Technical University of Denmark

DTU

Data Mining and Visualization of Large Human Behavior Data Sets

Cuttone, Andrea; Larsen, Jakob Eg; Jørgensen, Sune Lehmann

[Link back to DTU Orbit](#)

DTU Library
Technical Information Center of Denmark

Data Mining and Visualization of Large Human Behavior Data Sets

Andrea Cuttone

Kongens Lyngby 2016

Summary (English)

Traditional methods for studying human behavior such as surveys and manual collection are expensive, time-consuming and therefore cannot be easily applied at large scale. In recent years an explosive amount of digital traces of human activity – for example social network interactions, emails and credit card transactions – have provided us new sources for studying our behavior. In particular smartphones have emerged as new tools for collecting data about human activity, thanks to their sensing capabilities and their ubiquity. This thesis investigates the question of what we can learn about human behavior from this rich and pervasive mobile sensing data.

In the first part, we describe a large-scale data collection deployment collecting high-resolution data for over 800 students at the Technical University of Denmark using smartphones, including location, social proximity, calls and SMS. We provide an overview of the technical infrastructure, the experimental design, and the privacy measures.

The second part investigates the usage of this mobile sensing data for understanding *personal* behavior. We describe two large-scale user studies on the deployment of self-tracking apps, in order to understand the patterns of usage and non-usage. Moreover we provide some design guidelines for facilitating reflection in self-tracking systems. Finally we propose a model for inferring sleep patterns from smartphone interactions.

In the third part, we focus on a specific aspect of *collective* behavior: human mobility. We perform an experiment to verify the feasibility of inferring places from location traces using mobile sensing data. We develop a hierarchical model for human mobility, which is able to measure mobility properties at multiple scales. We perform a study on the factors influencing the accuracy of next-place prediction models. Finally we present an open-source tool for creating geographical visualizations.

Summary (Danish)

Traditionelle metoder til at studere menneskelig adfærd, som f.eks. spørgeskema-er og manuelle observationer, er bekostelige og tidskrævende og derfor vanskelige at anvende i stor målestok. I de senere år har der imidlertid været en eksplosiv vækst i digitale spor som hidrører fra menneskelig aktivitet — eksempelvis interaktion på sociale netværk, email, og kreditkort transaktioner — som giver adgang til nye måder at studere adfærd på. I særdeleshed fremstår smartphones, med deres mange sensorer og store udbredelse, som et nyt værktøj til indsamling af data om menneskelig aktivitet. Denne afhandling undersøger hvad vi kan lære om menneskelig adfærd ud fra disse omfattende og vidt udbredte mobile datakilder.

I den første del beskriver vi indsamling af høj-opløselige data i stor skala fra smartphones fra over 800 studerende ved Danmarks Tekniske Universitet, in-klusive data om geografisk placering, sociale interaktioner, opkald og SMS. Vi giver et overblik over den tekniske infrastruktur, det eksperimentelle design og hvad der er gjort for at sikre privatlivet.

I den anden del undersøger vi hvordan disse indsamlede data kan bruges til at forstå *personlig* adfærd. Vi beskriver to stor-skala studier af apps til at registre-re egen adfærd (self-tracking), for at forstå mønstre i brug og ikke-brug heraf. Yderligere angiver vi design-retningslinier som hjælper til selvreflektion i syste-mer til registrering af egen adfærd. Vi foreslår også en model der kan bruges til at udlede søvnmønstre ud fra brugen af smartphones.

I den tredje del fokuserer vi på et specifik aspekt af *kollektiv* adfærd: menneskelig mobilitet. Vi udfører et eksperiment for at verificere muligheden for at udlede information om steder ud fra geografiske spor i mobile data. Vi udvikler en hierarkisk model for menneskelig mobilitet som kan måle mobile egenskaber i varierende målestoksforhold. Vi studerer faktorer som påvirker nøjagtigheden af at kunne forudsige den næste forventede placering. Endelig præsenterer vi open-source værktøjer som kan bruges til at skabe geografiske visualiseringer.

Preface

This thesis was prepared at DTU Compute, Cognitive Systems section under the supervision of Associate Professor Jakob Eg Larsen and Associate Professor Sune Lehmann in fulfillment of the requirements for acquiring a Ph.D. in Engineering. The thesis deals with data mining and visualization of large human behavior data sets, and includes four published papers and five upcoming papers.

Lyngby, 31-July-2016

Andrea Cuttone

Acknowledgments

The results presented in this thesis are not only my own doing, but they have been produced thanks to the support, ideas, conversations, discussions and work of many many people.

My first thanks go to my supervisors Jakob Eg Larsen and Sune Lehmann who have guided me through this extremely complex process that is science, and have always been an invaluable resource for discussing ideas and finding solutions. I would also like to thank all my collaborators and colleagues that worked with me over the years: Arkadiusz Stopczynski, Vedran Sekara, Radu Gatej, Piotr Sapiezynski, Enys Mones, David Kofoed Wind, Per Bækgaard, Michael Kai Petersen, Georgios Chatzigeorgakidis.

I am grateful to Marta González for inviting me at her HumNet lab, and for collaborating with me on a number of projects. Thanks also to Serdar, Edward, Riccardo, Yingxiang, Shan, Philip, Fahad, Sid, Brad and Luis for the nice time in Boston.

I am also grateful to Otto Mønsted Fonden, Oticon Fonden and Fisker Fonden for the generous grants supporting my external stay and travels.

A huge thanks to my parents, my sister and my girlfriend Klaudia who supported me with love over all these years.

Finally, I would like to thank you reader for taking a look at my thesis. I apologize in advance for writing such a long dissertation, but I didn't have time to write a short one.

Contents

List of Figures

Introduction

Imagine there was a device that could know everything about you. It could measure your physical activity, quantify the length and quality of your sleep, and predict if you are sick or depressed. It would know where your home is, the location of your work and the other places of your life, if you bike or take the bus, and could even recommend new destinations. It would keep track of the people you spend time with, you talk to, and store every single conversation you have.

This device is right here in your pocket: your smartphone. Modern smartphones are packed with sensors for location, audio, movement, and can record a plethora of information about their usage, including phone calls and SMS, and screen interactions. They also provide enough computational power to run sophisticated algorithms for extracting information from the sensor data.

This thesis investigates the question of what we can learn about human behavior from this rich and pervasive mobile sensing data.

At the individual level, this data has a personal significance and can therefore be used to understand and maybe even improve personal behavior, health and wellbeing. How to extract meaningful behavior from raw mobile sensing data? What can we learn about ourselves from this data? How can we design systems and tools that help us understand our own behavior?

At the collective level, this data can be used to study societies as a whole and try to understand complex phenomena such as social interactions and human mobility. Can we find laws that govern human behavior, such as patterns of communication, dynamics of human movement, fundamental structures of social interactions?

This thesis will address some of these questions.

1.1   Outline

The remainder of this thesis is structured as follows.

Chapter 2 introduces the idea measuring human behavior through data. The chapter gives an overview of the many sources of data, with particular focus on mobile sensing. An introduction to the field of Computational Social Science is provided, including goals, major results and challenges. Finally our work on large-scale measurement of human behavior is described.

Chapter 3 concerns the usage of personal data to understand individual behavior. The ideas of self-tracking and Quantified-Self are discussed. An overview of key Personal Informatics literature is then provided, including different theoretical frameworks, the issue of self-reflection, and the role of data visualization. Several contributions are then presented: the study of a large scale deployment of a Personal Informatics system, an analysis of the relation between personality traits and self-tracking, a set of design guidelines to facilitate reflection and finally a model for inferring sleep patterns from smartphones interactions.

Chapter 4 describes the usage of human behavior data for modeling one specific type of collective behavior: human mobility. The chapter starts by providing an overview of positioning systems, with focus on smartphones. Our work on a tool for visualizing geographical data is then presented. The rest of the chapter is then divided into three research areas: place recognition, mobility properties and next-place prediction. For each of the areas, the highlights of the relevant literature are provided. Our work in each area is then presented: a verification of the feasibility of inferring mobility properties from low-accuracy smartphones data, a study of mobility properties at multiple scales and finally an analysis on the factors influencing the accuracy of next-place prediction.

Chapter 5 summarizes the contributions of this thesis, discusses the challenges encountered, and finally suggests open questions and areas for future work.

1.2 Summary of Papers

This dissertation consists of 4 published papers, 4 papers in submission and one manuscript in preparation. Here we provide a short summary for each of them. The papers are provided in Appendix.

[A] Measuring Large-Scale Social Networks with High Resolution describes a large-scale data collection deployment that aims at measuring human behavior with unprecedented depth and temporal resolution. Data is collected using smartphones and includes location, Bluetooth proximity, calls and SMS, personality traits and Facebook activity. The data is collected approx. 800 first-year university students at the Technical University of Denmark and for more than 2 years. We describe the research agenda and the technical, experimental and privacy challenges and solutions. Finally we provide an overview of the initial results and show the importance of multi-channel high-resolution data.

[B] The Long Tail Issue in Large Scale Deployment of Personal Informatics reports on the large-scale deployment (N=136) of a Personal Informatics tool providing visualizations of mobility and social interactions. We discuss the limited uptake of the app and the limitation of user engagement, and how this can be a significant issue for any large-scale study on Personal Informatics.

[C] Who Wants to Self-Track Anyway? Measuring the Relation between Self-Tracking Behavior and Personality Traits reports on a follow-up large-scale deployment (N=796) of a Personal Informatics system providing feedback on personal mobility. In particular we focus on the relation between the app usage and personality traits. We find that only conscientiousness may have a significant impact, while we find no significant difference due to narcissism, in contrast to popular views.

[D] Four Data Visualization Heuristics to Facilitate Reflection in Personal Informatics discusses how data visualization can be used to facilitate reflection on personal data. From a review of Personal Informatics and data visualization literature we propose four design heuristics: make data interpretable at a glance, enable exploration of patterns in time series data, enable discovery of trends in multiple data streams and turn key metrics into affordances for action. Finally we consider as a case study two popular self-tracking tools: Basis and Fitbit.

[E] SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events proposes a Bayesian model for inferring sleep patterns from simple smartphone usage activity. Our model allows us to estimate the probability of sleep and awake times, and can encode prior knowledge and

dependencies among parameters. We fit the model to over 400 participants from two different datasets, and we verify the results against ground truth collected with sleep trackers. Finally we show that the model is able to extract reasonable individual and collective sleep patterns.

[F] geoplotlib: a Python Toolbox for Visualizing Geographical Data describes an open-source python toolbox for visualizing geographical data, developed as part of this thesis. geoplotlib implements many common tools such as dot maps, heatmaps, Voronoi tesselation and shapefiles. We describe geoplotlib design, use cases and features.

[G] Inferring Human Mobility from Sparse Low Accuracy Mobile Sensing Data studies the feasibility of inferring human mobility from sparse, low accuracy data collected opportunistically with smartphones. We compare different techniques for extracting significant places from mobility traces, and we validate the results using ground truth collected in form of travel diaries. Our results suggest that it is possible to reasonably infer mobility patterns even from this low-resolution data.

[H] Measuring Human Mobility at Multiple Scales describes a hierarchical model for human mobility that is able to measure mobility properties at multiple scales. We quantify a number of properties including the heavy-tailed staying time, the exploration patterns during vacations periods, the daily and weekly periodic returns, and the navigation in the spatial hierarchy between different scales.

[I] Understanding Predictability and Exploration in Human Mobility investigates which factors influence the accuracy of next-place prediction in human mobility. We show that the problem formulation, and the spatial and temporal resolution have a strong effect in the accuracy of prediction. Finally we uncover the mechanism of exploration of new locations and we measure than on average 20-25% of transitions are to new places, and over 70% of places are visited only once. We suggest that exploration is another important limiting factor for predicting human mobility.

CHAPTER 2

Measuring Human Behavior

"Measure what is measurable,
and make measurable what is not so"
GALILEO GALILEI

In this chapter we discuss the idea of measuring human behavior using data. Section 2.1 provides motivation for collecting data about human behavior, and the limitations of traditional Social Sciences. Section 2.2 shows how data generation and storage is now pervasive, and lists some examples of the many data sources available. Section 2.3 discusses how smartphones have become one of the most accurate measurement tools for human behavior. Section 2.4 introduces the field of Computational Social Science, and Section 2.5 provides a review of the largest data collection studies for human behavior. Finally Section 2.6 presents our contribution to the field, the largest data collection deployment in terms of channels and number of participants to date.

2.1 Social Sciences

Social Sciences aim to understand human behavior, both at the individual and at the collective level. For example the discipline of economics tries to understand

how people exchange goods and services, human geography seeks to understand how humans move and behave spatially, and sociology studies the ways people interact and function as a society.

Traditionally, data on human behavior has been collected manually by researchers either by observing people, or asking them directly using surveys, questionnaires, phone or in-person interviews, or focus groups. These methods present a number of limitations. They are time-consuming and costly, as they need to be carried out manually by researchers. Consequently, the samples are typically limited to small sizes. Due to these logistic problems, data collection is limited to few points in time, sometimes even just a single measurement. This gives only a static view of reality, and makes it impossible to study the changes of behavior over time. Moreover questionnaires and surveys can introduce a large number of biases [1]: people may misinterpret questions and answers, omit private or embarrassing details, simply forget or distort the past experiences, or be too fatigued to truthfully complete the responses.

2.2   Big Data Revolution

In recent years there has been a dramatic increase in our capability to produce, store and process data in digital format. Most information is now digitalized, from bank transactions to medical records, from phone bills to purchase history. Some of this data is collected passively and automatically, without us even noticing. Every time we make a phone call, we check-in on the bus or we perform a search in our browser, we leave behind some information about ourselves. At the same time, our ability to share digital content has also multiplied: we email colleagues at work, keep in touch with our friends using multiple instant messengers, upload pictures of our food, *tweet* about our vacations, record our run with GPS and share it on Facebook, and post videos on YouTube. And the increasingly cheaper and faster storage options mean that all this data can be stored and retrieved for unlimited amount of time. All these pieces of information constitute "digital breadcrumbs" [2] that reveal some facets of our behavior, or an approximation of it. The contacts on online social networks may represent some of our real friendships. The geo-tags capture part of our real patterns of mobility. Our status updates may reveal our political orientation.

While traditional Social Sciences studies can provide a very detailed picture of human behavior but typically only at small-scale, these new big data sources can provide a simpler view but of a size many orders of magnitude larger. We will now describe a number of these digital sources for data on human behavior.

Email is one of the oldest forms of digital communication, but it has stood the test of time surprisingly well. Using the senders and receivers of emails it is possible to build a graph of social contacts, while timestamps can be used to infer rhythms of communications. Email datasets have been used to predict the quality of collaborative work [3], characterize communication patterns [4, 5], and study the structure of social networks [6, 7].

Online Social Networks (OSNs) are websites that allow users to interact with each other in different ways such as messaging, sharing content, or playing games. Probably the most famous OSN is Facebook, which in August 2015 reached over 1.5 billion monthly active users [8]. Twitter comes at a distant second place with over 300 million monthly active users [9]. OSNs provide an excellent data source for studying how users communicate with each other, how they form social relationships, and how they share contents and ideas. OSNs data has been used for a myriad of purposes including: location prediction [10, 11], modeling of human mobility [12], social network analysis [13, 14], detection of catastrophic events [15], stock market forecast [16], election forecast [17], and information spreading [18].

Call Detail Records (CDRs) are the metadata collected by telecom operators about phone calls, including receivers and senders, time and rough geographical location at cell tower level. These datasets are typically very comprehensive, covering entire countries, with millions of users and for very long periods of time. CDRs have been used to study human mobility patterns [19, 20, 21], urban structures [22, 23, 24], social networks [25, 26] and privacy of metadata [27].

Open-source repositories such as *sourgeforge.com* and *github.com* allow programmers to share personal code, and to collaborate on open-source projects. From the workflow on the code bases it is possible to infer collaborative patterns, and their effect on software quality [28, 29, 30].

Scientific repositories such as *arXiv*, *Scopus* and *Web of Science* record the co-authorship information for thousands of scientific papers. This information has been used for measuring scientific impact [31, 32] and predicting scientific collaborations [33].

RFID are small devices that contain a radio emitter and receiver, and therefore are able to detect when they are in close proximity. These devices can be used to detect face-to-face interaction or co-presence among people wearing them, typically in form of badges. They have been for example used at scientific conferences, in offices [3] or even in hospitals [34] as proxies for close interaction.

2.3   Smartphones as Measuring Tools

The data sources discussed so far present several limitations: they capture a single facet of behavior such as social interactions on a specific channel, and they are sparse in time as they capture only specific events such as sending an email or making a phone call. Human behavior is of course much more complex: it is composed by a large number of facets expressed in multiple ways through multiple channels, and unfolds both in a very short timescale (multiple events in a short periods), and a very long timescale (behavioral changes over long stretches of time). In order to have a better measure of human behavior we need tools that are able to capture multiple channels, with high temporal resolution and for long periods of time. Smartphones have in recent years emerged as this new tool for capturing human behavior [35].

First of all, smartphones provide a multitude of sensors that can capture different aspects of behavior. The location sensor through a combination of GPS, Wi-Fi and GSM can determine the user location with accuracy of meters. The Bluetooth sensor (a short-range wireless communication device) can be used to detect close proximity to other people, and therefore as a proxy for person-to-person social interactions. Logs of calls and SMS record two channels of communication between people. The accelerometer sensor is able to measure the relative movement of the phone in terms of intensity, direction and orientation, and this data can be used to infer physical activity, mode of transport and even sleeping patterns. Smartphone apps usage can be used to infer contextual information, or even demographics data. Just knowing whether a user is interacting with his phone or not may give indication of his activity.

Smartphones are personal and always on, therefore supporting a continuous, pervasive data collection. Sensor data can be collected passively, without the user even noticing thus helping to ensure *ecological validity* [35] – that is, make sure that the experimental data is as close as possible to the real-life. Sensors can perform measurements even hundreds of times per second, and the data can be collected and stored indefinitely.

Finally, smartphones facilitate really large-scale studies because they are inexpensive and ubiquitous: in 2012 the global number of smartphones in use surpassed one billion [36], and 72% of people in USA own a smartphone in 2016 [37].

The usage of smartphone and other personal devices as tools for collecting data has produced a shift in the sensing paradigm, from traditional fixed sensors to people-centric sensing [38]. People themselves have become a fundamental part of the sensing framework, either in opportunistically collecting data on their

environment, or in a participatory fashion submitting the data themselves.

Data collected using smartphones does not however perfectly reflect human activity. Smartphone sensors still represent a coarse view of real human behavior, as sensors have limited accuracy and sensor data needs to be processed to extract any meaningful information. For example the Android location service can typically provide a position estimation within a radius from 5 to 100 meters, depending on the availability of GPS and Wi-Fi. The Bluetooth sensor is limited to scan a finite number of devices within a short distance, and the signal is affected by walls and other obstacles. The microphone sensor has limited capacity in terms of volume and frequency. At the same time, data can be missing or corrupted due to software and hardware issues, including hardware failure, sensors failures, software crashes both at the app and operating system level.

Finally human factors play a role. People may forget bringing the phone with them, temporarily – for example when going to another room – or for long periods of time – for example leaving the phone at home when going for holidays. The phone may be turned off either intentionally or because the battery dies. Users may want to disable some sensors for privacy reasons, to save battery (e.g. GPS), or simply by mistake. And as for any observational study, the mere knowledge of being observed may cause some changes in behavior for participants.

In recent years there has been an increased interest in using smartphones as sensing tools in the academic world, and a number of frameworks have been developed to automate the process of collecting, preprocessing and storing mobile sensing data. We cite as examples *Funf* [39], *SocioXensor* [40], *ContextPhone* [41], *MyExperience* [42], *CenceMe* [43], and *DarwinPhones* [44]. Typical features of such frameworks are automated and passive sensor collection, configurable scheduling of sensing behavior to save battery life, automatic upload of data to server, possibility to run questionnaires directly on the phone, encryption and secure communication.

2.4 Computational Social Science

The large-scale availability of these digital traces of human behavior has led to the formation of a new scientific field: Computational Social Science (CSS) [2]. The new discipline proposes to investigate many of the classical Social Sciences questions (i.e. social relations, mobility, economics) using a data-driven and computationally intensive approach. The vision is the capability of studying much larger populations, with a much deeper level of details, for longer periods

of time, and without self-reporting biases. But in order to achieve these results, CSS has to face many challenges.

One of the main issues is to protect user privacy while still allowing researchers to carry out scientific studies. The pervasive availability of personal data presents an excellent opportunity for advancing our understanding of society and human behavior, but at the same time poses a great threat to personal privacy – the right of a person to have control over his personal information. Privacy is recognized as a human right by the United Nations [45], and after the revelation of a mass surveillance program carried out by the NSA in the United States, privacy has become a hot topic in the public debate as well. A great deal of the personal data that is routinely collected is highly sensitive, for example: location, social interactions, political and religious ideology, economic status. To preserve privacy, data may be anonymized by removing or scrambling some fields [46], aggregated by substituting individual records to group averages [47], or perturbed with noise [48]. Unfortunately data analysis and privacy are often conflicting: data analysts may benefit from detailed individual data, while users may prefer their data to be as anonymized as possible.

Another issue intimately related to privacy is the reproducibility of studies. Limiting access to the data often means the impossibility to reproduce the data analysis. For example, a recent study made by Facebook on their social network has reported that the famous number of degrees of separation between people in the world is approximately four [14]. The authors could not disclose the raw data for privacy reasons, and this has sparked much controversy on the reproducibility of such studies, where the information is owned by private companies.

Another challenge for CSS is interdisciplinarity. Due to its very nature, the discipline needs both Computer Scientists who have expertise in sophisticated data analysis and computational tools, and Social Scientists that have deep knowledge of the main social and societal research questions. It is often the case that the two schools remain separated, but it is evident that a collaboration would greatly benefit both parties. Fortunately in recent years a number of conferences and journals focusing on CSS have appeared, and this has helped the collaboration among the multiple disciplines.

Finally these new data sources require completely new methods and tools for analysis: from algorithms and software to analyze huge amount of data, to statistical methods for testing large amounts of hypotheses, to new ways of thinking about dynamical social networks.

2.5 Large-Scale Data Collection Studies

In the previous section we have described some tools for collecting data about human behavior. But having the tools is only one part of the puzzle. In order to be able to draw reasonable conclusions, complex experimental studies are needed.

In many cases, it is not sufficient to study some individuals separately. Often researchers need to collect data for a large number of people, and for long periods of time.

Moreover a typical requirement is that the individuals in the population somehow interact with each other. For example, in order to study social behavior the population under study should contain individuals that normally have social contacts such as families, friends or students at the same university; or for studying metropolitan mobility patterns, it is required to have many users from a single city.

Finally, it is desirable for researchers to measure as many information channels as possible, for example using a combination of mobile sensing, questionnaires, and Facebook activity.

Due to the very large effort and resources needed to organize such experiments, only a few large-scale studies have been performed so far. We describe here the most significant studies of human behavior, in terms of multiplicity of information channels, number of participants and duration.

The **Reality Mining** project [49] pioneered large-scale studies using mobile phones. The project collected smartphone data including Bluetooth proximity, cell-towers location, call log and app usage for 100 MIT students and staff members over 9 months. The researchers report how different types of users (staff, professors, students) have different values of location entropy. They also propose a Hidden Markov Model for labeling locations as home/work/other. Finally they describe how the Bluetooth proximity can be used to build social networks, classify people as friends or colleagues and measure how social interaction and work rhythms change in response to particular events.

The **Lausanne Data Collection Campaign (LDCC)** [50] collected data using Nokia N95 smartphones for 170 participants in the Lausanne region (Switzerland) for approximately 1 year. The goal of the project was to collect a richer set of sensors compared to Reality Mining, including location from GPS and Wi-Fi, app usage, activity detection based on acceleration sensor, call and SMS logs, and metadata about media usage. Moreover the project aimed at collect-

ing data for a more heterogeneous population from a whole city, instead of a sample from a university. Their system proposes a data collection client based on a state machine, which determines the optimal trade-off between sensor and battery usage. In subsequent work [51], the data is shared with other researchers in order to compete in three different tasks. In the task of recognition of place categories, places were tagged with a semantic label, and the goal was to develop a classification model to assign the right label to unlabeled places, given a set of features. In the next place prediction task, the goal was to predict the next location given the past location history and the current contextual information. In the estimation of demographic attributes task, the goal was to predict basic demographic groups based on behavioral indicators of phone usage. Finally an open track allowed researchers to propose their own data analysis task.

The **SocialFMRI** [52] project recruited 130 participants within a social community in USA for over 15 months. The participants collected mobile sensing data (location, Bluetooth, apps, etc.), and periodically filled out surveys. Researchers had also access to Facebook activity and credit card transactions. The project produced a number of findings. By looking at the interplay between social interaction diversity and financial status, the researchers found that individuals lose social interaction diversity when their economic status gets worse, and vice versa gain diversity when their economic status gets better. The researchers also studied how the spreading of ideas is influenced by face-to-face interactions. They found in particular that people spending more time in proximity have higher chances of having more common apps. Finally the researchers report on an intervention experiment for increasing fitness activity. People were divided into different reward schemes, and it was found that the social schema had the strongest effect for influencing behavior.

In **Sensing the "Health State" of a community** [53], researchers collected mobile sensing data for a community of 70 students at an undergraduate university residence over an entire academic year. Participants were also asked to fill out monthly self-report surveys on diet and exercise, political opinions during the presidential election campaign, and daily symptoms reports on influenza. The researchers show how the behavioral change of people infected by influenza can be measured using mobile sensing (different call patterns, reduced mobility, less contact entropy). They also showed a correlation between Body-Mass Index of social contacts, and a mechanism for propagation of political opinions.

The **NetSense Smartphone Study** [54] describes the deployment of smartphones for mobile sensing at the University of Notre Dame. The study documents the difficulties in setting up and running such a large scale experiment, in particular participants retention: the number of active users started from around 200, and over the months fell to less than 80.

In **StudentLife** [55] data was collected for 48 students at Dartmouth College
across 10 weeks using mobile sensing, questionnaires and periodic ecological
momentary assessments [56] to probe mental state (stress, mood). The mobile
sensing app performed physical activity detection, sleep detection, and social
conversation detection. Significant correlations were found between sleep du-
ration, conversation frequency and duration, Bluetooth co-location and depres-
sion, stress levels, loneliness, academic performance. The researchers showed
that it is possible to detect effect of the increasing academic workload: stress
raises, mobility decreases, people stop going to the gym, class attendance de-
creases.

2.6 Paper Summary: "Measuring Large-Scale Social Networks with High Resolution"

The chapter so far has set the context for the work described in this section.
Previous studies have been limited either by the small number of participants,
by the short duration, or by the limited number of data channels recorded –
which give an incomplete view of collective behavior.

In this paper (Appendix A) we describe SensibleDTU, a large-scale data col-
lection project where we obtain data for more than 800 university students at
the Technical University of Denmark, for a period of over two years. Partic-
ipants were recruited on a voluntary basis, with the incentive of obtaining a
Nexus 4 smartphone that they could keep as long as they participated in the
experiment. Data was collected using smartphones running an app that pas-
sively captures sensors (Wi-Fi, GPS, Bluetooth, calls and SMS, screen on/off).
Moreover participants were requested to fill a number of surveys regarding per-
sonality, self-esteem, narcissism, and depression. Participants also shared their
Facebook activity and friendship networks.

To date the SensibleDTU project represents the richest study of its kind, thanks
to the combination of a population of around 800 participants, for a period
longer than two years, and a large multitude of data channels. Table 2.1 shows
the order of magnitude of the collected data.

The vision of the project is to understand human behavior, in particular social
interactions at a whole new level of details. For long time social scientists have
tried to understand social connections but their efforts have been limited by the
available data, typically self-reports from a few participants. Using smartphone
data we can directly measure social proximity using Bluetooth and location
using GPS, thus enabling us to infer when social gathering happen, who are

Channel	Samples collected
Location	258,000,000
Phone calls (sent/received)	1,000,000
SMS (sent/received)	4,500,000
Bluetooth proximity	48,000,000
Wi-Fi access points	285,000,000

Table 2.1: Number of samples collected for the SensibleDTU deployment

the participants, and where they take place. The automatic and pervasive data collection enabled by smartphones allows us to monitor this behavior 24 hours a day, 7 days a week, and for a timespan of years. The multitude of channels (phone calls, SMS, face-to-face, Facebook) allows us to capture the multiple facets of social interactions across different time periods (Fig. 2.1). We show that our approach of collecting behavioral traces on multiple channels and at multiple time scales manages to capture many more details of human behavior than single-channel low-resolution data sources.

This data opens the way for many exciting research possibilities including: identifying new structures for understanding social networks [57], studying the mechanisms of epidemics on a real-world network [58, 59], quantifying the privacy risks of different channels [60], and as we will see in Chapter 4 understanding human mobility at a new level of detail.

The paper also introduces a new standard for participants' privacy in such studies. The infrastructure is built on top of the OpenPDS framework [61], and guarantees anonymization, encryption, and secure communication. Participants are granted full control over their collected data thanks to the implementation of the Living Informed Consent principle [62]. Participants can at any time see the raw data collected about them, see who and how often their data is accessed, or even drop out of the study. Finally, presentations were periodically organized to share with participants how their data was used and which findings were produced.

The study is not only a passive data collection process but constitutes a *living lab*: a real-world testbed for running experiments and dynamically interacting with the population. For example, participants were asked questions about their current location using an Experience Sampling Method [63] app. In another instance, a virtual virus was spread among participants to study the mechanisms of epidemics and vaccination.

Finally the study aimed at keeping participants engaged by providing them

services based on their data. Participants had access to a number of web-based
and mobile apps, which were providing visualizations and statistics on personal
behavior such as mobility and social interactions [64].

The SensibleDTU project constitutes the foundation for the remainder of the
work in this thesis, as it provides the behavioral data and the experimental
framework.
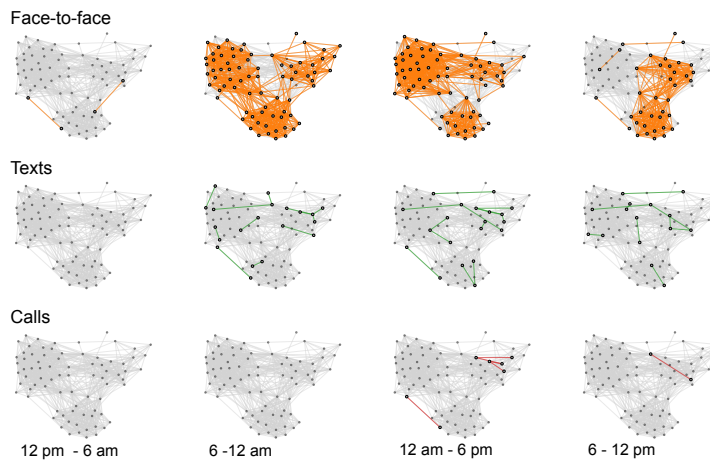
Figure 2.1: Social contacts are measured on multiple channels (calls, SMS,
face-to-face) and across different time periods (Figure from Ap-
pendix A).

Understanding Individual Behavior: Personal Informatics

"Know thyself"

INSCRIPTION IN THE TEMPLE OF APOLLO, DELPHI

In Chapter 2 we discussed how the explosive amount of data that we can collect represents an excellent opportunity for understanding human behavior. This data is first of all personal, and can therefore reveal insights about individual behavior.

The idea of self-tracking – collecting and using personal data for learning about personal behavior – is not completely new. Benjamin Franklin, one of the Founding Fathers of the United States, describes in his autobiography the practice of tracking his daily adherence to 13 "virtues" for self-improvement [65]. English statistician Francis Galton carried a homemade "registrator" to count different mundane aspects of life, including how many times people yawned during his talks [66]. In modern days, graphics designer Nicholas Felton produces an annual report of his own life, quantifying and visualizing its most detailed aspects

such as drinks consumed, places visited, and people met [67].

Self-tracking is common for regular people too: counting calories and measuring weight, tracking expenses and budgets, keeping diaries of mood and thoughts. A recent survey has shown that 7 out of 10 Americans keep track of a health indicator for themselves or a loved one [68]. While in the past pen, paper and self-reports were the only tools available, nowadays technology provides new powerful tools to capture, store, and analyze personal data. Using small and relatively cheap electronic devices, we can now measure signals that in the past required professional-grade equipment: sleep phases, body temperature, caloric expenditure, heartbeat. The rise of self-tracking has led to the parallel development of two movements: the Quantified-Self in popular culture, and the Personal Informatics in the academics circles[1].

This chapter is structured as follows. Section 3.1 describes the development of the Quantified-Self movement. Section 3.2 provides an overview of the Personal Informatics field, including a model for self-tracking, goals and challenges of self-tracking, the problem of reflection, and a number of related works. Section 3.3 provides a brief introduction to data visualization, one of the main tools for understanding data. Section 3.4 describes prior work done on a large-scale deployment of a Personal Informatics system. Section 3.5 summarizes a paper where we discuss the issues of uptake and usage on a large-scale deployment of a Personal Informatics system. Section 3.6 describes a paper where we analyze the relation between self-tracking and personality traits. Section 3.7 summarizes a paper where we propose four data visualization heuristics for facilitating reflection using data visualization. Finally Section 3.8 describes a paper where we propose a model for inferring individual sleep patterns from smartphones usage.

3.1   The Quantified-Self Movement

In 2007 Wired editor Gary Wolf coined the term *Quantified Self* (QS) to indicate the practice of self-tracking with the goal of "knowledge through numbers". From its inception, the QS phenomenon has grown dramatically. As today, the QS website *quantifiedself.com* lists over 500 between apps, websites and devices dedicated to self-tracking. The list includes fitness trackers from companies like *Fitbit* [69], *Jawbone* [70], *mybasis* [71] and *Nike* [72], which can measure a combination of distance, steps, calories burned, physical activity, and in some

---

[1]Although sometimes these terms are used interchangeably, for the rest of this thesis I will use Quantified-Self for the movement, and self-tracking otherwise.

cases heartbeat and body temperature. The Zeo[2] sleep tracking band provided in-depth sleep phases analysis. Websites like *TrackYourHappiness* [73], *Mood-Scope* [74] and *Mappiness* [75] allow users to track and share daily mood and feelings. Apps like *Daytum* [76], *Mycrocosm* [77] and *MercuryApp* [78] can track, capture and visualize any aspect of everyday life. On communities like *CureTo-gether* [79] and *PatientsLikeMe* [80] users can share their self-tracked symptoms and conditions, with the hope of finding solutions from other people's experiences.

There are now over a hundred QS groups in 34 countries, which periodically meet to exchange ideas, tools and results of their self-tracking efforts using a "show and tell" format. For the last few years, QS conferences have attracted hundreds of people from all over the world, and it has been sponsored by major IT companies such as Autodesk, Intel and Philips. The QS movement has generated a new market niche, with dozen of start-ups having huge growth. One example is the fitness tracking device Fitbit, which has sold nearly 11 million units in 2014 alone [81]. Big tech companies also joined the market: Microsoft, Apple, Google now produce smartbands and smart watches capable of self-tracking. It is estimated that health and fitness self-tracking devices, smartphones and tablets will reach 515 million units in 2017 [82].

Despite the claims of improving life through self-knowledge, some people question the effectiveness of the self-tracking process (see Section 3.2.3). However many individuals report satisfaction with their self-tracking effort, as the mere act of observing our own behavior may help us behave better. Moreover the QS movement has in some cases also benefited science. The data collected by the large QS companies represents some of the most extensive datasets on human behaviors ever recorded. For example Zeo, the manufacturers of a sleep tracking device, have over the year built an unprecedentedly large database of anonymized sleep patterns, and using this data researchers have found that women get less REM sleep than men [83]. *PatientsLikeMe*, a community where self-trackers can share and discuss their health status with others, has allowed researchers to compare symptoms with a new medication at a much larger scale than traditional clinical trials [83].

3.2   Personal Informatics

In the academic circles, self-tracking tools are referred to as Personal Informatics, which in the definition of Li et al. [84] are "systems [...] that help people collect personally relevant information for the purpose of self-reflection

---

[2]the company has shut down in 2013

and gaining self-knowledge". The creation of tools for collecting and interacting with personal data is a multi-disciplinary effort: from the hardware used for capturing the raw data, to the data mining algorithms for extracting knowledge from the data, to the user interfaces and visualizations to present the results to the user. In the rest of this chapter we will focus mostly on the Human-Computer Interaction aspects of Personal Informatics, except for Section 3.8, which instead describes an application of data mining to infer sleep patterns from smartphone interactions.

3.2.1   The Five-Stages Model

A widely used model for Personal Informatics was proposed by Li et al. [84], which developed a stage-based model with 5 stages:

1. **Preparation:** the user determines which information to collect, which tools to use and the goals for self-tracking

2. **Collection:** data is collected either manually (e.g. pen and paper) or using sensors (e.g. smartphones, fitness trackers)

3. **Integration:** data is processed to extract higher-level knowledge (patterns, trends, anomalies)

4. **Reflection:** results from the knowledge discovery are presented to the user in form of text summaries and data visualizations, in order to obtain insights about personal behavior

5. **Action:** recommended actions are suggested to correct or improve the current behavior

The stages are iterative, so it is typical to go through them multiple times to refine the process. Each stage presents some barriers, which are problems that may arise and slow or halt the self-tracking progress. These barriers cascade, which means that a problem in an early stage tends to create more problems in the subsequent stages. Example of barriers include: not knowing what to track, not having a specific goal in mind, choosing inappropriate tools for data collection, having to perform most of the collection or processing manually, lack of expertise in data analysis.

In subsequent work, Li et al. [85] identify six types of questions that people ask about their data: Status (what is the situation now?), History (what are the long-term trends?), Goals (what is my target?), Discrepancies (what are the

differences between my status and my goal?), Context (what is the context of my behavior?), and Factors (what influences my behavior?).

3.2.2   Goals of Self-Tracking

What are the reasons for tracking personal data? Surveys in literature [84, 86, 87], as well as personal reports of practitioners reveal a variety of motivations for the usage of self-tracking tools:

- **weight management** In order to lose, maintain of gain weight, self-trackers need to measure food intake and caloric expenditure, using a combination of food diary tools, fitness trackers, and old-fashioned scales.

- **health conditions** Another common use case is for people dealing with a specific sickness, from small issues like migraine to debilitating conditions like diabetes or autoimmune diseases. In this case the self-tracker looks for patterns that trigger or worsen their condition, such as specific foods, physical activity, blood pressure, body temperature, or sleep duration.

- **fitness tracking** Self-trackers can use a variety of tools for measuring physical activity, caloric expenditure, and store and share personal fitness records such as bicycle rides.

- **productivity** By collecting information about sleep, diet, or supplements the self-tracker can study the effect on productivity at work, and try to optimize it. Alternatively, the self-tracker can identify or block unproductive behavior (e.g. checking emails or Facebook too frequently).

- **social sharing** Some users may be interested in capturing and sharing personal life events, or achievements such a particularly long run. For a long time online social networks such as Facebook or Foursquare have provided a tool for sharing messages and activity. Self-tracking tools give even more options for sharing data.

- **technology enthusiasm** Some self-trackers simply enjoy playing with technology and gadgets such as self-tracking software and devices.

- **self-discovery** In some cases self-trackers may collect personal data without any specific purpose, but only a curiosity about self. It is interesting for some self-trackers to simply see a quantification of their own life, from the nightly hours of sleep, to the daily number of steps. The data may be collected and stored just in case one day could be used, in a kind of data hoarding.

3.2.3   Challenges

There are many challenges that can limit the effectiveness of the self-tracking effort.

Lack of scientific rigor Although it is now relatively easy to capture the raw signals (for example of steps, calories, heartbeat, sleep), it is much more difficult to make inferences about their relation to behavioral patterns. As an example, in order to infer the role of a specific food on the quality of sleep, one would need to collect the data on the food intake and the sleep patterns, and then analyze the data to arrive to a conclusion. But this requires knowledge of experimental design, data analysis, statistical testing and all issues related to the scientific method. Most self-trackers lack such scientific finesse and their analyses are limited to simple correlation between input and output [86]. And since in self-tracking the scientist is also the subject, placebo effects are somewhat inevitable: the simple belief that a treatment may work can already influence the behavior enough to produce a change.

Lack of integration Another issue in self-tracking is the lack of integration between different services [84]. Each tool is typically specialized in one area, for example sleep, physical activity or mood. So in order to combine multiple channels for more sophisticated computations, it is needed to export the data from each tool, make it compatible between each other, and re-aggregate it for the final analysis. This typically requires quite advanced programming and data analysis skills, often out of reach for most self-trackers.

Lapsing Despite the automation provided by most self-tracking tools, the self-tracking process can be time and effort consuming, so it is common for practitioners to stop tracking at some point. Reasons for this lapsing include: forgetting to track, tracking too many things at once, skipping entries, not knowing what to track, not tracking context [88].

3.2.4   Reflection

Reflection is one of the fundamental goals of Personal Informatics, and the fourth stage in Li's model [84]. However it is hard to arrive to a precise definition of what reflection really is [89, 90]. A review by Baumer et al. [91] has found that in a set of 76 papers using reflection in their title, only 30 tried to explain what reflection really means.

One common definition used is from Boud et al. [90], which define reflection as

"those intellectual and affective activities in which individuals engage to explore their experiences in order to lead to new understandings and appreciations". Several frameworks are based on this model. For example Rivera et al. [92] propose a tracking-triggering-recalling model built on top of Boud's. Fleck and Fitzpatrick [89] describe a five layers reflection framework (Description, Reflective Description, Dialogic Reflection, Transformative Reflection, Critical Reflection), where each layer represents a more in-depth understanding. Other reviews [93, 94, 95] note the cognitive, psychological and sociological issues related to reflection.

Given the difficulties on agreeing on a theoretically-grounded definition of reflection, more pragmatic evaluations are typically done. For example discoveries on personal behavior ("I am very sedentary", "I sleep too little"), noting correlations between events ("I sleep better after a workout"), or reminiscence of the past ("I remember visiting that place") are considered part of the reflection process. Another common way to evaluate reflection is in terms of concrete outcomes of the self-tracking process such as improved education, better design or improved understanding [91, 94].

3.2.5   User Interfaces for Reflection

The Personal Informatics field has significant overlap with lifelogging, ubiquitous and pervasive computing, so it is not surprising to find relevant work in these fields as well.

The first tools attempting to manage personal information were dedicated to organize files and documents, following Bush's vision of the *Memex*, "a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility" [96]. We cite as main examples *LifeStreams* [97] and *MyLifeBits* [98], which propose file managers to organize personal documents in a flexible way.

Lifelogging [99, 100] proposes the idea of systems able to capture and retrieve life experiences, including what people see, hear and think. In this context, user interfaces such as [101, 102, 103] focus on presenting personal data as life stories, under different contexts (spatial, temporal, social) and around different events.

In order to promote persuasion and behavior change, a number of avatar-based solution have been proposed. In *Fish'n'steps* [104], physical activity is represented by a virtual fish that becomes happier the more active the user is, while in *UbiFit* [105] a blossoming garden is used instead. In *UbiGreen* [106], polar bears and trees are used to encourage environmental-friendly choices. Evalua-

tion of these systems has shown that users respond well to the virtual avatars, and are motivated by them.

Finally, even abstract art [107, 108] has been proposed as a form of representation for personal data.

3.3   Data Visualization

Many of the user interfaces for Personal Informatics and commercial self-tracking tools try to facilitate the reflection process by providing feedback in some form of data visualization. In this section we provide a brief overview of the ideas behind data visualization.

Data visualization is the process of using graphical elements such as lines, shapes, colors and text to represent data [109]. Visualizations define an encoding between the data and the graphical elements, enabling data to be understood visually. For example a continuous variable can be encoded in a bar chart by the length of the bars, or in a pie chart by the angle of the sections. Data visualization exploits the human visual information system, which is extremely good at spotting patterns, trends and regularities in large amount of visual information.

One great example of the power of visualization is due to Anscombe [110], which describes four datasets having the same mean, variance and correlation, and then asks: what should we conclude about these datasets? A simple scatter plot of the data is then provided, which immediately shows how the four datasets are radically different.

A vast array of visualization techniques is available [111], depending on the data to visualize:

- **distributions:** barcharts, pie charts, histograms, box plots

- **two variables:** line charts, scatter plots

- **multivariate:** heatmaps, parallel coordinates, radial charts, correlogram

- **geographical data:** dot density maps, choroplet, cartograms, flow maps

One purpose of data visualization is Exploratory Data Analysis [112], where the data is represented in multiple ways to help the analyst explore the dataset. Often it is not clear in advance which questions to ask, or which hypothesis to

test. By visually exploring the data it is possible to find interesting patterns, spot outliers and generate questions. This process is very dynamic: there is a constant iteration between data visualization and hypothesis generation. To support this exploration, data visualizations are typically very dynamic and support interactivity. Some of the main techniques for interaction are [113, 114]: providing an overview, zoom and filtering, showing details on demand, sorting, selecting elements, providing focus+context, linking multiple views.

Another role for data visualization is the explanatory usage. In this case, we already have achieved some results and we want to represent them visually in order to communicate this knowledge. Data visualization is once again one of the most effective ways to do so. Creating explanatory visualization is a design task, and multiple factors should be taken into account. What is the message of the visualization? Which questions are being answered? Who is the audience of the visualization? What is their technical expertise in the subject? What is their competency in statistics and visualization? How much time and effort should the readers put for using the visualization? A number of design principles have been developed in the data visualization community, most notably the work of Tufte [115], who recommends minimizing decorative elements and the distortion of graphical representation.

The evaluation of a given data visualization is a very complex task, since as in the case of reflection (Section 3.2.4) it is not clear what effectiveness of visualization really means. Even choosing the "best" visualization for a given task or dataset is not simple, since the interpretation varies from person to person. The practical methods used for evaluating novel visualization techniques are measuring the efficiency for specific tasks, counting the number of insights that users obtain, and analyzing qualitative feedback [116]. The evaluation of visualization techniques remains one of the top open problems in the field [117].

3.4 SensibleJournal

In previous work [64] we described the large-scale deployment of SensibleJournal, a Personal Informatics mobile app that provides visualizations of mobility and social interactions. The app was deployed for approx. 6 months to N=136 first year university students at the Technical University of Denmark. The app provided a number of user interfaces and visualizations based on the mobile sensing data:

- a feed-based user interface displaying a daily summary of the visited places, modes of transport and distance traveled.

- a map view displaying an animation of daily trajectories and visited places.

- a spiral visualization [118] of the time series of places visits, which highlights recurring patterns and anomalies in mobility (Fig. 3.1).

- a bubble chart of social contacts and social communities inferred from Bluetooth proximity (Fig. 3.2).
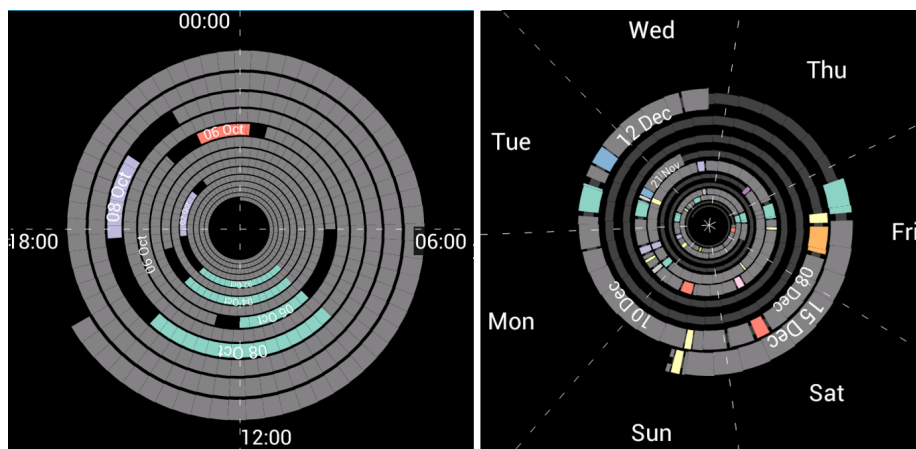
Figure 3.1: A view from SensibleJournal: a spiral visualization of the places visits (Figure from [64]).
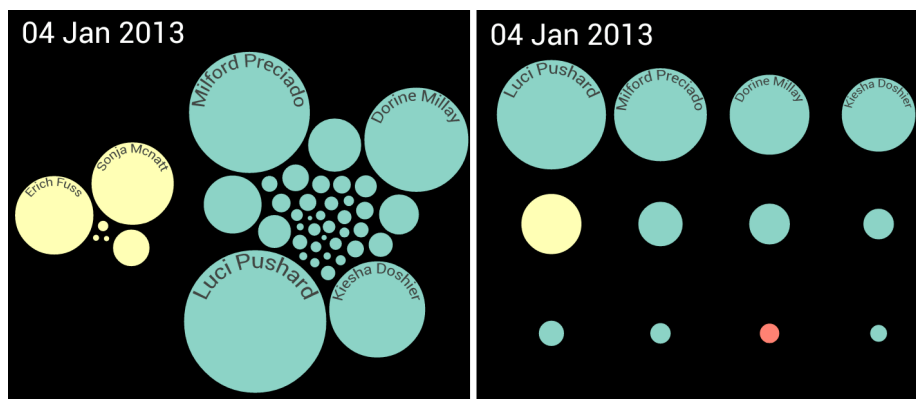
Figure 3.2: Another view from SensibleJournal: a bubble chart of social contacts and communities (Figure from [64]).

The deployment of the app was done in phases, so that new visualizations were added over time. The app collected usage logs by recording the timestamped user interactions with the app. Finally, users were asked to answer a questionnaire, to which 45 participants responded (33%). The goal of this deployment was to collect data on the usage patterns of such self-tracking tools. In the next section we will describe the results of this analysis.

3.5 Paper Summary: "The Long Tail Issue in Large Scale Deployment of Personal Informatics"

Despite the commercial success and the widespread adoption of Personal Informatics and self-tracking tools, large-scale controlled studies are somewhat lacking in literature. In particular most studies focus on small selected group of participants that are requested to use the Personal Informatics tools during the experiment, and are often already interested in self-tracking.

In this work (Appendix B) we study the voluntary uptake and usage patterns of SensibleJournal, a Personal Informatics application deployed to a large population (N=136), as described in Section 3.4. We analyze the usage logs and the questionnaire answers to investigate the patterns of usage and non-usage of the SensibleJournal app, as an instance of a large-scale deployment for a Personal Informatics system. Our main finding is that despite the large scale of the deployment, only a small percentage of users were engaged by the tool.

We quantify for each day the number of active users as users having at least one interaction of 10 seconds or longer. We find that the number of active users fluctuates over time, and sharp peaks can be measure in the periods following the release of new visualizations (Fig. 3.3). We also measure the total time spent on the app per user.

We find that the distribution of time user has a long tail, with most users spending little time on the app, and only a few users being active (Fig. 3.4). We discuss how this phenomenon of the long tail of user adoption is a critical issue in Personal Informatics, since after considerable time and costs of the development of such systems only a small percentage of users is engaged.
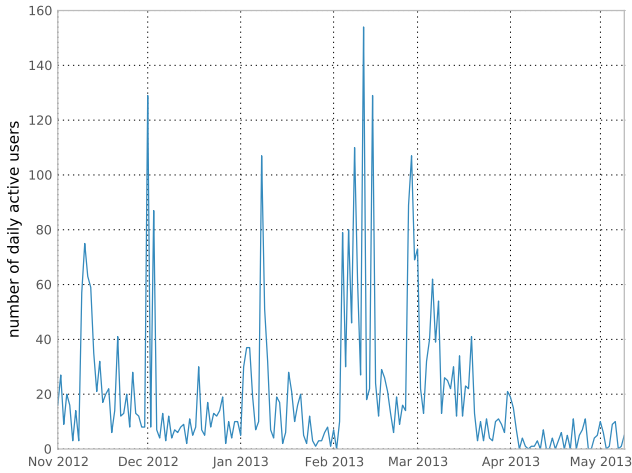
Figure 3.3: The number of active users of the SensibleJournal app over time (Figure from Appendix B).
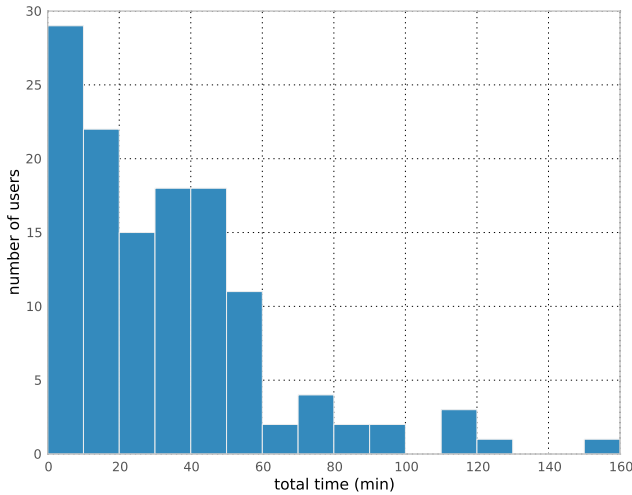
Figure 3.4: The distribution of total time of usage of the SensibleJournal app (Figure from Appendix B).

3.6   Paper Summary: "Who Wants to Self-Track Anyway? Measuring the Relation between Self-Tracking Behavior and Personality Traits"

The work presented in the previous section has opened a number of questions around the disparity between the frequent users and the rest. Who are the self-trackers? What are the reasons and the goals for people that used the self-tracking app? Section 3.2.2 has shown how the literature in this area has described a number of concrete use cases, but the fundamental reasons that lead some people to self-track and some others to not self-track are yet to be understood.

Personality is one possible candidate for interest or non-interest in self-tracking. In particular narcissism – the (excessive) interest in oneself – seems a possible candidate for the self-tracking behavior. Indeed the popular media has suggested the link between self-tracking and narcissism [119, 120]. Also in the research literature the self-tracking practices are often related to narcissism [93, 121, 122] although some authors have criticized this view [123]. To date only a small-scale study (N=36) has made an attempt to quantify these hypotheses [124].

In this paper (Appendix C) we perform a large-scale deployment (N=796) of a Personal Informatics mobile app that provides feedback on personal mobility patterns. The app was made available to participants of the SensibleDTU study (Section 2.6), together with a collector app that captures a number of mobile sensing channels, including location data. Each participant was required to install the app, but the subsequent usage was left optional. At the beginning of the experiment participants were also requested to fill a questionnaire including the Big Five inventory [125] and Narcissism NAR-Q [126], from which the following six personality traits were obtained: extraversion, agreeableness, conscientiousness, neuroticism, openness and narcissism. The app provided a number of views of personal mobility, for example the number of places visited, the distance travelled or the mode of transport, including maps and graphs. Fig. 3.5 shows an example of the views provided by the app.

A usage log of the interaction with the app was collected and periodically uploaded to our server. The analysis of the usage patterns confirms the finding reported in our prior work. The majority of users had little interaction with the app, and only a small minority used the system on a regular basis, with less 5% launching it more than 20 times. The number of active users slowly decayed from the start of the experiment, with a peak in the beginning of September – the start of the new university semester.
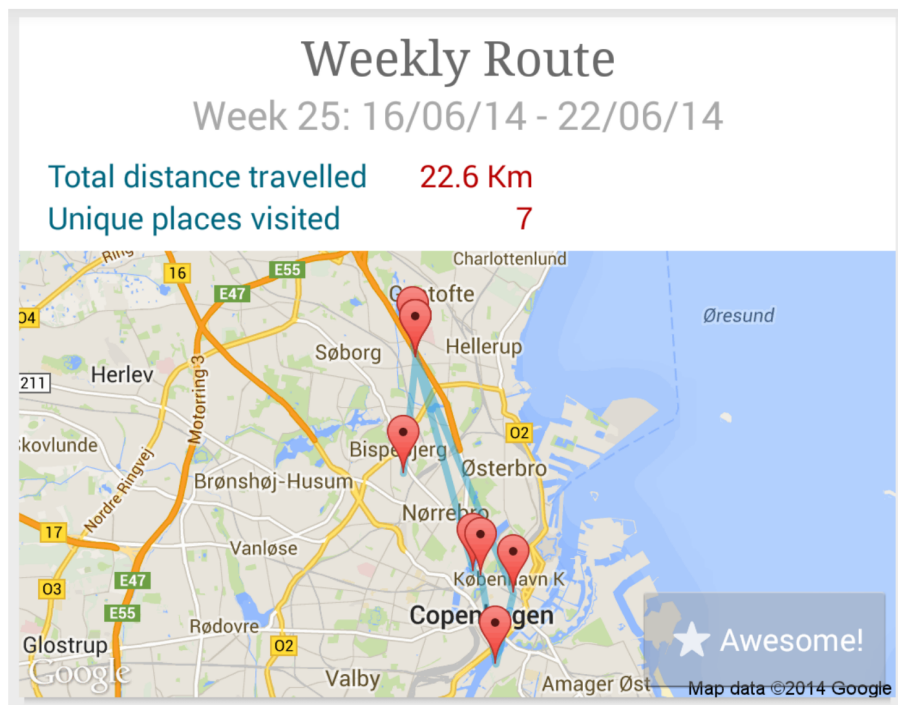
Figure 3.5: A screenshot from SensibleJournal2014 (Figure from Appendix C).

In order to quantify the relation between personality and self-tracking we compute four usage-related measures: number of days with at least one launch, total time interacting with the app, total number of launches and mean session duration. For each of the usage measures we divide the population into the top 10% and the rest 90%, and we compare the mean for each of the personality traits using t-tests (4 usage measures x 6 personality traits = 24 tests). In order to correct for multiple comparisons, we adjust the p-values using the Holm-Bonferroni procedure [127]. The result of the analysis is that only conscientiousness with total time produces a statistically significant difference. Notably narcissism does not produce any statistically significant difference, contrary to the previous hypotheses. Although more work is needed to confirm the effect of personality traits on self-tracking behavior, this work provides some first indications on the possible causes.

3.7 Paper Summary: "Four Data Visualization Heuristics to Facilitate Reflection in Personal Informatics"

As discussed in the previous sections, Personal Informatics systems try to facilitate reflection using different forms of feedback. In this work (Appendix D) we discuss how data visualization can be applied to the personal data domain to facilitate the reflection process. Exploratory data analysis using data visualization can be a powerful tool for exploring personal data, and can support many of the self-tracking goals: finding correlations, spotting outliers, understanding trends. From a review of the data visualization and Personal Informatics literature, we identify four design heuristics.

Make data interpretable at a glance. Users want to obtain insights on their behaviors with minimal time and effort. Given the complexity of multi-channel longitudinal self-tracking data, visualization should support interpretation at a glance. Visualizations could for example provide simple overviews such as dashboards or summaries, and show more details on demand.

Enable exploration of patterns in time series data. Time series data are one of the fundamental formats for self-tracking information. Two time-related behavior are of main interest: global trends (does my weight decrease over time?) and periodic patterns (do I smoke more in the weekend?). Most systems provide support for analyzing global trends, but often lack visualizations of periodic ones. Visualization techniques such as calendars and spirals [118] can facilitate the reflection process on periodic patterns.

Enable discovery of trends in multiple data streams. In most cases it is desirable not only to inspect trends for individual self-tracking variables, but also to understand the relation between them. For example it is indeed interesting to see that my weight has gone down in the last few months, but it would be even more interesting to see that the weight loss has happened since I started to walk more than 10,000 steps per day. The analysis of relation between variables can be facilitated by appropriate visualization techniques for multivariate data, for example scatter plots, corrgrams [128] or small multiples [115].

Turn key metrics into affordances for action. As for exploratory data analysis, reflection on personal data is an iterative process where a feedback form produces more questions which in turn prompt the user to ask new questions. Visualization tools for Personal Informatics should be able to support this iterative process implementing some common interaction techniques such as filtering, details on demand, annotations, and history.

Finally we consider as representative case studies two popular self-tracking tools (Basis and Fitbit) and we discuss how they offer limited support for more advanced data exploration.

3.8   Paper Summary: "SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events"

In this section we focus on understanding another aspect of personal behavior: sleep patterns. Sleep is a fundamental aspect of life, as sleep quality have significant impact of health and well-being. Large-scales sleep studies have relied on questionnaires and self-reports, which are not completely reliable. More accurate measurements require expensive and inconvenient equipments, and therefore do not scale well to large populations. In this context, smartphones are once again a precious tool for the measurement of human behavior. A number of smartphone apps such as *Sleep Cycle* [129], *SleepBot* [130], and *Sleep as Android* [131] can track sleep patterns. Moreover several scientific studies have shown that it is possible to infer reasonable sleep patterns using smartphone sensors such as audio from the microphone [132], light [133], and accelerometer movement [134].

Our contribution to this area (Appendix E) is to develop a model for inferring sleep patterns using an even weaker signal: just knowing when the smartphone screen is turned on or off. The development of this idea started from considering how much smartphones have become part of our life. We wake up and as first thing in the morning we snooze our smartphone alarm clock. On our way to work we read emails and check our calendar. During the day we keep checking our phone for chats and text messages. Finally the last thing we do for the day is to set the alarm clock for the next morning. Only when we sleep we leave our phone alone, and even then we may check it when we wakeup once or twice in the middle of the night. Our hypothesis is that smartphones interactions really become a proxy for our sleep, or lack thereof.

We develop a Bayesian model based on the idea that the number of screen-on events are generated by two separate distributions: a Poisson distribution with rate λ_{awake} when being awake, and a Poisson distribution with rate λ_{sleep} while being asleep. We divide each day into 15-minutes timebins, and assume that the rate changes depending on the timebin of the day. The rate is λ_{awake} between the time of wakeup t_{awake} and the time of sleep t_{sleep}, while the rate is λ_{sleep} before t_{awake} and after t_{sleep}. Fig. 3.6 shows a conceptual representation of the relation between the parameters.
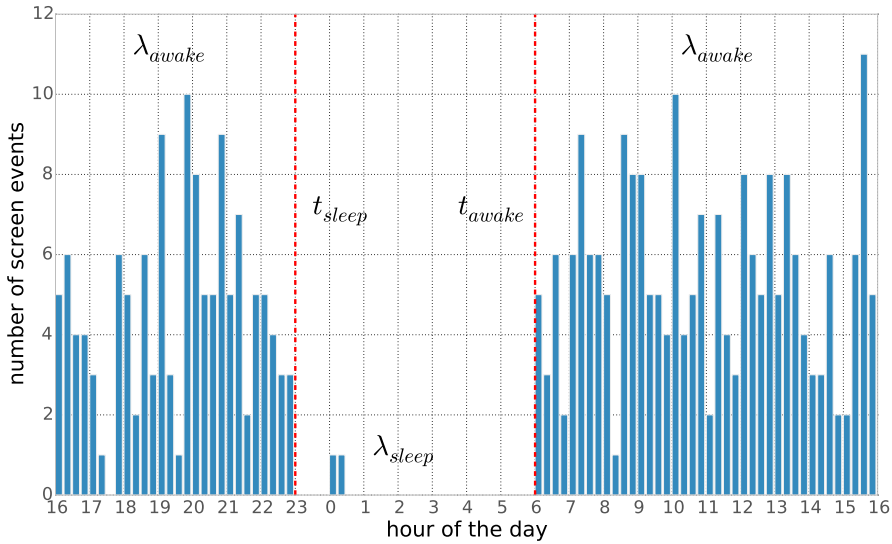
Figure 3.6: A conceptual representation of the sleep model (Figure from Appendix E).

We fit our model to two datasets, one courtesy of Sony and another one from
SensibleDTU (Section 2.6), for a total of more than 400 users. The fit is done
using Markov Chain Monte Carlo (MCMC), which searches for the most likely
values of the parameters guided by the log likelihood. Compared to other approaches using ad-hoc rules for inferring sleep patterns from smartphone interactions [135], our model has several advantages. The Bayesian approach can relate
the parameters in multiple ways, allowing us to test different models for sleep
mechanisms. We can build a model supposing that for each day the time of sleep
are totally independent, or we can assume that they are related to a hierarchical
parameter representing the baseline times of sleep/awake for a person. Similarly
we can assume that the awake rate is completely independent for each day, or
we can assume that it comes from an underlying distribution characterized by
a base rate. Moreover the Bayesian approach can incorporate prior knowledge
for the rates and the typical times of sleep/awake into the model, and provides
an estimation of the uncertainty for the parameters. Using the model we can
produce detailed sleep periods estimates for all users. We compare the inferred
sleep schedules with the ground truth taken from armband sleep trackers worn
by the Sony subjects. We find that the model is able to determine sleep patterns
with a mean accuracy of 0.89, and a mean F1 score of 0.83.

Fig. 3.7 illustrates the sleep schedule inferred by our model for 4 users, representative of different sleep behaviors. Each row represents one day, and each column represents one time bin. The blue shading represents the probability that the user is asleep in that timebin: darker color indicates higher probability. The radius of the red dots shows the number of screen-on events per bin. User A has a quite regular schedule, waking up around 07:00 except on some days (presumably weekends). User B has also a pretty regular schedule, waking up around 06:00 most days, but he has a few smartphone interactions at night – presumably when he temporarily wakes up. User C has a much more unstable schedule, with no fixed pattern for waking up, many long nights, and a lot of smartphone interactions during the day. Finally User D has very noisy data, with few interactions interleaved during day and night – the model cannot really conclude much about his sleep schedule.

From a large number of individual sleep schedules we can derive collective patterns of sleep. Fig. 3.8 shows the aggregated sleep times during the week. It is possible to notice the normal routine of sleep during weekdays, and the delayed sleep and wakeup schedule for weekends.

This section has shown how the inference of behavior for many single individual can be aggregated to understand trends in collective behavior. The next chapter will focus precisely on how to infer collective behavior in a specific context: human mobility.
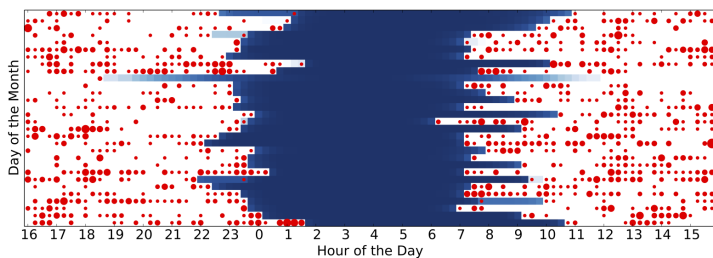
(a) User A
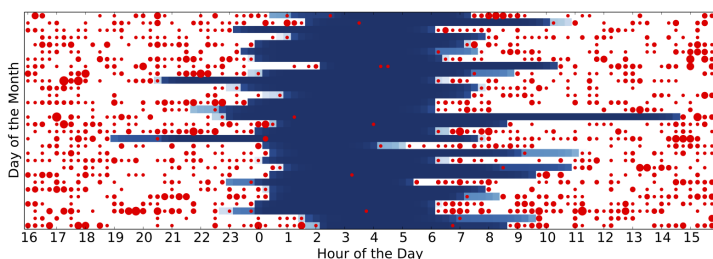
(b) User B
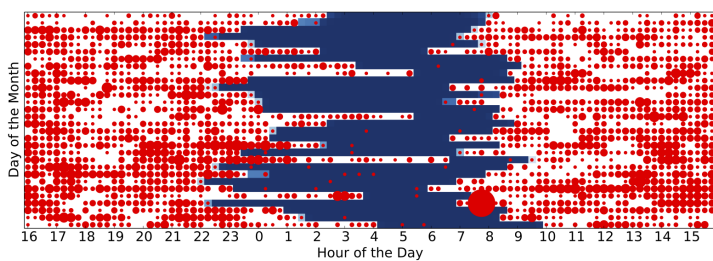
(c) User C
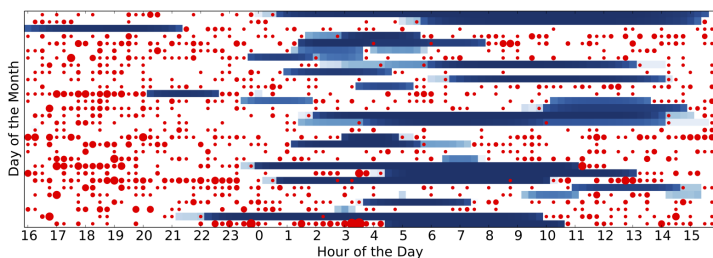
(d) User D

Figure 3.7: The inferred sleep schedules for 4 users. Each row represents one day, and each column represents one time bin. The blue shading represents the probability that the user is asleep in that timebin: darker color indicates higher probability. The radius of the red dots shows the number of screen-on events per bin.
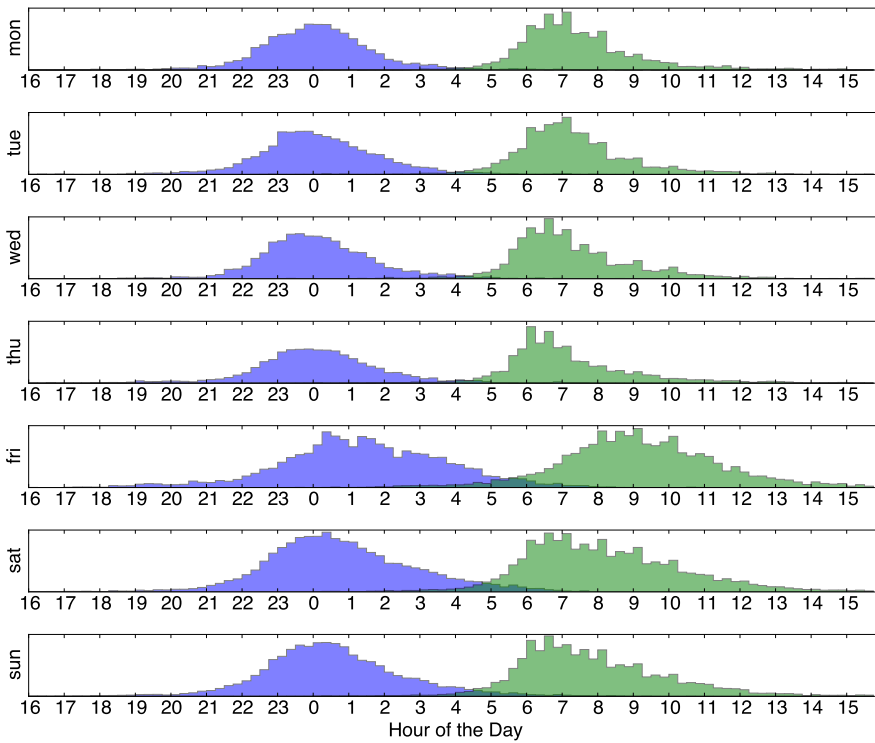
Figure 3.8: The aggregated sleep times for all users during the days of the week (Figure from Appendix E).

Understanding Collective Behavior: Human Mobility

"Two roads diverged in a wood,
and I – I took the one less traveled by"
ROBERT FROST

In the previous chapter we focused on using personal data to understand individual behavior. In this chapter we try instead to look at the trends in aggregated individual behaviors, and we focus on one specific aspect: human mobility. This facet of behavior really has a special importance: everything happens in a specific geographical location. The spatial context often influences other aspects of our behavior. We meet different people depending on our location: colleagues at work, family at home, friends at a bar. Our activities are also place-dependent: we only workout at the gym, sleep at home, play with our smartphone on the bus. Movement patterns are also a very personal behavior, which can be used to uniquely identify people [27].

Other than the significance at the personal level, understanding large-scale mobility patterns has important applications in many fields. In traffic management and urban planning, mobility models can be used to forecast and optimize the traffic within cities, or better understand the urban structure [136, 137, 138].

In anticipatory computing, a predictive model could guess your next location and provide context-related information such as a commute timetable, opening hours, traffic estimation, reminders or alternative routes [139]. Advertisers may be interested in such systems for providing ads targeted to the next place [140]. In mobile computing environments, predicting the next user location is used to optimize the allocation of infrastructural resources [141]. Co-location is a fundamental requirement for spreading of epidemics, and consequently a better understanding of mobility could lead to better prevention and containment strategies [142, 143, 144, 145, 146]. Movement of populations is critical during and immediately after natural disasters, therefore being able to model changes in mobility can lead to better emergency response [147, 148].

Human mobility is a very vast field, and here we focus on three areas: place recognition, mobility properties and next-place prediction. The rest of the chapter is organized as follows. Section 4.1 gives an overview on how the data can be acquired using different systems for measuring location. Section 4.2 focuses on smartphones as positioning systems, and discusses the challenges of working with location data. Section 4.3 describes an open-source toolbox to visualize geographical data, developed as part of this dissertation. Section 4.4 provides an overview of the problem of identifying places from location traces, including the most relevant techniques from literature. Section 4.5 describes a paper where we study the feasibility of inferring human mobility from sparse low accuracy data collected with smartphones. Section 4.6 provides an overview of the statistical physics approaches for understanding human mobility properties. Section 4.7 describes a paper where we propose a hierarchical model for human mobility and we analyze human mobility properties at multiple scales. Section 4.8 gives an overview of the field of next-place prediction, and the most important models from literature. Finally Section 4.9 describes a paper where we investigate which factors influence the accuracy of next-place prediction.

4.1   Positioning Systems

The geographic coordinate system is the reference system for describing any point on Earth using latitude, longitude (and altitude). The latitude represents the north/south angle from equator, while the longitude represents the west/east angle from the Greenwich meridian. A positioning system estimates the location on Earth using beacons, which are reference points with known location. This measurement is typically done by calculating the time needed for the beacon signal to travel, or the strength of the signal. There are three main technologies for beacons: GPS satellites, cell towers, and Wi-Fi.

GPS (Global Positioning System) is a technology that provides a position estimation based on satellites communication. The system is composed of 24 satellites orbiting Earth, and continuously transmitting their position and reference time. Any GPS receiver needs to obtain the signal from at least 4 satellites, and can then calculate its own position based on the time that took the signals to propagate[1]. GPS estimations can be very precise, up to a few meters for consumer devices. However GPS requires line of sight to satellites, which is not available indoors and for some weather conditions. Moreover obtaining an initial fix on the satellites requires some initial time.

Cell towers are the antennas used to transmit the phone network signal. Since mobile phones need to be connected to cell towers all the time to make and receive phone calls, cell towers have become a pervasive positioning system. The issue with cell tower positioning is that is often very inaccurate, since a single cell tower may cover an area of hundreds of meters, or even kilometers in less populated regions. On the other hand, the cell towers usage for millions of users routinely recorded by telecommunication providers in form of Call Details Records (CDRs) presents an incredibly rich dataset for studying human mobility, and in fact some of the most significant mobility studies have been done on CDRs.

Wi-Fi is a technology for connecting to a wireless Local Area Network. Virtually any modern electronic communication device (phone, tablet, watch, laptop, PC) has a Wi-Fi receiver, and most buildings (homes, offices, libraries, restaurants, etc.) have Wi-Fi routers, each one broadcasting a unique identifier. This means that in densely populated areas there are always some Wi-Fi access points present. A proof of the pervasive coverage of Wi-Fi has been given by Sapiezynski et al. [60], who show that only using 20 Wi-Fi access points it is possible to reconstruct people's location 90% of the time. Wi-Fi positioning has the advantage over GPS to work indoors, and compared to cell-towers Wi-Fi has typically a much shorter range (tens of meters), therefore providing a more accurate estimation of position.

4.2   Positioning Using Smartphones

All modern smartphones provide tools for estimating the current user location. In this section we will refer to the *Android Location Service* [149] that has been used in the SensibleDTU study (Section 2.6), but similar principles apply to

---

[1]this calculation is quite complex, and it has to even take into consideration the Special and General Relativity due to the fact that satellites are at a significant distance from Earth's gravity.

other families of smartphones.

The Android API provides methods for requesting the last known location, and to receive updates when the estimation improves or the location changes. The location estimation is the guess that the phone can provide using either GPS or a *fused* sensor based on Wi-Fi and cell towers positioning. Google does not officially provide information on how the estimation for the fused sensor is done. However it has been reported that Google (and other major mobile phone companies) have been building for years an extensive map of the Wi-Fi and cell towers positions by cross-referencing them with the GPS data from mobile phones [150, 151], so this is the likely mechanism behind it.

One location sample from the Android API contains a latitude and longitude in decimal degrees, a provider (GPS or fused) and an accuracy r in meters. The accuracy is defined so that if we draw a circle of radius r around the current estimation, there is 68% probability that the true location is inside the circle [152]. Therefore each location sample has an intrinsic uncertainty of positioning, which can make it difficult to determine the exact location for example among different buildings. We analyze here a representative subsample of over 20,000,000 locations in order to illustrate the characteristics of the collected data.

Due to battery consumption, the Android API prefers to use the fused sensor instead of GPS whenever possible. In this subset of records, only 9% of the samples are from GPS. Fig. 4.1 shows the distribution of reported accuracy, divided by provider. Most of the GPS samples have accuracy better than 20-30 meters, and most of the fused samples have accuracy better than 40 meters.

As discussed before, GPS is available only outdoors while the Wi-Fi positioning is available almost everywhere. The accuracy of estimations has a spatial meaning, as location samples collected indoors tend to less accurate (higher radius of uncertainty). Fig. 4.2 shows a map of the median accuracy of the Android location samples over the campus at the Technical University of Denmark. The white color represents good accuracy of < 10 meters for the street and outdoors. The orange color represents accuracy of 20-30 meters in the outer parts of the buildings. The red color represents accuracy of > 30 meters, in the inner parts of buildings.

It is interesting to see the behavior of the location sensor over time. Fig. 4.3 shows the accuracy and the movement over one day for one user. The top subplot shows how the accuracy over time fluctuates around 20-60 meters. The bottom subplot shows the sample-by-sample displacements. Even during night hours when the phone is presumably stationary, the location samples are randomly distributed around due to noise. Around 9.00, 12.00 and 18.00 bigger travel distances are measured, indicating that the user was probably commuting.
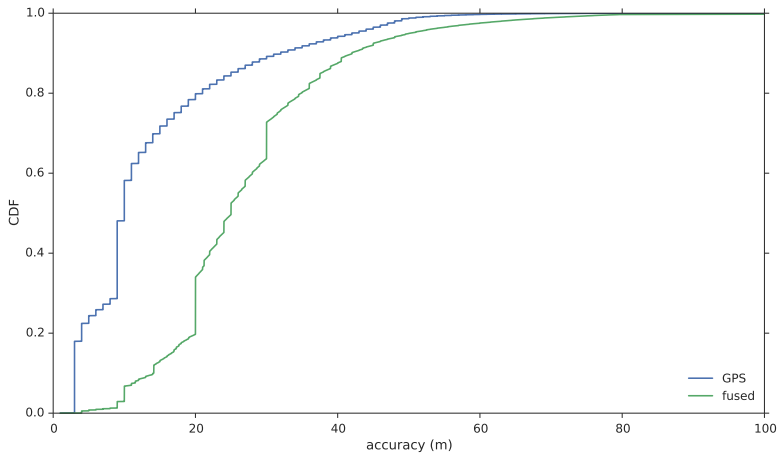
Figure 4.1: Accuracy (in meters) of over 20,000,000 location samples.
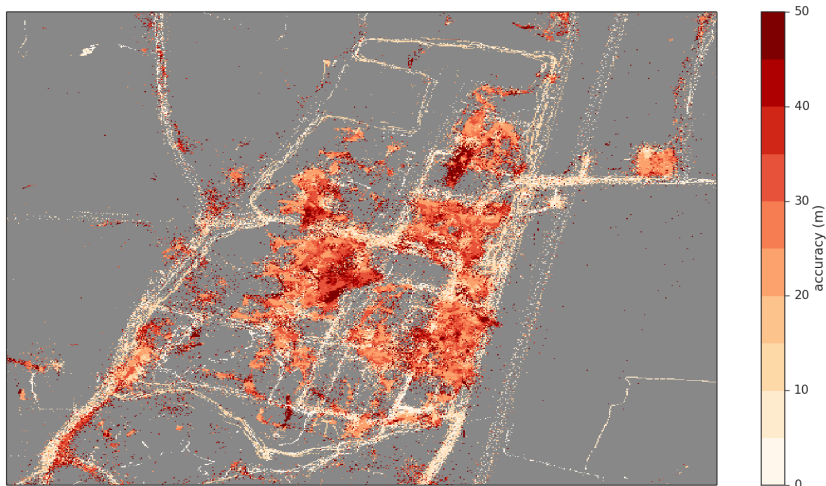
Figure 4.2: Map of the median accuracy (in meters) over the campus at the Technical University of Denmark.
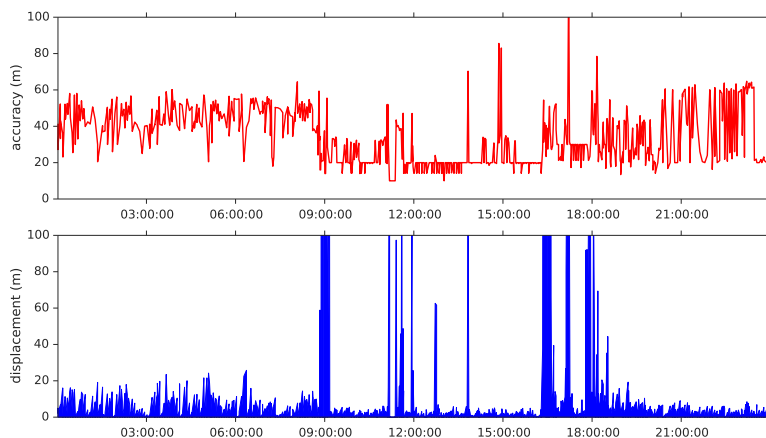
Figure 4.3: Accuracy and sample-by-sample displacement over one day for one
user.

4.3 Paper Summary: "geoplotlib: a Python Toolbox for Visualizing Geographical Data"

Analyzing large-scale data is a complex task and in many cases the analysis can
be facilitated by creating exploratory visualizations [112, 153]. In the context
of geographical data, maps are the natural visual representations. During our
work with geographical data we found that surprisingly there is a lack of tools
for creating maps that integrate well with the python programming language,
the tool of choice for many data analysts (including us).

Looking to related work, the matplotlib [154] library provides many visualization
tools, but it does not support geographical maps by default. Some extensions
such as Basemap [155] and Cartopy [156] provide geographical projections but
do not support map tiles. Another alternatives is to generate custom code in
Javascript/HTML either manually or using some third-party libraries such as
Google Maps [157], Leaflet [158] and OpenLayers [159]. Finally, Geographic
Information Systems such as QGIS [160] and ARCGIS [161] could also be con-
sidered. These solutions however are sub-optimal, since they require the data
to be exported and do not have full integration with python.

To fill this need we implemented geoplotlib, an open-source python toolbox

for visualizing geographical data. geoplotlib is designed according to three key principles:

- **simplicity:** geoplotlib provides a set of built-in visualization techniques such as dot maps, kernel density estimation, spatial graphs, Voronoi tessellation and shapefiles. The API is inspired by the matplotlib [154] library (the de-facto standard for data visualization in python) to make it easier for python users to get started.

- **integration:** geoplotlib visualizations are scripts in pure python, so they can use any other python code or package. This supports full integration with the many available libraries for scientific computing, machine learning and numerical analysis.

- **performance:** geoplotlib uses numpy for fast numerical computations and OpenGL for hardware-accelerated graphical rendering, and is able to visualize millions of points in real time.

The software has been released as open-source on github [162], and has received a moderate attention with hundreds of downloads and over 360 favorites as of July 2016. The paper in Appendix F describes in details the architecture, the use cases and the features of geoplotlib.

4.4   Place Recognition

Location expressed as latitude and longitude can be a very precise measurement, but this representation is difficult for humans to understand. Saying that you are at position (48.858342, 2.294449) is not as eloquent as to say that you are under the Eiffel tower. Humans tend to think spatially in terms of places, that is a discrete set of spatial units with well-defined contextual meaning. The transformation from raw mobility traces to discrete places is also a needed step for other tasks in mobility analysis. In order to predict transitions between places, it is needed to define how to extract places from mobility traces. Another example is that for studying the mobility properties such as frequency of visits or staying times, it is firstly needed to identify what the places are. Finally, context-aware applications must have a way to recognize different places and distinguish for example your home from your favorite coffee shop. Place recognition algorithms can be divided into two main categories: geometry-based and fingerprinting-based.

4.4.1 Geometry-based

Geometry-based place recognition takes a history of absolute locations (typically in latitude/longitude format) and defines places in terms of set of points or geometric shapes such as rectangles or circles.

One very simple way to distinguish between places is to define an arbitrary discretization of the surface of Earth, and consider each unit as a different place. The most elementary way to do this is to define a rectangular grid on geographical coordinates space, and consider each grid cell as a different place. Fig. 4.4 shows an example of converting mobility traces into places using a grid. This solution has a number of problems, since is not clear how large the grid cells should be. If cells are too large, then one single cell may contain multiple places; if cells are too small they may split a single place into two. No single choice of size can really fit everything, as places can be very small (a single house) or very large (a shopping mall, a stadium). Despite these problems, the grid binning solution is in some case a good enough approximation, and in fact it has been used in many works [163, 164, 165].

timestamp	lat	lon	cell_id
1379412900	55.784555	12.517756	55784,12517
1379431800	55.738143	12.449360	55738,12449
1379502900	55.783958	12.518652	55783,12518
1379519100	55.763086	12.477459	55763,12477
1379523600	55.738135	12.449390	55738,12449
1379553300	55.739645	12.444714	55739,12444
1379565000	55.738165	12.449322	55738,12449
1379588400	55.783921	12.517496	55783,12517

Figure 4.4: An example of converting mobility traces into places using a grid (taken from Appendix I).

An improvement over the naive grid-binning has been suggested by Zheng et al. [166]. The proposed algorithms starts by defining as before a grid over the map and assigning points to the grid cells. Then the algorithm considers each of the 8 neighboring cells, and it merges the non-empty ones. This method has the advantage of being able to capture places of different sizes and shapes, although it can only have rectangular boundaries.

A heuristic for identifying places is based on the GPS limitation of losing signal within building. Therefore significant places can be identified as these spots

where signal is lost for a significant amount of time. This technique has been suggested by Marmasse et al. [167] and Ashbrook et al. [168].

The problem of assigning locations to places can be quite naturally framed as a data clustering problem. What we want to achieve is to assign locations to places so that locations belonging to one place are more similar to each other than to locations belonging to a different place. In other words, given a set of points in the latitude-longitude coordinate space, we want to assign each point to a cluster so that points in the same cluster are more similar than to points in other clusters. Not surprisingly, a lot of literature has been dedicated to apply clustering techniques for identifying places from location traces.

A clustering method that would be suitable for this task should not need any prior knowledge of the number of clusters, since the number of significant places changes by person and over time. For this reason, the density-based clustering algorithms have been one of tools of choice, in particular DBSCAN [169] and OPTICS [170]. The principle behind density-base clustering is that each point is at the beginning assigned to his own cluster, and then clusters "grow" and merge together as long as the points are within a distance threshold ϵ. The main parameter needed is the distance threshold that specify how far can points be to still belong to the same cluster. Numerous studies have adopted this density-based clustering method, and we cite as examples [171, 172, 173, 174, 175]. Related to this approach is the work of Kang et al. [176], who propose an online algorithm for place recognition that sequentially groups locations over time, and determines the membership of place according to a spatial distance d and a time difference t.

Fig. 4.5 shows an example of identifying places as clusters of stops using density-based clustering (taken from Appendix I). In the left panel each stop is represented as a red dot. In the right panel the result of the clustering is shown: each cluster corresponding to place is represented by a different color.

4.4.2 Fingerprinting-based

Fingerprinting-based place recognition uses a completely different approach in defining places. Instead of requiring the precise geographical positioning in terms of latitude and longitude, this approach defines a place with a unique fingerprint in form of beacons: cell towers, Wi-Fi, or Bluetooth.

For cell towers, the simplest assumption is that there is a one-to-one correspondence between each place and a cell tower. This effectively divides the geographical area into a Voronoi tessellation [177], so that each point on the

Figure 4.5: An example of identifying places as clusters of stops using density-based clustering (taken from Appendix I).
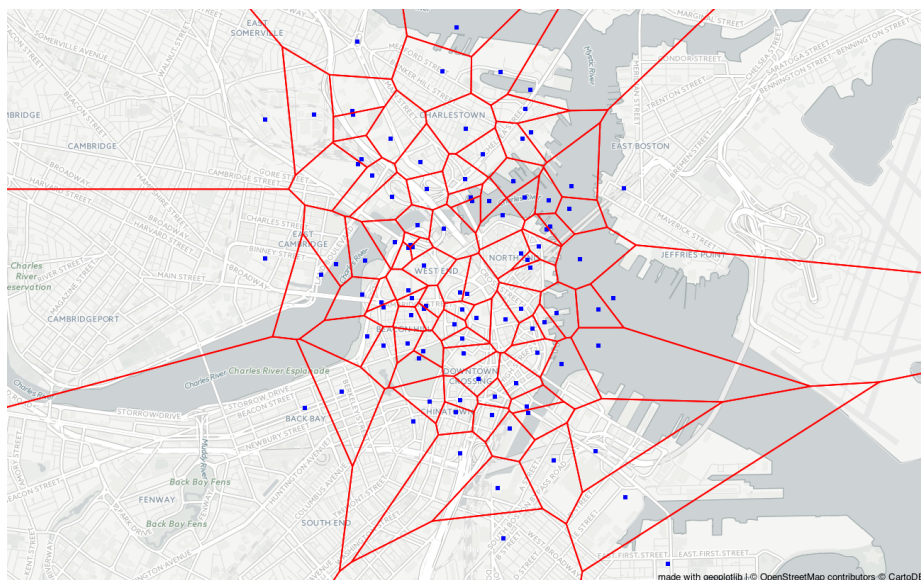
Figure 4.6: An example of assigning location samples (blue dots) to Voronoi cells (red lines).

map belongs to the nearest cell tower. Fig. 4.6 shows an example of the assignment of locations to Voronoi cells. Laasonen et al. [178] have suggested an improvement by grouping nearby towers together, to link towers that belong to the same logical place.

Fingerprinting techniques for Wi-Fi typically define each place by the set of visible access points, or their signal strengths. For example Hightower et al. [179] propose scanning the Wi-Fi access points in a fixed time window to identify changes in the set of visible access points. A similar approach is proposed by Kim et al. [180]. Wind et al. [181] show how a greedy search for selecting the most frequently seen Wi-Fi access points is able to detect significant places.

Bluetooth is a short-distance radio transmission technology, commonly available in electronic communication devices. Bluetooth devices have also been used as beacons for positioning [182], especially indoors. However Bluetooth as positioning system is much less common, due to the Bluetooth devices being much less pervasive than Wi-Fi and cell towers.

4.4.3   Evaluation

One major challenge for place recognition literature is the validation of the results. The set of places and visits inferred by a place recognition system must be validated against the actual behavior of people, and this typically requires to collect the ground truth in form of travel diaries, or to interview participants and ask feedback about the results. This presents all the usual limitations related to manual data collection: it is costly, time-consuming, and error-prone. Therefore experimental results in literature are limited to studies on tens to hundreds of people at most, and for typical periods of few weeks.

4.5   Paper Summary: "Inferring Human Mobility from Sparse Low Accuracy Mobile Sensing Data"

As discussed before, the problem of place recognition is well studied in literature, and many different strategies have been proposed. However in many experimental settings the sampling rate is typically as high as one sample every few minutes [171, 172], or even every few seconds [168, 174, 167]. Moreover much of the literature focuses on high-accuracy GPS estimations from mobile phones or even dedicated devices.

Such measurement settings may not be realistic in a large-scale data collection study. In the SensibleDTU study (Section 2.6), location samples were acquired using Android smartphones. The location sensing was limited to one sample

every 5-15 minutes, to avoid excessive battery drain and to reduce the size of the collected data. Moreover the location collection system provided by the Android smartphones has typically less accuracy than high-precision GPS (see Section 4.2).

In this paper (Appendix G) we therefore investigate the feasibility of inferring human mobility patterns from this more sparse, lower accuracy mobile sensing data. In particular we investigate the performance of several techniques to detect stop locations and identify places. We recruited six students at our campus, and we gave them a Nexus 4 with a data collector app installed (the same setup as the SensibleDTU experiment). Participants were instructed to write down a travel diary of the places visited during the day, with the name, the time of arrival and departure. One researcher also collected his own location data and kept his own diary.

We test a number of algorithms for place recognition: distance grouping, speed thresholding, Gaussian Mixtures Model and DBSCAN clustering. The performance of the place recognition techniques is evaluated using the f_1 score between the inferred places and the ground truth places the diaries. Finally, we qualitatively inspect the geographical distribution of the inferred places and the patterns of visitations. Our results indicate that the place recognition task produces reasonable results, despite the lower quality of the data.

4.6   Mobility Properties

Human behavior can be very complex, and finding general laws that perfectly describe individual choices may be an almost impossible task. However if we look at aggregated trends of behavior of a large population, some general properties emerge. This is the approach taken by statistical physics [183], which was originally applied for describing physical phenomena (such as the interaction between particles) using statistical tools. Statistical physics applied to human mobility looks at properties at the aggregated level, and tries to describe mobility behaviors such as travel distance, stay duration, diffusion and return patterns as stochastic processes. In this section we describe some of the key results obtained by this approach.

4.6.1   Travel Distance and Diffusion

In **The Scaling Laws of Human Travel** [184] Brockmann et al. analyze over 1 million reports of geographical location of bank notes, and calculate their spatial displacements. The distribution of displacements decays as a power law, therefore the authors suggest that human movement is well approximated with Lévy flights[2]. A Lévy flight [188] is a random walk where the step size Δd follows a power-law distribution:

$$P(\Delta d) \sim \Delta d^{-(1+\beta)} \tag{4.1}$$

with $\beta < 2$.
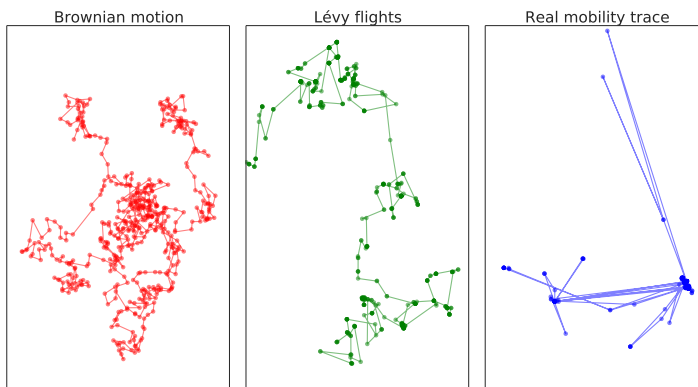
Figure 4.7: Comparison between simulated Brownian motion, simulated Lévy flights and a real mobility trace from one individual. The Brownian motion is composed by regular steps of approximate same size. The Lévy flights exhibit both short and long jumps. The real mobility trace also has both short and long displacements, but is also characterized by multiple returns to the same places.

This model for human behavior also makes a lot of intuitive sense: we typically travel very short distances (home-work) but sometimes we travel farther (to another city) and occasionally very far (to another country).

Moreover the authors find that the dispersion of notes is much slower than

---

[2]Interestingly enough, Lévy flights also describe trajectories of monkeys [185], albatrosses [186] and marine predators [187].

expected for a Lévy flight. Therefore they investigate different possible reasons, and finally suggest that this is caused by the heavy-tailed waiting time between movements, which slows down the diffusion process.

In **On the Levy-walk Nature of Human Mobility** [189] Rhee et al. confirm the results from Brockmann, applying a similar analysis on GPS traces from 101 volunteers in five different environments: two university campuses, the New York City metropolitan area, and one theme park and one state fair. The mobility traces of people are found to fit well as Lévy flights. The distribution of displacements and waiting times are both heavy tailed, and a number of distributions fits are tested. Moreover it is shown that users have super-diffusion up to 30-60 min and subdiffusion after that.

In **Understanding Individual Human Mobility Patterns** [19] González et al. analyze the position of 100,000 users from Call Detail Records for six months. They find that the displacements between consecutive locations follow a truncated power-law. Moreover they define the radius of gyration r_g as measure of the characteristic travel distance for a user:

$$r_g = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (r_i - r_{cm})^2} \qquad (4.2)$$

where r_i is the i-th position, and r_{cm} is the center of mass position. The radius of gyration is also distributed as a truncated power-law, and over time it increases more slowly than the random walk model would predict.

The authors suggest that the reason of this behavior is the regularity of human mobility: people tend to return to a few familiar locations with high probability, and with specific frequency of days or weeks. In particular the probability of returning to a location ranked L in the frequency of visits follows Zipf's law [190]: $P(L) \sim 1/L$. Finally the paper shows that after rescaling and correcting for anisotropy, the individual spatial probability distributions collapse into a single distribution.

4.6.2   Predictability and Exploration

In **Modelling the Scaling Properties of Human Mobility** [21], the authors discuss how modeling human mobility as Lévy flights or Continuous-Time Random Walk (CTRW) fails to reproduce some key properties: a slower than expected diffusion, the Zipf's law for frequency of visits, and the saturation of

the number of newly explored locations over time. They propose a mechanism to explain this behavior. At each step, a person decides whether to explore a new location that has never been visited before with probability $P_{new} = \rho S^{-\gamma}$, or return to an already visited location with the complementary probability $P_{ret} = 1 - \rho S^{-\gamma}$. When returning to a known location, the probability of choosing each location is proportional to the frequency of previous visits. This way frequently locations tend to be visited more and more – a mechanism similar to preferential attachment in networks [191]. The authors fit this model to a Call Detail Records dataset of 50,000 users for three months. By combining this mechanism with the heavy-tailed travel distances and waiting times, the model is able to fit much better the mobility traces.

In **Limits of Predictability in Human Mobility** [20], Song et al. propose a new method, based on Information Theory, for quantifying the regularity and predictability of human mobility. The authors apply this framework to a Call Detail Records dataset of 50,000 users for three months. Each user's mobility trace is composed by the sequence of hourly time bins, each of them containing the cell tower ID that the user is connected to. Their idea is to measure the Shannon entropy, that is a quantification of the (dis)order of the series of locations visits. In particular they define three types of entropy: S^{rand}, S^{unc}, and S:

$$S^{rand} = \log_2 n \tag{4.3}$$

$$S^{unc} = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{4.4}$$

$$S = \left(\frac{1}{n} \sum_i \Lambda_i \right)^{-1} \ln n \tag{4.5}$$

S^{rand} represents the entropy when choosing one of the n locations uniformly at random, thus ignoring both the temporal sequence information and the frequency of visits. S^{unc} represents the time-uncorrelated entropy, obtained by ignoring just the temporal sequence information and applying the entropy formula on the frequencies of locations visits. Finally S represents the real entropy of the time series, which accounts both for the temporal sequence information and the frequency of visits. S is estimated using a process similar to the Lempel-Ziv data compression [192]: for each of the n steps we find Λ_i, the length of the shortest substring starting at position i that does not previously appear from position 1 to $i-1$. By construction $0 \leq S \leq S^{unc} \leq S^{rand} < \infty$.

The entropy of mobility has an interesting interpretation. $S = 0$ indicates no

uncertainty at all, meaning that the next location is completely predictable. In this work the authors show that for most people $S^{rand} \approx 6$, indicating that on average one user location update represents six bits of information – or in other words a user can be found on average in one of $2^6 = 64$ locations. In contrast $S \approx 0.8$, indicating that a user can be found in fewer than two ($2^{0.8} \approx 1.74$) locations.

Finally, the authors show how to transform the entropy value into the upper bound of the predictability performance of an ideal algorithm using Fano's inequality. Surprisingly, the distribution for all users of this upper bound of predictability is tightly centered at around 93%, indicating that most people's mobility patterns are very regular and predictable.

The idea of the limits of predictability for human mobility has been further explored by Lu et al. [148] who study the predictability of the population of Haiti after the earthquake in 2010, by Lin et al. [193] who investigate the effects of spatial and temporal resolution, and by Smith et al. [194] who consider the spatial constraints for next-place transitions.

In **Returners and Explorers Dichotomy in Human Mobility** [195], Pappalardo et al. discover the existence of two types of mobility profiles: returners and explorers. The former limit their mobility to a few places, while the latter do not. In order to quantify this difference the authors define k-gyration $r_g^{(k)}$ as the radius of gyration limited to the top k places. The ratio $r_g^{(k)}/r_g$ across the population under study shows two peaks: around 1 for returners whose total gyration corresponds to the gyration of their top k places, and around 0 for explorers whose gyration cannot be described in such as way. Finally a modified exploration and preferential return model [21] is proposed to incorporate the explorers/returners dichotomy, which better fits the data.

In **Unravelling Daily Human Mobility Motifs** [196], Schneider et al. apply the concept of network motifs – subnetworks that occur more often than in randomized versions of the entire network – to individual transitions between places. The authors find that only 17 unique network configurations are sufficient to explain 90% of the mobility patterns of the population, and each individual is characterized by a number of motifs that remain stable over several months. The most common motif is composed by two visited locations and transitions between them; the second most common motif with only one location; the third and fourth most common motifs are characterized by three locations and four transitions, either starting and ending at the same location, or with one stop in the middle.

4.7   Paper Summary: "Measuring Human Mobility at Multiple Scales"

In Section 4.4 we discussed the problem of extracting places from mobility data, and how this is typically solved by using clustering methods or by imposing geometric boundaries. Much of the literature in this area has pragmatically focused on finding places depending on the problem formulation or the available data sources, without much discussion about the actual definition of a place.

But what is a *place*? If we start to think about it, the answer is not so simple. Right now I am sitting in my office, but I am also inside my university campus, in Copenhagen, and in Denmark. The definition of place is complex, and depends on the context and ultimately on the scale we are considering. Places are spatially nested and form a hierarchy. Scales can be found even on old-fashioned paper maps: there are specific maps for country, for cities and so on.

The idea of the hierarchical organization of human places has been proposed in other fields, such as Geography [197] and Environmental Psychology [198]. As described in Section 4.6, much work has been done to measure the many properties of human mobility but little attention has been given to the effect of geographical scale. In this paper (Appendix H) we build a model to capture the hierarchical organization of human mobility and to measure mobility properties at multiple scales.

We analyze the location data from SensibleDTU (Section 2.6). For each user individually we find stops as sequence of locations where the user has been approximately stationary, that is within a distance threshold approximately corresponding to the GPS accuracy. We then group stops into places with a recursive clustering based on DBSCAN [169]. We first cluster stops at a very large distance ($\epsilon = 150$ kilometers), which produces clusters corresponding to country-level partitions. We then recursively apply clustering at a distance $\epsilon = 5000$ meters to find city-level partitions, and then again at distance $\epsilon = 500$ meters for neighborhood-level and $\epsilon = 20$ meters for building-level. The clusters do not precisely correspond to the exact administrative boundaries, but they capture well the geographical scales.

Fig. 4.8 shows an example for one user. Each cluster is depicted as a circle, and each scale has a different color: blue for building, green for neighborhoods, orange for cities and red for countries. The left panel shows the several buildings (blue) within one neighborhood (green). The center panel shows the neighborhood (green) within a city (orange). Finally the right panel shows the city (orange) within a country (red).

Figure 4.8: An example of the hierarchical structure identified by the model. Each cluster is represented by a circle, and each color represents a different scale: blue for building, green for neighborhood, orange for city and red for country.

The model allows us to analyze mobility properties at multiple scales. We can for example measure the fraction of time spent at different places (Fig. 4.9). The fraction of time rapidly decreases as a function of place rank, and the decay is stronger for larger scales. These results are in line with the well-known properties of long-tailed staying time and preferential return [21], but in addition show the effect of different scales.
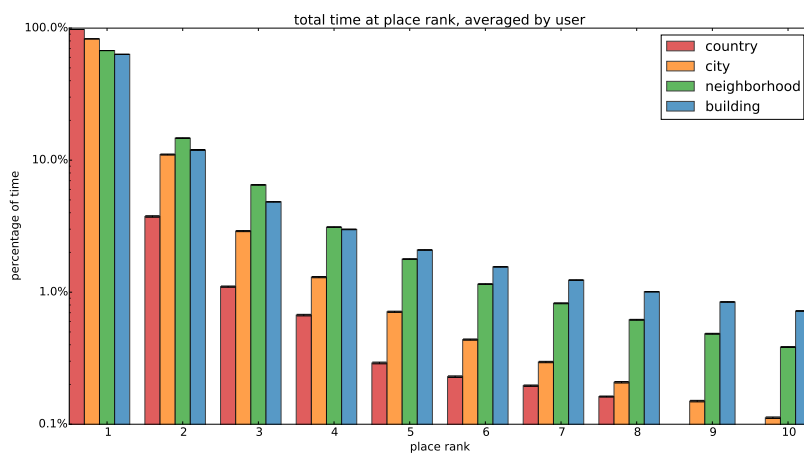

Figure 4.9: The fraction of time versus place rank for multiple scales (from Appendix H).

We also measure a number of other mobility properties. The cumulative fraction of time show that, depending on the scale, a different number of places are needed to cover the majority of the time. The probability of returning to places displays two strong periodic patterns, one every 24 hours and one every 7 days. We measure the exploration behavior over the year, and show how it is possible to detect periods of higher exploration in correspondence to holidays. We can quantify the navigation in the spatial hierarchy, for example distinguishing the probability of transitions between buildings in the same neighborhood and across neighborhoods. In all cases we show how the different scales have an effect on the mobility properties, and sometimes the hierarchical model can capture details that would be lost in a flat model.

4.8   Next-Place Prediction

Predictive models for human mobility try to answer the question "where will you go next?". The underlying assumption for any predictive model is that modeled behavior is deterministic, that is the same conditions always produce the same output. In the context of human mobility this means that your past location history and your context perfectly determine your next movement. Even though human behavior is extremely complex and it may not be considered completely deterministic, human mobility is indeed driven by routine. People tend to follow the same schedule week after week: go to the same workplace on weekdays, to the same gym on specific days of the week, to the supermarket near home after work, to their favorite bar on Fridays. Yet it is easy to image how there are many "exceptions to the rule" in this routine. Unlike robots, we never follow exactly the same plan. We may come to work later on day, and leave early another, we may skip the gym a few times, or change after a while our favorite restaurant. Despite this uncertainty, mobility models have managed to obtain accuracy as high as 90% in some contexts.

Typically the prediction is done by choosing the next place among a finite set of possible places. The discretization of mobility traces into places can be done in various ways, as discussed in Section 4.4. Predicting the next place could mean predicting the next grid cell, a geometric boundary, or a cell tower for example. The problem is then formulated as a supervised learning task: the model is trained on a subset of the data in form of sequence of locations and contextual information (time, sensor data, etc.), and tries to learn which patterns are predictive of the next location.

We now review some of the key ideas in predictive models for next-place prediction.

4.8.1   Markov Chain Models

One family of predictive models is based on the idea of Markov chains. A Markov chain is a random process that transitions between discrete states, and the probability of each transition depends only on the current state, and not on any of the previous ones. This "absence of memory" is called Markov property. In other words, for a sequence of random variables X_1, X_2, X_3, ... the probability of X_{n+1} depends only on X_n:

$$\Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} = x \mid X_n = x_n)$$
$$(4.6)$$

An extension to the simple Markov chain is to have an order m Markov model – or Markov(m). A Markov(m) model has a memory of size m, so that the current state depends on the previous m states. The simple Markov chain model can then be seen as a Markov(1) model.

In the context of human mobility, a Markov chain model assumes that the next visited location depends only on the current location. This assumption may seem an excessive simplification at first – our mobility patterns are for sure much more complex! – but in practice it has worked well on many different datasets.

Although the reasons for the success of the Markov models are not completely clear, it is possible to speculate why it is so. One reason is probably that the model can correctly predict the most frequent transitions home-work-home. Going back home is also the most likely transition for the majority of the other places. Moreover there are many occasions where people follow the Markov property: we may visit a cinema only after going to a specific restaurant, or go to our parents' house only after stopping for a while at the station on the way.

The simplest way to build a Markov chain model from a sequence of transitions between places is the following. Let each state $x_1, ..., x_m$ represent a place. The transition probability from x_i to x_j is the number of times that the user has traveled from x_i to x_j over the total transitions from x_i. Fig. 4.10 visualizes as a graph the Markov chain model fitted on an individual's mobility trace, limited to transitions appearing more than once. Each place is represented by a node, and the size of each node is proportional to the number of visits to that place. Each link represents a non-zero probability of a transition between two places, so the absence of a link indicates that no transitions are possible between two places. The large blue node represents the home location, where most visits take

place. The home node has most of the links both incoming and outgoing, since most trips are from and to home, and the other nodes are mainly connected to the home location.
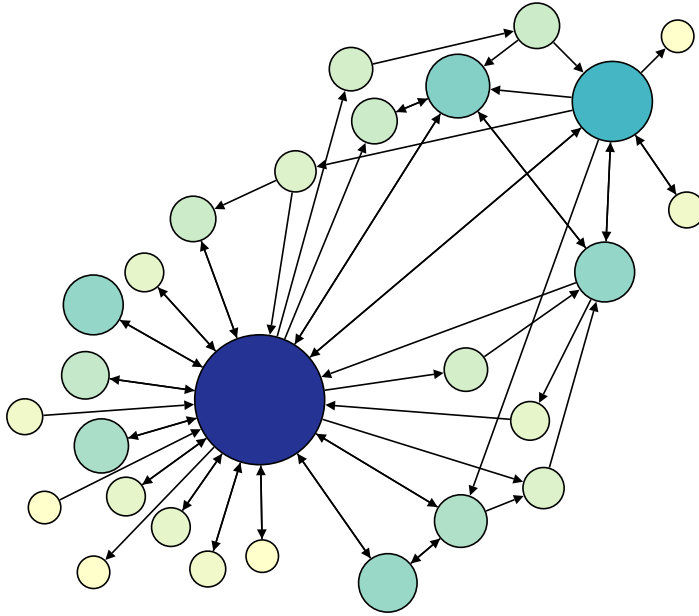
Figure 4.10: A Markov chain model of transitions between places for a single individual.

We now review some of the main prior work on next-place prediction using Markov chains.

Evaluating Next-cell Predictors with Extensive Wi-Fi Mobility Data [199] by Song et al. was one of the first large-scale studies on next-place prediction. The authors study the problem of predicting the next location for over 6000 users at the Dartmouth University campus. Locations are represented by the Wi-Fi access points that users are connected to. The authors test Markov chain models of different orders, and find that the Markov(2) model has the best performance (72% median accuracy) – better than higher order models. The Markov(2) model performs even better than more complex models based on Lempel-Ziv, Prediction by Partial Matching and Sampled Pattern Matching predictors. Finally the authors show that the accuracy increases for longer trace lengths, and is negatively correlated with the entropy of the location history.

In **Approaching the Limit of Predictability in Human Mobility** [200], Lu et al. apply the Markov chain model to Call Detail Records data from 500,000 users in Ivory Coast, West Africa. Their problem formulation consists in predicting for each day the last visited prefecture. In this context the Markov(1) model performs extremely well with accuracy as high as 87-95%, better than higher-order models. The authors however note that a large part of this predictive power comes from the high stationarity of mobility, since most people spent the vast majority of their time in the most visited location.

In **Next Place Prediction using Mobility Markov Chains** [201], Gambs et al. develop a Markov chain model with 3 states: home, work and other. The Markov(2) model performs best achieving an accuracy ranging from 70% to 95%.

In **A Variable Order Markov Model Approach for Mobility Prediction** [202], Bapierre et al. propose a Variable-Order Markov Model (VOMM) that supports a flexible number of states for prediction. The model consider as states combinations of location labels, day of the week and time of the day (morning, afternoon, evening). The model is evaluated on the Geolife [203] and Reality Mining [49] datasets and achieves a mean accuracy of 0.65 without temporal context, and 0.78 with temporal context.

4.8.2   Conditional Probability Models

The Markov chain models we have discussed make a very strong simplification: they assume that the next location depends only on the previous locations. However our experience tells us that this is rarely the case. A large number of other factors play a role in our decision of where to go next. The most obvious one is time: depending on our schedule we may go to one place or another. But other factors may also play a role: are we with friends or alone? Did we plan something special for today? Are we sick? Of course not everything can be promptly measured, but mobile sensing does provide some means for measuring at least a subset of these factors, for example:

- Current location can be extracted as discussed before

- Semantic information about location (home, work, shop) can be inferred from the data, or mined from Points of Interest databases

- Time can be determined by the smartphone clock

- Social proximity (are we alone or in a group?) can be determined by Bluetooth, Wi-Fi or GPS proximity

- Social activity in terms of communication can be recorded by calls and SMS

It is reasonable to expect that a combination of these features can provide better predictions than location alone. We can formalize the prediction using this context information as follows. At each step, we want to infer the most probable next location given the context variables discussed above. In other words we want to compute $P(\hat{L}|c_1, c_2, c_3, ..., c_n)$, where $\hat{L}$ is the next location, and c_1, c_2, c_n are the variables representing different contexts. One way to solve this is by using Bayes' rule:

$$P(\hat{L}|c_1, c_2, ..., c_n) = \frac{P(\hat{L})P(c_1, c_2, ..., c_n|\hat{L})}{P(c_1, c_2, ..., c_n)} \propto P(\hat{L})P(c_1, c_2, ..., c_n|\hat{L}) \quad (4.7)$$

and then assuming conditional independence:

$$P(\hat{L})P(c_1, c_2, ..., c_n|\hat{L}) = P(\hat{L})P(c_1|\hat{L})P(c_2|\hat{L})...P(c_n|\hat{L}) \quad (4.8)$$

which corresponds to using a naive Bayes classifier [204]. The conditional independence assumption is almost always wrong, as predictors are often correlated. In this context it is easy to imagine that for example social activity, place, time are all somehow related. Yet naive Bayes models often perform surprisingly well despite the limitation of this assumption [205]. We describe here a few representative papers utilizing the naive Bayes model, or a similar conditional probability framework.

In **Predestination: Inferring Destinations from Partial Trajectories** [165], Krumm and Horvitz consider the problem of predicting car trips destinations for a rectangular grid in the Seattle area. The data is collected for 169 subjects for two weeks. They use a naive Bayes model that combines the prior probability trips lengths, driving efficiency and grid cell destinations. The model is able to predict the destination with a median error of two kilometers.

In **Mobile Location Prediction in Spatio-Temporal Context** [206], Gao et al. propose a model that estimates the probability of visiting the next place by combining the probability of coming from the current place, the probability of visit by hour of the day and day of the week. To compensate for the sparsity of the data when estimating the conditional probabilities, the model assumes

that users visits follow a Gaussian distribution over time. The model is tested
on location data from 80 users for one year from the Nokia Mobile Data Chal-
lenge dataset [51]. The proposed model achieves an accuracy of approx. 50%,
outperforming the Markov chain models.

In **Contextual Conditional Models for Smartphone-based Human Mo-
bility Prediction** [207], Do et al. consider an ensemble method to calculate the
join probability distribution for the next location by assigning a weight to each
individual conditional probabilities. The model considers as context features
the current location, time of the day, day of the week, weekend, frequency and
durations of visits, number of nearby Bluetooth devices, and SMS/call informa-
tion. Their model is fit on data for 153 users during 17 months from the Nokia
Mobile Data Challenge dataset [51]. The best performing model has accuracy
of approx. 60% by using the location and hour features.

In **A Probabilistic Kernel Method for Human Mobility Prediction
with Smartphone** [208], Do et al. consider two novel prediction tasks: esti-
mating the most probable location at a specific time in the future, and within
a fixed time window. The proposed model combines the previous location in-
formation with temporal information such as time of the day, day of the week,
weekend, etc. In this case in order to overcome the sparsity of the data when es-
timating the conditional probabilities, a kernel density estimation is performed
with different types of kernels. The model is tested on locations from 133 users
from the Nokia Mobile Data Challenge dataset [51]. The proposed method
achieves accuracy of 84% for the next hour, and an accuracy of 77% for the
next three hours, in both cases outperforming the Markov models.

4.8.3   Social Co-location Models

Another category of models is based on the idea that mobility and social inter-
actions are tightly connected. Our location often determines whom we interact
with: we meet colleagues at work, relatives at home, friends at a bar. Therefore
our social contacts can become a proxy for our own mobility: using the location
of your friends, your partner or relatives could be used to infer something about
your location. This methodology is particularly attractive for studies made on
location-enabled social networks (such as Twitter, Foursquare, Facebook), where
users can report both their location and friendship status with other users. In
this context, a location is a venue where the user *checks-in*, that is updates its
status to say that he has been at that location. The task is then to predict
the next check-in location. We now provide a few examples of papers utilizing
social co-location for mobility prediction.

In **Friendship and Mobility: User Movement in Location-based Social Networks** [209], Cho et al. analyze data from Gowalla and Brightkite location-based social networks. The authors first note that only 4.1% of all check-ins in Brightkite and 9.6% of all check-ins in Gowalla were first visited by a friend and then by the user, suggesting the limitations of social-based mobility prediction alone. The paper then proposes to model user locations with two elements: one individual part based on routine and a social part based on the location of friends. The individual part is modeled as a mixture of two Gaussians centered at home and work locations. The social component is modeled to choose a friend's location depending on the time elapsed from the last check-in and the distance of the venue. The proposed model is able to predict the correct check-in location approximately 40% of the time.

In **Finding Your Friends and Following Them to Where You Are** [10], Sadilek et al. analyze location and friendship data from Twitter for one month of in New York City and Los Angeles, USA. The location prediction is performed using a Dynamic Bayesian Network, where the hidden state is the current user location, and the observable states represent the locations of n friends, the time of the day and day of the week. The accuracy for the prediction is around 50% using location from only one friend, and approx. 80% using location from two or more friends.

In **Location Prediction: Communities Speak Louder than Friends** [210], Pang et al. analyze 18 months of Gowalla data in New York City, Los Angeles and San Francisco, USA. The novelty of their work is to detect communities in the social friendship graph, and then use the communities to define predictive features, including: geographical distance between centroids of communities, number of users in the community, number of frequent movement areas, total number of check-ins, ratio between the number of edges in the community and the maximal number of possible edges. The authors apply a logistic regression model using the described features and obtain an accuracy of 67-81%. Finally they conclude how communities have a stronger impact on users' mobility than individual friends, and different communities have influence in different spatial and temporal contexts.

4.8.4   Other Models

Finally we cite here some other notable approaches for next-place prediction that do not directly implement an intuitive mobility mechanism, but use more sophisticated models.

In **Eigenbehaviors: Identifying Structure in Routine** [211], Eagle and

Pentland study the locations of 100 users from the Reality Mining dataset. Locations are encoded into a number of days times 24 hours matrix, where each element encodes one of the following labels: Home, Elsewhere, Work, No Signal, Off. The authors apply Principal Component Analysis (PCA) to the individual location matrices. The PCA components are referred as *eigenbehaviors*, the fundamental building blocks of human activity. Just six of these components are sufficient to approximate with 90% accuracy the original data. The six main components are used for prediction, by fitting the PCA on the first 12 hours of each day and then predicting the labels for the next 12 hours. This method achieves average accuracy of 79%.

In **Far Out: Predicting Long-Term Human Mobility** [212], Sadilek et al. consider the problem of predicting mobility patterns in the far future. For each user, the possible visited locations are the 10 most visited cells (400 meter size) plus one extra cell for the rest of the visits. Each day is represented by a vector encoding for each hour the probability of being in cell i, for a total of $24 \cdot 11$ cells. Seven binary indicators for the weekday and one binary indicator for holidays are finally added for each day. Principal Component Analysis is applied on the days vectors to find *eigendays*, fundamental days representative of all mobility. Ten components are sufficient for reconstructing the original data with 90% accuracy. The model is then used to perform a prediction by learning weights of the components on the train data, and predicting for the test set. The model is tested on 307 people and 396 vehicles, and shows accuracy of approx. 80%.

In **NextPlace: A Spatio-Temporal Prediction Framework for Pervasive Systems** [213], Scellato et al. propose a method based on non-linear time series analysis. The model has the particular feature of not predicting the next location, but instead the time to the next visit and the duration of the visit. The model is tested on a variety of datasets including mobile GPS locations, cabs positions San Francisco, and Wi-Fi access points. The accuracy varies depending on the time horizon, from approx. 60-90% for 5 minutes in the future, to 40-70% in one hour, to 30-50% in 4 hours. In all cases the non-linear model outperforms the linear version and the Markov chain baselines.

4.9 Paper Summary: "Understanding Predictability and Exploration in Human Mobility"

As we have seen, the performance of predictive models in literature varies quite broadly, from as high as 93% to as low as under 40%. Even though we started this section by defining the next-place prediction problem as simply described by

the question "where will you go next?", the literature review actually shows the many complexities of this problem. There are in fact many aspects to take into account. One issue is once again the definition of *place*: are we predicting the next grid cell, the next check-in, the next county, the next Wi-Fi access point? And moreover are we considering only the most commonly visited places, and if so, how many? Another aspect is temporal: are we predicting your position in the next few minutes? Or in hours? Or days? Or the next transition? And also: are we considering periods long enough to capture the long-term trends of mobility? Finally, what kind of data is being used: cell tower positions, Wi-Fi connections, social network check-ins, GPS? Due to these many different experimental settings, it is often not possible to directly compare different model performances, and it is not clear how the actual prediction performance is affected by different factors.

In this paper (Appendix I) we therefore propose to measure the effects of different factors on the prediction performance. We consider the data from the SensibleDTU study, in particular we select 454 users that have between 3 months and one year of contiguous location data.

We start by distinguishing between two fundamentally different problem formulations: *next-cell* and *next-place*. In *next-cell* we consider grid cells as different places, and the task is to predict the grid cell in the next 15-minute timebin. In the *next-place* formulation, we extract the sequence of visits at places and the task is to predict the transition to the next place.

For the *next-cell* formulation, we implement three models: a "toploc" model that always predicts the most common grid cell seen so far, a Markov(1) model, and a stationary model that predicts to stay in the current cell. The models obtain average accuracy of 0.4, 0.7 and 0.7 respectively. In particular the stationary model performance is very strongly correlated with the Markov model performance, suggesting that much of the predictive power in this setting is due to self-transitions. We study the impact of spatial resolution on the predictive performance by considering different cell sizes, and we find that prediction performance is better for larger cell sizes, suggesting that the choice of the spatial units has a role in the prediction performance. We also study the effect of the temporal resolution by considering larger timebins, and we find that the predictive performance is better for smaller timebins due to the increase in self-transitions.

We then consider the *next-place* formulation, where for each user we extract the sequence of visits at places using a density-based clustering. In this case we are therefore not interested in the next timebin location, but in the next transition to another place. When fitting this model to the data, we obtain an accuracy of approx. 0.4 for the Markov(1), considerably lower than the performance on the

next-cell formulation. This underlines how strong the impact of the problem statement is, and that the task of predicting the next transition is quite more difficult than predicting the next timebin. We also measure the predictive power of different mobile sensing features using a logistic regression model, which in the best case still achieves a low accuracy (less than 0.45).

We want then to investigate what are the reasons of this difficulty in prediction. We find that exploration plays a significant role in the mobility of the population under study. We can measure that on average users discover a new location in 20-25% of the transitions (Fig. 4.11). This represents a twofold problem. First, transitions to newly explored places cannot be predicted by any model that selects the candidate next place from the already visited ones, which is the case for most of the models described in literature. Second, the consequence of the frequent exploration is that the pool of possible places grows over time reaching almost 200 places for one year. We also measure that, as a consequence of preferential return to the most visited locations, approx. 70% of the locations are visited only once (Fig. 4.12). The combination of these factors makes the problem of selecting of the next place very difficult even for non-explorations. Finally, given the importance of exploration in human mobility, we consider the novel task of predicting when an exploration will happen, and we show that a logistic regression model is able to predict exploration with f_1 score of 0.4.
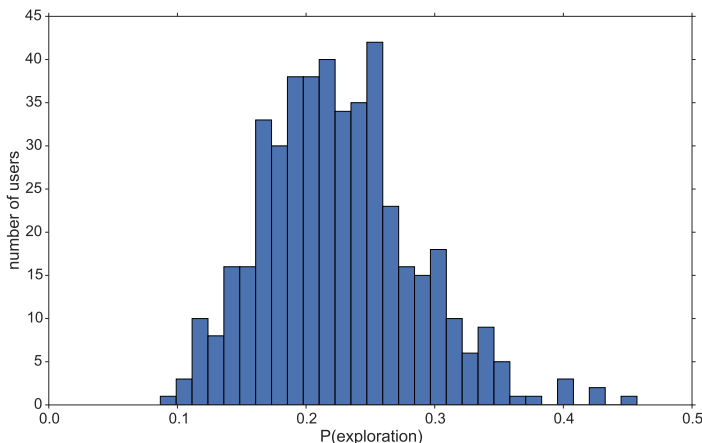
Figure 4.11: The probability of exploration of a new place is approx. 20-25%, that is one out of 4-5 transitions is an exploration (Figure from Appendix I).
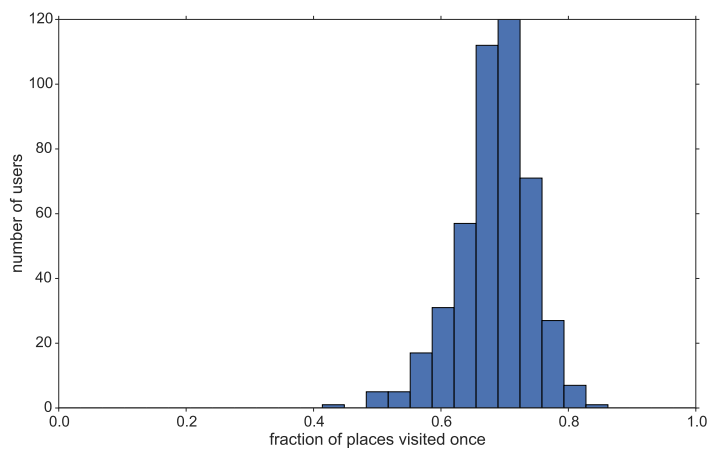
Figure 4.12: About 70% of places are visited only once, due to the preferential return to the most visited locations (Figure from Appendix I).

Conclusions

This chapter contains a summary of the contributions, the lesson learned and finally some suggestions for future work.

5.1 Contributions

In this thesis I have presented a number of tools and approaches for understanding human behavior through data. Here is a short summary of the contributions:

- I have described the SensibleDTU project, a large-scale deployment for collecting high-resolution data on human behavior using smartphones. This experiment has pushed forward the state of the art in terms of number of participants, quality of the data, and longitudinality. We have described the methods, the challenges and the solutions adopted, and we have made available the whole software stack as open-source. We hope that this work may serve as inspiration for future studies, and facilitate the deployments of even larger projects.

- I have described two user studies of large-scale deployment of self-tracking apps, in order to understand the patterns of usage and non-usage. Our work is one of the first to look at large-scale deployment of self-tracking tools, and contributes to better understand the real world challenges of developing and deploying Personal Informatics systems in the wild.

- I have described a model to capture human mobility at multiple scales, and a study of the factors influencing predictability in human mobility. These studies provide some novel views of human mobility, namely the importance of considering the scales when studying mobility properties, and a description of the many factors determining mobility behavior.

- As part of this work I have also developed a tool for visualizing geographical data using python, released as open-source software.

5.2 Lesson Learned

The path leading to the results presented in this dissertation was not a straightforward one, but it was instead full of challenges. In part these issues are personal and specific to my own experience, but I think many of these problems are common for those who work on the topics covered in this dissertation, in particular Data Science and Computational Social Science. Here I provide a brief summary of such challenges and lesson learned.

I believe the first lesson is that human behavior is highly complex. Most mathematical models are based on the idea that the system under study is deterministic and obeys specific laws, or stochastic and can be described by certain statistical properties. Elementary particles, chemical reactions, physical phenomena, and biological processes can be precisely described by mathematical formulas and laws. But human behavior cannot be modeled in this way (at least with our current understanding), since there is a large number of factors that comes into play to determine our behavior. This makes the creation of quantitative models for human behavior a much more complex task. Another reason for the difficulty of modeling human behavior is that our sources of information are often limited. For example studying communication patterns using only email or phone calls gives us a really limited view on the actual social behavior. Another example is social network check-ins or antenna positioning, as these represent only a coarse approximation of mobility behavior. Given these limitations, it is actually surprising that we indeed find a lot of regularity in human behavior, and we can (to an extent) describe or even predict its patterns.

A second lesson learned is that interdisciplinarity is fundamental. The topic of this dissertation is the use of computational tools for modeling human behavior. The work therefore falls into two fields: from one hand the development of tools and methods for analyzing data, and on the other the development and verification of theories for human behavior. Coming from a Computer Science background, my expertise was much more in the computational methods than in the Social Sciences theories. On the other hand I know several students from the Social Sciences that were bringing to the table a lot of expertise in terms of social theories and ideas, but had some problems working with data and developing software. This challenge is very much a central issue in Computational Social Science. It is important therefore for scholars from both fields to collaborate as much as possible to push forward the field.

5.3   Future Work

As always in science, obtaining answers to a problem opens the door to new and more complex questions. From the work presented in this thesis, I see two main issues for future work.

Our work on Personal Informatics systems has discussed the challenges and issues related to the usage and non-usage of self-tracking tools. In particular we have described how the systems failed to engage a large part of the population, and those who were engaged quickly lost their interest. Nowadays we have access to incredibly rich behavioral data, from movement to social contacts, physical activity, heartbeat, sleep patterns and much more. Therefore we have the potential of quantifying with unprecedented precision many aspects of life. Unfortunately as it has emerged from our surveys, human behavior is intrinsically repetitive and therefore in a way *boring*. For example people tend to travel home-work-home on most days, and meet the same people over and over. The main question that remains is: how to *design* self-tracking systems that are both interesting and useful, despite this repetitiveness? I believe this question is of fundamental importance for the success of any self-tracking tool, from the small-scale research prototypes to billion-dollar Quantified-Self apps.

For the field of human mobility, I think the most interesting question that emerged is the problem of scales. In this dissertation we have shown that mobility properties do change at different scales, and in some cases mobility behavior can be captured only thinking in terms of these scales. We have seen that the issue of scales (spatial resolution) is also present when defining the predictive power of next-place prediction models. For this work we have however set the size and number of scales manually, according to our human intuition of

the right scales. A fascinating question that remains is how to find the scales of human mobility in the data. Are there characteristic behaviors for each scale? Other than the administrative boundaries, what mobility patterns define a neighborhood or a city? Answering these questions can lead to understanding the underlying mechanisms generating the scale-free properties of human mobility, and provide fundamentally new views about our mobility patterns.

Measuring Large-Scale Social Networks with High Resolution

Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, **Andrea Cuttone**, Mette My Madsen, Jakob Eg Larsen, Sune Lehmann.

Measuring Large-Scale Social Networks with High Resolution

Arkadiusz Stopczynski[1]*, Vedran Sekara[1], Piotr Sapiezynski[1], Andrea Cuttone[1], Mette My Madsen[3], Jakob Eg Larsen[1], Sune Lehmann[1,2]

1 DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark, 2 The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark, 3 Department of Anthropology, University of Copenhagen, Copenhagen, Denmark

Abstract

This paper describes the deployment of a large-scale study designed to measure human interactions across a variety of communication channels, with high temporal resolution and spanning multiple years—the Copenhagen Networks Study. Specifically, we collect data on face-to-face interactions, telecommunication, social networks, location, and background information (personality, demographics, health, politics) for a densely connected population of 1 000 individuals, using state-of-the-art smartphones as social sensors. Here we provide an overview of the related work and describe the motivation and research agenda driving the study. Additionally, the paper details the data-types measured, and the technical infrastructure in terms of both backend and phone software, as well as an outline of the deployment procedures. We document the participant privacy procedures and their underlying principles. The paper is concluded with early results from data analysis, illustrating the importance of multi-channel high-resolution approach to data collection.

Introduction

Driven by the ubiquitous availability of data and inexpensive data storage capabilities, the concept of big data has permeated the public discourse and led to surprising insights across the sciences and humanities [1,2]. While collecting data may be relatively easy, it is a challenge to combine datasets from multiple sources. This is in part due to mundane practical issues, such as matching up noisy and incomplete data, and in part due to complex legal and moral issues connected to data ownership and privacy, since many datasets contain sensitive data regarding individuals [3]. As a consequence, most large datasets are currently locked in 'silos', owned by governments or private companies, and in this sense the big data we use today are 'shallow'—only a single or very few channels are typically examined.

Such shallow data limit the results we can hope to generate from analyzing these large datasets. We argue below (in Motivations Section) that in terms of understanding of human social networks, such shallow big data sets are not sufficient to push the boundaries in certain areas. The reason is that human social interactions take place across various communication channels; we seamlessly and routinely connect to the same individuals using face-to-face communication, phone calls, text messages, social networks (such as Facebook and Twitter), emails, and many other platforms. Our hypothesis is that, in order to understand social networks, we must study communication across these many channels that are currently siloed. Existing big data approaches have typically

concentrated on large populations ($\mathcal{O}(10^5) - \mathcal{O}(10^8)$), but with a relatively low number of bits per participant, for example in call detail records (CDR) studies [4] or Twitter analysis [5]. Here, we are interested in capturing deeper data, looking at multiple channels from sizable populations. Using big data collection and analysis techniques that can scale in number of participants, we show how to start deep, i.e. with detailed information about every single study participant, and then scale up to very large populations.

We are not only interested in collecting deep data from a large, highly connected population, but we also aim to create a dataset that is collected interactively, allowing us to change the collection process. This enables us to rapidly adapt and change our collection methods if current data, for example, have insufficient temporal resolution with regard to a specific question we would like to answer. We have designed our data collection setup in such a way that we are able to deploy experiments. We have done this because we know that causal inference is notoriously complicated in network settings [6]. Moreover, our design allows us to perform continuous quality control of the data collected. The mindset of real-time data access can be extended beyond pure research, monitoring data quality and performing interventions. Using the methods described here, we can potentially use big data in real time to observe and react to the processes taking place across entire societies. In order to achieve this goal, researchers must approach the data in the same way large Internet services do—as a

resource that can be manipulated and made available in real time as this kind of data inevitably loses value over time.

In order to realize the interactive data collection, we need to build long-lasting testbeds to rapidly deploy experiments, while still retaining access to all the data collected hitherto. Human beings are not static; our behavior, our networks, our thinking change over time [7,8]. To be able to analyze and understand changes over long time scales, we need longitudinal data, available not just to a single group of researchers, but to changing teams of researchers who work with an evolving set of ideas, hypotheses, and perspectives. Ultimately, we aim to be able to access the data containing the entire life-experience of people and look at their lives as dynamic processes. Eventually, we aim to even go beyond the lifespan of individuals and analyze the data of the entire generations. We are not there yet, but we are moving in this direction. For example, today, all tweets are archived in the Library of Congress (https://blog.twitter.com/2010/tweet-preservation), a person born today in a developed country has a good chance of keeping every single picture they ever take, the next generation will have a good chance of keeping highly detailed life-log, including, for example, every single electronic message they have ever exchanged with their friends. The status quo is that we need to actively opt out if we want to prevent our experiences from being auto-shared: major cloud storage providers offer auto-upload feature for pictures taken with a smartphone, every song we listen to on Spotify is remembered and used to build our profile—unless we actively turn on private mode.

In this paper, we describe a large-scale study that observes the lives of students through multiple channels—the Copenhagen Network Study. With its iterative approach to deployments, this study provides an example of an interdisciplinary approach. We collect data from multiple sources, including questionnaires, online social networks, and smartphones handed out to the students. Data from all of these channels are used to create a multi-layered view of the individuals, their networks, and their environments. These views can then be examined separately, and jointly, by researchers from different fields. We are building the Copenhagen Networks Study as a framework for long-lived extensible studies. The 2012 and 2013 deployments described here are called *SensibleDTU* and are based at the Technical University of Denmark. They have been designed as part of the *Social Fabric* project (see Acknowledgements for details) in close collaboration with researchers from the social sciences, natural sciences, medicine (public health), and the humanities. We are currently in the second iteration where we have deployed phones to about 1 000 participants, enabling us to compile a dataset of unprecedented size and resolution. In addition to the core task of collecting deep behavioral data, we also experiment with creating rich services for our participants and improving privacy practices.

Human lives, especially when seen over a period of months and years, take place in multiple dimensions. Capturing only a single channel, even for the entire life of an individual, limits the knowledge that can be applied to understand a human being. True interdisciplinary studies require deep data. Anthropologists, economists, philosophers, physicists, psychologists, public health researchers, sociologists, and computational social science researchers are all interested in distinct questions, and traditionally use very different methods. We believe that it is when these groups start working together, qualitatively better findings can be made.

Here we give a brief overview of the related work, in the domains of data collection and analysis, extend the description of the motivation driving the project, and outline the experimental plan and data collection methodology. We report on privacy and informed consent practices that are used in the study, emphasizing

how we went beyond the usual practice in such studies and created some cutting edge solutions in the domain. We also report a few initial results from the project, primarily in the form of an overview of collected data, and outline future directions. We hope the work presented here will serve as a guideline for deploying similar massive sensor-driven human-data collection studies. With the overview of the collected data, we extend an invitation to researches of all fields to contact the authors for the purpose of defining novel projects around the Copenhagen Networks Study testbed.

Related Work

Lazer et al. introduced computational social science (CSS) as a new field of research that studies individuals and groups in order to understand populations, organizations, and societies using big data, i.e. phone call records, GPS traces, credit card transactions, webpage visits, emails, and data from social networks [9]. CSS focuses on questions that can now be studied using data-driven computational analyses of datasets such as the ones mentioned above, and which could only previously be addressed as self-reported data or direct observations, for example dynamics in work groups, face-to-face interactions, human mobility, or information spreading. The hope is that such a data-driven approach will bring new types of insight that are not available using traditional methods. The challenges that emerge in this set of new approaches include wrangling big data, applying network analysis to dynamic networks, ensuring privacy of personal information, and enabling interdisciplinary work between computer science and social science, to name just a few.

In this section we describe related work in terms of the central methods of data collection. Furthermore, we provide a brief overview of results obtained from the analysis of CSS data, and finally, mention some principles regarding privacy and data treatment.

Data collection

Many of the CSS studies carried out to date have been performed on call detail records (CDRs), which are records of phone calls and messages collected by mobile phone operators. Although CDRs can be a proxy for mobility and social interaction [10], much of the social interaction happens face-to-face, and may therefore be difficult to capture with CDRs or other channels such as social networks (Twitter, Facebook, etc.) [11]. To gain a fuller view of participants' behavior, some CSS studies have developed an approach of employing Radio Frequency Identification (RFID) devices [12], sociometric badges [13,14], as well as smartphones for the data collection [15–18]. Smartphones are unobtrusive, relatively cheap, feature a plethora of embedded sensors, and tend to travel nearly everywhere with their users. They allow for automatic collection of sensor data including GPS, WiFi, Bluetooth, calls, SMS, battery, and application usage [19]. However, collecting data with smartphones presents several limitations as sensing is mainly limited to pre-installed sensors, which may not be of highest quality. Furthermore, off-the-shelf software and hardware may not be sufficiently robust for longitudinal studies.

A large number of solutions for sensor-driven human data collection have been developed, ranging from dedicated software to complete platforms, notably ContextPhone [20], SocioXensor [21], MyExperience [22], Anonysense [23], CenceMe [24], Cityware [25], Darwin phones [26], Vita [27], and ContextTool-box [28].

Running longitudinal rich behavioral data collection from large populations presents multiple logistical challenges and only few

studies have attempted to do this so far. In the Reality Mining study, data from 100 mobile phones were collected over a nine-month period [29]. In the Social fMRI study, 130 participants carried smartphones running the Funf mobile software [30] for 15 months [31]. Data was also collected from Facebook, credit card transactions, and surveys were pushed to the participants' phones. The Lausanne Data Collection Campaign [32,33] featured 170 volunteers in the Lausanne area of Switzerland, between October 2009 and March 2011. In the SensibleOrganization study [34], researchers used RFID tags for a period of one month to collect face-to-face interactions of 22 employees working in a real organization. Preliminary results from the OtaSizzle study covering 20 participants from a large university campus have been reported [35]. Finally, in the Locaccino study [36], location within a metropolitan region was recorded for 489 participants for varying periods, ranging from seven days to several months.

Data analysis

In the following, we provide selected examples of results obtained from analysis of CSS datasets in various domains.

Human Mobility. Gonzales et al. analyzed six months of CDRs of 100 000 users. Their results revealed that human mobility is quite predictable, with high spatial and temporal regularity, and few highly frequented locations [37]. Their findings were further explored by Song et al., who analyzed three months of CDRs from 50 000 individuals and found a 93% upper bound of predictability of human mobility. This figure applies to most users regardless of different travel patterns and demographics [38]. Sevtsuk et al. focused instead on the aggregate usage of 398 cell towers, describing the hourly, daily, and weekly patterns and their relation to demographics and city structure [39]. Bagrow et al. analyzed 34 weeks of CDRs for 90 000 users, identifying habitats (groups of related places) and found that the majority of individuals in their dataset had between 5 and 20 habitats [40]. De Domenico et al. showed in [41] how location prediction can be performed using multivariate non-linear time series prediction, and how accuracy can be improved considering the geo-spatial movement of other users with correlated mobility patterns.

Social Interactions. Face-to-face interactions can be used to model social ties over time and organizational rhythms in response to events [29,42,43]. Comparing these interactions with Facebook networks, Cranshaw et al. found that meetings in locations of high entropy (featuring a diverse set of visitors) are less indicative than meetings in locations visited by a small set of users [36]. Clauset et al. found that a natural time scale of face-to-face social networks is 4 hours [44].

Onnela et al. analyzed CDRs from 3.9 million users [45] and found evidence supporting the weak ties hypothesis [46]. Lambiotte et al. analyzed CDRs from 2 million users and found that the probability of the existence of the links decreases as d^{-2}, where d is the distance between users [47]. In another study with CDRs from 3.4 million users, the probability was found to decrease as $d^{-1.5}$ [48]. Analyzing CDRs for 2 million users, Hidalgo et al. found that persistent links tend to be reciprocal and associated with low degree nodes [49].

Miritello et al. analyzed CDRs for 20 million people and observed that individuals have a finite limit of number of active ties, and two different strategies for social communication [50,51]. Sun et al. analyzed 20 million bus trips made by about 55% of the Singapore population and found distinct temporal patterns of regular encounters between strangers, resulting in a co-presence network across the entire metropolitan area [52].

Health and Public Safety. Using CDRs from the period of the 2008 earthquake in Rwanda, Kapoor et al. created a model for detection of the earthquake, the estimation of the epicenter, and determination of regions requiring relief efforts [53]. Aharony et al. performed and evaluated a fitness activity intervention with different reward schemes, based on face-to-face interactions [31], while Madan et al. studied how different illnesses (common cold, depression, anxiety) manifest themselves in common mobile-sensed features (WiFi, location, Bluetooth) and the effect of social exposure on obesity [54]. Salathé et al. showed that disease models simulated on top of proximity data obtained from a high school are in good agreement with the level of absenteeism during an influenza season [55], and emphasize that contact data is required to design effective immunization strategies.

Influence and Information Spread. Chronis et al. [16] and Madan et al. [56] investigated how face-to-face interactions affect political opinions. Wang et al. reported on the spread of viruses in mobile networks; Bluetooth viruses can have a very slow growth but can spread over time to a large portion of the network, while MMS viruses can have an explosive growth but their spread is limited to sub-networks [57]. Aharony et al. analyzed the usage of mobile apps in relation to face-to-face interactions and found that more face-to-face interaction increases the number of common applications [31]. Using RFID for sensing face-to-face interactions, Isella et al. estimated the most probable vehicles for infection propagation [58]. Using a similar technique, however applied to 232 children and 10 teachers in a primary school, Stehle et al. described a strong age homophily in the interactions between children [59].

Bagrow et al. showed how CDR communications, in relation to entertainment events (e.g. concerts, sporting events) and emergencies (e.g. fires, storms, earthquakes), have two well-distinguishable patterns in human movement [60]. Karsai et al. analyzed CDR from six millions users and found that strong ties tend to constrain the information spread within localized groups of individuals [61].

Studies of Christakis and Fowler on the spread of obesity and smoking in networks [62,63] prompted a lively debate on how homophily and influence are confounded. Lyons was critical toward the statistical methods used [64]. Stelich et al. discussed how friendship formation in a dynamic network based on homophily can be mistaken for influence [65], and Shalizi and Thomas showed examples of how homophily and influence can be confounded [6]. Finally, Aral et al. provided a generalized statistical framework for distinguishing peer-to-peer influence from homophily in dynamic networks [66].

Socioeconomics and Organizational Behavior. For employees in a real work environment, face-to-face contact and email communication can be used to predict job satisfaction and group work quality [34]. Having more diverse social connections is correlated with economic opportunities, as found in the study containing CDRs of over 65 million users [67]. A similar result was reported in a study of economic status and physical proximity, where a direct correlation between more social interaction diversity and better financial status was found [31]. Or, as shown in a study of Belgian users, language regions in a country can be identified based solely on CDRs [68].

Privacy

Data collected about human participants is sensitive and ensuring privacy of the participants is a fundamental requirement—even when participants may have limited understanding of the implications of data sharing [69,70]. A significant amount of literature exists regarding the possible attacks that can be performed on personal data, such as unauthorized analysis [71] with a view to decoding daily routines [72] or friendships [42] of

the participants. In *side channel information* attacks, data from public datasets (e.g. online social networks) are used to re-identify users [73–75]. Even connecting the different records of one user within the same system can compromise privacy [73]. Specific attacks are also possible in network data, as nodes can be identified based on the network structure and attributes of the neighbors [76,77].

Various de-identification techniques can be applied to the data. *Personally Identifiable Information* (PII) is any information that can be used to identify an individual, such as name, address, social security number, date and place of birth, employment, education, or financial status. In order to avoid re-identification and consequent malicious usage of data, PII can be completely removed, hidden by aggregation, or transformed to be less identifiable, resulting in a trade-off between privacy and utility [78]. Substituting PII with the correspondent one-way hash allows removal of plaintext information and breaks the link to other datasets. This method, however, does not guarantee protection from re-identification [79–82]. $K-$anonymity is a technique of ensuring that it is not possible to distinguish any user from at least $k-1$ other in the dataset [83]; studies have shown that this method often may be too weak [72]. $L-$diversity [84] and $t-$closeness [85] have been proposed as extensions of $k-$anonymity with stronger guarantees.

Another approach to introducing privacy is based on perturbing the data by introducing noise, with the goal of producing privacy-preserving statistics [86–90]. *Homomorphic encryption*, on the other hand, can be used to perform computation directly on the encrypted data, thus eliminating the need of exposing any sensitive information [91–94]; this technique has been applied, for example, to vehicle positioning data [95] and medical records [96].

The flows of data—creation, copying, sharing—can be restricted. *Information Flow Control* solutions such as [97–99] attempt to regulate the flow of information in digital systems. *Auditing* implementations such as [100–102] track the data flow by generating usage logs. *Data Expiration* makes data inaccessible after a specific time, for example by self-destruction or by invalidating encryption keys [103–106]. *Watermarking* identifies records using hidden fingerprints, to allow traceability and identification of leaks [107–109].

Motivation

Here we describe our primary motivation for deploying the Copenhagen Networks Study, featuring deep and high-resolution data and a longitudinal approach.

Multiplexity

The majority of big data studies use datasets containing data from a single source, such as call detail records (CDRs) [4], RFID sensors [110], Bluetooth scanners [111], or online social networks activity [2]. Although, as we presented in the Related Work section, analyzing these datasets has led to some exciting findings, we may however not understand how much bias is introduced in such single-channel approaches, particularly in the case of highly interconnected data such as social networks.

We recognize two primary concerns related to the single-source approach: incomplete data and limitation with respect to an interdisciplinary approach. For social networks, we intuitively understand that people communicate on multiple channels: they call each other on the phone, meet face-to-face, or correspond through email. Observing only one channel may introduce bias that is difficult to estimate [11]. Ranjan et al. investigated in [112] how CDR datasets, containing samples dependent upon user activity and requiring user participation, may bias our under-

standing of human mobility. The authors used data activities as the ground truth; due to applications running in the background, sending and requesting data, smartphones exchange data with the network much more often than typical users make calls and without the need for their participation. Comparing the number of locations and significant locations [113], they found that the CDRs reveal only a small fraction of users' mobility, when compared with data activity. The identified home and work locations, which are considered the most important locations, did not, however, differ significantly when estimated using either of the three channels (voice, SMS, and data).

Domains of science operate primarily on different types of data. Across the sciences, researchers are interested in distinct questions and use very different methods. Similarly, as datasets are obtained from different populations and in different situations, it is difficult to cross-validate or combine findings. Moreover, the single-channel origin of the data can be a preventive factor in applying expertise from multiple domains. If we collect data from multiple channels in the same studies, on the same population, we can work together across field boundaries and draw on the different expertise and results generated by the studies and thereby achieve more robust insights.

Social networks are 'multiplex' in the sense that many different types of links may connect any pair of nodes. While recent work [114,115] has begun to explore the topic, a coherent theory describing multiplex, weighted, and directed networks remains beyond the frontier of our current understanding.

Sampling

In many big data studies, data sampling is uneven. CDRs, for example, only provide data when users actively engage, by making or receiving a phone call or SMS. Users can also have different patterns of engagement with social networks, some checking and interacting several times a day, while others only do so once a week [116]. Further, CDRs are typically provided by a single provider who has a finite market share. If the market share is 20% of the population and you consider only links internal to your dataset, this translates to only 4% of the total number of links, assuming random network and random sampling [4]. Thus, while CDRs might be sufficient when analysing of mobility, it is not clear that CDRs are a useful basis for social network analysis. Such uneven, sparse sampling decreases the resolution of data available for analysis. Ensuring the highest possible quality of the data, and even sampling, is possible with primarily passive data gathering, focusing on digital traces left by participants as they go through their lives, for example by using phones to automatically measure Bluetooth proximity, record location, and visible WiFi networks [9,29,31]. In cases where we cannot observe participants passively or when something simply goes wrong with the data collection, we aim to use the redundancy in the channels: if the participant turns off Bluetooth for a period, we can still estimate the proximity of participants using WiFi scans (as described in the Results section).

Uneven sampling not only reduces the quality of available data, but also—maybe more importantly—may lead to selection bias when choosing participants to include in the analysis. As investigated in [112], when only high-frequency voice-callers are chosen from a CDR dataset for the purpose of analysis, this can incur biases in Shannon entropy values (measure of uncertainty) of mobility, causing overestimation of the randomness of participants' behavior. Similarly, as shown in [116], choosing users with a large network and many interactions on Facebook may lead to overestimation of diversity in the ego-networks. Every time we have to discard a significant number of participants, we risk introducing bias in the data. Highly uneven sampling that cannot

be corrected with redundant data, compels the researcher to make mostly arbitrary choices as part of the analysis, complicating subsequent analysis, especially when no well-established ground truth is available to understand the bias. Our goal here is to collect evenly sampled high-quality data for all the participants, so we do not have to discard anyone; an impossible goal, but one worth pursuing.

Since we only record data from a finite number of participants, our study population is also a subset, and every network we analyze will be sampled in some way, see [117] for a review on sampling. While the 2013 deployment produces a dataset that is nearly complete in terms of communication between the participants, it is clear that it is subject to other sampling-related issues. For example, a relatively small network embedded in a larger society has a large 'surface' of links pointing to the outside world, creating a *boundary specification problem* [118].

Dynamics

The networks and behaviors we observe are not static; rather they display dynamics on multiple time-scales. Long-term dynamics may be lost in big data studies when the participants are not followed for a sufficiently long period, and only a relatively narrow slice of data is acquired. Short-term dynamics may be missed when the sampling frequency is too low.

It is a well-established fact that social networks evolve over time [8,119]. The time scale of the changes varies and depends on many factors, for example the semester cycle in students' life, changing schools or work, or simply getting older. Without following such dynamics, and if we focus on a single temporal slice, we risk missing an important aspect of human nature. To capture it, we need long-term studies, that follow participants for months or even years.

Our behavior is not static, even when measured for very short intervals. We have daily routines, meeting with different people in the morning and hanging out with other people in the evening, see Figure 1. Our workdays may see us going to places and interacting with people differently than on weekends. It is easy to miss dynamics like these when the quality of the data is insufficient, either because it has not been sampled frequently enough or because of poor resolution, requiring large time bins.

Because each node has a limited bandwidth, only a small fraction of the network is actually 'on' at any given time, even if the underlying social network is very dense. Thus, to get from node A to node B, a piece of information may only travel on links that are active at subsequent times. Some progress has been made on the understanding of dynamic networks, for a recent review see [120]. However, in order to understand the dynamics of our highly dense, multiplex network, we need to expand and adapt the current methodologies, for example by adapting the link-based viewpoint to dynamical systems.

Feedback

In many studies, the data collection phase is separated from the analysis. The data might have been collected during usual operation, before the idea of the study had even been conceived (e.g. CDRs, WiFi logs), or access to the data might have not been granted before a single frozen and de-identified dataset was produced.

One real strength of the research proposed here is that, in addition to the richness of the collected data, we are able to run controlled experiments, including surveys distributed via the smartphone software. We can, for example, divide participants into sub-populations and expose them to distinct stimuli, addressing the topic of causality as well as confounding factors both of which have proven problematic [64,121] for the current state-of-the-art [122,123].

Moreover, we monitor the data quality not only on the most basic level of a participant (number of data points) but also by looking at the entire live dataset to understand if the quality of the collected data is sufficient to answer our research questions. This allows us to see and fix bugs in the data collection software, or learn that certain behaviors of the participants may introduce bias in the data: for example after discovering missing data, some interviewed students reported turning their phones off for the night to preserve battery. This allowed us to understand that, even if in terms of the raw numbers, we may be missing some hours of data per day for these specific participants, there was very little information in that particular data anyway.

Building systems with real-time data processing and access allows us to provide the participants with applications and services. It is an important part of the study not only to collect and analyze the data but also to learn how to create a feedback loop, directly feeding back extracted knowledge on behavior and interactions to the participants. We are interested in studying how personal data can be used to provide feedback about individual behavior and promote self-awareness and positive behavior change, which is an active area of research in Personal Informatics [124]. Applications for participants create value, which may be sufficient to allow us to deploy studies without buying a large number of smartphones to provide to participants. Our initial approach has included the development and deployment of a mobile app that provides feedback about personal mobility and social interactions based on personal participant data [125]. Preliminary results from the deployment of the app, participant surveys, and usage logs suggest an interest in such applications, with a subset of participants repeatedly using the mobile app for personal feedback [126]. It is
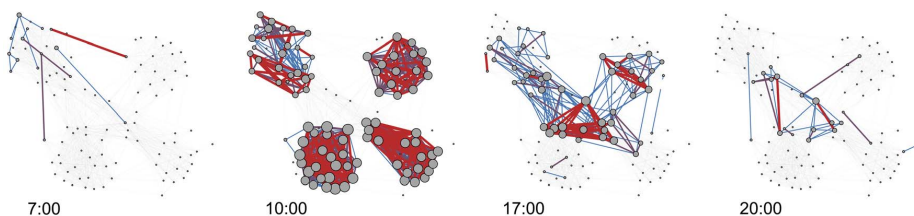

Figure 1. Dynamics of face-to-face interactions in the 2012 deployment. The participants meet in the morning, attend classes within four different study lines, and interact across majors in the evening. Edges are colored according to the frequency of observation, ranging from low (blue) to high (red). With 24 possible observations per hour, the color thresholds are respectively: blue ($0 < $ observations ≤ 6), purple ($6 < $ observations ≤ 12), and red (< 12 observations). Node size is linearly scaled according to degree.
doi:10.1371/journal.pone.0095978.g001

clear that feedback can potentially influence the study results: awareness of a certain behavior may cause participants to want to change that behavior. We believe, however, that such feedback is unavoidable in any study, and studying the effects of such feedback (in order to account for it) is an active part of our research.

New Science

The ability to record the highly dynamic networks opens up a new, microscopic level of observation for the study of diffusion on the network. We are now able to study diffusion of behavior, such as expressions of happiness, academic performance, alcohol and other substance abuse, information, as well as real world infectious disease (e.g. influenza). Some of these vectors may spread on some types of links, but not others. For example, influenza depends on physical proximity for its spread, while information may diffuse on all types of links; with the deep data approach we can study differences and similarities between various types of spreading and the interplay between the various communication channels [127,128].

A crucial step when studying the structure and dynamics of networks is to identify communities (densely connected groups of nodes) [129,130]. In social networks, communities roughly correspond to social spheres. Recently, we pointed out that communities in many real world networks display *pervasive overlap*, where each and every node belongs to more than one group [131]. It is important to underscore that the question of whether or not communities in networks exhibit pervasive overlap has great practical importance. For example, the patterns of epidemic spreading change, and the optimal corresponding societal countermeasures are very different, depending on the details of the network structure.

Although algorithms that detect disjoint communities have operated successfully since the notion of graph partitioning was introduced in the 1970s [132], we point out that most networks investigated so far are highly incomplete in multiple senses. Moreover, we can use a simple model to show that sampling could cause pervasively overlapping communities to appear to be disjoint [133]. The results reveal a fundamental problem related to working with incomplete data: *Without an accurate model of the structural ordering of the full network, we cannot estimate the implications of working with incomplete data.* Needless to say, this fact is of particular importance to studies carried out on (thin) slices of data, describing only a single communication channel, or a fraction of nodes using that channel. By creating a high-quality, high-resolution data set, we are able to form accurate descriptions of the full data set needed to inform a proper theory for incomplete data. A deeper understanding of sampling is instrumental for unleashing the full potential of data from the billions of mobile phones in use today.

Methods: Data Collection

The Copenhagen Networks Study aims to address the problem of single-modality data by collecting information from a number of sources that can be used to build networks, study social phenomena, and provide context necessary to interpret the findings. A series of questionnaires provides information on the socioeconomic background, psychological traces, and well-being of the participants; Facebook data enables us to learn about the presence and activity of subjects in the biggest online social networking platform [134]; finally, the smartphones carried by all participants record their location, telecommunication patterns, and face-to-face interactions. Sensor data is collected with fixed intervals, regardless of the users' activity, and thus the uneven sampling issue, daunting especially CDR-based studies, is mainly

overcome. Finally, the study is performed on the largest and the most dense population to date in this type of studies. The physical density of the participants helps to address the problem of missing data, but raises new questions regarding privacy, since missing data about a person can, in many cases, be inferred from existing data of other participants. For example, if we know that person A, B, and C met at a certain location based on the data from person A, we do not need social and location data from B and C to know where and with whom they were spending time.

Below we describe the technical challenges and solutions in multi-channel data collection in 2012 and 2013 deployments. Data collection, anonymization, and storage were approved by the Danish Data Protection Agency, and comply with both local and EU regulations.

Data Sources

The data collected in the two studies were obtained from questionnaires, Facebook, mobile sensing, an anthropological field study, and the WiFi system on campus.

Questionnaires. In 2012 we deployed a survey containing 95 questions, covering socioeconomic factors, participants' working habits, and the Big Five Inventory (BFI) measuring personality traits [135]. The questions were presented as a Google Form and participation in the survey was optional.

In 2013 we posed 310 questions to each participant. These questions were prepared by a group of collaborating public health researchers, psychologists, anthropologists, and economists from the Social Fabric project (see Acknowledgements). The questions in the 2013 deployment included BFI, Rosenberg Self Esteem Scale [136], Narcissism NAR-Q [137], Satisfaction With Life Scale [138], Rotters Locus of Control Scale [139], UCLA Loneliness scale [140], Self-efficacy [141], Cohens perceived stress scale [142], Major Depression Inventory [143], The Copenhagen Social Relation Questionnaire [144], and Panas [145], as well as number of general health- and behavior-related questions. The questions were presented using a custom-built web application, which allowed for full customization and complete control over privacy and handling of the respondents' data. The questionnaire application is capable of presenting different types of questions, with branching depending on the answers given by the participant, and saving each participant's progress. The application is available as an open source project at github.com/MIT-Model-Open-Data-and-Identity-System/SensibleDTUData-Apps-Questionaires. Participation in the survey was required for taking part in the experiment. In order to track and analyze temporal development, the survey (in a slightly modified form) was repeated every semester on all participating students.

Facebook Data. For all participants in both the 2012 and 2013 deployment, it was optional to authorize data collection from Facebook, and a large majority opted in. In the 2012 deployment, only the friendship graph was collected every 24 hours, until the original tokens expired. In the 2013 deployment, data from Facebook was collected as a snapshot, every 24 hours. The accessed scopes were birthday, education, feed, friend lists, friend requests, friends, groups, hometown, interests, likes, location, political views, religion, statuses, and work. We used long-lived Facebook access tokens, valid for 60 days, and when the tokens expired, participants received notification on their phones, prompting them to renew the authorizations. For the academic study purposes, the Facebook data provided rich demographics describing the participants, their structural (friendship graph) and functional (interactions) networks, as well as location updates.

Sensor Data. For the data collection from mobile phones, we used a modified version of the Funf framework [31] in both

deployments. The data collection app was built using the framework runs on Android smartphones, which were handed out to participants (Samsung Galaxy Nexus in 2012 and LG Nexus 4 in 2013). All the bugfixes and the improvement of the framework are public and available under the OpenSensing github organization at github.com/organizations/OpenSensing.

In the 2012 deployment, we manually kept track of which phone was used by each student, and identified data using device IMEI numbers, but this created problems when the phones were returned and then handed out to other participants. Thus, in the 2013 deployment, the phones were registered in the system by the students in an OAuth2 authorization flow initiated from the phone; the data were identified by a token stored on the phone and embedded in the data files. The sensed data were saved as locally encrypted sqlite3 databases and then uploaded to the server every 2 hours, provided the phone was connected to WiFi. Each file contained 1 hour of participant data from all probes, saved as a single table. When uploaded, the data was decrypted, extracted, and included in the main study database.

Qualitative Data. An anthropological field study was included in the 2013 deployment. An anthropologist from the Social Fabric project was embedded within a randomly selected group of approximately 60 students (August 2013–august 2014). A field study consists of participant observation within the selected group, collecting qualitative data while simultaneously engaging in the group activities. The goal is to collect data on various rationales underlying different group formations, while at the same time experiencing bodily and emotionally what it was like to be part of these formations [146]. The participant observation included all the student activities and courses, including extracurricular activities such as group work, parties, trips, and other social leisure activities. All participants were informed and periodically reminded about the role of the anthropologist.

In addition to its central purpose, the anthropological data adds to the multitude of different data channels, deepening the total pool of data. This proved useful for running and optimizing the project in a number of ways.

Firstly, data from qualitative social analysis are useful—in a very practical sense—in terms of acquiring feedback from the participants. One of the goals of the project is to provide value to the participants; in addition to providing quantified-self style access to data, we have also created a number of public services: a homepage, a Facebook page, and a blog, where news and information about the project can be posted and commented on. These services are intended to keep the students interested, as well as to make participants aware of the types and amounts of data collected (see Privacy section). Because of the anthropologist's real-world engagement with the students, the qualitative feedback contains complex information about participants' interests and opinions, including what annoyed, humored, or bored them. This input has been used to improve existing services, such as visualizations (content and visual expression), and to develop ideas for the future services. In summary, qualitative insights helped us understand the participants better and, in turn, to maintain and increase participation.

Secondly, the inclusion of qualitative data increases the potential for interdisciplinary work between the fields of computer science and social science. Our central goal is to capture the full richness of social interactions by increasing the number of recorded communication channels. Adding a qualitative social network approach makes it possible to relate the qualitative observations to the quantitative data obtained from the mobile sensing, creating an interdisciplinary space for methods and theory. We are particularly interested in the relationship between the observations made by the embedded anthropologist and the data recorded using questionnaires and mobile sensing, to answer questions about the elements difficult to capture using our high-resolution approach. Similarly, from the perspective of social sciences, we are able to consider what may be captured by incorporating quantitative data from mobile sensing into a qualitative data pool—and what can we learn about social networks using modern sensing technology.

Finally, these qualitative data can be used to ground the mathematical modeling process. Certain things are difficult or impossible to infer from quantitative measurements and mathematical models of social networks, particularly in regard to understanding *why* things happen in the network, as computational models tend to focus on *how*. Questions about relationship-links severing, tight networks dissolving, and who or what caused the break, can be very difficult to answer, but they are important in regard to understanding the dynamics of the social network. By including data concerned with answering *why* in social networks, we add a new level of understanding to the quantitative data.

WiFi Data. For the 2012 deployment, between August 2012 and May 2013, we were granted access to the campus WiFi system logs. Every 10 minutes the system provided metadata about all devices connected to the wireless access points on campus (access point MAC address and building location), together with the student ID used for authentication. We collected the data in a de-identified form, removing the student IDs and matching the participants with students in our study. Campus WiFi data was not collected for the 2013 deployment.

Backend System

The backend system, used for data collection, storage, and access, was developed separately for the 2012 and 2013 deployments. The system developed in 2012 was not designed for extensibility, as it focused mostly on testing various solutions and approaches to massive sensor-driven data collection. Building on this experience, the system for the 2013 deployment was designed and implemented as an extensible framework for data collection, sharing, and analysis.

The 2012 Deployment. The system for the 2012 deployment was built as a Django web application. The data from the participants from the multiple sources, were stored in a CouchDB database. The informed consent was obtained by presenting a document to the participants after they authenticated with university credentials. The mobile sensing data was stored in multiple databases inside a single CouchDB instance and made available via an API. Participants could access their own data, using their university credentials. Although sufficient for the data collection and research access, the system performance was not adequate for exposing the data for real-time application access, mainly due to the inefficient de-identification scheme and insufficient database structure optimization.

The 2013 Deployment. The 2013 system was built as an open Personal Data System (openPDS) [147] in an extensible fashion. The architecture of the system is depicted in Figure 2 and consisted of three layers: platform, services, and applications. In the platform layer, the components common for multiple services were grouped, involving identity provider and participant-facing portal for granting authorizations. The identity provider was based on OpenID 2.0 standard and enabled single sign-on (SSO) for multiple applications. The authorizations were realized using OAuth2 and could be used with both web and mobile applications. Participants enroll into studies by giving informed consent and subsequently authorizing application to submit and access data from the study. The data storage was implemented
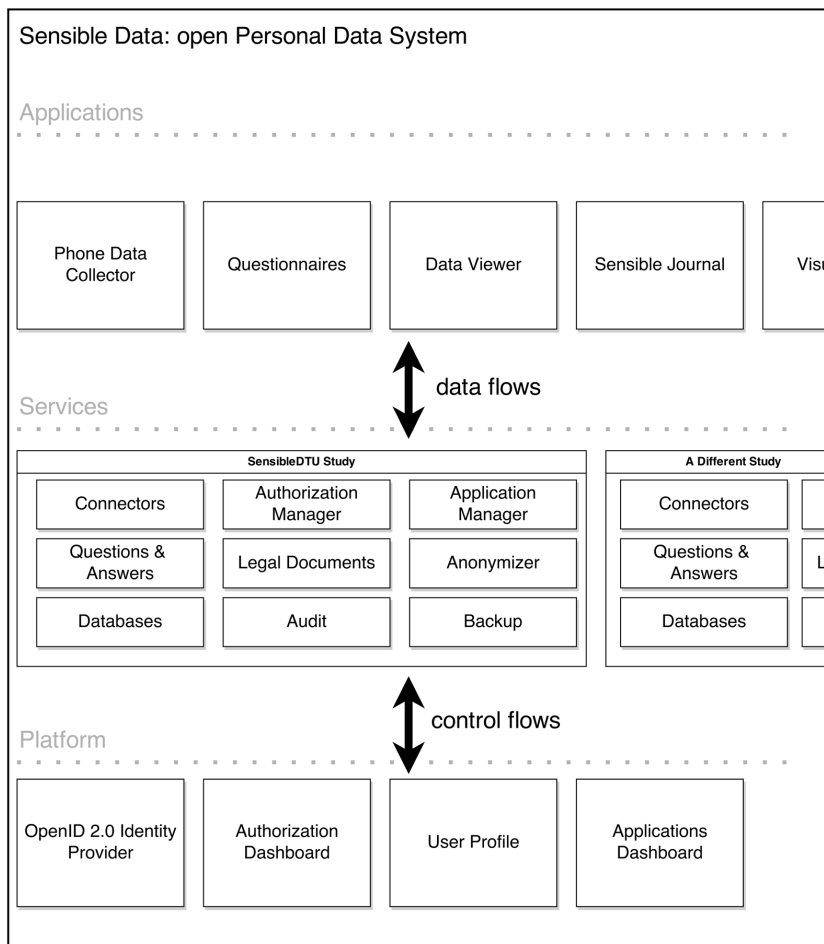
Figure 2. Sensible Data openPDS architecture. This system is used in the 2013 deployment and consists of three layers: platform, services, and applications. The platform contains element common for multiple services (in this context: studies). The studies are the deployments of particular data collection efforts. The applications are OAuth2 clients to studies and can submit and access data, based on user authorizations.
doi:10.1371/journal.pone.0095978.g002

using MongoDB. Participants can see the status and change their authorizations on the portal site, the system included an implementation of the Living Informed Consent [3].

Deployment Methods

Organizing studies of this size is a major undertaking. All parts from planning to execution have to be synchronized, and below we share some considerations and our approaches. While their main purpose was identical, the two deployments differed greatly in size and therefore also in the methods applied for enrolling and engaging the participants.

SensibleDTU 2012. In 2012 approximately 1,400 new students were admitted to the university, divided between two main branches of undergraduate programs. We focused our efforts on the larger branch containing 900 students, subdivided into 15 study lines (majors). For this deployment we had ~ 200 phones

available to distribute between the students. To achieve maximal coverage and density of the social connections, we decided to only hand out phones in a few selected majors that had a sufficient number of students interested in participating in the experiment. Directly asking students about their interest in the study was not a good approach, as it could lead to biased estimates and would not scale well for a large number of individuals. Instead, we appealed to the competitive element of human nature by staging a competition, running for two weeks from the start of the semester. All students had access to a web forum, which was kept separate for each major, where they could post ideas that could be realized by the data we would collect, and subsequently vote for their own ideas or three seed ideas that we provided. The goal of the competition was twofold; first we wanted students to register with their Facebook account, thereby enabling us to study their online social network, and second we wanted to see which major could

gain most support (percentage of active students) behind a single idea. Students were informed about the project and competition by the Dean in person and at one of 15 talks given—one at each major. Students were told that our choice of participants would be based on the support each major could muster behind their strongest idea before a given deadline. This resulted in 24 new research ideas and 1 026 unique votes. Four majors gained >93% support for at least one idea and were chosen to participate in the experiment.

The physical handing out of the phones was split into four major sessions, in which students from the chosen majors were invited; additional small sessions were arranged for students that were unable to attend the main ones. At each session, participants were introduced to our data collection methods, de-identification schemes, and were presented with the informed consent form. In addition, the participants were instructed to fill out the questionnaire. A small symbolic deposit in cash was requested from each student; this served partially as compensation for broken phones, but was mainly intended to encourage participants take better care of the phones, than if they had received them for free [148]. Upon receiving a phone, participants were instructed to install the data collector application. The configuration on each phone was manually checked when participants were leaving—this was particularly important to ensure high quality of data.

This approach had certain drawbacks; coding and setting up the web fora, manually visiting all majors and introducing them to the project and competition, and organizing the handout sessions required considerable effort and time. However, certain aspects were facilitated with strong support from the central administration of the university. A strong disadvantage of the outlined handout process is that phones were handed out 3–4 weeks into the semester, thus missing the very first interactions between students.

SensibleDTU 2013. The 2013 deployment was one order of magnitude larger, with 1 000 phones to distribute. Furthermore, our focus shifted to engaging the students as early as possible. Pamphlets informing prospective undergraduate students about the project were sent out along with the official acceptance letters from the university. Early-birds who registered online via Facebook using the links given in the pamphlet were promised phones before the start of their studies. Students from both branches of undergraduate programs were invited to participate (approximately 1 500 individuals in total), as we expected an adoption percentage between 30% and 60%. Around 300 phones were handed out to early-birds, and an additional 200 were handed out during the first weeks of semester. As the adoption rate plateaued, we invited undergraduate students from older years to participate in the project.

The structure of the physical handout was also modified, the participants were requested to enroll online before receiving the phone. Moreover, the informed consent and the questionnaire were part of the registration. Again, we required a symbolic cash deposit for each phone. We pre-installed custom software on each phone to streamline the handout process; students still had to finalize set up of the phones (make them Bluetooth-discoverable, activate WiFi connection, etc.).

For researchers considering similar projects with large scale handouts, we recommend that the pool of subjects are engaged in the projects as early as possible and be sure to keep their interest. Make it easy for participants to contact you, preferably through media platforms aimed at their specific age group. Establish clear procedures in case of malfunctions. On a side note, if collecting even a small deposit, when multiplied by a factor of 1 000, the total

can add up to significant amount, which must be handled properly.

Methods: Privacy

When collecting data of very high resolution, over an extended period, from a large population, it is crucial to address the privacy of the participants appropriately. We measure the privacy as a difference between what a participant understands and consents to regarding her data, and what in fact happens to these data.

We believe that ensuring sufficient privacy for the participants, in large part, is the task of providing them with tools to align the data usage with their understanding. Such privacy tools must be of two kinds: to inform, ensuring participants understand the situation, and to control, aligning the situation with the participant's preferences. There is a tight loop where these tools interact: as the participant grows more informed, she may decide to change the settings, and then verify if the change had the expected result. By exercising the right to information and control, the participant expresses Living Informed Consent as described in [3].

Not all students are interested in privacy, in fact we experienced quite the opposite attitude. During our current deployments the questions regarding privacy were rarely asked by the participants, as they tended to accept any terms presented to them without thorough analysis. It is our—the researchers'—responsibility to make the participants more aware and empowered to make the right decisions regarding their privacy: by providing the tools, promoting their usage, and engaging in a dialog about privacy-related issues.

In the 2012 deployment, we used a basic informed consent procedure with an online form accepted by the participants, after they authenticated with the university account system. The accepted form was then stored in a database, together with the username, timestamp, and the full text displayed to the participant. The form itself was a text in Danish, describing the study purpose, parties responsible, and participants' rights and obligations. The full text is available at [149] with English translation available at [150].

In the 2013 deployment, we used our backend solution (described in Backend System Section) to address the informed consent procedure and privacy in general. The account system, realized as an OpenID 2.0 server, allowed us to enroll participants, while also supporting research and developer accounts (with different levels of data access). The sensitive Personally Identifiable Information attributes (PIIs) of the participants were kept completely separate from the participant data, all the applications identified participants based only on the pseudonym identifiers. The applications could also access a controlled set of identity attributes for the purpose of personalization (e.g. greeting the participant by name), subject to user OAuth2 authorization. In the enrollment into the study, after the participant had accepted the informed consent document—essentially identical to that from 2012 deployment—a token for a scope *enroll* was created and shared between the platform and service (see Figure 2). The acceptance of the document was recorded in the database by storing the username, timestamp, hash of the text presented to the participant, as well as the git commit identifying the version of the form.

All the communication in the system was realized over HTTPS, and endpoints were protected with short-lived OAuth2 bearer tokens. The text of the documents, including informed consent, was stored in a git repository, allowing us to modify everything, while still maintaining the history and being able to reference

Figure 3. Authorizations page. Participants have an overview of the studies in which they are enrolled and which applications are able to submit to and access their data. This is an important step towards users' understanding what happens with their data and to exercising control over it. This figure shows a translated version of the original page that participants saw in Danish.
doi:10.1371/journal.pone.0095978.g003

which version each participant has seen and accepted. A single page overview of the status of the authorizations, presented in Figure 3, is an important step in moving beyond lengthy, incomprehensible legal documents accepted by the users blindly and giving more control over permissions to the participant.

In the 2013 deployment, the participants could access all their data using the same API as the one provided for the researchers and application developers. To simplify the navigation, we developed a data viewer application as depicted in Figure 4, which supports building queries with all the basic parameters in a more user-friendly way than constructing API URLs. Simply having access to all the raw data is, however, not sufficient, as it is really high-level inferences drawn from the data that are important to understand, for example *Is someone accessing my data to see how fast I drive or to study population mobility*? For this purpose, we promoted the development of a *question & answer* framework, where the high-

level features are extracted from the data before leaving the server, promoting better participant understanding of data flows. This is aligned with the vision of the open Personal Data Store [147].

Finally, for the purposes of engaging the participants in the discussion about privacy, we published blogposts (e.g. https://www.sensible.dtu.dk/?p = 1622), presented relevant material to students, and answered their questions via the Facebook page(https://www.facebook.com/SensibleDtu).

Results and Discussion

As described in the previous sections, our study has collected comprehensive data about a number of aspects regarding human behavior. Below, we discuss primary data channels and report some early results and findings. The results are mainly based on the 2012 deployment due to the availability of longitudinal data.
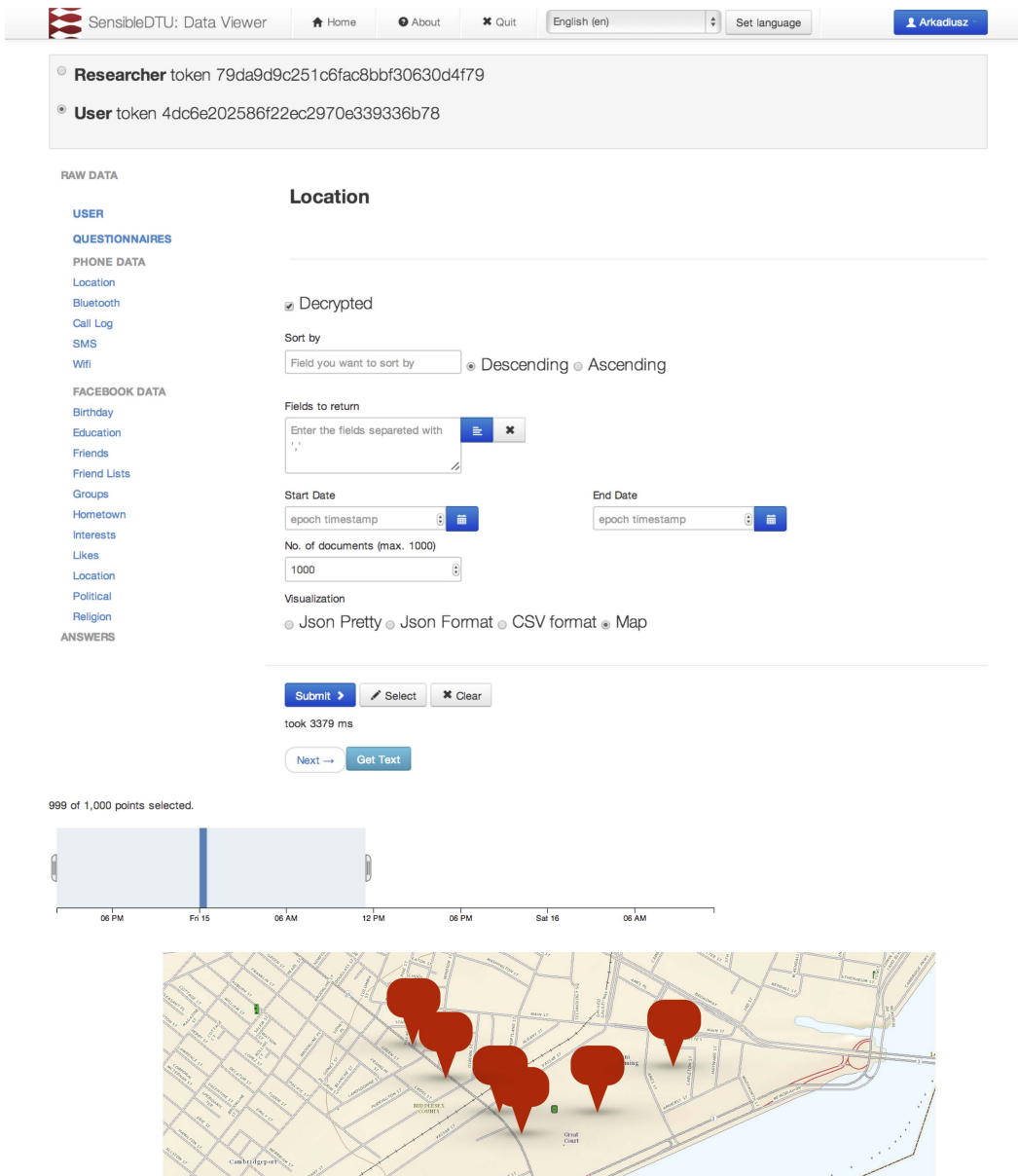
Figure 4. Data viewer application. All the collected data can be explored and accessed via an API. The API is the same for research, application, and end-user access, the endpoints are protected by OAuth2 bearer token. Map image from USGS National Map Viewer, replacing original image used in the deployed application (Google Maps).
doi:10.1371/journal.pone.0095978.g004

Bluetooth and Social Ties

Bluetooth is a wireless technology ubiquitous in modern-day mobile devices. It is used for short-range communication between devices, including smartphones, hands-free headsets, tablets, and other wearables. As the transmitters used in mobile devices are primarily of very short range—between 5 and 10 m ($16-33$ feet)—detection of the devices of other participants (set in 'visible' mode) can be used as a proxy for face-to-face interactions [29]. We take the individual Bluetooth scans in the form (i,j,t,σ), denoting that device i has observed device j at time t with signal strength σ.
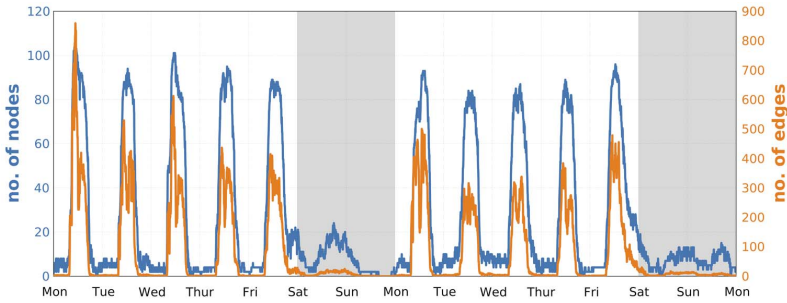
Figure 5. Weekly temporal dynamics of interactions. Face-to-face interaction patterns of participants in 5-minute time-bins over two weeks. Only active participants are included, i.e. those that have either observed another person or themselves been observed in a given time-bin. On average we observed 29 edges and 12 nodes in 5-minute time-bins and registered 10 634 unique links between participants.
doi:10.1371/journal.pone.0095978.g005
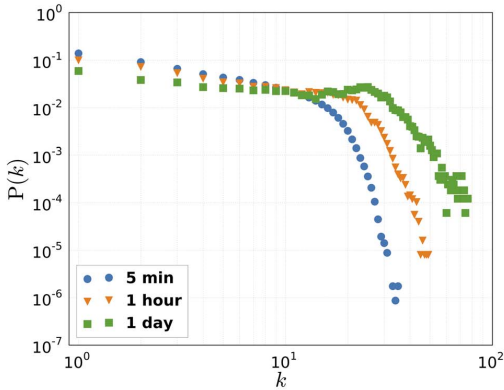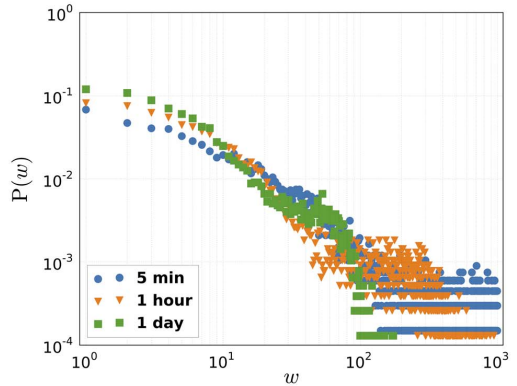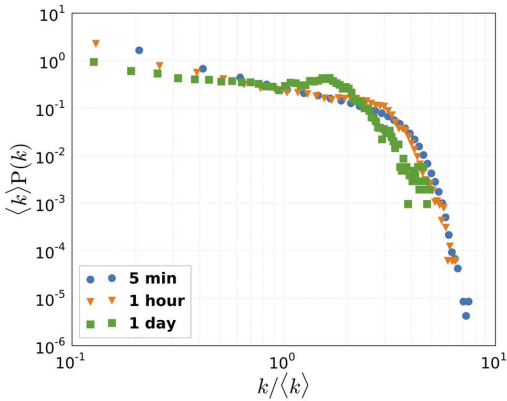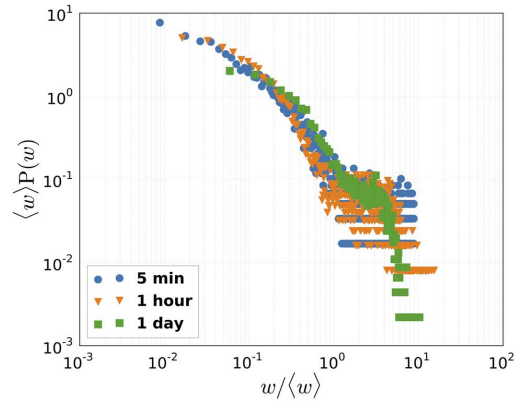
Figure 6. Face-to-face network properties at different resolution levels. Distributions are calculated by aggregating sub-distributions across temporal window. Differences in rescaled distributions suggest that social dynamics unfold on multiple timescales.
doi:10.1371/journal.pone.0095978.g006

Figure 7. WiFi similarity measures. Positive predictive value (precision, ratio of number of true positives to number of positive calls, marked with dashed lines) and recall (sensitivity, fraction of retrieved positives, marked with solid lines) as functions of parameters in different similarity measures. **A**) In 98% of face-to-face meetings derived from Bluetooth, the two devices also sensed at least one common access point. **D**) Identical strongest access point for two separate mobile devices is a strong indication of a face-to-face meeting.
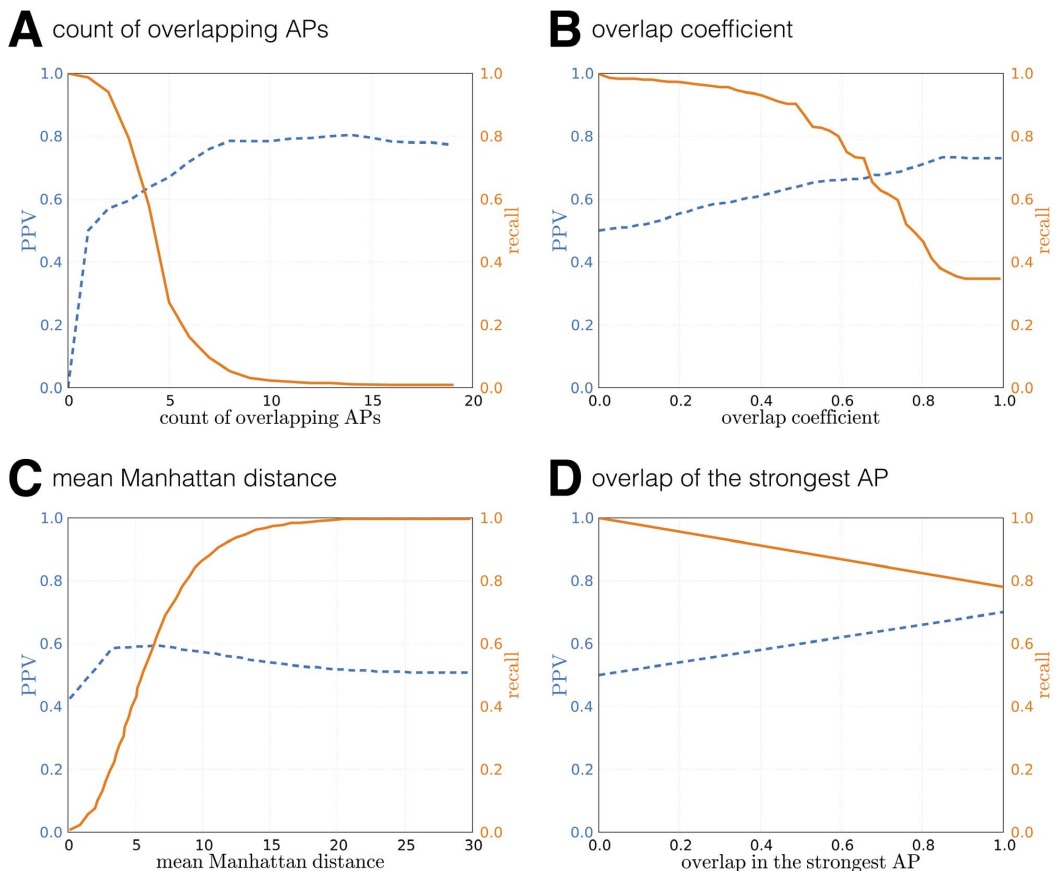doi:10.1371/journal.pone.0095978.g007

Bluetooth scans do not constitute a perfect proxy for face-to-face interactions [151], since a) it is possible for people within 10 m radius not to interact socially, and b) it is possible to interact socially over a distance greater than 10 m, nevertheless, they have been successfully used for sensing social networks [31] or crowd tracking [152].
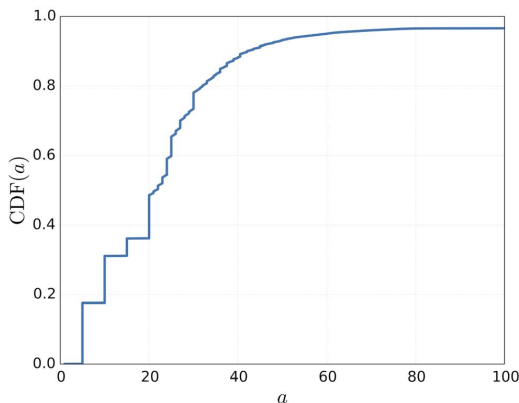
Between October 1st, 2012 and September 1st, 2013, we collected 12 623 599 Bluetooth observations in which we observed 153 208 unique devices. The scans on the participants' phones were triggered every five minutes, measured from the last time the phone was powered on. Thus, the phones scanned for Bluetooth in a desynchronized fashion, and not according to a global schedule. To account for this, when extracting interactions from the raw Bluetooth scans, we bin them into fixed-length time windows, aggregating the scans within them. The resulting adjacency matrix, $W_{\Delta t}$ does not have to be strictly symmetric, meaning that participant i can observe participant j in time-bin t, but not the other way around. Here we assume that Bluetooth scans do not produce false positives (devices are not discovered unless they are

really there), and in the subsequent network analysis, we force the matrix to be symmetric, assuming that if participant i observed participant j, the opposite is also true.

The interactions between the participants exhibit both daily and weekly rhythms. Figure 1 shows that the topology of the network of face-to-face meetings changes significantly within single day, revealing academic and social patterns formed by the students. Similarly, the intensity of the interactions varies during the week, see Figure 5.

Aggregating over large time-windows blurs the social interactions (network is close to fully connected) while a narrow window reveals detailed temporal structures in the network. Figure 6A shows the aggregated degree distributions for varying temporal resolutions, with $P(k)$ being shifted towards higher degrees for larger window sizes; this is an expected behavior pattern since each node has more time to amass connections. Figure 6B presents the opposite effect, where the edge weight distributions $P(w)$ shift towards lower weights for larger windows; this is a consequence on definition of a link for longer time-scales or, conversely, of links

A estimated cumulative distribution function for the accuracy of samples *a* (in meters)

B Gaussian kernel density estimation for the radius of gyration r_g (in kilometers)

C histogram of the locations, with hex binning and logarithmic color scale
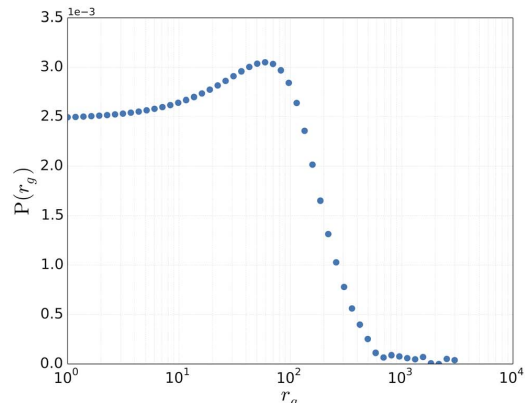
D cumulated transitions between stop locations
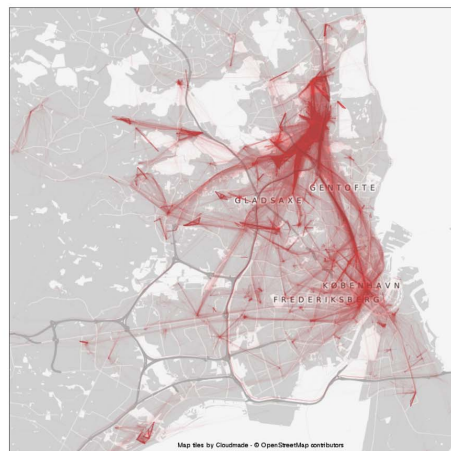
Figure 8. Location and Mobility. We show the accuracy of the collected samples, radius of gyration of the participants, and identify patterns of collective mobility.
doi:10.1371/journal.pone.0095978.g008

appearing in each window on shorter timescales. To compare the distribution between timescales, we rescale the properties according to Krings et al. [153] as $Q(x) = \langle x \rangle P(x/\langle x \rangle)$ with $\langle x \rangle = \sum x P(x)$ (Figure 6C and 6D). The divergence of the rescaled distributions suggest a difference in underlying social dynamics between long and short timescales, an observation supported by recent work on temporal networks [44,153,154].

WiFi as an Additional Channel for Social Ties

Over the last two decades, wireless technology has transformed our society to the degree where every city in the developed world is now fully covered by mobile [155] and wireless networks [156]. The data collector application for mobile phones was configured to scan for wireless networks in constant intervals, but also to record the results of scans triggered by any other application

running on the phone ('opportunistic' sensing). Out of the box, Android OS scans for WiFi every 15 seconds, and since we collected these data, our database contains 42 692 072 WiFi observations, with 142 871 unique networks (SSIDs) between October 1ˢᵗ, 2012 and September 1ˢᵗ, 2013 (i.e. the 2012 deployment). Below we present the preliminary result on WiFi as an additional data-stream for social ties, to provide an example of how our multiple layers of information can complement and enrich each other.

For computational social science, using Bluetooth-based detection of participants' devices as a proxy for face-to-face interactions is a well-established method [19,29,31]. The usage of WiFi as a social proxy has been investigated [157], but, to our knowledge, has not yet been used in a large-scale longitudinal study. For the method we describe here, the participants' devices do not sense

each other, instead they record the visible beacons (in this instance WiFi access points) in their environment. Then, physical proximity between two devices—or lack thereof—can be inferred by comparing results of the WiFi scans that occurred within a sufficiently small time window. Proximity is assumed if the lists of access points (APs) visible to both devices are similar according to a similarity measure. We establish the appropriate definition of the similarity measure in a data-driven manner, based on best fit to Bluetooth data. The strategy is to compare the lists of results in 10-minute-long time bins, which corresponds to the forced sampling period of the WiFi probe as well as to our analysis of Bluetooth data. If there are multiple scans within the 10-minute bin, the results are compared pair-wise, and proximity is assumed if at least one of these comparisons is positive. The possibility of extracting face-to-face interactions from such signals is interesting, due to the ubiquitous nature of WiFi and high temporal resolution of the signal.

We consider four measures and present their performance in Figure 7. Figure 7A shows the positive predictive value and recall as a function of minimum number of overlapping access points ($|X \cap Y|$) required to assume physical proximity. In approximately 98% of all Bluetooth encounters, at least one access point was seen by both devices. However, the recall drops quickly with the increase of their required number. This measure favors interactions in places with a high number of access points, where it is more likely that devices will have a large scan overlap. The result confirms that lack of a common AP has a very high positive predictive power as a proxy for *lack* of physical proximity, as postulated in [158]. Note, that for the remaining measures, we assume at last one overlapping AP in the compared lists of scan results.

The overlap coefficient defined as $\mathrm{overlap}(X, Y) = \dfrac{|X \cap Y|}{\min(|X|, |Y|)}$ penalizes encounters taking place in WiFi-dense areas, due to higher probability of one device picking up a signal from a remote access point that is not available to the other device, see Figure 7B.
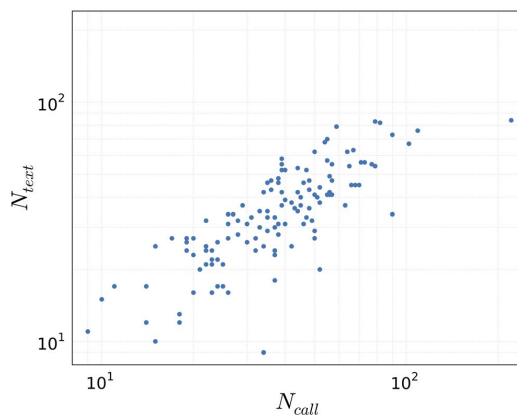

Figure 9. Diversity of communication logs. Diversity is estimated as the set of unique numbers that a person has contacted or been contacted by in the given time period on a given channel. We note a strong correlation in diversity (Pearson correlation of 0.75, $p \ll 0.05$), whereas the similarity of the sets of nodes is fairly low (on average $\langle \sigma \rangle = 0.37$).
doi:10.1371/journal.pone.0095978.g009

Next, we compare the received signal strengths between overlapping routers using the mean ℓ_1-norm (mean Manhattan distance, $\dfrac{||X \cap Y||_1}{|X \cap Y|}$). Received signal strength (RSSI) is measured in *dBm* and the Manhattan distance between two routers is the difference in the RSSI between them, measured in *dB*. Thus, the mean Manhattan distance is the mean difference in received signal strength of the overlapping routers in the two compared scans.

Finally, we investigate the similarity based on the router with the highest received signal strength—the proximity is assumed whenever it is the same access point for both devices, $\max(X) = \max(Y)$. This measure provides both high recall and positive predictive value and, after further investigation for the causes for errors, is a candidate proxy for face-to-face interactions.

The performance of face-to-face event detection based on WiFi can be further improved by applying machine-learning approaches [158,159]. It is yet to be established, by using longitudinal data, whether the errors in using single features are caused by inherent noise in measuring the environment, or if there is a bias that could be quantified and mitigated. Most importantly, the present analysis is a proof-of-concept and further investigation is required to verify if networks inferred from WiFi and Bluetooth signals are satisfyingly similar, before WiFi can be used as an autonomous channel for face-to-face event detection in the context of current and future studies. Being able to quantify the performance of multi-channel approximation of face-to-face interaction and to apply it in the data analysis is crucial to address the problem of missing data, as well as to estimate the feasibility and understand the limitations of single-channel studies.

Location and Mobility

A number of applications ranging from urban planning, to traffic management, to containment of biological diseases rely on the ability to accurately predict human mobility. Mining location data allows extraction of semantic information such as points of interest, trajectories, and modes of transportation [160]. In this section we report the preliminary results of an exploratory data analysis of location and mobility patterns.

Location data was obtained by periodically collecting the best position estimate from the location sensor on each phone, as well as recording location updates triggered by other applications running on the phone (opportunistic behavior). In total we collected 7 593 134 data points in 2012 deployment in the form (userid, timestamp, latitude, longitude, accuracy). The best-effort nature of the data presents new challenges when compared with the majority of location mining literature, which focuses on high-frequency, high-precision GPS data. Location samples on the smartphones can be generated by different providers, depending on the availability of the Android sensors, as explained in developer.android.com/guide/topics/location/strategies.html. For this reason, accuracy of the collected position can vary between a few meters for GPS locations, to hundreds of meters for cell tower location. Figure 8A shows the estimated cumulative distribution function for the accuracy of samples; almost 90% of the samples have a reported accuracy better than 40 meters.

We calculate the radius of gyration r_g as defined in [38] and approximate the probability distribution function using a gaussian kernel density estimation, see Figure 8B. We select the appropriate kernel bandwidth through leave-one-out cross-validation scheme from Statsmodels KDEMultivariate class [161]. The kernel density peaks around 10^2 km and then rapidly goes down, displaying a fat-tailed distribution. Manual inspection of the few participants with r_g around 10^3 km revealed that travels abroad can amount to
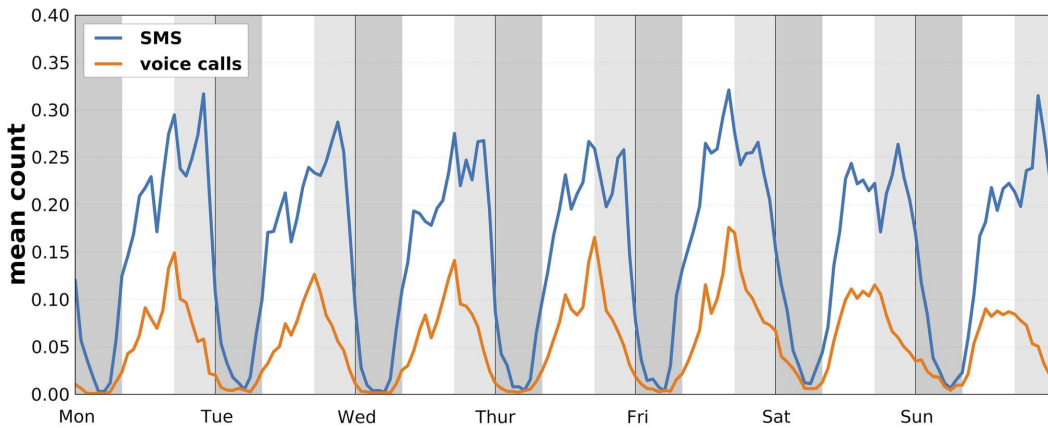
Figure 10. Weekly temporal dynamics of interactions. All calls and SMS, both incoming and outgoing, were calculated over the entire dataset and averaged per participant and per week, showing the mean number of interactions participants had in a given weekly bin. Light gray denotes 5pm, the time when lectures end at the university, dark gray covers night between 12 midnight and 8am. SMS is used more for communication outside regular business hours.
doi:10.1371/journal.pone.0095978.g010

such high mobility. Although we acknowledge that this density estimation suffers due to the low number of samples, our measurements suggest that real participant mobility is underestimated in studies based solely on CDRs, such as in [38], as they fail to capture travels outside of the covered area.

Figure 8C shows a two-dimensional histogram of the locations, with hexagonal binning and logarithmic color scale (from blue to red). The red hotspots identify the most active places, such as the university campus and dormitories. The white spots are the frequently visited areas, such as major streets and roads, stations, train lines, and the city center.

From the raw location data we can extract stop locations as groups of locations clustered within distance D and time T [162–165]. By drawing edges between stop locations for each

participant, so that the most frequent transitions stand out, we can reveal patterns of collective mobility (Figure 8D).

Call and Text Communication Patterns

With the advent of mobile phones in the late 20th century, the way we communicate has changed dramatically. We are no longer restricted to landlines and are able to move around in physical space while communicating over long distances.

The ability to efficiently map communication networks and mobility patterns (using cell towers) for large populations has made it possible to quantify human mobility patterns, including investigations of social structure evolution [166], economic development [67], human mobility [37,38], spreading patterns [57], and collective behavior with respect to emergencies [60]. In
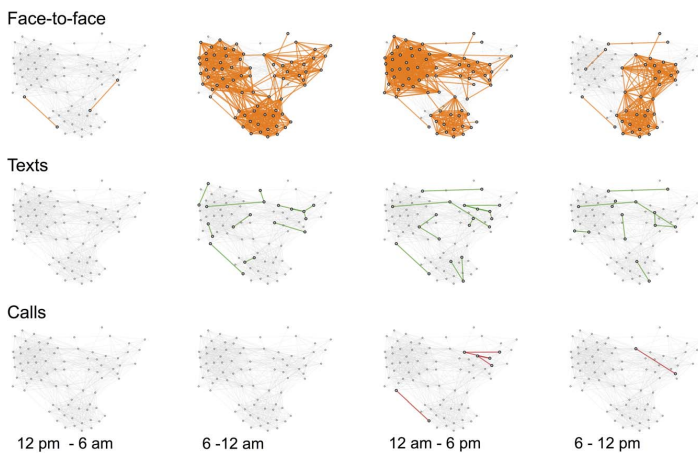
Figure 11. Daily activations in three networks. One day (Friday) in a network showing how different views are produced by observing different channels.
doi:10.1371/journal.pone.0095978.g011

Online

Face-to-face

Extra Info. Online

Extra Info. F2F
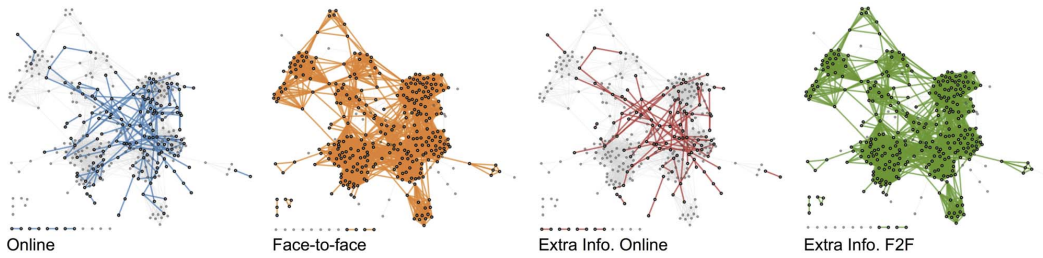
Figure 12. Face-to-face and online activity. The figure shows data from the 2013 deployment for one representative week. **Online**: Interactions (messages, wall posts, photos, etc.) between participants on Facebook. **Face-to-Face**: Only the most active edges, which account for 80% of all traffic, are shown for clarity. **Extra Info. F2F**: Extra information contained in the Bluetooth data shown as the difference in the set of edges. **Extra Info. Online**: Additional information contained in the Facebook data.
doi:10.1371/journal.pone.0095978.g012

A face-to-face and Facebook interactions
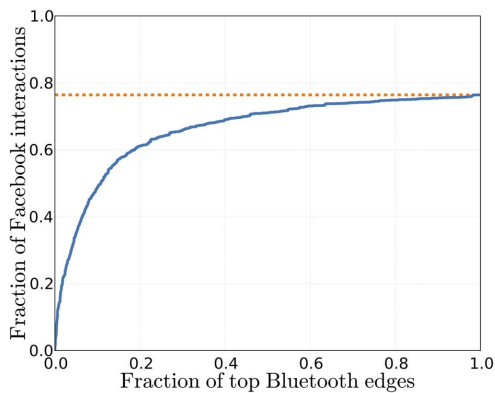
B face-to-face and voice calls
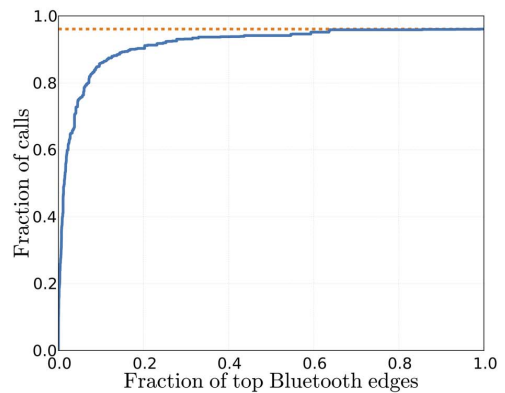
C voice calls and Facebook interactions
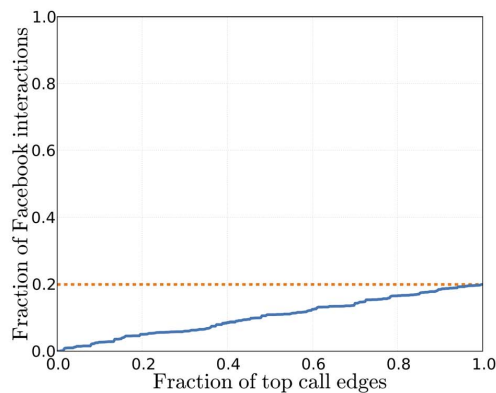
Figure 13. Network similarity. Defined as the fraction of ties from one communication channel that can be recovered by considering the top k fraction of edges from a different channel. Orange dashed line indicates the maximum fraction of ties the network accounts for. The strongest 10% of face-to-face interactions account for $>50\%$ of online ties and $\sim 90\%$ of call ties, while 23.58% of Facebook ties and 3.85% of call ties are not contained in the Bluetooth data. Between call and Facebook, the 10% strongest call ties account for $<3\%$ while in total $>80\%$ of Facebook ties are unaccounted. All values are calculated for interactions that took place in January 2014.
doi:10.1371/journal.pone.0095978.g013
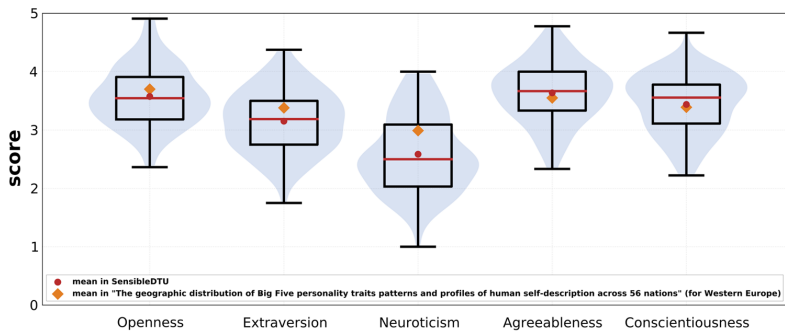
Figure 14. Personality traits. Violin plot of personality traits. Summary statistics are: **openness** $\mu_O = 3.58$, $\sigma_O = 0.52$; **extraversion** $\mu_E = 3.15$, $\sigma_E = 0.53$; **neuroticism** $\mu_N = 2.59$ $\sigma_N = 0.65$; **agreeablenes** $\mu_A = 3.64$ $\sigma_A = 0.51$; **conscientiousness** $\mu_C = 3.44$ $\sigma_C = 0.51$. Mean values from our deployment (red circles) compared with mean values reported for Western Europe (mixed student and general population) [170] (orange diamonds).
doi:10.1371/journal.pone.0095978.g014

this study, we have collected call logs from each phone as (caller, callee, duration, timestamp, call type), where the call type could be incoming, outgoing, or missed. Text logs contained (sender, recipient, timestamp, incoming/outgoing, one-way hash of content).

In the 2012 deployment we collected 56 902 incoming and outgoing calls, of which 42 157 had a duration longer than zero seconds. The average duration of the calls was $\langle d \rangle = 142.04s$, with a median duration of $48.0s$. The average ratio between incoming and outgoing calls for a participant was $r_{in/out} = 0.98$. In the same period, we collected 161 591 text messages with the average ratio for a participant $r_{in/out} = 1.96$.

We find a Pearson correlation of 0.75 ($p \ll 0.05$) between the number of unique contacts participants contacted via SMS and voice calls, as depicted in Figure 9. However, the similarity $\sigma = |N_{call} \cap N_{text}|/|N_{call} \cup N_{text}|$ between the persons a participant contacts via calls (N_{call}) and SMS (N_{text}) is on average $\langle \sigma \rangle = 0.37$, suggesting that even though participants utilize both forms of communication in similar capacity, those two are, in fact, used for distinct purposes.

Figure 10 shows the communication for SMS and voice calls (both incoming and outgoing, between participants and with the external world) as a time series, calculated through the entire year and scaled to denote the mean count of interactions participants had in given hourly time-bins in the course of a week. Also here, we notice differences between the two channels. While both clearly show a decrease in activity during lunch time, call activity peaks around the end of the business day and drops until next morning. In contrast, after a similar decrease that we can associate with commute, SMS displays another evening peak. Also at night, SMS seems to be a more acceptable form of communication, with message exchanges continuing late and starting early, especially on Friday night, when the party never seems to stop.

We point out that the call and SMS dynamics display patterns that are quite distinct from face-to-face interactions between participants as seen in Figure 5. Although calls and SMS communication are different on the weekends, the difference is not as dramatic as in the face-to-face interactions between the participants. This indicates that the face-to-face interactions we observe during the week are driven primarily by university-related activities, and only few of these ties manifest themselves during the

weekends, despite the fact that the participants are clearly socially active, sending and receiving calls and messages.

In Figure 11, we focus on a single day (Friday) and show activation of links between participants in three channels: voice calls, text messages, and face-to-face meetings. The three networks show very different views of the participants' social interactions.

Online friendships

The past years have witnessed a shift in our interaction patterns, as we have adapted new forms of online communication. Facebook is to date the largest online social community with more than 1 billion users worldwide [167]. Collecting information about friendship flows and communication flows allows us to construct a comprehensive picture of the online persona. Combined with other recorded communication channels we have an unparalleled opportunity to piece together an almost complete picture of all major human communication channels. In the following section we consider Facebook data obtained from the 2013 deployment. In contrast to the first deployment, we also collected interaction data in this deployment. For a representative week (Oct. 14–Oct. 21, 2013), we collected 155 interactions (edges) between 157 nodes, yielding an average degree $\langle d \rangle = 1.98$, average clustering $\langle c \rangle = 0.069$, and average shortest path in the giant component (86 nodes) $\langle l \rangle = 6.52$. The network is shown in the left-most panel of Figure 12. By comparing with other channels we can begin to understand how well online social networks correspond to real life meetings. The corresponding face-to-face network (orange) is shown in Figure 12, where weak links, i.e. edges with fewer than 147 observations (20%) are discarded. Corresponding statistics are for the 307 nodes and 3 217 active edges: $\langle d \rangle = 20.96$, $\langle c \rangle = 0.71$, and $\langle l \rangle = 3.2$. Irrespective of the large difference in edges, the online network still contains valuable information about social interactions that the face-to-face network misses—red edges in Figure 12.

A simple method for quantifying the similarity between two networks is to consider the fraction of links we can recover from them. Sorting face-to-face edges according to activity (highest first) we consider the fraction of online ties the top k Bluetooth links correspond to. Figure 13A shows that 10% of the strongest Bluetooth ties account for more than 50% of the Facebook interactions. However, as noted before, the Bluetooth channel does not recover all online interactions—23.58% of Facebook ties are unaccounted for. Applying this measure between Bluetooth
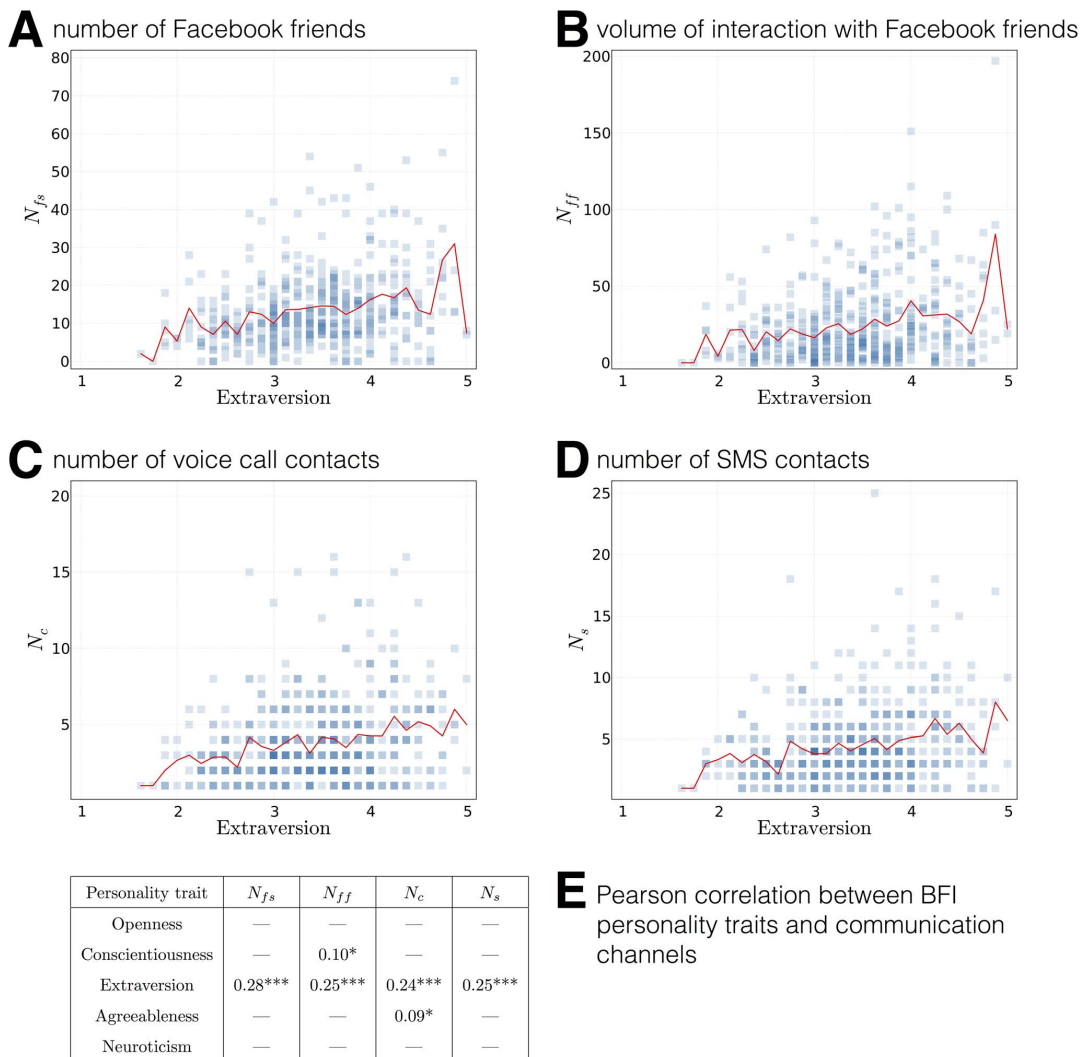
Figure 15. Correlation between personality traits and communication. Data from the 2013 deployment for N = 488 participants, showing communication only with other study participants. Extraversion, the only significant feature across all networks is plotted. The red line indicates mean value within personality trait. Random spikes are due to small number of participants with extreme values. **E)** Pearson correlation between Big Five Inventory personality traits and number of Facebook friends N_{fs}, volume of interactions with these friends N_{ff}, number of friends contacted via voice calls N_c and via SMS N_s. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.
doi:10.1371/journal.pone.0095978.g015

and voice calls (Figure 13B) shows a similar behavior, while there is low similarity between voice calls and Facebook ties (Figure 13C).

Personality traits

While the data from mobile sensing and online social networks provide insights primarily into the structure of social ties, we are also interested in the demographics, psychological and health traits, and interests of the participants. Knowing these characteristics, we can start answering questions about the reasons for the

observed network formation; why are ties created and what drives their dynamics? For example, homophily plays a vital role in how we establish, maintain, and destroy social ties [168].

Within the study, participants answered questions covering the aforementioned domains. These questions included the widely used *Big Five Inventory* [135] measuring five broad domains of human personality traits: openness, extraversion, neuroticism, agreeableness, and conscientiousness. The traits are scored on a 5-point Likert-type scale (low to high), and the average score of

questions related to each personality domain are calculated. As Big Five has been collected for various populations, including a representative sample from Germany [169] and a representative sample covering students mixed with the general population from Western Europe [170], we report the results from the 2012 deployment in Figure 14, suggesting that our population is unbiased with respect to these important traits.

Following the idea that personality is correlated with the structure of the social networks, we examine how the Big Five Inventory traits relate to the communication ego networks of the participants: number of Facebook friends, amount of communication with these friends, number of people ever contacted over voice calls or SMS. We only consider communication within the study, in the 2013 deployment for $N = 488$ participants for whom complete and longitudinal data was available. It is worth noting that participants answered the questions very early in the semester, and that we anecdotally know that a vast majority of the friendships observed between participants are 'new' in that they are between people who met when they started studying. Thus, we mainly observe the effect of personality on the network structure, not the other way around. The results are consistent with the literature, where Extraversion was shown to be correlated with number of Facebook friends [171]. Extending this result, Figure 15 depicts the correlation between Extraversion and number of Facebook friends (structural network) N_{fs} (Figure 15A), volume of interactions with these friends (functional network) N_{ff} (Figure 15B), number of friends contacted via voice calls N_c (Figure 15C), and number of friends contacted via SMS N_s (Figure 15D). In Table 15E, we show the (Pearson) correlation between all five traits and the aforementioned communication channels, reporting only significant results. The values of correlation for Extroversion are consistent across the networks, and are close to those reported in [171,172] (~ 0.2). Following the result from Call & Text Communication Patterns Section, where we showed that the communication in SMS and call networks are similar in volume, however have limited overlap in terms of who participants contact, both those channels show similar correlation with Extraversion. Here, we only scratched the surface with regard to the relation between personality and behavioral data. The relation between different behavioral features, network structure, and personality has been studied in [173–176]. By showing the impact of Extraversion on the network formed with participants inside the study is consistent with values reported for general populations, we indicate that within the Copenhagen Networks Study, we capture a true social system, with different personalities positioned differently in the network.

Perspectives

We expect that the amount of data collected about human beings will continue to increase. New and better services will be offered to users, more effective advertising will be implemented, and researchers will learn more about human nature. As the complexity and scale of studies on social systems studies grows, collection of high-resolution data for studying human behavior will become increasingly challenging on multiple levels, even when offset by the technical advancements. Technical preparations, administrative tasks, and tracking data quality are a substantial effort for an entire team, before even considering the scientific work of data analysis. It is thus an important challenge for the scientific community to create and embrace re-usable solutions, including best practices in privacy policies and deployment

procedures, supporting technologies for data collection, handling, and analysis methods.

The results presented in this paper—while still preliminary considering the intended multi-year span of the project—clearly reveal that a single stream of data rarely supplies a comprehensive picture of human interactions, behavior, or mobility. At the same time, creating larger studies, in terms of number of participants, duration, channels observed, or resolution, is becoming expensive using the current approach. The interest of the participants depends on the value they get in return and the inconvenience the study imposes on their lives. The inconvenience may be measured by decreased battery life of their phones, annoyance of answering questionnaires, and giving up some privacy. The value, on the other hand, is classically created by offering material incentives, such as paying participants or, as in our case, providing smartphones and creating services for the participants. Providing material incentives for thousands or millions of people, as well as the related administrative effort of study management, may simply not be feasible.

In the not-so-distant future, many studies of human behavior will move towards accessing already existing personal data. Even today we can access mobility of large populations, by mining data from Twitter, Facebook, or Flickr. Or, with participants' authorizations, we can track their activity levels, using APIs of self-tracking services such as Fitbit or RunKeeper. Linking across multiple streams is still difficult today (the problem of data silos), but as users take more control over their personal data, scientific studies can become consumers rather than producers of the existing personal data.

This process will pose new challenges and amplify the existing ones, such as the replicability and reproducibility of the results or selection bias in the context of full end-user data control. Still, we expect that future studies will increasingly rely on the existing data, and it is important to understand how the incomplete view we get from such data influences our results. For this reason, we need research testbeds—such as the Copenhagen Networks Study— where we study 'deep data' in the sense of multi layered data streams, sampled with high temporal resolution. These deep data will allow us to unlock and understand the future streams of big data.

Acknowledgments

Christensen, and Agnete Skovlund Dissing. From the Niels Bohr Institute (Physics): Joachim Mathiesen, and Mogens Høgh Jensen.

References

1. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2008) Detecting influenza epidemics using search engine query data. Nature 457: 1012–1014.
2. Aral S,Walker D (2012) Identifying influential and susceptible members of social networks. Science 337: 337–341.
3. Stopczynski A, Pietri R, Pentland A, Lazer D, Lehmann S (2014) Privacy in Sensor-Driven Human Data Collection: A Guide for Practitioners. arXiv preprint arXiv:14035299.
4. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, et al. (2007) Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104: 7332–7336.
5. Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in Twitter: The million follower fallacy. ICWSM 10: 10–17.
6. Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. Sociological Methods & Research 40: 211–239.
7. Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, et al. (2002) Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications 311: 590–614.
8. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. Science 311: 88–90.
9. Lazer D, Pentland A, Adamic L, Aral S, Barabási A, et al. (2009) Life in the network: the coming age of computational social science. Science (New York, NY) 323: 721.
10. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2013) The impact of biases in mobile phone ownership on estimates of human mobility. Journal of The Royal Society Interface 10: 20120986.
11. Madrigal A (2013) Dark social: We have the whole history of the web wrong. The Atlantic.
12. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, et al. (2010) Dynamics of personto-person interactions from distributed RFID sensor networks. PloS one 5: e11596.
13. Wu L, Waber B, Aral S, Brynjolfsson E, Pentland A (2008) Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task. Available at SSRN 1130251.
14. Polastre J, Szewczyk R, Culler D (2005) Telos: enabling ultra-low power wireless research. In: Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on. IEEE, pp. 364–369.
15. Raento M, Oulasvirta A, Eagle N (2009) Smartphones an emerging tool for social scientists. Sociological methods & research 37: 426–454.
16. Chronis I, Madan A, Pentland AS (2009) SocialCircuits: the art of using mobile phones for modeling personal interactions. In: Proceedings of the ICMI-MLMI'09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing. ACM, p. 1.
17. Pentland AS (2008) Honest Signals: how they shape our world. MIT Press.
18. Olguín D, Madan A, Cebrian M, Pentland A (2011) Mobile sensing technologies and computational methods for collective intelligence. Next Generation Data Technologies for Collective Computational Intelligence: 575–597.
19. Miller G (2012) The smartphone psychology manifesto. Perspectives on Psychological Science 7: 221–237.
20. Raento M, Oulasvirta A, Petit R, Toivonen H (2005) ContextPhone: A prototyping platform for context-aware mobile applications. Pervasive Computing, IEEE 4: 51–59.
21. Mulder I, Ter Hofte G, Kort J (2005) SocioXensor: Measuring user behaviour and user experience in context with mobile devices. In: Proceedings of Measuring Behavior. pp. 355–358.
22. Froehlich J, Chen MY, Consolvo S, Harrison B, Landay JA (2007) MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In: Proceedings of the 5th international conference on Mobile systems, applications and services. ACM, pp. 57–70.
23. Cornelius C, Kapadia A, Kotz D, Peebles D, Shin M, et al. (2008) AnonySense: privacy-aware people-centric sensing. In: Proceedings of the 6th international conference on Mobile systems, applications, and services. ACM, pp. 211–224.
24. Miluzzo E, Lane N, Fodor K, Peterson R, Lu H, et al. (2008) Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. In: Proceedings of the 6th ACM conference on Embedded network sensor systems. ACM, pp. 337–350.
25. Kostakos V, O'Neill E (2008) Cityware: Urban computing to bridge online and real-world social networks. Handbook of research on urban informatics: The practice and promise of the real-time city: 195–204.
26. Miluzzo E, Cornelius C, Ramaswamy A, Choudhury T, Liu Z, et al. (2010) Darwin phones: the evolution of sensing and inference on mobile phones. In: Proceedings of the 8th international conference on Mobile systems, applications, and services. ACM, pp. 5–20.

27. Hu X, Leung VC, Chu TH, Chan HC (2013) Vita: A crowdsensing-oriented mobile cyber-physical system. IEEE Transactions on Emerging Topics in Computing 1: 148–165.
28. Larsen JE, Jensen K (2009) Mobile Context Toolbox. In: Smart Sensing and Context, Springer. pp. 193–206.
29. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10: 255–268.
30. (2013). Funf Open Sensing Framework. URL http://funf.org/. [Online; accessed 19-March-2014].
31. Aharony N, Pan W, Ip C, Khayal I, Pentland A (2011) Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing 7: 643–659.
32. Kiukkonen N, Blom J, Dousse O, Gatica-Perez D, Laurila J (2010) Towards rich mobile phone datasets: Lausanne data collection campaign. Proc ICPS, Berlin.
33. Laurila J, Gatica-Perez D, Aad I, Blom J, Bornet O, et al. (2012) The mobile data challenge: Big data for mobile computing research. In: Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK.
34. Olguín D, Waber B, Kim T, Mohan A, Ara K, et al. (2009) Sensible organizations: Technology and methodology for automatically measuring organizational behavior. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 39: 43–55.
35. Karikoski J, Nelimarkka M (2011) Measuring social relations with multiple datasets. International Journal of Social Computing and Cyber-Physical Systems 1: 98–113.
36. Cranshaw J, Toch E, Hong J, Kittur A, Sadeh N (2010) Bridging the gap between physical location and online social networks. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing. ACM, pp. 119–128.
37. González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453: 779–782.
38. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. Science 327: 1018–1021.
39. Sevtsuk A, Ratti C (2010) Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. Journal of Urban Technology 17: 41–60.
40. Bagrow JP, Lin YR (2012) Mesoscopic structure and social aspects of human mobility. PLoS One 7: e37676.
41. De Domenico M, Lima A, Musolesi M (2013) Interdependence and predictability of human mobility and social interactions. Pervasive and Mobile Computing 9: 798–807.
42. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences 106: 15274–15278.
43. Eagle N, Pentland AS (2009) Eigenbehaviors: Identifying structure in routine. Behavioral Ecology and Sociobiology 63: 1057–1066.
44. Clauset A, Eagle N (2007) Persistence and periodicity in a dynamic proximity network. In: DyDAn Workshop on Computational Methods for Dynamic Interaction Networks.
45. Onnela JP, Saramäki J, Hyvönen J, Szabó G, De Menezes MA, et al. (2007) Analysis of a largescale weighted network of one-to-one human communication. New Journal of Physics 9: 179.
46. Granovetter MS (1973) The strength of weak ties. American Journal of Sociology: 1360–1380.
47. Lambiotte R, Blondel VD, de Kerchove C, Huens E, Prieur C, et al. (2008) Geographical dispersal of mobile communication networks. Physica A: Statistical Mechanics and its Applications 387: 5317–5325.
48. Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA (2011) Geographic constraints on social network groups. PLoS One 6: e16939.
49. Hidalgo CA, Rodriguez-Sickert C (2008) The dynamics of a mobile phone network. Physica A: Statistical Mechanics and its Applications 387: 3017–3024.
50. Miritello G, Lara R, Cebrian M, Moro E (2013) Limited communication capacity unveils strategies for human interaction. Scientific reports 3.
51. Miritello G, Moro E, Lara R, Martínez-López R, Belchamber J, et al. (2013) Time as a limited resource: Communication strategy in mobile phone networks. Social Networks 35: 89–95.
52. Sun L, Axhausen KW, Lee DH, Huang X (2013) Understanding metropolitan patterns of daily encounters. Proceedings of the National Academy of Sciences 110: 13774–13779.
53. Kapoor A, Eagle N, Horvitz E (2010) People, quakes, and communications: Inferences from call dynamics about a seismic event and its influences on a population. In: AAAI Spring Symposium: Artificial Intelligence for Development.

54. Madan A, Cebrian M, Moturu S, Farrahi K, Pentland S (2012) Sensing the health state of a community. IEEE Pervasive Computing 11: 36–45.

55. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, et al. (2010) A high-resolution human contact network for infectious disease transmission. Proceedings of the National Academy of Sciences 107: 22020–22025.

56. Madan A, Farrahi K, Gatica-Perez D, Pentland A (2011) Pervasive sensing to model political opinions in face-to-face networks. In:Lyons K, Hightower J, Huang E, editors, Pervasive Computing, Springer Berlin Heidelberg, volume 6696 of Lecture Notes in Computer Science. pp.214–231. doi: 10.1007/978-3-642-21726-514. URL http://dx.doi.org/10.1007/978-3-642-21726-5_14.

57. Wang P, González MC, Hidalgo CA, Barabási AL (2009) Understanding the spreading patterns of mobile phone viruses. Science 324: 1071–1076.

58. Isella L, Romano M, Barrat A, Cattuto C, Colizza V, et al. (2011) Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. PLoS One 6: e17144.

59. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, et al. (2011) High-resolution measurements of face-to-face contact patterns in a primary school. PLoS One 6: e23176.

60. Bagrow JP, Wang D, Barabási AL (2011) Collective response of human populations to large-scale emergencies. PLoS One 6: e17680.

61. Karsai M, Perra N, Vespignani A (2013) The emergence and role of strong ties in time-varying communication networks. arXiv preprint arXiv:13035966.

62. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. New England Journal of Medicine 357: 370–379.

63. Christakis NA, Fowler JH (2008) The collective dynamics of smoking in a large social network. New England Journal of Medicine 358: 2249–2258.

64. Lyons R (2011) The spread of evidence-poor medicine via awed social-network analysis. Statistics, Politics, and Policy 2.

65. Steglich C, Snijders TA, Pearson M (2010) Dynamic networks and behavior: Separating selection from influence. Sociological Methodology 40: 329–393.

66. Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences 106: 21544–21549.

67. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328: 1029–1031.

68. Blondel V, Krings G, Thomas I (2010) Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. Brussels Studies 42.

69. Mahato H, Kern D, Holleis P, Schmidt A (2008) Implicit personalization of public environments using Bluetooth. In: CHI'08 extended abstracts on Human factors in computing systems. ACM, pp. 3093–3098.

70. Klasnja P, Consolvo S, Choudhury T, Beckwith R, Hightower J (2009) Exploring privacy concerns about personal sensing. In: Proceedings of the 7th International Conference on Pervasive Computing. Berlin, Heidelberg: Springer-Verlag, Pervasive '09, pp. 176–183. doi: 10.1007/978-3-642-01516-8_13. URL http://dx.doi.org/10.1007/978-3-642-01516-8_13.

71. Altshuler Y, Aharony N, Elovici Y, Pentland A, Cebrian M (2011) Stealing reality: when criminals become data scientists (or vice versa). Security and Privacy in Social Networks: 133–151.

72. Shokri R, Theodorakopoulos G, Le Boudec J, Hubaux J (2011) Quantifying location privacy. In: Security and Privacy (SP), 2011 IEEE Symposium on. IEEE, pp. 247–262.

73. Lane N, Xie J, Moscibroda T, Zhao F (2012) On the feasibility of user de-anonymization from shared mobile sensor data. In: Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones. ACM, p. 3.

74. Srivatsa M, Hicks M (2012) Deanonymizing mobility traces: using social network as a side-channel. In: Proceedings of the 2012 ACM conference on Computer and communications security. ACM, pp. 628–637.

75. Mislove A, Viswanath B, Gummadi KP, Druschel P (2010) You are who you know: inferring user profiles in online social networks. In: Proceedings of the third ACM International Conference on Web search and data mining. ACM, pp. 251–260.

76. Zhou B, Pei J (2008) Preserving privacy in social networks against neighborhood attacks. In: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. IEEE, pp. 506–515.

77. Cheng J, Fu AWc, Liu J (2010) K-isomorphism: privacy preserving network publication against structural attacks. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, pp. 459–470.

78. Li T, Li N (2009) On the tradeoff between privacy and utility in data publishing. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 517–526.

79. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: Security and Privacy, 2008. SP 2008. IEEE Symposium on. IEEE, pp. 111–125.

80. Sweeney L (2000) Simple demographics often identify people uniquely. Health (San Francisco): 1–34.

81. Barbaro M, Zeller T, Hansell S (2006) A face is exposed for AOL searcher no. 4417749. New York Times 9: 8For.

82. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: The privacy bounds of human mobility. Scientific reports 3.

83. Sweeney L (2002) k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10: 557–570.

84. Machanavajjhala A, Kifer D, Gehrke J, VenkitasubramaniamM(2007) l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1: 3.

85. Li N, Li T, Venkatasubramanian S (2007) t-closeness: Privacy beyond k-anonymity and l-diversity. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE, pp. 106–115.

86. Dinur I, Nissim K (2003) Revealing information while preserving privacy. In: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, pp. 202–210.

87. Dwork C, Nissim K (2004) Privacy-preserving datamining on vertically partitioned databases. In: Advances in Cryptology–CRYPTO 2004. Springer, , pp. 134–138.

88. Blum A, Dwork C, McSherry F, Nissim K (2005) Practical privacy: the SuLQ framework. In: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, pp. 128–138.

89. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M (2006) Our data, ourselves: Privacy via distributed noise generation. Advances in Cryptology-EUROCRYPT 2006: 486–503.

90. Chawla S, Dwork C, McSherry F, Smith A, Wee H (2005) Toward privacy in public databases. Theory of Cryptography: 363–385.

91. Rivest RL, Adleman L, Dertouzos ML (1978) On data banks and privacy homomorphisms. Foundations of secure computation 4: 169–180.

92. Gentry C (2009) A fully homomorphic encryption scheme. Ph.D. thesis, Stanford University.

93. Tebaa M, El Hajji S (2012) Homomorphic encryption applied to the cloud computing security. In: Proceedings of the World Congress on Engineering. volume 1, pp. 4–6.

94. Naehrig M, Lauter K, Vaikuntanathan V (2011) Can homomorphic encryption be practical? In: Proceedings of the 3rd ACM workshop on Cloud computing security workshop. ACM, pp. 113–124.

95. Popa R, Balakrishnan H, Blumberg A (2009) VPriv: protecting privacy in location-based vehicular services. In: Proceedings of the 18th conference on USENIX security symposium. USENIX Association, pp. 335–350.

96. Molina A, Salajegheh M, Fu K (2009) HICCUPS: health information collaborative collection using privacy and security. In: Proceedings of the first ACM workshop on Security and privacy in medical and home-care systems. ACM, pp. 21–30.

97. Zdancewic SA (2002) Programming languages for information security. Ph.D. thesis, Cornell University.

98. Sfaxi L, Abdellatif T, Robbana R, Lakhnech Y (2010) Information flow control of componentbased distributed systems. Concurrency and Computation: Practice and Experience.

99. Zeldovich N, Boyd-Wickizer S, Mazières D (2008) Securing distributed systems with information flow control. In: Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation. USENIX Association, pp. 293–308.

100. Mundada Y, Ramachandran A, Feamster N (2011) Silverline: data and network isolation for cloud services. In: Proceedings of the 3rd USENIX conference on Hot topics in cloud computing. USENIX Association, pp. 13–13.

101. Pappas V, Kemerlis V, Zavou A, Polychronakis M, Keromytis AD (2012) CloudFence: Enabling users to audit the use of their cloud-resident data.

102. Ganjali A, Lie D (2012) Auditing cloud administrators using information ow tracking. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS). pp. 79–84.

103. Boneh D, Lipton R (1996) A revocable backup system. In: USENIX Security Symposium. pp. 91–96.

104. Perlman R (2005) The ephemerizer: Making data disappear. Technical report, Mountain View, CA, USA.

105. Perlman R (2005) File system design with assured delete. In: Security in Storage Workshop, 2005. SISW'05. Third IEEE International. IEEE, pp. 6–pp.

106. Geambasu R, Kohno T, Levy A, Levy HM (2009) Vanish: Increasing data privacy with selfdestructing data. In: Proc. of the 18th USENIX Security Symposium. p. 56.

107. Agrawal R, Haas PJ, Kiernan J (2003) Watermarking relational data: framework, algorithms and analysis. The VLDB journal 12: 157–169.

108. Cox IJ, Miller ML, Bloom JA (2000) Watermarking applications and their properties. In: Information Technology: Coding and Computing, 2000. Proceedings. International Conference on. IEEE, pp. 6–10.

109. Cox IJ, Linnartz JP (1998) Some general methods for tampering with watermarks. Selected Areas in Communications, IEEE Journal on 16: 587–593.

110. Cattuto C, Van den Broeck W, Barrat A, Colizza V, Pinton JF, et al. (2010) Dynamics of personto-person interactions from distributed RFID sensor networks. PLoS One 5: e11596.

111. Larsen JE, Sapiezynski P, Stopczynski A, Mørup M, Theodorsen R (2013) Crowds, Bluetooth, and Rock'N'Roll: Understanding music festival participant behavior. In: Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia. New York, NY, USA: ACM, PDM '13, pp. 11–18. doi:10.1145/2509352.2509399. URL http://doi.acm.org/10.1145/2509352.2509399.

112. Ranjan G, Zang H, Zhang ZL, Bolot J (2012) Are call detail records biased for sampling human mobility? SIGMOBILE Mob Comput Commun Rev 16: 33–44.

113. Isaacman S, Becker R, Cáceres R, Kobourov S, Martonosi M, et al. (2011) Identifying important places in peoples lives from cellular network data. In: Pervasive Computing, Springer. pp. 133–151.

114. Mucha P, Richardson T, Macon K, Porter M, Onnela JP (2010) Community structure in timedependent, multiscale, and multiplex networks. Science 328: 876–878.

115. Szell M, Lambiotte R, Thurner S (2010) Multirelational organization of large-scale social networks. Proceedings of the National Academy of Sciences USA 107: 13636–13641.

116. Madden M, Lenhart A, Cortesi S, Gasser U, Duggan M, et al. (2013). Teens, Social Media, and Privacy. URL http://www.pewinternet.org/~/media//Files/Reports/2013/PIP_TeensSocialMediaandPrivacy.pdf. [Online; accessed 19-March-2014].

117. Kossinets G (2006) Effects of missing data in social networks. Social Networks 28: 247–268.

118. Laumann E, Marsden P, Prensky D (1983) The boundary specification problem in network analysis, Sage Publications. pp. 18–34.

119. Saramäki J, Leicht E, López E, Roberts SG, Reed-Tsochas F, et al. (2014) Persistence of social signatures in human communication. Proceedings of the National Academy of Sciences 111: 942–947.

120. Holme P, Saramäki J (2012) Temporal networks. Physics reports 519: 97–125.

121. Shalizi C, Thomas A (2011) Homophily and contagion are generically confounded in observational social network studies. Sociological Methods & Research 40: 211–239.

122. Fowler J, Christakis N (2008) Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham heart study. British Medical Journal 337: a2338.

123. Christakis N, Fowler J (2009) Connected: the surprising power of our social networks and how they shape our lives. Little, Brown and Company.

124. Li I, Dey A, Forlizzi J (2010) A stage-based model of personal informatics systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 557–566.

125. Larsen JE, Cuttone A, Jørgensen SL (2013) QS Spiral: Visualizing periodic quantified self data. In: CHI 2013 Workshop on Personal Informatics in the Wild: Hacking Habits for Health & Happiness.

126. Cuttone A, Lehmann S, Larsen JE (2013) A mobile personal informatics system with interactive visualizations of mobility and social interactions. In: Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia. ACM, pp. 27–30.

127. Rocha L, Liljeros F, Holme P (2011) Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. PLoS Computational Biolology 7: e1001109.

128. Lee S, Rocha LE, Liljeros F, Holme P (2012) Exploiting temporal network structures of human interaction to effectively immunize populations. PLoS One 7: e36439.

129. Fortunato S (2010) Community detection in graphs. Physics Reports 486: 75–174.

130. Gulbahce N, Lehmann S (2008) The art of community detection. BioEssays 30: 934–938.

131. Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466: 761–764.

132. Fiedler M (1975) A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslovak Mathematical Journal 25: 619.

133. Bagrow JP, Lehmann S, Ahn YY (2011) Robustness and modular structure in networks. arxiv/1102.5085.

134. (2013). Facebook reports first quarter 2013 results. URL http://investor.fb.com/releasedetail.cfm?ReleaseID=761090. [Online; accessed 19-March-2014].

135. John OP, Srivastava S (1999) The big five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research 2: 102–138.

136. Rosenberg M (1989) Society and the adolescent self-image (rev). Wesleyan University Press.

137. Back MD, Küfner AC, Dufner M, Gerlach TM, Rauthmann JF, et al. (2013) Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. Journal of Personality and Social Psychology 105: 1013.

138. Diener E, Emmons RA, Larsen RJ, Griffin S (1985) The satisfaction with life scale. Journal of personality assessment 49: 71–75.

139. Rotter JB (1966) Generalized expectancies for internal versus external control of reinforcement. Psychological monographs: General and applied 80: 1.

140. Russell DW (1996) UCLA loneliness scale (version 3): Reliability, validity, and factor structure. Journal of personality assessment 66: 20–40.

141. Sherer M, Maddux JE, Mercandante B, Prentice-Dunn S, Jacobs B, et al. (1982) The self-efficacy scale: Construction and validation. Psychological reports 51: 663–671.

142. Cohen S, Kamarck T, Mermelstein R (1983) A global measure of perceived stress. Journal of health and social behavior: 385–396.

143. Bech P, Rasmussen NA, Olsen LR, Noerholm V, Abildgaard W (2001) The sensitivity and specificity of the major depression inventory, using the present state examination as the index of diagnostic validity. Journal of affective disorders 66: 159–164.

144. Lund R, Nielsen LS, Henriksen PW, Schmidt L, Avlund K, et al. (2014) Content validity and reliability of the Copenhagen Social Relations Questionnaire. Journal of aging and health 26: 128–150.

145. Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of personality and social psychology 54: 1063.

146. Ellen RF, Firth R (1984) Ethnographic research: A guide to general conduct. Academic Press London.

147. de Montjoye YA, Wang SS, Pentland A, Anh DTT, Datta A, et al. (2012) On the trusted use of large-scale personal data. IEEE Data Eng Bull 35: 5–8.

148. Shampanier K, Mazar N, Ariely D (2007) Zero as a special price: The true value of free products. Marketing Science 26: 742–757.

149. (2013). SensibleDTU informed consent form (da). URL https://github.com/MIT-Model-Open-Data-and-Identity-System/SensibleData-Service/blob/production_sensibledtu1k/sensible_data_service/documents/service_informed_consent_da.txt.

150. (2013). SensibleDTU informed consent form (en). URL https://github.com/MIT-Model-Open-Data-and-Identity-System/SensibleData-Service/blob/production_sensibledtu1k/sensible_data_service/documents/service_informed_consent_en.txt.

151. Sekara V, Lehmann S (2014) Application of network properties and signal strength to identify face-to-face links in an electronic dataset. arXiv preprint arXiv:14015836.

152. Stopczynski A, Larsen JE, Lehmann S, Dynowski L, Fuentes M (2013) Participatory Bluetooth sensing: A method for acquiring spatio-temporal data about participant mobility and interactions at large scale events. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on. IEEE, pp. 242–247.

153. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. EPJ Data Science 1: 1–16.

154. Ribeiro B, Nicola P, Baronchelli A (2013) Quantifying the effect of temporal resolution on timevarying networks. Scientific reports 3.

155. Whitehead M, Phillips T, Page M, Molina M, Wood C (2012). European mobile industry observatory 2011. URL http://www.gsma.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf. [Online; accessed 19-March-2014].

156. LaMarca A, Chawathe Y, Consolvo S, Hightower J, Smith I, et al. (2005) Place Lab: Device positioning using radio beacons in the wild. In: Pervasive Computing, Springer. pp. 116–133.

157. Kjærgaard MB, Nurmi P (2012) Challenges for Social Sensing Using WiFi Signals. In: Proceedings of the 1st ACM workshop on Mobile systems for computational social science. New York, NY, USA: ACM, MCSS '12, pp. 17–21. doi:10.1145/2307863.2307869. URL http://doi.acm.org/10.1145/2307863.2307869.

158. Carlotto A, Parodi M, Bonamico C, Lavagetto F, Valla M (2008) Proximity classification for mobile devices using Wi-Fi environment similarity. In: Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments. ACM, pp. 43–48.

159. Carreras I, Matic A, Saar P, Osmani V (2012) Comm2Sense: Detecting proximity through smartphones. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on. IEEE, pp. 253–258.

160. Lin M, Hsu WJ (2013) Mining GPS data for mobility patterns: A survey. Pervasive and Mobile Computing.

161. (2013). statsmodels.nonparametric.kde.kdeuniivariate. URL http://statsmodels.sourceforge.net/devel/generated/statsmodels.nonparametric.kde.KDEUnivariate.html. [Online; accessed 19-March-2014].

162. Hariharan R, Toyama K (2004) Project Lachesis: parsing and modeling location histories. In: Geographic Information Science, Springer. pp. 106–124.

163. Zheng Y, Zhang L, Xie X, Ma WY (2009) Mining interesting locations and travel sequences from GPS trajectories. In: Proceedings of the 18th international conference on World wide web. ACM, pp. 791–800.

164. Montoliu R, Gatica-Perez D (2010) Discovering human places of interest from multimodal mobile phone data. In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia. ACM, p. 12.

165. Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with GPS history data. In: Proceedings of the 19th international conference on World Wide Web. ACM, pp. 1029–1038.

166. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. Nature 446: 664–667.

167. (2013). "Facebook Reports Third Quarter 2013 Results". URL http://investor.fb.com/releasedetail.cfm?ReleaseID=802760. [Online; accessed 19-March-2014].

168. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annual review of sociology: 415–444.

169. Blüml V, Kapusta ND, Doering S, Brähler E, Wagner B, et al. (2013) Personality factors and suicide risk in a representative sample of the german general population. PLoS One 8: e76646.

170. Schmitt DP, Allik J, McCrae RR, Benet-Martínez V (2007) The geographic distribution of big five personality traits patterns and profiles of human self-description across 56 nations. Journal of Cross-Cultural Psychology 38: 173–212.

171. Quercia D, Lambiotte R, Stillwell D, Kosinski M, Crowcroft J (2012) The personality of popular facebook users. In: Proceedings of the ACM 2012 conference on computer supported cooperative work. ACM, pp. 955–964.

172. Swickert RJ, Rosentreter CJ, Hittner JB, Mushrush JE (2002) Extraversion, social support processes, and stress. Personality and Individual Differences 32: 877–891.

173. Staiano J, Lepri B, Aharony N, Pianesi F, Sebe N, et al. (2012) Friends don't lie: inferring personality traits from social network structure. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, pp. 321–330.

174. Chittaranjan G, Blom J, Gatica-Perez D (2011) Who's who with big-five: Analyzing and classifying personality traits with smartphones. In: Wearable Computers (ISWC), 2011 15th Annual International Symposium on. IEEE, pp. 29–36.

175. Kalish Y, Robins G (2006) Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. Social Networks 28: 56–84.

176. de Montjoye YA, Quoidbach J, Robic F, Pentland AS (2013) Predicting personality using novel mobile phone-based metrics. In: Social Computing, Behavioral-Cultural Modeling and Prediction, Springer. pp. 48–55.

The Long Tail Issue in Large Scale Deployment of Personal Informatics

Andrea Cuttone, Jakob Eg Larsen.

The Long Tail Issue in Large Scale Deployment of Personal Informatics

Andrea Cuttone
Tech. University of Denmark
Dept. of Applied Mathematics
and Computer Science
Cognitive Systems Section
Matematiktorvet, Bld. 324
2800 Kgs Lyngby, Denmark
ancu@dtu.dk

Jakob Eg Larsen
Tech. University of Denmark
Dept. of Applied Mathematics
and Computer Science
Cognitive Systems Section
Matematiktorvet, Bld. 324
2800 Kgs Lyngby, Denmark
jaeg@dtu.dk

Abstract
We describe the challenges and the open questions arising
during the design and deployment of SensibleJournal, a
mobile personal informatics system with interactive
visualizations of mobility and social interactions based on
data acquired from embedded smartphone sensors. The
SensibleJournal system was evaluated in a large scale
(N=136) mobile sensing field study. We report issues in
deployment, limitations in user engagement and uptake,
and the challenges in measuring the effect of the system.

Author Keywords
Large-scale mobile sensing, self-tracking, deployment,
mobility, social interaction, feedback interfaces, behavior
change, personal informatics, quantified self

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g.,
HCI)]: Miscellaneous.

Introduction
SensibleJournal [3] is a Personal Informatics system for
Android smartphones with visualizations of personal
mobility and social interactions. The mobile app was
deployed to N=136 first year university students at our
university campus on an opt-in basis for approximately 6
months from October 2012, as part of the larger

Copenhagen Networks Study project [6], in which we measure mobility and social interactions with high resolution. The SensibleJournal mobile app provides four different feedback interfaces [3]:

- Stats: a daily summary of mode of transportation, top places and distance travelled

- Movement: a map with animated playback of daily movements and places visited

- TimeSpiral: a spiral visualization [4] of time series of places visited to highlight periodic patterns and reoccuring events

- Bubbles: a bubble chart of your social contacts and communities inferred from Bluetooth proximity

Here we evaluate the usage of SensibleJournal as a personal informatics system, thus participants used a smartphone with the SensibleJournal app as their primary phone, while detailed usage of the mobile app feedback interfaces were logged.

Personal Informatics for Self-Reflection

One of the main objectives in Personal Informatics is to facilitate self-reflection through a deeper understanding of personal behavior [5] [1]. However, it is challenging to quantify this as a precise metric. What does it mean to have a deeper understanding and what exactly causes behavior change (if it occurs)?

In our case we applied qualitative analysis using questionnaires asking the participants how often the app was used, how interested the participants were in the feedback, and if they discovered something new about their own behavior. As it is challenging to convert the qualitative data from the answers into unambigous measures, we also analyzed the usage patterns of the system from the usage logs. This allow us to infer the interests in the different features in the app. From the analysis of the usage logs we were able to quantify the frequency of use and time spent on the different feedback interfaces. However, it is not clear how much these usage patterns are indicative of increased understanding of personal behavior and whether it facilitates self-reflection.

Personal Informatics systems typically support the reflection stage [5] through some type of data visualization. A possible approach of the evaluation is to measure the effectiveness of the chosen visualization for representing the considered data. In our system we evaluated the proposed feedback interfaces in terms of which insights they were offering. However, the evaluation of the visualization as such is still an open problem in research [2], and often there is no conclusive answer to which visualization type is optimal for a given dataset. Even choosing an appropriate visualization is open to interpretation, as this representation may vary from person to person depending on cultural and technical background.

Challenges in the Large Scale Deployment

From our survey with 45 respondents (33%) [3] we found that keeping users engaged was very challenging, and our system failed to engage but a limited subset of the participants on a regular basis.

Many respondents commented that the app was interesting and provided new insights about personal behavior, however 58% of the respondents reported using the app less than once a week, 31% about once a week, and 11% more than once a week.
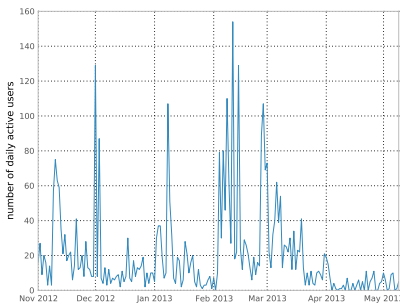
Figure 1: Number of active users per day over the duration of the full study. The aggregate usage data contains many peaks and valleys, and some peaks happened immediately after the release of a new version of the app that introduced new features (1 January and 4 February).

For quantifying user engagement, we analyzed the usage logs in the form <userid, timestamp, event>, with events generated for each different view of the app, and when the app is sent to the background. For each day, we count a user as active if he/she had at least one event of duration ≥ 10 seconds.

Figure 1 shows that the number of active users and contains many peaks and valleys. Some peaks happened immediately after the release of a new version of the app (1 Jan and 4 Feb). This may indicate that after releases of new versions of the app, there are spikes of interest. However, the usage rapidly declined in the days and weeks following. We suggest that one possible explanation is that the provided feedback was not interesting enough to

justify frequent usage. We suspect that one reason for this is that human activity is highly periodic and regular, so the provided feedback such as daily movement patterns would become repetitive and uninformative after a while. This suggests that highlighting deviations from routine might provide higher value to the users.
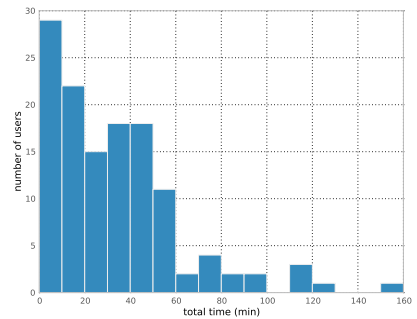
Figure 2: Distribution of the total time spent by number of users. The distribution has an exponential-like decay, where the large majority of participants spent a limited amount of time with the app, while only 10-15% of the participants were more active.

From the usage logs, we could also deduce the cumulative time that each user spent on the app and particular feedback interfaces. Figure 2 shows that there is a very large variation of usage patterns among the users. The time distribution has an exponential-like decay, where the large majority of participants spent a small amount of time with the app, while a few were much more active.

This "long tail" effect illustrates a challenge with limited sustained uptake among this broad population in the large scale deployment. The limited sustained usage suggests that this approach is quite costly in terms of evaluating the personal informatics feedback system as such.

We found that participants preferred simpler but less informative feedback to more complex but more informative one. Usage logs showed a preference for the simple Movement view over the TimeSpiral view, and several participants reported not understanding how to interpret the TimeSpiral. This level of quantitative data about own behavior is a novel concept, thus a suitable level of complexity in the feedback interface is necessary.

The goal of SensibleJournal was limited to facilitate reflection. However, promoting behaviour change is often the goal of a Personal Informatics system, and it can be unclear how self-reflection may actually promote such behavior change. An increased awareness of limited mobility or social behavior does not necessarily push towards a more active or social lifestyle. An open question is how to design a feedback loop that uses self-reflection to promote positive behavior change. In our survey a few participants reported increased awareness of their sedentary behavior, but it was not clear if this would lead to actual behavior change or the sustainability of this increased awareness longer-term.

Conclusions

The deployment of a mobile personal informatics system as part of a large scale mobile sensing study has been described. The challenges in identifying a suitable metric for evaluation beyond usage logs was discussed and especially the difficulty in creating a system that engage participants and keep a sustained interest. Measuring

actual behavior change caused by usage of the system was not possible. So although a large scale deployment may appear promising for obtaining substantial data and novel insights, the "long tail" issue of limitations in uptake and sustained usage lead to limited results considering the scale and cost of the approach.

Acknowledgments

References

[1] Baumer, E. P., Khovanskaya, V., Matthews, M., Reynolds, L., Sosik, V. S., and Gay, G. K. Reviewing reflection: On the use of reflection in interactive system design. In *Proceedings of DIS'14*, ACM (2014).

[2] Chen, C. Top 10 unsolved information visualization problems. *Computer Graphics and Applications, IEEE 25*, 4 (2005), 12–16.

[3] Cuttone, A., Lehmann, S., and Larsen, J. E. A mobile personal informatics system with interactive visualizations of mobility and social interactions. In *Proc. of the 1st ACM Int. Workshop on Personal Data Meets Distributed Multimedia*, ACM (2013), 27–30.

[4] Larsen, J. E., Cuttone, A., and Jørgensen, S. L. Qs spiral: Visualizing periodic quantified self data. In *CHI 2013 Workshop on Personal Informatics in the Wild: Hacking Habits for Health & Happiness* (2013).

[5] Li, I., Dey, A., and Forlizzi, J. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2010), 557–566.

[6] Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., and Lehmann, S. Measuring large-scale social networks with high resolution. *PLoS ONE 9*, 4 (04 2014), e95978.

Who Wants to Self‑Track Anyway? Measuring the Relation between Self‑Tracking Behavior and Personality Traits

Georgios Chatzigeorgakidis, **Andrea Cuttone**, Sune Lehmann, Jakob Eg Larsen.

Who Wants to Self-Track Anyway? Measuring the Relation between Self-Tracking Behavior and Personality Traits

Georgios Chatzigeorgakidis[1,4] Andrea Cuttone[2*], Sune Lehmann[2,3], Jakob Eg Larsen[2]

1 Athena Research and Innovation Center, Athens, Greece
2 DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark
3 The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
4 Department of Informatics and Telecommunications, University of Peloponnese, Tripolis, Greece

* ancu@dtu.dk

Abstract

We describe an empirical study of the usage of a mobility self-tracking app, *SensibleJournal 2014*, which provides personal mobility information to $N=796$ participants as part of a large mobile sensing study. Specifically, we report on the app design, as well as deployment, uptake and usage of the app. The latter analysis is based on logging of user interactions as well as answers gathered from a questionnaire provided to the participants. During the study enrollment process, participants were asked to fill out a questionnaire including a Big Five inventory and Narcissism NAR-Q personality tests. A comparison of personality traits was conducted to understand potential differences among the users and non-users of the app. We found a relation between self-tracking and conscientiousness, but contrary to the view in popular media, we found no relation between self-tracking behavior and narcissism.

Introduction

Recently, the area of lifelogging, Quantified Self and Personal Informatics have gained substantial attention and uptake due to the availability of smartphones, low-cost wearable sensors and, more recently, smart watches. This development has significantly lowered the barrier for people to engage in a wide range of self-tracking activities, with monitoring of exercise, physical activity, and step counting being widely adopted. Also in research, the area has gained increased attention in recent years, with international workshops on Personal Informatics [1–3], as well as conferences having sessions on Quantified Self. However, paradoxically, empirical research describing the self-tracking phenomenon is somewhat limited [4, 5].

In this paper we describe an empirical study of mobility self-tracking, using a smartphone app that has been developed as part of our research. We measured the usage of the self-tracking app on a population of almost 800 individuals – bachelor level university students from all technical sciences, for a duration of four months. In an ongoing mobile sensing study [6] we offered all participants a self-tracking app that provides a feedback interface reporting on personal mobility patterns on a daily basis. All participants were informed about the existence of the app through a notification system. While all participants were instructed to install the app on their smartphone, usage of the app was optional.

The motivation behind the study was to measure the uptake of a Quantified Self app in order to see how many would be interested in self-tracking, as well as the duration of the interest in the app, in terms of usage patterns over time. As the participants were enrolled in the mobile sensing study at the time of the introduction of the app, mobility and social interaction data were already being collected. Thus, the participants were merely provided with a mobile feedback interface providing access to personal data that was already being collected about them.

As it has been debated whether specific personality traits are distinctive among self-trackers, we wanted to measure the relation between self-tracking behavior and personality traits. In particular, the *narcissism* label has been associated with self-tracking in popular media and has been debated in research literature too [5, 7].

Related work

Li et al. [8] describe Personal Informatics using a model that involves a five stage process, where a key stage is *self-reflection*, which is often supported through visualization of the collected personal data. Data visualization is seen as a mean to gain insights into personal behaviors, which can shape the basis for achieving behavior change.

There is a plethora of examples, both scientific and commercial, that attempt to leverage the potential of data visualization as the means for individuals to interact with and gain insights from personal data. The most relevant to this study is the 2013 version of SensibleJournal, providing visualizations of mobility patterns [9], as well as visualizations on social interactions of its users [10]. Other examples that involve visualizing information on a map are *Personal Driving Diary* [11], which presents images of the detected events during driving along with pinpointed locations, and *Now Let Me See Where I Was* [12], which creates mobility-based visualizations for each participant by pinpointing locations on a map.

The self-tracking phenomenon has attracted attention in popular media, with articles sometimes suggesting a tendency of self-trackers towards narcissism [13]. Similarly, in recent report from Symantec on Quantified Self data security, Quantified Self is described as part of "*a trend towards [...] narcissism*" [14].

Research literature has discussed this popular media view on the practice of self-tracking as obsessive or narcissistic [7, 15, 16] and in [17], the initial encounter with self-tracking is being described as something that would appear to be "*just another example of technology stretching the limits of narcissism*". However, this viewpoint on self-tracking has been criticized by Bode and Kristensen [5], with a call for a more varied description of the self-tracking phenomenon.

An attempt to measure narcissism among self-trackers has been made on a small scale (N=36) in an online survey in the Quantified Self community, using the NPI-16 test [18] with 16 questions related to narcissism [19]. The result was a 0.38 score on narcissism compared to the mean scores in five American studies, which reported to be in the range 0.31-0.41 [18]. While concluding that there was no correlation between self-tracking and narcissism, it is also suggested that the definition of narcissism may not be clear [19].

Method

This work is part of the *SensibleDTU* project, a large-scale study of high-resolution social networks, which is described in detail in [6]. Data collection was approved by the Danish Data Protection Agency, and informed consent has been obtained for all study

all participants. We study N=796 first year students provided with an Android smartphone. The phone is equipped with a data collector app running continuously in the background. The latter collects and periodically uploads data to a server from multiple sources: location, Bluetooth, calls, SMS, and WiFi. The participants were also asked to fill out a questionnaire including the Big Five inventory [20] and Narcissism NAR-Q [21], from which we can deduce the following six personality traits: *extraversion*, *agreeableness*, *conscientiousness*, *neuroticism*, *openness* and *narcissism*. We have found that our population under study is unbiased with respect to the personality traits of the general population [6].

All participants that had joined the study were requested to install SensibleJournal 2014, a self-tracking mobile app designed to support self-reflection [8] on personal mobility, in terms of places visited and movements between them. The locations are extracted by clustering groups of consecutive locations within a predetermined distance, as described in [22].

The SensibleJournal 2014 App

The mobile app uses a card-based user interface, which shows mobility related information on cards that appear on a continuous timeline from most to least recent. Each card contains a static mini-map, which pinpoints locations of interest, along with specific informative text. We provide six different types of cards, each presenting different information about personal mobility: "*My Current Location*", "*Last Visited Place*", "*Latest Journey*", "*Daily Route*", "*Weekly Route*", and "*Most Visited Places*". The cards contain a static map along with descriptive text. An example card is shown in Fig. 1.

A card can be tapped in order to open the corresponding detailed view, which offers a more informative visualization through an interactive map which offers the ability to pan and zoom. Users can also access their history through an archive view that chronologically lists specific older detailed views, accessible from a "navigation drawer" (by tapping on the app title area). In order for the participants to be reminded to check for newly available cards, the app sends a periodic reminder using a notification on users' devices. To avoid intrusive behavior, the notification is sent once every three days at noon.

Usage Data Collection

SensibleJournal 2014 gathers data about the way users interact with the app. In particular, the app logs interaction events, such as when a user launches/pauses the app, or navigates through the interface. Each card contains an "*Awesome!*" button (see Fig. 1), which can be tapped in order to provide feedback about the cards. The usage log is periodically uploaded and stored on our server.

Results

We analyze the usage logs between late June and October 2014. For each participant we consider events that lasted at least 5 seconds and at most 10 minutes to avoid considering accidental app launches, or anomalies in the usage collection. In total, 242 (30.4%) of all 796 participants had no interaction with the app at all. Even though they had to install the app when joining the study, they never launched it. The cumulative distribution function of the total number of times that the app was launched per user, illustrated in Fig. 2, shows that the usage decays exponentially: around 60% of the participants launched the app less than 5 times or not at all and less than 5% launched it more than 20 times. This is in line with our findings in previous work [23].
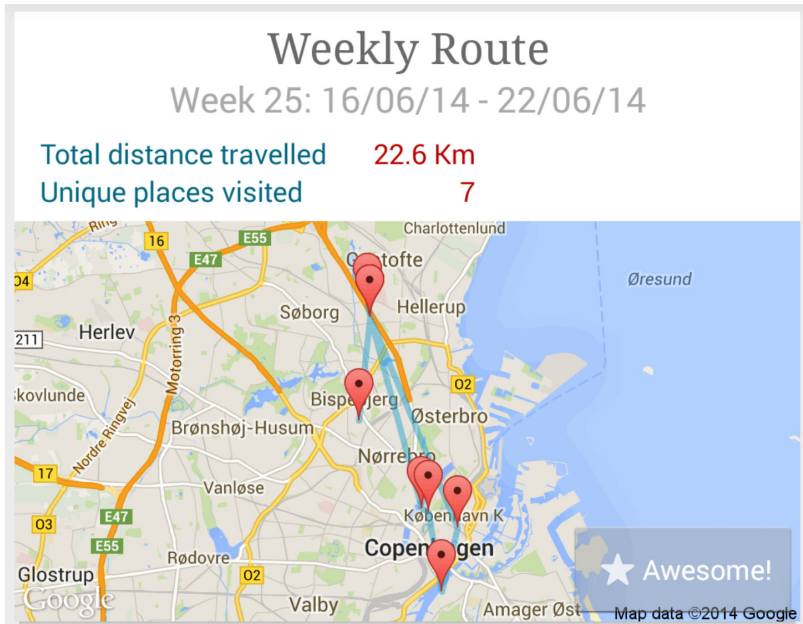
Fig 1. An example card showing mobility information in the SensibleJournal 2014 mobile app. The app user interface contains a list of such cards showing different views of personal mobility

Additionally, we count the per-day number of users with at least one launch (Fig. 3). The number of active users slowly decays from the start of the experiment. There is a peak in the beginning of September, which coincides with the start of the new university semester. A similar decay over time was also reported in previous work [23].

From the 796 participants, only 16 individuals (2%) used SensibleJournal 2014 more systematically. We defined as more regular users, the ones that used the app at least 20 times and at least once per month during the experiment.

Usage and Notifications

As mentioned, a notification system alerted the participants about new cards every three days, at noon. Fig. 4 shows that the total number of the app launches was significantly higher between 12:00 and 14:00, which suggests that the notification was an important factor in engaging users (Fig. 3). This observation is in line with [24], which reported a notably higher usage of a self-tracking mobile app with reminding through notifications.

Survey

In December 2014, all the participants were contacted and asked to fill out an electronic survey with the following questions:

1. Have you discovered something new or interesting about yourself? If yes, what? (open answer)
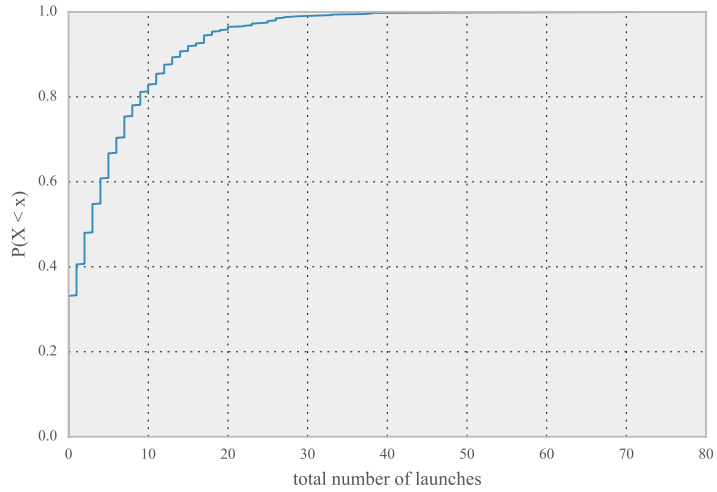
Fig 2. Cumulative distribution function of usage (number of app launches). The usage decays exponentially, with only a small percentage of participants using the app on a regular basis
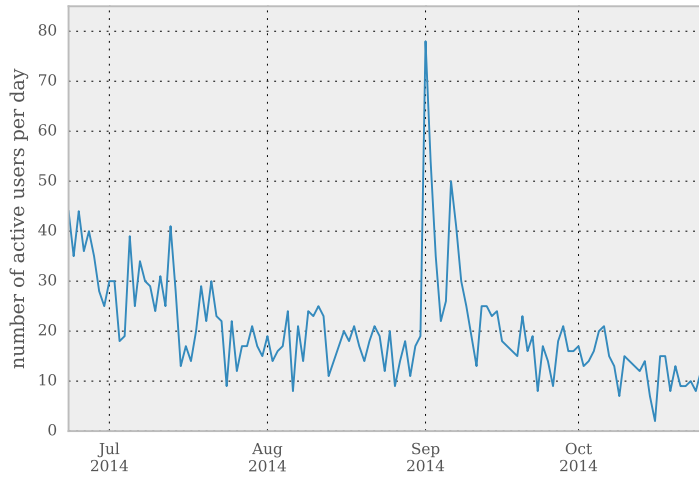
Fig 3. The number of active users as a function of time illustrates the decreasing trend from the beginning of the experiment, with an exception being early September, which coincides with the beginning of a new semester

2. If you no longer use SensibleJournal 2014 app, why not? (multiple options)
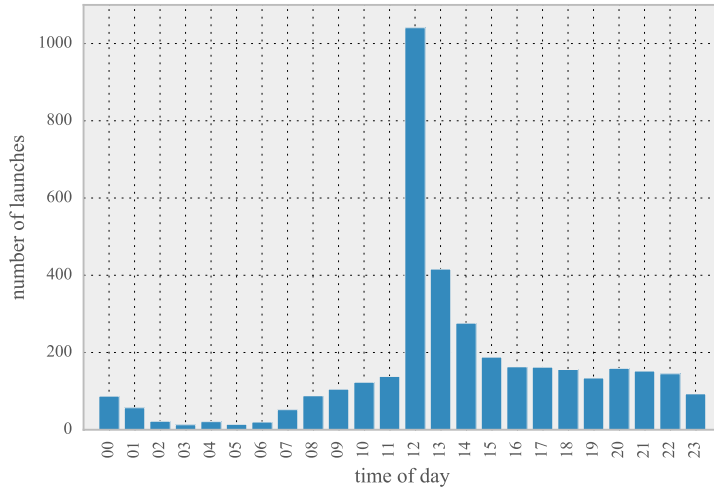
Fig 4. Usage per time of day. A clear peak happens between 12 and 14, probably due to the notifications scheduled at noon

    (a) The app is too slow

    (b) The visualizations are confusing

    (c) I am not interested in my location data

    (d) I do not know the app

    (e) The app always shows similar information

    (f) I do not learn anything new from my data

3. What do you think of the notifications?

    (a) There should be more

    (b) There should be less

    (c) They should be removed

4. Have you clicked "Awesome!"? If yes, why?

    (a) I liked the visualizations

    (b) I liked the information shown in the visualizations

5. How do you use the app? (open answer)

6. Do you have any other comments or suggestions? (open answer)

A total of 51 participants answered the survey (6% response rate). The majority reported not learning anything interesting, however some reported being surprised to gain new knowledge about their daily patterns: "*I am very surprised about my monotonous patterns, home-work-home*", "*Although I thought that I move around a lot, I basically spend my time with a specific friend of mine*", "*there is a whole routine in our lives, that is surprising!*" and "*I go out at the same places (more or less)*". One

participant reported checking the app after a change in his/her routine, and several reported using the app for learning about the time spent between home and university. The repetitiveness of daily routine and consequently of the app feedback, was one of the prominent causes for many users to stop using it: 43% reported not learning anything new and 22% reported that the app shows similar information every time. One participant even suggests that the app should provide feedback only when "*something new happens*" such as visiting a place never seen before. Moreover, 16% complained that the visualizations were confusing and 28% said not to be interested in their location data. Regarding the "*Awesome!*" button, 18% reported to have clicked it because they liked the visualizations and 20% because they found the information shown in the visualizations useful. 16% reported that there should be more notifications, 30% preferred less and 24% would prefer not to have notifications at all.

Personality

For each participant we compute the following six personality traits based on the answers to the questionnaire including the Big Five inventory [20] and Narcissism NAR-Q [21]: extraversion, agreeableness, conscientiousness, neuroticism, openness and narcissism. Additionally, for each participant we determine a number of features based on the collected usage data:

- number of days with at least one launch
- total time interacting with the app
- total number of launches
- mean session duration

For the users who have no usage data we assigned the value 0 to all the above mentioned features. We split the population into the top 10% and remaining 90% quantiles according to each usage feature and compared the distribution of each personality trait between the top and remaining quantiles using t-tests (6 personality traits x 4 usage features = 24 tests). Table 1 contains the corresponding p-values for each trait and feature pair.

Table 1. p-values for the t-tests for each trait-feature pair

	extra-version	agreeable-ness	conscien-tiousness	neuroti-cism	openness	narcissism
total events	0.6970	0.5653	0.0342	0.5141	0.2301	0.0677
total time	0.6943	0.7673	0.0009	0.5214	0.4531	0.3882
mean session dur.	0.3382	0.3485	0.4042	0.4401	0.0614	0.7776
active days	0.0416	0.3702	0.3418	0.2676	0.0319	0.1147

However, after correcting for multiple comparisons using the Holm-Bonferroni method [25], we find that the only statistically significant difference is between conscientiousness and total time. Table 2 displays the corrected p-values.

The Holm-Bonferroni correction is quite strict, therefore we provide another view of the results using bootstrap to illustrate the differences. We calculate the bootstrapping distributions of the means of the top 10% quantiles, for each personality trait-usage feature pair. In particular:

Table 2. p-values for each trait-feature pair, after the Holm-Bonferroni correction

	extra-version	agreeable-ness	conscien-tiousness	neuroti-cism	openness	narcissism
total events	1.0000	1.0000	0.7515	1.0000	1.0000	1.0000
total time	1.0000	1.0000	**0.0210**	1.0000	1.0000	1.0000
mean session dur.	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
active days	0.8726	1.0000	1.0000	1.0000	0.7338	1.0000
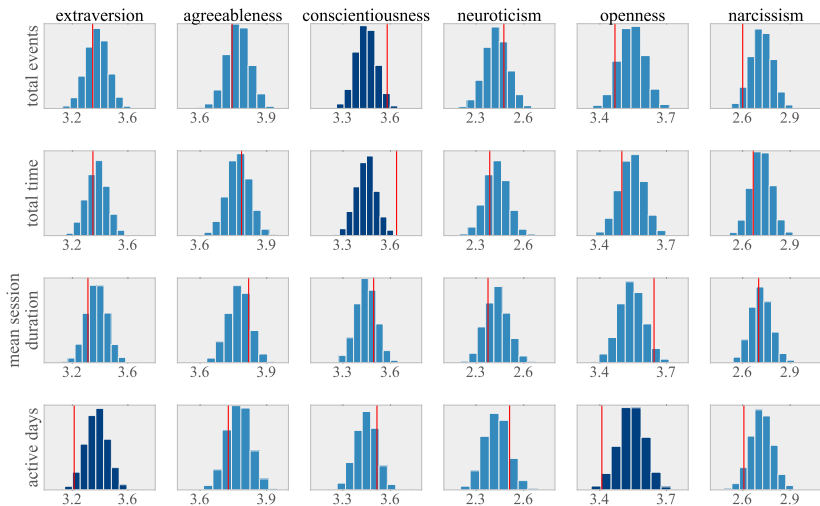
Fig 5. Distributions of the bootstrapped means of the subsamples. The red line is the measured mean of the top 10% quantile for each trait. The cases where the mean of top 10% are more extreme than 95% of the bootstrapped samples are highlighted in dark blue.

- We calculate the mean of the trait for the top 10% according to the usage feature

- We bootstrap $n = 5000$ subsamples from the initial population and obtain their means

- We compare the distribution of the means of the subsamples and the mean of the top 10%

Fig. 5 depicts the bootstrapping distributions of each personality trait and usage feature. The red line indicates the measured mean of the top 10%. We note that, as before, conscientiousness with total time have the most visible difference from the bootstrapped means, but also some for other feature-trait pairs (highlighted in dark blue) there are quite large differences.

Discussion

A key area of interest in this work is to understand the reasons for usage and non-usage of the SensibleJournal 2014 app as a Personal Informatics self-tracking system. Our population consists of well-educated, tech-savvy young adults, a fact that may introduce some biases. However, we suggest that this population would be possible adopters of self-tracking tools.

The total usage varies greatly from participant to participant. About 30% of the participants did not use the app at all and only a small fraction used it every few days. Since significant amount of time and resources are spent in developing such Personal Informatics systems, this inability to engage the potential users can be a concern for any such system, either within a research project in academia, or a product in commercial settings. A possible factor for the limited uptake of the app is the fact that the participants might receive similar knowledge from other commercial and more visually polished apps or services, thus resulting in rapid loss of interest.

The interest was higher in the initial phase, and then declined as time progressed. The decrease over time may have multiple explanations, such as the repetitive nature of the feedback. As many respondents reported, the app tends to report similar information over time, that is time spent usually at the same few places, like home and work. Therefore, users may not be able to learn anything new and would eventually abandon the app. The repetitive feedback information is due to the inherent nature of human mobility, which tends to be habitual and predictable [26]. Habitual living suggests that once initial insights have been obtained, limited new significant knowledge can be gained from the data itself. This problem may potentially affect many Personal Informatics systems measuring periodic behavior such as fitness activity, heart rates and sleep patterns. Any app reporting about the status of the user will soon produce repetitive feedback and may risk to become uninteresting for the user.

One possible solution is to generate feedback only when new or deviating information is available, such as something that has not happened before, something that is different or something requiring user attention. A goal-setting feature could also stimulate the interest in self-reflection and facilitate the process of behavior change.

One hypothesis was that personality traits are a factor in the adoption of Quantified Self tools. We find, however, that narcissism makes no significant difference in respect to adoption, in contrast to conscientiousness, which is the only trait making a statistically significant difference. The present data is insufficient for a full understanding of the casual relationship, but we hypothesize that the organization and self-discipline characteristics of conscientiousness could be an important driver for the usage of such self-tracking apps.

Conclusions

We have presented results from an empirical study on the usage of a mobility self-tracking smartphone app, among N=796 participants in an ongoing mobile sensing study. The app was offered to all participants, but whether they would use it or not was left optional and not enforced in any way.

A relatively low uptake of the app was observed, as 30% of the participants never used the app. Only 16 participants (2%) ended up using the app on a regular basis and among that group, a decline in usage over time was observed. A questionnaire on the participants' interests and attitudes towards the self-tracking app provided indications that the recurring data patterns lead to a drop in interest, once initial insights on personal mobility had been obtained. The fact that the app does not provide radically different information over time, or suggestions based on the data obtained, is a possible

reason for the participants' gradual loss of interest in the app, as well as in their personal mobility data.

Personality traits of the $N=796$ participants were corresponding to those of the general population. In order to understand potential differences between those who adopted the self-tracking behavior by using the app and those who decided not to use it (or only use it short-term), we compared their personality traits. Through this comparison, we found a relation between self-tracking and conscientiousness, an observation that is in contrast with the view in popular media, which suggest a tendency towards narcissism among people that adopt self-tracking behavior.

Acknowledgments

References

1. Froehlich JE, Larsen JE, Kay M, Thomaz E. Disasters in Personal Informatics: The Unpublished Stories of Failure and Lessons Learned. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. ACM; 2014. p. 673–678.

2. Li I, Froehlich J, Larsen JE, Grevet C, Ramirez E. Personal Informatics in The Wild: Hacking Habits for Health & Happiness. In: CHI'13 Extended Abstracts on Human Factors in Computing Systems. ACM; 2013. p. 3179–3182.

3. Li I, Medynskiy Y, Froehlich J, Larsen J. Personal Informatics in Practice: Improving Quality of Life Through Data. In: CHI'12 Extended Abstracts on Human Factors in Computing Systems. ACM; 2012. p. 2799–2802.

4. Ruckenstein M. Visualized and Interacted Life: Personal Analytics and Engagements with Data Doubles. Societies. 2014;4(1):68–84.

5. Bode M, Kristensen DB. The Digital Doppelgänger Within. A Study on Self-tracking and The Quantified Self Movement. In: Domen Bajde and Robin Canniford Assembling Consumption. Oxford/ New York: Routledge; 2015.

6. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, et al. Measuring Large-Scale Social Networks with High Resolution. PLoS ONE. 2014;9(4):e95978. doi:10.1371/journal.pone.0095978.

7. Lupton D. Self-tracking Cultures: Towards a Sociology of Personal Informatics. In: Proceedings of the OzCHI conference. ACM; 2014.

8. Li I, Dey A, Forlizzi J. A Stage-Based Model of Personal Informatics Systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2010. p. 557–566.

9. Larsen JE, Cuttone A, Lehmann S. QS Spiral: Visualizing Periodic Quantified Self Data. In: Proceedings of the CHI 13 Personal Informatics in the Wild: Hacking Habits for Health & Happiness; 2013.

10. Cuttone A, Lehmann S, Larsen JE. A Mobile Personal Informatics System for Visualizing Mobility and Social Interactions. In: Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia. ACM; 2013. p. 27–30.

11. Ryoo MS, Lee JY, Joung JH, Choi S, Yu W. Personal Driving Diary: Constructing a Video Archive of Everyday Driving Events. In: Proceedings of IEEE Workshop on Applications of Computer Vision (WACV); 2011. p. 628–633.

12. Kalnikaite V, Sellen A, Whittaker S, Kirk D. Now Let Me See Where I Was: Understanding How Lifelogs Mediate Memory. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2010. p. 1–10.

13. Wolf G. The Data-driven Life; 2010. Available from: `http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html`.

14. Barcena MB. How Safe is Your Quantified Self?; 2014. Available from: `https://www.symantec.com/content/dam/symantec/docs/white-papers/how-safe-is-your-quantified-self-en.pdf`.

15. Lupton D. Self-tracking Modes: Reflexive Self-monitoring and Data Practices. Available at SSRN 2483549. 2014;.

16. Lupton D. You Are Your Data: Self-tracking Practices and Concepts of Data. Available at SSRN. 2014;.

17. Waltz E. How I Quantified Myself. Spectrum, IEEE. 2012;49(9):42–47.

18. Ames DR, Rose P, Anderson CP. The NPI-16 as a Short Measure of Narcissism. Journal of Research in Personality. 2006;40(4):440–450.

19. Wolf G. Are Self-Trackers Narcissists?; 2009. Available from: `http://quantifiedself.com/2009/02/are-self-trackers-narcissists/`.

20. John OP, Srivastava S. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. Handbook of personality: Theory and research. 1999;2(1999):102–138.

21. Back MD, Küfner AC, Dufner M, Gerlach TM, Rauthmann JF, Denissen JJ. Narcissistic Admiration and Rivalry: Disentangling the Bright and Dark Sides of Narcissism. Journal of Personality and Social Psychology. 2013;105(6):1013.

22. Cuttone A, Lehmann S, Larsen JE. Inferring Human Mobility from Sparse Low Accuracy Mobile Sensing Data. In: 3rd ACM Workshop on Mobile Systems for Computational Social Science(MCSS 2014). ACM; 2014.

23. Cuttone A, Larsen JE. The Long Tail Issue in Large Scale Deployment of Personal Informatics. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. ACM; 2014. p. 691–694.

24. Bentley F, Tollmar K. The Power of Mobile Notifications to Increase Wellbeing Logging Behavior. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2013. p. 1095–1098.

25. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics. 1979; p. 65–70.

26. Song C, Qu Z, Blumm N, Barabási AL. Limits of Predictability in Human Mobility. Science. 2010;327(5968):1018–1021.

Four Data Visualization Heuristics to Facilitate Reflection in Personal Informatics

Andrea Cuttone, Michael Kai Petersen, Jakob Eg Larsen.

Four Data Visualization Heuristics to Facilitate Reflection in Personal Informatics

Andrea Cuttone[1], Michael Kai Petersen[1], and Jakob Eg Larsen[1]

Technical University of Denmark (DTU)
Department of Applied Mathematics and Computer Science
Cognitive Systems Section
Matematiktorvet 303B, 2800 Kgs Lyngby, Denmark
`ancu,mkai,jaeg@dtu.dk`

Abstract. In this paper we discuss how to facilitate the process of reflection in Personal Informatics and Quantified Self systems through interactive data visualizations. Four heuristics for the design and evaluation of such systems have been identified through analysis of self-tracking devices and apps. Dashboard interface paradigms in specific self-tracking devices (Fitbit and Basis) are discussed as representative examples of state of the art in feedback and reflection support. By relating to existing work in other domains, such as event related representation of time series multivariate data in financial analytics, it is discussed how the heuristics could guide designs that would further facilitate reflection in self-tracking personal informatics systems.

Keywords: personal informatics, quantified self, self-tracking, information visualization, feedback, reflection, heuristics

1   Introduction

In recent years self-tracking and lifelogging have received increased interest with the introduction of a wide variety of low-cost mobile apps, wearable computers, and sensors. These devices allow easy collection of data that can describe various aspects of human behavior. However, making sense of the ever increasing amounts of everyday self-tracking data retrieved across multiple domains create new demands for turning data points and trends into affordances for action. The reflection stage is a fundamental component in modeling and using Personal Informatics (PI) systems [18, 8, 16] to facilitate an understanding of self-tracking data reflecting daily habitual patterns and to make such data actionable for behavioral change [9]. Different solutions have been suggested to facilitate self-reflection, including the usage of charts [18, 20, 21, 6], avatars [22, 12, 14], notifications [1], narrative [23] and abstract art [7, 10]. Although these solutions may facilitate increased awareness due to the fact that behavioural aspects are being observed, the process of turning observations and insights into actions remains a challenge. Even a recent review of activity trackers in *The*

New York Times[1] emphasized that while self-tracking devices enable the user to collect behavioral data, they fall short of assisting the user in learning how to change habits.

One element of personal informatics is the iterative process with self-reflection questions phrased by a user and feedback provided by a self-tracking system to answer those questions. However, we suggest that state-of-the art systems offer fairly limited flexibility in terms of the types of questions that can be phrased, and the possible feedback that can be provided. In a broader perspective, the self-tracking data obtained might be characterized as quantitative time series data which combines behavioral data with associated discrete events. Similar to how financial analytics like those provided by Bloomberg might combine a vertical flow of business related earning reports or corporate news updates, with distinct time stamped markers, outlined horizontally within the continuous timeline fluctuations of stock values. Thus hierarchically adding layers of relevant information embedded both within the chart and in adjacent panels linked to external events, that may facilitate interpretation or be a direct cause of rising or falling trends visualized in the data [26]. The emphasis of integrating distinct interpretable events in continuous flows of quantitative data within financial analytics reflects a need for these interfaces to provide a foundation for taking concrete action related to aspects of optimizing profits or avoiding a loss. We suggest that these advances within financial analytics software for interpretation of complex data may both provide underlying design patterns for long term comparison of trends in multivariate flows, as well as defining thresholds which could likewise turn quantified self generated data into actionable parameters for optimizing lifestyle in personal informatics systems.

2   Related Work

Several frameworks have been proposed to formalize the reflection process in personal informatics. Li et al. [19] identify six kinds of questions for reflection: *Status* (what is my situation now?), *History* (what was my situation in the past?), *Goals* (what future status should I aim for?), *Discrepancies* (how does my status compare with my goals?), *Context* (what affects my status?), *Factors* (how are different attributes related?). Moreover, two alternating phases are defined: *Maintenance* (known relation between status and behavior) and *Discovery* (not known goals or effect of behavior). Fleck et al. [8] define a multi-layer reflection framework: *Description*, *Reflective Description*, *Dialogic Reflection*, *Transformative Reflection*, *Critical Reflection*. Each layer builds on top of the previous, and corresponds to a deeper understanding of personal data. Guidelines for facilitating reflection are proposed, including supporting questions and providing multiple perspectives on the data. Rivera et al. [24] apply Boud's reflective learning framework to personal informatics, and identifies two levels

---

[1] http://www.nytimes.com/2014/01/30/technology/personaltech/review-the-fitbit-force-activity-tracker.html Last accessed Jan 29, 2014

of reflection: *Triggering* (active notification or passive feedback) and *Recalling* (aggregating, contextualizing, visualizing).

Several feedback schemes have been suggested including avatar-based feedback that employs a virtual object to represent a judgment on behavior. These solutions exploit participants' empathy with the virtual avatars to persuade them in adopting positive behavioral changes. For example, *Fish'n'Steps* [22] provides feedback about daily step count as a virtual fish, *Ubigreen* [12] using virtual trees and polar bears to provide feedback on green transportation habits, and *UbiFit Garden* [5] represents fitness activity as a virtual garden. *Spark* [7] visualizes physical activity as abstract art through an ambient display, whereas *Lifestyle Stories* [23] provides feedback about mobile sensing personal data in form of stories composed by events of various categories. Many commercial self-tracker systems employ a combination of traditional charts, maps and dashboards (see for example *Nike+*[2], *Fitbit*[3], *Basis*[4], *Jawbone UP*[5], *Mint*[6], *DailyBurn*[7], *Moves*[8]).

3   Reflection as Data Analysis

We suggest to treat reflection as data analysis on personal information. What are the crucial questions or answers that analytics should provide? In Bloomberg financial analytics the ability to couple external events to timeline charts appears crucial for interpreting the causality behind the data. One might in a more general context consider online news media like *The Wall Street Journal* or Twitter feeds an expanded version of this paradigm, providing not only the current market data but also a highly curated selection of background material as well as live updates on events, that would provide the necessary foundation for making informed decisions. Even in admirably simple single sensor quantified self apps like Fitbit, limited to measuring the number of steps taken during the day or week, data might provide valuable insights into user behavior. But it would require that annotations are added automatically with calendar events or smartphone location data, thereby enriching the representation beyond the current ability of manually attaching labels. We see a similar potential for advanced self-tracking devices with multiple sensors like Basis, which likewise translates complex patterns of behaviors into singular goal oriented habits to be fulfilled on a daily basis at regular hours. Coupling calendar events for monitoring heart rate related to specific physical tasks, indicating how this sensor data is correlated to differences in sleep patterns, or influenced by levels of exercise across weeks, might provide additional value.

---

[2] http://nikeplus.nike.com/plus/
[3] http://www.fitbit.com/
[4] http://www.mybasis.com/
[5] http://jawbone.com/up
[6] http://www.mint.com/
[7] http://dailyburn.com/
[8] http://www.moves-app.com/

A user may want to retrieve specific information from his own dataset (current status, progress), or explore it for finding interesting patterns. In order to facilitate this analysis, we propose to use data visualization, that is the representation of data using position, size, shape, color, and text [3]. Visualization facilitates analysis by exploiting the human visual system, which is extremely good at processing large quantities of information and spotting patterns. Visualization is widely used in Exploratory Data Analysis [28], a statistical technique for exploring datasets, in order to gain insights, obtain a better understanding, spot patterns, trends, correlation, and outliers. In many cases, the data analyst does not know in advance which specific question to ask, so he can explore the data in order to find interesting patterns. This process is highly iterative, as once a question has been posed, its answer often leads to more questions to be asked. Similarly, in PI systems the user may not know which questions to ask, or may not be interested in a specific question but in exploring his own data for curiosity. Indeed, one of the barriers for reflection is not knowing which questions to ask to personal data [18]. We can represent this iterative process as a cycle between questions asked, and feedback provided. We define *question space* the set of all possible questions, and *feedback space* the set of all visualizations. Each feedback type can answer one or more questions, and each question can be answered by one or more type. In data analysis, there are a number of common questions that can be asked: distribution of values (mostly around a central value and gradually less on the sides? Mostly for a value and very rapidly decaying? Multiple peaks?), grouping and outliers (are there values much different than most of the others? Are items clustered into groups?), correlation (what is the relation between x and y? Is there a linear, quadratic, exponential, sinusoidal trend?), geographical (how are values related to locations? Are there locations with similar values?), connectivity (are there items related together? Are there items which are more tightly connected? Are there non-connected items?) We identify the most relevant questions for the goal of self-reflection, and we summarize them into heuristics.

4 Design Heuristics

In this section we introduce four design heuristics that can be applied as a guideline for creating and evaluating interactive visualizations of self-tracking data with the aim to facilitate effective exploration of personal data and make such data actionable for behavior change. Throughout the discussion of the heuristics we relate to existing state of the art self-tracking systems using Fitbit and Basis as examples of personal informatics systems with interactive visualizations. We do not intend to criticize these two systems in particular, but rather consider them as representative and illustrative examples of state of the art in personal self-tracking systems. The scope of the discussion is limited to facilitating the reflection process, while acknowledging that further discussion is needed in terms of providing actionable items as well as other aspects of reflection in personal informatics as mentioned in the Related Work Section.

4.1 Make Data Interpretable at a Glance

Often users want to obtain answers to a question with the minimal effort and time. For this reason, data visualizations optimized for interpretation at a glance are needed, in order to provide a swift overview of personal tracking activities, and to augment and support subjective recollection. Quantified self apps like Fitbit or Basis may aim to simplify visualization of complex patterns by transforming the collected measurements into single activity dashboard dials or progress bars reflecting goal oriented accomplishments. Figure 1 shows the personal dashboard provided by Fitbit, which reduces the collected data to simple indications of (daily) goal fulfillment (percentage) and an overview of daily activity levels. Although it provides an overview of the level of goal fulfillment this division of activities into separate silos makes it a challenge to interpret the data in a larger context. In contrast financial analytics interfaces like those provided by Bloomberg [26] may contain large amounts of data which is nevertheless made interpretable based on established conventions for using dynamically changing font colors to signify up- or downward moving prices, or positive negative outlooks based on earning reports, which when collapsed form independent parallel layers of color coded trend lines that remain surprisingly legible on top of contrasting neutral background screens.

Fig. 1. The Fitbit dashboard show daily goal fulfillment (percentage) and an overview of daily activity

While data visualizations can be very useful they may be complex and difficult to interpret and understand. The complexity of visualizations can range from data-poor, informal infographics for the general audience to complex, rigorous scientific visualization aimed at scholars. Several issues should be considered when designing a visualization, including the technical and domain knowledge of the users, the goals of reflection (exploring data, asking specific questions, testing hypothesis), the time and effort expected from the user (a quick glance or a long interaction?). We suggest to provide a simple visualization as the starting

point, and allow advanced users to dynamically increase the level of complexity and details. The usage of explanatory elements, such as text, axes, legends, annotations can greatly facilitate the comprehension. Visualizing data is often compared to storytelling, especially in the fields of journalism and business reporting. Some authors prefer to present the data as raw as possible, with little or no annotations and highlighting, in order to give the reader full freedom in the interpretation of the facts behind the data. Others prefer to editorialize the data to various degrees, by marking samples, comparing with other distributions, providing comments. We argue that a certain editorialization is good for PI system, as it can act as a persuasive force towards positive behavior. For instance a fitness tracker system that encourages the user to be more active by visualizing and forecasting the positive consequences.

4.2 Enable Exploration of Patterns in Time Series Data

There are two fundamental patterns to analyze: global trends (does the variable increase, decrease or remain constant over a period of time?) and periodic patterns (does the variable value change with a repeating pattern, for example hourly, weekly or yearly?).

Although several self-tracking app interfaces emphasize simplified dashboard representations of accumulated data which limits exploration, the recently added Basis sleep monitoring goes far beyond the previous single modality heatmaps by breaking down total sleep duration into the different phases of rapid eye movement (REM) and deep sleep, thereby making it possible to translate these periodic patterns into quantifiable aspects of mind and body refresh, see Figure 2.

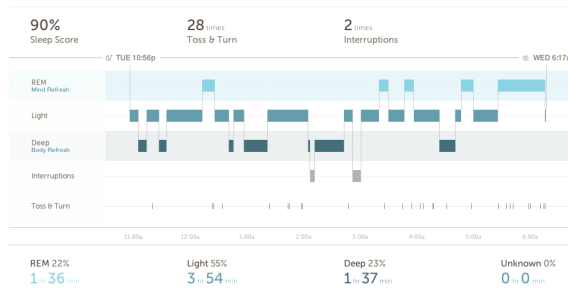**Fig. 2.** The Basis sleep visualization interface also break continuous sleep data into discrete sleep phases including rapid eye movement (REM), light, and deep sleep

Time series analysis is a common task for self-trackers, which may be interested in observing changes over time, periodic patterns, rate of change, and time left to reach goals. The most common representation for time series data

is *line plots*, which allow to easily see the overall change over time of a variable. Due to the unavoidable noise, it is useful to add *trend lines* such a LOWESS [4] or least squares. This enable the support of forecasting on future status if the current behavior is kept or modified, thus a prediction of future values can be visualized [15]. As an example Figure 3 shows the amount over time of money in a bank saving account with a clear trend of increase among the individual month-to-month fluctuations. When the duration of time periods is a factor, *timelines* may be used to represent events in a linear layout to captures the temporal sequence.
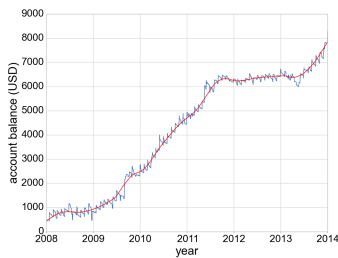
Fig. 3. Amount of money in a bank saving account, visualized as line plot with LOWESS trend line
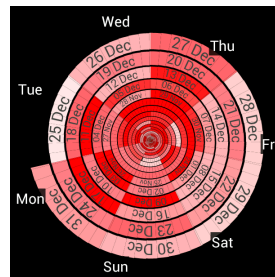
Fig. 4. Number of steps per day represented as a spiral heatmap, with colors from white (low) to red (large)

Line plots and linear timelines do not however facilitate the exploration of periodic patterns, which are characteristic of human behavior. A familiar metaphor for displaying regular patterns is a calendar. A *calendar heatmap* represents each day as a cell, and the variable value as the color shade of the cells. Cells can be aligned for example by day of the week to allow to spot weekly patterns. Spirals have been recently proposed [17] as another representation to facilitate the exploration of periodic patterns in the quantified-self domain. A *spiral heatmap* represents each time unit as an arc in the spiral, and variable value as the color shade of the arcs. By choosing different periods, periodic patterns at various time scale can emerge. Figure 4 show a step count value over time as calendar as spiral heatmaps, with a color scale ranging from white (small count) to red (large count).

4.3 Enable Discovery of Trends in Multiple Data Streams

Financial analytics may also offer inspiration in terms of comparison of key performance indicators in multivariate data. As an example *The Wall Street Journal* allows for extensive personalization when exploring moving averages for smoothing fluctuating trends in time series data, high and low relative to

previous values, weighted blends of values and their variation over time, which may be further customized based on choice of display graphics, adjustable time frames or sensitivity of measures.

Fig. 5. The Basis Activity Details visualization allows the user to explore the relations between multiple biometric time series data (heart rate, steps, calories, skin temperature, perspiration, and activities) in an adjustable time interval

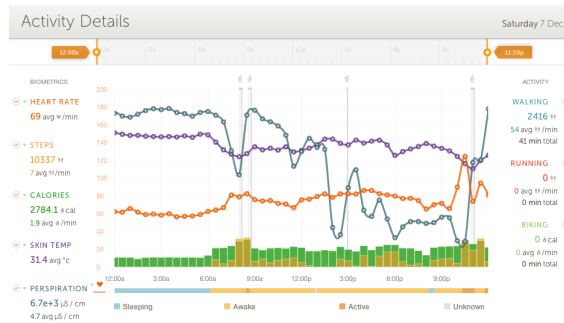The Basis activity details visualization only allows the user limited possibilities of exploring the relations between multiple time series data in an adjustable time interval, as shown in Figure 5. Multivariate analysis is the process of analyzing multiple variables together, in order to find and understand their relation. In the simplest case, two variables x and y are to be compared, and the following relations can exist: direct correlation (y increases as x increases), inverse correlation (y decreases as x increases), or no correlation. For example, fitness trackers may be interested in how the weight loss is affected by exercise and food intake, or productivity trackers may be interested in how coffee intake, sleep patterns, and exercise affect productivity.

The relation between two variable can be visualized as a *scatterplot*, where each variable is represented on one axis. Scatterplots allow to easily spot trends and outliers. If more than two variables are to be compared, a *scatterplot matrix* allows to inspect all possible combinations. A scatterplot matrix is a array of $n \times n$ scatterplots, where the scatterplot S_{ij} displays the relation between variables X_i and X_j. Scatterplot matrices are a very powerful tool, but can also be intimidating for users. A simpler version of multivariate visualization is the *Corrgram* [11] which distinguishes positively and negatively correlation between pairs of variables using color-coding. As a constructed example Figure 6 and 7 show the relation between coffee intake, hours at work, hours of sleep, and steps. The coffee intake appears directly correlated with working hours, and inversely correlated with sleep hours, and no correlation seem to be present with number of steps.
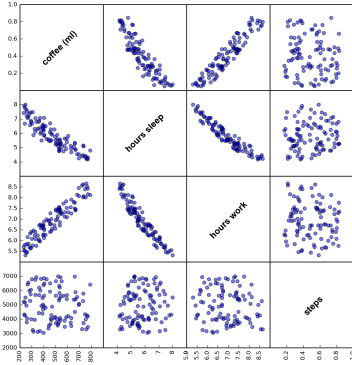
Fig. 6. Scatterplot matrix: all pairs of 4 variables are plotted. The alignment of the data points helps to identify direct, inverse or no correlations.
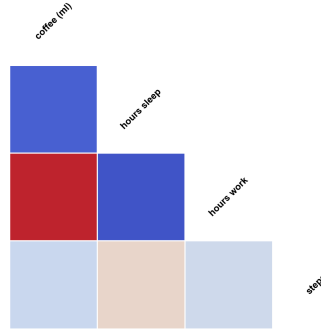
Fig. 7. Corrgram: the correlation between all pair of 4 variables is represented in a color scale from blue (inverse correlation), grey (no correlation) and red (direct correlation)
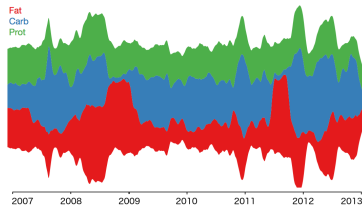
Fig. 8. The composition of dietary intake over time as a streamgraph
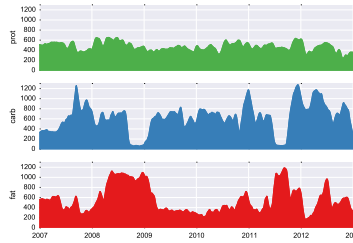
Fig. 9. Small multiples: the three variables of dietary intake are shown side-to-side in separate sub-views

Often users do not need to explore relations of variables between each others, but they are interested in the change of multiple variables over time. A *Streamgraph* [2] can be used when it is important to show the contribution of each variable to a total. Each variable generates a section of different height, and the resulting areas are stacked to form a stream. *Small multiples* [27] allow to display multiple facets of a dataset, often in comparison to time. Each variable is displayed separately in his own subview, and subviews are layed out side-by-side to facilitate comparison. Figures 8 and 9 show the composition of dietary intake over time. The small multiples enable multivariate comparison, and the streamgraph facilitates the understanding of the total caloric intake.

4.4 Turn Key Metrics into Affordances for Action

The emphasis on interaction in data visualization [13] is reflected in a typical analyst workflow including the generation of a data view, exploration of the result, adjustment of parameters or creation of a completely different visualization in order to further explore the data. This process may lead to new insights or the identification of key metrics. As an example, financial portfolio management often requires a specific action in response to events that cause values to get out of bounds due to regulatory terms that trigger alarms for reevaluation. Likewise in currency trading applications one may need to quickly buy or sell when prices transcend previously set values indicating pain or gain thresholds. In a similar fashion we suggest that the self-tracking workflow involves feedback provided by a personal informatics system which may generate insights about personal data. This may lead to phrasing new questions or directly imposing threshold values that proactively trigger responses to be considered, based on general monitoring of health issues known to be of general concern. This iterative process may be one approach to identify key metrics that can be turned into affordances for actions related to changing behavior.

With the complexity of multi channel self-tracking data sets it may have limited utility to try to visualize all data at once. A user may want to slice the data in various ways, such as by time or category. A user may also want to select a specific set of elements that match a given criteria, such as points inside a geographical region, or values between some thresholds. *Filtering* allows to focus on a specific subset of the data. One of the recommended interaction pattern is "Overview first, zoom and filter, then details-on-demand" [25]. *Navigation* may be supported by allowing scroll and zoom views. When focusing on a subset of the data, the context for the current details could hold valuable information to understand behavior.

The long sequence of interactions with a visualization system, such as filtering, zooming, transforming can be recorded in form of *history*. This log helps the user remember the steps he took, navigate in his interaction sequence, undo eventual mistakes and facilitate a trial-and-error exploration. Providing a visual representation of this history (such a timeline or snapshots of the views) can help the user to orient in his own workflow. In the process of reflection, the user may want to document his findings, write down questions to be investigated, add notes to self. To this end, a visualization can support *annotation* with text and sketches.

These interaction techniques can be readily applied in data visualization in personal systems in order to facilitate the reflection process. Imagine a fitness tracking system, where the user is provided with an overview of his activities. He filters the activity log to a specific part of the year. He views his activity both as a timeline of step counter and as a breakdown of his caloric intake and annotated activities.

5 Conclusions

In this paper we have discussed the support of reflection in state of the art personal informatics systems arguing that it is limited in terms of making observations and insights obtained from interactive visualizations of self-tracking data actionable. We have proposed four heuristic principles for the design and evaluation of interactive data visualization feedback that could further facilitate the process of reflection in self-tracking personal informatics systems. Each design heuristic has been discussed on the basis of an analysis of visualization feedback available in state of the art personal informatics systems.

References

1. Bentley, F. and Tollmar, K: The power of mobile notifications to increase wellbeing logging behavior. In Proc. of the SIGCHI Conf on Human Factors in Computing Systems, 1095-1098, 2013
2. Byron, L., & Wattenberg, M.: Stacked graphs-geometry & aesthetics. IEEE Transactions on Visualization and Computer Graphics, 14(6), 1245-1252, 2008
3. Card, Stuart K., Mackinlay, J. D. & Schneiderman, B.: Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, 1999
4. Cleveland, W. S.: Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association, 74(368), 829-836, 1979
5. Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., & Landay, J. A.: Activity sensing in the wild: a field trial of ubifit garden. In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, 1797-1806, 2008
6. Cuttone, A., Lehmann, S. and Larsen, J. E.: A mobile personal informatics system with interactive visualizations of mobility and social interactions In Proc. of the 1st ACM Int. workshop on Personal data meets distributed multimedia, 2013
7. Fan, C., Forlizzi, J., and Dey, A. A.: A spark of activity: Exploring informative art as visualization for physical activity. In Proc. of the 2012 ACM Conf. on Ubiquitous Computing, 81-84, 2012
8. Fleck, R., and Fitzpatrick, G.: Reflecting on reflection: framing a design landscape. In Proc. of the 22nd Conf. of the Computer-Human Interaction SIG of Australia on Computer-Human Interaction 216-223, 2010
9. Fogg, B. J.: Persuasive technology: Using computers to change what we think and do. Ubiquity no. 5, 2002
10. Frick, L.: Experiments in self tracking. http://www.lauriefrick.com/category/work/, Last accessed Dec 13, 2013
11. Friendly, M. Corrgrams: Exploratory displays for correlation matrices. The American Statistician, 56(4), 316-324, 2002
12. Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B., & Landay, J. A.: UbiGreen: Investigating a mobile tool for tracking and supporting green transportation habits. In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems,1043-1052, 2009
13. Heer, J. and Shneiderman, B.: Interactive dynamics for visual analysis. Queue, 10(2), 30, 2012

14. Hekler, E. B., King, A. C., Banerjee, B., Robinson, T., and Alonso, M.: A case study of BSUED: Behavioral Science-informed User Experience Design. In Proc. of the CHI11 Workshop on Personal Informatics, 2011
15. Koeman, L. and Rogers, Y. Enabling Foresight and Reflection: Interactive Simulations to Support Behaviour Change In Proc. of the CHI13 Personal Informatics in the Wild: Hacking Habits for Health & Happiness, 2013
16. Konrad, A., Whittaker, S., and Isaacs, E.: Short and long-term benefits of reflective technologies. In Proc. of the CHI13 Workshop on Personal Informatics, 2013
17. Larsen, J. E., Cuttone, A., and Lehmann, S.: QS Spiral: Visualizing Periodic Quantified Self Data. In Proc. of the CHI 13 Personal Informatics in the Wild: Hacking Habits for Health & Happiness, 2013
18. Li, I., Dey, A., and Forlizzi, J.: A stage-based model of personal informatics systems. In Proc. of the 28th Int. Conf. on Human Factors in Computing Systems, 557-566, 2010
19. Li, I., Dey, A., and Forlizzi, J.: Understanding my data, myself: supporting self-reflection with ubicomp technologies. In Proc. of UbiComp'11, 405-414, 2011.
20. Li, I., Medynskiy, Y., Froehlich, J., and Larsen, J. E.: Personal informatics in practice: improving quality of life through data, 2012. In CHI12 Extended Abstracts on Human Factors in Computing Systems, CHI EA 12, 2799-2802, 2012
21. Li, I., Froehlich, J., Larsen, J. E., Grevet, C., and Ramirez, E.: Personal informatics in the wild: hacking habits for health & happiness. In CHI'13 Extended Abstracts on Human Factors in Computing Systems, 3179-3182, 2013
22. Lin, J. J. et al.: FishnSteps: Encouraging physical activity with an interactive computer game. UbiComp 2006: Ubiquitous Computing, 261-278, 2006
23. Pavel, D., Trossen, D., Holweg, M., and Callaghan, V.: Lifestyle stories: correlating user information through a story-inspired paradigm. In Pervasive Computing Technologies for Healthcare, 412-415, 2013
24. Rivera-Pelayo, V., Zacharias, V., Müller, L. and Braun, S.: A framework for applying quantified self approaches to support reflective learning, Mobile Learning, 2012
25. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In Proc. of IEEE Symposium on Visual Languages, 336-343, 1996
26. Sorenson, E., and Brath, R. Financial Visualization Case Study: Correlating Financial Timeseries and Discrete Events to Support Investment Decisions. IEEE Information Visualisation, 2013
27. Tufte, E. R., & Graves-Morris, P. R.: The Visual Display of Quantitative Information (Vol. 2). Cheshire, CT: Graphics press, 1983
28. Tukey, J. W.: Exploratory data analysis. Reading, Ma, 231, 1977

SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events

Andrea Cuttone, Per Bækgaard, Vedran Sekara, Håkan Jonsson, Jakob Eg Larsen, Sune Lehmann.

SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events

Andrea Cuttone[1,❂,*], Per Bækgaard[1,❂], Vedran Sekara[1,3], Håkan Jonsson[3], Jakob Eg Larsen[1], Sune Lehmann[1,2]

1 DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark
2 The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
3 Sony Mobile, Nya Vattentornet, Mobilvägen, 221 88 Lund, Sweden.

❂These authors contributed equally to this work.
* ancu@dtu.dk

Abstract

We propose a Bayesian model for extracting sleep patterns from smartphone events. Our method is able to identify individuals' daily sleep periods and their evolution over time, and provides an estimation of the probability of sleep and wake transitions. The model is fitted to more than 400 participants from two different datasets, and we verify the results against ground truth from dedicated armband sleep trackers. We show that the model is able to produce reliable sleep estimates with an accuracy of 0.89, both at the individual and at the collective level. Moreover the Bayesian model is able to quantify uncertainty and encode prior knowledge about sleep patterns. Compared with existing smartphone-based systems, our method requires only screen on/off events, and is therefore much less intrusive in terms of privacy and more battery-efficient.

Introduction

Sleep is an important part of life, and quality of sleep has a significant impact on individual well-being and performance. This calls for methods to analyze sleep patterns in large populations, preferably without laborious or invasive consequences, as people typically disapprove of the use of intrusive technologies [1].

Large scale studies of human sleep patterns are typically carried out using questionnaires, a method that is known to be unreliable. It is possible to perform more accurate studies, but these are currently carried out within small controlled environments, such as sleep labs. In order to perform accurate measurements of sleep in large populations—consisting of thousands of individuals—without dramatically increasing costs, alternative methods are needed.

Smartphones have become excellent proxies for studies of human behavior [2,3], as they are able to automatically log data from built-in sensors (GPS, Bluetooth, WiFi) and on usage patterns (phone calls, SMS and screen interaction), from which underlying user behavioral patterns can be derived.

Smartphone data has been used to infer facets of human behavior such as social interactions [4], communication [5], mobility [6], depression [7] and also sleep patterns [8]. Either paired with additional sensors or on their own, mobile app solutions are able – sometimes very ingeniously – to track individual sleep patterns and visualize them. We cite as examples *Smart Alarm Clock* [9], *Sleep Cycle* [10], *SleepBot* [11], and *Sleep as Android* [12].

Using mobile phone data to derive sleep patterns has thus already been demonstrated and verified, and offers advantages (i.e. reduced cost) as an alternative to dedicated sleep monitoring devices. In this paper we suggest extending previous approaches, using a Bayesian model to infer rest and wake periods based on smartphone screen activity information. The advantages of our proposed Bayesian approach *SensibleSleep*, as compared to previous work, are that it:

- is less sensitive to "noisy" data, for instance infrequent phone usage during sleep interruptions (such as checking the phone at night)

- is able to quantify not only specific rest and wake times but also characterize their distributions and thus uncertainty

- can encode specific prior beliefs, for instance on expected rest periods (when desirable)

- can capture complex dependencies between model variables, and possibly even detect and relate patterns that are common to a group of people with diverging individual patterns (when using one of the proposed hierarchical models), such as detecting how available daylight may modulate sleep patterns across an otherwise heterogeneous group of users

Our method, moreover, only needs screen on/off events and is thus *non-intrusive*, *privacy-preserving*, and has *lower battery cost* than microphone or accelerometer based ones.

We start by providing an overview of the related work. We then describe the collected data, and introduce the Bayesian model. We compare the model results with ground truth obtained by sleep trackers, and show how the model is able to infer the sleep patterns with high accuracy. Finally we describe the individual and collective sleep patterns inferred from the data.

Related Work

A key finding by Zhang et al. [13] shows a global prevalence of sleep deprivation in a group of students, partly linked to heavy media usage. In this study sleep patterns are largely deduced from the teachers' perception or based on individual self-reports, lacking more direct measurements.

Corroborating this finding, Orzech et al. [14] report that digital media usage before bedtime is common among university students, and negatively impacts sleep. The findings are based on studies involving self-reports through (online) sleep diaries and digital media surveys, and also lacks more direct measurements of sleep patterns. Additionally, this would make it possible to increase the scale of the experiment and enable the study of larger populations.

Abdullah et al. [8] have previously demonstrated using 9 subjects how a simple rule-based algorithm is able to infer sleep onset, duration and midpoint based on a (filtered) list of screen on-off patterns with the help of previously learned individual corrective terms, and further analyzed behavioral traits of the inferred *circadian rhythm* [15, 16]. The algorithm uses an initial two weeks of data with journal self-reported sleep for learning key corrective terms in order to improve the accuracy and compensate for differences between *actual* sleep and *inferred* nightly rest period. The method has been verified against a daily online sleep journal and results in differences less than 45 minutes of average sleep duration over the entire analysed period. While our proposed Bayesian model, which has been applied to more than 400 users, may be more complex, it increases the robustness and allows us to better quantify

the uncertainties of the inferred resting periods as well as offer the possibility of building more advanced models across heterogeneous groups of users. In particular, our model may better be able to handle short midnight interruptions, which appear to be not uncommon, without any additional filtering.

In contrast to Abdullah et al. using (only) screen on-off events, a fine-grained sleep monitoring by "hearing" and analyzing breathing through the earphone of a smartphone is suggested by Ren et al. [17]. Here six users tested the system over a period of 6 months, demonstrating the feasibility of using smartphones for the purpose of analysing breathing patterns, using a Respiration Monitor Logger as ground truth. Sleep estimates are not directly inferred in this paper, however. This technology is also non-invasive, although it does requires capturing and analyzing large samples of audio data.

iSleep [18] proposes detecting sleep patterns by means of a decision tree model, also based on audio features. The system was evaluated with 7 users for a total of 51 days, and shows high accuracy in detecting snoring and coughing as well as sleep periods, but report drops in performance due to ambient noise.

Increasing the number of features, the *Best Effort Sleep model* [19] is based on a linear combination of phone usage, accelerometer, audio, light, and time features using a self-reporting sleep journal, and subsequently achieved a 42 minutes mean error on 8 subjects in a test period of 7 days.

Other work also tries to estimate sleep quality, for example *Intelligent Sleep Stage Mining Service with Smartphones* [20], which uses Conditional Random Fields on a similar set of features trained on 45 subjects over 2 nights, and reports over 65% accuracy of detection of sleep phases, compared to EEG ground truth on 15 test subjects over 2 nights.

Candy Crushing Your Sleep [21] uses the longest period of phone usage inactivity as heuristic for sleep, with some ad-hoc rules for merging multiple periods, and proceeds to quantify the sleep quality and to identify aspects of daily life that may affect sleep. The inferred sleep period was however not validated against any ground truth.

The *Sleep Well* framework [22] deploys a Bayesian probabilistic change-point detection, in parallel with an unsupervised classification, of features extracted from accelerometer data, in order to identify fine-grained sleep state transitions. It then uses an active learning process to allow users to incrementally label sleep states, improving accuracy over time. It was evaluated both on existing datasets with clinical ground truth, and on 17 users for 8-10 days with user diary data as ground truth, reaching an average sleep stage classification accuracy approaching 79%.

In comparison, even though sleep quality is not estimated, our non-intrusive model only needs screen on/off events and has been tested on a large user-base, and can suitable for very large-scale deployment.

Methods

Data Collection

We have analyzed two datasets in this work.

The first dataset (A) was provided by Sony Mobile, and contains smartphone app launches coupled with sleep tracking data from the SWR10 and SWR30 fitness tracking armbands [23]. For each user we have a set of records containing an anonymized unique user identifier, a timestamp and the unique app package name. Note that the model only uses the app launch timestamp and completely ignores the app identifier, therefore no privacy risks related to app names are present. The sleep tracking data indicates when each user is detected asleep or awake with a granularity of one minute, serving as ground truth that we will compare our results against. From this dataset we select 126

users that have at least 3 hours of tracked sleep per day, and have between 2 and 4 weeks of contiguously tracked sleep.

The second dataset (B) originates from the *SensibleDTU* project [24], which collected smartphone sensor data for more than 800 students at the Technical University of Denmark. In this dataset we focus on the screen interaction sensor that records whenever the smartphone screen is turned on or off, either by user interaction or by notifications. Each record contains a unique user identifier, a timestamp, and the event type (on or off). From this dataset we select 324 users in November 2013 that have at least 10 events per day, thus filtering out users with gaps in the collected data or with very sparse data. There is on average ≈ 76 screen-on activations pr. day pr. user in this period.

Data collection for the SensibleDTU dataset was approved by the Danish Data Protection Agency, and informed consent has been obtained for all study all participants. For the data from Sony, written consent has been obtained from all the participants.

Model Assumptions

The underlying assumptions of the model are (1) that the user is in one of two modes: being *awake* or *sleeping*, and (2) that mobile phone usage differs between the two modes. In particular a user will have many screen interactions when awake, and very few or even no interactions when sleeping.

Sleeping is here considered as an extended *resting* period that typically takes place once every 24 hours at roughly similar times, as governed by the users circadian rhythm and influenced by socio-dynamic structures, during which the owner physically rests and/or sleeps. Resting periods, however, might be interrupted by short periods of activity, such as checking the time on the phone or responding to urgent messages. This behavior leads to two different activity levels, which we label λ_{awake} and λ_{sleep}, one for each mode.

If we can deduce when the switchpoint between the two distributions occur during each 24 hour period, we can also infer the time during which the owner is *resting* for the night, and thereby also the period within which sleeping takes place.

Short of using the more invasive EEG or polysomnographic methods, properly differentiating the resting period and actual sleep is difficult; even sleep diaries may easily contain reporting bias or be somewhat inaccurate. To remove self-reporting bias and to study a larger population we have therefore decided on using a motion-based detector (Sony fitness tracking armbands) as ground truth.

If higher accuracy would be required, applying individual corrective terms (i.e. average sleep/rest time differences) learned from an initial period by more accurate means (polysomnography, external observer or possibly a careful user diary) might be possible, similar to what as demonstrated by Abdullah et al. [8].

Model Structure

Each user is considered independently. We divide time into $24-$hour periods starting at 16:00 and ending at 15:59 on the next calendar day, so that the night period and the expected sleep midpoint is in the middle, for convenience. Each day is divided into $n = 24 * 4 = 96$ time bins of size 15 minutes. We count the number of events that start within each time bin, where an event is an app launch for dataset A and a screen-on for dataset B. Information about the duration of the events is purposely discarded, as phone usage typically takes place in short bursts. This is supported by the median duration of screen events in dataset B, which is ≈ 26.5 seconds.

It is reasonable to assume that the count of events k in each time bin follows a Poisson distribution:

$$P(k) = \text{Poisson}(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

with $\lambda = \lambda_{awake}$ or $\lambda = \lambda_{sleep}$, depending on the mode of the user. It is, furthermore, assumed that the user mode, and consequently the value for λ, is determined by two switchpoint variables t_{sleep} and t_{awake}, both assuming values from 0 to n:

$$\lambda = \begin{cases} \lambda_{sleep} & \text{if } t_{sleep} \leq t < t_{awake} \\ \lambda_{awake} & \text{if } t < t_{sleep} \vee t \geq t_{awake} \end{cases}$$

For simplicity, all models assume that λ_{sleep} is identical for all days of a given user. It can be expected that users have a very low number of screen events during sleep mode, which is encoded in this prior belief:

$$\lambda_{sleep} \sim \text{Exponential}(10^4)$$

Here Exponential represents the exponential distribution:

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The rate parameter is set to a very large value to encode our prior belief that almost no events should happen during the sleep time.

Fig. 1 shows an illustration of the model idea.
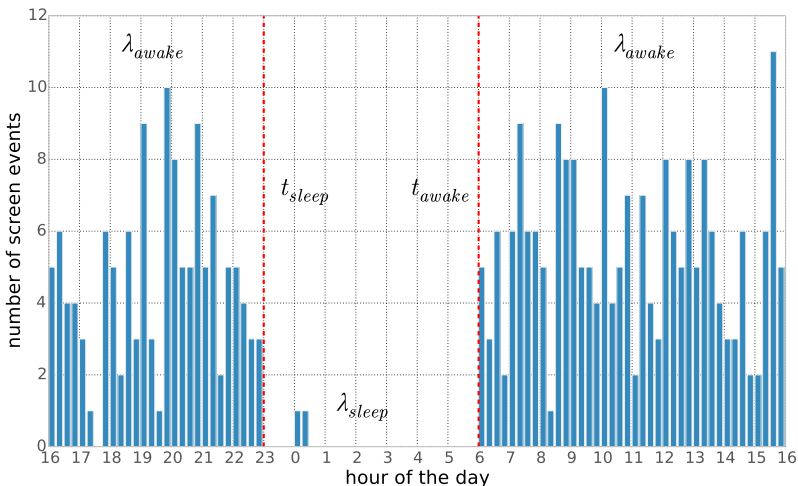
Fig 1. Conceptual illustration of the model. We assume that for each day the event counts follow two different Poisson distributions: one for sleep periods (rate λ_{sleep}) and one for awake periods (rate λ_{awake}). Furthermore we assume that two switchpoints t_{sleep} and t_{awake} determine the rate (i.e. the Poisson distribution) that generates the events.

We now propose four different models, which differ in the assumptions made on the relation of the rate and sleep/awake time parameters for different days.

Pooled-Pooled Model: Pooled Times and Rates

The simplest model assumes that for a given user there is a single λ_{awake}; i.e. the user has very similar phone interaction patterns each day. Also t_{sleep} and t_{awake} are each identical for all days, that is: the user goes to sleep, and wakes up, at the same times each day:

$$t_{sleep} \sim \text{DiscreteUniform}(0, n)$$
$$t_{wake} \sim \text{DiscreteUniform}(0, n)$$
$$\lambda_{awake} \sim \text{Gamma}(2.5, 1)$$

Here $\text{DiscreteUniform}(0, n)$ represents a uniform probability to choose a timebin between 0 and $n = 96$. No additional prior knowledge of t_{sleep} and t_{awake} is assumed; there is equal probability of any bin value. In other words, sleep and awake time are equally probable at any time of the day. The prior for λ_{awake} is chosen to represent our prior belief of a reasonable rate of events, specifically with both mean and variance = 2.5 (events/bin) and a longer tail than a normal distribution.

Independent-Pooled Model: Independent Times

A somewhat more realistic model would assume that each day has independent t_{sleep} and t_{awake} times, while still sharing λ_{awake} rates. Therefore in this model there are t_{sleep}^{i} and t_{awake}^{i}, with $i = 1...m$, one for the each of the considered days:

$$t_{sleep}^{i} \sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m$$
$$t_{wake}^{i} \sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m$$
$$\lambda_{awake} \sim \text{Gamma}(2.5, 1)$$

The rest of the model remains as above.

Independent-Independent Model: Independent Times and Rates

It may further be assumed that each day could have its own specific activity rate. We modeled this as separate λ_{awake}^{i} for each of the m days, in addition to t_{sleep} and t_{awake} for each of the m days:

$$t_{sleep}^{i} \sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m$$
$$t_{wake}^{i} \sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m$$
$$\lambda_{awake}^{i} \sim \text{Gamma}(2.5, 1) \text{ for } i = 1...m$$

Independent-Hyper Model: Hierarchical Rates

The assumption that each day's interaction rate is completely independent may not be correct. It may not be unreasonable to imagine that the daily rate(s) arise from an underlying user-specific rate; i.e. the user may have certain habits that varies from day to day but share some similarities specific to that user. This is modeled by adding α_{λ} and β_{λ} hyperparameters to the Gamma priors for λ_{awake}^{i}:

$$t^i_{sleep} \sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m$$
$$t^i_{wake} \sim \text{DiscreteUniform}(0, n) \text{ for } i = 1...m$$
$$\alpha_\lambda \sim \text{Exponential}(1)$$
$$\beta_\lambda \sim \text{Exponential}(1)$$
$$\lambda^i_{awake} \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda) \text{ for } i = 1...m$$

We do not have strong prior beliefs for α and β, so we set their prior distributions to generic exponential distribution with rate parameter $= 1$, Exponential(1).

Hyper-Hyper Model: Hierarchical Times and Rates

Finally we could assume that each day's sleep and awake times derive from an underlying circadian rhythm that is specific to the user, but still modulated by events that take place during the week. This can be modeled by changing the t^i_{sleep} and t^i_{awake} priors to a normal distribution, with hyperparameters α_t, β_t and τ_t as follows:

$$\alpha_t \sim \text{Exponential}(1)$$
$$\beta_t \sim \text{Exponential}(1)$$
$$\tau_t \sim \text{Gamma}(\alpha_t, \beta_t)$$
$$t^i_{sleep} \sim \text{Normal}(8 * (n/24), \tau_t) \text{ for } i = 1...m$$
$$t^i_{wake} \sim \text{Normal}(15 * (n/24), \tau_t) \text{ for } i = 1...m$$
$$\alpha_\lambda \sim \text{Exponential}(1)$$
$$\beta_\lambda \sim \text{Exponential}(1)$$
$$\lambda^i_{awake} \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda) \text{ for } i = 1...m$$

The t^i_{sleep} are here chosen to be centered at the bin corresponding to 23:00, while the t^i_{awake} are centered at the bin corresponding to 07:00. Also in this case we have no strong prior knowledge of the τ_t, α_t and β_t parameters, so we set their prior distribution to a non-informative Exponential and Gamma respectively.

Model Fitting and Selection

The models are fitted using Markov Chain Monte Carlo (MCMC) sampling [25], where the parameter values are estimated by a random walk in the parameter space guided by the log likelihood. We use the *pymc3* python library [26, 27] for running the sampling, but any MCMC framework could be used to implement our model. The result of the Bayesian inference is a trace that captures the most probable values of the parameters, and also gives an indication of the uncertainty of the estimation.

It is important to note that the models are unsupervised, which means that they are fitted only to the number of events without having access to the ground truth of the actual sleep patterns. This allows the model to be fit to other datasets where we do not have ground truth of sleep patterns, which is desirable if the sleep inference has to be deployed on a large scale. For dataset A we verify the fit by comparing with the sleep patterns from sleep trackers, while for dataset B we evaluate the fit by inspecting the inferred sleep patterns.

In order to find the model that provides the best overall fit for the intended purpose without introducing too many degrees of freedom, we compare the log posterior from the traces of the models, logp, and see how they converge.
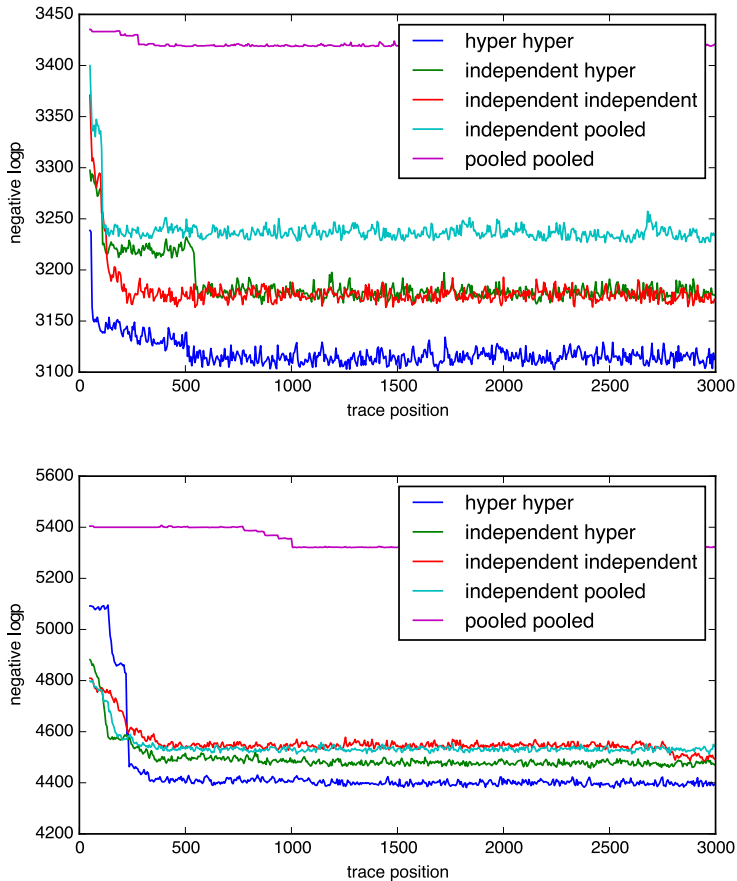
Fig 2. Typical logp traces (A top, B bottom)

One example of a plot of logp traces for the five models is shown in Fig. 2, which shows that the hyper-hyper model (blue) has the highest (least negative) logp, followed by the independent-hyper model for dataset B. The three other models appear with lower logp. In 76% of the analyzed cases of dataset A (84% for dataset B), the hyper-hyper model has the highest logp score, followed by the independent-hyper model with the highest logp in 11% (13%) of the cases.

The logp estimation does not, however, take into account the added complexity of the more advanced models. An attempt to do so is the Deviance Information Criterion (DIC) [28], which penalizes the increased degrees of freedom (more model parameters) that usually result in a model that is easier to fit to the data. Fig. 3 shows the Relative DIC score (vs. the simplest model, pooled-pooled). The order is identical for both datasets.

Further, Table 1 compares the 5 models by ranking the calculated DIC for all 126 and 324 users. The *median* rank shows that the hyper-hyper model is the "best" model;
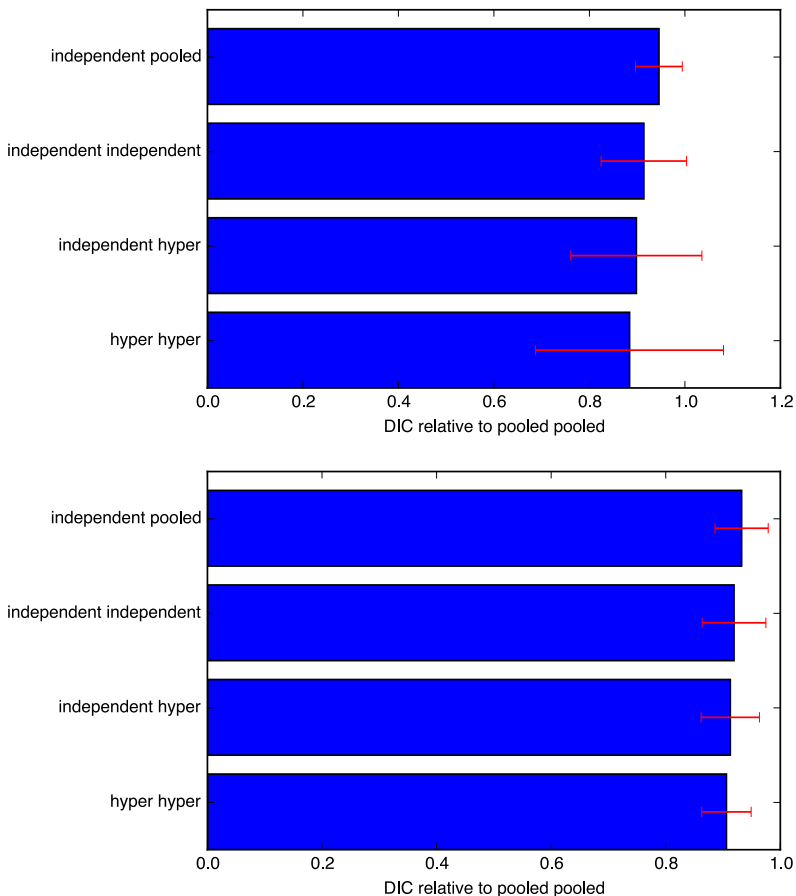
Fig 3. Relative DIC scores (A top, B bottom), sorted by their mean value (error bars represent one standard deviation). For both datasets the order is the same, with the hyper-hyper model having the lowest mean DIC.

it has a probability of being the best ranked model ($p(Best)$) in 62% of the cases for dataset A (69% for dataset B). The independent-hyper model follows as a somewhat distant 2nd best, ranking highest in 17% (19%) of the cases.

It should be noted that, in addition to their different abilities to reflect the underlying assumptions and provide varying levels of fit to the actual data, the models also differ in their runtime; the most complex model typically takes 15 times longer to execute than the simplest. In particular, the hyper-hyper model on average had a runtime that is 60% longer than the independent-hyper model, so there may be cases where the latter would be a better model to use despite the slightly worse DIC ranking.

9

	Model Ranks	Median	Mean		p(Best)	Mean Relative DIC	
			Value	(StdDev)		Value	(StdDev)
A	pooled-pooled	5	4.27	(1.37)	0.10	0.96	(0.16)
	independent-pooled	4	3.82	(0.85)	0.03	0.95	(0.05)
	independent-independent	3	2.86	(1.08)	0.08	0.91	(0.09)
	independent-hyper	2	2.29	(0.83)	0.17	0.90	(0.14)
	hyper-hyper	1	1.76	(1.11)	0.62	0.88	(0.20)
B	pooled-pooled	5	4.70	(0.89)	0.02	0.99	(0.01)
	independent-pooled	4	3.75	(0.66)	0.02	0.93	(0.05)
	independent-independent	3	2.92	(1.02)	0.09	0.92	(0.06)
	independent-hyper	2	2.06	(0.69)	0.19	0.91	(0.05)
	hyper-hyper	1	1.56	(0.94)	0.69	0.91	(0.04)

Table 1. Model DIC comparisons

Results

All five models have been run on both datasets, producing an estimation of the times of sleep and wake up for each day, as well as estimates for the other hyperparameters, for each user. Moreover, we calculated logp and DIC as discussed in the previous section. We firstly verify the accuracy our method using the ground truth from the sleep trackers. We then provide a qualitative analysis of some key examples of individual sleep patterns, and a description of the aggregated sleep patterns for both datasets. For the remainder of the paper we restrict our analysis to the model with the best fit, the hyper-hyper model.

Comparison to Related Work and to Ground Truth

To assess the results, we compare the sleep periods inferred by our model and those inferred by a previously suggested rule-based method to the ground truth collected by the Sony sleep trackers.

For each day we calculate the time of sleep and time of awake inferred by our model as the mean of the t_{sleep}^i and t_{wake}^i respectively, and we consider the user asleep ($Z = 1$) for all time bins between t_{sleep}^i and t_{wake}^i, and awake ($Z = 0$) for the remaining bins.

For a representative and comparable method, we chose to implement a rule-based algorithm similar to what is proposed by Abdullah et. al. [8] to derive sleep data for dataset A. This rule-based method essentially works by finding the longest contiguous sleep period, with a prior assumption that sleep must start after 10 PM and before 7 AM next morning. Note that the original algorithm is based on screen on-off events and furthermore discards events of short duration during the night; in our case we use app launches with no available duration, and thus cannot discard events of short duration.

For the sleep trackers we can directly mark each time bin as sleep ($Z = 1$) if the trackers have detected at least one sleep status in that bin, and awake ($Z = 0$) otherwise.

We again consider one user at a time. For each user we now have three binary matrices: two inferred sleep status values per time bin from either model, and one measured sleep status value per time bin (ground truth) . We evaluate this as two binary classification problems, and calculate accuracy, precision, recall and F1 for each model and for each user according to the definitions:

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{predictions}}$$

$$\text{precision} = \frac{\text{true positives}}{\text{predicted positives}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{all positives}}$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Fig. 4 shows the resulting distribution of accuracy, precision, recall and F1 scores for the proposed method. The SensibleSleep method achieves a mean accuracy of 0.89, and a mean F1 score of 0.83. The below-average scores for some users are expected, since it is likely that among the large population under study there will be people having irregular sleep schedule or noisy sleep ground truth.
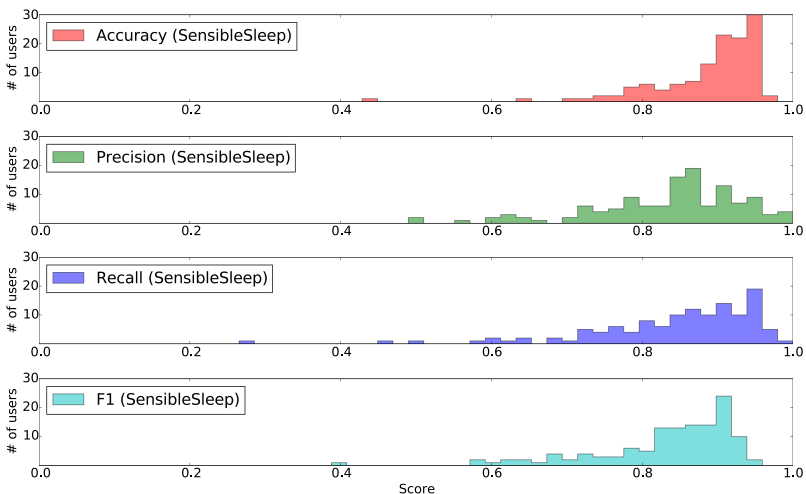

Fig 4. Histogram of the calculated accuracy, precision, recall and F1 score for users in dataset A, comparing the proposed method to the sleep tracker ground truth.

Fig. 5 shows the corresponding *complementary cumulative distributions* of the accuracy, precision, recall and F1 scores of the proposed SensibleSleep model vs that of the rule-based model [8]. The results are generally comparable between the two models, on this particular dataset. Our model has slightly better accuracy and precision whereas the previously suggested rule-based model has a slightly better recall. The F1 scores, which weights precision and recall equally, are comparable. This particular dataset has only very limited sleep interruptions during the night. For populations with more interrupted sleep, we expect our model to maintain a high score.

Individual Sleep Patterns

We now analyze individual sleep patterns to show the results of the model in details. For each user we create a visualization of sleep schedules. We call this the *sleep matrix*. Each row represents one day, and each column represents one time bin. The blue color shows the probability that sleep takes place within the interval; the darker the color the higher the probability. The red dots show activity count per bin; the larger the radius the more events are registered within that particular bin. This compact representation is able to capture at a glance the sleep patterns of individuals over time. We have
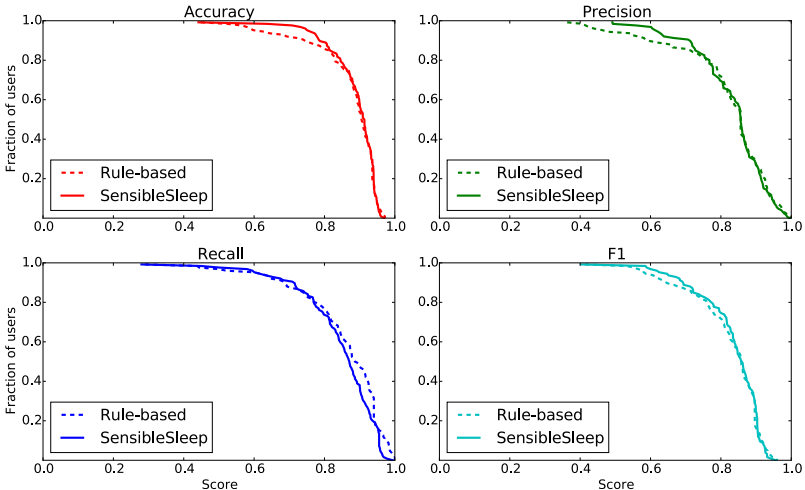
Fig 5. Complementary cumulative distribution of accuracy, precision, recall and F1 scores for users in dataset A, comparing the proposed model (solid line) to the rule-based model (dashed line), showing the proportion of users (y-axis) having a score less than or equal to a specific value (x-axis).

created one such sleep matrix for each of the users, which allows us to inspect hundreds of sleep patterns quickly. Large individual variability both in sleep schedules (regular, irregular) and in phone activity (low, high, during day or night) are noticeable. Still, in most cases it is evident that the model is able to capture a reasonable sleep period, even if it may have been somewhat interrupted.

Let us consider the inferred sleep patterns for two example users in Fig. 6. The top user has a pretty regular schedule, waking up around 5:30 except every few days, when he/she wakes up later – presumably due to vacation or weekends. Notice the light blue sections that indicate how the model is less confident about the probability of sleep due to events that do not follow the usual patterns. The bottom user instead has a much more unstable app usage, therefore the model infers a correspondingly more unstable sleep schedule. The bottom user has also some events in the middle of the night throughout many days (which is presumably checking the phone at night) yet the model is still able to correctly infer this being a sleep phase. Finally notice how the two users have significantly different intensity of app usage (the bottom one uses the phone much more than the top one), yet this is not a problem since the model learns individual activity rates.

Aggregated Sleep Schedules

In this section we also quantify the aggregated sleep patterns. From the posterior probability distribution functions (PDFs), $P_{t_{sleep}}(t)$ and $P_{t_{awake}}(t)$, the probability that the user is sleeping can be estimated as follows:

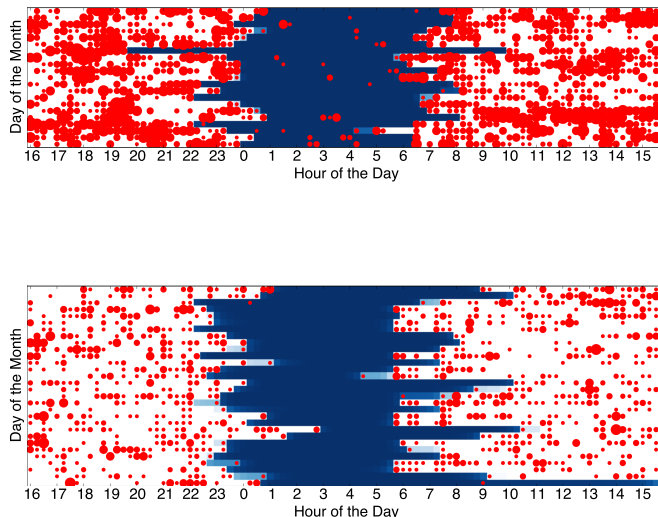$$P_{sleep}(t) = P_{t_{sleep}}(t) - P_{t_{awake}}(t)$$

Fig 6. Sleep matrix of two sample users (21 days from dataset A top, 30 days from dataset B bottom)

	Sleep Time		Wake Time	
	Mean	(Std)	Mean	(Std)
A	23:38	(2h 16m)	7:40	(2h 2m)
B	0:35	(2h 6m)	7:55	(2h 15m)

Table 2. Aggregated sleep and wake times

This is equivalent to stating that a user is currently sleeping if he has passed the time of falling asleep but has not yet passed the time waking up.

The derived values of sleep-length $t_{sleeplength}$ and mid-sleep time $t_{midsleep}$ can be calculated directly from the values of t_{sleep} and t_{awake} for each sample of the trace, and the posterior density can be estimated for these derived values in a similar way as for the model parameters. Fig. 7 shows the aggregate posterior probability density functions for t_{sleep} and t_{awake} for the 126 users of dataset A over $15 - 30$ days, and for the 324 users of dataset B over a selected period of 30 days (just after semester start). It may not be entirely meaningful to average the sleep patterns from all users, but it serves to illustrate the distribution of t_{sleep} and t_{awake} for a larger population. Table 2 summarizes the sleep and wake times.

Across the 30 (14-28) analyzed days for the 324 (126) users of the study, the distribution of sleep durations are as shown in Fig 8. The model allows us to easily compute such metrics. The mean value is around 8:02 ($\pm$2h 36m) for dataset A and 7:20 ($\pm$2h 28m) for dataset B. Notice how the distributions are not completely similar; this is likely due to the fact that the larger dataset B captures the sleeping behavior of students as opposed to dataset A that may have a more diverse demographic distribution.

Fig. 9 shows the probability density functions for the t_{sleep} and t_{awake} times for all users of dataset B, grouped according to weekday. Mondays to Thursdays appear quite

13
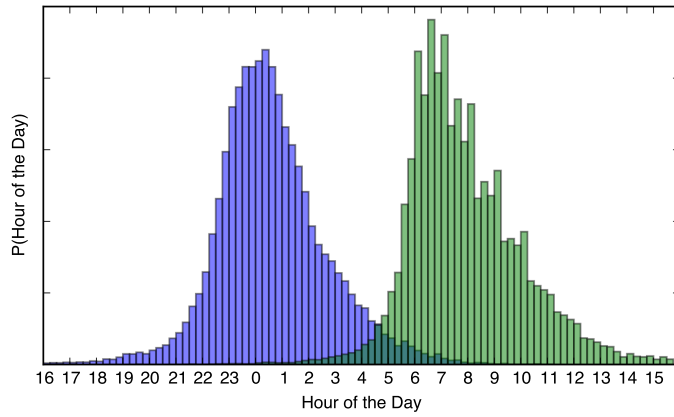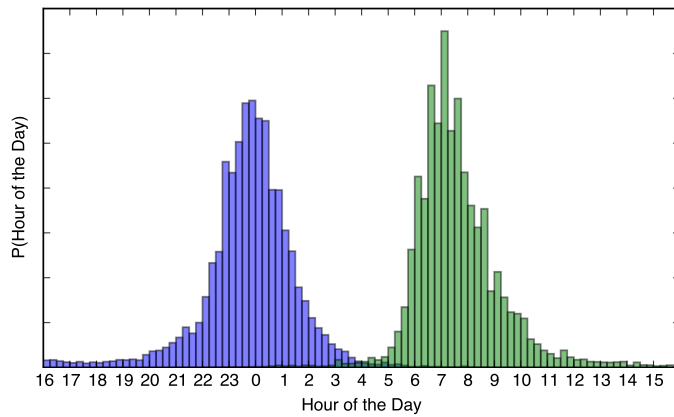
Fig 7. Aggregate Posterior Probability Distributions of t_{sleep} (blue) and t_{awake} (green) (A top, B bottom), showing what the probability is for the specific population to go to sleep or wake up at the specified time.
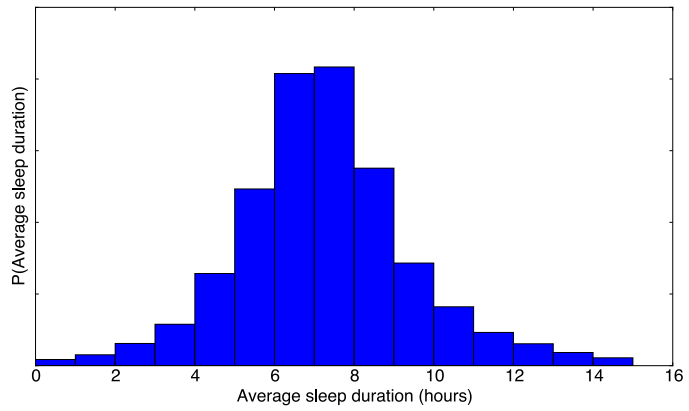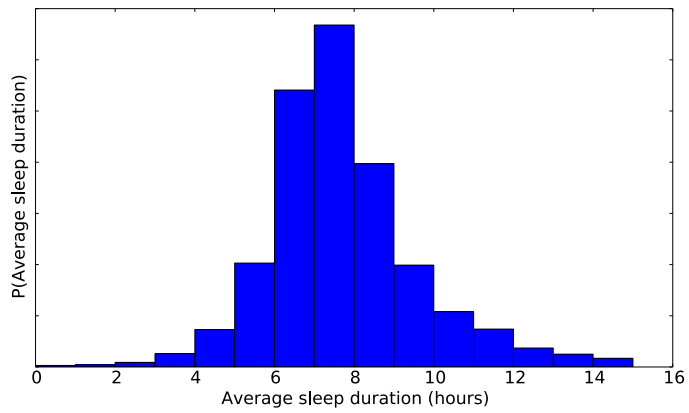
Fig 8. Aggregated Sleep Durations (A top, B bottom), based on the Posterior Probability Functions. This illustrates the probability of the length of a nights sleep within the population within the datasets.
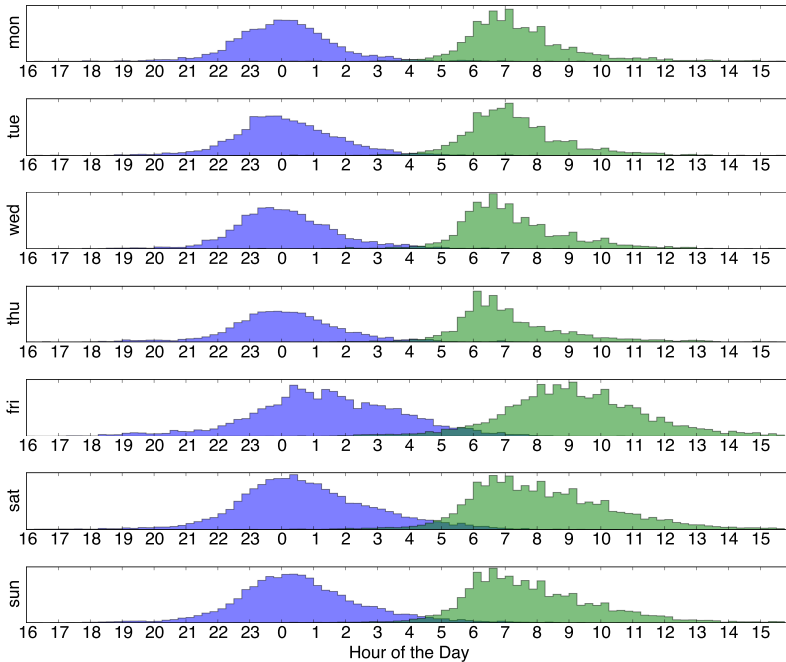
Fig 9. t_{sleep} (blue) and t_{awake} (green) over weekdays for dataset B

similar, but Friday shows a much wider distribution; users typically go to bed much later on Friday and sleep in on Saturday. The distributions start to narrow down Saturday and Sunday but are more "week-like" only from Tuesday morning again.

Discussion

The main contribution of this work is to show how simple counts of smartphone interactions can be used to infer sleep patterns with reasonably high accuracy. We have demonstrated how the seemingly weak signal of screen events carry significant information of the user status. Our method has several advantages:

- The method requires only a smartphone and can therefore be deployed without the need for special equipment or methods, such as fitness or sleep tracking bands, or sleep diaries.

- The data collection is completely automated, as no action is required from the user in setting up the tracking or remembering to log his/her activity.

- Since the model requires only screen interactions, it is absolutely non-intrusive and privacy-preserving. Although in this work we stored the data on a central server for analysis purposes, the data could remain on the phones and the sleep analysis could in principle be run directly on the phones as well.

- Compared to accelerometer or microphone-based methods, using only screen events is much more battery-efficient.

16

Although solutions using screen events have been proposed before [8, 21], our model provides a number of key improvements:

- It is more robust to noise such as screen events generated by checking the phone at night.

- Using a Bayesian formulation allows us to provide confidence intervals for the sleep and awake times, instead of point estimates only.

- It does not depend on ad-hoc rules, but it is based on a well-defined statistical formulation.

- It is fitted and verified on a much larger userbase of over 400 users, and a longer time duration (between 2 and 4 weeks).

Demonstrating the feasibility of inferring reasonable sleep patterns from simple event counts opens the way for new exciting research directions. In particular we believe that similar methods can be applied to large datasets of user activity. For example on social network (such as Twitter, Facebook, Meetup, Gowalla) users leave a trace of their activity in the form of messages, posts, likes, etc. Another great example is Call Detail Records, the logging information kept by telecom providers about user calls and SMS. These events could be treated again as a proxy for sleep and wake cycles.

The main drawback of the proposed method is that it requires that users periodically interact with their phones during their wake time. In line with other recent polls (see for example [29–31]), we show that in most cases this does happen, as the population of users analyzed here tend to check their phone from the early morning to the late night when awake. Different populations, however, such as elderly people less accustomed to smartphone usage, may not show similar usage patterns. There is therefore a need for additional work in order to understand how increased sparsity would affect sleep pattern reconstruction.

Conclusions

We have presented a Bayesian model to infer sleep patterns from smartphone interactions, which we have applied to two datasets of more than 400 users in total. We have compared the model output with ground truth from sleep trackers, and we have shown how the model is able to recover the sleep state with a mean accuracy of 0.89 and a mean F1 score of 0.83. Furthermore, we have shown how the model is capable of producing very reasonable individual and aggregated sleep patterns. Our method represents a cost-effective, non-intrusive and automatic alternative for inferring sleep patterns, and can pave the way for large-scale studies of sleep rhythms.

Acknowledgments

References

1. Choe EK, Consolvo S, Watson NF, Kientz JA. Opportunities for Computing Technologies to Support Healthy Sleep Behaviors. In: Proceedings of the SIGCHI

Conference on Human Factors in Computing Systems. CHI '11. New York, NY, USA: ACM; 2011. p. 3053–3062. Available from: `http://doi.acm.org/10.1145/1978942.1979395`.

2. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási AL. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical. 2008;41(22):1–11. doi:10.1088/1751-8113/41/22/224015.

3. González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature. 2008;453(7196):779–782. doi:10.1038/nature06958.

4. Sekara V, Stopczynski A, Lehmann S. The fundamental structures of dynamic social networks. arXiv preprint arXiv:150604704. 2015;.

5. Onnela JP, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, et al. Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(18):7332–7336. doi:10.1073/pnas.0610245104.

6. Zheng Y, Zhang L, Xie X, Ma WY. Mining interesting locations and travel sequences from GPS trajectories. Proceedings of the 18th international conference on World wide web - WWW '09. 2009;(49):791. doi:10.1145/1526709.1526816.

7. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. Journal of Medical Internet Research. 2015;17(7):e175. doi:10.2196/jmir.4273.

8. Abdullah S, Matthews M, Murnane EL, Gay G. Towards Circadian Computing: "Early to Bed and Early to Rise" Makes Some of Us Unhealthy and Sleep Deprived. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014; p. 673–684. doi:10.1145/2632048.2632100.

9. Smart Alarm Clock;. `https://itunes.apple.com/us/app/smart-alarm-clock-sleep-cycles/id586910133?mt=8`.

10. Sleep Cycle;. `https://itunes.apple.com/us/app/sleep-cycle-alarm-clock/id320606217?mt=8`.

11. SleepBot;. `https://play.google.com/store/apps/details?id=com.lslk.sleepbot&hl=en`.

12. Sleep as Android;. `https://play.google.com/store/apps/details?id=com.urbandroid.sleep&hl=en`.

13. Zhang M, Tillman DA, An SA. Global prevalence of sleep deprivation in students and heavy media use. Education and Information Technologies. 2015; p. 1–16.

14. Orzech KM, Grandner MA, Roane BM, Carskadon MA. Digital media use in the 2 h before bedtime is associated with sleep variables in university students. Computers in Human Behavior. 2016;55:43–50.

15. Richter CP. Biological Clocks in Medicine and Psychiatry: Shock-Phase Hypothesis. Proceedings of the National Academy of Sciences of the United States of America. 1960;46:1506–1530. doi:10.1073/pnas.46.11.1506.

16. Aschoff J. Circadian Rhythms in Man. Science (New York, NY). 1965;148(3676):1427–1432. doi:10.1126/science.148.3676.1427.

17. Ren Y, Wang C, Yang J, Chen Y. Fine-grained sleep monitoring: Hearing your breathing with smartphones. In: Computer Communications (INFOCOM), 2015 IEEE Conference on. IEEE; 2015. p. 1194–1202.

18. Hao T, Xing G, Zhou G. iSleep: unobtrusive sleep quality monitoring using smartphones. In: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems. ACM; 2013. p. 4.

19. Chen Z, Lin M, Chen F, Lane ND, Cardone G, Wang R, et al. Unobtrusive sleep monitoring using smartphones. In: 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth). IEEE; 2013. p. 145–152. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6563918.

20. Gu W, Yang Z, Shangguan L, Sun W, Jin K, Liu Y. Intelligent Sleep Stage Mining Service with Smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. UbiComp '14. New York, NY, USA: ACM; 2014. p. 649–660. Available from: http://doi.acm.org/10.1145/2632048.2632084.

21. Jayarajah K, Radhakrishnan M, Hoi S, Misra A. Candy Crushing Your Sleep. In: Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers. UbiComp/ISWC'15 Adjunct. New York, NY, USA: ACM; 2015. p. 753–762. Available from: http://doi.acm.org/10.1145/2800835.2804393.

22. Hossain HMS, Roy N, Khan MAAH. Sleep Well: A Sound Sleep Monitoring Framework for Community Scaling. In: Proceedings of the 2015 16th IEEE International Conference on Mobile Data Management - Volume 01. MDM '15. Washington, DC, USA: IEEE Computer Society; 2015. p. 44–53. Available from: http://dx.doi.org/10.1109/MDM.2015.42.

23. Sony SmartWear;. http://www.sonymobile.com/global-en/products/smartwear/.

24. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, et al. Measuring large-scale social networks with high resolution. PloS one. 2014;9(4):e95978. doi:10.1371/journal.pone.0095978.

25. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. vol. 2. Taylor & Francis; 2014.

26. Patil A, Huard D, Fonnesbeck CJ. PyMC: Bayesian stochastic modelling in Python. Journal of statistical software. 2010;35(4):1.

27. Fonnesbeck CJ. PyMC version 3; 2015. Available from: https://github.com/pymc-devs/pymc3.

28. Berg A, Meyer R, Yu J. Deviance information criterion for comparing stochastic volatility models. Journal of Business & Economic Statistics. 2004;22(1):107–120.

29. Tecmark survey finds average user picks up their smartphone 221 times a day; 2014. http://www.tecmark.co.uk/smartphone-usage-data-uk-2014/.

30. Newport F. Most U.S. Smartphone Owners Check Phone at Least Hourly; 2015.
    `http://www.gallup.com/poll/184046/`
    `smartphone-owners-check-phone-least-hourly.aspx`.

31. BankOfAmerica. Trends in Consumer Mobility Report; 2015.
    `http://newsroom.bankofamerica.com/files/doc_library/additional/`
    `2015_BAC_Trends_in_Consumer_Mobility_Report.pdf`.

geoplotlib: a Python Toolbox for Visualizing Geographical Data

Andrea Cuttone, Sune Lehmann, Jakob Eg Larsen.

geoplotlib: a Python Toolbox for Visualizing Geographical Data

Andrea Cuttone[1*], Sune Lehmann[1,2], Jakob Eg Larsen[1]

1 DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark
2 The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

* ancu@dtu.dk

Abstract

We introduce geoplotlib, an open-source python toolbox for visualizing geographical data. geoplotlib supports the development of hardware-accelerated interactive visualizations in pure python, and provides implementations of dot maps, kernel density estimation, spatial graphs, Voronoi tesselation, shapefiles and many more common spatial visualizations. We describe geoplotlib design, functionalities and use cases.

Introduction

Geographical data visualization is a fundamental tool for communicating results related to geospatial analyses, and for generating hypotheses during exploratory data analysis [1, 2]. The constantly increasing availability of geolocated data from social media, mobile devices and spatial databases implies that we need new tools for exploring, mining and visualizing large-scale spatial datasets.

The python programming language [3] has been gaining attention as a data analysis tool in the scientific community [4, 5] thanks to the clarity and simplicity of its syntax, and due to an abundance of third-parties libraries e.g. within many disciplines including scientific computing [6, 7], machine learning [8], bayesian modeling [9], neuroscience [10], and bioinformatics [11]. Currently, however, there is limited support for geographical visualization.

Here, we introduce geoplotlib, a python toolbox for visualizing geographical data. geoplotlib provides a simple yet powerful API to generate geographical visualizations on OpenStreetMap [12] tiles. We release geoplotlib as open-source software [13], accompanied by a rich set of examples and documentation.

In the remainder of this paper, we discuss existing tools for geographical visualization and document the geoplotlib functionalities in detail, and finally we evaluate the computational performance on a large-scale dataset.

Related work

In this section we compare existing tools for visualizing geographical data using python. We divide the related work into three categories: pure python packages, HTML-based packages and Geographical Information System plug-ins.

Pure-python packages

The matplotlib [14] library has become the de-facto standard for data visualization in python and provides a large array of visualization tools including scatter and line plots, surface views, 3D plots, barcharts, and boxplots, but it does not provide any support for visualization on a geographical map by default.

The Basemap [15] and Cartopy [16] packages support multiple geographical projections, and provide several visualizations including point plots, heatmaps, contour plots, and shapefiles. PySAL [17] is an open-source library of spatial analysis functions written in Python and provides a number of basic plotting tools, mainly for shapefiles. These libraries however do not allow a user to draw on map tiles, and have limited support for custom visualizations, interactivity, and animation.

HTML-based packages

There is a very rich ecosystem for data visualization for the web. A number of frameworks allow users to generate plots and charts: we cite as representative Protoviz [18], d3 [19], Google Charts [20], sigmajs [21]. There is also a large number of libraries for displaying online tile maps, including Google Maps [22], Bing Maps [23], Leaflet [24], OpenLayers [25], ModestMaps [26], PolyMaps [27].

In order to generate a HTML visualization from python code, it is needed to generate the HTML and JavaScript code that maps the data to the graphical elements. A number of libraries attempt to automate the conversion, such as Folium [28], Vincent [29] and mplleaflet [30]. This process however is often complex, error-prone and time consuming. The complexity can be even greater if some support for animation or interaction is needed. Finally, the JavaScript rendering performance may not be adequate for plotting very large datasets.

Geographical Information System plugins

Geographic Information Systems (GIS) such as QGIS [31], GrassGIS [32], ARCGIS [33], MapInfo [34] provide very powerful tools for spatial data analysis and visualization. GIS tools usually provide some support for python scripting, although the availability varies from one to another. The main limitation of GIS products is their complexity, requiring a significant amount of training to be used effectively, and as discussed before, the need to export the data from python.

Overview

An overview of the geoplotlib architecture is given in Fig. 1. geoplotlib builds on top of numpy [6] and scipy [7] for numerical computations, and OpenGL/pyglet [35] for graphical rendering. geoplotlib implements the map rendering, the geographical projection, the user interface interaction and a number of common geographical visualizations.

Design principles

geoplotlib is designed according to three key principles:

- *simplicity*: geoplotlib tries to minimize the complexity of designing visualizations by providing a set of built-in tools for the most common tasks such as density visualization, spatial graphs, and shapefiles. The geoplotlib API is inspired by the
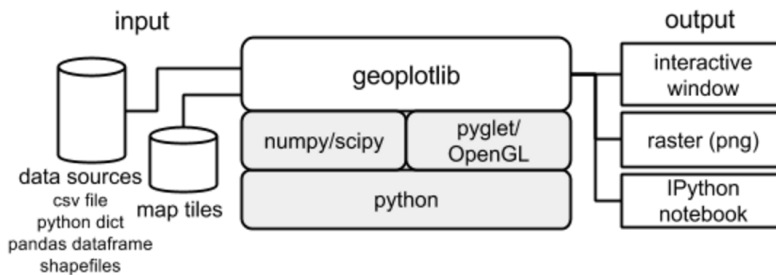
Fig 1. Conceptual overview of the geoplotlib architecture. geoplotlib builds on top of numpy, scipy, and OpenGL/pyglet. It allows to generate geographical visualization as raster, in an interactive window or inside IPython notebooks.

matplotlib [14] programming model and syntax, the de-facto standard for data visualization in python; this makes it easier for matplotlib users to get started.

- *integration*: geoplotlib visualizations are standard python scripts, and may contain any arbitrary python code and use any other package. There is no need to export to other formats (e.g. shapefiles, HTML) or use external programs. This supports a complete integration with the rich python data analysis ecosystem such as scientific computing, machine learning and numerical analysis packages. The visualization can even run within an IPython [36] session, supporting interactive data analysis and facilitating the iterative design for visualizations.

- *performance*: under the hood, geoplotlib uses numpy/scipy for fast numerical computations, and pyglet/OpenGL for hardware-accelerated graphical rendering. This allows the visualizations to scale to millions of datapoints in realtime.

A first script

A simple geoplotlib script looks like this:

```
data = read_csv('data/bus.csv')
geoplotlib.dot(data)
geoplotlib.show()
```

This script launches the geoplotlib window and shows a dot map of the data points, in this example the location of bus stops in Denmark (Fig. 2). geoplotlib automatically determines the map bounding box, downloads the map tiles, perform the geographical projection, draws the base map and the visualization layers (the dots in this example). The map is interactive and allows a user to zoom and pan with mouse and keyboard.

As discussed above, the usage of the geoplotlib API is very similar to matplotlib. The visualization canvas is initially empty, and each command adds a new layer of graphics. The geoplotlib window is displayed when `show()` is called. Alternatively, the map can be rendered to image file using `savefig('filename')`, or displayed inline in an IPython notebook using `inline()`.

Layers

The geoplotlib package provides several common geographical visualizations in form of layers. The API provides convenient methods for quickly adding a new visualization

layer. In this section we provide a summary of the built-in visualizations. The data for all examples is available on the project website [13].

Dot Map

An elementary operation in geographical visualization is to display "what is where", that is to place a graphic element on the map for each of the objects in consideration. This provides an immediate idea of the absolute and relative locations of objects. Moreover, the density of points directly maps to the density of objects on geographical surface, identifying zones of higher and lower density. An example of dot map is shown in Fig. 2. The dot map shows the spatial distribution of bus stops in Denmark at a glance. The zones of higher density – corresponding to the Copenhagen metropolitan area and to the other major cities are immediately recognizable. The `dot` method allows users to configure points size, color and transparency, and optionally to attach a dynamic tooltip to each point.
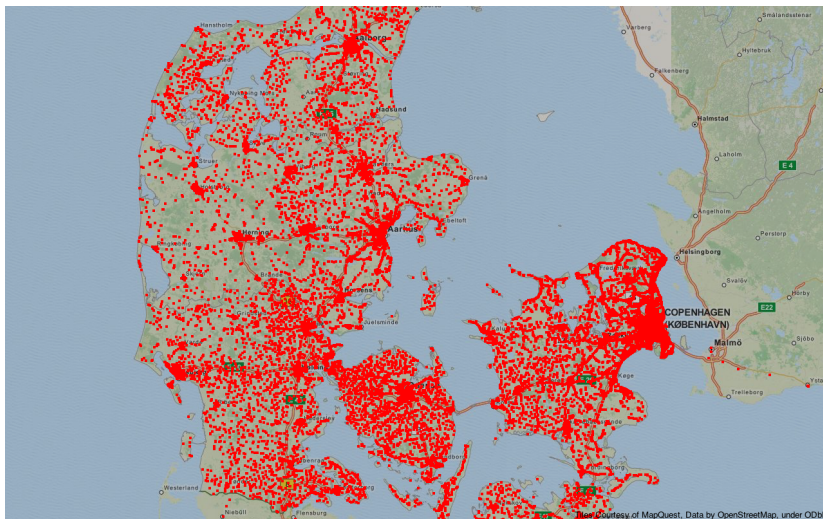
Fig 2. A dot map of bus stops in Denmark, where each sample is represented by a point.

2D Histogram

One limitation of dot maps is that it is hard to distinguish between areas of high density, as the number of point is so high that they uniformly cover the visualization canvas. A more direct visualization of density is to compute a 2D histogram of point coordinates. A uniformly spaced grid is placed on the map, and the number of samples within each cell is counted. This value is an approximation of the density, and can be visualized using a color scale. In geoplotlib we can generate the 2D histogram of the data using `hist`:

```
data = read_csv('data/opencellid_dk.csv')
geoplotlib.hist(data, colorscale='sqrt', binsize=8)
geoplotlib.show()
```

Here `binsize` refers to the size in pixels of the histogram bins.

4

Fig 3. A 2D histogram of cell tower locations in Denmark, using the 'hot' colormap (dark red is lower, yellow-white is higher).

The example above loads some data related to cell tower positions in Denmark, and then generates a histogram with a specific colorscale and bin size (Fig. 3). Compared to the dot map example, the histogram provides a clearer depiction of the density distribution.

Heatmap

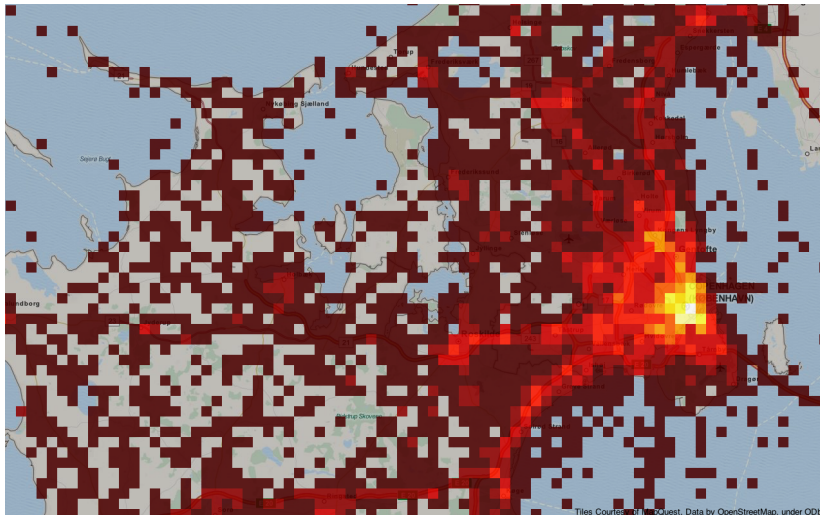The main deficiency of histogram visualizations is that they are discrete approximations of a (effectively continuous) density function. This creates a dependence on the bin size and offset, rendering histograms sensitive to noise and outliers. To generate a smoother approximation, a kernel density estimator approximates the true density function applying kernel functions in a window around each point [37]. The size of this window depends on the bandwidth parameter: a smaller bandwidth will produce more detailed but also noisier estimation, while a larger bandwidth will produce a less detailed but smoother estimation. A kernel estimation function can then be visualized by a surface where the color encodes the density value (this visualization is often called a "heatmap"). In geoplotlib, the `kde` method generates a kernel density estimation visualization:

```
data = read_csv('data/opencellid_dk.csv')
geoplotlib.kde(data, bw=[5,5])
geoplotlib.show()
```

Fig. 4 shows the kernel density estimation applied to the cell tower data. Comparing the histogram from Fig. 3 with the kernel density estimation in Fig. 4, it is evident how the latter produces a smoother and consequently clearer visualization of density. The kernel bandwidth (in screen coordinates) can be configured to regulate the smoothness. The density upper bound can be set to clip density values over a threshold. Also the density lower bound can be set, to avoid rendering areas of very low density:

```
# lowering clip_above changes
# the max value in the color scale
geoplotlib.kde(data, bw=[5,5],
               cut_below=1e-6, clip_above=1)
```
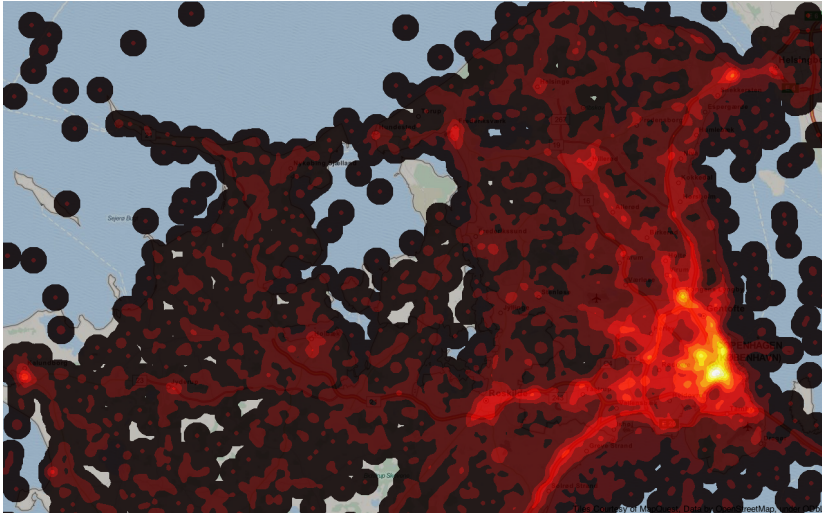
5

Fig 4. A heatmap (kernel density estimation) of the cell tower locations in Denmark, using a hot colormap (dark red is lower, yellow-white is higher). The kernel density produces a much smoother estimation and therefore a clearer visual representation of the density, if compared with a histogram (Fig. 3).

```
# different bandwidths
geoplotlib.kde(data, bw=[20,20],
               cmap='coolwarm', cut_below=1e-6)
geoplotlib.kde(data, bw=[2,2],
               cmap='coolwarm', cut_below=1e-6)

# linear colorscale
geoplotlib.kde(data, bw=[5,5], cmap='jet',
               cut_below=1e-6, scaling='lin')
```

Markers

In some cases it is useful to represent objects on the map using custom symbols with specific meaning. The `markers` method allows a user to place customs markers on the map:

```
metro = read_csv('./data/metro.csv')
s_tog = read_csv('./data/s-tog.csv')

geoplotlib.markers(metro, 'data/m.png',
               f_tooltip=lambda r: r['name'])
geoplotlib.markers(s_tog, 'data/s-tog.png',
               f_tooltip=lambda r: r['name'])
geoplotlib.show()
```

Fig. 5 shows an example of custom markers for metro and train stops in Copenhagen. Markers graphics can be any common raster format (png, jpeg, tiff), and can be rescaled to a custom size. Optionally a dynamic tooltip can be attached to each marker.

Spatial Graph

Spatial graphs are a special type of graphs where nodes have a well-defined spatial configuration. Examples includes transport networks (bus routes, train tracks, flight
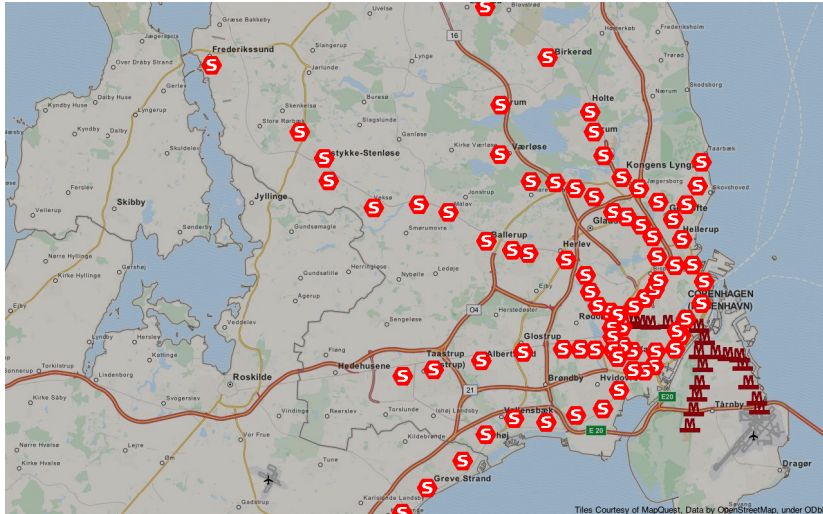
Fig 5. Markers showing the location of train and metro stations in the Copenhagen area. geoplotlib can use any raster image (png, jpg) as markers.

paths), supply chain networks, phone call networks and commute networks. In geoplotlib `graph` renders a spatial graph:

```
data = read_csv('./data/flights.csv')
geoplotlib.graph(data,
                 src_lat='lat_departure',
                 src_lon='lon_departure',
                 dest_lat='lat_arrival',
                 dest_lon='lon_arrival',
                 color='hot_r',
                 alpha=16,
                 linewidth=2)
geoplotlib.show()
```

Fig. 6 shows the resulting spatial graph of airport locations, where each node represents an airport and each edge represents a flight connection. Edges are colored using a colormap encoding the edge length.

Voronoi Tessellation

A Voronoi tessellation [38] is a partition of space into regions induced by some seed points, so that each region (called a Voronoi cell) consists of all points closer to a specific seed than to any others. The analysis of Voronoi tessellation is used in numerous fields including ecology, hydrology, epidemiology, mining and mobility studies.

In geoplotlib `voronoi` can be used to generate a Voronoi tessellation visualization. Voronoi cell fill, shading and colors can be configured.

```
data = read_csv('data/bus.csv')
geoplotlib.voronoi(data, line_color='b')
geoplotlib.show()
```

Fig. 7 provides an example of Voronoi tessellation of bus stops in Denmark. Voronoi cells provide a measure of the space closer to one stop than any others. The density of points is also captured by the size of Voronoi cells, as smaller cells indicate more densely covered areas.
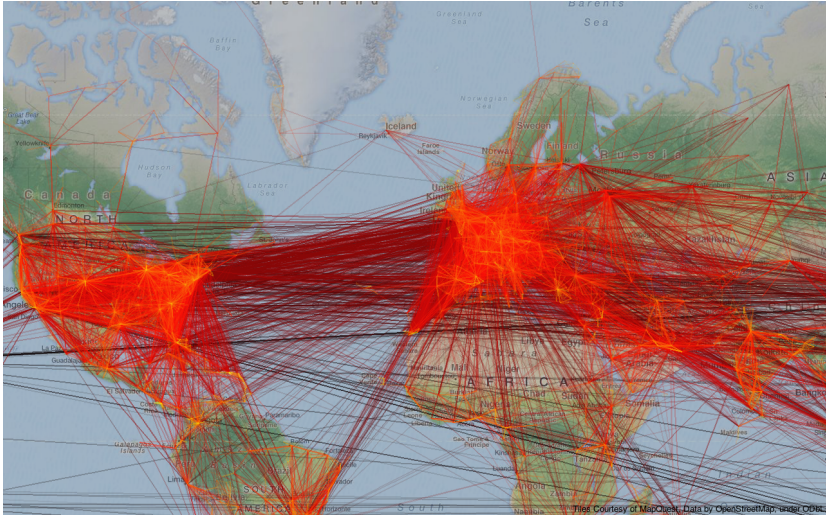
Fig 6. Spatial graph of airport locations, where each node represents an airport and each edge represent a flight connection. Edges are colored using a colormap encoding the edge length.

Delaunay triangulation

A Delaunay triangulation [39] is a convenient method for generating triangles meshes from a set of points. In geoplotlib the `delaunay` method can be used for this purpose.
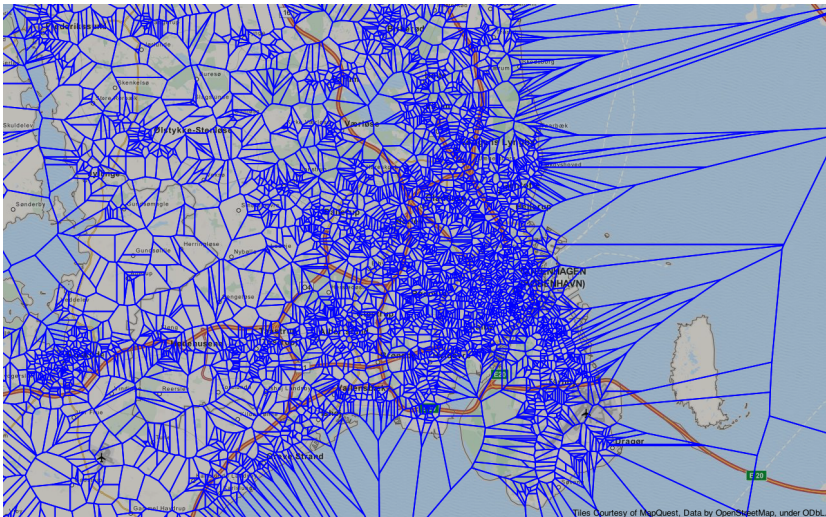
Fig 7. Voronoi tessellation of bus stops in Denmark. Voronoi cells provide an estimation of the space closer to one stop than any others. The density of points is also captured by the size of Voronoi cells, as smaller cells indicate more densely covered areas.

The edge color can be configured to a fixed value, or to encode the length of the edges.

```
data = read_csv('data/bus.csv')
geoplotlib.delaunay(data, cmap='hot_r')
geoplotlib.show()
```

Fig. 8 shows the Delaunay triangulation of bus stops, with edges colored according to length.
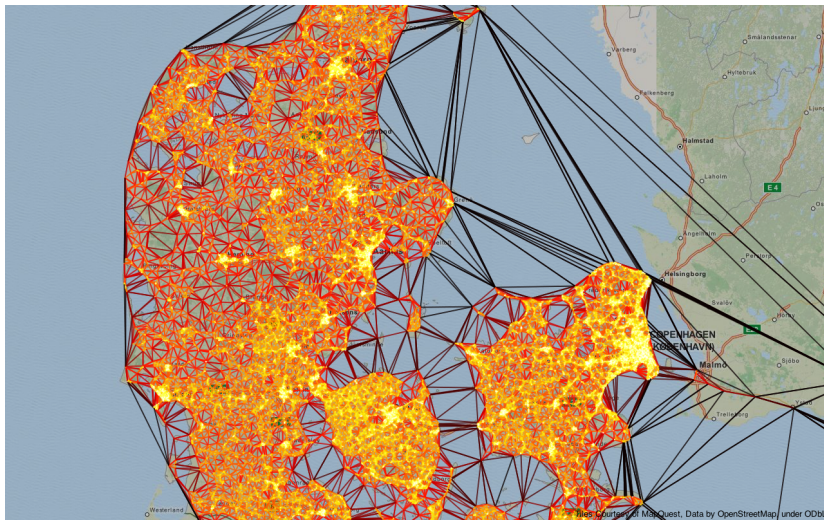
Fig 8. A Delaunay triangulation of the bus stops in Denmark, with edges colored according to length

Convex Hull

A convex hull [39] of a set of finite points is the smallest convex polygon that contains all the points. Convex hulls can be used for example to visualize the approximate area corresponding to a set of points. In geoplotlib:

```
geoplotlib.convexhull(data, color, fill=True)
```

Fig. 9 shows the bus stops points split into 6 groups, and each group is represented by a differently colored convex hull.

Shapefiles

Shapefile [40] is a popular file format for describing vector graphics for geographical information systems. geoplotlib uses pyshp [41] to parse the shapefiles. The line color can be configured and an optional tooltip can be attached to each shape. In the following example we display the *kommuner* administrative regions in Denmark (Fig. 10):

```
geoplotlib.shapefiles('data/dk_kommune/dk_kommune',
                      f_tooltip=lambda attr: attr['STEDNAVN'],
                      color=[0,0,255])
geoplotlib.show()
```

GeoJSON

GeoJSON [42] is a human-readable format for encoding geographical data, such as polygons and lines. geoplotlib can render shapes from the GeoJSON format, and shape color and tooltip can be dynamically altered to encode data. For instance GeoJSON shapes can be used to generate a choropleth where each geographic unit is colored to encode a continuous variable. In the following example (Fig. 11) we generate a choropleth of unemployment in USA [43]:

```python
def get_color(properties):
    key = str(int(properties['STATE']))
    key += properties['COUNTY']
    if key in unemployment:
        return cmap.to_color(unemployment.get(key),
                             .15, 'lin')
    else:
        return [0, 0, 0, 0]

with open('data/unemployment.json') as fin:
    unemployment = json.load(fin)

cmap = ColorMap('Blues', alpha=255, levels=10)
geoplotlib.geojson('data/gz_2010_us_050_00_20m.json',
    fill=True, color=get_color,
    f_tooltip=lambda properties: properties['NAME'])
geoplotlib.geojson('data/gz_2010_us_050_00_20m.json',
    fill=False, color=[255, 255, 255, 64])
geoplotlib.show()
```

Advanced Functionalities

Data access

The `DataAccessObject` class is the fundamental interface between the raw data and all the geoplotlib visualizations. A `DataAccessObject` is conceptually similar to a table
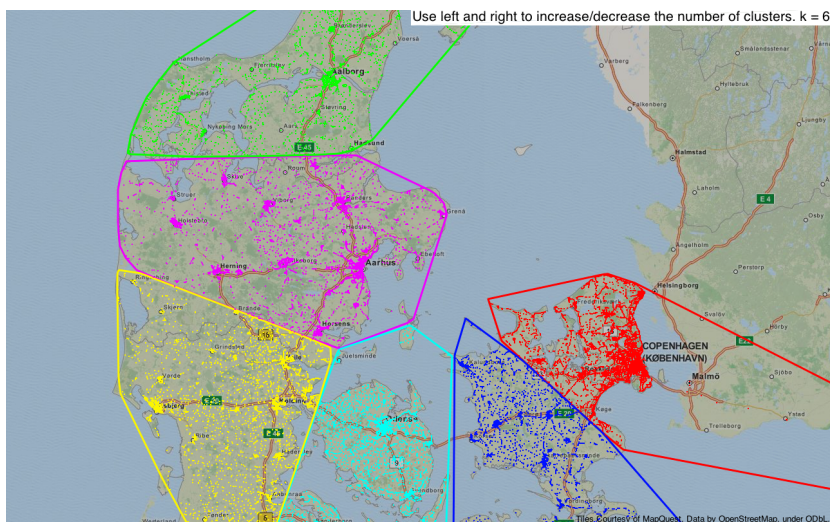
Fig 9. The bus stops points are split into 6 groups, and each group is represented by a different colored convex hull.
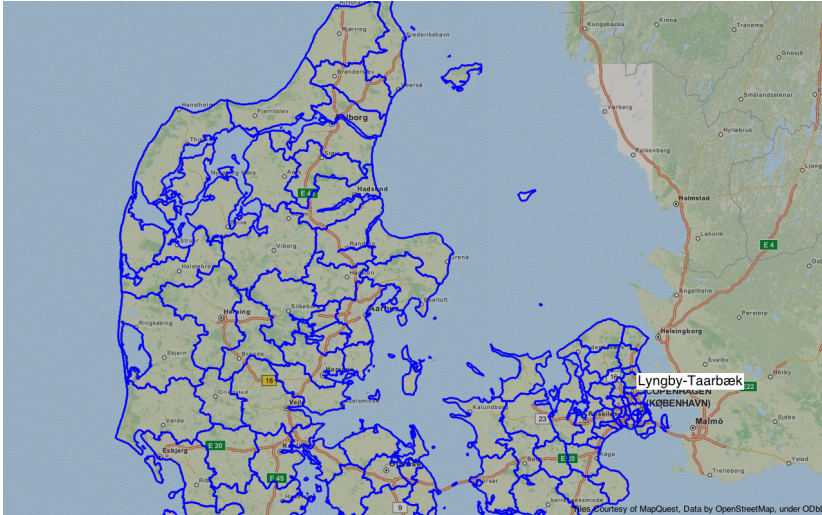
Fig 10. Rendering the shapefiles for the *kommuner* administrative regions in Denmark.

with one column for each field and one row for each sample. This paradigm is very common in data analysis terminology, and is equivalent to ndarrays in numpy, and dataframes in pandas and R. A `DataAccessObject` can be initialized by reading a comma-separated values (CSV) file with the built-in `read_csv` method, or can be constructed from a python dict, or from a pandas [44] dataframe:

```
dao1 = DataAccessObject({'field1': somevalues,
                         'field2': othervalues})
dao2 = DataAccessObject(mydataframe)
```
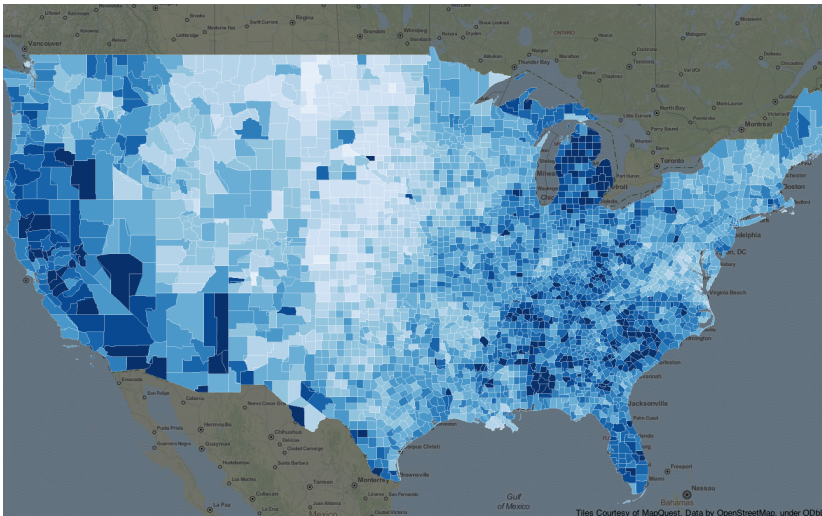

Fig 11. Choropleth of unemployment in USA using GeoJSON shapefiles

11

```
dao3 = read_csv('somefile.csv')
```

The only two fields required are `lat` and `lon`, which represent to the geographic coordinates. Most of the built-in visualization implicitly refer to these two fields to locate entities in space. `DataAccessObject` also provides a few method for basic data wrangling, such as filtering, grouping, renaming and deleting rows and columns.

Tile providers

Any OpenStreetMap tile server can be configured using the `tile_provider` method (users are kindly asked to check the tile usage policy for the selected server, and make sure to provide attribution as needed). A number of common free tiles providers are supported, including Stamen Watercolor and Toner [45], MapQuest [46], CartoDB Positron and DarkMatter [47].

Defining custom layers

The built-in visualizations provide various commonly used tools for geographical data visualization. Multiple layers can be combined into a single visualization for richer display. For even more complex visualizations, geoplotlib allows users to define custom layers. In order to generate a new visualization, a new class extending `BaseLayer` must be defined. The custom layer must at least define an `invalidate` and a `draw` method. The `invalidate` method is called each time the map projection must be recalculated, which typically happens each time that the map zoom-level changes. The `invalidate` method receives a `Projection` object, which provides methods for transforming the data points from the geographic coordinates to screen coordinates. The screen coordinates can then be passed to a `BatchPainter` object for the rendering. A `BatchPainter` can efficiently draw OpenGL primitives such as points, lines and polygons. The `draw` method is called at each frame, and typically calls the `batch_draw` method of the painter prepared during `invalidate`. The following is a complete example of a custom layer, which simply draws samples as points:

```
class CustomLayer(BaseLayer):

    def __init__(self, data):
        self.data = data

    def invalidate(self, proj):
        x, y = proj.lonlat_to_screen(self.data['lon'],
                                     self.data['lat'])
        self.painter = BatchPainter()
        self.painter.points(x, y)

    def draw(self, proj, mouse_x, mouse_y, ui_manager):
        self.painter.batch_draw()
```

The final step needed is to add the layer to the visualization using `add_layer`, then call `show`:

```
geoplotlib.add_layer(CustomLayer(mydata))
geoplotlib.show()
```

Animation

A custom layer can be also used for creating animated visualizations. Each time the draw method is called, the custom layer can update its state to the next frame. As an example, let us imagine having data containing the position of an object over time. A simple animation can use a frame counter, and at each frame render only the datapoint at the current instant:

```
class AnimatedLayer(BaseLayer):

    def __init__(self, data):
        self.data = data
        self.frame_counter = 0

    def invalidate(self, proj):
        self.x, self.y = proj.lonlat_to_screen(
            self.data['lon'], self.data['lat'])

    def draw(self, proj, mouse_x, mouse_y, ui_manager):
        self.painter = BatchPainter()
        self.painter.points(self.x[self.frame_counter],
                            self.y[self.frame_counter])
        self.painter.batch_draw()
        self.frame_counter += 1
```

Notice that in this case we do not initialize the `BatchPainter` inside `invalidate`, but we create a new one at each frame. We also keep track of the current frame with the `frame_counter` variable. Even this very simple code is able to visualize a non-trivial animation of an object moving over time. To produce a movie from the animation, individual frames can be captured using the `screenshot` method, and then combined together.

Colormaps

Colors can be used as additional mapping for encoding information into a visualization. Continuous variables (for example points density or the edges distances) can be mapped to a continuous color scale. The `ColorMap` class allows a user to perform this conversion. A `ColorMap` object is constructed by passing any of the matplotlib colorscales, and optionally an alpha value and a number of discretization levels. The `to_color` method performs the conversion from real value to color:

```
# hot colormap
cmap = ColorMap('hot')

# Reds colormap with transparency
cmap = ColorMap('Reds', 128)

# coolwarm colormap with 4 levels
cmap = ColorMap('coolwarm', levels=4)

# linear scaling
cmap.to_color(10, 100, 'lin')

# logarithmic scaling
cmap.to_color(10, 100, 'log')

# square-root scaling
cmap.to_color(10, 100, 'sqrt')
```

Discrete variables such as categories can be represented using categorical colormaps. The `colorbrewer` method provides access to the ColorBrewer [48] colors. Categorical colormaps can be also generated from regular colormaps using using `create_set_cmap`:

```
cmap1 = colorbrewer([1,2,3])
cmap2 = create_set_cmap('hsv', [1,2,3])
```

Controlling the map view

The map view is determined by the projection parameters: the latitude offset, the longitude offset and the zoom level. By default, the projection is chosen so to fit all selected points, with the maximum zoom level possible. The view can changed to a specific portion of the map by passing a `BoundingBox` object to the `set_bbox` method.

A `BoundingBox` object defines the map view boundaries, and can be constructed in multiple ways. The most direct way is to specify two ranges of latitudes and longitudes. Alternatively, a `BoundingBox` can be constructed to fit a subset of points using the `from_points` methods.

```
bbox1 = BoundingBox(north=51.3, west=-124.3,
                    south=14.8, east=-56.8)

bbox2 = BoundingBox.from_points(lons, lats)
```

Interactivity

Finally, geoplotlib allows users to create interactive visualizations by provides support for rendering a user interface, and dynamically changing the visualization on user input:

- on-screen text such as information or status can be added using the `UiManager` class.

- mouseover tooltips can be configured on arbitrary graphical elements or screen regions using the `HotspotManager` class.

- layers can be configured to react to specific key presses by defining a `on_key_release` method

Performance

We test the performance of geoplotlib by generating some of the described visualization on a dataset consisting of one million samples, using the default visualization parameters. All tests consider only the time needed for the actual rendering of the visualization, excluding the time for loading the data. The measurements are repeated 10 times for each visualization type. The experiments were performed on a MacBook Pro 2012 with an Intel 2.3 GHz i7 CPU, 8 GB RAM and nVidia GeForce GT 650M GPU. Table 1 shows that in all cases the visualizations require only a few seconds, thus demonstrating that geoplotlib is suitable even for large-scale datasets.

Table 1. Execution time for one million samples

visualization type	mean time [s]	SD [s]
dot	1.57	0.08
graph	1.99	0.09
hist	8.12	0.55
kde	5.37	0.37
voronoi	3.08	0.66

Conclusion

We have presented geoplotlib, a python toolbox for generating geographical visualizations. We demonstrated how geoplotlib provides a simple yet powerful API to visualize geographical data, greatly facilitating exploratory data analysis of geographical information. We believe that geoplotlib can become a powerful tool in the data analyst toolbox, both for analyzing complex spatial patterns and for communicating results in forms of geographical visualizations. Future work includes the addition of more visualization tools, and the integration of spatial analysis methods.

Acknowledgments

References

1. Tukey JW. Exploratory Data Analysis. Addison-Wesley; 1977.

2. Andrienko N, Andrienko G. Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach. Springer Science & Business Media; 2006.

3. Van Rossum G, Drake Jr FL. Python Reference Manual. Centrum voor Wiskunde en Informatica Amsterdam; 1995.

4. Millman KJ, Aivazis M. Python for Scientists and Engineers. Computing in Science & Engineering. 2011;13(2):9–12. doi:http://dx.doi.org/10.1109/MCSE.2011.36.

5. Oliphant TE. Python For Scientific Computing. Computing in Science and Engineering. 2007;9(3):10–20.

6. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering. 2011;13(2):22–30.

7. Jones E, Oliphant T, Peterson P, et al.. SciPy: Open Source Scientific Tools for Python; 2001–. Available from: `http://www.scipy.org/`.

8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research. 2011;12:2825–2830.

9. Patil A, Huard D, Fonnesbeck CJ. PyMC: Bayesian Stochastic Modelling in Python. Journal of statistical software. 2010;35(4):1.

10. Peirce JW. PsychoPy — Psychophysics Software in Python. Journal of neuroscience methods. 2007;162(1):8–13.

11. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. BioPython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. Bioinformatics. 2009;25(11):1422–1423.

12. Haklay M, Weber P. OpenStreetMap: User-Generated Street Maps. Pervasive Computing, IEEE. 2008;7(4):12–18.

13. Cuttone A. geoplotlib; 2015-2016. Available from: `https://github.com/andrea-cuttone/geoplotlib`.

14. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science and Engineering. 2007;9(3):90–95.

15. Whitaker J. Basemap;. Available from: `http://matplotlib.org/basemap/`.

16. Met Office. Cartopy: A Cartographic Python Library with a Matplotlib Interface; 2010 - 2015. Available from: `http://scitools.org.uk/cartopy`.

17. Rey SJ, Anselin L. In: Fischer MM, Getis A, editors. PySAL a Python Library for Spatial Analytical Methods. Berlin: Springer-Verlag; 2010. p. 175–193.

18. Bostock M, Heer J. Protovis: A Graphical Toolkit for Visualization. Visualization and Computer Graphics, IEEE Transactions on. 2009;15(6):1121–1128.

19. Bostock M, Ogievetsky V, Heer J. D^3 Data-Driven Documents. Visualization and Computer Graphics, IEEE Transactions on. 2011;17(12):2301–2309.

20. Google Charts;. Available from: https://developers.google.com/chart/.

21. Alexis Jacomy GP. sigmajs;. Available from: http://sigmajs.org/.

22. Google Maps;. Available from: https://developers.google.com/maps/.

23. Bing Maps;. Available from: https://www.bingmapsportal.com/.

24. leafletjs;. Available from: http://leafletjs.com/.

25. openlayers;. Available from: http://openlayers.org/.

26. modestmaps;. Available from: http://modestmaps.com/.

27. polymaps;. Available from: http://polymaps.org/.

28. folium;. Available from: https://github.com/wrobstory/folium.

29. vincent;. Available from: https://github.com/wrobstory/vincent.

30. mplleaflet;. Available from: https://github.com/jwass/mplleaflet.

31. QGis. Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project. 2011;.

32. GrassGIS;. Available from: http://grass.osgeo.org/.

33. ArcGIS;. Available from: http://www.arcgis.com/.

34. MapInfo;. Available from: http://www.mapinfo.com/.

35. pyglet;. Available from: https://bitbucket.org/pyglet/pyglet/wiki/Home.

36. Pérez F, Granger BE. IPython: A System for Interactive Scientific Computing. Computing in Science & Engineering. 2007;9(3):21–29.

37. Parzen E. On Estimation Of a Probability Density Function and Mode. The annals of mathematical statistics. 1962; p. 1065–1076.

38. Aurenhammer F. Voronoi Diagrams — A Survey of a Fundamental Geometric Data Structure. ACM Computing Surveys (CSUR). 1991;23(3):345–405.

39. De Berg M, Van Kreveld M, Overmars M, Schwarzkopf OC. Computational Geometry. Springer; 2000.

40. ESRI. ESRI Shapefile Technical Description; 1998.

41. pyshp;. Available from: https://github.com/GeospatialPython/pyshp.

42. Butler H. geoJSON;. Available from: http://geojson.org/geojson-spec.html.

43. Bostock M. Unemployment in USA;. Available from:
    `http://bl.ocks.org/mbostock/4060606`.

44. McKinney W. Data Structures for Statistical Computing in Python. In: van der
    Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference;
    2010. p. 51 – 56.

45. Stamen;. Available from: `http://maps.stamen.com/`.

46. MapQuest;. Available from:
    `http://developer.mapquest.com/web/products/open/map`.

47. CartoDB;. Available from: `http://cartodb.com/basemaps/`.

48. Harrower M, Brewer CA. ColorBrewer.org: An Online Tool for Selecting Colour
    Schemes for Maps. The Cartographic Journal. 2003;40(1):27–37.

Inferring Human Mobility from Sparse Low Accuracy Mobile Sensing Data

Andrea Cuttone, Sune Lehmann, Jakob Eg Larsen.

Inferring Human Mobility from Sparse Low Accuracy Mobile Sensing Data

Andrea Cuttone
DTU Compute
Dept. of Applied Mathematics
and Computer Science
Richard Petersens Plads, Bldg.
324
2800 Kgs Lyngby, Denmark
ancu@dtu.dk

Sune Lehmann
DTU Compute
Dept. of Applied Mathematics
and Computer Science
Richard Petersens Plads, Bldg.
324
2800 Kgs Lyngby, Denmark
sljo@dtu.dk

Jakob Eg Larsen
DTU Compute
Dept. of Applied Mathematics
and Computer Science
Richard Petersens Plads, Bldg.
324
2800 Kgs Lyngby, Denmark
jaeg@dtu.dk

Abstract

Understanding both collective and personal human
mobility is a central topic in Computational Social
Science. Smartphone sensing data is emerging as a
promising source for studying human mobility. However,
most literature focuses on high-precision GPS positioning
and high-frequency sampling, which is not always feasible
in a longitudinal study or for everyday applications
because location sensing has a high battery cost. In this
paper we study the feasibility of inferring human mobility
from sparse, low accuracy mobile sensing data. We
validate our results using participants' location diaries,
and analyze the inferred geographical networks, the time
spent at different places, and the number of unique places
over time. Our results suggest that low resolution data
allows accurate inference of human mobility patterns.

ACM Classification Keywords

H.2.8 [Database Management]: Database Applications –
data mining

Author Keywords

Spatial data mining; mobile sensing; place discovery;
location sensing; mobility;

Introduction

Understanding human mobility has a number of important applications, ranging from containment of infection diseases [4] to urban planning and traffic management [8]. Moreover mobility plays an important role in understanding face-to-face [1] and long distance social interactions [12].

Most studies on mobility have been using Call Detail Records (CDR) as proxy for location [15, 7, 3], inferring people's positions depending on the cell towers that mobile phones are connected to. Such data sources can only produce very rough location estimates with space granularity of kilometers and time granularity of hours. Thanks to the explosive adoption of smartphones equipped with location sensors, mobile sensing data is becoming a promising alternative for inferring patterns of human mobility and social interaction [14].

An extensive literature on mining location data has been produced, but a large amount of this work focuses on high-precision GPS positioning and high-frequency sampling, with location collected every few minutes [20, 21] or even every few seconds [2, 19, 10, 18, 11, 5, 17]. Such collection methods are not realistic in a longitudinal study based on smartphone data, since high frequency GPS sampling would lead to rapid battery drain. For this reason, in the *Copenhagen Network Study* [16] we collect location data in an opportunistic fashion using Android smartphones and the Google Location API[1], which provides an improved battery life at the price of lower sampling frequency and more variable accuracy. Understanding data collected at a fixed rate is a necessary baseline and point-of-comparison

---

[1]http://developer.android.com/google/play-services/location.html

for more advanced low energy approaches, such as adaptive sampling [9]. Thus, in this paper we focus on the specific problem of inferring human mobility from such sparse, low accuracy mobile sensing data. We perform a small experiment using diaries collected by 6 participants and one researcher, and we motivate that it is possible to infer mobility patterns with reasonable accuracy.

Related work

A number of recent studies on inferring mobility from mobile sensing data have been performed, and they differ for number of participants, frequency of sampling, and data mining techniques. Here we cite a few representative examples. Ashbrook et al. [2] collect data for 6 users at 1 sample per second; places are extracted using k-means and evaluated by user interviews. Zhou et al. [21] collect data for 28 users at 1 sample per minute, and extract places using DBSCAN clustering. Palma et al. [13] propose a speed-based DBSCAN clustering, but give no user evaluation of the results. Montoliu et al. [11] collect data for 8 participants using continuous sampling; locations are identified using grid clustering, and the results evaluated comparing with participants diaries. Cao et al. [6] analyze the GPS traces of 119 cars, and exploit GPS signal loss for detecting stops; the inferred network of places is analyzed. Yan et al. [17] determine stops using a dynamic speed thresholding and places are extracted using intersection with geometry; results are evaluated for 6 users.

Problem definition

We collect location data as a sequence of samples represented by tuples in the format *(timestamp, latitude, longitude, accuracy)* ordered by non-decreasing *timestamp.* A point of interest (POI) is a location of relevance for a person, such as his home, his workplace or a gym he frequents. A POI can be described as a tuple

(POI-id, POI-latitude, POI-longitude), where *POI-latitude* and *POI-longitude* refer to the centroid of the POI. A stop at a POI is a specific occasion when the person has visited a POI in a given day and hour. A stop is thus described by a tuple *(timestamp-arrival, timestamp-departure, POI-id)*. By definition, each stop is related exactly to one POI, but one POI can be related to many stops. Stops are non-overlapping, that is $timestamp\text{-}arrival_{i+1} > timestamp\text{-}departure_i$ for every stop. A stops extraction algorithm accepts as input a sequence of location samples, and produces as output a set of stops related to POI.

Given the discovered stops and the ground truth stops, we need to compute how many actually match. In several related studies [21, 11, 17, 2] the participants were interviewed and asked to manually match the discovered locations with the personal diaries. In our case we set threshold parameters to define when a match occurs, trying to emulate human judgment. A person asked to decide whether two stops are matching would probably say yes if they are reasonably close to each other and the time interval was approximately the same. Therefore we set the following threshold parameters: we use a distance of 150 meters (roughly the average size of one city block), and we adopt a time threshold equal to 25% of the total stop duration, so that shorter stops must be identified more precisely than longer stops. Although it is possible to argue for different choices of these parameters, we found that different values (distance threshold = 50 and 100 meters, and time threshold = 10%) have only minor effect on the results. Let B_{stops} be the baseline stops (stops in the diary), D_{stops} the discovered stops (stops extracted by an algorithm), and $B_{stops} \cap D_{stops}$ the baseline stops

which are discovered and recognized as matching. Let:

$$recall_{stops} = |B_{stops} \cap D_{stops}|/|B_{stops}|$$
$$precision_{stops} = |B_{stops} \cap D_{stops}|/|D_{stops}|$$
$$f1_{stops} = \frac{2 \cdot precision_{stops} \cdot recall_{stops}}{precision_{stops} + recall_{stops}}$$

In a similar fashion we evaluate the performance of identifying POI by calculating $f1_{POI}$, considering a match if the centroids are at distance < 150 meters.

Experimental settings

We recruited a total of 6 participants, all of them students at our campus. Participants were provided with a Samsung Galaxy Nexus smartphone, with a collector app based on the Funf Open Sensing framework [1]. Participants were instructed to use the provided smartphone as their main phone. We asked participants to keep a diary of their daily movements, and we provided them an electronic spreadsheet where they could fill entries in the format: date, hour, place description. We instructed them to keep the diary updated as accurately as possible, and we sent weekly reminders via email. One author also carried a Samsung Galaxy Nexus smartphone for collecting his own data and kept his own location diary. Table 1 shows the summary statistics of diary entries for each participant (P1-P6) and for the researcher (R). The participants' data was collected in October and November 2013, while the researcher's data was collected in four periods between September and January 2014. At the end of the study, the participants were asked to create a list of POI with description, latitude and longitude obtained using Google Maps[2].

---

[2]http://maps.google.com

User	Stops	μ(Stops/day)	σ(Stops/day)	Unique POI
P1	120	2.9	1.3	33
P2	127	3.2	1.5	34
P3	187	4.0	1.9	37
P4	130	2.9	1.3	49
P5	61	2.3	0.9	18
P6	111	4.4	2.3	22
R	227	3.5	1.4	39

Table 1: Summary statistics of diary entries for each participant (P1-P6) and for the researcher (R).

The usage of diaries as ground truth presents some challenges, as participants' compliance tends to decrease over time, and the process of filling a diary every day can be tedious, so participants may avoid or simply forget to do it. Moreover the concepts of stops and POI are complex and often ambiguous. Finally, any human task is error-prone, and by manually inspecting the diary entries we noticed self-evident errors such as sequence of stops in wrong temporal order and typos in descriptions, dates and times.

Data collection

Location samples are collected with a custom version of the Funf Open Sensing framework [1], which requests one sample every 15 minutes using the Android location API. We have indications that these settings do not significantly impact power consumption, since participants did not report reduced battery life using our collector app. The location is provided either by GPS positioning, Wi-Fi or cellular networks, depending on availability. Location is acquired in an opportunistic manner, so every time another app requests a location, this sample is also recorded in our system. The collected data is temporarily stored on the phone, and periodically uploaded to our servers.

This sampling method presents several challenges, including the variability of accuracy, the unpredictability of samples arrival, and the presence of outliers and duplicates. We firstly investigate the quality of the raw location data. We calculate the empirical cumulative distribution function (CDF) for the samples accuracy, and we find that the vast majority ($> 90\%$) of samples have accuracy better than 60 meters.

We then analyze the time distribution of samples. We calculate the CDF of the time in seconds between samples $\Delta t = timestamp_{i+1} - timestamp_i$. Figure 1 shows that around 60% of the intervals of time between samples are under 10 seconds, 80% are under 60 sec, and 90% are under 200 seconds.
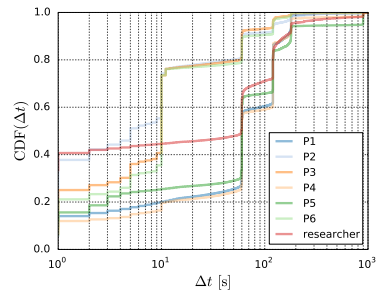

Figure 1: Cumulative Distribution Function for the time between samples Δt. Around 60% of the intervals of time between samples are under 10 seconds, 80% are under 60 sec, and 90% are under 200 seconds.

Even though many samples are collected with high frequency, they are unevenly distributed in time. In fact the opportunistic sensing settings produce a pattern of burst collection, where a rapid sequence of samples is collected, and then none for a longer period of time. To investigate what is the probability of finding at least one sample in a given time interval, we divide the samples into time bins, and count what fraction of bins are empty for different bin sizes. Figure 2 shows that the fraction of empty bins decreases for larger bin sizes, with a large fraction of empty bins for smaller sizes. Starting from 900 seconds the fraction of empty bins becomes approximately constant.

Figure 2: Fraction of empty time bins versus time bin size. The fraction of empty bins decreases for larger bin sizes, with a large fraction of empty bins for smaller sizes. Starting from 900 seconds the fraction of empty bins becomes approximately constant.
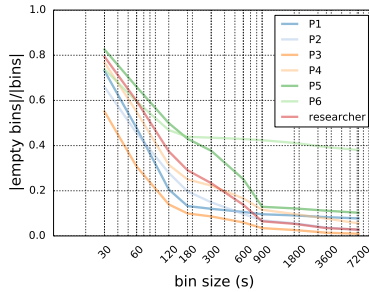
This shows that in most cases the samples collected opportunistically tend to be highly redundant as they capture approximately the same location in a very short timespan.

Stops detection algorithms

In this section we describe three simple algorithms for extracting stops from location data, based on methods from the literature: distance grouping, speed thresholding and Gaussian Mixtures Model (GMM) clustering.

Each algorithm accepts as input a sequence of location samples and produces as output a sequence of stops. We discard stops with duration shorter than a minimum time, since very short stops may be inferred from the location data but they are not meaningful in human terms. Although the choice of what is the minimum time duration of a stop is quite subjective, we chose the minimum time as 15 minutes, since the very large majority of diaries entries have duration greater than 30 minutes.

Each algorithm has parameters that influence the stop extraction, and ultimately determine the $f1$ performance. We are interested in how estimating the parameters on a set of participants would generalize to all others, therefore we perform cross-validation. We select the parameters that perform best for one participant and we calculate the scores for the remaining participants using these parameters. We repeat the operation for a each participant, and we calculate the average $f1$. This procedure is used for each algorithm.

Distance grouping
The distance grouping algorithm is built on the idea that a stop corresponds to a temporal sequence of locations within a maximal distance d_{max} from each other. Locations are examined sequentially by non-decreasing timestamp. Each stop initially contains only one location loc_i, and each subsequent location loc_{i+k} is added to the

stop until $distance(loc_{i+k}, loc_i) < d_{max}$. Then the process starts again from loc_{i+k+1}. The choice of the parameter d_{max} influences the detection of the stops. A large d_{max} tends to merge more location samples and produce less stops of longer duration, while a smaller d_{max} tends to merge less locations and produce more stops of shorter duration. An initial increase in d_{max} leads to better $f1$ performance since smaller stops are correctly merged into more meaningful larger stops. For a range of values the $f1$ score stabilizes around an optimal value, as increasing d_{max} does not result in merging more stops. Finally for larger values of d_{max}, even very far stops are merged together, resulting in worse performance shown by the lower $f1$ score.

Speed thresholding
Speed thresholding uses the calculated speed of movement to classify location samples as stops and moves. Given two successive geographical positions we calculate $speed_{i+1} = \frac{distance(pos_{i+1}, pos_i)}{timestamp_{i+1} - timestamp_i}$. In our dataset however the sample-by-sample speed oscillates widely, due to the variability of the frequency and accuracy of samples. We are instead interested in a smoother speed estimate, in order to detect transitions between places. Therefore we create time bins of size T, and for each bin we consider the position as the median of all the samples in the bin. We then calculate the speed between bins. Using the speed information, we can discard samples with $speed > speed_{max}$, and then group consecutive static locations into stops. The choice of the bin size T and of the $speed_{max}$ threshold influence the performance of the algorithm. A longer period produces smoother transitions but may fail to detect shorter stops. A lower $speed_{max}$ threshold will result in discarding a larger number of location samples and consequently to more information loss.

Gaussian Mixtures Model
The distance grouping and the speed thresholding algorithms take a sequential approach to the stops detection, as they examine samples one by one in temporal order. A different approach is to look at the overall distribution of the samples independently of time, and identify POI as clusters of location samples with higher density. A Gaussian Mixtures Model (GMM) assigns samples to clusters modeled as a finite number of Gaussian distributions with unknown parameters. Once each sample is assigned to a cluster, we group temporally consecutive samples with the same cluster label into stops. The condition that determines the cluster assignment is the *min_covariance* parameter. A larger *min_covariance* tends to produce larger clusters, while a smaller *min_covariance* produces smaller clusters.

Results and Discussion
In this section we provide the results of the $f1_{stops}$ and $f1_{POI}$ scores. Since our work focuses on sparse opportunistic sampling, we cannot compare directly with results in the existing literature, which are based on data sets with high-frequency, high-precision sampling. Additionally, most studies provide either an informal performance assessment based on aggregated data [6, 13, 5, 17], or limit their evaluations to the overall identification of POI [20, 21, 19, 18, 11]. Here, we directly evaluate how well we can infer the full stop-by-stop sequence. This is a more challenging task, which shifts the focus to evaluating the feasibility of inferring mobility.

Evaluation of stop extraction
We apply the three algorithms for stops extraction for each participant (P1-P6) and for the researcher (R), and we obtain the $f1_{stops}$ scores (Figure 3).
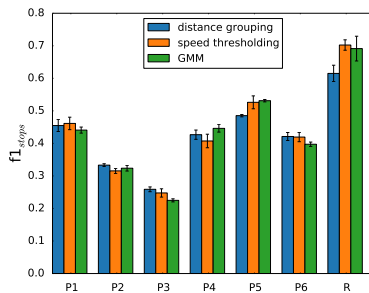
Figure 3: $f1_{stops}$ scores for each participant (P1-P6) and for the researcher (R).

For each participant, the three algorithm have a substantially similar performance. The scores suggest that it is possible to infer mobility patterns with reasonable accuracy, despite the very sparse nature of the collected data.

There is an evident difference between the researcher and participants, with the researcher's score being much higher than any participants' scores. To estimate an upper limit of how well diaries can be matched against the location data, we compare the sequence of locations inferred from the data against the sequence inferred from the diary. In order to generate a sequence of positions from the location data we create time bins of size 900 seconds and take the median position for each bin, while for each diary entry we generate one position at the corresponding POI for every 900 seconds. We compare the two sequences as follows: for each time bin we count a match if the distance

between the diary position and the inferred position is $<$ 100 meters. The we calculate an overlap score as: $overlap = \frac{|matches|}{|bins|}$ We find a significant correlation between $f1_{stops}$ score and $overlap$ ($\rho_{distance} = 0.939$, $\rho_{speed} = 0.944$ and $\rho_{GMM} = 0.941$). We suggest that the higher $f1_{stops}$ score for the researcher is probably due to a better quality of the ground truth. The researcher's data was curated to be as precise as possible and record every single stop, while it is quite likely that participants have not been as consistent in their diaries.

Evaluation of POI extraction
The stop extraction algorithms produce a sequence of stops, which can then be grouped into clusters corresponding to POI. In the case of GMM, the stops are automatically marked with a *poi-id* associated with the corresponding cluster. The distance grouping and speed thresholding algorithms instead do not provide any information about groups of stops. In this case we infer the POI membership for stops applying DBSCAN clustering to the stops extracted by distance grouping. We set $minpts = 1$ and we determine the optimal eps by cross-validation, selecting the parameter value that performs best for one participant and calculating the $f1_{POI}$ scores for the remaining participants using this value.

Figure 4 shows the $f1_{POI}$ for the GMM and the DBSCAN algorithms. The results are consistent with our findings about $f1_{stops}$. The two algorithms have similar performance for the same participant, but there is a quite significant difference between participants' and researcher's scores. Also in this case there is a correlation with overlap, which again suggests that the lower participants score is due to inaccurate ground truth. Finally if we compare the $f1_{stops}$ and $f1_{POI}$ for each

participant, we see that the $f1_{POI}$ is better, which is probably because it is easier to infer POI than stops.
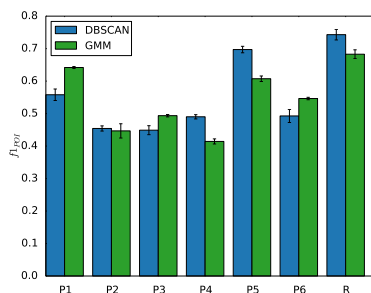

Figure 4: Comparison of the $f1_{POI}$ for the GMM and the DBSCAN algorithms.

Mobility analysis
In this section we further validate the discovery of stops and POI by performing an analysis of the mobility patterns. We consider the results obtained using the GMM algorithm, although similar results are obtained using the other algorithms.

We first look at the geographical network of the 20 POI where most time is spent. We draw a node to represent the geographical coordinates of each stop at a POI, and one arc for each transition between stops. We find that the two networks are quite similar, although some POI are notably missing from the discovered network. Figure 5 shows the comparison of the geographical network as inferred from the diary (left) versus the one from the discovered stops (right) for one participant.
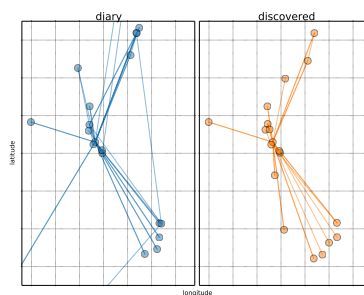
Figure 5: Comparison of the geographical network from the diary (left) versus the one from the discovered stops (right) for one participant.

We then look at the distribution of time versus POI. It is common knowledge that we spend the most of our lives at home and at work, and rest the of our time is shared between all other locations, some of which are regularly visited. Several studies have confirmed this notion, as they found that the distribution of time spent at different locations is heavy-tailed [7]. For each participant, we compute the time spent at POI both from the diary stops and from the discovered stops. The two time distributions are remarkably close, and both show heavy-tailed distributions, with the 2 most visited POI (presumably home and work) corresponding to the majority of time. Figure 6 shows as example the comparison for one participant.

Figure 6: Comparison of the time spent at POI as inferred from the diary and from the discovered stops for one participant.
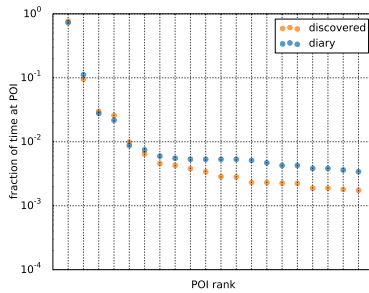
Finally, we calculate the number of unique POI over time. For each $timestamp_i$ we count the unique POI visited by the stops between $timestamp_0$ and $timestamp_i$. We find that for all participants the two distributions follow a similar trend, although in some periods the discovered POI are fewer than the POI in the diary.

Conclusions

We have evaluated the feasibility of inferring human mobility from sparse low accuracy mobile sensing data. We have performed a comparison with ground truth in form of location diaries, and evaluated in details the inferred mobility in terms of geographical networks, time spent at different places, and the number of unique places over time. Our results indicate that it is possible to infer the mobility patterns accurately, despite the sparse nature of the data. As future work, different algorithms could be tested to see if the detection performance can be improved. We also find a significant difference in performance between the researcher' and the participants' scores, and we suggest that it may be caused by the inaccuracies in the participants' diaries which seem to be a limiting factor for accurately studying human mobility.

Having indication of the feasibility of inferring mobility, we intend to apply these procedures to the dataset of nearly one thousand participants that we are actively collecting [16]. We believe that replacing raw location data with meaningful points of interest will allow us to understand human mobility at a level much closer to the human point of view.

References
[1] Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing 7*, 6 (2011), 643–659.

[2] Ashbrook, D., and Starner, T. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing 7*, 5 (2003), 275–286.

[3] Bagrow, J. P., and Lin, Y.-R. Mesoscopic structure and social aspects of human mobility. *PLoS One 7*, 5 (2012).

[4] Belik, V., Geisel, T., and Brockmann, D. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X 1*, 1 (2011), 011001.

[5] Boytsov, A., Zaslavsky, A., and Abdallah, Z. Where

have you been? using location clustering and context awareness to understand places of interest. In *Internet of Things, Smart Spaces, and Next Generation Networking*. Springer, 2012, 51–62.

[6] Cao, X., Cong, G., and Jensen, C. S. Mining significant semantic locations from GPS data. *Proc. of the VLDB Endowment 3*, 1-2 (2010), 1009–1020.

[7] Gonzalez, M. C., Hidalgo, C. A., and Barabási, A.-L. Understanding individual human mobility patterns. *Nature 453*, 7196 (2008), 779–782.

[8] Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment 2009*, 07 (2009), L07003.

[9] Lu, H., Yang, J., Liu, Z., Lane, N. D., Choudhury, T., and Campbell, A. T. The Jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, ACM (2010), 71–84.

[10] Marmasse, N., and Schmandt, C. Location-aware information delivery with ComMotion. In *Handheld and Ubiquitous Computing*, Springer (2000), 157–171.

[11] Montoliu, R., and Gatica-Perez, D. Discovering human places of interest from multimodal mobile phone data. In *Proc. of the 9th Int. Conf. on Mobile and Ubiquitous Multimedia*, ACM (2010), 12.

[12] Onnela, J.-P., Arbesman, S., González, M. C., Barabási, A.-L., and Christakis, N. A. Geographic constraints on social network groups. *PLoS one 6*, 4 (2011), e16939.

[13] Palma, A. T., Bogorny, V., Kuijpers, B., and Alvares, L. O. A clustering-based approach for discovering interesting places in trajectories. In *Proc. of the 2008 ACM symposium on Applied computing*, ACM

(2008), 863–868.

[14] Raento, M., Oulasvirta, A., and Eagle, N. Smartphones an emerging tool for social scientists. *Sociological methods & research 37*, 3 (2009), 426–454.

[15] Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. Limits of predictability in human mobility. *Science 327*, 5968 (2010), 1018–1021.

[16] Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., and Lehmann, S. Measuring large-scale social networks with high resolution. *PLoS One 9*, 4 (04 2014), e95978.

[17] Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., and Aberer, K. Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology (TIST) 4*, 3 (2013), 49.

[18] Zheng, V. W., Zheng, Y., Xie, X., and Yang, Q. Collaborative location and activity recommendations with GPS history data. In *Proc. of the 19th Int. Conf. on World Wide Web*, ACM (2010), 1029–1038.

[19] Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. Mining interesting locations and travel sequences from GPS trajectories. In *Proc. of the 18th Int. Conf. on World Wide Web*, ACM (2009), 791–800.

[20] Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L. Discovering personal gazetteers: an interactive clustering approach. In *Proc. of the 12th annual ACM Int. workshop on Geographic information systems*, ACM (2004), 266–273.

[21] Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L. Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS) 25*, 3 (2007), 12.

Measuring Human Mobility at Multiple Scales

Andrea Cuttone, Marta C. González, Jakob Eg Larsen, Sune Lehmann.

Measuring Human Mobility at Multiple Scales

Andrea Cuttone, Marta C. González, Jakob Eg Larsen, Sune Lehmann

Understanding human mobility is of fundamental importance for many applications, including traffic management and urban planning[1–3], epidemics containment[4–8] and disaster response[9,10]. Some of the key results in the human mobility field are the high degree of predictability[11] and periodicity[12] of movement patterns, the heavy-tailed staying times[13] and the Zipf's law of frequency versus place rank[14], which have been explained with the mechanism of preferential return[15]. Moreover human movements have been described by scale-free random walks known as Lévy flights[13]. Despite many of the mobility properties are scale-free, human understanding of mobility is based on multiple discrete scales within a hierarchical organization, such as buildings, neighborhood and cities. Here we develop a model capturing the hierarchical organization of human mobility, and measure mobility properties at multiple scales. We measure a number the heavy-tailed staying time, the daily and weekly periodic behavior, the exploration patterns, and the navigation in the spatial hierarchy. Our results provide a new view of human mobility properties, and show the importance of considering the scales of human mobility.

1 Introduction

Human mobility patterns have been described by scale-free random walks known as Lévy flights[13], and human diffusion have been characterized by a heavy-tailed radius of gyration[14]. Despite mobility properties are typically scale-free, human understanding of mobility is based on multiple

discrete scales within a hierarchical organization, for example in terms of buildings within cities. The contrast between the two views has generated a large amount of work trying to explain the origin of the heavy-tailed distribution for human properties. For example it has been shown that the power-law behavior may emerge due to a time-dependent activity rate[16], the mixture of multiple individual rates[17] and behaviors[18], and the mixture of different transportation modes[19,20].

The concept of hierarchical and multi-scale organization of human mobility has been proposed in several fields: in Geography, the Central Place Theory[21] assumes the existence of a hierarchy of spatial regions to explain the distribution of human settlements; in Environmental Psychology the Bio-Ecological Systems Theory[22] suggests the existence of nested spatial structures; in Spatial Analysis the Modifiable Area Unit Problem (MAUP)[23] warns about the influence of the size of the spatial units of the statistical properties. But surprisingly no model has been developed to capture such hierarchical organization and measure the emerging mobility properties at multiple scales. In this work we fill the gap by developing a model which is able to capture the hierarchical organizations of human mobility, and therefore study mobility properties at multiple scales.

2   Method

We analyze a dataset from the Copenhagen Network Study[24], which collected mobile sensing data for more than 800 students at the Technical University of Denmark. In particular we focus on the location data, which is collected with frequency one sample every 15 minutes and has a median accuracy of ≈ 20 meters, with more than 90% of the samples having an accuracy better than 40

meters. The data has been collected mainly in Copenhagen and in Denmark, but covers also many

other countries as participants kept collecting their data also during travels abroad. In order to filter

users with good data quality, we select for each users the longest period with at least one sample

in 90% of the 15-minutes time bins. Moreover we keep only users for which we have at least 3

months of data. We remain with 454 users, which have data for periods ranging from three months

to one year.

We define our model in the following way. For each user individually we consider the se-

quence of location samples in time order. We identify *stops* as sequence of locations where the user

has been approximately stationary, which means that the distance between the location at time t

and $t + \Delta t$ is less than a threshold $\delta = 50$ meters, roughly corresponding to the GPS accuracy. For

each stop we calculate a centroid as the median latitude and longitude of all its location samples,

and a duration equal to the difference between the timestamp of its last and first locations. To

remove brief stops generated when the user is moving, we discard stops with duration less than 15

minutes.

We now want to group stops into *places*, where a place is group of stops representing a

spatial unit at a given scale (e.g. building scale). Therefore we apply a recursive clustering using

the DBSCAN[25] clustering. We apply the clustering with a very large grouping distance ($\epsilon = 150$

kilometers), which produces clusters of very large size, roughly corresponding to country-scale

partitions. For each cluster we then apply the clustering to its members with a smaller grouping

distance ($\epsilon = 5000$ meters), therefore producing sub-clusters of roughly at city scale. We repeat

once again the clustering with $\epsilon = 500$ meters (roughly a neighborhood-size scale) and $\epsilon = 20$ meters, roughly a building-size scale. At the end of this process, each stop is marked as being member of a cluster at four scales: country, city, neighborhood and building. Although the clusters do not necessarily correspond to the exact administrative boundaries, they very often capture well the geographical scales. From now on we will refer to the clusters with their level names.

Fig. 1 represents an example of the model results for one user. Each cluster is depicted as a circle enclosing all its points, and each scale of clusters has a different color: blue for building, green for neighborhoods, orange for cities and red for countries. The leftmost panel shows the several building clusters (blue) within one neighborhood cluster (green) at the Technical University of Denmark campus. The center panel shows the previous neighborhood cluster (green) within a city cluster (orange), and other two city clusters around the Copenhagen metropolitan area. Finally the rightmost panel shows the previous city clusters (orange) within a country cluster (red), together with other two country clusters. Another view of the result of our model is that we can directly reconstruct the whole hierarchy of clusters and sub-clusters which represents the hierarchy of personal mobility in terms of countries, cities, neighborhoods and buildings partitions. From the world-scale point of view, we can measure the division into country clusters, and for each country cluster we can measure the division into cities clusters, and for cities into neighborhoods, and finally neighborhoods into buildings. We now proceed to use our hierarchical model to measure a number of mobility properties.

3 Staying time

To investigate the scaling of staying times and visits, we measure staying time at the four scales. We considered each user individually, and one scale at the time. We calculate the total time at a place as the sum of the individual stops durations at that place. We sort the places by the decreasing total time spent, so that the most visited place has rank 1, the second most visited rank 2 and so on. We then divide the individual total times at places by their sums, to obtain the fraction of time spent at each place. After repeating this operation for each user, for each rank we calculate the average by user thus obtaining the average fraction of time at rank k. We repeat this procedure for each scale separately, so at the end we finally obtain the average fraction of time spend at place rank k for each scale. Fig. 2 displays the results, limited to the top 10 ranks. The figure shows how at each scale there is a sharp decrease of the fraction of time with increasing rank. At all scales, the most visited place captures the vast majority of the time, and the rest of the time is split among the other places. Moreover the scale determines the degree of this decay, with larger scales having a much stronger drop in the fraction of time. For example, the most and second most visited countries capture on average 97.6% and 3.7% of the time respectively, while the most and second most visited building capture on average 63.3% and 11.9% of the time respectively. This result confirms previous findings that staying times are heavy-tailed[13–15], but it also shows how this property holds at multiple scales. One interpretation of this result is that the preferential return to places is strongly influenced by the scale: it is common to have a building where we spend most of the time (home), and another one that we visit a bit less (work), and then the rest. But it is typical to have one country where we spend virtually all of our time, and very rarely spend time in another

country.

4 Cumulative time

A question related to the distribution of time at places is the number of significant places which capture the majority of mobility patterns. In other words, how many places are needed to capture most – say 90% – of your time? In order to investigate this question we start similarly as before. Again we considered each user individually, and one scale at the time. We calculate the total time at a place as the sum of the individual stops durations at that place. We sort the places by the decreasing total time spent, and we divide the individual total times at places by their sums, to obtain the fraction of time spent at each place. This time however we consider the cumulative fraction of time at rank k, which is the total amount of time captured considering places at rank $1, 2, ..., k - 1, k$. For example the cumulative fraction of time at rank 1 is simply equal to the fraction of time for rank 1, while the cumulative fraction of time at rank 2 is the sum of the fraction of time at rank 1 and the fraction of time at rank 2, and so on. We repeat this operation for each user, and for each rank we calculate the average by user thus obtaining the average cumulative fraction of time up to rank k. This procedure is done for each scale separately, so at the end we finally obtain the average cumulative fraction of time spend at place up to rank k for each scale. Fig. 3 displays the results, limited to the top 10 ranks. Looking at the figure we can now answer our original question: how many places are needed to capture 90% of the time? The answer is that on average up to 8 buildings are needed for capturing 90% of time, but only 3-4 neighborhoods, and only 1 city and 1 country are enough.

6

5 Periodicity

We consider another characteristic aspect of human mobility: periodicity[14]. The times of arrival and departure at different places are not random, but follow specific patterns due to the human circadian rhythm and the individual schedules: we return home everyday, go to work during week-days, visit our favorite restaurant in the weekend and so on. In order to characterize the periodicity of mobility behavior, we measure the distribution of the inter-event times between arrivals at the same place. We consider the union of all inter-event times for all places of all users at a given scale, and we repeat this operation for the four scales. We obtain four probability density functions (one for each scale) which describe the probability of returning to the same place after a given interval of time (Fig. 4). The probability functions display two strong periodic patterns: one every 24 hours, and one every 7 days, in agreement with our intuition of human periodicity. Interestingly these periodic patterns are evident at the building scale, but also present at the neighborhood scale, and in minor measure at the city scale, while the pattern breaks down at the country scale due to lack of data.

6 Navigation in the spatial hierarchy

The hierarchical organization of mobility captured by the model can measure details which are lost in a flat model. Consider for example the probability of arriving at any building over the hour of the week. Since we have information about the hierarchical relation between spatial units we can investigate the following question (which would be impossible in a flat model): when are there

transitions between buildings in the same neighborhood, and when are there transitions between different neighborhoods? Fig. 5 shows the probability functions for the two cases. It is evident how the two probabilities differ significantly around 12pm-1pm, when the probability of changing building in the same neighborhood is much higher than changing neighborhoods. This is likely explained by the fact that it is common for the population under study (mainly student) to change buildings within the university neighborhood during lunch break. Although this is an example specific to our population, it illustrates how our model is able to capture novel aspects of mobility behavior.

7   Exploration

Exploration of new places over time is a fundamental part of mobility[15], and so is the number of unique places visited over time. We quantify this aspect of mobility at multiple scales in the following way. We consider again each user individually and one scale at the time. For each user we count the number of unique places visited for each week of the year, and we calculate a baseline exploration as the median over the year. We divide each week value by the baseline exploration thus obtaining a normalized value for each week, and we call this value *exploration score*. We repeat this operation for each user, then we average by user the exploration scores for each week thus obtaining an average exploration score for each week of the year. Finally we repeat this process for each of the four scales obtaining average exploration scores for each week of the year and for each scale (Fig. 6). The result shows how the exploration score changes over time at all scales – especially notable are the peaks in middle October (Autumn break), December (Christmas

and New Year's Eve), February (Winter break), middle of April (Easter) and from July (Summer break). Moreover in the same periods of the year the changes in exploration scores depend on the scale: in particular notice how in February the peak for exploration score is much stronger at the country scale – this is likely because students in our population typically go abroad for ski vacations during that break. Again we show here that some mobility properties details may be lost without considering the right scale.

8 Discussion

The proposed model can measure mobility properties at multiple scales, and demonstrates the universality of many mobility properties: the same patterns apply at multiple scales. Moreover this work raises some questions about the scale-free nature of mobility: are there really scales in human mobility, and if so, can they be detected from the data alone? Although in the current work the scales are somewhat arbitrarily set, and sometimes do not correspond to real administrative boundaries, it is evident that different behaviors happen at different scales.

References

1. Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 89–98, ACM, 2011.

2. J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference*

on Knowledge discovery and data mining, pp. 186–194, ACM, 2012.

3. S. Çolak, A. Lima, and M. C. González, "Understanding congested travel in urban areas," *Nature communications*, vol. 7, 2016.

4. L. Hufnagel, D. Brockmann, and T. Geisel, "Forecast and control of epidemics in a globalized world," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 42, pp. 15124–15129, 2004.

5. V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani, "Modeling the world-wide spread of pandemic influenza: baseline case and containment interventions," *PLoS Med*, vol. 4, no. 1, p. e13, 2007.

6. S. Ni and W. Weng, "Impact of travel patterns on epidemic dynamics in heterogeneous spatial metapopulation networks," *Physical review E*, vol. 79, no. 1, p. 016111, 2009.

7. D. Balcan and A. Vespignani, "Phase transitions in contagion processes mediated by recurrent mobility patterns," *Nature physics*, vol. 7, no. 7, pp. 581–586, 2011.

8. V. Belik, T. Geisel, and D. Brockmann, "Natural human mobility patterns and spatial spread of infectious diseases," *Physical Review X*, vol. 1, no. 1, p. 011001, 2011.

9. J. P. Bagrow, D. Wang, and A.-L. Barabasi, "Collective response of human populations to large-scale emergencies," *PloS one*, vol. 6, no. 3, p. e17680, 2011.

10. X. Lu, L. Bengtsson, and P. Holme, "Predictability of population displacement after the 2010 haiti earthquake," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11576–11581, 2012.

11. C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

12. C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130246, 2013.

13. D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

14. M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

15. C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

16. R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. Amaral, "A poissonian explanation for heavy tails in e-mail communication," *Proceedings of the National Academy of Sciences*, vol. 105, no. 47, pp. 18153–18158, 2008.

17. C. A. Hidalgo R, "Conditions for the emergence of scaling in the inter-event time of uncorrelated and seasonal systems," *Physica A: Statistical Mechanics and its Applications*, vol. 369, no. 2, pp. 877–883, 2006.

18. X.-Y. Yan, X.-P. Han, B.-H. Wang, and T. Zhou, "Diversity of individual mobility patterns and emergence of aggregated scaling laws," *Scientific reports*, vol. 3, 2013.

19. K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, "Explaining the power-law distribution of human mobility through transportation modality decomposition," *Scientific reports*, vol. 5, 2015.

20. R. Gallotti, A. Bazzani, S. Rambaldi, and M. Barthelemy, "How transportation hierarchy shapes human mobility," *arXiv preprint arXiv:1509.03752*, 2015.

21. L. J. King, "Central place theory," 1985.

22. U. Bronfenbrenner and P. A. Morris, "The bioecological model of human development," *Handbook of child psychology*, 2006.

23. S. Openshaw and S. Openshaw, "The modifiable areal unit problem," Geo Abstracts University of East Anglia, 1984.

24. A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann, "Measuring large-scale social networks with high resolution," *PloS one*, vol. 9, no. 4, p. e95978, 2014.

25. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.
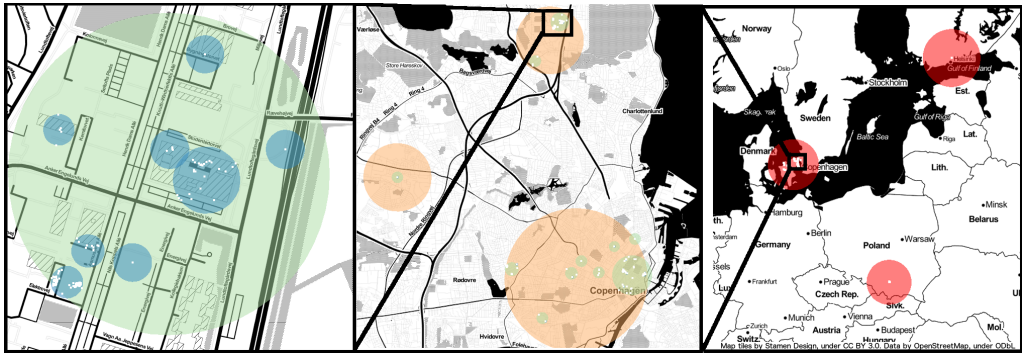
Figure 1: An illustration of the hierarchal spatial structure identified by the model. Each cluster is represented by a circle, and each color represents a different scale. The leftmost panel shows several building clusters (blue) within a neighborhood cluster (green). The center panel shows the previous neighborhood cluster within one of the city clusters (orange). The right panel shows the city clusters within one of the country clusters (red).
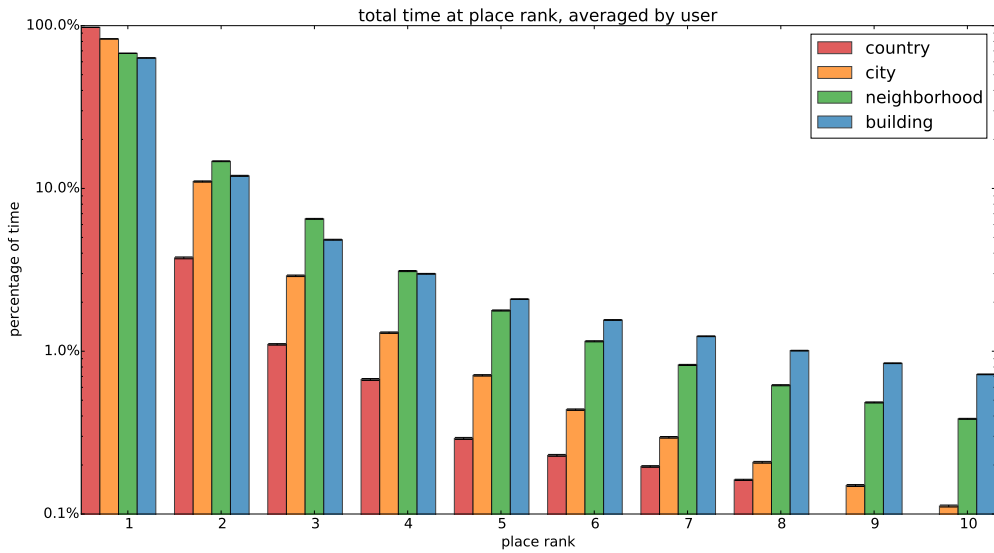
Figure 2: The fraction of time spent at places for different ranks and scales. The fraction of time rapidly decays as a function of rank, and this decay is stronger for larger scales. For example the difference in fraction of time between the first and second top country is much larger than between the first and second top building.
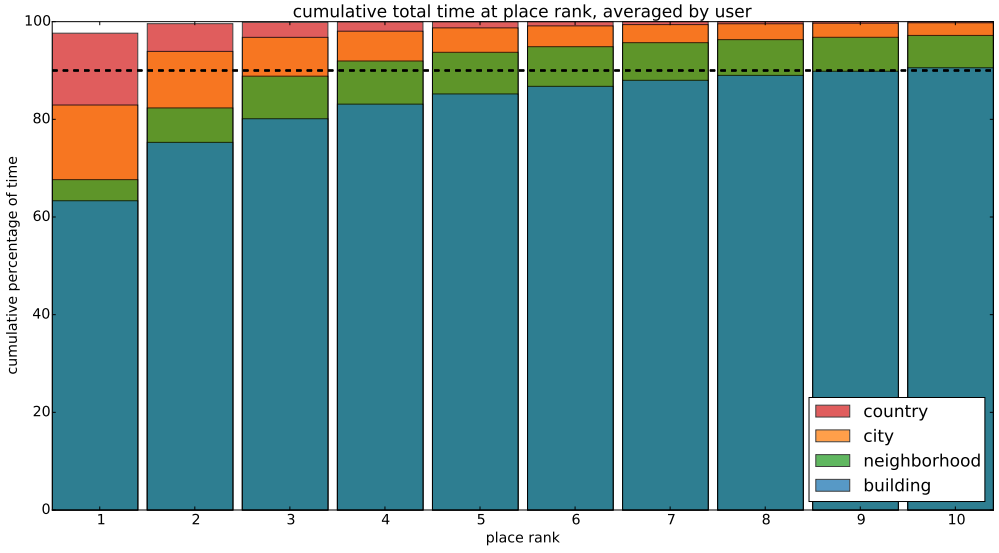
Figure 3: The cumulative fraction of time spent for different ranks and scales (the black line at 90% is a guide for the eyes). It is possible to see how many places are needed to capture the majority of the time: on average up to 8 buildings are needed for capturing 90% of time, but only 3-4 neighborhoods, and only 1 city and 1 country are enough.
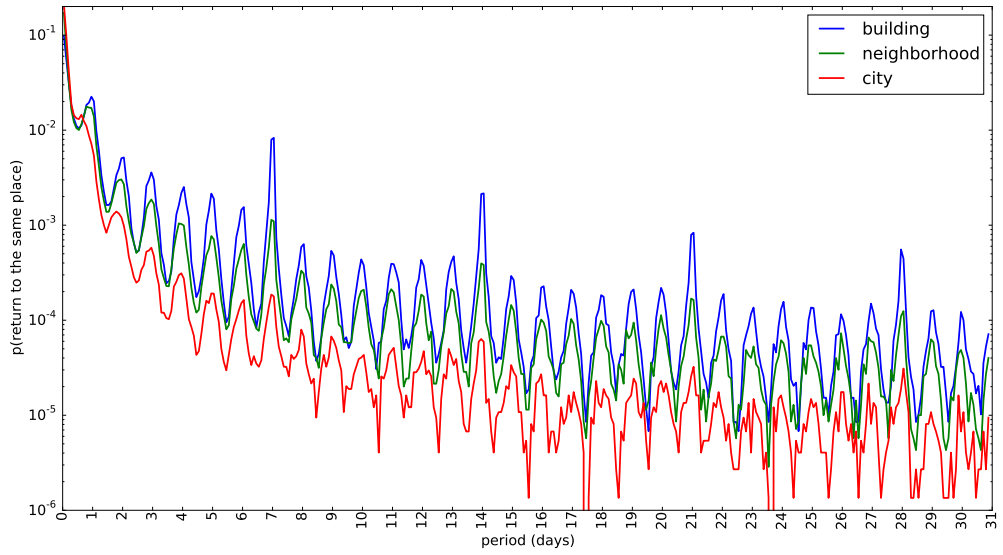
Figure 4: The probability of returning to the same place. It is possible to see two strong periodic trends: one every 24 hours and one every 7 days.
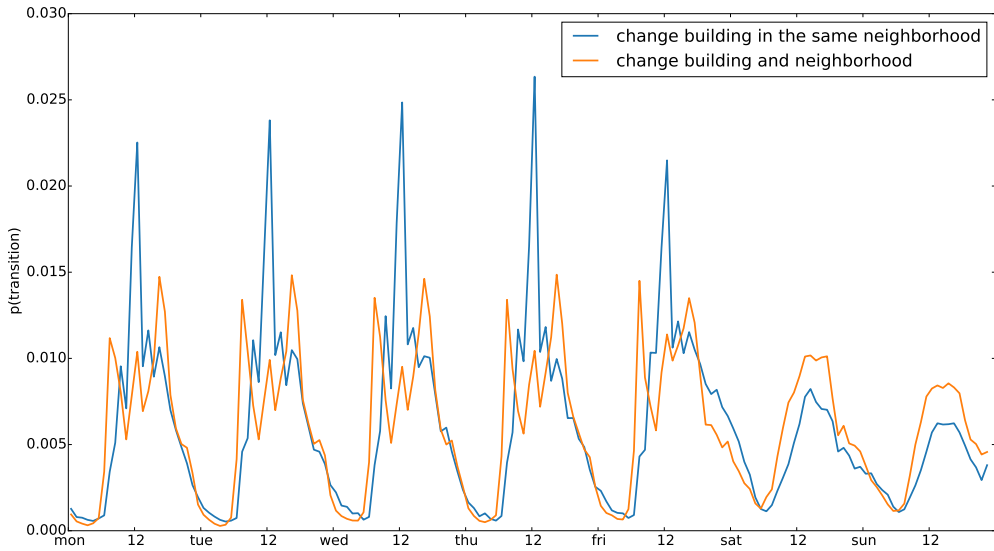
Figure 5: The hierarchical model is able to answer complex questions which would be impossible in a flat model. In this example we compute the probability of changing building within the same neighborhood versus the probability of changing neighborhood (and building). The peak between 12pm-1pm is probably because typically students need to change building within the university campus for lunch break.
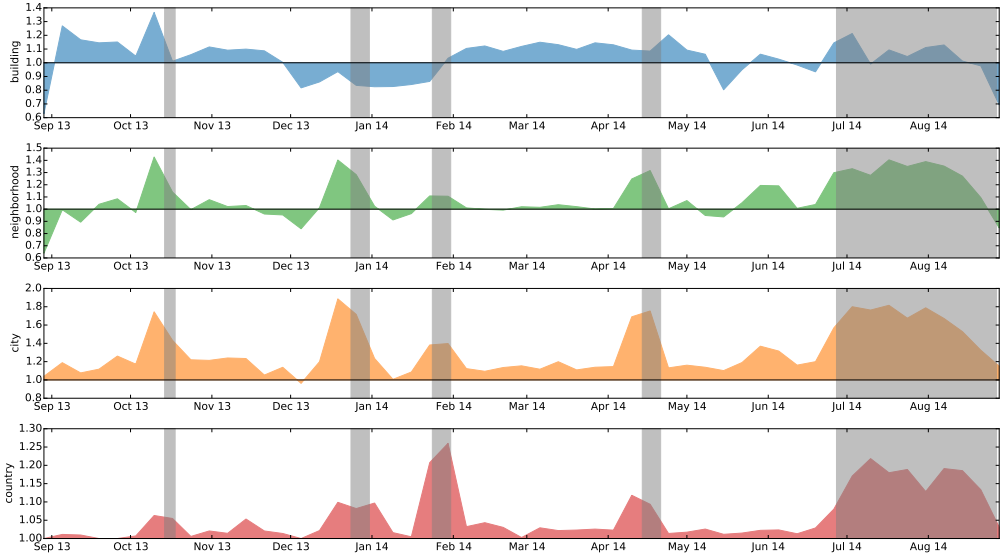
Figure 6: A quantification of exploratory behavior over the year. For each week of the year we compute an *exploration score* as the number of visited places normalized by the yearly median. We repeat this operation for every scale. We can measure how some periods of the year are more exploratory than others, in particular around holidays such as Christmas, Easter, winter and summer breaks.

Understanding Predictability and Exploration in Human Mobility

Andrea Cuttone, Sune Lehmann, Marta C. González.

Understanding Predictability and Exploration in Human Mobility

Andrea Cuttone[1*], Sune Lehmann[1,2], Marta C. González[3]

1 DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark
2 The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
3 Department of Civil and Environmental Engineering and Engineering Systems, Massachusetts Institute of Technology, Cambridge, 02139 Massachusetts, United States of America

* ancu@dtu.dk

Abstract

Predictive models for human mobility have important applications in many fields such as traffic control, ubiquitous computing and contextual advertisement. The predictive performance of models in literature varies quite broadly, from as high as 93% to as low as under 40%. In this work we investigate which factors influence the accuracy of next-place prediction, using a high-precision location dataset of more than 400 users for periods between 3 months and one year. We show that it is easier to achieve high accuracy when predicting the time-bin location than when predicting the next place. Moreover we demonstrate how the temporal and spatial resolution of the data can have strong influence on the accuracy of prediction. Finally we uncover that the exploration of new locations is an important factor in human mobility, and we measure that on average 20-25% of transitions are to new places, and approx. 70% of locations are visited only once. We discuss how these mechanisms are important factors limiting our ability to predict human mobility.

Introduction

Billions of personal devices, ranging from in-car GPS to mobile phones and fitness bracelets, connect us to the cloud. These ubiquitous interconnections of the physical and the digital world are opening up a host of new opportunities for predictive mobility models. Each user of these devices produces rich information that can help us to capture their daily mobility routine. This core knowledge, when obtained from massive number of individuals, impacts a wide range of areas such as health monitoring [1], ubiquitous computing [2,3], disaster response [4] or smart traffic management [5].

In the age of ubiquitous computing, recent contributions to mobility modeling have flourished in computer science [6–8], transportation engineering [9,10], geographic information sciences [11,12], and complexity sciences [13–15]. While these findings have enhanced our level of understanding of mobility modeling we need further work to tackle the problem of individual predictability.

Human mobility has been studied using a multitude of proxies (for example Call Detail Records (CDR), GPS, WiFi, travel surveys), and a variety of techniques have been suggested for predictive models, such as Markov chains, Naive Bayes, artificial neural networks, time series analysis. Studies report varying results for the predictive

power of these models, with accuracy as high as 93% and as low as under 40%. In this paper we set out to uncover the reasons behind these differences in performance by a thorough investigation of the factors that may influence an estimation of mobility predictability. The key contributions of this paper are:

1. We describe the factors that have influenced the various ranges when estimating predictability. These include: (a) Does the analysis concern the upper limit of predictability, or actual next-place prediction? (b) What is the specific formulation of the prediction problem? E.g. is the goal to predict the next location, or is the goal to identify location in the next time-bin? (c) What is the spatial resolution? E.g. is the analysis based on GPS vs. CDR data? (d) What is the temporal resolution e.g. minutes, hours?

2. We quantify the amount of explorations and locations visited only once, and show that these are key limiting factors in the accuracy of predictions for individual mobility.

3. We measure the predictive power of a number of contextual features (e.g. social proximity, time, call/SMS).

4. We study the problem of predictability of human mobility using a novel, longitudinal, high-precision location dataset for more than 400 users.

The rest of the paper is organized as follows. We first provide an overview of related work in the field of human mobility prediction. Next, we introduce the dataset and describe the preprocessing steps. In the subsequent section we describe the baseline models, and compare their performances. Finally we introduce the exploration prediction problem and report the performance of the predictive models.

Related work

In a seminal paper Song *et al.* [13] investigate the limits of predictability of human mobility, using Call Detail Records (CDR) as proxy for human movement. In their analysis, the authors discretize location into a sequence of places, and estimate an upper limit for the predictive performance using Fano's inequality on the temporal entropy of visits. Their results show that for a majority of users, this upper bound is surprisingly high (93%). This framework has been further explored to refine the upper limit. Specifically, Lin et al. [16] study the effects of spatial and temporal resolution on the predictability limit, Smith et al. [17] consider the spatial reachability constraints when selecting the next place to visit, and obtain a tighter upper bound of 81-85%, and Lu et al. [4] analyze the predictability of the population of Haiti after the earthquake in 2010, and show that the upper limit of predictability remains as high as 85%.

The work described above focuses on the upper limit of predictability based on estimating the entropy of trajectories. When the topic is actual prediction performance, some of the most studied models are Markov chains, where the probability of the next location is assumed to depend only on the current location. Markov chains have been applied to a variety of data sets. Lu et al. [18] applied Markov chain models to CDR-based locations in Cote D'Ivore, with a prediction goal of estimating the last location of the day at the prefecture (county) level. Under these conditions the models perform extremely well, reaching an accuracy of over 90%. In [19] the authors apply the Markov models to WiFi traces at Darthmouth campus and find that the best performing model is order 2 and has a median accuracy of about $65 - 72\%$. Finally, Bapierre et al. [20] applied a variable-order Markov chain to the Reality Mining [6] and Geolife [21] datasets.

Another frequently used category of models is naive Bayes, where the probability of next location is factorized as independent probabilities for a number of context variables. Gao et al. [22] applied this approach to the Nokia Data Challenge dataset [23] using time and location features, and obtained an accuracy of approximately 50%. Do et al. [24] applied the same technique but used a larger number of features including also SMS, calls and Bluetooth proximity, and obtained an accuracy of approximately 60%. In a subsequent paper [25] the same authors then explore a kernel density estimation approach for improving performance.

A number of more complex methods have also been explored in the literature, including non-linear time series [26], Principal Component Analysis [27], Gaussian Mixtures [28] and Dynamic Bayesian Networks [29].

While recent work on predictability has resulted in richer methods and incorporated interesting new features such as social contacts, they have not deeply characterized the intrinsic characteristics of human mobility that form the basis for the limitations in predicting the next visited location. In this paper we focus on that aspect, showing that in 53 weeks, individuals visit on average 200 unique locations, of which 70% of them are visited only once. Despite most of the trips being among 30% of their recurrent locations; the occurrence of an exploration can be predicted with at best 41% of accuracy. Separating the two types of visited locations and improving the ways to predict an exploration would advance the methods in this area.

Materials and Methods

Data description

In this study we analyze a dataset from the Copenhagen Network Study [30]. The project has collected mobile sensing data from smartphones for more than 800 students at the Technical University of Denmark (DTU). The data sources include GPS location, Bluetooth, SMS, phone contacts, WiFi, and Facebook friendships. Data collection was approved by the Danish Data Protection Agency, and informed consent has been obtained for all study all participants.

For this study we focus on the location data, which is collected by the smartphone with frequency of one sample every 15 minutes. Each location sample contains a timestamp, a latitude and longitude, and an accuracy value. The location is determined by the best available provider, either GPS or WiFi, with a median accuracy of ≈ 20 meters; more than 90% of the samples are reported to have an accuracy better than 40 meters. For individual participants, there may be periods missing data. These periods can occur for various reasons, for example due to a drained battery, the phone being switched off, the location probe being disabled, or due to software issues. Since we are interested in reconstructing mobility histories without large gaps, we select the longest period that has at least one sample in 90% of the 15-minutes time-bins for each participant. Moreover we consider only participants that have at least 3 months long period of such contiguous data. We are left with 454 users, with data collection periods of data ranging from three months to one year. Fig. 1 shows the distribution of period lengths.

The data is mainly concentrated in Denmark where the study takes place, but because students use the phones during travel, the dataset spans several other countries as well. Fig. 2 shows a map of the locations in the world (left pane) and in Denmark (right pane).

In this work we are interested in the location prediction task. This task can be broadly stated as follows: given your location history, how well can we predict your future location? The specific details of how this question is implemented have a
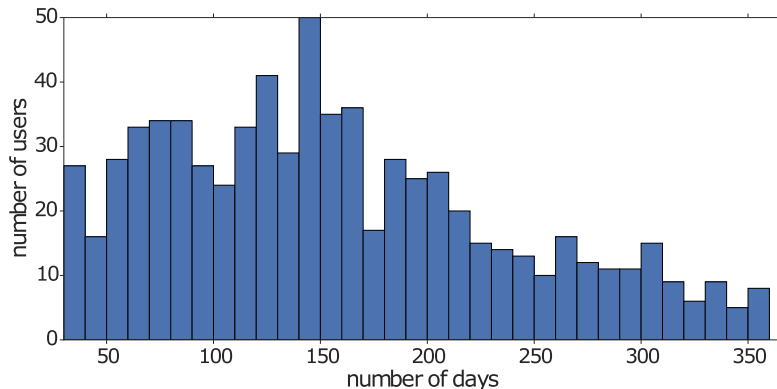
Fig 1. Durations for the periods of collected data for all 454 users. For each user we select the longest period that has at least one sample in 90% of the 15-minutes time-bins.
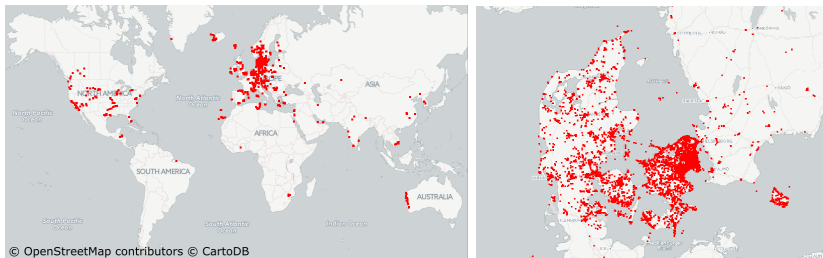

Fig 2. Map of recorded locations across the world (left pane) and in Denmark (right pane). Each red marker corresponds to a location sample.

profound impact on the prediction accuracy. Below we investigate how various factors, e.g. spatial and temporal data resolution play a role in determining the reported accuracy for a single underlying dataset.

Because the prediction task can be stated in many different ways, we start the discussion by analyzing different problem formulations. In terms of spatial prediction it is possible to discretize space in grid cells, Voronoi cells or define places using a clustering method. In terms of temporal prediction we could decide to predict a location in the next time-bin, or within a time horizon, or as the next visited place. In this paper we select two of the most common problem formulations: *next-cell* and *next-place*. In the *next-cell* formulation we discretize space into grid cells, and we predict the cell in the next time-bin. In the *next-place* formulation we detect visits to places and we predict the next visited place. The following sections provide details on the two alternative formulations, and show how each formulation affects the prediction task.

next-cell prediction

In the first problem formulation, we convert geographical coordinates *(lon,lat)* into discrete symbols by placing a uniform grid on the map and retrieving the grid cell id

associated with the coordinates. Specifically, we start by considering a grid of approximate size 50 meters x 50 meters. At each timestep $\Delta t = 15$ minutes, we convert the current *(lon,lat)* into a cell id, therefore producing a sequence of symbols through which we can represent a user's location history. Fig. 3 illustrates the process.

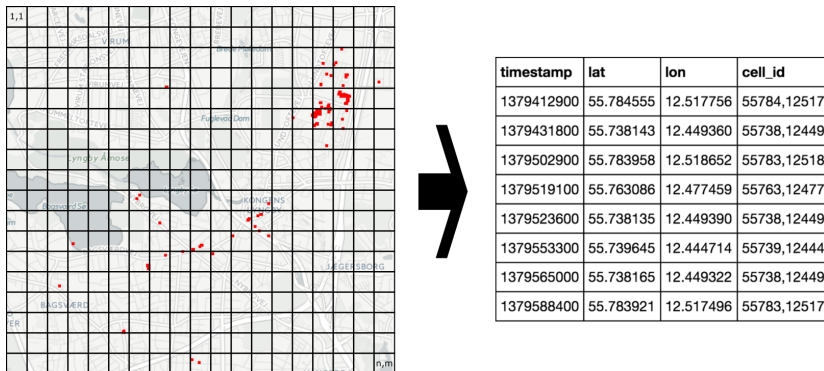| timestamp	lat	lon	cell_id
1379412900	55.784555	12.517756	55784,12517
1379431800	55.738143	12.449360	55738,12449
1379502900	55.783958	12.518652	55783,12518
1379519100	55.763086	12.477459	55763,12477
1379523600	55.738135	12.449390	55738,12449
1379553300	55.739645	12.444714	55739,12444
1379565000	55.738165	12.449322	55738,12449
1379588400	55.783921	12.517496	55783,12517

Fig 3. Process for converting raw geographical coordinates into sequence of grid cells. An approximately uniform grid is placed on the map. For each timestep, the geographical coordinates are converted into the corresponding grid cell ID. The mobility trace becomes the sequence of visited cell IDs.

In this formulation, the problem can be restated as follows: given your past cell sequence up to time t, which cell will you visit at time $t + \Delta t$? Before trying to perform any prediction at all, following the process suggested in [13], we calculate the theoretical upper limit for the predictability of the cells sequence. Fig. 4 shows how the maximum predictability for the grid cell formulation is peaked at around 0.95.

We now consider different baseline strategies for next grid cell prediction. For each of the strategies, we perform prediction in an online manner, by training the algorithm on the data up to timestep t, and predicting cell at timestep $t + \Delta t$. We measure the accuracy as number of correct predictions over the number of total predictions.

We first consider the *toploc* strategy, where at each timestep we predict the most frequent symbol in the history so far. Given the highly stationary nature of most human mobility trajectories, we expect this simple strategy to achieve a relatively high accuracy. Fig. 5 top panel shows the distribution of accuracies for all the users. The accuracy of the *toploc* is indeed reasonable, peaking at around 0.4.

We now consider the Markov chain model. In this model, the prediction of next state depends only on the current state. The transition probabilities between locations are estimated based on past transitions in the location history. For making a prediction, we then consider the transition that has the highest probability among all possible transitions from the current cell. If the current state has never been seen before, then we have no information about the transition probability to other states. In this case we fall back and predict the most frequent state. Again we fit the model in an online manner, updating at each step the transition probabilities and then making a prediction for the next timestep. Fig. 5 middle panel shows the distribution of accuracies for all the users. The accuracy of the *Markov* model is much higher than toploc, peaking at around 0.7.

Considering the highly stationary nature of typical trajectories, we hypothesize that a significant part of the Markov prediction power in this formulation comes from self-transitions, that is, the model predicting the user to remain in the same state as in the previous time-bin. To test this hypothesis, we consider the *stationary* strategy: at
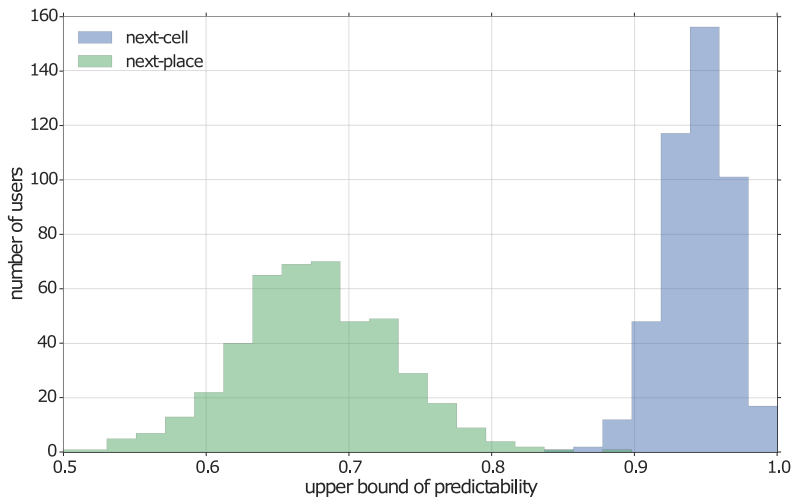
Fig 4. Upper bound of predictability for all users for the next-cell and next-place formulations.
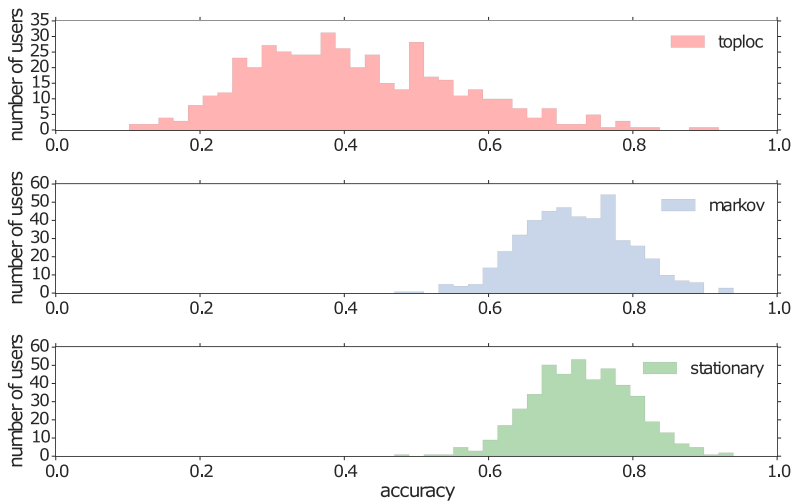
Fig 5. Accuracy of the prediction in the next-cell formulation. The top panel shows the results of the toploc strategy, that is predicting the most common location at each step. The middle panel shows the accuracy for the Markov chain model. The bottom panel shows the accuracy for the stationary strategy, that is predicting remaining in the previous cell.

each step we predict that the user will remain in the current cell. Fig. 5 bottom panel shows that the distribution of accuracies for *stationary* closely matches the one for

Markov. Furthermore Fig. 6 shows how the two are very strongly correlated (Pearson's $r = 0.993$, $p < 0.001$). This strongly suggests that, in this formulation, most of the Markov model power comes from self-transitions, as suspected.

Fig 6. Correlation between the accuracy for the Markov model and the stationary model in the next-cell formulation.

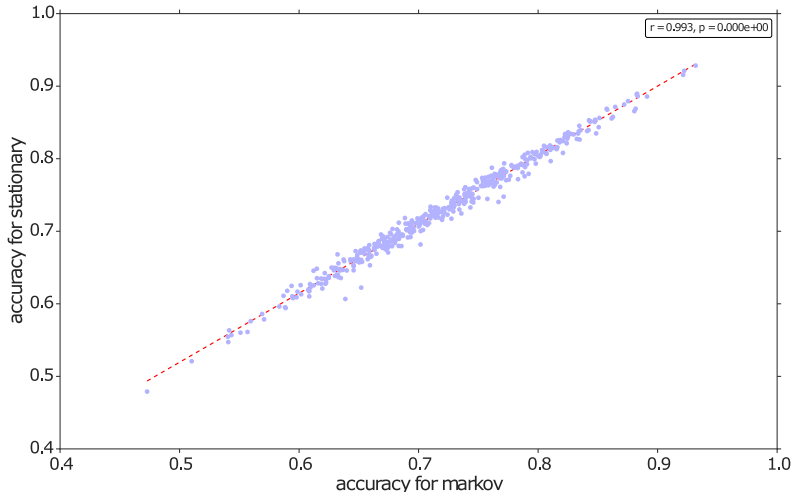We now investigate another issue related to this problem formulation. Intuitively, we expect that the size of our spatial units will influence the accuracy of prediction. Predicting a user's location with the precision of few meters is intuitively much more difficult that predicting with precision of several kilometers. In order to examine the effect of spatial resolution, we also consider results for cell size 500 meters and 5000 meters, and apply the *Markov* model. Fig. 7 compares the accuracy for different spatial resolutions. As expected the accuracy dramatically improves as the spatial size increases.

Finally we investigate the effect of temporal resolution within this problem formulation. Our findings above suggest that using a very fine-grained temporal resolution will increase the number of self-transitions, thus driving up the accuracy of the prediction that is mainly able to capture stationarity. We achieve this by discretizing the location at 50 meters cell size, but varying the temporal time binning to 15 minutes, 30 minutes and 60 minutes, and then running the *Markov* model for each scenario. Fig. 8 compares the accuracy for different temporal resolutions. As expected, the accuracy is decreased as the time-bins grow larger due fewer self-transitions.

next-place prediction

We now consider an alternative problem formulation. Instead of predicting the cell in the next timestep, we want to predict only when we observe a transition between places, eliminating the possiblity of self-transitions. In order to do so, we convert the raw GPS locations into a sequence of stops at places. A large amount of literature has been dedicated to the problem of place detection, such as methods based on WiFi fingerprint [31], grid clustering [32], and kernel density estimation [33].
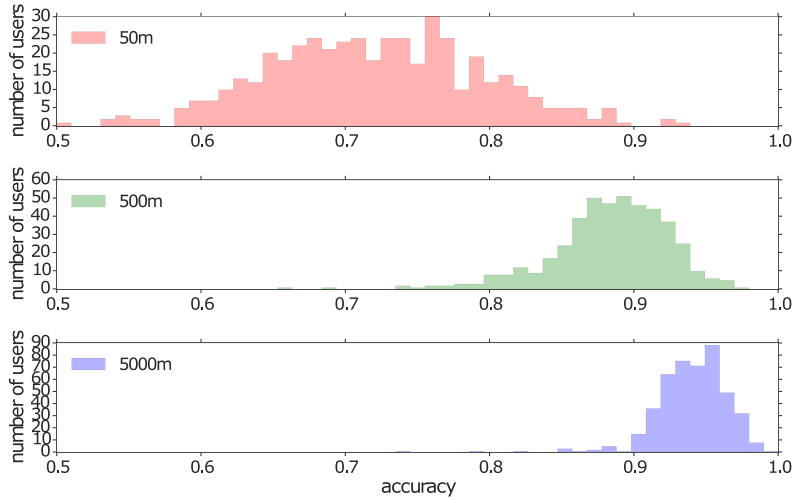
Fig 7. Effect of spatial granularity of the accuracy for the Markov model in the next-cell formulation. Each panel shows the accuracy for a different spatial bin size: 50m, 500m and 5000m. Increasing the size of the spatial bins increases the accuracy prediction.

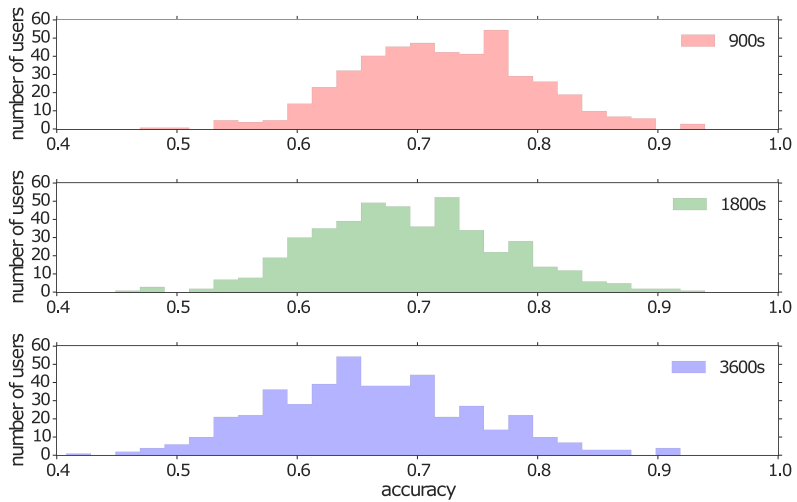**Fig 8.** Effect of temporal sampling of the accuracy for the Markov model in the next-cell formulation. Each panel shows a different temporal bin resolution: 900s, 1800s and 3600s. Decreasing the temporal resolution in this problem formulation increases the accuracy, since there are more self-transitions.

In this paper we consider the following process, based on density-based clustering approaches such as [34–36]. Each user is treated individually. We define a "stop" as sequence of location-observations where the user has been approximately stationary, that is, the distance between position at time t and $t + \Delta t$ is less than a threshold $\delta = 50$ meters, roughly corresponding to the GPS accuracy. This produces a sequence of stops, each one with a centroid calculated as the median of the locations coordinates, and a duration equal to the time between the last location and the first location sample. In order to filter out the short stops during commute, we consider only stops with duration greater than 15 minutes. The average number of stops per user per day is 2.89 with standard deviation 0.89.

We are now interested in grouping stops into places, where a "place" is a group of spatially related stops representing a self-contained area such as a building. In order to do so, we apply the DBSCAN [37] clustering to the stops in the geographical coordinate space, using the haversine distance. We set as parameter the grouping distance $\epsilon = 50$ meters, and $min\_pts = 2$. This distance threshold is set to produce places of the approximate size of a large building. The result of the clustering is an assignment of a cluster label to each stop, where the label represents the place that the stop belongs to. Finally, in order avoid artifacts due to missing samples or noise, we merge multiple consecutive stops at the same place into one. This process converts the raw location history into a sequence of stops at places. Fig. 9 and 10 illustrate the complete process of stop detection.
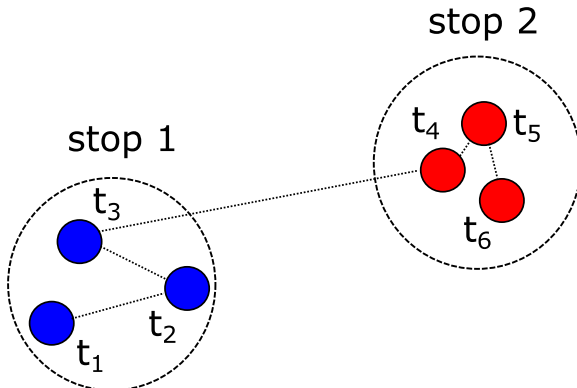
Fig 9. Extraction of stops. The sequence of location samples $t_1, ..., t_6$ are examined sequentially, and are grouped into a stop as long as they are within a distance threshold. In the example, t_1, t_2 and t_3 are assigned to the first stop but t_4 does not, since it too far away. Subsequently t_4, t_5 and t_6 are assigned to stop 2.

As example result of this process, let us consider the stops and places extracted for a user. The sequence of stops at places can be represented as a weekly schedule capturing the user's movement patterns. In Fig. 11 each row represents a week from Monday to Sunday; each place is encoded as a different color. Inspecting this visualization it immediately possible to spot the periodic patters characterizing human mobility, such as evening returns to the home location, and morning trips to class. We can also spot many irregularities however, that deviate from the normal schedule: small stops, new explorations, and day-by-day variability. Finally we can also see a large change in routine starting week 20, where the home location changes. Each user can be characterized by a similar plot.

The prediction task can now be re-formulated as follows: given a sequence of stops

Fig 10. Stops are spatially clustered into places using the DBSCAN clustering algorithm, which groups stops according to their haversine distanc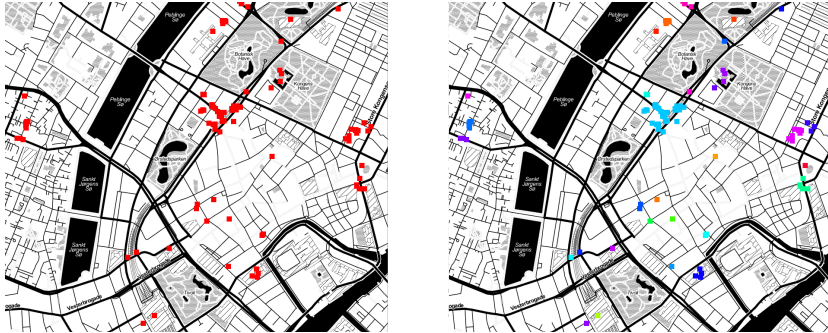e. The left panel shows all stops before the algorithm is run, and the right panel shows the assignment of places labels, with each place represented by a different color.

up to step n, can we predict your next stop at step $n + 1$? Notice that a key difference from the cell grid formulation is that in this case there are (by definition) no self-transitions; we are interested in the place changes only.

As before, we start by investigating the upper predictability limit bound. Fig. 4 shows how the maximum predictability for the stops formulation is peaked at 0.68, significantly lower than what we observe in the grid cells formulation.

We now apply the two prediction strategy *toploc* and *Markov* to this new formulation. The two models remain conceptually the same, but instead of trying to predict the grid cell at each step, they try to predict the next stop (note that in this formulation we cannot use the *stationary* strategy, as by construction we are interested in transitions to new places). In this case we also fit each user separately, and we perform the prediction in an online manner. Fig. 12 shows the accuracy for both models. It is evident that the accuracy for these models (around 0.3 for toploc and 0.4 for Markov) is significantly lower in the next-place formulation, indicating that this problem formulation presents a more difficult task.

Importance of contextual features

We have investigated how the details of the problem formulation strongly impact the reported accuracy for location prediction tasks. We now focus on next-place prediction and study the influence of different contextual features on the prediction task. We formalize the problem as follows. At each step, we want to compute the most probable next location given the current location. We may also want to include other context variables, such as time of the day, day of the week, call activity, or distance from home for example. In other words we want to compute $P(\hat{L}|c_1, c_2, c_3, ..., c_n)$, where $\hat{L}$ is the next location, and c_1, c_2, c_n are the variables representing different contexts. For this purpose we use a logistic regression model, and we study the usefulness of various predictor features. The goal of the model is not to suggest a new state-of-the-art method, but rather to evaluate the importance of individual contextual features. Specifically, we consider the current location, the time metadata (hour of the day, day of the week, hour of the week, weekend), a 'home' binary indication, distance from home, call and SMS activity, and Bluetooth proximity. Table 1 provides a summary of the features.

We model each user separately since we want to perform next-place prediction at the
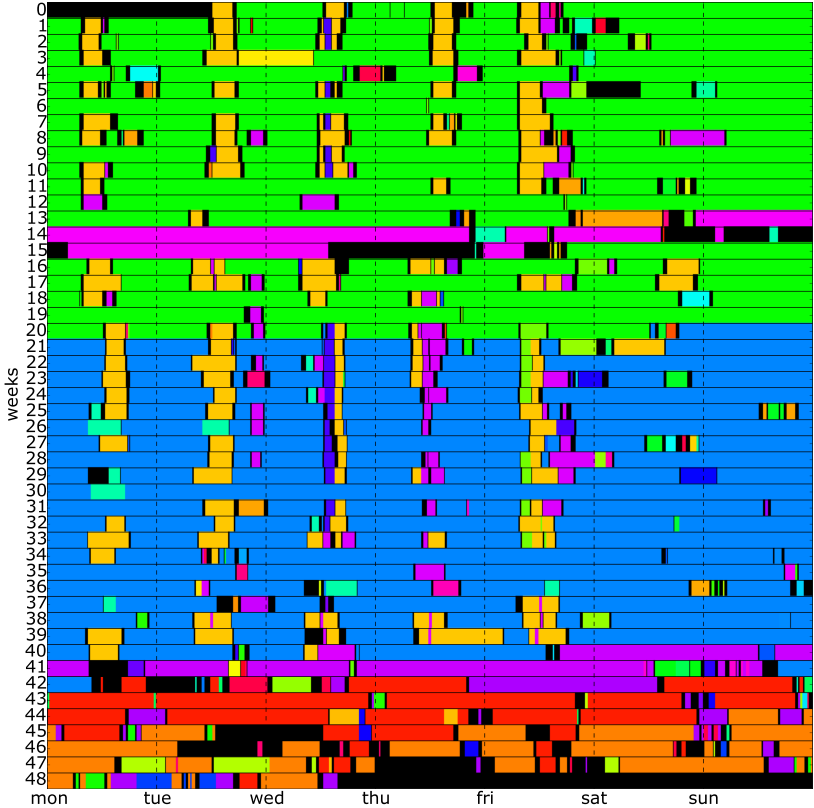
Fig 11. Example of the sequence of stops from one user. Each row represents a week, from Monday to Sunday. Each rectangle represents a stop, and its color encodes the corresponding place. This visualization highlights the complexity of human mobility, with a weekly schedule, periodic returns and irregularities.

individual level. As before, we perform an online prediction where we fit the data up to step n, and we predict the next location at step $n + 1$. For each user, we fit a logistic regression model using all the individual features separately, and then a model with all features. Fig. 13 shows the accuracy for each of the models, averaged by user.

The Markov chain model baseline is highlighted in red. Using the current location and time features, the logistic regression model outperforms the Markov chain based model. Even using the current location only (which is conceptually very similar to a Markov chain model), the logistic regression shows stronger performance, likely due to the explicit optimization of the model. It is also interesting that other context variable such as call and SMS data have little predictive power in this model formulation. The most complex model that considers all features is practically identical in performance to the model using only current location and hour of the week.

Although the logistic regression model does improve the accuracy over the Markov model, the absolute value of accuracy is remains low (below 45%). We therefore investigate the possible reasons of this difficulty in prediction.
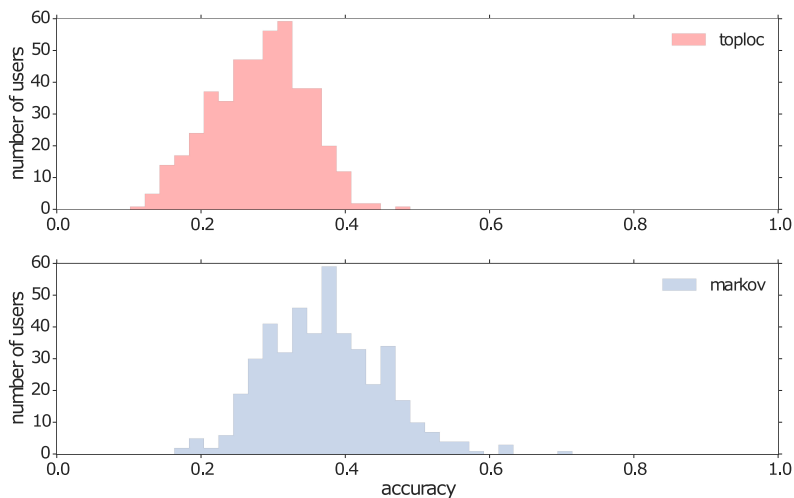
Fig 12. Accuracy for the toploc and Markov models in the next-place formulation. The accuracy in this formulation is considerably lower than in the next-cell formulation.
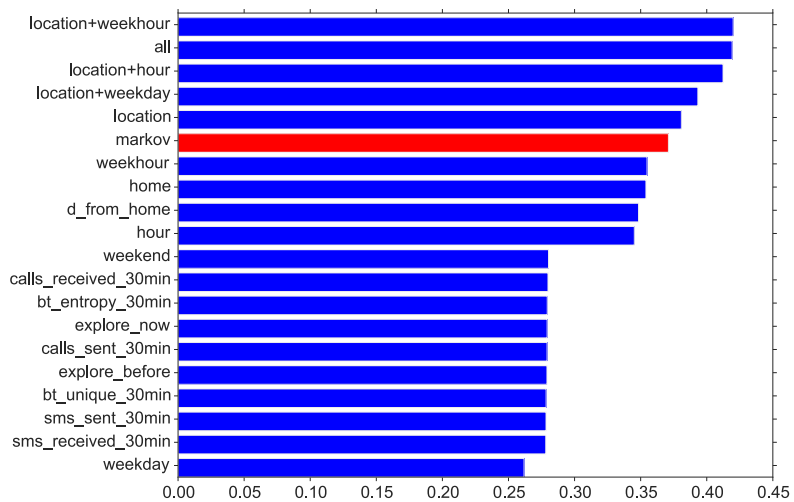
Fig 13. Summary of next-place prediction accuracy for all logistic regression models. The location and time-related features are the most predictive ones, and outperform the Markov model baseline.

Understanding the set of location states

It is well known that the majority of individuals tend to spend most of the times at very few places such as home and work, and only sporadically visit other places. This

| feature name	description
location	location ID
hour	hour of the day (0-23)
weekhour	hour of the week (0-167)
weekday	day of the week (0-6)
weekend	sat/sun (1) or Mon-Fri (0)
explore_before	1 if the previous stop is an exploration, 0 otherwise
explore_now	1 if the current stop is an exploration, 0 otherwise
home	1 if the current stop is at the home location (most visited place), 0 otherwise
d_from_home	distance from the current stop to the home location
sms_received_30min	number of SMS received in the 30 min before the current stop timestamp
sms_sent_30min	number of SMS sent in the 30 min before the current stop timestamp
calls_received_30min	number of phone calls received in the 30 min before the current stop timestamp
calls_sent_30min	number of phone calls sent in the 30 min before the current stop timestamp
bt_entropy_30min	entropy of Bluetooth devices scanned in the 30 min before the current stop timestamp
bt_unique_30min	number of unique Bluetooth devices scanned in the 30 min before the current stop timestamp

Table 1. Description of the features used for the logistic regression models.

phenomenon has been described using concepts such as preferential return [38], heavy-tailed stay times and return rate based on the number of visits [39]. For the location prediction tasks, the consequence is that the target classes are very unbalanced, which implies that most records belong to very few classes and most classes are represented by only few records. To illustrate this issue, we consider the extreme case of places visited only once. Fig. 14 shows that surprisingly this fraction is quite large (0.7). This fact is, in large part, the central reason behind the difficulty of the prediction task.

As we shall see below, another central challenge is not just that our population visits a large number of different places, but also that many new places are discovered over time. We consider a stop at a location as "exploration" if this place has not been seen in the location history so far for a given user. In other words, this place is being visited for the first time by the user. To express this formally, we consider the sequence of stops $s_1, s_2, s_3, ..., s_n$ for each user . We consider a stop s_i as *return* ($Y = 0$) if s_i has been seen before in the location history, that is there exists a stop $s_j = s_i$ for $1 \leq j < i$. Otherwise we consider stop s_i an *exploration* ($Y = 1$), that is the place s_i is visited for the first time at step i. For example given a location sequence A B A C B C, the target variable exploration would have values 1 1 0 1 0 0.

We can then estimate the probability of exploration as fraction of explorations over the number of stops. To our surprise, this probability is particularly large: between 0.2 and 0.25 (Fig. 15). This implies that most users discover a new place every 4 or 5 stops.

The fact that a large fraction of stop-locations have never been seen before poses a challenge for the prediction task, since by construction any model that tries to predict a place from an alphabet of previously visited places will be unable to predict new, unseen symbols. Moreover, another consequence of this frequent exploration is that the pool of possible places constantly grows over time and, given the longitudinal nature of our dataset, ends up being very large. Fig. 16 shows how the average number of new
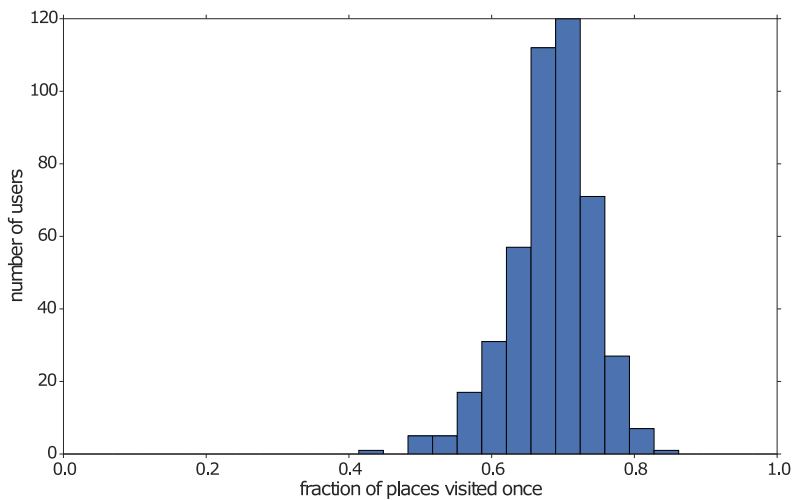
Fig 14. For each user we measure the fraction of places visited only once. This fraction is surprisingly large, as for each user on average 70% of the places were visited only once.
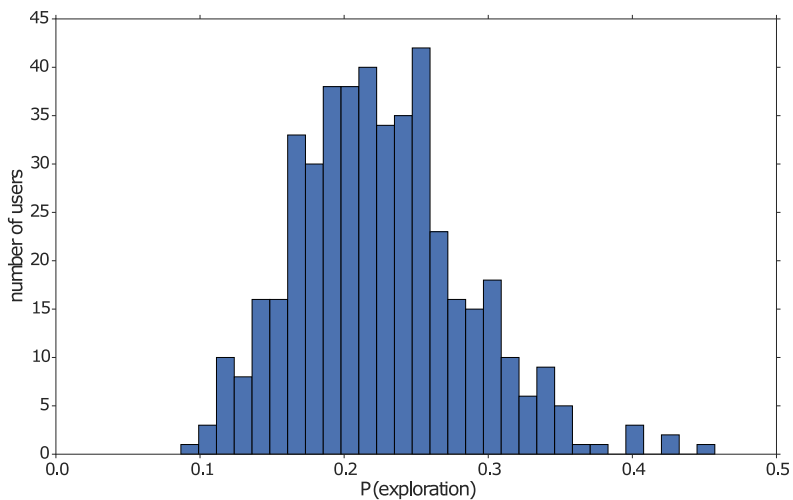
Fig 15. The probability of exploration estimated as fraction of explorations over the number of stops per user. Surprisingly this probability is quite large, meaning that on average users discover a new place every 4 or 5 stops.

place explored per week remains approximately constant around 4, and consequently the total number of places keeps growing to hundreds of places (Fig. 17). This is a problem for the prediction task, as the number of possible places that the classifier needs to choose from increases constantly.
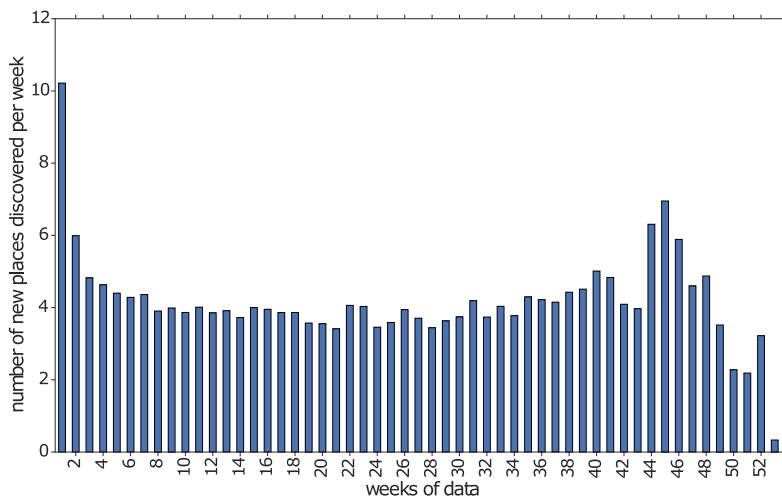
Fig 16. Number of new visited (explored) places for each week, average by user. Surprisingly, the number of explored places does not decrease over time, but remains around 4. This highlights the highly exploring behavior of our population.
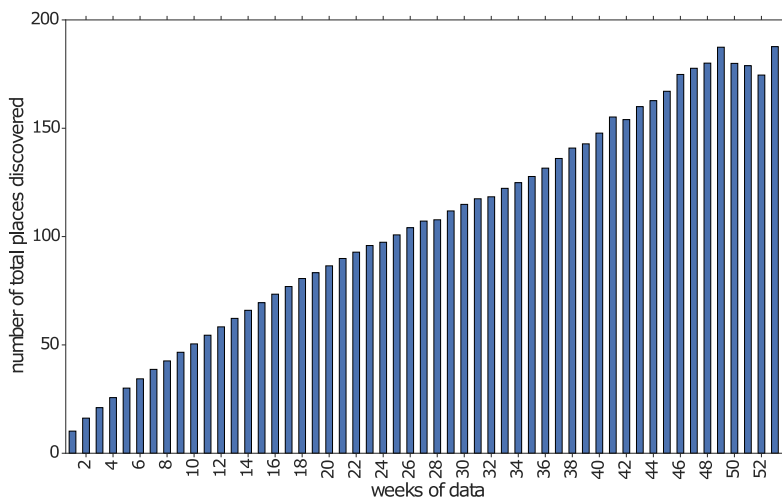
Fig 17. Cumulated number of new visited (explored) places for each week, average by user. As consequence of the large amount of exploration, the number of possible places to visit increases steadily over time, reaching on average almost 200 in one year.

In fact if we measure the relation between the number of unique places per user and the performance of the best performing logistic regression model using Pearson's correlation coefficient, we find a quite strong negative correlation ($r = -0.478$,

$p < 0.001$). On the other hand we find no significant correlation for accuracy with period length or number of stops.

These facts suggest that the exploration phenomenon is a key reason for the relatively low accuracy of mobility prediction tasks at high spatial resolution. Given the importance of exploration, we now consider a novel task in mobility prediction: exploration prediction.

Exploration prediction

The exploration prediction task can be stated as follows: given a user's location history up to step n, will the stop at step $n + 1$ be an exploration or a return?

The first question is: what should be the baseline model for the exploration prediction task? Surprisingly, most literature on human mobility prediction has focused on next location prediction but has overlooked the exploration prediction problem, and to the best of our knowledge no suitable solution has been proposed for this task. We therefore suggest, as a reasonable baseline, random guessing with probability equal to our prior knowledge of the fraction of explorations: $P(exploration) \approx 0.2$.

For our main model we use as before the logistic regression model with the same features constructed for the next place prediction model. We also add two additional features: *explore_now* and *explore_before*, which capture if the current stop or the previous stops were explorations, respectively. The intuition for these is that multiple explorations may occur in a row, and therefore the current exploration may increase the likelihood for an exploration at the next stop. As before, we fit each individual separately, and we perform an online prediction, that fits based on the data up to step n, and predicts exploration at step $n + 1$. We fit one logistic regression model for each of the single features, and a more complex model with all the features at once.

Measuring the performance of these models requires a few considerations. In this case, the classification problem is imbalanced, that is the number of positive cases (exploration) is much smaller than negative cases (return), as shown in Fig. 15. This implies that accuracy is therefore not a good metric, since a classifier predicting always return (the most probable class) would have good performance, but would not be useful. Instead we employ the f_1 score, which is the harmonic mean of precision (the fraction of correctly predicted explorations over all predicted explorations) and recall (the fraction of correctly predicted explorations over all true explorations). Fig. 18 shows the results of the exploration prediction.

As we would expect, the model with the most complete set of features outperforms the others. Among the single feature models, perhaps not surprisingly, the current location feature has the best performance. This finding can be explained by the role of some places as "gateways" for exploration such as public transport hubs (e.g. central train station). The individual features that perform also well are the time-related ones, in agreement with the intuition that exploration tends to happen to at specific times of the day or week. The *explore_now* and *explore_before* also perform well, suggesting an element of burstiness in the exploratory behavior. If we consider our best performing model, we find that it has average precision of 0.3 and recall of 0.65. Overall the performance of this model is far from perfect, showing that the exploration prediction problem is a challenging one.

Discussion

In this paper we first show that when interpreting results of predictive performance there are a number of factors that must be taken into consideration. The problem formulation is the central factor what should be taken into account when interpreting
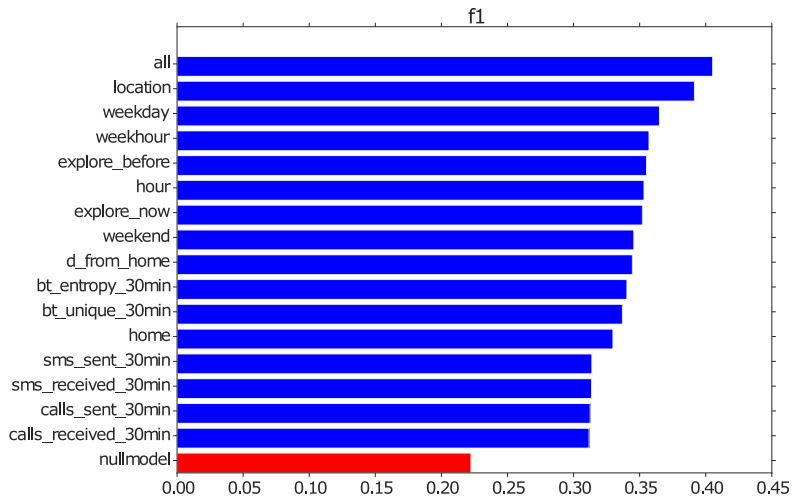
Fig 18. Exploration prediction: f_1 score for all models.

predictability results, since e.g. predicting the next time-bin is a very different (and much easier) task than predicting the next transition. We show that the most challenging problem is the next-place prediction, which is arguably the most useful task for practical applications such as travel recommendations. Another issue to be taken into account is the spatial resolution of the prediction, here we show how more coarse spatial precision results in an easier task. Similarly the time resolution also has an effect on the predictive power. We suggest that the factors described in this paper should be taken in consideration as context when comparing results from prediction models.

Other than the factors discussed above, we believe that one further reason for performance differences could be the demographics of the dataset. The population under study is here composed by students that have no single workplace but tend to change multiple classes per week, even multiple times per day. Moreover a younger population may have a more irregular schedule and more exploratory behavior. Certainly more work is needed to conclusively link demographics and predictability. For future directions, we suggest considering demographic factors when trying to characterize human mobility, as it has been done, for example, by linking changes in mobility patterns with unemployment status [40].

We also discussed the issue of exploration, and we show how frequently new places are discovered. Based on that, we show that the mechanism of exploration is an important part of human mobility and plays a role in next-place prediction. Because any model that tries to predict a next place from a set of visited place will fail when an exploration occurs. This problem has rarely been addressed in mobility prediction literature, which almost always assumes that the next place can be determined from the past history. Providing a full solution for next explored place prediction is beyond the scope of this work, and here we simply aim to stress the fact that the prediction of explorations is very different from the predictions to returns to known places. Some previous work on next-place prediction using social information [28,29] or nearby Points Of Interest [41] may be the starting point for investigating this problem.

In this sense, we raise the question on whether the simple location history is enough

for accurate next-place prediction. As we have discussed, there are indeed a lot of regularities both in the sequence of visits, and in the daily and weekly temporal patterns of visitation. However there are a lot of "exceptions to the rules", where schedules change, plans are canceled, and people run late. We speculate that other channels such as email, social media, calendar, class schedule may be needed for achieving a satisfying accuracy in the prediction task.

Acknowledgments

References

1. Lane ND, Mohammod M, Lin M, Yang X, Lu H, Ali S, et al. BeWell: A Smartphone Application to Monitor, Model and Promote Wellbeing. In: 5th International ICST Conference on Pervasive Computing Technologies for Healthcare; 2011. p. 23–26.

2. Quercia D, Lathia N, Calabrese F, Di Lorenzo G, Crowcroft J. Recommending Social Events from Mobile Phone Location Data. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE; 2010. p. 971–976.

3. Aalto L, Göthlin N, Korhonen J, Ojala T. Bluetooth And WAP Push Based Location-aware Mobile Advertising System. In: Proceedings of the 2nd International Conference on Mobile systems, Applications, and Services. ACM; 2004. p. 49–58.

4. Lu X, Bengtsson L, Holme P. Predictability of Population Displacement after the 2010 Haiti Earthquake. Proceedings of the National Academy of Sciences. 2012;109(29):11576–11581.

5. Çolak S, Lima A, González MC. Understanding Congested Travel in Urban Areas. Nature Communications. 2016;7.

6. Eagle N, Pentland A. Reality Mining: Sensing Complex Social Systems. Personal and Ubiquitous Computing. 2006;10(4):255–268.

7. Liao L, Patterson DJ, Fox D, Kautz H. Learning and Inferring Transportation Routines. Artificial Intelligence. 2007;171(5):311–331.

8. Zheng Y, Zhou X. Computing with Spatial Trajectories. Springer Science & Business Media; 2011.

9. Arentze T, Timmermans H. Albatross: A Learning Based Transportation Oriented Simulation System. Eirass Eindhoven; 2000.

10. Balmer M, Meister K, Rieser M, Nagel K, Axhausen KW, Axhausen KW, et al. Agent-based Simulation of Travel Demand: Structure and Computational Performance of MATSim-t. ETH, Eidgenössische Technische Hochschule Zürich, IVT Institut für Verkehrsplanung und Transportsysteme; 2008.

11. Goodchild MF. Citizens as Sensors: The World of Volunteered Geography. GeoJournal. 2007;69(4):211–221.

12. Batty M. The New Science of Cities. MIT Press; 2013.

13. Song C, Qu Z, Blumm N, Barabási AL. Limits of Predictability in Human Mobility. Science. 2010;327(5968):1018–1021.

14. Simini F, González MC, Maritan A, Barabási AL. A Universal Model for Mobility and Migration Patterns. Nature. 2012;484(7392):96–100.

15. Schneider CM, Rudloff C, Bauer D, González MC. Daily Travel Behavior: Lessons from a Week-long Survey for the Extraction of Human Mobility Motifs Related Information. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. ACM; 2013. p. 3.

16. Lin M, Hsu WJ, Lee ZQ. Predictability of Individuals' Mobility with High-resolution Positioning Data. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM; 2012. p. 381–390.

17. Smith G, Wieser R, Goulding J, Barrack D. A Refined Limit on the Predictability of Human Mobility. In: Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on. IEEE; 2014. p. 88–94.

18. Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L. Approaching the Limit of Predictability in Human Mobility. Scientific Reports. 2013;3.

19. Song L, Kotz D, Jain R, He X. Evaluating Next-cell Predictors with Extensive Wi-fi Mobility Data. Mobile Computing, IEEE Transactions on. 2006;5(12):1633–1649.

20. Bapierre H, Groh G, Theiner S. A Variable Order Markov Model Approach for Mobility Prediction. Pervasive Computing. 2011; p. 8–16.

21. Zheng Y, Xie X, Ma WY. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. IEEE Data Eng Bull. 2010;33(2):32–39.

22. Gao H, Tang J, Liu H. Mobile Location Prediction in Spatio-temporal Context. In: Nokia Mobile Data Challenge Workshop. vol. 41; 2012. p. 44.

23. Laurila JK, Gatica-Perez D, Aad I, Bornet O, Do TMT, Dousse O, et al. The Mobile Data Challenge: Big Data for Mobile Computing Research. In: Pervasive Computing. EPFL-CONF-192489; 2012.

24. Do TMT, Gatica-Perez D. Contextual Conditional Models for Smartphone-based Human Mobility Prediction. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM; 2012. p. 163–172.

25. Do TMT, Dousse O, Miettinen M, Gatica-Perez D. A Probabilistic Kernel Method for Human Mobility Prediction with Smartphones. Pervasive and Mobile Computing. 2015;20:13–28.

26. Scellato S, Musolesi M, Mascolo C, Latora V, Campbell AT. Nextplace: A Spatio-temporal Prediction Framework for Pervasive Systems. In: Pervasive Computing. Springer; 2011. p. 152–169.

27. Sadilek A, Krumm J. Far Out: Predicting Long-Term Human Mobility. In: AAAI; 2012.

28. Cho E, Myers SA, Leskovec J. 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. p. 1082–1090.

29. Sadilek A, Kautz H, Bigham JP. 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. p. 723–732.

30. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, et al. Measuring Large-scale Social Networks with High Resolution. PloS one. 2014;9(4):e95978.

31. Kang JH, Welbourne W, Stewart B, Borriello G. Extracting Places from Traces of Locations. In: Proceedings of the 2nd ACM International workshop on Wireless mobile applications and services on WLAN hotspots. ACM; 2004. p. 110–118.

32. Zheng VW, Zheng Y, Xie X, Yang Q. Collaborative Location and Activity Recommendations with GPS History Data. In: Proceedings of the 19th International Conference on World Wide Web. ACM; 2010. p. 1029–1038.

33. Thierry B, Chaix B, Kestens Y. Detecting Activity Locations from Raw Gps Data: A Novel Kernel-based Algorithm. International Journal of Health Geographics. 2013;12(1):1.

34. Zhou C, Frankowski D, Ludford P, Shekhar S, Terveen L. Discovering Personally Meaningful Places: An Interactive Clustering Approach. ACM Transactions on Information Systems (TOIS). 2007;25(3):12.

35. Zheng Y, Zhang L, Xie X, Ma WY. Mining Interesting Locations and Travel Sequences from GPS Trajectories. In: Proceedings of the 18th International Conference on World Wide Web. ACM; 2009. p. 791–800.

36. Montoliu R, Gatica-Perez D. Discovering Human Places of Interest from Multimodal Mobile Phone Data. In: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia. ACM; 2010. p. 12.

37. Ester M, Kriegel HP, Sander J, Xu X. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. AAAI Press; 1996. p. 226–231.

38. Song C, Koren T, Wang P, Barabási AL. Modelling the Scaling Properties of Human Mobility. Nature Physics. 2010;6(10):818–823.

39. Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding Individual Human Mobility Patterns. Nature. 2008;453(7196):779–782.

40. Toole JL, Lin YR, Muehlegger E, Shoag D, González MC, Lazer D. Tracking Employment Shocks Using Mobile Phone Data. Journal of The Royal Society Interface. 2015;12(107):20150185.

41. Yu C, Liu Y, Yao D, Yang LT, Jin H, Chen H, et al. Modeling User Activity Patterns for Next-Place Prediction. IEEE Systems Journal. 2015;PP(99):1–12. doi:10.1109/JSYST.2015.2445919.

Bibliography

[1] B. C. Choi and A. W. Pak, "A catalog of biases in questionnaires," *Prev Chronic Dis*, vol. 2, no. 1, p. A13, 2005.

[2] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Life in the network: the coming age of computational social science," *Science (New York, NY)*, vol. 323, no. 5915, p. 721, 2009.

[3] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland, "Sensible organizations: Technology and methodology for automatically measuring organizational behavior," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 1, pp. 43–55, 2009.

[4] J.-P. Eckmann, E. Moses, and D. Sergi, "Entropy of dialogues creates coherent structures in e-mail traffic," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 40, pp. 14333–14337, 2004.

[5] R. D. Malmgren, J. M. Hofman, L. A. Amaral, and D. J. Watts, "Characterizing individual communication patterns," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 607–616, ACM, 2009.

[6] H. Ebel, L.-I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," *Physical review E*, vol. 66, no. 3, p. 035103, 2002.

[7] P. S. Dodds, R. Muhamad, and D. J. Watts, "An experimental study of search in global social networks," *science*, vol. 301, no. 5634, pp. 827–829, 2003.

[8] "Facebook Reports Fourth Quarter and Full Year 2015 Results." `http://investor.fb.com/releasedetail.cfm?ReleaseID=952040`, 2016. [Online; accessed 27 May 2016].

[9] "Twitter Usage." `https://about.twitter.com/company`, 2016. [Online; accessed 27 May 2016].

[10] A. Sadilek, H. Kautz, and J. 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*, pp. 723–732, ACM, 2012.

[11] D. Jurgens, "That's what friends are for: inferring location in online social media platforms based on social relationships," *ICWSM*, vol. 13, pp. 273–282, 2013.

[12] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare," *ICwSM*, vol. 11, pp. 70–573, 2011.

[13] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, "Structural diversity in social contagion," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 5962–5966, 2012.

[14] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four degrees of separation," in *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 33–42, ACM, 2012.

[15] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, pp. 851–860, ACM, 2010.

[16] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[17] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment.," *ICWSM*, vol. 10, pp. 178–185, 2010.

[18] K. Lerman and R. Ghosh, "Information contagion: an empirical study of the spread of news on digg and twitter social networks," *ICWSM*, vol. 10, pp. 90–97, 2010.

[19] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[20] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[21] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[22] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, "Redrawing the map of great britain from a network of human interactions," *PloS one*, vol. 5, no. 12, p. e14248, 2010.

[23] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 0018–26, 2011.

[24] T. Louail, M. Lenormand, O. G. C. Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, "From mobile phone data to the spatial structure of cities," *Scientific reports*, vol. 4, 2014.

[25] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[26] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1100–1108, ACM, 2011.

[27] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, 2013.

[28] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *Proceedings of the 2006 international workshop on Mining software repositories*, pp. 137–143, ACM, 2006.

[29] F. Thung, T. F. Bissyandé, D. Lo, and L. Jiang, "Network structure of social coding in github," in *Software Maintenance and Reengineering (CSMR), 2013 17th European Conference on*, pp. 323–326, IEEE, 2013.

[30] B. Vasilescu, V. Filkov, and A. Serebrenik, "Stackoverflow and github: Associations between software development and crowdsourced knowledge," in *Social Computing (SocialCom), 2013 International Conference on*, pp. 188–195, IEEE, 2013.

[31] D. E. Acuna, S. Allesina, and K. P. Kording, "Future impact: Predicting scientific success," *Nature*, vol. 489, no. 7415, pp. 201–202, 2012.

[32] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[33] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[34] L. Isella, M. Romano, A. Barrat, C. Cattuto, V. Colizza, W. Van den Broeck, F. Gesualdo, E. Pandolfi, L. Ravà, C. Rizzo, *et al.*, "Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors," *PloS one*, vol. 6, no. 2, p. e17144, 2011.

[35] M. Raento, A. Oulasvirta, and N. Eagle, "Smartphones an emerging tool for social scientists," *Sociological methods & research*, vol. 37, no. 3, pp. 426–454, 2009.

[36] "Strategy Analytics: Worldwide Smartphone Population Tops 1 Billion in Q3 2012." `http://www.businesswire.com/news/home/20121017005479/en/Strategy-Analytics-Worldwide-Smartphone-Population-Tops-1`, 2012. [Online; accessed 27 May 2016].

[37] J. Poushter, "Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies." `http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/`, 2016. [Online; accessed 27 May 2016].

[38] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, R. A. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn, "The rise of people-centric sensing," *Internet Computing, IEEE*, vol. 12, no. 4, pp. 12–21, 2008.

[39] "Funf Open Sensing Framework." `https://http://funf.org/`. [Online; accessed 25 July 2016].

[40] I. Mulder, G. Ter Hofte, and J. Kort, "Socioxensor: Measuring user behaviour and user experience in context with mobile devices," in *Proceedings of Measuring Behavior*, pp. 355–358, 2005.

[41] M. Raento, A. Oulasvirta, R. Petit, and H. Toivonen, "Contextphone: A prototyping platform for context-aware mobile applications," *Pervasive Computing, IEEE*, vol. 4, no. 2, pp. 51–59, 2005.

[42] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay, "Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones," in *Proceedings of the 5th international conference on Mobile systems, applications and services*, pp. 57–70, ACM, 2007.

[43] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell, "Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pp. 337–350, ACM, 2008.

[44] E. Miluzzo, C. T. Cornelius, A. Ramaswamy, T. Choudhury, Z. Liu, and A. T. Campbell, "Darwin phones: the evolution of sensing and inference on mobile phones," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pp. 5–20, ACM, 2010.

[45] U. Nations, "The Universal Declaration of Human Rights." `http://www.un.org/en/universal-declaration-human-rights/index.html`, 2016. [Online; accessed 27 May 2016].

[46] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 517–526, ACM, 2009.

[47] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[48] E. Oikarinen, "Revealing information while preserving privacy," in *Survey for the Seminar T-79.514: Special Course on Cryptology/Privacy Preserving Data Mining at Helsinki University of Technology*, 2003.

[49] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," *Personal and ubiquitous computing*, vol. 10, no. 4, pp. 255–268, 2006.

[50] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila, "Towards rich mobile phone datasets: Lausanne data collection campaign," *Proc. ICPS, Berlin*, 2010.

[51] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, *et al.*, "The mobile data challenge: Big data for mobile computing research," in *Pervasive Computing*, no. EPFL-CONF-192489, 2012.

[52] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643–659, 2011.

[53] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, *et al.*, "Sensing the" health state" of a community," *IEEE Pervasive Computing*, no. 4, pp. 36–45, 2012.

[54] A. Striegel, S. Liu, L. Meng, C. Poellabauer, D. Hachen, and O. Lizardo, "Lessons learned from the netsense smartphone study," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 51–56, 2013.

[55] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 3–14, ACM, 2014.

[56] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[57] V. Sekara, A. Stopczynski, and S. Lehmann, "The fundamental structures of dynamic social networks," *arXiv preprint arXiv:1506.04704*, 2015.

[58] A. Stopczynski, A. S. Pentland, and S. Lehmann, "Physical proximity and spreading in dynamic social networks," *arXiv preprint arXiv:1509.06530*, 2015.

[59] E. Mones, A. Stopczynski, A. Pentland, N. Hupert, and S. Lehmann, "Vaccination and complex social dynamics," *arXiv preprint arXiv:1603.00910*, 2016.

[60] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann, "Tracking human mobility using wifi signals," *PloS one*, vol. 10, no. 7, p. e0130824, 2015.

[61] Y.-A. de Montjoye, S. S. Wang, A. Pentland, D. T. T. Anh, A. Datta, *et al.*, "On the trusted use of large-scale personal data," *IEEE Data Eng. Bull.*, vol. 35, no. 4, pp. 5–8, 2012.

[62] A. Stopczynski, R. Pietri, A. Pentland, D. Lazer, and S. Lehmann, "Privacy in sensor-driven human data collection: A guide for practitioners," *arXiv preprint arXiv:1403.5299*, 2014.

[63] M. Csikszentmihalyi and R. Larson, "Validity and reliability of the experience-sampling method," in *Flow and the Foundations of Positive Psychology*, pp. 35–54, Springer, 2014.

[64] A. Cuttone, S. Lehmann, and J. E. Larsen, "A mobile personal informatics system with interactive visualizations of mobility and social interactions," in *Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia*, pp. 27–30, ACM, 2013.

[65] B. Franklin, "Autobiography of Benjamin Franklin." `https://www.gutenberg.org/ebooks/20203`. [Online; accessed 12 July 2016].

[66] N. W. Gillham, *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics.* Oxford University Press, 2001.

[67] N. Felton, "Feltron annual report." `http://feltron.com/`, 2016. [Online; accessed 17 May 2016].

[68] S. Fox and M. Duggan, "Tracking for Health." `http://www.pewinternet.org/2013/01/28/tracking-for-health/`, 2013. [Online; accessed 11 May 2016].

[69] "Fitbit." `https://www.fitbit.com`, 2016. [Online; accessed 17 May 2016].

[70] "Jawbone." `https://jawbone.com/`, 2016. [Online; accessed 17 May 2016].

[71] "mybasis." `http://www.mybasis.com/`, 2016. [Online; accessed 17 May 2016].

[72] "nikeplus." `https://secure-nikeplus.nike.com/plus/products/fuelband/`, 2016. [Online; accessed 17 May 2016].

[73] "TrackYourHappiness." `https://www.trackyourhappiness.org/`, 2016. [Online; accessed 17 May 2016].

[74] "MoodScope." `https://www.moodscope.com/`, 2016. [Online; accessed 17 May 2016].

[75] "Mappiness." `http://www.mappiness.org.uk/`, 2016. [Online; accessed 17 May 2016].

[76] "Daytum." `http://daytum.com/`, 2016. [Online; accessed 17 May 2016].

[77] Y. Assogba and J. Donath, "Mycrocosm: visual microblogging," in *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pp. 1–10, IEEE, 2009.

[78] "MercuryApp." `https://www.mercuryapp.com/`, 2016. [Online; accessed 17 May 2016].

[79] "CureTogether." `http://curetogether.com/`, 2016. [Online; accessed 17 May 2016].

[80] "PatientsLikeMe." `https://www.patientslikeme.com/`, 2016. [Online; accessed 17 May 2016].

[81] B. Dolan, "Fitbit files for IPO, sold nearly 11 million fitness devices in 2014." `http://mobihealthnews.com/43412/fitbit-files-for-ipo-sold-nearly-11-million-fitness-devices-in-2014`, 2015. [Online; accessed 17 May 2016].

[82] C. C. Mareca Hatler, Darryl Gurganious, "Mobile sensing health & wellness." `http://onworld.com/mobilesensing/health/`, 2013. [Online; accessed 12 May 2016].

[83] E. Singer, "The measured life." `https://www.technologyreview.com/s/424390/the-measured-life/`, 2011. [Online; accessed 10 May 2016].

[84] I. Li, A. Dey, and J. Forlizzi, "A stage-based model of personal informatics systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 557–566, ACM, 2010.

[85] I. Li, A. K. Dey, and J. Forlizzi, "Understanding my data, myself: supporting self-reflection with ubicomp technologies," in *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 405–414, ACM, 2011.

[86] E. K. Choe, N. B. Lee, B. Lee, W. Pratt, and J. A. Kientz, "Understanding quantified-selfers' practices in collecting and exploring personal data," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 1143–1152, ACM, 2014.

[87] J. Rooksby, M. Rost, A. Morrison, and M. C. Chalmers, "Personal tracking as lived informatics," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pp. 1163–1172, ACM, 2014.

[88] D. A. Epstein, A. Ping, J. Fogarty, and S. A. Munson, "A lived informatics model of personal informatics," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 731–742, ACM, 2015.

[89] R. Fleck and G. Fitzpatrick, "Reflecting on reflection: framing a design landscape," in *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, pp. 216–223, ACM, 2010.

[90] D. Boud, R. Keogh, and D. Walker, *Reflection: Turning experience into learning.* Routledge, 2013.

[91] E. P. Baumer, V. Khovanskaya, M. Matthews, L. Reynolds, V. Schwanda Sosik, and G. Gay, "Reviewing reflection: on the use of reflection in interactive system design," in *Proceedings of the 2014 conference on Designing interactive systems*, pp. 93–102, ACM, 2014.

[92] V. Rivera-Pelayo, V. Zacharias, L. Müller, and S. Braun, "A framework for applying quantified self approaches to support reflective learning," *Mobile Learning*, vol. 2012, p. 123, 2012.

[93] D. Lupton, "Self-tracking cultures: towards a sociology of personal informatics," in *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*, pp. 77–86, ACM, 2014.

[94] E. P. Baumer, "Reflective informatics: conceptual dimensions for designing technologies of reflection," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 585–594, ACM, 2015.

[95] A. Ayobi, P. Marshall, and A. L. Cox, "Reflections on 5 years of personal informatics: rising concerns and emerging directions," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2774–2781, ACM, 2016.

[96] V. Bush, "As We May Think." `http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/`, 1945. [Online; accessed 16 May 2016].

[97] E. Freeman and D. Gelernter, "Lifestreams: A storage model for personal data," *ACM SIGMOD Record*, vol. 25, no. 1, pp. 80–86, 1996.

[98] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong, "Mylifebits: fulfilling the memex vision," in *Proceedings of the tenth ACM international conference on Multimedia*, pp. 235–238, ACM, 2002.

[99] A. L. Allen, "Dredging up the past: Lifelogging, memory, and surveillance," *The University of Chicago Law Review*, vol. 75, no. 1, pp. 47–74, 2008.

[100] A. J. Sellen and S. Whittaker, "Beyond total capture: a constructive critique of lifelogging," *Communications of the ACM*, vol. 53, no. 5, pp. 70–77, 2010.

[101] K. Tollmar, F. Bentley, C. Viedma, and I. Libertyville, "Mobile health mashups," *Proc. PervasiveHealth*, vol. 12, 2012.

[102] D. Pavel, D. Trossen, M. Holweg, and V. Callaghan, "Lifestyle stories: Correlating user information through a story-inspired paradigm," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pp. 412–415, IEEE, 2013.

[103] D. Epstein, F. Cordeiro, E. Bales, J. Fogarty, and S. Munson, "Taming data complexity in lifelogs: exploring visual cuts of personal informatics data," in *Proceedings of the 2014 conference on Designing interactive systems*, pp. 667–676, ACM, 2014.

[104] J. J. Lin, L. Mamykina, S. Lindtner, G. Delajoux, and H. B. Strub, "Fish'n'steps: Encouraging physical activity with an interactive computer game," in *UbiComp 2006: Ubiquitous Computing*, pp. 261–278, Springer, 2006.

[105] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, *et al.*, "Activity sensing in the wild: a field trial of ubifit garden," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1797–1806, ACM, 2008.

[106] J. Froehlich, T. Dillahunt, P. Klasnja, J. Mankoff, S. Consolvo, B. Harrison, and J. A. Landay, "Ubigreen: investigating a mobile tool for tracking and supporting green transportation habits," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1043–1052, ACM, 2009.

[107] C. Fan, J. Forlizzi, and A. K. Dey, "A spark of activity: exploring informative art as visualization for physical activity," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 81–84, ACM, 2012.

[108] L. Frick, "Experiments in self-tracking." `http://www.lauriefrick.com/category/work/`, 2016. [Online; accessed 16 May 2016].

[109] L. Wilkinson, *The grammar of graphics*. Springer Science & Business Media, 2006.

[110] F. J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[111] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo," *Commun. Acm*, vol. 53, no. 6, pp. 59–67, 2010.

[112] J. W. Tukey, "Exploratory data analysis," 1977.

[113] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pp. 336–343, IEEE, 1996.

[114] J. Heer and B. Shneiderman, "Interactive dynamics for visual analysis," *Queue*, vol. 10, no. 2, p. 30, 2012.

[115] E. R. Tufte and P. Graves-Morris, *The visual display of quantitative information*, vol. 2. Graphics press Cheshire, CT, 1983.

[116] C. North, "Toward measuring visualization insight," *Computer Graphics and Applications, IEEE*, vol. 26, no. 3, pp. 6–9, 2006.

[117] C. Chen, "Top 10 unsolved information visualization problems," *Computer Graphics and Applications, IEEE*, vol. 25, no. 4, pp. 12–16, 2005.

[118] J. E. Larsen, A. Cuttone, and S. L. Jørgensen, "Qs spiral: Visualizing periodic quantified self data," in *CHI 2013 Workshop on Personal Informatics in the Wild: Hacking Habits for Health & Happiness*, 2013.

[119] G. Wolf, "The data-driven life." `http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html?_r=0`, 2010. [Online; accessed 27 June 2016].

[120] M. B. Barcena, "How safe is your quantified self?." `https://www.symantec.com/content/dam/symantec/docs/white-papers/how-safe-is-your-quantified-self-en.pdf`, 2014. [Online; accessed 27 June 2016].

[121] D. Lupton, "Self-tracking cultures: Towards a sociology of personal informatics," in *Proceedings of the OzCHI conference*, ACM, 2014.

[122] D. Lupton, "You are your data: Self-tracking practices and concepts of data," *Available at SSRN*, 2014.

[123] M. Bode and D. B. Kristensen, "The digital doppelgänger within. a study on self-tracking and the quantified self movement," in *Domen Bajde and Robin Canniford Assembling Consumption*, Oxford/ New York: Routledge, 2015.

[124] G. Wolf, "Are self-trackers narcissists?," 2009.

[125] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[126] M. D. Back, A. C. Küfner, M. Dufner, T. M. Gerlach, J. F. Rauthmann, and J. J. Denissen, "Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism.," *Journal of Personality and Social Psychology*, vol. 105, no. 6, p. 1013, 2013.

[127] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[128] M. Friendly, "Corrgrams: Exploratory displays for correlation matrices," *The American Statistician*, vol. 56, no. 4, pp. 316–324, 2002.

[129] "SleepCycle." `https://itunes.apple.com/en/app/sleep-cycle-alarm-clock/id320606217?mt=8`, 2016. [Online; accessed 17 May 2016].

[130] "SleepBot." `https://play.google.com/store/apps/details?id=com.lslk.sleepbot&hl=en`, 2016. [Online; accessed 17 May 2016].

[131] "Sleep As Android." `https://play.google.com/store/apps/details?id=com.urbandroid.sleep&hl=en`, 2016. [Online; accessed 17 May 2016].

[132] T. Hao, G. Xing, and G. Zhou, "isleep: unobtrusive sleep quality monitoring using smartphones," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, p. 4, ACM, 2013.

[133] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive sleep monitoring using smartphones," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, pp. 145–152, IEEE, 2013.

[134] H. Sajjad Hossain, N. Roy, and M. A. Al Hafiz Khan, "Sleep well: a sound sleep monitoring framework for community scaling," in *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, vol. 1, pp. 44–53, IEEE, 2015.

[135] S. Abdullah, M. Matthews, E. L. Murnane, G. Gay, and T. Choudhury, "Towards circadian computing: early to bed and early to rise makes some of us unhealthy and sleep deprived," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 673–684, ACM, 2014.

[136] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proceedings of the 13th international conference on Ubiquitous computing*, pp. 89–98, ACM, 2011.

[137] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 186–194, ACM, 2012.

[138] S. Çolak, A. Lima, and M. C. González, "Understanding congested travel in urban areas," *Nature communications*, vol. 7, 2016.

[139] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft, "Recommending social events from mobile phone location data," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 971–976, IEEE, 2010.

[140] L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala, "Bluetooth and wap push based location-aware mobile advertising system," in *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, pp. 49–58, ACM, 2004.

[141] G. Yavaş, D. Katsaros, Ö. Ulusoy, and Y. Manolopoulos, "A data mining approach for location prediction in mobile environments," *Data & Knowledge Engineering*, vol. 54, no. 2, pp. 121–146, 2005.

[142] L. Hufnagel, D. Brockmann, and T. Geisel, "Forecast and control of epidemics in a globalized world," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 42, pp. 15124–15129, 2004.

[143] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani, "Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions," *PLoS Med*, vol. 4, no. 1, p. e13, 2007.

[144] S. Ni and W. Weng, "Impact of travel patterns on epidemic dynamics in heterogeneous spatial metapopulation networks," *Physical review E*, vol. 79, no. 1, p. 016111, 2009.

[145] D. Balcan and A. Vespignani, "Phase transitions in contagion processes mediated by recurrent mobility patterns," *Nature physics*, vol. 7, no. 7, pp. 581–586, 2011.

[146] V. Belik, T. Geisel, and D. Brockmann, "Natural human mobility patterns and spatial spread of infectious diseases," *Physical Review X*, vol. 1, no. 1, p. 011001, 2011.

[147] J. P. Bagrow, D. Wang, and A.-L. Barabasi, "Collective response of human populations to large-scale emergencies," *PloS one*, vol. 6, no. 3, p. e17680, 2011.

[148] X. Lu, L. Bengtsson, and P. Holme, "Predictability of population displacement after the 2010 haiti earthquake," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11576–11581, 2012.

[149] Google, "Google Location Services for Android." `https://developer.android.com/guide/topics/location/strategies.html`, 2016. [Online; accessed 29 April 2016].

[150] C. Arthur, "Android phones record user-locations according to research." `https://www.theguardian.com/technology/2011/apr/22/iphone-android-location-based-services`, 2011. [Online; accessed 29 April 2016].

[151] P. Sapiezynski, R. Gatej, A. Mislove, and S. Lehmann, "Opportunities and challenges in crowdsourced wardriving," in *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pp. 267–273, ACM, 2015.

[152] Google, "Android Location class documentation." `https://developer.android.com/reference/android/location/Location.html#getAccuracy()`, 2016. [Online; accessed 2 May 2016].

[153] N. Andrienko and G. Andrienko, *Exploratory analysis of spatial and temporal data: a systematic approach.* Springer Science & Business Media, 2006.

[154] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science and engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[155] J. Whitaker, "Basemap." `http://matplotlib.org/basemap/`, 2016. [Online; accessed 17 May 2016].

[156] M. Office, "Cartopy: a cartographic python library with a matplotlib interface." `http://scitools.org.uk/cartopy`, 2016. [Online; accessed 17 May 2016].

[157] Google, "Google Maps." `https://developers.google.com/maps/`, 2016. [Online; accessed 17 May 2016].

[158] "LeafletJS." `http://leafletjs.com/`, 2016. [Online; accessed 17 May 2016].

[159] "OpenLayers." `http://openlayers.org/`, 2016. [Online; accessed 17 May 2016].

[160] QGis, "Quantum gis geographic information system," *Open Source Geospatial Foundation Project*, 2011.

[161] "ArcGIS." `http://www.arcgis.com/`, 2016. [Online; accessed 17 May 2016].

[162] A. Cuttone, "geoplotlib." `https://github.com/andrea-cuttone/geoplotlib`, 2016. [Online; accessed 17 May 2016].

[163] Y. Wang, N. J. Yuan, D. Lian, L. Xu, X. Xie, E. Chen, and Y. Rui, "Regularity and conformity: location prediction using heterogeneous mobility data," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1275–1284, ACM, 2015.

[164] Z. Fan, X. Song, and R. Shibasaki, "Cityspectrum: A non-negative tensor factorization approach," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 213–223, ACM, 2014.

[165] J. Krumm and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," in *UbiComp 2006: Ubiquitous Computing*, pp. 243–260, Springer, 2006.

[166] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with gps history data," in *Proceedings of the 19th international conference on World wide web*, pp. 1029–1038, ACM, 2010.

[167] N. Marmasse and C. Schmandt, "Location-aware information delivery with commotion," in *Handheld and Ubiquitous Computing*, pp. 157–171, Springer, 2000.

[168] D. Ashbrook and T. Starner, "Using gps to learn significant locations and predict movement across multiple users," *Personal and Ubiquitous Computing*, vol. 7, no. 5, pp. 275–286, 2003.

[169] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[170] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," in *ACM Sigmod Record*, vol. 28, pp. 49–60, ACM, 1999.

[171] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personal gazetteers: an interactive clustering approach," in *Proceedings of the 12th annual ACM international workshop on Geographic information systems*, pp. 266–273, ACM, 2004.

[172] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personally meaningful places: An interactive clustering approach," *ACM Transactions on Information Systems (TOIS)*, vol. 25, no. 3, p. 12, 2007.

[173] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A clustering-based approach for discovering interesting places in trajectories," in *Proceedings of the 2008 ACM symposium on Applied computing*, pp. 863–868, ACM, 2008.

[174] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th international conference on World wide web*, pp. 791–800, ACM, 2009.

[175] X. Cao, G. Cong, and C. S. Jensen, "Mining significant semantic locations from gps data," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1009–1020, 2010.

[176] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," in *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pp. 110–118, ACM, 2004.

[177] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," *ACM Computing Surveys (CSUR)*, vol. 23, no. 3, pp. 345–405, 1991.

[178] K. Laasonen, M. Raento, and H. Toivonen, "Adaptive on-device location recognition," in *Pervasive Computing*, pp. 287–304, Springer, 2004.

[179] J. Hightower, S. Consolvo, A. LaMarca, I. Smith, and J. Hughes, "Learning and recognizing the places we go," in *UbiComp 2005: Ubiquitous Computing*, pp. 159–176, Springer, 2005.

[180] D. H. Kim, J. Hightower, R. Govindan, and D. Estrin, "Discovering semantically meaningful places from pervasive rf-beacons," in *Proceedings of the 11th international conference on Ubiquitous computing*, pp. 21–30, ACM, 2009.

[181] D. K. Wind, P. Sapiezynski, M. A. Furman, and S. Lehmann, "Inferring stop-locations from wifi," *PloS one*, vol. 11, no. 2, p. e0149105, 2016.

[182] J. Hallberg, M. Nilsson, and K. Synnes, "Positioning with bluetooth," in *Telecommunications, 2003. ICT 2003. 10th International Conference on*, vol. 2, pp. 954–958, IEEE, 2003.

[183] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Reviews of modern physics*, vol. 81, no. 2, p. 591, 2009.

[184] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[185] G. Ramos-Fernández, J. L. Mateos, O. Miramontes, G. Cocho, H. Larralde, and B. Ayala-Orozco, "Lévy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi)," *Behavioral Ecology and Sociobiology*, vol. 55, no. 3, pp. 223–230, 2004.

[186] G. M. Viswanathan, V. Afanasyev, S. Buldyrev, E. Murphy, P. Prince, H. E. Stanley, *et al.*, "Lévy flight search patterns of wandering albatrosses," *Nature*, vol. 381, no. 6581, pp. 413–415, 1996.

[187] D. W. Sims, E. J. Southall, N. E. Humphries, G. C. Hays, C. J. Bradshaw, J. W. Pitchford, A. James, M. Z. Ahmed, A. S. Brierley, M. A. Hindell, *et al.*, "Scaling laws of marine predator search behaviour," *Nature*, vol. 451, no. 7182, pp. 1098–1102, 2008.

[188] J. Klafter, M. F. Shlesinger, and G. Zumofen, "Beyond brownian motion," *Physics today*, vol. 49, no. 2, pp. 33–39, 1996.

[189] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM transactions on networking (TON)*, vol. 19, no. 3, pp. 630–643, 2011.

[190] M. E. Newman, "Power laws, pareto distributions and zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[191] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[192] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to english text," *Information Theory, IEEE Transactions on*, vol. 44, no. 3, pp. 1319–1327, 1998.

[193] M. Lin, W.-J. Hsu, and Z. Q. Lee, "Predictability of individuals' mobility with high-resolution positioning data," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 381–390, ACM, 2012.

[194] G. Smith, R. Wieser, J. Goulding, and D. Barrack, "A refined limit on the predictability of human mobility," in *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, pp. 88–94, IEEE, 2014.

[195] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility," *Nature communications*, vol. 6, 2015.

[196] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130246, 2013.

[197] L. J. King, "Central place theory," 1985.

[198] U. Bronfenbrenner and P. A. Morris, "The bioecological model of human development," *Handbook of child psychology*, 2006.

[199] L. Song, D. Kotz, R. Jain, and X. He, "Evaluating next-cell predictors with extensive wi-fi mobility data," *Mobile Computing, IEEE Transactions on*, vol. 5, no. 12, pp. 1633–1649, 2006.

[200] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, 2013.

[201] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility markov chains," in *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, p. 3, ACM, 2012.

[202] H. Bapierre, G. Groh, and S. Theiner, "A variable order markov model approach for mobility prediction," *Pervasive Computing*, pp. 8–16, 2011.

[203] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: a collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.

[204] S. Russell, P. Norvig, and A. Intelligence, "A modern approach," *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, vol. 25, p. 27, 1995.

[205] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46, IBM New York, 2001.

[206] H. Gao, J. Tang, and H. Liu, "Mobile location prediction in spatio-temporal context," in *Nokia mobile data challenge workshop*, vol. 41, p. 44, 2012.

[207] T. M. T. Do and D. Gatica-Perez, "Contextual conditional models for smartphone-based human mobility prediction," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 163–172, ACM, 2012.

[208] T. M. T. Do, O. Dousse, M. Miettinen, and D. Gatica-Perez, "A probabilistic kernel method for human mobility prediction with smartphones," *Pervasive and Mobile Computing*, vol. 20, pp. 13–28, 2015.

[209] E. Cho, S. A. Myers, and J. 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*, pp. 1082–1090, ACM, 2011.

[210] J. Pang and Y. Zhang, "Location prediction: communities speak louder than friends," in *Proceedings of the 2015 ACM on Conference on Online Social Networks*, pp. 161–171, ACM, 2015.

[211] N. Eagle and A. S. Pentland, "Eigenbehaviors: Identifying structure in routine," *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1057–1066, 2009.

[212] A. Sadilek and J. Krumm, "Far out: predicting long-term human mobility," in *AAAI*, 2012.

[213] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "Nextplace: a spatio-temporal prediction framework for pervasive systems," in *Pervasive computing*, pp. 152–169, Springer, 2011.