8 (2009).

[160] Sam GB Roberts, Robin IM Dunbar, Thomas V Pollet, and Toon Kuppens. "Exploring variation in active network size: Constraints and ego characteristics." In: *Social Networks* 31.2 (2009), pp. 138–146.

[161] Simone Centellegher, Eduardo López, Jari Saramäki, and Bruno Lepri. "Personality traits and ego-network dynamics." In: *PloS one* 12.3 (2017), e0173110.

[162] Stefan Wehrli et al. "Personality on social network sites: An application of the five factor model." In: *Zurich: ETH Sociology (Working Paper No. 7)* (2008).

[163] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. "Structure and evolution of online social networks." In: *Link mining: models, algorithms, and applications*. Springer, 2010, pp. 337–357.

[164] Mark EJ Newman. "Clustering and preferential attachment in growing networks." In: *Physical review E* 64.2 (2001), p. 025102.

[165] Alan Mislove, Hema Swetha Koppula, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. "Growth of the flickr social network." In: *Proceedings of the first workshop on Online social networks*. ACM. 2008, pp. 25–30.

[166] Robin IM Dunbar and Matt Spoors. "Social networks, support cliques, and kinship." In: *Human nature* 6.3 (1995), pp. 273–290.

[167] Tasuku Igarashi, Jiro Takai, and Toshikazu Yoshida. "Gender differences in social network development via mobile phone text messages: A longitudinal study." In: *Journal of Social and Personal Relationships* 22.5 (2005), pp. 691–713.

[168] Cornelia Wrzus, Martha Hänel, Jenny Wagner, and Franz J Neyer. "Social network changes and life events across the life span: A meta-analysis." In: *Psychological bulletin* 139.1 (2013), p. 53.

[169] Juan Carrasco, Eric Miller, and Barry Wellman. "How far and with whom do people socialize?: Empirical evidence about distance between social network members." In: *Transportation Research Record: Journal of the Transportation Research Board* 2076 (2008), pp. 114–122.

[170] Pauline van den Berg, Theo Arentze, and Harry Timmermans. "Size and composition of ego-centered social networks and their effect on geographic distance and contact frequency." In: *Transportation Research Record: Journal of the Transportation Research Board* 2135 (2009), pp. 1–9.

[171] Karen E Campbell, Peter V Marsden, and Jeanne S Hurlbert. "Social resources and socioeconomic status." In: *Social networks* 8.1 (1986), pp. 97–117.

[172] Miller McPherson, Lynn Smith-Lovin, and Matthew E Brashears. "Social isolation in America: Changes in core discussion networks over two decades." In: *American sociological review* 71.3 (2006), pp. 353–375.

[173] Harry T Reis, John Nezlek, and Ladd Wheeler. "Physical attractiveness in social interaction." In: *Journal of Personality and Social Psychology* 38.4 (1980), p. 604.

[174] James J Jaccard. "Predicting social behavior from personality traits." In: *Journal of Research in Personality* 7.4 (1974), pp. 358–367.

[175] Walter Mischel. "Toward a cognitive social learning reconceptualization of personality." In: *Psychological review* 80.4 (1973), p. 252.

[176] Sam GB Roberts, Ruth Wilson, Pawel Fedurek, and RIM Dunbar. "Individual differences and personal social network size and structure." In: *Personality and individual differences* 44.4 (2008), pp. 954–964.

[177] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. "Friends don't lie: inferring personality traits from social network structure." In: *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM. 2012, pp. 321–330.

[178] Yu-En Lu, Sam Roberts, Pietro Lio, Robin Dunbar, and Jon Crowcroft. "Size matters: variation in personal network size, personality and effect on information transmission." In: *Computational Science and Engineering, 2009. CSE'09. International Conference on*. Vol. 4. IEEE. 2009, pp. 188–193.

[179] Jens B Asendorpf and Marcel AG Van Aken. "Personality–relationship transaction in adolescence: Core versus surface personality characteristics." In: *Journal of personality* 71.4 (2003), pp. 629–666.

[180] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Pentland. "Predicting Personality Using Novel Mobile Phone-Based Metrics." In: *SBP*. Springer. 2013, pp. 48–55.

[181] Maarten Selfhout, William Burk, Susan Branje, Jaap Denissen, Marcel Van Aken, and Wim Meeus. "Emerging late adolescent friendship networks and Big Five personality traits: A social network approach." In: *Journal of personality* 78.2 (2010), pp. 509–538.

[182] Jens Asendorpf and Jaap JA Denissen. "Predictive validity of personality types versus personality dimensions from early childhood to adulthood: Implications for the distinction between core and surface traits." In: *Merrill-Palmer Quarterly* 52.3 (2006), pp. 486–513.

[183] Susan JT Branje, Cornelis FM van Lieshout, and Marcel AG van Aken. "Relations between Big Five personality characteristics and perceived support in adolescents' families." In: *Journal of personality and social psychology* 86.4 (2004), p. 615.

[184] Olle Järv, Rein Ahas, and Frank Witlox. "Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records." In: *Transportation Research Part C: Emerging Technologies* 38 (2014), pp. 122–135.

[185] Stefan Schönfelder and Kay W Axhausen. *Urban rhythms and travel behaviour: spatial and temporal phenomena of daily travel*. Ashgate Publishing, Ltd., 2010.

[186] Mei-Po Kwan. "Gender differences in space-time constraints." In: *Area* 32.2 (2000), pp. 145–156.

[187] Gonzalo M Vazquez-Prokopec, Donal Bisanzio, Steven T Stoddard, Valerie Paz-Soldan, Amy C Morrison, John P Elder, Jhon Ramirez-Paredes, Eric S Halsey, Tadeusz J Kochel, Thomas W Scott, et al. "Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment." In: *PloS one* 8.4 (2013), e58802.

[188] Chaogui Kang, Song Gao, Xing Lin, Yu Xiao, Yihong Yuan, Yu Liu, and Xiujun Ma. "Analyzing and geo-visualizing individual human mobility patterns using mobile call records." In: *Geoinformatics, 2010 18th International Conference on*. IEEE. 2010, pp. 1–7.

[189] Shannon N Zenk, Amy J Schulz, Stephen A Matthews, Angela Odoms-Young, JoEllen Wilbur, Lani Wegrzyn, Kevin Gibbs, Carol Braunschweig, and Carmen Stokes. "Activity space environment and dietary and physical activity behaviors: a pilot study." In: *Health & place* 17.5 (2011), pp. 1150–1161.

[190] Mei-Po Kwan and Jiyeong Lee. "Geovisualization of human activity patterns using 3D GIS: a time-geographic approach." In: *Spatially integrated social science* 27 (2004).

[191] Veronique Van Acker, Bert Van Wee, and Frank Witlox. "When transport geography meets social psychology: toward a conceptual model of travel behaviour." In: *Transport Reviews* 30.2 (2010), pp. 219–240.

[192] Martin J Chorley, Roger M Whitaker, and Stuart M Allen. "Personality and location-based social networks." In: *Computers in Human Behavior* 46 (2015), pp. 45–56.

[193] Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. "Geographic constraints on social network groups." In: *PLoS one* 6.4 (2011), e16939.

[194] David Jurgens. "That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships." In: *ICWSM* 13.13 (2013), pp. 273–282.

[195] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. "Human mobility, social ties, and link prediction." In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, pp. 1100–1108.

[196] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. "Exploiting place features in link prediction on location-based social networks." In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, pp. 1046–1054.

[197] Huy Pham, Cyrus Shahabi, and Yan Liu. "Ebm: an entropy-based model to infer social strength from spatiotemporal data." In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM. 2013, pp. 265–276.

[198] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. "Bridging the gap between physical location and online social networks." In: *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM. 2010, pp. 119–128.

[199] Thomas T Hills, Peter M Todd, David Lazer, A David Redish, Iain D Couzin, Cognitive Search Research Group, et al. "Exploration versus exploitation in space, mind, and society." In: *Trends in cognitive sciences* 19.1 (2015), pp. 46–54.

[200] Jonathan D Cohen, Samuel M McClure, and J Yu Angela. "Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1481 (2007), pp. 933–942.

[201] Tiziana Casciaro. "Seeing things clearly: Social structure, personality, and accuracy in social network perception." In: *Social Networks* 20.4 (1998), pp. 331–351.

[202] Mohammadreza Hojat. "Loneliness as a function of selected personality variables." In: *Journal of Clinical Psychology* 38.1 (1982), pp. 137–141.

[203] Paul R Amato. "Personality and social network involvement as predictors of helping behavior in everyday life." In: *Social Psychology Quarterly* (1990), pp. 31–43.

[204] Yair Amichai Hamburger and Elisheva Ben-Artzi. "The relationship between extraversion and neuroticism and the different uses of the Internet." In: *Computers in human behavior* 16.4 (2000), pp. 441–449.

[205] Ulrike Grömping et al. "Relative importance for linear regression in R: the package relaimpo." In: *Journal of statistical software* 17.1 (2006), pp. 1–27.

[206] Anders Mollgaard, Sune Lehmann, and Joachim Mathiesen. "Correlations between human mobility and social interaction reveal general activity patterns." In: *PloS one* 12.12 (2017), e0188973.

[207] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis." In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[208] Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. "Kernel PCA and de-noising in feature spaces." In: *Advances in neural information processing systems*. 1999, pp. 536–542.

[209] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

[210] Robert R McCrae and Oliver P John. "An introduction to the five-factor model and its applications." In: *Journal of personality* 60.2 (1992), pp. 175–215.

[211] Renaud Lambiotte and Michal Kosinski. "Tracking the digital footprints of personality." In: *Proceedings of the IEEE* 102.12 (2014), pp. 1934–1939.

[212] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. "Daily stress recognition from mobile phone data, weather conditions and individual traits." In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 477–486.

[213] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. "Mining large-scale smartphone data for personality studies." In: *Personal and Ubiquitous Computing* 17.3 (2013), pp. 433–450.

[214] Stuart C Aitken. "Person-environment theories in contemporary perceptual and behavioural geography I: personality, attitudinal and spatial choice theories." In: *Progress in Human Geography* 15.2 (1991), pp. 179–193.

[215] Gordon Willard Allport. "Personality: A psychological interpretation." In: (1937).

[216] Mark E J Newman. *Networks. An Introduction*. Oxford Univesity Press, 2010.

[217] Matthew O Jackson et al. *Social and economic networks*. Vol. 3. Princeton university press Princeton, 2008.

[218] Bruno Gonçalves and Nicola Perra. *Social phenomena: From data analysis to models*. Springer, 2015.

[219] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge Univesity Press, 2008.

[220] Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. "Activity driven modeling of time-varying networks." In: *Scientific Reports* 2 (2012), p. 469.

[221] Marton Karsai, Nicola Perra, and Alessandro Vespignani. "Time varying networks and the weakness of strong ties." In: *Scientific Reports* 4 (2014), p. 4001.

[222] Enrico Ubaldi, Nicola Perra, Marton Karsai, Alessandro Vezzani, Raffaella Burioni, and Alessandro Vespignani. "Asymptotic theory of time-varying social networks with heterogeneous activity and tie allocation." In: *Scientific Reports* 6 (2016).

[223] Guillaume Laurent, Jari Saramäki, and Márton Karsai. "From calls to communities: a model for time-varying social networks." In: *The European Physical Journal B* 88.11 (2015), pp. 1–10.

[224] Giovanna Miritello, Esteban Moro, and Rubén Lara. "Dynamical strength of social ties in information spreading." In: *Physical Review E* 83.4 (Apr. 2011), p. 045102. DOI: 10.1103/PhysRevE.83.045102. URL: http://dx.doi.org/10.1103/PhysRevE.83.045102.

[225] Aaron Clauset and Nathan Eagle. "Persistence and periodicity in a dynamic proximity network." In: *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*. 2007, pp. 1–5.

[226] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter Van den Broeck. "What's in a crowd? Analysis of face-to-face behavioral networks." In: *J. Theor. Biol* 271 (2011), p. 166.

[227] Jari Saramäki and Esteban Moro. "From seconds to months: an overview of multi-scale dynamics of mobile telephone calls." In: *The European Physical Journal B* 88.6 (2015), pp. 1–10.

[228] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. "Fundamental structures of dynamic social networks." In: *Proceedings of the National Academy of Sciences* 113.36 (2016), pp. 9977–9982. DOI: 10.1073/pnas.1602803113. eprint: http://www.pnas.org/content/113/36/9977.full.pdf. URL: http://www.pnas.org/content/113/36/9977.abstract.

[229] Mario V Tomasello, Nicola Perra, Claudio J. Tessone, Marton Karsai, and Frank Schweitzer. "The role of endogenous and exogenous mechanisms in the formation of R&D networks." In: *Scientific reports* 4 (2014).

[230] Alain Barrat and Ciro Cattuto. "Face-to-face interactions." In: *Social Phenomena*. Springer International Publishing, 2015, pp. 37–57.

[231] N. Perra, A. Baronchelli, D Mocanu, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. "Random walks and search in time varying networks." In: *Phys. Rev. Lett.* 109 (2012), p. 238701.

[232] Bruno Ribeiro, Nicola Perra, and Andrea Baronchelli. "Quantifying the effect of temporal resolution on time-varying networks." In: *Scientific Reports* 3 (2013), p. 3006.

[233] Suyu Liu, Nicola Perra, Márton Karsai, and Alessandro Vespignani. "Controlling Contagion Processes in Activity Driven Networks." In: *Phys. Rev. Lett.* 112 (11 2014), p. 118702. DOI: 10.1103/PhysRevLett.112.118702. URL: http://link.aps.org/doi/10.1103/PhysRevLett.112.118702.

[234] Su-Yu Liu, Andrea Baronchelli, and Nicola Perra. "Contagion dynamics in time-varying metapopulation networks." In: *Phys. Rev. E* 87 (3 2013), p. 032805. DOI: 10.1103/PhysRevE.87.032805. URL: http://link.aps.org/doi/10.1103/PhysRevE.87.032805.

[235] Guangming Ren and Xingyuan Wang. "Epidemic spreading in time-varying community networks." In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 24.2, 023116 (2014), pp. –. DOI: http://dx.doi.org/10.1063/1.4876436. URL: http://scitation.aip.org/content/aip/journal/chaos/24/2/10.1063/1.4876436.

[236] M. Starnini, A. Machens, C. Cattuto, A. Barrat, and R. Pastor-Satorras. "Immunization strategies for epidemic processes in time-varying contact networks." In: *Journal of Theoretical Biology* 337 (2013), pp. 89–100.

[237] Michele Starnini, Andrea Baronchelli, Alain Barrat, and Romualdo Pastor-Satorras. "Random walks on temporal networks." In: *Phys. Rev. E* 85 (5 2012), p. 056115.

[238] Eugenio Valdano, Luca Ferreri, Chiara Poletto, and Vittoria Colizza. "Analytical computation of the epidemic threshold on temporal networks." In: *Physical Review X* 5.2 (2015), p. 021005.

[239] Ingo Scholtes, Nicolas Wider, Rene Pfitzner, Antonios Garas, Claudo J Tessone, and Frank Schweitzer. "Causality driven slow-down and speed-up of diffusion in non-Markovian temporal networks." In: *Nature communications* 5 (2014).

[240] Matthew J. Williams and Mirco Musolesi. "Spatio-temporal networks: reachability, centrality and robustness." In: *Royal Society Open Science* 3.6 (2016). DOI: 10.1098/rsos.160196.

[241] Luis EC Rocha and Naoki Masuda. "Random walk centrality for temporal networks." In: *New Journal of Physics* 16.6 (2014), p. 063023.

[242] Taro Takaguchi, Nobuo Sato, Kazuo Yano, and Naoki Masuda. "Importance of individual events in temporal networks." In: *New Journal of Physics* 14.9 (2012), p. 093003.

[243] Luis EC Rocha and Vincent D Blondel. "Bursts of vertex activation and epidemics in evolving networks." In: *PLoS Comput Biol* 9.3 (2013), e1002974.

[244] Gourab Ghoshal and Petter Holme. "Attractiveness and activity in Internet communities." In: *Physica A: Statistical Mechanics and its Applications* 364 (2006), pp. 603–609.

[245] Kaiyuan Sun, Andrea Baronchelli, and Nicola Perra. "Contrasting effects of strong ties on SIR and SIS processes in temporal networks." In: *The European Physical Journal B* 88.12 (2015), pp. 1–8.

[246] René Pfitzner, Ingo Scholtes, Antonios Garas, Claudio J Tessone, and Frank Schweitzer. "Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks." In: *Physical review letters* 110.19 (2013), p. 198701.

[247] T. Takaguchi, N. Sato, K. Yano, and N. Masuda. "Importance of individual events in temporal networkss." In: *New J. Phys.* 14 (2012), p. 093003.

[248] Taro Takaguchi, Naoki Masuda, and Petter Holme. "Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics." In: *PloS one* 8.7 (2013), e68629.

[249] Petter Holme and Fredrik Liljeros. "Birth and death of links control disease spreading in empirical contact networks." In: *Scientific reports* 4 (2014).

[250] Petter Holme and Naoki Masuda. "The basic reproduction number as a predictor for epidemic outbreaks in temporal networks." In: *PloS one* 10.3 (2015), e0120567.

[251] Mikko Kivela, Raj Kumar Pan, Kimmo Kaski, Janos Kertesz, Jari Saramaki, and Marton Karsai. "Multiscale Analysis of Spreading in a Large Communication Network." In: *J. Stat. Mech.* 03005 (2012).

[252] Till Hoffmann, Mason A Porter, and Renaud Lambiotte. "Generalized master equations for non-Poisson dynamics on networks." In: *Physical Review E* 86.4 (2012), p. 046102.

[253] Zhen Wang, Chris T Bauch, Samit Bhattacharyya, Alberto d?Onofrio, Piero Manfredi, Matjaž Perc, Nicola Perra, Marcel Salathé, and Dawei Zhao. "Statistical physics of vaccination." In: *Physics Reports* (2016).

[254] Julie Fournet and Alain Barrat. "Contact patterns among high school students." In: *PloS one* 9.9 (2014), e107878.

[255] Antoine Moinet, Michele Starnini, and Romualdo Pastor-Satorras. "Burstiness and aging in social temporal networks." In: *Physical review letters* 114.10 (2015), p. 108701.

[256] Marton Karsai, Kimmo Kaski, Albert-L Barabási, and Janos Kertész. "Universal features of correlated bursty behavior." In: *Scientific Reports* (2012), p. 397.

[257] Enrico Ubaldi, Alessandro Vezzani, Marton Karsai, Nicola Perra, and Raffaella Burioni. "Burstiness and tie reinforcement in time varying social networks." In: *Scientific Reports* 7.46225 (2017).

[258] Jukka-P. Onnela, Jorkki Saramaki Jari .and Hyvonen, Gabor Szabo, David Lazer, Kimmo Kaski, Janos Kertesz, and Albert-L. Barabasi. "Structure and tie strengths in mobile communication networks." In: *Proceedings of the National Academy of Sciences* 104.18 (2007), pp. 7332–7336. URL: http://www.pnas.org/content/104/18/7332.abstract.

[259] Mark Granovetter. "The strength of weak ties." In: *Am. J. Sociol.* 78 (1973), pp. 1360–1380.

[260] Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. "Modeling human dynamics of face-to-face interaction networks." In: *Physical Review Letters* 110 (2013), p. 168701.

[261] Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. "Model reproduces individual, group and collective dynamics of human contact networks." In: *Social Networks* 47 (2016), p. 130.

[262] Manuel Sebastian Mariani, Matúš Medo, and Yi-Cheng Zhang. "Ranking nodes in growing networks: When PageRank fails." In: *Scientific reports* 5 (2015).

[263] Albert-Laszlo Barabasi. *Network science*. Cambridge University Press, 2016.

[264] Renaud Lambiotte, Lionel Tabourier, and Jean-Charles Delvenne. "Burstiness and spreading on temporal networks." In: *The European Physical Journal B* 86.7 (2013), pp. 1–4.

[265] Renaud Lambiotte, Vsevolod Salnikov, and Martin Rosvall. "Effect of memory on the dynamics of random walks on networks." In: *Journal of Complex Networks* 3.2 (2015), pp. 177–188.

[266] Michele Starnini, Mattia Frasca, and Andrea Baronchelli. "Emergence of metapopulations and echo chambers in mobile agents." In: *Scientific Reports* 6 (2016), p. 31834.

[267] Daniel T Gillespie. "Exact stochastic simulation of coupled chemical reactions." In: *The journal of physical chemistry* 81.25 (1977), pp. 2340–2361.

[268] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. "On the evolution of user interaction in facebook." In: *Proceedings of the 2nd ACM workshop on Online social networks*. ACM. 2009, pp. 37–42.

[269] *Facebook wall posts network dataset – KONECT*. Oct. 2016. URL: http://konect.uni-koblenz.de/networks/facebook-wosn-wall.

[270] *Linux kernel mailing list replies network dataset – KONECT*. Oct. 2016. URL: http://konect.uni-koblenz.de/networks/lkml-reply.

[271] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. "The architecture of complex weighted networks." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.11 (2004), pp. 3747–3752.

[272] Jae D Noh and Heiko Rieger. "Random Walks on Complex Networks." In: *Phys. Rev. Lett.* 92 (2004), p. 118701.

[273] Alessandro Vespignani. "Modeling dynamical processes in complex socio-technical systems." In: *Nature Physics* 8 (2012), pp. 32–30.

[274] Sidney Redner. *A Guide To First-Passage Processes*. Cambridge: Cambridge University Press, 2001.

[275] Andrea Baronchelli and Vittorio Loreto. "Ring structures and mean first passage time in networks." In: *Physical Review E* 73.2 (2006), p. 026103.

[276] Albert-Laszlo Barabasi. "The origin of bursts and heavy tails in human dynamics." In: *Nature* 435.7039 (2005), pp. 207–211.

[277] Kwang-I. Goh and Albert L Barabási. "Burstiness and memory in complex systems." In: *EPL (Europhysics Letters)* 81.4 (2008), p. 48002. URL: http://stacks.iop.org/0295-5075/81/i=4/a=48002.

[278] Alexei Vázquez, João Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. "Modeling bursts and heavy tails in human dynamics." In: *Phys. Rev. E* 73 (3 2006), p. 036127. DOI: 10.1103/PhysRevE.73.036127. URL: http://link.aps.org/doi/10.1103/PhysRevE.73.036127.

[279] Márton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. "Universal features of correlated bursty behaviour." In: *Sci. Rep.* 2 (May 2012). URL: http://dx.doi.org/10.1038/srep00397.

[280] Márton Karsai, Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, A-L Barabási, and Jari Saramäki. "Small but slow world: How network topology and burstiness slow down spreading." In: *Phys. Rev. E* 83 (2 2011), p. 025102. DOI: 10.1103/PhysRevE.83.025102. URL: http://link.aps.org/doi/10.1103/PhysRevE.83.025102.

[281] Márton Karsai, Kimmo Kaski, and János Kertész. "Correlated Dynamics in Egocentric Communication Networks." In: *PLoS ONE* 7.7 (July 2012), e40612. DOI: 10.1371/journal.pone.0040612. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0040612.

[282] Antoine Moinet, Michele Starnini, and Romualdo Pastor-Satorras. "Burstiness and Aging in Social Temporal Networks." In: *Phys. Rev. Lett.* 114 (10 2015), p. 108701. DOI: 10.1103/PhysRevLett.114.108701. URL: http://link.aps.org/doi/10.1103/PhysRevLett.114.108701.

[283] *The Dynamics of Complex Urban Systems: An Interdisciplinary Approach*. 2008th ed. Physica, Dec. 2007. ISBN: 3790819360. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/3790819360.

[284] Yosef Sheffi. *Urban transportation networks*. Prentice-Hall, Englewood Cliffs, NJ, 1985.

[285] Michael G. H. Bell and Yasunori Iida. *Transportation Network Analysis*. 1st ed. Wiley, Apr. 1997. ISBN: 047196493X. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/047196493X.

[286] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. "Small-world properties of the Indian railway network." In: *Phys. Rev. E* 67 (Mar. 2003), p. 036106. DOI: 10.1103/PhysRevE.67.036106. URL: http://link.aps.org/doi/10.1103/PhysRevE.67.036106.

[287] Christopher Kennedy, Eric Miller, Amer Shalaby, Heather Maclean, and Jesse Coleman. "The Four Pillars of Sustainable Urban Transportation." In: *Transport Reviews* 25.4 (July 2005), pp. 393–414. ISSN: 0144-1647. DOI: 10.1080/01441640500115835. URL: http://dx.doi.org/10.1080/01441640500115835.

[288] Christoph E. Mandl. "Evaluation and optimization of urban public transportation networks." In: *European Journal of Operational Research* 5.6 (1980), pp. 396–404. DOI: http://dx.doi.org/10.1016/0377-2217(80)90126-5. URL: http://www.sciencedirect.com/science/article/pii/0377221780901265.

[289] Jayanth R. Banavar, Amos Maritan, and Andrea Rinaldo. "Size and form in efficient transportation networks." In: *Nature* 399.6732 (May 1999), pp. 130–132. ISSN: 0028-0836. DOI: 10.1038/20144. URL: http://dx.doi.org/10.1038/20144.

[290] John Bates, John Polak, Peter Jones, and Andrew Cook. "The valuation of reliability for personal travel." In: *Transportation Research Part E: Logistics and Transportation Review* 37.2-3 (Apr. 2001), pp. 191–229. ISSN: 13665545. DOI: 10.1016/s1366-5545(00)00011-9. URL: http://dx.doi.org/10.1016/s1366-5545(00)00011-9.

[291] Malachy Carey. "Optimizing scheduled times, allowing for behavioural response." In: *Transportation Research Part B: Methodological* 32.5 (June 1998), pp. 329–342. ISSN: 01912615. DOI: 10.

1016/s0191-2615(97)00039-8. URL: http://dx.doi.org/10.1016/s0191-2615(97)00039-8.

[292] Niels Van Oort and Rob van Nes. "Regularity analysis for optimizing urban transit network design." In: *Public Transport* 1.2 (2009), pp. 155–168.

[293] P. Lambert J.H.and Sarda. "Terrorism Scenario Identification by Superposition of Infrastructure Networks." In: *Journal of Infrastructure Systems* 11.4 (2005), pp. 211–220.

[294] *TCRP Synthesis 115 - Transportation Research Board*. http://onlinepubs.trb.org/Onlinepubs/tcrp/tcrp_syn_115.pdf. Accessed: 2016-02-16.

[295] Marie-Pier Pelletier, Martin Trépanier, and Catherine Morency. "Smart card data use in public transit: A literature review." In: *Transportation Research Part C: Emerging Technologies* 19.4 (2011), pp. 557–568.

[296] Peter Gregory Furth, Brendon J Hemily, T Muller, and James G Strathman. *Uses of archived AVL-APC data to improve transit performance and management: Review and potential*. Transportation Research Board Washington, DC, USA, 2003.

[297] John Steenbruggen, Maria Teresa Borzacchiello, Peter Nijkamp, and Henk Scholten. "Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities." In: *GeoJournal* 78.2 (2013), pp. 223–243.

[298] *Google Transit Data Feed*. https://code.google.com/archive/p/googletransitdatafeed/wikis/PublicFeeds.wiki. Accessed: 2016-02-16.

[299] *General Transit Feed Spec Changes*. https://groups.google.com/forum/forum/gtfs-changes. Accessed: 2016-02-16.

[300] Yilin Zhao. "Mobile phone location determination and its impact on intelligent transportation systems." In: *Intelligent Transportation Systems, IEEE Transactions on* 1.1 (2000), pp. 55–64.

[301] Ahmed M El-Geneidy, Jessica Horning, and Kevin J Krizek. "Analyzing transit service reliability using detailed data from automatic vehicular locator systems." In: *Journal of Advanced Transportation* 45.1 (2011), pp. 66–79.

[302] Yindong Shen, Jia Xu, and Zhongyi Zeng. "Public transit planning and scheduling based on AVL data in China." In: *International Transactions in Operational Research* (2015).

[303] Mahmoud Mesbah, Johnny Lin, and Graham Currie. ""Weather" transit is reliable? Using AVL data to explore tram performance in Melbourne, Australia." In: *Journal of Traffic and Transportation Engineering (English Edition)* 2.3 (2015), pp. 125–135.

[304] E Mazloumi, G Currie, and M Sarvi. "Assessing measures of transit travel time variability and reliability using AVL data." In: *Transportation Research Board Annual Meeting, 87th, 2008, Washington, DC, USA*. 2008.

[305] Sybil Derrible and Christopher Kennedy. "Network Analysis of World Subway Systems Using Updated Graph Theory." In: *Transportation Research Record: Journal of the Transportation Research Board* 2112 (2009), pp. 17–25. DOI: 10.3141/2112-03. eprint: http://dx.doi.org/10.3141/2112-03. URL: http://dx.doi.org/10.3141/2112-03.

[306] Julian Sienkiewicz and Janusz A. Holyst. "Statistical analysis of 22 public transport networks in Poland." In: *Phys. Rev. E* 72 (Oct. 2005), p. 046127. DOI: 10.1103/PhysRevE.72.046127. URL: http://link.aps.org/doi/10.1103/PhysRevE.72.046127.

[307] David Levinson. "Network Structure and City Size." In: *PLoS ONE* 7.1 (2012). DOI: 10.1371/journal.pone.0029721. URL: http://dx.doi.org/10.1371\%2Fjournal.pone.0029721.

[308] Rémi Louf, Camille Roth, and Marc Barthelemy. "Scaling in Transportation Networks." In: *PLoS ONE* 9.7 (2014). DOI: 10.1371/journal.pone.0102007. URL: http://dx.doi.org/10.1371\%2Fjournal.pone.0102007.

[309] Remi Louf and Marc Barthelemy. "How congestion shapes cities: from mobility patterns to scaling." In: 4 (July 2014). DOI: 10.1038/srep05561. URL: http://dx.doi.org/10.1038/srep05561.

[310] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. "Multilayer networks." In: *Journal of Complex Networks* 2.3 (Sept. 2014), pp. 203–271. ISSN: 2051-1329. DOI: 10.1093/comnet/cnu016. URL: http://dx.doi.org/10.1093/comnet/cnu016.

[311] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I. del Genio, Jesus Gómez-Gardeñes, Miguel Romance, Irene Sendiña Nadal, Zhen Wang, and Massimiliano Zanin. "The structure and dynamics of multilayer networks." In: *Physics Reports* 544.1 (Nov. 2014), pp. 1–122. ISSN: 03701573. DOI: 10.1016/j.physrep.2014.07.001. URL: http://dx.doi.org/10.1016/j.physrep.2014.07.001.

[312] Dimitrios Tsiotas and Serafeim Polyzos. "Decomposing multilayer transportation networks using complex network analysis: a case study for the Greek aviation network." In: *Journal of Complex Networks* (Feb. 2015), cnv003+. ISSN: 2051-1329. DOI: 10.1093/comnet/cnv003. URL: http://dx.doi.org/10.1093/comnet/cnv003.

[313] Riccardo Gallotti and Marc Barthelemy. "The multilayer temporal network of public transport in Great Britain." In: *Scientific Data* 2 (Jan. 2015), pp. 140056+. ISSN: 2052-4463. DOI: 10.1038/sdata.2014.56. arXiv: 1501.02159. URL: http://dx.doi.org/10.1038/sdata.2014.56.

[314] Petter Holme and Jari Saramäki. "Temporal networks." In: *Physics Reports* (2012).

[315] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Second Edition*. 2nd. The MIT Press, Sept. 2001. ISBN: 0262032937. URL: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0262032937.

[316] T. Asperges, E. Cornelis, T. Steenberger, and Others. "Déterminants des choix modaux dans les chaines de déplacements [Determinants of transport mode choice.]" In: *Résumé, Plan d'Appui scientifique à une politique de Développement Durable (PADD II), [Summary, Plan of scientific support to a a sustainable development policy]* 1 (2007).

[317] T. Le Jeanic, J. Armoogum, E. Bouffard-Savary, and Others. "La mobilité des Français, panorama issu de l'enquête nationale transports et déplacements 2008. [Mobility in France, an overview from the national survey of transsport and mobility 2008]." In: *Paris: ministère de l'Écologie, du Développement durable, des Transports et du Logement [The Ministry for Ecology, Sustainable Development, Transport and Housing]* (2010).

[318] Felipe E. Lillo Viedma. "Coloured-edge graph approach for the modelling of multimodal networks." PhD thesis. Auckland University of Technology, 2011.

[319] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. "Small-world properties of the Indian railway network." In: *Physical Review E* 67.3 (2003), p. 036106.

[320] Julian Sienkiewicz and Janusz A. Hołyst. "Statistical analysis of 22 public transport networks in Poland." In: *Physical Review E* 72.4 (2005), p. 046127.

[321] C. Von Ferber, T. Holovatch, Yu Holovatch, and V. Palchykov. "Public transport networks: empirical analysis and modeling." In: *The European Physical Journal B-Condensed Matter and Complex Systems* 68.2 (2009), pp. 261–275.

[322] JK Jolliffe and TP Hutchinson. "A behavioural explanation of the association between bus and passenger arrivals at a bus stop." In: *Transportation Science* 9.3 (1975), pp. 248–282.

[323] Daniel Csikos and Graham Currie. "Investigating Consistency in Passenger Arrivals–Insights from Longitudinal Ticket Validations." In: *Conference of Australian Institue of Transport Research (CAITR), 29th, 2007, Adelaide, South Australia, Australia*. 2007.

[324] S Chang and Chun-Lin Hsu. "Modeling passenger waiting time for intermodal transit stations." In: *Transportation Research Record: Journal of the Transportation Research Board* 1753 (2001), pp. 69–75.

[325] *Enquête Nationale Transports et Déplacements 2007-2008*. 2008.

[326] Ioannis Psorakis, Stephen Roberts, Mark Ebden, and Ben Sheldon. "Overlapping community detection using bayesian non-negative matrix factorization." In: *Physical Review E* 83.6 (2011), p. 066114.

[327] Fei Wang, Tao Li, Xin Wang, Shenghuo Zhu, and Chris Ding. "Community discovery using nonnegative matrix factorization." In: *Data Mining and Knowledge Discovery* 22.3 (2011), pp. 493–521.

[328] Shihua Zhang, Rui-Sheng Wang, and Xiang-Sun Zhang. "Uncovering fuzzy community structure in complex networks." In: *Physical Review E* 76.4 (2007), p. 046103.

[329] Ioannis Psorakis, Stephen Roberts, and Ben Sheldon. "Efficient bayesian community detection using non-negative matrix factorisation." In: *arXiv preprint arXiv:1009.2646* (2010).

[330] Zhong-Yuan Zhang, Yong Wang, and Yong-Yeol Ahn. "Overlapping community detection in complex networks using symmetric binary matrix factorization." In: *Physical Review E* 87.6 (2013), p. 062803.

[331] Dongxiao He, Di Jin, Carlos Baquero, and Dayou Liu. "Link Community Detection Using Generative Model and Nonnegative Matrix Factorization." In: *PloS one* 9.1 (2014).

[332] Xiaochun Cao, Xiao Wang, Di Jin, Yixin Cao, and Dongxiao He. "The (un) supervised detection of overlapping communities as well as hubs and outliers via (bayesian) NMF." In: *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. International World Wide Web Conferences Steering Committee. 2014, pp. 233–234.

[333] Art B. Owen and Patrick O. Perry. "Bi-cross-validation of the SVD and the nonnegative matrix factorization." In: *The Annals of Applied Statistics* (2009), pp. 564–594.

[334] *Fichier Mobilités professionnelles des individus : déplacements commune de résidence / commune de travail [Home to work commuting flows by transportation modes]*. 2010.

[335] *Scipy Stats Statistical Functions*. http://docs.scipy.org/doc/scipy/reference/stats.htmlmodule-scipy.stats. Accessed: 2016-05-22.

[336] Leendert Cornelis Elisa Struik. "Physical aging in amorphous polymers and other materials." PhD thesis. TU Delft, Delft University of Technology, 1977.

[337] Animesh Mukherjee, Francesca Tria, Andrea Baronchelli, Andrea Puglisi, and Vittorio Loreto. "Aging in language dynamics." In: *PLoS One* 6.2 (2011), e16677.

[338] Google Developers. *General Transit Feed Specification Reference $https://developers.google.com/transit/gtfs/reference?hl=it$*. 2012. URL: https://developers.google.com/transit/gtfs/reference?hl=it.

[339] *Enquête Nationale Transports et Déplacements 2007-2008*.

[340] *Google Maps APIs*.

[341] *Base communale des aires urbaines 2010*.

[342] Daniel D. Lee and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." In: *Advances in neural information processing systems*. 2001, pp. 556–562.

[343] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. "Metagenes and molecular pattern discovery using matrix factorization." In: *Proceedings of the national academy of sciences* 101.12 (2004), pp. 4164–4169.