From Raw Data to Social Systems - Separating the Signal from the Noise in Smartphone Sensor Measurements

Sapiezynski, Piotr

Publication date:
2017

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

Link back to DTU Orbit

Citation (APA):
From Raw Data to Social Systems
Separating the Signal from the Noise in Smartphone Sensor Measurements

Piotr Sapiezynski

DTU

Kongens Lyngby 2016
Abstract

Digital tools for communication and information exchange are now deeply ingrained in our everyday routines. This fact, however, does not mean that our lives have fully migrated to the digital domain. Even though the Internet makes it possible to survive without ever leaving the confines of our bedrooms, we still choose to meet our friends in person or to travel through physical, rather than virtual, space. There is a richness to personal contact and direct experience that has not yet been replaced by digital services. Until this shift happens, we continue to analyze and investigate our offline lives in the pursuit for deepening our understanding of humans and societies. The digital traces, which we leave behind with every online activity, are relatively easy to collect. On the other hand, capturing our offline behaviors remains a challenge. Scientists often rely on data that approximates only one facet of our lives. For example, mobile operator records reveal who we call, but not who we meet. An alternative approach is to derive proxies of certain behaviors from smartphone sensor readings. Copenhagen Networks Study (CNS) employs this method, among others, to build the largest dataset of the kind available to academics.

From a bird’s eye view, the thesis shows a path from collecting raw smartphone data for CNS, through extracting increasingly meaningful information, to gaining novel insights into human behavior. Step by step, the goal is to turn a cryptic collection of hardware identifiers and received signal strengths into a detailed record of people’s mobility, person-to-person interactions, and social ties. The methods for location and interaction sensing that I propose in the thesis constitute a more privacy-aware alternative to currently employed approaches. At the same time, the findings presented in this work emphasize the fragility of our privacy: the data we today consider as safe to share, might prove to carry unexpected and rich information about our lives tomorrow.
Digitale tjenester til formidling af kommunikation og informationsudveksling er i dag en uadskillelig del af vores dagligliv. Dette er dog ikke ensbetydende med, at vores liv fuldstændigt er migrerede til det digitale domæne. Selv om internettet muliggør overlevelse uden nogensinde at skulle forlade soveværelset, vælger vi stadig at møde vores venner personligt eller rejse gennem det fysiske, frem for det virtuelle rum. Der er en dybere rigdom forbundet med personlig kontakt og direkte oplevelse, som endnu ikke er blevet erstattet af digitale tjenester. Indtil dette skift finder sted, fortsætter vi med at undersøge og analysere vores offline liv i jagten på at uddybe vores forståelse af mennesker og samfund. Hvor indsamling af digitale spor fra vores online aktiviteter er ligetil, er vores offline liv fortsat en empirisk udfordring. Forskere begrænser sig ofte til data vedrørende et enkelt aspekt af vores liv. For eksempel kan et teleselskabs opkaldsdata fortælle, hvem vi ringer til, men ikke hvem vi mødes med. En alternativ tilgang er at benytte sensorer i smartphones til at udlede visse adfærdsmønstre. Copenhagen Networks Study (CNS) har anvendt denne metode, blandt andet, til at skabe det største datasæt af sin slags i den akademiske verden.

Overordnet set viser denne afhandling vejen fra indsamling af rå data fra smartphones for CNS, til udvindingen og destilleringen af meningsfuld information, resulterende i nye indsigter i menneskelig adfærd. Skridt for skridt er målet at omdanne en kryptisk samling af hardware ID’er og opsamlede signalstyrker, til en detaljeret optegnelse af menneskers mobilitet, person-til-person-interaktioner og sociale bånd. De metoder jeg foreslår til at detektere lokation og interaktion, udgør et mere privatlivsbeskyttende alternativ til de for tiden anvendte metoder. De fund der fremlægges i denne afhandling understreger ydermere skrøbeligheden af vores privatlivs fortrolighed: de data som vi betragter det som sikkert at dele i dag, kan indeholde uventet og rig information om vores liv i morgen.
List of publications


Personal preface

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

T.S. Eliot

Three years ago I decided not to pursue a safe career path in the industry and do Science instead. I set out to broaden our understanding of how an individual’s social ties and position in the network affect measurable life outcomes, for example their success in higher education, health problems, or general well-being. Three years later, an answer to the question seems to be even further away than it seemed back then and yet I have learned so much.

What has happened?

I decided to do things properly. If we were to understand the influence of ties on life outcomes, we should first understand what constitutes a tie. Even the richest behavioral data will not reveal the meaning and the nature of human relationships. However, we can start with the well established assumption that prolonged interactions induce emotional evaluations and responses [97]. Propinquity (psychical and psychological proximity between people) is a key factor in forming bonds between humans. The need for face to face contact has not been replaced by the ever-growing possibility of remote interaction [7]. The data available to me as a researcher working on the Copenhagen Networks Study included two years worth of records of face to face interactions among a large cohort of students. I had reasonable means to measure face to face interactions, I had the necessary time scale, and I could have (should have?) just built on this to predict life outcomes.
I didn’t.

What if our reasonable means to measure face to face interaction were not there? Given the volatile approach of smartphone manufacturers to the technology we relied on (Bluetooth), we could at any moment lose access to the crucial building block of our study. Moreover, in general context outside of our controlled experiment, it is not reasonable to expect Bluetooth data to be available to researchers (for reasons I describe in Chapter 7). Therefore, I decided to explore alternative channels and I focused on investigating the potential of WiFi signals. The initial findings were very promising, with WiFi being quite a well performing proxy for the information obtained via Bluetooth. I had reasonable means to measure face to face interactions using two independent channels, I had the necessary time scale, and I could have (should have?) just built on this to predict life outcomes.

I didn’t.

It is fairly intuitive that the interpretation of an interaction depends on the context in which the interaction happens. Bluetooth, and the simple WiFi approach we implemented, gave us the social context (who is present) and the temporal context (when does the interaction happen and how long does it last) of each interaction. What was missing from the picture was the location context. The straightforward way to enrich our understanding of each meeting is to look at the GPS data, which we were also collecting. I had reasonable means to measure face to face interactions using two independent channels, I had an idea of how to interpret the context, I had the necessary time scale, and I could have (should have?) just built on this to predict life outcomes.

I didn’t.

The temporal resolution of the location data we were collecting was constrained by the battery life. It was noisy too: while trajectories of students traveling faster than the speed of light were easy to filter out, smaller errors could still influence the interpretation of the context. Working on cleaning the location data was not of interest to me: I intended to work on social ties and life outcomes! Location was just a means to an end, not a goal on its own! Given the much higher temporal resolution of WiFi data, we decided to use it to enrich the location traces. First, however, we needed to find out where access points were located. As the students explored Copenhagen, their phones constantly scanned for WiFi networks and occasionally for GPS estimations. All we had to do is to combine the two data sources to estimate the locations of routers. That seemed easy.

It wasn’t.

It took one student’s Master’s thesis to conclude that the problem was not trivial but had a solution. After a few more months of research and one publication later, we arrived at a solution that was close to satisfactory. In the process, we realized that WiFi data reveals much more about the owner of a phone collecting it than we initially expected. We know that the access points can be used to aid location estima-
tion. However, it is not obvious that the WiFi scans alone, even without combining with GPS, reveal rich mobility patterns and leak potentially sensitive information, for example: when people are away from their homes, or how much time they spend at work. We found that these results can be achieved in a more straightforward manner than using actual location data. I had reasonable means to measure face to face interactions using two independent channels, I had a good idea of how to interpret the context, I had the necessary time scale, and I could have (should have?) just built on this to predict life outcomes.

I didn’t.

In the course of the research on the connection between WiFi and location we also showed that 80% of applications on Android Play Market had access to the rich WiFi data, even without the permission to obtain the location of users. The methods we described made it feasible to extract a small number of routers to represent the detailed routine of individuals, without having to know the location of thousands of access points each of them passes by everyday. Publicizing this immediate privacy threat seemed like a more timely pursuit than social ties and life outcomes. Soon after the publication of the article and a release of our mobile proof of concept app, Google announced the relevant change in privacy policy and has since solved the problem (coincidence? I think... it very well might have been). And so, as a side effect of our research, we drew the attention of multiple people to important privacy problems. With the privacy aspect of WiFi information out of the way, I had reasonable means to measure face to face interactions using two independent channels, I had a good idea of how to interpret the context, I had the necessary time scale, and I could have (should have?) just built on this to predict life outcomes.

I didn’t.

Dissatisfied with the performance of the simple WiFi approach to detect face to face interactions, I continued exploring alternatives. But even with a satisfying model in place, I was still missing the key ingredient - the social ties. We know from Krackhardt that proximity is the fundamental requirement for the ties to form, but clearly not all repeated proximity leads to bonding. In the university context students spend hours and days together but they do not become friends with all their classmates. Before exploring the academic outcomes, I first investigated how to reliably recognize the few meaningful relationships in a huge crowd of constantly interacting people.

And it is at this stage that this thesis ends. At times it feels like I did not get any closer to answering my initial question. I do know, however, I walked a long way and after the three years I realize just how much further I was from any answers in the beginning. Then again, as Einstein (might have) said: If we knew what we were doing, it would not be called research, would it?1

1While the quote is now popularly attributed to Albert Einstein, it had initially circulated without an author and its origin is unclear. See https://goo.gl/FFI58N for more details.
None of the work described in this thesis would have been possible were it not for the incredible people I had the pleasure to interact with along the way. I am forever grateful!

To Sune Lehmann for pulling me into academia (just when I thought I was out), and for creating an environment full of good science, psychological safety, and burgers.

To Jure Leskovec for inviting me to Stanford, as well as to Michal Kosinski and Bruno Abrahao for helping Jure make this decision.

To my co-authors for giving me the chance to learn from them (in order of appearance): Arek Stopczynski, Jakob Eg Larsen, Alan Mislove, Vedran Sekara, Andrea Cuttone, Mette My Madsen, Radu Gatej, Alex ‘Sandy’ Pentland, David Kofoed Wind, Magdalena Anna Furman, Jure Leskovec, Laura Alessandretti, and Andrea Baronchelli.

To the people I have collaborated with on projects that have not yet seen the light of day: Yves-Alexandre de Montjoye, Laura Radaeli, Emilio Ferrara, Bjarke Mønsted, Ioanna Psylla, and Valentin Kassarnig.

To those whose work cleared my path of obstacles and enabled me to focus on research: Lars Kai Hansen, Otto Monsteds Foundation, and the Villum Foundation.

To Wojciech and Ask of The Shittens, who made sure I was never too preoccupied with my work, and to Arek “Ping?” Stopczynski, who made sure I could never take my mind off it either.

To Alina who pushed me to start writing this thesis, and to the patient readers for making the outcome more digestible: Piotr, Anna, Laura, Milos, Ancsa, Andrea, Bjarke, Snorre, Daniel (and his father!), and Nathaniel. You guys rock.

To my extended family of The Royal Palace for being themselves and helping me do so too. And for understanding why I do not put their names here.

To my parents for raising me to be curious, and to my sister for teaching me how to read, write, and show things on maps. I have been using these skills quite a lot recently.

Finally, I thank Anna for not just bearing with me through all the stages of my short academic career (which, I imagine, was hard enough) but for being a partner. You challenge me at every step but also give me support, inspiration, and a reason to pull through. And ice cream. Thanks.

Thank you.
Contents

Abstract i
Summary (Danish) iii
List of publications v
Personal preface vii
Acknowledgements xi

1 Introduction 1
  1.1 Computational Social Science ........................................... 1
  1.2 Life in the network of sensors ........................................... 2
  1.3 Contributions in this thesis ............................................. 3

2 A brief introduction to networks 5
  2.1 Basic definitions ............................................................ 5
    2.1.1 Degree distributions .................................................. 6
    2.1.2 Clustering coefficient ................................................ 6
  2.2 Types of networks .......................................................... 7
    2.2.1 (Un)directed networks ............................................... 7
    2.2.2 Weighted networks ................................................... 8
    2.2.3 Temporal networks ................................................... 8
    2.2.4 Multiplex networks .................................................. 8
  2.3 Random vs. real world networks ....................................... 9

3 A brief introduction to classification 13
  3.1 The basics ................................................................. 14
  3.2 Performance curves ...................................................... 16
# Contents

3.3 Other metrics .................................................. 17  
3.4 Choosing the threshold ....................................... 18  
3.5 The class imbalance problem .................................. 19  

4 Data collection ..................................................... 21  
4.1 Data types ....................................................... 22  
  4.1.1 Background information .................................. 22  
  4.1.2 Online activity ............................................. 23  
  4.1.3 Sensor data ................................................ 23  
4.2 Sources of potential bias ....................................... 24  
4.3 Incentivisation ................................................... 24  

5 Location ........................................................... 27  
5.1 Location techniques and their reference points .......... 27  
  5.1.1 Trilateration ............................................... 28  
  5.1.2 Radio Frequency Fingerprinting ......................... 29  
5.2 Learning the beacon positions ................................. 29  
  5.2.1 Manual reporting .......................................... 30  
  5.2.2 Anchored inference ....................................... 31  
  5.2.3 Wardriving ................................................. 31  
5.3 Precision, deployment complexity, energy efficiency ... 32  
5.4 Summary of “Opportunities and Challenges in Crowd-sourced Wardriving” ...................................... 35  

6 Mobility ............................................................. 37  
6.1 Properties of human mobility .................................. 37  
6.2 Representing mobility ........................................... 39  
  6.2.1 Stop locations .............................................. 40  
6.3 Summary of “Tracking Human Mobility using WiFi Signals” .................................................. 43  
6.4 Summary of “Inferring Stop-locations from WiFi” ...... 46  
6.5 Summary of “Conservation Laws in Human Mobility” ... 47  

7 Person-to-person interactions ................................. 49  
7.1 Measurement methods .......................................... 50  
  7.1.1 Peer-to-peer sensing ....................................... 50  
  7.1.2 System-to-peer sensing .................................... 52  
  7.1.3 Peer-to-system sensing .................................... 53  
7.2 Epidemic spreading .............................................. 55  
7.3 Summary of “Inferring Person-to-person Proximity from WiFi Signals” ........................................ 57  
7.4 Summary of “Temporal Fidelity in Dynamic Social Networks” .................................................. 58  

8 Offline social networks and their online representations 59  
8.1 Inferring social ties from co-presence events ............ 62  
8.2 Communication networks as proxies for real-world relationships .................................................. 63  
8.3 Conservation of ties over time ................................. 64  
8.4 Summary of “Offline behaviors of online friends” ...... 65
## CONTENTS

9 Contributions to privacy 67  
9.1 “I have nothing to hide” 68  
9.2 Outreach 68  
9.3 Impact on future studies 70  

10 Summary 71  

A Measuring Large-scale Social Networks with High Resolution 73  
B Opportunities and Challenges in Crowdsourced Wardriving 99  
C Tracking Human Mobility using WiFi signals 107  
D Inferring Stop Locations from WiFi 127  
E Conservation laws in human mobility 143  
F Temporal Fidelity in Dynamic Social Networks 197  
G Inferring Person-to-person Proximity Using WiFi Signals 205  
H Offline Behaviors of Online Friends 217  

Bibliography 227
Chapter 1

Introduction

Progressively more aspects of our lives are now being recorded and stored for later analysis. There is a lot of hope and plenty of fear in the anticipation of the outcomes of this process. It will indubitably provide insights into human nature and the structure of societies at an unprecedented scale. At the same time, it has already started to enable unchecked intrusion into our private lives.

1.1 Computational Social Science

The new, data-driven, Computational Social Science aims at answering important questions about our societies using the now-available massive datasets [104]. The deep scientific insights, however, are arriving slowly due to a number of problems concerning data quality and scale. One of the major issues is the fact that data collected from different channels resides in separate, disconnected databases. Our mobile phone operator knows whom we call, but not whom we interact with online. Our bank knows how much money we spent in a restaurant, but not whom we shared the food with. No single channel can convey the different facets of our lives, and especially not so when it is not understood how the different channels relate to each other. Furthermore, the data collection is usually event-driven: we learn about the state of an individual when they perform a specific type of action (call a friend, use a credit card), but we remain in the dark as to what happens with them in the interim. Finally, the majority of the studies are performed on datasets that only span periods from a day
to a few months. Such time spans might not be enough for the researchers to observe long-term changes in human behavior. The Copenhagen Networks Study [170] is an attempt to address these three problems. We collected two years of data from different channels about a group of \(\sim\)1000 people: location of individuals, online exchanges, calls and messages, person-to-person interactions, and rich personal characteristics. The collection happened at specified intervals, regardless of the activity of the person. We, therefore, facilitated the creation of probably the most comprehensive dataset about human behavior available in academia. This endeavour provided enough technical challenges to occupy multiple people for a number of years. We shall see in the coming years whether it begets the scientific breakthrough we strived to enable.

1.2 Life in the network of sensors

Our world is progressively more enclosed in infrastructure whose primary goal is to support communication, mobility, payments, and advertising. To achieve this goal, systems built on top of this infrastructure collect the necessary data and, as a side effect, create immensely rich logs of human activity at an unprecedented scale. Mobile operators store the information of who-calls-whom along with the time and location of each subscriber to enable billing and make informed decisions about network expansion. In hands of data analysts, these very same logs are transformed into diaries of people’s social networks and travel routines. As the developed countries move away from cash in favor of credit cards and mobile payments, high resolution traces of every individual’s spending behaviors are created. On one hand, people do not lose wads of cash on a night out anymore. On the other, your bank will reconstruct and remember the unfolding of that night’s events; events which you might prefer to forget, if you didn’t already. Public transportation systems transition from relying on paper tickets to using smart cards with RFID chips, thus enabling the storage and analysis of every trip. Again, the main reason is to enable billing and transportation network analysis. As a side effect, however, the databases are flooded with identities of people who go out on a weekday and end up late for work the day after. Car navigation systems were created to assist people in getting to their destinations quickly. In the times of data economy, your navigation system sells the information about how quickly exactly you got there to the insurance companies or the highest bidder.

Finally, the WiFi systems. Their primary purpose is to provide Internet connectivity. However, it is now a recognized fact that they are used to track people by various organizations: the Police [90] and intelligence agencies [159], cities [115], airports [139], and advertisers in shopping centers [145]. Our modern-day smartphones constantly search for WiFi networks in an active way. This means that as often as every 15 seconds each phone sends a probe request — it advertises its presence and its unique identifier to all nearby WiFi routers (unless the user explicitly configures the phone not to do it). Access point listen to those requests and some access points are configured to store this information. When a single entity controls multiple WiFi systems or a system that spans a significant geographical area, it can reconstruct detailed mobility traces based on those collected requests.
In this thesis I present my work on the data that was collected in a complementary fashion. Once a router receives the probe request from a mobile phone, it responds with its own unique identifier. This response is then received and stored by the phone. As a consequence, each phone stores the list of nearby WiFi access points as often as every 15 seconds. And, as I learned during the three years of my PhD research, there is a whole lot one can do with this information. This realization becomes even more striking once you know that access to this information is granted to virtually all applications running on Android versions older than version 6 (Marshmallow). So, what can be inferred from the list of WiFi routers?

1.3 Contributions in this thesis

Location. Nearby WiFi access points are used by mobile phones to coarsely estimate the location of the user. However, as I show in Chapter 5, the estimations are far more precise than what one would refer to as coarse. Knowing the list of routers scanned by the person’s phone we can estimate that person’s location within 15 meters. To collect the information about the routers’ locations we did not have to send out specialized cars or use professional-grade equipment. Instead, the participants of the Copenhagen Networks Study collected this data as they stumbled upon access points across the city.

Mobility. Because today’s phones scan the WiFi environment so often, it becomes possible to track the mobility of individuals with a sub-minute temporal resolution. This would not be possible if we were to rely on GPS due to the detrimental influence of this sensor on the battery life. In Chapter 5 I describe how, through a relatively simple analysis of WiFi data, the spatio-temporal details of our daily routines become apparent: where and how long we work; whether we use public transportation; where we eat lunch and how much time we spend doing so; when exactly we are at home and when we are not; how active we are throughout the day, etc. These details are crucial from the perspective of privacy and we should be able to control who can access them. At the same time, being able to track human mobility with such an unprecedented spatial and temporal resolution can potentially have a significant impact on urban planning, public health studies, and deepen our understanding of such aspects of mobility as predictability\(^1\), stability\(^2\) and exploration\(^3\).

---
\(^1\)predictability is a quantification of the extent to which movements of an individual can be predicted based on the past
\(^2\)stability is the tendency of an individual to return to the same locations over long time periods
\(^3\)exploration is a behavioral pattern in which an individual chooses to seek a new location/experience instead of repeating an action they had already tried
**Interactions.** Person-to-person interactions are indispensable for creating social bonds and, at the same time, instrumental for spreading certain diseases. In Chapter 7, I explain how close proximity interactions can be inferred by comparing the lists of nearby routers from two individuals. Furthermore, because of our progressing envelopment in the WiFi infrastructure, there are dozens of visible WiFi routers at any point in urban space. This allows us to accurately estimate proximity between two people even in environments where we do not know the locations of the routers. Finally, contrary to previous attempts at using WiFi to infer person-to-person encounters, we do not rely on the properties of the signal propagation alone. We also exploit the properties of human interactions, for example their circadian and weekly rhythms.

**Social connections.** Social ties constitute a fundamental building block of our society. The relationships we form influence our health and well-being, make or break our careers, help us quit smoking (while making us obese), and tell more about ourselves than we would care to share. In Chapter 8, I show how one can infer the social ties between a group of people from a longitudinal log of person-to-person proximity measured using WiFi. Based on the Copenhagen Networks Study data, we show that it is possible to find interacting friends among the vast noise of serendipitous encounters between familiar strangers. Furthermore, there are certain behaviors observable through the WiFi scan results that hint at the nature of the relationship. For example, we can discern whether two people are just friends on Facebook or whether they also call each other.

**Privacy.** The fact that countless app developers have had the chance to collect WiFi data for years and learn about each of multiple aspects of our lives has tremendous privacy implications. Since our publication, Google partially remodeled the permission system in Android. However, only a small fraction of devices will have access to the updated software. It is, therefore, crucial to disseminate the knowledge on how one can protect their personal data. As I point out in Chapter 9, our results were covered in mass media and our proof-of-concept application has been downloaded more than 3000 times. Moreover, we have developed methods that make it possible to deploy organize computational social science experiments without relying on technologies that inadvertently share the participants’ data with third parties. In the task of detecting interactions, Bluetooth can be successfully replaced by comparing WiFi scans. Google’s WiFi routers location database can be substituted with a solution developed in-house. I hope that, through the work I have been doing over the last three years, I have contributed more to raising awareness than to making malicious tracking more commonplace.

---

4 In Chapter 8, I support these claims with appropriate citations.
5 At the time of writing, 5.6% of devices ran the Android version with the updated permission model.
In this part of the thesis I offer a crash course into fundamentals of graph theory and network science. In the chapters to follow I employ the network paradigm to represent the complex system of the Copenhagen Network Study participants and their interactions. Basic understanding of network properties and related vocabulary will facilitate reading the rest of the manuscript.

I have been inspired primarily by Chapters 2 and 9 of “Network Science” by Albert László-Barabási [15], Sune Lehmann’s course “Social graphs and interactions” [105], during which I worked as a teaching assistant, and Peter Holme’s “Modern temporal network theory: a colloquium” [78].

2.1 Basic definitions

A network is a powerful way of representing a complex system. It describes the elements the system is built from and the interactions, or dependencies, between these elements. In the network representation, each element of the system is a node. An interaction between two nodes is represented by a link. A network has $N$ nodes and $L$ links. The number of links that connect a node to other nodes is the degree of that node.
A brief introduction to networks

Figure 2.1: A network with five nodes and six links. Nodes are elements of the system, links symbolize the interactions between them.

A network can be visualized, see Figure 2.1. The network in the example has \( N = 5 \) nodes and \( L = 6 \) links. Remember that a visualization like this is symbolic and the placement of the nodes does not necessarily correspond to the placement of elements they represent in physical space. Furthermore, the distances between nodes in the visualization do not necessarily correspond to physical distances between the elements.

2.1.1 Degree distributions

In the previous section I introduced the notion of a node degree, \( k \) - the number of links that connect to the node. Degree distribution, \( p_k \), is a function that specifies the probability that a randomly selected node in the network has a degree \( k \). Figure 2.2 shows the degree distribution of our first example network. The degree distribution affects a number of phenomena, from network robustness to spreading of viruses \[15\]. These distributions in real-world social networks are very skewed: most of the nodes have very few links, while a few nodes have a lot of links.

Figure 2.2: A network and its corresponding degree distribution. There are a total of five nodes in the network. One node (20%) has a degree of one, two (40%) have degree of two, one of three and one of four.

2.1.2 Clustering coefficient

Clustering coefficient \( C_i \) for a node \( i \) describes the extent to which the neighbors of this node link to each other. It is the fraction between the number of existing links among the neighbors of that node and the total number of possible links among them.
If a node has $k$ neighbors, there can be $\frac{k(k-1)}{2}$ links between them: each node can link to all nodes apart from itself (hence $k(k - 1)$ in the numerator) and links are not directional (hence 2 in the denominator). Figure 2.3 shows three networks, along with the clustering coefficient of the red node.

Figure 2.3: Clustering coefficient of a node quantifies how well the node’s neighbors are connected to each other. The red node $i$ has four neighbors. Therefore, there can be at most six links between them. There are six links in the left panel (hence $C_i = 1$), three links in the center panel (hence $C_i = 0.5$), and no links in the right panel (hence $C_i = 0$).

2.2 Types of networks

In this section I present a number of types of networks. Importantly, the types I list here are not mutually exclusive. A network be directed, weighted, temporal, and multiplex at the same time.

2.2.1 (Un)directed networks

The links in the network can be undirected or directed (see Figures 2.1 and 2.4 respectively). For example, the Facebook friendship graph is undirected — it means that if $A$ is a friend of $B$, then $B$ is also a friend of $A$. On the other hand, the Facebook interaction network is directed — the fact that user $A$ sends a message to $B$, does not mean that $B$ will necessarily reply. The links in a directed network can be bi-directional. In directed networks, each node has an in-degree (number of incoming links) and an out-degree (number of outgoing links). The direction of the link is marked in the network visualization using arrowheads, see Figure 2.4.
2.2.2 Weighted networks

The interactions between different pairs of elements in a system can vary in intensity. We symbolize this principle using weighted links. The weight of the link can, for example, describe the frequency of sending messages on Facebook, or the total time two people spend in physical proximity. In weighted networks, the degree of the node is the sum of the weights of its links. The weight of a link can, for example, be represented by varying the width of the line that symbolizes it, see Figure 2.5.

2.2.3 Temporal networks

In temporal networks, the presence of nodes and links can change over time. They can be used to represent, for example, the dynamic nature of human interactions: in the morning we interact with the family, then with co-workers during the day, and with friends in the evening. On a longer timescale, they can also symbolize the formation and dissolution of social ties, etc. In temporal networks, the degree of a node changes dynamically and is only defined for a specified time period. There is a number of ways to represent a temporal network. I present a few examples in Figure 2.6.

2.2.4 Multiplex networks

Multiplex networks (also referred to as multilayer or multidimensional networks) are used to express systems in which different kinds of relations can occur between the same set of nodes. For example, a group of people can use a number of communication
channels: email, phone calls, or face-to-face meetings. We use different layers to represent these channels. When visualizing a multilayer network, one can, for example, mark the links at each layer with a different color (see Figure 2.7), or present each layer on a separate graph.

**Figure 2.7: A multiplex (multilayer) network.** There can be multiple types of interactions among the same set of nodes. Each network layer can correspond to a different communication channel, and each pair of nodes can interact on multiple layers.

### 2.3 Random vs. real world networks

Figure 2.8 shows a visualization of three networks. They all have the same number of nodes and edges. Node radius in the graphs corresponds to the degree of the node.
The network in panel (a) is a real world network. Each node represents one of my Facebook friends (I am not in the graph). There is an edge between two nodes if the two friends are also friends with one another. The degree distribution is skewed (most of the nodes have very few connections, there are a few nodes with a lot of connections). The distribution of local clustering coefficient is quite even across the range.

The network in panel (b) is a random network generated by shuffling the links from the real network in panel (a): we move each link in the network so that it connects two random nodes. In each random selection every node has the same probability of being selected. The number of links and nodes is the same as in panel (a), but note that the degree distribution is very different: it is centered around the mean, and there are no nodes with a very high degree (these nodes are referred to as hubs).

The network in panel (c) is also a random network, but in this case each node had a different probability of being selected. This probability corresponded to the degree of the node in the network in panel (a). The degree distribution, as well as node and edge counts remain the same as in panel (a). Note, however, that the distribution of the clustering coefficient resembles the distribution from panel (b). Contrary to the network in panel (a), the two random networks do not have clusters, or groups of highly interconnected nodes.

There is structure in real world networks beyond the degree distribution. The phenomena driving this structure include, among others, assortativity (the preference of nodes to form links with similar nodes; for example, nodes with high degree connect to other nodes with high degree), preferential attachment (the process through which nodes that already have many links are more likely to form more), and transitivity (the tendency to close triads: if node $A$ is linked to $B$, and $B$ is linked to $C$, $A$ will likely connect to $C$). Detailed description of these phenomena is beyond the scope of this short introduction.
Figure 2.8: Three networks with the same number of nodes, edges, and average degree. Panel (a) shows the connections between my Facebook friends. Panel (b) is a random network with the same number of nodes and edges but a normal degree distribution. Panel (c) is a random network with the same number of nodes and edges, as well as equal average degree, and degree distribution as the network in panel (a). The clusters do not form in the randomized versions of the real network.
A brief introduction to networks
Chapter 3

A brief introduction to classification

In the later chapters of the thesis I assume that the reader is familiar with the problem of binary classification and the measures of classifier performance. I employ machine learning methods to solve classification problems such as detection of the user state (stationary/moving) or link prediction in networks. In this chapter, I cover the essential definitions and concepts.

Supervised machine learning is a task of inferring a function from labeled training data [132]. In practice, a dataset is divided into mutually exclusive training and testing subsets. A supervised machine learning method tries to learn a function that best describes the dependency between the variables and the outcomes in the training set. The learned function is then applied to the observations in the test set and the predictions it makes are compared to the true outcomes. The problems pertaining to the supervised machine learning can be divided into two groups, based on the type of the predicted variable. Regression is a task of inferring the value of a continuous variable, for example, predicting air humidity or stock price. Classification is a task of inferring the category (class) to which an element belongs, for example, predicting the gender of a person, or what kind of physical activity a person is undertaking. In my work I used binary classification, i.e. one that has only two possible classes, to infer the state of a person (moving/stationary) and to infer friendships between participants of the Copenhagen Networks Study (friend/not friend). In this chapter I briefly explain how we measure the performance of binary classification methods. Then, I introduce the problem of class imbalance and its influence on the performance metrics.


3.1 The basics

Let’s imagine we operate a sonar and we are on a lookout for submarines. The image on the sonar is not clear but every time we see an object, we must decide whether it is a submarine or another, uninteresting object. We cannot refrain from making a decision. We are rarely completely convinced which class the object belongs to. However, before we make each decision, we are able to quantify how certain we are that the object is a submarine. We have the object and our estimation of the likelihood that it is a submarine; now comes the time to make a decision — should we sound the alarm or just assume the object is a fish? If we only sound the alarm when we are completely certain, we are not going to cause any false alarms. However, we also risk missing an actual enemy submarine and thus endangering our fellows. On the other hand, if we do alarm them when we are not completely certain, we minimize the probability of missing the enemy, while increasing the chance of sounding false alarms. The decision is a difficult one to make, but, fortunately, this is just a training in a simulator. Every time we make a decision, we note it down along with how certain we were that it was a submarine (e.g: object 10: 70%, submarine; object 11: 20%; not a submarine\(^1\)). After a few hundred trials, we get feedback on how well we did. Each of the decisions we made falls into one of these four categories:

- if the object was a submarine and we classified it as such (we made a positive decision), it is a true positive, \(TP\)
- if the object was not a submarine and we classified it as not a submarine (we made a negative decision), it is a true negative, \(TN\)
- if the object was not a submarine, but we thought it was (we made a positive decision), it is a false alarm, false positive, \(FP\)
- if the object was a submarine, but we missed it (we made a negative decision), it is a false negative, \(FN\)

Table 3.1 summarizes these definitions.

<table>
<thead>
<tr>
<th>Actual condition</th>
<th>Predicted condition</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True positive, (TP)</td>
<td></td>
<td>False negative, (FN)</td>
</tr>
<tr>
<td>Negative</td>
<td>False positive, (FP)</td>
<td>True negative, (TN)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Basic definitions in binary classification.

Each of our decision falls into one (and only one) of these categories. The other metrics of performance that I introduce in this chapter are based on comparing the number of decisions in each category. In the following definitions I replace the word “submarine” with target and “not a submarine” with distractor.

\(^1\)Here “20%, not a submarine” means we are 20% certain it was a submarine. It is equivalent to saying we are 80% certain it was not a submarine.
3.1 The basics

Accuracy, $ACC$ is the fraction of decisions that were correct:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$ (3.1)

It assumes a value between 0 (when all decisions, both positive and negative, are wrong) and 1 (when all decisions, both positive and negative, are correct). While it might seem that accuracy is a good metric of performance, it is rarely used in real applications. In situations with far fewer targets than distractors, a classifier can achieve a high accuracy by always making a negative decision. For example, if the data has 100 targets and 900 distractors, a classifier that always indicates “negative” has a high accuracy of 0.9.

Precision, or positive predictive value, $PPV$, is the fraction of positive decisions that are correct:

$$PPV = \frac{TP}{TP + FP}$$ (3.2)

It assumes a value between 0 (all positive decisions are false positives) and 1 (all positive decisions are true positives). Since it only evaluates positive decisions, precision alone could give a distorted picture of the performance. For example, the classifier makes just one positive call, and it is correct, but there are also 99 other targets that the classifier mislabeled as distractors. In such case $PPV = 1$ but, depending on the application, one would not describe this performance as satisfactory. To paint a fuller picture, precision is usually reported together with recall, also referred to as True Positive Rate, $TPR$ - the fraction of all targets that the classifier identified:

$$recall = \frac{TP}{TP + FN} = TPR$$ (3.3)

Recall also assumes a value between 0 (none of the actual targets were correctly identified) and 1 (all targets were correctly identified), and the value is not sensitive to how many distractors were classified as targets. In our example with 99 missed targets, recall would only be 0.01. Similarly to $PPV$, recall alone might give a distorted impression of the performance: a classifier that always makes a positive prediction, has a recall of 1, but its $PPV$ suffers.

The fraction of distractors that a classifier incorrectly classifies as targets is called the false positive rate, $FPR$:

$$FPR = \frac{FP}{FP + TN}$$ (3.4)

Like the previous measures, it assumes a value between 0 (none of the distractors were misclassified as targets) to 1 (all distractors were misclassified as targets). This value is not sensitive to how many targets were correctly identified.
3.2 Performance curves

Let us now return to the submarines. During the training in the simulator we noted down how certain we were about each sample. Now, that we know the correct answers, we can plot the probability distributions of our certainty for each class. Figure 3.1 shows an example of what these distributions can look like.

![Probability distributions and the definitions of TN, FN, FP, TP. On average, our estimate of certainty is higher when we are looking at a target, than when we are looking at a distractor. However, these distributions overlap. As we lower the threshold, we increase the number of the targets we identify (the light blue area of TP increases), but we also cause more false alarms (the dark red area of FP increases as well).](image_url)

We apply a threshold $t$ on our certainty: if we are at least $t$-certain, we assume the object is a target, otherwise, we assume it is a distractor. As we vary the threshold, the values of $TN$, $FN$, $FP$, $TP$ change, and so do $PPV$, $TPR$, and $FPR$. With a high threshold, we are strict when making a positive decision. This results in high $PPV$ and low $FPR$ (both of which are wanted), but also in low recall, $TPR$ (unwanted). By lowering the threshold, we achieve a higher recall, but also make more false positive mistakes, resulting in a lower $PPV$ and a higher $FPR$.

The curve, which shows how $TPR$ and $FPR$ are affected by lowering the threshold is called the **Receiver Operating Characteristic**, ROC, see the left panel of Figure 3.2. The function $y = x$ shows the performance of a random classifier (a classifier that does not know the difference between a submarine and a fish and guesses at random). A classifier whose curve is above the $y = x$ function performs better than random. The area under the classifier’s curve is called **Area Under Curve, AUC**, and is commonly used to summarize the performance of a classifier using one number.\(^2\)

\(^2\)AUC is approximated from the ROC by fitting a number of trapezoids under the curve and computing their area, see https://goo.gl/wHgwVi for an illustration.
Area Under Curve can be interpreted as the fraction of times a classifier scores the
target higher (has a higher confidence associated with the target) if presented with a
target and a distractor.

The curve showing the function of $PPV$ against $TPR$ is called **Precision-Recall Curve**, or P-R Curve, see the right panel of Figure 3.2. The performance of a random classifier is marked with the function $y = \frac{P}{P+N}$. This means that the precision of the random classifier is equal to the ratio between positive samples and all samples in the data. It and does not depend on the $TPR$ of the classifier at any threshold.

Because the curves in Figure 3.2 are above the random classifier lines, we say that we (as the classifier of submarines) perform better than random. However, our performance is not perfect. If it was, precision ($PPV$) would remain at 1 regardless of recall ($TPR$) and recall ($TPR$) would remain equal to 1 regardless of $FPR$.

**Figure 3.2:** Receiver Operating Characteristic curve (left) and the **Precision-Recall curve** (right). The performance of a random classifier is marked with the dashed line in each plot. Because the curves of our classifier are above the random classifier curves, we can say that it performs better than random. In the case of a perfect classifier, precision would remain at 1 regardless of recall and recall would remain equal to 1 regardless of $FPR$.

### 3.3 Other metrics

So far I introduced two metrics which express the performance of a classifier using one number: accuracy and area under receiver operating curve. Accuracy can give a misleading picture in imbalanced datasets and $AUC$ summarizes the performance
of the classifier for all thresholds. In any practical application, there a particular threshold needs to be set and used for classification. Here I present two metrics which indicate the performance at a certain threshold and, as opposed to accuracy, penalize classifiers that always predict the negative class.

\[ F_1 \text{ score} = \frac{2}{\frac{1}{TPV} + \frac{1}{TPR}} = \frac{2}{\frac{TP + FP}{TP} + \frac{TP + FN}{TP}} = \frac{2TP}{2TP + FP + FN} \] (3.5)

Matthews correlation coefficient, \( MCC \), indicates the correlation between the predicted and true values. It assumes a value between \(-1\) (predictions completely opposite to the truth), through \(0\) (not better than random), to \(+1\) (perfect predictions). It is defined as follows:

\[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \] (3.6)

### 3.4 Choosing the threshold

The question that remains is how to choose the right threshold? There is no general answer — it depends on the application. The most straightforward approach is to choose a threshold that maximizes one of the metrics I described, for example the \( F_1 \) score or \( MCC \). Another approach is to assume that a fixed number of test samples are targets and the rest are distractors, a method originally proposed in the context of link prediction [109]. Each predicted sample is assigned a probability of being a target. The samples are then sorted in descending order and the top \( k \) samples are assumed to be targets. \( k \) can be chosen to be proportional to the fraction of the target samples in the training data, i.e. if 10% of samples in the training data are targets and we are trying to predict 1000 samples, we predict the top 100 as targets. Alternatively, one can design a cost function — assign a cost to missing a target and another cost to causing a false alarm — and find the threshold at which the function reaches its minimum. For example, in case of classification problems in the financial domain, the costs might be straightforward to estimate, as each decision has a measurable monetary outcome. Sometimes, however, assigning an objective cost to making mistakes might be difficult or impossible, for example, when those mistakes concern human life. In such cases, researchers settle on an arbitrary accepted value of precision or recall. For example, in the problem of identifying potential terrorists, NSA has to balance a trade off between finding the highest fraction of them and minimizing the risk of targeting innocent citizens [9]. They use metadata of phone calls among 100,000 inhabitants of Pakistan. Seven people in the dataset are known terrorists. The classifier is trained using the information about six of them and has the task of finding the seventh. NSA, for the lack of a well defined cost function, settle for 50% recall[7]. At this fixed recall,
their classifier has 0.18% false positive rate, which is a great performance in most usual use cases. However, because of the class imbalance, it means that only one out of 51 “terrorists” picked by the model is an actual terrorist. Given the size of the Pakistani population, targeting half of the terrorists selected this way would also imply targeting 99,000 civilians.

3.5 The class imbalance problem

In our submarine example, there was an equal number of targets and distractors. In real-world applications, this is rarely the case — usually the distractors outnumber the targets, like in the case of the NSA program. A friendship detection task on real data is also highly imbalanced: one could be friends with any of the 7.125 billion people in the world4 but most people bond only with up to 150 individuals76.

Each performance measure reacts differently to the class imbalance. Therefore, it is crucial to carefully interpret the reported values. Figure 3.3 shows how the values of the different metrics change, when the class balance is varied from the same number of targets and distractors (imbalance equal to 1) to more distractors than targets (imbalance greater than 1). Clearly, the area under receiver operating characteristic is not affected by the class imbalance, see Figure 3.3a. This property of AUC is often argued to be its advantage both by research articles and in course books173. Remember, the interpretation of the AUC value is the answer to “how much does a typical negative example vary from a typical positive example?” or “what fraction of the time when a classifier is presented with one negative and one positive example is it able to identify the positive one?” Arguably, these questions do not reflect the reality of classification tasks in imbalanced datasets. A typical target might be very different from a typical distractor. However, a small fraction of distractors typically resemble targets. If the number of distractors is much higher than the number of targets, many will be misclassified as targets. Case in point, the NSA documents show that a journalist (a distractor) is the sample with the highest probability of being a terrorist in the data. In a real-world scenario, a classifier is not presented with two samples of different classes — that would require actually knowing the right answer a priori, thus defeating the point of classification.

Resampling is one of the methods popularly employed to handle imbalanced datasets. In this approach, one creates an alternative training set by choosing only a subset of distractors (under-sampling), or selecting the same targets multiple times (over-sampling). Resampling can be applied in cases where there is just not enough examples of targets for the classifier to learn their characteristics. The thresholds chosen to maximize the $F_1$ score or $MCC$ in the resampled training datasets remain optimal for the imbalanced test data. However, it is crucial to remember that the values of these performance metrics will be much lower in a real, imbalanced test set compared to the balanced training set. This problem is illustrated in Figure 3.3. Figure 3.3b-d shows

---

4estimate of world population according to the World Bank
three performance metrics measured at three thresholds: with 50% recall (as used in Ref. [62]) and with top k examples treated as positive (as proposed in Ref. [109]), and a threshold chosen to maximize the value of the particular metric (referred to as optimal). Figure 3.3b shows that accuracy grows with the class imbalance. This emphasizes why one should not optimize for accuracy in such datasets - it leads to classifying all samples as negative. MCC and $F_1$ appear to give realistic estimates: the reported value is the highest for a balanced dataset and drops in case of imbalance (Figure 3.3c). The figure only shows the class imbalance in which there are more distractors than targets. Arguably, this is the most common scenario in real world datasets.

![Figure 3.3: Impact of class imbalance on the reported performance of the classifier.](image)

The Area Under Receiver Operating Characteristic curve is commonly used to summarize the performance of models and to select the models that perform best on a given dataset. However, it is important to realize that a high value of this measure might be misleading if taken out of context. To be able to really judge the applicability of the model to a given problem, it is important to know its performance in a dataset with an expected class ratio and at a certain threshold. Only then can we decide whether the performance is satisfactory and the costs outweigh the gains.
Until recently, social scientists have not had access to large-scale quantitative methods of observing behavioral patterns and interactions among large populations. Instead, they relied on participant observation \cite{13} or information reported by subjects through surveys. This method of collecting data enables the researchers to ask in-depth, personal questions, possibly leading to thorough understanding of some aspects of the investigated individuals’ lives. However, the questionnaire data can be biased by low temporal resolution of the responses as well as by the limited capacity of subjects to objectively recall and report past events. For example, it has been shown that people fail to order their social contacts by the frequency of interactions \cite{19} and are biased towards recalling more recent events \cite{48}. Moreover, because of the cost and organizational complexity, surveys can not be administered to large-scale cohorts.

On the other side of the spectrum, “traditional” big data approaches, as for example using call detail records (CDR) or public online activity data, promise to alleviate some of these problems. They enable investigating populations that are orders of magnitude larger, with studies reporting between 2.5 million \cite{102} and 25 million \cite{188} participants in mobile phone experiments and as many as 41 million in one of the studies of Twitter \cite{100}. Furthermore, data collected this way is about events, not subjective perceptions, and its time resolution is not, in principle, limited. However, such data only offers access to a thin slice of human activities and, in practice, still suffers from non-contiguity, with samples of information created only when a subject performs a certain activity.

Using smartphones as social sensors appears to combine the best of both worlds — the
richness of survey information, and the scale of big data studies — to ultimately address the major problems. These personal devices can capture a multitude of information channels, from activity and mobility of individuals, encounters in the physical space, to communication events across a number of media. Moreover, smartphones are becoming ever more widespread with 60% of the market penetration in the developed world, ranging from 51% in Europe to 70% in North America at the end of 2014 [3].

The use of smartphones for social sensing has been pioneered by Alex ‘Sandy’ Pentland and his collaborators in the seminal Reality Mining experiment [47]. Notable deployments based on the reality mining idea include for example:

- Reality Mining [47], with 9 months of data about 100 participants.
- Lausanne data collection campaign [93], with 12 months and 170 participants.
- Friends and Family study [6], with 15 months of data about 130 individuals. The experiment was designed to investigate decision making and social influence.
- OtaSizzle [88], with 20 participants and an unreported duration.
- Social Evolution [117], with eight months of data about 60 individuals. The experiment was designed to study the adoption of political opinions and habits, spread of diseases, and the formation of social ties.
- NetSense [171], with two years of data about 200 individuals.
- Phone-Lab [134], a continuously developing deployment, initially with 288 devices, focused more on using smartphones for measuring telecommunication infrastructure than on social networks.
- StudentLife Study [177], with 10 weeks of data about 48 participants

To the best of my knowledge, Copenhagen Networks Study [170] gave rise to the dataset that covers the largest number of participants, spans the longest time period, and offers the highest spatio-temporal resolution of the data, among the reality mining datasets available to academics.

4.1 Data types

Our article “Measuring Large-Scale Social Networks with High Resolution” describes the experimental design of Copenhagen Networks Study in great detail. Here, I offer an overview of the data we collected.

4.1.1 Background information

Each student filled out a survey of 310 questions before participating in the experiment. The questionnaire was prepared by an interdisciplinary group of public health
4.1 Data types

Researchers, anthropologists, psychologists, and economists. It was served to the participants using a browser app I implemented. It included a number of questions pertaining to well established measures: Big Five Inventory (BFI), Rosenberg Self Esteem Scale [147], Narcissism NAR-Q [11], Satisfaction With Life Scale [45], Rotters Locus of Control Scale [148], UCLA Loneliness scale [149], Self-efficacy [163], Cohens perceived stress scale [35], Major Depression Inventory [17], The Copenhagen Social Relation Questionnaire [116], and Panas [178]. Furthermore, the participants were asked about their exercise habits, sleep patterns, and substance use.

Answering the questionnaire was a necessary condition of participating in the experiment. Additionally, at the end of every semester the students were asked to answer another wave of surveys. However, the response rates dropped over time.

4.1.2 Online activity

The participants of the study were given a choice of sharing their Facebook data with us. Every 24 hours we collected every type of information visible about the users who opted in: their wall content, birthday, education, friend lists, friends, groups, hometown, interests, likes, location, declared political views and religion, work, and gender. We did not collect the content nor the meta data of private messages exchanged between the users, since Facebook does not offer such access to developers.

4.1.3 Sensor data

Each participant of the study was provided with an LG Google Nexus 4 smartphone. We based our data collection application on the Funf framework [6]. All the probes were configured to periodically query for updates of their respective data types. Additionally, the probes collected all the samples that were requested by other applications running on the phone. For example, the WiFi probe was configured to scan for WiFi routers only every 10 minutes. However, the Android system performs a scan every 15 seconds unless instructed otherwise by the user. Therefore, we have obtained WiFi scan results every 15 seconds for a vast majority of users.

Location and mobility. The collector app measured the user’s position using GPS and network based location estimator every five minutes. Furthermore, every 10 minutes the collector also stored the identifier of the cell tower the phone was connected to. The location of users can be also inferred from their WiFi scan results, which, in our dataset, have a sampling frequency varying from 15 seconds to 10 minutes. In Chapter 5 I describe the differences between the different sources of location data.
Data collection

Interactions. We used Bluetooth as a proxy for person-to-person contact. Each phone was configured to be discoverable by other phones and to scan for nearby Bluetooth devices every five minutes. Because of its relatively short range (up to \(\sim 10\) meters), Bluetooth is often used in similar deployments to track face-to-face interactions [47, 6]. Furthermore, as I show in Chapter 7, the person-to-person encounters between participants can be inferred from their WiFi scan results. Finally, we also collected the logs of metadata about phone calls and short messages: the phone numbers involved and timing, without the actual content.

What we did not collect. There are a few data sources in modern smartphones that we decided not to gather. Firstly, we resigned from collecting the accelerometer data, which could have proved useful for activity analysis. The main reason behind this decision was the volume of the raw data: Nexus 4 phones can produce up to 180 accelerometer readings per second, far more than we would be able to store. Furthermore, we did not collect the actual content of messages and calls, audio features, or browsing histories. Such data is associated with a direct infringement of privacy. In fact, access to the microphone of the device was one of the few Android permissions that the participants voiced their concerns about.

4.2 Sources of potential bias

Pursuing a bachelor degree at the Technical University of Denmark was a necessary condition to participating in the experiment. The sample is representative of the population of freshmen at DTU but not of any larger, more general population. The participants are young (\(\mu_{\text{age}} = 21.5, \sigma_{\text{age}} = 2.2\)) and predominantly male (77%). They trusted us with their personal data and might have been convinced to participate in the study by the prospect of receiving a free smartphone (at least partially). Moreover, they interact much more than a random sample of 1000 people in a city would - they take classes together and attend social activities organized after hours by the university. Many of the subjects live on the campus that is located away from the city center and offers only a few places suited for socializing. On the other hand, their responses to surveys show that they are not significantly different from general public in terms of their personality traits. Finally, we collected the behavioral data with fixed rates and thus avoided many of the biases inherent to event-driven data collection.

4.3 Incentivisation

The students were incentivised to participate in the Copenhagen Networks Study in a number of ways. Firstly, they were given a smartphone and that smartphone would be

\[1\] I explain the implications of such setting in Chapter 7.
4.3 Incentivisation

replaced for them in case they broke or lost it. Furthermore, each student had access to the data collected about themselves, both in the raw format but also through custom built quantified-self applications. We also organized a lottery with movie theater tickets as a way to coax the participants to respond to surveys. Finally, we tried to convey the importance of the research we were doing through outreach and presentations for the students. Some of them found motivation in the prospect of contributing to science [119].

It is important to stress that the smartphone incentive appeared to be the most important one. The majority of students that dropped out from the study motivated their decision by having obtained a newer device. Furthermore, the response rates to the questionnaires were dropping wave-to-wave, regardless of the cinema tickets raffle. Finally, we failed to incetivise the students to periodically re-enable our access to their Facebook data. As a consequence, we only have high quality Facebook data for approximately six out of 24 months of the study.

Even after convincing the subjects to join an experiment, it is crucial to provide the right incentives for continued participation. In our experience data sources that require any action from the subjects (even as seemingly effortless as clicking “I agree” instead of swiping the screen away) are the most difficult to collect. On the other hand, the behavioral data collected passively regardless of the users’ activity maintained high quality throughout the experiment. Other researchers experimented with paying the participants for each question they answer [117] or for each bit of behavioral data they reveal [106].

Another important issue to consider is providing the participants with feedback based on their data. On one hand, it is vital that the decision to participate is really an informed one. Showing what can be derived from one’s data can help to ensure that the subjects understand the implications of participation. Moreover, if the feedback is valuable to the user, it can incetivise them to stay in the study. On the other hand, the subjects can be influenced by what they learn from the data about themselves and their peers, and alter their behavior as a consequence. Furthermore, the feedback can appear uncanny and discourage prolonged participation. One possible way of estimating the impact of feedback tools on the participants’ actions is tracking the usage of these instruments and comparing the behavior before and after. In any case, the researchers need to strike the right balance between ensuring that the participants’ consent is truly informed and trying to minimize the influence of the observation mechanism on the participants.

\footnote{we did not ask the students to install the data collector application on their own devices.}
Data collection
A user’s physical location drives the personalization of a number of applications, for example search engines [95], recommendations of places of interest [187], or online dating [8]. Furthermore, the set of locations that we often visit characterizes us uniquely. This fact can be used both for behavioral authentication [65] and for deanonymization attacks [42]. Finally, knowing the location of a user is an important factor in determining and understanding the context of their actions. For example, meeting somebody at work on a Monday morning has a very different meaning than meeting that person in a bar on a Friday night. I show in Chapter 8 just how important different locations are in the process of forming social ties.

In this chapter I discuss some of the methods of estimating the user’s location. I focus on location in the sense of a set of geographical coordinates (for example 43.645, -115.992), not its semantic meaning (for example home).

5.1 Location techniques and their reference points

The localization approaches I describe here follow roughly these steps:

1. Identify signals from reference points.
2. Estimate the position relative to the reference points.
3. Translate the estimated relative position to global coordinates.

In the following sections I discuss each of these steps in more detail.

5.1.1 Trilateration

Estimating the device’s location using its distance from the reference points is called trilateration. GPS, arguably the positioning technology with the widest popular recognition, is based on trilateration. A GPS receiver needs to identify signals from at least four satellites to estimate its position. It measures the time the signal takes to reach the receiver from each of the satellites to calculate the corresponding distances. Because the satellites transmit their own position, it is then possible to solve the system of four equations with four unknowns (three coordinates in space, and time offset of the receiver). Other satellite based positioning systems using trilateration include Russian GLONASS and European Galileo.

The GPS satellites transmit their local time. The receiver compares it to its local time to compute the time of flight, and thus the distance (the travel speed is constant and equal to the speed of light). However, in principle, the trilateration method does not require that the distance is estimated using the time of flight. The distance can be calculated from the amount of attenuation a radio signal undergoes as it travels through air. There are a number of models describing the attenuation of radio frequency signals as functions of distance. One of the simplest is the log-distance path loss (LDPL).

1 The reference points in GPS are 24 satellites located in six orbits in such a way that at any given moment at least eight can be seen from almost any place on Earth. They are not geostationary. Instead, they rotate the Earth twice per day. This means that their reference position is not fixed with respect to the Earth surface. Therefore, each satellite constantly has to transmit its current position as well as time. The GPS receiver can calculate the distance from each of the satellites it sees by measuring the time it took for the signal to propagate at the speed of light. The problem now appears to have three unknowns and should therefore require information from three satellites to solve. However, this is not the case. While the atom clocks in the satellites are highly precise, the clocks in the consumer grade GPS receivers are not. The calculation of the propagation time requires measuring the bias $b$ of the receiver’s internal clock. The satellites’ clocks are synchronized and their ticking rates partially compensate for the relativistic effects. Because the satellites are constantly revolving around the Earth at a significant distance, the implications of the Special and General Relativity theories must be accounted for. Their motion relative to the surface causes their clocks to fall behind the surface clocks by seven microseconds per day. On the other hand, satellites are further away from the Earth’s center of mass than the observer, which causes their clocks to appear to tick faster than those on the surface, and thus get ahead of the surface clocks by 45 microseconds a day. Ignoring the relativistic effects would cause an accumulated error of 10 kilometers per day. Each satellite’s bias is constant and constitutes the fourth unknown. Therefore, the information from at least four satellites is necessary to estimate the location of the receiver. In practice, however, the estimations can be affected by a number of noise sources and using information from more than four satellites can partially alleviate the problem.
model [70], from which the distance can be calculated as follows:

\[ d_i = 10^{\frac{P_i - p_i}{10\gamma_i}} \]  

(5.1)

In Equation 5.1 the receiver is at distance \( d_i \) (meters) from the radio beacon \( i \) and receives the signal strength of \( p_i \) (dBm). \( P_i \) (dBm) is the power transmitted by the beacon. The path loss exponent \( \gamma_i \) (unit-less) captures the rate of fall of the received signal strength (RSSI) around the beacon \( i \) and it depends on the environment the beacon is in [73]. If the transmitted power and path loss exponent are known, non-collinear measurements of three beacons should theoretically be enough to determine the position of the receiver using trilateration. Similarly to the situation with GPS, using information from more than the minimum number of beacons can help overcome some of the noise inherent in the measurement and achieve a more precise estimation.

5.1.2 Radio Frequency Fingerprinting

Fingerprinting is another approach that employs the radio frequency beacons. A fingerprint of a location is, for example, the list of beacons that are typically visible in that place. Contrary to trilateration, location estimation through fingerprinting does not rely on the routers with known positions. Instead, it requires a database of location fingerprints and their corresponding coordinates. The task of estimating the location of the device amounts to finding the most similar fingerprint in the database and using its associated coordinates. Fingerprinting has been shown to successfully identify separate locations both indoors [74, 91] and outdoors [101].

5.2 Learning the beacon positions

There is a wide range of radio frequency beacons that can be used for trilateration. Some of them are dedicated to the task of enabling navigation; for example Non-directional Beacons used in aviation or Apple iBeacons deployed in places of interest to enable semantic interpretation of location data. The beacons I use in this thesis are setup as telecommunication infrastructure. They can still be used for trilateration because they also transmit radio frequency signals.

Notable examples of radio frequency beacons used both for trilateration and fingerprinting include GSM base stations and WiFi access points. Because GSM cell towers can have a range of a few hundred meters to a few kilometers they cover large areas but offer low precision of the location estimations. WiFi access points, on the other hand, have a smaller coverage (up to 200 meters) but enable high precision estimations. Nevertheless, a location system based on using WiFi routers as beacons can have a comprehensive coverage due to the ubiquity of these devices (see Figure 5.1).
Figure 5.1: The estimated router locations are not evenly distributed in space. Each red dot represents the location estimation of one access point. The estimations are drawn to the streets as an artifact of the data collection process that relies on people scanning the routers while walking around the city.

Regardless of the technology used, using trilateration or fingerprinting for localization requires knowing the positions of the relevant beacons. Mobile phone operators know the exact coordinates of their GSM base stations and they might opt to release this information to the public, as is the case in Denmark. Unlike GSM base stations, WiFi routers are not under control of a few selected entities. Anyone can buy and deploy an access point and there is no official central registry of where each of them is located. Therefore, in order to use them as beacons for localization, we need to find out where they are. Here, I describe three methods commonly used to achieve this goal.

5.2.1 Manual reporting

In their seminal work on indoor localization Bahl and Padmanabhan introduce a system called RADAR [13]. They propose a data collection scenario in which a person equipped with a WiFi scanning device traverses a floor of a building. The person stops at each predefined location and performs multiple WiFi scans to create a fingerprint. Because the system knows where each of these locations is, it can associate the fingerprints with the corresponding coordinates. Whenever a user wants to find out their location, they upload their WiFi scan results and the system reports the coordinates of the place characterized by the most similar (best matching) fingerprint. This cumbersome and expensive, albeit precise method has been employed by
5.2 Learning the beacon positions

multiple other studies, which mostly focused on optimizing the fingerprint matching function \[26, 61, 98\].

5.2.2 Anchored inference

A number of approaches were proposed to leverage the positions of the routers relative to each other: if we know the location of a few anchors and the signal strengths of the unknown access points, we can calculate the unknown positions \[63, 111\]. Implementing this approach requires knowing the positions of the anchors, i.e. controlling at least a few access points in a given building. To enable WiFi mapping in buildings where none of the routers' positions are known, Chintalapudi introduced the EZ Localization algorithm \[32\]. In their method, the anchors are replaced by intermittent GPS updates. Even though GPS does not generally work inside buildings, a users who find themselves near a window can still obtain a GPS location estimate. This way, the locations of routers located close to windows are learned first. Then, these devices serve as anchors to infer the positions of the remaining access points.

5.2.3 Wardriving

Manual reporting at predefined locations does not scale for outdoor applications. The reason why predefined locations are used indoors is because of the difficulties with obtaining GPS estimates inside of the buildings. However, this is not the case with outdoor scenarios. Outside of the buildings it becomes possible to simultaneously collect WiFi scan results and GPS estimations in the process called wardriving. Traditionally, a person equipped with the sensors would drive around a city specifically with the purpose of building a map of fingerprints and their locations. This approach has been employed both by academic \[30, 66\] and commercial projects, for example Google Street View \[51\]. While it might be possible to drive through each street of a mid-sized city over the course of a few days, we now know that a database of router locations needs to be constantly updated \[153\]. Therefore, these databases are nowadays created in a distributed, crowd-sourced way. Smartphone users report the locations of routers as they stumble upon them. This reporting is done in the background without any user input, often even when the users explicitly disable WiFi (see Chapter \[9\]). In our article “Opportunities and Challenges in Crowd-sourced Wardriving” \[153\] we show that creation of such databases is far from trivial, due to noisy measurements of signal strengths, sparse GPS sampling, and the prevalence of mobile routers.
5.3 Precision, deployment complexity, energy efficiency

GPS offers the precision of approximately 5-10 meters under optimal weather conditions and with multiple satellites in direct view. However, the direct view requirement means that GPS cannot be used reliably inside buildings. Moreover, the energy requirements of GPS modules are too high for constant operation in mobile devices.

On the other hand, using radio based trilateration introduces a much smaller energy overhead: mobile phones scan their surroundings both for cell and WiFi signals regardless of the localization needs. To save even more energy, the actual location estimation can be done by a third party, a scenario depicted in Figure 5.2. The disadvantages of this approach include the need for constructing and maintaining databases of beacon locations as well as lower precision than that achievable with GPS. I describe the challenges of building such a database in more detail in Ref. [153].

Figure 5.2: Mobile devices rely on third party services to perform trilateration. Android offers a trilateration service which translates a given list of proximate WiFi access points and cell towers into a location estimate based on Google’s internal database of beacon positions. This figure is adapted from my presentation at Internet Measurement Conference 2015 in Tokyo, Japan.

Figure 5.3 illustrates how the same underlying mobility is seen through the different location approximation systems. The data shows the movements of one participant of the Copenhagen Networks Study during one day. Figure [5.3] and [5.3b] contain location estimations from the Android Location API — which combines GPS with WiFi and
cell information — collected at five and one minute intervals, respectively. The details of the paths the person took are lost when sampling every five minutes. Figure 5.3b shows the same trace reconstructed using the location of cell towers to which the phone connected. The two important locations (home and a shopping center) are correctly recognized, but the trace also contains erroneous locations that the person did not visit. Finally, Figure 5.3d shows the trace reconstructed using the locations of WiFi routers sampled every 15 seconds. The trace is as rich as the one in panel b, but it does not require exchanging information with Google servers every minute, thus leading to energy and financial savings, and avoiding the need to reveal the person’s location to Google.
Figure 5.3: The same underlying mobility is seen differently depending on the location estimation method and sampling frequency. a) and b) are reported by the Android location API (with median sampling period of 338 and 60 seconds respectively), which leverages GPS and known locations of radio beacons. c) is the trace estimated based on the locations of GSM towers the phone connected to (610 seconds median sampling period), and d) is the trace based on the locations of WiFi routers the person passed by (15 seconds median sampling period).
5.4 Summary of “Opportunities and Challenges in Crowd-sourced Wardriving”

As described in this chapter, trilateration requires knowing where the radio frequency beacons are located. GPS satellites send their location to the receiver; traditional beacons are installed at known locations; the placement of cell towers is known to their respective mobile network operator. The location of WiFi routers, on the other hand, needs to be discovered, and one possible approach is the process called wardriving. To date, most of the research regarding this method was focused on datasets collected both artificially (driving with the intent to collect data, rather than collecting data while performing everyday activities) and over short periods of time (e.g., during a one-day drive around a city). In contrast, most in-use databases are collected by mobile devices automatically, and are maintained by large mobile operating systems providers. As a result, the research community has a poor understanding of the challenges in creating and using large-scale WiFi localization databases. We leverage the Copenhagen Networks Study dataset to address this situation. We identify a number of challenges in using such data to build a WiFi localization database (e.g., mobility of access points), and introduce techniques to mitigate them. We also explore the level of coverage needed to accurately estimate a user’s location, showing that only a small subset of the database is needed to achieve high accuracy. We found that mobility of access points was a key challenge in ensuring that the database is accurate; a significant fraction (30%) of APs are actually non-static. However, we found that using just the APs that we are confident are static, we can provide a location estimate for 73% of all scans with a median accuracy of 15 meters. Overall, our results provide the largest-scale investigation of WiFi localization databases that we know of in the research community.
In the previous chapter I described the means for estimating the location of individuals at any given moment. Time ordered series of location estimations are referred to as mobility traces. Their growing availability opens the opportunities for a myriad of avenues of scientific inquiry. Research based on CDR data shed light on the basic laws describing our daily habits: their regularity [55, 83, 140], stability [114], and predictability [165]. Mobility data is personal in that each person's mobility trace (and even its small subsets) is unique [42] and can reveal their personal attributes [43]. Furthermore, comparing the traces of multiple people can reveal latent ties among them, such as friendship [39, 38, 175] or even undercover spying [5]. At the societal level, mobility traces can aid research in modeling of epidemic spreading across large distances [50, 172, 108, 87, 81] as well as in optimization of traffic in road [92] and telecommunication networks [24].

6.1 Properties of human mobility

Human mobility traces have been shown to be regular [55, 83, 140], stable over time [114], and predictable [165].

Regularity of mobility means that it is, to a large extent, governed by daily and weekly rhythms: we go to work every morning and return home every evening; we visit our friends and family on the weekends. As a consequence, we return to the same few
locations and we do so at certain times of day and week \[55\]. It is also possible to deduce the semantic meaning of these locations \[83\].

Stability of mobility means that the set of these important locations remains relatively unchanged over long periods of time. Lu et al. analyzed the changes in mobility traces of 1.9 million people in the wake of a major earthquake \[114\]. They showed that even people displaced by a disaster return to places they had previously visited instead of changing their patterns altogether. In our work “Tracking Human Mobility using WiFi Signals” we show that the most important locations visited by a person in the first week of observation are still visited six months later. At the same time, as we show in “Conservation laws in human mobility”, people also continue exploring new locations.

Predictability quantifies the regularity and stability of mobility: it describes the extent to which our destinations can be anticipated based on our past travels. Perhaps surprisingly, the results of this quantification rely, in a large part, on how the prediction problem is defined, and how granular the spatio-temporal information is. For example, some researches attempt to predict the location of an individual in the next timebin. A null model that always predicts stationarity —given than the individual is at place $p$ in time $t$, the model predicts they will also be at place $p$ at time $t + 1$—performs well in this task. Song et al. in their seminal work \[165\] achieve predictability of 93% in the problem of predicting the location in the next one-hour time bin. As pointed out by Sekara et al. \[161\], the smaller the time bin, the higher the reported performance of such a model: it is easier to predict my location next minute, than next hour. Figure 6.1 shows how impactful the choice of the time bin is: the mean predictability with a time bin of 30 minutes is around 0.98 and only about 0.6 for a time bin of 24 hours. Furthermore, the lower the spatial resolution of the data, the higher the performance: it is easier to guess which city an individual is going to be during the next hour than to identify the building they will be in. Another formulation of the problem is to predict the location of an individual after the next transition, which means that the null model described above does not apply. The results are independent of the temporal resolution of the data. An example application of such model, possibly familiar to the reader, is the Google Now app. Among other functionality, it predicts the next location of the user based on their current location \[50\]. Anecdotally, in my case it only uses two locations. When I am at home, it suggests I should go to work. When I am at work, it remains silent.

Most of the research on properties of human mobility has been performed using CDR data, which suffers from low spatiotemporal resolution. Some of the properties identified by the scientists can possibly be attributed to the limited fidelity of the data. For example, exploration of multiple nearby locations can be misinterpreted as returning to a known place due to the large size of spatial bins in the CDR data. The growing availability of higher fidelity data is now enabling us to verify the phenomena previously observed in coarse mobility traces.
6.2 Representing mobility

A mobility trace is formed from multiple time ordered records of an individual’s momentary location estimations. Therefore, each record consists of a timestamp and a longitude and latitude pair. Some properties of mobility can be calculated directly from data in such format. One of the ways to quantify the mobility of a person is to use the radius of gyration:

\[
r_g = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\vec{r}_i - \vec{r}_{cm})^2},
\]

(6.1)

where \( \vec{r}_i \) represents the \( i = 1, \ldots, n \) positions of a user and \( \vec{r}_{cm} \) is the center of mass of the trajectory \[54\]. Consider, however, the estimation of distance traveled by a person during a day. The intuitive approach would be to calculate the distance between each two consecutive estimations and sum the results. In practice, each record is encumbered with measurement errors. Therefore, there will be some distance between any two estimations even if they are taken at the exact same location. With high sampling rates, this noise can accumulate to seemingly large distances traveled by a stationary person. I illustrate this problem on the example of the person whose trace I showed in Figure \[5.3\]. Here, I focus on the night hours of their daily mobility trace, see \[6.2\]. Visual inspection leads us to assume that the person remained in a single building during the observation period. Clearly, the traveled distance should be close to 0. When we apply the first intuition we obtain a very different result (see Table \[6.1\]):

Figure 6.1: Effects of binning on predictability bounds. The distribution of individual predictability depends on the window size. The shorter the window, the more predictable people appear to be. The metric drops as we increase the window size. The inset shows a close up for high values of predictability. The figure is adapted from Ref. \[161\].
Figure 6.2: Location estimations of a person during night hours. The subfigures present the data obtained using different sources and sampling intervals (a) Google Location API, sampled every minute, (b) Google Location API sampled every 10 minutes, (c) WiFi estimation, every 15 seconds, (d) WiFi estimation, every 10 minutes.

<table>
<thead>
<tr>
<th>Raw data source</th>
<th>sampling interval</th>
<th>estimated distance</th>
<th>radius of gyration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location API</td>
<td>60 s</td>
<td>2931 m</td>
<td>14 m</td>
</tr>
<tr>
<td>Location API</td>
<td>600 s</td>
<td>982 m</td>
<td>13 m</td>
</tr>
<tr>
<td>WiFi trilateration</td>
<td>60 s</td>
<td>4540 m</td>
<td>6 m</td>
</tr>
<tr>
<td>WiFi trilateration</td>
<td>600 s</td>
<td>450 m</td>
<td>6 m</td>
</tr>
<tr>
<td>GSM association</td>
<td>600 s</td>
<td>10 m</td>
<td>2 m</td>
</tr>
</tbody>
</table>

Table 6.1: A naive approach to calculating traveled distance results in a great variation of estimation depending on the source of the data and the sampling frequency. Radius of gyration is not significantly affected by changing the sampling frequency.

the person appears to have traveled up to 4.5 kilometers! Two scenarios that appear closer to what we expected involve temporal subsampling or using GSM towers for estimating location. However, both of these approaches cause significant estimation errors when the person is actually moving (see Figure 5.3).

6.2.1 Stop locations

Instead of lowering the spatial or temporal resolution (by using GSM towers or subsampling, respectively) one can group nearby estimations into stop locations (places where a user remains stationary). In our scenario, all points belong to one stop location and that stop location corresponds to the building in which the person resides. The distance traveled between stop locations in this case would be 0. Note, that an important decision has to be made: at which distance do we stop considering two points as being nearby? This question reveals the inherent problem of scale [20, 183, 120, 186]. The stop location of the person in our example can be reported as their room, their
6.2 Representing mobility

building, DTU Campus, Lyngby, Denmark, or Europe. The choice of scale depends on the goal of the application and its privacy safeguards. In my work, I relied on the building scale, as it is the smallest that can be reliably estimated using the location API or WiFi trilateration.

Here I illustrate the principle by computing the stop locations in a simplistic way:

1. Sort the location estimations by time.
2. Store the first estimation and calculate its distance from each consecutive estimation $2, 3, ..., n - 1, n$ until the distance is larger than $d$.
3. Compute the median of all the estimations from first to $n - 1$ and store the coordinates as the first stop location.
4. Repeat steps 2-3 starting from location $n$ instead of from the first

Table 6.2 shows the estimations of traveled distances between stops extracted using this procedure, with varying $d$. With the distance threshold of $d = 40$, WiFi trilateration groups all estimations into one stop location. Possibly because of higher noise, the threshold required to cluster all the estimations into one stop location in the case of Location API data is 60 meters.

<table>
<thead>
<tr>
<th>data source</th>
<th>raw distance [km]</th>
<th>20 m</th>
<th>30 m</th>
<th>40 m</th>
<th>50 m</th>
<th>60 m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location API</td>
<td>2.93</td>
<td>1.50</td>
<td>0.60</td>
<td>0.14</td>
<td>0.11</td>
<td>0</td>
</tr>
<tr>
<td>WiFi trilat.</td>
<td>4.53</td>
<td>0.14</td>
<td>0.09</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.2: We grouping points within a set radius into stop locations. Then, we summing the distances between consecutive stops. As we increase the grouping distance, we arrive at a more realistic estimation of a distance traveled.

Next, I continue the example with the entire day of mobility of the person in question. Grouping the nearby points into stop locations with $d = 50$ m cuts down the list of observations from 1055 to 42 in case of Location API (25:1 compression) and from 5931 to 67 in case of WiFi (88:1 compression). As shown in Figure 6.3 this compression does not have any negative impact on the precision of traces.

In scientific literature the stop location extraction usually requires one more step: discarding the groups of estimations where a person spent less time than $t$. Clusters of stop locations which are visited repeatedly are referred to as points of interest. Several ways of compressing traces, finding stop locations, and identifying points of interest are described in our article “Inferring Stop-Locations from WiFi”. One of these methods, which we introduced in “Tracking Human Mobility using WiFi Signals”, is to associate each point of interest with a WiFi access point. This way, to represent the trace of a person we only need to know the location of as many routers as the number of locations visited by a person. As we show in our work, this number is surprisingly
Figure 6.3: Grouping nearby points into stop locations allows for an easier analysis without compromising the resolution of the data. The number of data points is significantly smaller after extracting the stop locations but the routes are preserved.

low: on average, we need to know the location of 10 routers to describe a person’s location 80% of the time.
6.3 Summary of “Tracking Human Mobility using WiFi Signals”

The omnipresent WiFi networks, primarily intended for enabling connectivity, have now become a location tracking infrastructure. They are employed as such by various organizations: from the Police and intelligence agencies, to cities, airports, shopping centers, and advertisers. Apple and Google have recently responded to such tracking by frequently randomizing the unique identifier of each phone. However, such an approach does not address the threat presented in this paper, where data is collected by an application on the phone, instead of external devices. We found that the WiFi routers a phone detects can be mapped to physical locations with great precision and, due to the predictability and long term stability of human mobility patterns, just a few routers out of thousands that a phone scans every day are necessary to reveal where the owner of the phone lives, works, and spends their leisure time, as presented in Figure 6.4.

Specifically, we show that:

- **Important locations are stable over time for majority of the users.** Most of the users retain their top locations over the 6 months of observation.

- **Users spend a vast majority of their time in a few selected locations.** If we know the position of one representative router per location, then with 20 routers per person we can account for their location 90% of time.

- **People share locations.** Because we spend a major part of a day with others, the sets of our important routers overlap with those of our friends and coworkers. As a consequence, the number of routers necessary to describe the mobility of the entire population has a sub-linear relation to the size of this population.

- **Meta-data reveals context.** Just by looking at times of day when a router is seen we can infer the semantic meaning of the location it represents (e.g. home, work, restaurant, bus stop).

Furthermore, the paper reveals shortcomings of the Android permission policy. At the time of writing of the article, the WiFi scans were not considered a location signal in the Android ecosystem. As a consequence, applications were able to collect WiFi scan results and consequently transform them into mobility traces, even without the location permission. The WiFi information permission was accompanied by a description, which did not reveal the potential threats of installing such applications, see Figure 6.5.
Figure 6.4: My mobility trace of 48 hours. Four important locations are marked in blue: home, two offices, and a food market. My phone sensed 3,822 unique routers in this period, but my location can be estimated using only eight of them more than 90% of time. (a) Traces recorded with GPS; (b) traces reconstructed using all available data on WiFi routers locations – the transition traces are distorted, but all stop locations are visible and the location is known 97% of the time. (c) With 8 top routers it is still possible to discover stop locations in which I spent 95% of the time. However, in this scenario the transitions are lost. (d) Timeseries showing when in the given 48 hours period each of the top routers was seen. It can be assumed that AP 1 is home, as it is seen every night, while AP 2 and AP 3 are offices, as they are seen during working hours. The last row shows the combined coverage of the top 8 routers, totaling 95% of time.
6.3 Summary of “Tracking Human Mobility using WiFi Signals”

Figure 6.5: The description of the WiFi location permission does not indicate that the user is about to reveal their detailed mobility traces to the application. In fact, 17 out of 20 top applications on the market required the user to grant this permission at the time of writing the article.
6.4 Summary of “Inferring Stop-Locations from WiFi”

Mobility traces represented as series of geographical coordinates cannot easily be interpreted by humans and are often too large in volume to be effectively analyzed by machines. Extraction of stop-locations can alleviate these problems to some extent.

In this work we show that this process can be done not only on series of geographical coordinates but even directly on the series of WiFi scan results. Previously suggested approaches for stop location detection using WiFi fingerprinting rely on high sampling rates [74, 91] and thus significantly decrease the battery life. Other proposed methods cluster the locations estimated through WiFi trilateration, not WiFi scan results directly [87, 75]. Here, we introduce two methods for detecting stop locations in WiFi data in a sampling rate agnostic manner. We compare them to an approach used on geographical coordinates and to ground-truth data collected manually; see Figure 6.6.

**DBSCAN on WiFi data.** The first method relies on assigning each stop location with its characteristic set of routers — its WiFi fingerprint. A person is considered to be in this stop location as long as their scan result has a certain similarity to the characteristic fingerprint.

**Greedy router selection.** The second method is adapted from the greedy method described in “Tracking Human Mobility using WiFi Signals”. For each person we find the minimal set of routers that describe their mobility and assume each router corresponds to a different stop location.

We find that, possibly due to the higher spatiotemporal resolution, the WiFi based approaches outperform the stop location detection based on geographical coordinates in terms of recall, $F_1$ score, accuracy, and Matthews Correlation Coefficient.

![Figure 6.6](image-url): Stop locations are usually inferred from series of geographical coordinates. Here we show that it is possible to infer them directly from the WiFi data, even if we do not know where each router is.
6.5 Summary of “Conservation Laws in Human Mobility”

The seminal studies on the regularity and stability of human mobility were performed using CDR data. This regularity is mostly explained by our circadian patterns [158, 12] and weekly patterns [31, 40, 162]: the commute between home and work [40, 162, 144, 188], and socializing [33].

On the other hand, we instinctively feel that our lives are, in fact, evolving: for example we change our taste in restaurants or move to another district. In a recent article, Pappalardo et al. argue that in terms of visiting new locations, people can be divided into two groups: returners and explorers. Their findings are based on three months of mobility data. Arguably, one needs a dataset covering a longer time span to model the inevitable but slow changes in our routines. The data collected during the Copenhagen Networks Study covers two years of mobility and thus enables the analysis of long term changes in our spatial signatures. In this article, we show that exploration is not just a domain of a certain group of people, but is a general human trait. As shown in Figure 6.7a, the number of locations an average person discovers keeps growing during the entire two years of observation. Furthermore, their geographic signature changes as well, as shown in Figure 6.7b — the Jaccard similarity between the weekly sets of places is only 0.25 with a time difference of five weeks.

![Figure 6.7: Exploration](image)

Participants in our study keep discovering new locations over the two years of observation (a) and their set of visited places constantly evolves (b).
Our interactions with other people are fundamental building blocks of human societies. Propinquity\footnote{physical and psychological closeness} is an indispensable element in formation of social ties\cite{18, 39}. Co-presence allows for a fuller understanding between people, enabling body language and facial expressions to enrich our verbal communication\cite{128}. By investigating the interactions among a group of people we can better understand their social dynamics\cite{161}, predict their financial situation\cite{6} and grades\cite{44}, and quantify power play in organizations\cite{136}. Face-to-face networks have been shown to influence opinions of individuals on a number of topics: from app preferences\cite{6} to politics\cite{118}. Commercial applications of personal proximity networks range from distributed ad hoc networking\cite{106} to romantic matchmaking\cite{46}. At the same time, it is through person-to-person proximity that we spread not only information, opinions, and affection, but also mobile phone viruses\cite{176} and certain human diseases\cite{84}.

In this chapter, I first describe how researchers obtain traces of face-to-face interactions. Then, I show how the dynamics of epidemic spreading can be simulated on the collected data.
7.1 Measurement methods

Approaches currently used for social sensing can be divided into three categories, depending on who senses whom: peer-to-peer, system-to-peer, peer-to-system. In this section I discuss the pros and cons of each and provide an overview of relevant technologies.

7.1.1 Peer-to-peer sensing

In the peer-to-peer sensing scenario, devices carried by individuals sense each other directly, without the necessary involvement of a central entity, see Figure 7.1. Here I describe a number of technologies through which peer-to-peer sensing can be realized.

Sociometric badges. Sociometric badges, or motes, are wearable devices designed and produced specifically for the task of social sensing. They feature RFID (radio frequency identification) or IR (infra-red) transceivers to announce their presence and detect the presence of others. They can measure the interpersonal distance with a high precision, and even estimate the position of participants relative to one another (e.g. whether they are facing each other). Sociometric badges can also be equipped with a microphone to enable the meta analysis of conversations (e.g. emotion, turn-taking, etc.). All these properties make sociometric badges a great tool for sensing actual face-to-face interactions. They have been deployed in corporations with the goal of measuring the group dynamics and information flow [136] and in schools to investigate the networks of disease spreading [151]. The method, however, does not scale well. First, the devices are custom built for the purpose and unlike smartphones, they would not normally be carried by the participants. Second, they usually lack their own connectivity with WiFi or cellular networks. Instead, they rely on specialized
Hardware infrastructure, set up in places where the interactions happen, for uploading the data they collect. While successful deployments involved hundreds of participants in schools [151] and conferences [27], the measurements only happen over short time periods and in confined spaces.

**Bluetooth.** Bluetooth is a technology for transferring data over short to medium distances, usually up to 10 meters. While social sensing has never been its primary use case, the technology was adapted for this purpose in a number of studies [17, 6, 170]. Arguably, 10 meters is too large a distance to foster a social interaction. Many studies rely on RSSI (received signal strength indication) as a proxy for distance between devices [160]. This approach can further rectify the estimation to allow for classification of distance, for example as public, social, and personal [139]. Bluetooth sensing, even when supported by the RSSI estimation does not offer the precision in distance estimation of the sociometric badges, but it allows for long-term deployments without geographic limitations. Scanning is performed by the phones that the participants would carry anyway. Thus the entrance barrier for potential observation subjects is lowered.

One of the requirements that limits the applicability of Bluetooth for social sensing is that for the phone to be sensed by other devices, it has to be in the discoverable mode. Phone manufacturers make it difficult for the user to make their phones discoverable. For example, iOS devices can only be discoverable when the user is in the Bluetooth Settings screen - whenever another app is in the foreground the phone stops responding to Bluetooth requests. Therefore, iOS devices cannot be used for sensing with Bluetooth. In the most recent Android versions (6.0+) Google have mimicked this approach. However, it is still possible for the developers to programatically enable discoverability—an option on which we relied in the Copenhagen Networks Study. The reason for these restrictive rules, is that discoverability raises a number of privacy [184] and security concerns [155]. When a phone is in discoverable mode the location of its owner can be tracked seamlessly by all third parties, including advertisers [11]. Moreover, whenever a phone is discoverable, a malicious actor can attempt to pair to it in order to steal contact lists, content of short messages, or spread mobile phone viruses [176]. Apart from the privacy and security issues of using Bluetooth for sensing, another shortcoming is that Bluetooth data lacks location context. When co-presence of individuals is inferred through devices sensing each other, an additional step is usually required to estimate the location of the meeting, for example by associating Bluetooth scans with GPS signal [161].

**WiFi.** Modern smartphones can act as personal hot-spots. Then, they can be detected by other smartphones through WiFi scanning. The first model to use this approach was the Virtual Compass [14]. In the Virtual Compass, multiple devices can share each other’s relative position to improve the general distance estimation. Carreras et al. proposed a system in which each smartphone in the experiment duty-cycles between the two modes (hot-spot and scanner) [25]. By using RSSI as a proxy, they are able to classify the distance between devices into the three categories mentioned
before (public, social, personal). In contrast to the Virtual Compass, the smartphones do not need to exchange messages, Bluetooth is not necessary and the reported precision is higher. Unfortunately, this approach requires installing custom firmware on the devices, which is prohibitive in larger scale deployments. Furthermore, the constant duty cycling of the WiFi mode interferes with the normal usage of the phone and can incur additional bandwidth cost for the user. Finally, the method potentially introduces even more privacy and security problems than Bluetooth.

### 7.1.2 System-to-peer sensing

In the system-to-peer sensing scenario, presence of an individual is detected by a sensor which is out of their control, see Figure 7.2. A central entity (the system) gathers the data from the sensors. Co-location of two or more individuals is assumed whenever they are sensed by the same sensor (or by multiple nearby sensors) within a specified time interval. System-to-peer sensing can be implemented using a number of technologies.

![Figure 7.2: System-to-peer sensing](image)

**Figure 7.2: System-to-peer sensing**, in which the same central entity controls the sensing infrastructure and collects the data about encounters.

**WiFi networks from the system perspective.** Any campus, company, or urban wireless network can easily be turned into a social sensor with no need for additional hardware, or active involvement of the participants. The simplest approach involves periodically storing the list of IDs of devices connected to each router in the system. Some implementations do not even require the clients to connect to the network: they rely intercepting WiFi probe requests constantly sent by mobile devices. In the first approach, the location of the user is assumed to be the same as the location of the router they are connected to. As a consequence, co-location of individuals can also be inferred: users are co-located whenever connected to the same router or whenever they are within a certain distance from one another. These opportunities have already been exploited by “law” enforcement [90], airports [159], and universities [152].
7.1 Measurement methods

**GSM towers.** Similarly to the WiFi from system perspective, one could assume co-location of two people whenever they are connected to the same cell tower. This approach, however, fails to offer the recall and the precision of WiFi systems. The low recall is caused by two factors. Firstly, a proximity event can only be recorded if it happens between customers of the same mobile operator. Secondly, CDR data is event-driven, so the location of the user is only known when they initiate or receive a call or a short message. Whenever two persons are together but do not actively use their phones, their interaction is not recorded in the CDR data. In my Master’s thesis, I compared the co-locations inferred from the cell data and found the recall of 10% and precision (positive predictive value) of 3% [152]. In real life scenarios, these numbers can be expected to be even lower as the data I used was not even driven; instead, it had a fixed temporal resolution of 10 minutes.

**Bluetooth scanners.** A Bluetooth scanner is a device, such as a smartphone or a computer, configured to continuously send Bluetooth probe requests and collect the responses. Proximate devices in discoverable mode receive these requests and respond with their unique identifiers. These responses are then collected, and because the identifiers do not change over time, can be used to reconstruct the mobility traces of individuals as well as interactions among them. Such approach has been used in studies of urban space utilization [96, 138], traffic optimization [69], congestion in security-critical areas [22, 68], as well as group dynamics during festivals [103], conferences [82], and other events [167]. The system-to-peer approach with Bluetooth has been losing in popularity as it requires the devices in the observed population to be in Bluetooth discoverable mode — a requirement which the phone producers insist on making increasingly difficult to meet.

### 7.1.3 Peer-to-system sensing

This category encompasses methods in which a device of the individual senses its location (or a proxy thereof) and uploads the measurement to the central entity, see Figure 7.3. That entity then compares the measurements of individuals to discover the person-to-person interactions post factum.

**Location.** The precision of location estimation on smartphones might not be sufficient to infer face-to-face interactions. The GPS readings have an uncertainty as low as 5 meters but the energy footprint of GPS is too high to be used continuously. Network based measurements report the uncertainty of at least to 20 meters, making them unreliable for the task of inferring interactions over a distance of, at maximum, 10 meters. Nevertheless, proximity even at a coarse grained level can point to the existence of social ties [39, 38].
Person-to-person interactions

Figure 7.3: Peer-to-system sensing, in which the devices sense the infrastructure regardless of who controls it, they upload the data and the central entity verifies whether they were proximate.

WiFi. In this approach two devices sense their WiFi environment and upload the results to a central entity, which then compares the list of visible access points. If a similarity is above a certain threshold, the two devices are assumed to be in physical proximity. The idea of comparing WiFi signals to measure proximity was first explored more than a decade ago. Initially, researchers relied on single-feature measures of similarity, such as the Manhattan distance in signal strength space [129] or overlap between the sets of visible routers [125]. In the NearMe project [99] four features are used in conjunction to compare the lists of seen access points and estimate the distance between two persons: (1) the number of overlapping APs, (2) the Spearman rank correlation coefficient between the list of APs sorted by signal strength, (3) sum of squared differences in signal strength (Euclidean distance) (4) the number of non-overlapping APs. The authors found that using more than one feature does not further increase the accuracy of distance estimation and that the estimation error grows if previously unseen data is used for testing. In addition to short-distance proximity, where the two devices must share at least one access point, the authors introduce a notion of long-distance proximity. The long-distance proximity between two people can be sensed using a precomputed network of proximity between access points. For example, when $AP_1$ and $AP_2$ are often seen by users in a single scan, they are defined as proximate. Then, when user $a$ senses $AP_1$ and user $b$ senses $AP_2$, they are considered to be in long-distance proximity. The distance between the two users can be estimated using average travel times between the two routers.

Kjærgaard and Nurmi offer a comprehensive overview of factors which make the proximity sensing using WiFi difficult [94]. Among the key challenges they name body attenuation, the differences in sensing hardware, and the multitude of environments where the sensing takes place. They show how two features — Euclidean distance in received signal strength space and Jaccard coefficient — depend on the environment.

The method differs from the System WiFi approach in that it works in all environments, regardless of who is in control of the sensed WiFi routers. On the other hand, the entity interested in measuring the interactions between users needs to convince
them to install a sensing application on their devices. However, as I describe in Chapter 9, this task proves to be relatively easy. We explore the feasibility of a large scale WiFi peer-to-system sensing scenario in our article “Inferring Person-to-person Proximity from WiFi Signals”.

7.2 Epidemic spreading

One of the direct applications of tracking person-to-person interactions is the simulation of disease spreading through the proximity network. There are a number of models to represent the state of the nodes (agents) in the simulation. Here, I will focus on the Susceptible-Infectious-Recovered (SIR) model, described by the following rules:

1. Begin the simulation by assigning a subset of nodes with the **Infectious** status.
2. At each time step **Infectious** nodes infect each of their **Susceptible** neighbors with the probability $\beta$.
3. At each time step each **Infectious** nodes overcomes the disease and becomes **Recovered** with the probability $\Gamma$; once recovered, they cannot be infected again.

The simulation can be performed on a static network, but adding the temporal aspect makes it apparent how important the high resolution temporal data is for the outcomes of the simulation. Figure 7.4 shows an example. The network in the panel (a) is static and all links persist in each time step. In this case all the nodes in the connected component are reachable from the infected (red) node. The network in the panel (b) is a temporal network and the links only exist in the timesteps written next to each link. Contrary to case (a), nodes C and E are not reachable from node A, because the links DC and BE are only active before nodes D and B are exposed to the disease. Finally, the network in the panel (c) is the same temporal network as in (b), but now a different node is infected and, as a consequence, a different set of nodes is reachable.

In the real world data about human interactions, the activity of links changes very dynamically. Therefore, realistic simulations of the epidemic spreading require interaction data at a very high temporal resolution.
Figure 7.4: Adding the temporal dimension to a network changes the epidemic dynamics. Reachability of nodes in a static network (a) does not depend on which nodes are initially infected. On the other hand, in temporal networks edges exist only at specified time steps. As a consequence, the selection of initially infected nodes has a higher impact on spreading dynamics (panels b and c).
Understanding complex social systems requires studying not only the movement of individuals but also their interactions. Sensing social interactions is a technical challenge and many commonly used approaches—including RFID badges or Bluetooth scanning—offer limited scalability and resolution. In this work we show that it is possible, in a scalable and robust way, to accurately infer physical proximity between two people based only on the list of WiFi access points measured by their smartphones. We introduce a number of ways in which to compare the list of routers, but we also exploit the properties of human contact dynamics. For example, we notice that the prior probability of two people being in close proximity (within Bluetooth range) depends on the time of the potential interaction. This probability is high during work time (Monday to Friday, morning to early afternoon), but also during the times associated with social activity: Friday and Saturday nights, Sundays around lunch time (see Figure 7.6). We combine the insights into machine learning models and show that the models based on the features we propose outperform other known approaches. Our results demonstrate both the value of WiFi signals in social sensing as well as their potential impact on privacy.

Figure 7.5: Probability of close proximity between two people in far proximity changes with the day of the week and the time of day. Each square in the plot corresponds to the probability that two people whose phones scan at least one common router (far proximity) are within Bluetooth range (close proximity). There are similar patterns from Monday to Thursday; on Friday and Saturday nights the probability remains high until early morning hours. During Sunday people mostly interact until early afternoon hours.
Human social interactions display patterns that emerge at multiple time scales, from minutes to months. On a fundamental level, understanding of the network dynamics can be used to inform the process of measuring social networks. On the other hand, collection of physical proximity data at high temporal resolution is difficult and expensive. It is therefore crucial to understand the trade-offs between simplifying the data gathering process and retaining a high fidelity representation of interactions with a limited impact on the simulation results. We simulate two wide-spread sampling models by subsampling our high resolution data. We find that even with a fixed sampling frequency, the method in which the sampling is performed has a significant influence on the measured epidemic dynamics.

**Figure 7.6: The number of interacting pairs of people changes with high frequency.** Weekly and daily patterns of human interactions can be clearly seen from the graph, but the number of active pairs changes with a much higher frequency (see the inset). High temporal resolution is necessary in the interaction data to simulate the epidemic spreading accurately.
Our views on social ties of friendship and acquaintanceship have been shaped by anthropologists, psychologists, and sociologists. However, the methods they employed have been limited in scale and scope and could introduce serious biases in the collected data [18]. Recent technological advancements enabled collecting massive datasets about social ties and interactions with minimal effort from the participants. In the Copenhagen Networks Study, aside from the interaction and mobility data I described in the previous chapters, we also collected the communication networks between the participants: Facebook friendship graph, Facebook interactions, as well as call and SMS networks (see Figure 8.1). The communication networks are often used as proxies for social ties. As opposed to physical proximity, the communication is rarely incidental and requires some level of acquaintanceship to happen. There are a number of reasons to study the formation, preservation, and dissolution of social ties both online and in real life. In the following paragraphs I will briefly describe a selected few.

**Mental health and well-being.** Love and sense of belonging are the third level of the Maslow’s hierarchy of needs [123]. The inability to satisfy these yearnings may lead to loneliness [180], social anxiety [179], clinical depression [21], and even increased death-rates [80]. On the other hand, being embedded in a social structure gives individuals access to *social capital* and reduces the occurrence of such problems [143]. Studies have found that these effects persist also in the online context: engaging in social in-
teractions facilitates building of the social capital [49] [163], while over-consumption of content produced by others is associated with increased feeling of loneliness [23]. Other researchers indicated that social ties can carry influence detrimental to health, leading to the spread of obesity [34] or popularizing the use of cigarettes and other drugs [59].

Access to resources and professional performance. In his seminal work, Granovetter identifies the overlap of social networks of two individuals as a measure of tie strength and then emphasizes the importance of weak ties [61]. In his argument, it is the weak ties that serve a bridging role in social networks and, therefore, enable diffusion processes to reach a larger number of people. The important example brought up in the article is the dissemination of information about job opportunities. In the study of 54 people, who recently found jobs through their social contacts, the majority reported receiving the information from peers with whom they only interacted rarely. The fraction of subjects who were referred an opening by close friends was as low as 16.7%.

The importance of social ties in professional context does not end once the individuals land their jobs. Those who were referred are seen by their peers as more productive and are measurably so [60]. Furthermore, they have lower quit rates, even at jobs which do not seemingly rely on collaboration [52]. The ability of an individual to access the information (or advice, insight, etc.) depends on their centrality in the network [4] as well as the network structure itself [10]. Recently, the influence of inter- and intrateam social ties on the performance of a group of collaborators has also been studied, often based on their patterns of communication [150] and face-to-face interactions [185]. In a recent paper de Montjoye et al. found that the existence of strongest ties within and outside a team of highly skilled workers explains the team’s performance better than the competence or personality of the individual members measured before the
Influence. There are two popularly used models of dissemination: *simple contamination* and *complex contagion*. Simple contagion is well suited for example to analyze the spread of diseases [89]. In simple contagion the transmission probability $\beta$ is constant at each exposure. In the complex contagion model, a susceptible node has to be exposed by multiple infectious nodes — not to be confused with “exposed multiple times” — for the infection to happen. It has been argued that this model is better suited for describing the adoption of opinions [146], behaviors [28], urban legends [72], unproven technologies [36], social movements [122, 137, 124], etc (see Ref. [29] for a comprehensive overview). In short, individuals need social reinforcement to adopt processes that are difficult, risky, or costly. While we can contract a disease from any stranger we come in proximity with, the complex contagion is more likely to happen through people we trust. As pointed out by Granovetter, tie strength is correlated with the number of overlapping social ties — we don’t share as many ties with the acquaintances, the bridging nodes, as we do with our close friends. Since for the complex contagion to happen one needs to be exposed to an idea from multiple sources, it is more likely, that an individual will adopt under the influence of their friends. Complex contagion, in which the minimum number of sources of exposures is known and independent from the node, is referred to as the *threshold model*. Figure 8.2 illustrates the threshold model of complex contagion with a toy example: a node needs to be exposed by two neighbors to adopt an idea and become infectious. The network in panel (a) has a bridging node $C$. If an idea is adopted by nodes $A$ and $B$, $C$ will convert in the following timestep. Starting from time step $t_2$ no susceptible node neighbors with two or more infected nodes. Therefore, no additional nodes will adopt the idea being spread. In panel (b), the network is closely knit and node $C$ does not function as a bridge anymore. In each time step one node becomes infectious and the infection threshold of another node is surpassed. By the end of the simulation all nodes have adopted the idea.

Privacy. McPherson *et al.* describe the principle of homophily in the process of tie formation [126]. According to this principle, relationships of all types are struck mostly between people who are similar with regards to their socioeconomic backgrounds, value systems, behaviors, etc. One of the side effects of this phenomenon is that it is possible to learn the characteristics of an individual solely from the attributes of their peers, even if the peers do not share any information explicitly about the individual. Mislove *et al.* showed that, because of homophily and clustering in the networks, it is possible to accurately infer the college, year, and the major of students in a population where only 20% reveal this information. Other researchers presented successful attempts at inferring gender [189], political views [112], and sexual orientation [85] of users of online social networks based on the attributes of their peers. Horvát *et al.* go a step further and show that this inference is possible even for non-users [79].
8.1 Inferring social ties from co-presence events

Eagle et al. in their seminal work [48] were the first to explore the relationship between self-reported social ties and behavioral data collected through smartphones. They found that, while people fail to estimate the time they spend with others accurately, there are certain behaviors more indicative of friendship than just the total time spent together. Their analysis revealed a stronger correlation between the reported friendships and extra-role (off campus, off hours) than in-role meetings ($\rho = 0.35$ and $\rho = 0.08$ respectively).

Eagle et al. focused on the co-locations of dyads with a limited view on the spatio-temporal context. Crawshaw et al. extended the approach to include location context beyond the simple on/off campus indication [39]. By tracking the social entropy of each location of interaction, they showed that meetings at less popular locations are a strong signal of friendship (using Facebook friendship links as ground truth). They also found that the more unpredictable the meeting schedule of a pair is (in terms of temporal, spatial, and spatiotemporal entropy) the higher their probability of being friends. Furthermore, they exploited transitivity inherent in the social networks to show that the similarity of neighborhoods can further aid the inference. They found that a model incorporating the newly introduced features far outperforms the simple approach of Eagle et al.

Using smartphones for social sensing offers unparalleled resolution of the data but is currently limited with respect to the number of participants. However, behavioral data
at lower resolutions are available for significantly larger populations. Wang et al. have shown that even co-locations measured through comparatively low-resolution CDR data can be used to infer social ties \cite{175}. The behavioral features they propose (including the co-location rate weighted by the number of people present) have only marginally lower predictive power than the features based on network similarity between two people. As presented by Li et al., using trajectories instead of separate co-location events might further increase the performance of link inference \cite{107}.

In parallel to these developments, researchers have also worked on the link prediction problem in settings where the continuous behavioral data is unavailable. There, instead of being tracked in the background, the subjects explicitly check-in at different locations at will. Crandall et al. investigated the relationship between the number of unique locations visited by two people and the probability of them being friends in a photo-sharing service \cite{38}. Instead of defining the co-location as a simultaneous presence of two individuals in a confined location, they allowed for visiting the location at different times (ranging from one day to one year) and the location could be of arbitrary size (ranging from $80 \times 80$ meters to $800 \times 800$ kilometers). They found a sharp growth of the probability of the friendship link with the growing number of unique locations visited. Perhaps surprisingly, lowering the spatial resolution does not inherently introduce noise. By counting only the unique locations at this lower resolution, the authors impose that the meetings take place further apart; multiple encounters in distant locations are less likely to be coincidental and are thus a strong signal of friendship. The data Crandall investigated has very few observations per user and only a very small fraction of friends (1%) “meet”. The sparsity of the data and lowering of spatial resolution imply that their results are likely to have high precision but low recall. Scellato et al. extended this approach by introducing additional properties of locations shared among two people, such as the social entropy \cite{157}. Other works showed that probability of friendship decreases with growing geographical distance \cite{110}, that clusters of friends tend to live nearby \cite{156}, and that friends meet in diverse locations \cite{141}.

Other researchers investigated the coupling between the social and the mobility data beyond the task of link prediction. Intuitively, since maintaining a bond requires physical proximity, some of people’s mobility is driven by social factors. Several works argue that many non-routine travels observed in real data can be attributed to individuals seeking interaction with their social contacts \cite{57, 174, 83}.

8.2 Communication networks as proxies for real-world relationships

Wiese et al. \cite{181} compared phone networks and self-reported friendships of 40 subjects. They found that, while frequent communication indicates strong ties, the lack of communication does not necessarily indicate a weak tie. Among other contributing factors they list the realization that people use multiple channels of communication
Offline social networks and their online representations

(including face to face) and their phone networks do not fully describe their social networks. Furthermore, in the work I have already mentioned, Eagle et al. [48] found that the single feature most correlated with self-reported friendship is the phone communication between two people.

Golder et al. [53] and Wilson et al. [182] found that only a small fraction of pairs of Facebook friends actually interact in the online social network. Wilson recommends using the interaction graph instead of the declared friendships to better model the underlying social network. These insights were further confirmed by Jones et al. [86]. By comparing self-reported friendships with online interaction data they found that the strength of tie is correlated with the intensity of contact on Facebook, especially with commenting each other’s wall content. Furthermore, they found that private messages, to which we do not have access in Copenhagen Networks Study, do not constitute a better indicator of real word friendship than wall posts.

8.3 Conservation of ties over time

Research on face-to-face interactions indicates that the temporal networks of proximity undergo drastic structural changes even at the lowest temporal scales [161, 169]. The majority of the links are spurious and the overlap between active edges is very low even between two consecutive time steps. On the other hand, there is a body of research showing that the cognitive networks of social ties, while constantly evolving, are more stable. Surveys conducted among fixed populations every few years indicated that some relationships can persist over multiple years [121, 34]. Even given the popularization of instantaneous communication, social networks tend to evolve at low rates. Miritello et al. analyzed a CDR dataset of 20 million subscribers and found that 75% of links are retained over a seven-month observation period [131]. As they point out, some of the dyads who cease to call each other could still have a social tie, which would manifest itself if the period observation was extended. In contrast, online social networks with no cost to maintaining a link can introduce the opposite bias. There, friendships are rarely deactivated and individuals “accumulate” friends. As shown by Grabowicz et al. actual interactions reveal more realistic structures than the graph of friendships [58]. Finally, in line with the theory of limited social capacity introduced by Dunbar [76], humans tend to retain a fixed number of active contacts [58, 131, 154, 54].

[1] Jones et al. were able to gain access to the Facebook private messages because they asked the participants to export their own data from Facebook and hand over it over to the experiment organizers. Copenhagen Networks Study dataset does not contain the private messages because the data was collected through Facebook API, which does not offer access to this information.
8.4 Summary of “Offline behaviors of online friends”

In this work we analyze traces of mobility and co-location among the participants of Copenhagen Networks Study. We show that the signal of friendship is strong enough to be discovered in the noise of serendipitous offline interactions between familiar strangers. We find that the size of the groups people meet in is an important factor in accurately inferring social ties, with friends spending more time in smaller groups. We also show that interactions between friends tend not to follow a particular schedule, as opposed to interactions between non-friends. Furthermore, we show that there are subtle yet observable behavioral differences between friends depending on the definition of the social tie. Using a dataset of an unprecedented scale and richness of interactions, we are able to compare different proxies for friendship used in social networks research: face-to-face offline proximity, Facebook friendship graph, Facebook interaction network, as well as call and SMS networks. Our study also offers an overview of methods used in the problem of link inference from offline behavior and an important contribution in terms of new features in the prediction task.

There are three major contributions in this work. First, we verify that the proxies of friendship used in computational social science (Facebook interactions, calls, and short messages) are reflected in the offline behaviors and are positively correlated with the intensity of person-to-person contact. Second, we show that there is a surprisingly low overlap between dyads who communicate on Facebook and those who call each other, and that offline behaviors can indicate the means of communication adopted by each dyad. Finally, we compare the performance of a number of behavioral traits in discerning the actual social ties from the imposed structure of the studies, see Figure 8.3. The insights can be applied to aid research and empower social applications, but also raise important questions regarding privacy of millions of smartphone users.
Figure 8.3: Relative importance of features in predicting the links in Facebook friendship graph, Facebook interaction network, SMS, and call networks. The most important feature for predicting call/sms networks is the time spent in extra-role (off campus, off hours) weighted by the number of people present. In-role (on campus, during working hours) interactions are more important for inferring Facebook links than extra-role. Time at home is consistently the least important feature for inferring Facebook ties.
Chapter 9

Contributions to privacy

One of the most important insights of my work is that information bleeds through different data channels. Throughout the thesis I show how location, interactions, and social ties can be inferred using WiFi scan results. One could argue that showing these dependencies contributes to the erosion of smartphone users’ privacy: the methods I describe can be used by malicious actors to gain access to more information than the users would want to share. However, I believe that research like this brings about a level playing field. There are very few parties with access to the kind of data we analyzed in the Copenhagen Network Study and they have no incentives to inform the public of the implications. By publishing our findings and promoting them through mass media, we grant the public the access to information they would not have otherwise learned about. I believe that the work I present in this thesis makes a number of contributions to strengthening the privacy of all users of Android smartphones, and especially of participants in future computational social science deployments.

Not leaving any digital breadcrumbs while remaining a part of today’s society might prove impossible. However, we can take small actions to limit our footprints without resigning from too many comforts.
9.1 “I have nothing to hide”

Many people feel that, as long as they do nothing illegal, their lives are not affected by the data collection programs. In fact, this sentiment is so widespread, it even earned a name — it is known as the nothing to hide argument [164]. Eric Schmidt, a former CEO of Google, once said “If you have something that you don’t want anyone to know, maybe you shouldn’t be doing it in the first place” [2]. Edward Snowden disarms this argument with an even more memorable quote: “Arguing that you don’t care about the right to privacy because you have nothing to hide is no different than saying you don’t care about free speech because you have nothing to say”. There is a number of reasons to care about privacy even if one does not do anything illegal. Firstly, what others know—or infer from data—about us has a measurable impact on our lives already today. It can for example influence our chances of finding a job [130], being given a loan [71, 113], and it even changes the prices we pay for products [67]. Secondly, the inability to remain anonymous, or to withhold details about mobility or interactions, makes it more difficult for journalists to protect their sources. Thus, the erosion of privacy translates to limitations of freedom of expression. In fact, we already know that governments and intelligence agencies use the data collected in bulk to investigate or even target journalists [127, 62] who are not officially suspected of any wrongdoing. Furthermore, the risk of being found guilty by association is intensified by the bulk data collection. Even if a person did nothing illegal, they might, at some point in their life, have contacted or met somebody who becomes of interest. The mere presence at the wrong place in the wrong time can make a person appear suspicious. Finally, we do not know what future tools and analysts will be able to infer about us from the data that today appears to be safe to share. Throughout the thesis I showed how much information can be recovered from WiFi signals, which until recently were not considered sensitive.

9.2 Outreach

In our work on using WiFi routers as proxies for location, we surfaced the issue of misleading Android permissions. We argued that, contrary to the official description, apps which request the access to WiFi information do not only get the name of the connected router. Instead:

- They get a list of identifiers of all nearby routers whenever any app requests it, in practice as often as every 15 seconds.
- Each scan can be trivially transformed to location using Google Location API without the location permission.
- The developer does not have to transform each of the scans into location, since our mobility can be accurately represented using the location of just a few routers.
• Since the system starts scanning WiFi immediately after the booting sequence is completed, the app with WiFi information permission can be started in the background at boot, without the start at boot permission.

Moreover, Android phones continue scanning for WiFi in the background even when the user explicitly disables WiFi. The user can opt out from this “scanning when off” functionality but the relevant control is buried deeply in the settings. Our findings have been described in numerous news outlets. We also released a proof of concept application to the Android Play Store (see Figure 9.1). The app only requests WiFi information permission but also shows the current location of the user as well as their past stop locations on a map. It has been downloaded more than 3000 times.

Figure 9.1: Screenshots of the WiFi Watchdog App. We used the WiFi Watchdog App as a proof of concept to disseminate our findings. It only requires the WiFi information permission yet it shows the user’s current location and their past stops on a map.

Google has partially addressed the problem since. According to the documentation, starting Android Marshmallow, the WiFi scan results are only available to applications which are granted the location permission. Applications which listen to WiFi scan results but do not hold the permission should receive an empty list, as if there were no routers around. Unfortunately, at the time of writing, this was not exactly the case. The applications without the location permission can still receive the WiFi scan results as long as they are in the foreground, regardless of the screen state. This means

1see http://omnipresentwifi.com for a full list
that the application a person is currently using can still, without their knowledge or consent, track their location. Moreover, if the user turns the screen off without closing the offending app first, it will still receive WiFi scan results until another application is brought to the foreground.

9.3 Impact on future studies

As I described in Chapter 7, requiring the study participants to remain Bluetooth discoverable increases the risk of third parties tracking their locations or even attempting to hack their phones. Inferring person-to-person interactions from WiFi addresses both of these problems. As of now, most of the phones still reveal their identity while scanning for WiFi routers, but this situation has already started to change. Both Google and Apple implemented MAC address randomization at scan.

Furthermore, by relying on the Android Location API for network based location estimates, we inadvertently share the location data of participants with Google. In my work I showed that study organizers can use the participants’ data to build databases of router locations and thus avoid having to leak their data to Google.
Chapter 10

Summary

Throughout the thesis I presented a pathway of how data collected from a noisy channel can be analyzed and transformed into information meaningful on multiple levels. From a list of WiFi access points, through geographical coordinates, we derive detailed documentation of daily routines and exploration of new places. From two sets of routers, through measurements of person-to-person proximity and models of epidemic spreading, we identify friends in a complex system of interacting familiar strangers. In data like this, one can only expect any single measurement to be noise. However, with the adequate analysis and longitudinal observation, social signals begin to shine through.

One of the factors that enabled me and my collaborators to design the proposed approaches was our access to the multitude of data channels. We needed GPS to show that WiFi can work as a proxy for location. Bluetooth served as ground truth when we verified the applicability of WiFi in social sensing. Finally, to find which behaviors are robust indicators of friendship, we relied on information from Facebook as well as calls and messages. In each of these cases, information derived from the raw WiFi signals approximate the original with high, but not perfect fidelity. Indeed, we do not expect a single channel to accurately represent all the facets of life.

Being able to verify the hypotheses against multiple independent sources of data is crucial for researchers in the field of computational social science. At the same time, along with the powerful opportunities for research, collection of rich behavioral information brings about the increased risks of individual misconduct and external attacks. The decisions on what data to gather, how to collect it, store it, and grant others ac-
cess to it have a great impact on the privacy and security of the participants. I hope that my work, and the research of the entire SensibleDTU \textit{(aka Social Fabric)} research group, will contribute to responsible study designs in the future.
Appendix A

Measuring Large-scale Social Networks with High Resolution
Measuring Large-Scale Social Networks with High Resolution

Arkadiusz Stopczynski1*, Vedran Sekara1, Piotr Sapiezynski1, Andrea Cuttone1, Mette My Madsen3, Jakob Eg Larsen1, Sune Lehmann1,2

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 ($O(10^5)$–$O(10^8)$), 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 SmableDTU 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 researchers 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], Anoysense [23], CenceMe [24], Cityware [25], Darwin phones [26], Vita [27], and ContextToolbox [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 MRI 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 Sensible Organization 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 396 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]. Clanseau et al. found that a natural time scale of face-to-face social networks is 4 hours [44]. Ounela 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 RHD for sensing face-to-face interactions, Isella et al. estimated the most probable vehicles for infection propagation [38]. 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. 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 understanding 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 [116].

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 weekdays 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, rather 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
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-

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, statues, 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 bugs fixes 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 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 with 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

PLOS ONE | www.plosone.org 7 April 2014 | Volume 9 | Issue 4 | e95978
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 using the links given in the pamphlet were promised 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 5–4 weeks into the semester, thus missing the very first interactions between study participants.

**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 and agreed to phone terms and conditions 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 data collection process; students still had to manually check when participants were leaving—this was primarily important to ensure high quality of data.

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 the 2012 deployment—a token for a scope of 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 of 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].

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...
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.
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 \((i,j,t,s)\), denoting that device \(i\) has observed device \(j\) at time \(t\) with signal strength \(s\).
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 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
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, $W_t$, 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.

![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.](https://doi.org/10.1371/journal.pone.0095978.g007)
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 \(\langle x \rangle = \sum xP(x)\) (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 142 073 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

---

**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
each other, instead they record the visible beacons (in this instance WiFi access points) in their environment. Then, physical proximity of 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 overlap \( X, Y \) = \( \frac{|X \cap Y|}{\min(|X|, |Y|)} \) 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.

Next, we compare the received signal strengths between overlapping routers using the mean \( \ell_2 \)-norm (mean Manhattan distance). \( \sqrt{|X \cap Y|} \). 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 as the strong 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-frequence, 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 10^3 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
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].

![Figure 10. Weekly temporal dynamics of interactions.](https://example.com/fig10.png)

**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

![Figure 11. Daily activations in three networks.](https://example.com/fig11.png)

**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
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

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 ~90% of call ties, while 33.58% 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
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 $\langle d \rangle = 142.044s$, 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 $r_{sms} = 1.96$.

We find a Pearson correlation of 0.75 ($p < 0.05$) 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 ties 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 (46 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 & 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
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 number of Facebook friends

![Figure 13A](image1.png)

**A** number of Facebook friends

![Figure 13B](image2.png)

**B** volume of interaction with Facebook friends

![Figure 13C](image3.png)

**C** number of voice call contacts

![Figure 13D](image4.png)

**D** number of SMS contacts

![Figure 13E](image5.png)

**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. 

<table>
<thead>
<tr>
<th>Personality trait</th>
<th><em>N</em>&lt;sub&gt;f&lt;/sub&gt;</th>
<th><em>N</em>&lt;sub&gt;ff&lt;/sub&gt;</th>
<th><em>N</em>&lt;sub&gt;c&lt;/sub&gt;</th>
<th><em>N</em>&lt;sub&gt;s&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Openness</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>0.10*</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Extraversion</td>
<td>0.28***</td>
<td>0.25***</td>
<td>0.24***</td>
<td>0.29***</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>—</td>
<td>—</td>
<td>0.09*</td>
<td>—</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

* E Pearson correlation between BFI personality traits and communication channels

* *: p < 0.05, **: p < 0.01, ***: p < 0.001.

doi:10.1371/journal.pone.0095978.g015
questions related to each personality domain are calculated. As Big Five has been collected for various populations, including a 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 (Figure 15A) and a representative sample covering students mixed with the general population from Germany [171,172] and the number of Facebook friends (Figure 15B). 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 inconveniences may be mitigated 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 improve upon our work 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

The SensibleDTU project was made possible by a Young Investigator Grant from the Villum Foundation (High Resolution Networks, awarded to SL). Scaling the project up to 1 000 individuals in 2013 was made possible by an interdisciplinary UCPH 2016 grant, Social Fabric (PI David Dreyer Lassen, SL, is co-PI) focusing mainly on the social and basic science elements of the project. This grant has funded purchase of the smartphones, as well as remuneration of technical personnel. We are indebted to our University of Copenhagen partners on a number of levels: All instrumentation on the 2013 questionnaires, as well as the embedded anthropologist, are courtesy of the Social Fabric group, and most importantly the Social Fabric consortium has made valuable contributions through discussion and insight regarding nearly every aspect of the study. For an overview of the Social Fabric project, see http://socialfabric.ku.dk/. We thank group leaders in the Social Fabric Project: Professor David Dreyer Lassen, Professor Morten Axel Pedersen, Associate Professor Anders Blok, Assistant Professor Jesper Dammeyer, Associate Professor Joachim Mathiesen, Assistant Professor Julie Zahle, and Associate Professor Rikke Lund. From Institute of Economics: David Dreyer Lassen, Andreas Bjerre Nielsen, Anne Folke Larsen, and Nikolaj Hamon. From Institute of Psychology: Jesper Dammeyer, Lars Lundmann, Lasse Meinek Jensen, and Patrick Bender. From the Institute of Anthropology: Morten Axel Pedersen. From the Institute of Sociology: Anders Blok, Tobias Bornakke. From the Institute of Philosophy: Jule Zahle. From the Institute of Public Health: Rikke Lund, Ingelise Andersen, Naja Halvø Rod, Ulla.
References


Author Contributions

Conceived and designed the experiments: AS VS PS AC MMM JEL SL. Performed the experiments: AS VS PS AC MMM JEL SL. Analyzed the data: AS VS PS AC MMM JEL SL. Wrote the paper: AS VS PS AC MMM JEL SL.


Opportunities and Challenges in Crowdsourced Wardriving
Opportunities and Challenges in Crowdsourced Wardriving

Piotr Sapiezynski
Technical University of Denmark
Radu Gatej
University of Copenhagen
Alan Mislove
Northeastern University
Sune Lehmann
Technical University of Denmark

ABSTRACT
Knowing the physical location of a mobile device is crucial for a number of context-aware applications. This information is usually obtained using the Global Positioning System (GPS), or by calculating the position based on proximity of WiFi access points with known location (where the position of the access points is stored in a database at a central server). To date, most of the research regarding the creation of such a database has investigated datasets collected both artificially and over short periods of time (e.g., during a one-day drive around a city). In contrast, most in-use databases are collected by mobile devices automatically, and are maintained by large mobile OS providers.

As a result, the research community has a poor understanding of the challenges in creating and using large-scale WiFi localization databases. We address this situation using the deployment of over 800 mobile devices to real users over a 1.5 year period. Each device periodically records WiFi scans and its GPS coordinates, reporting the collected data to us. We identify a number of challenges in using such data to build a WiFi localization database (e.g., mobility of access points), and introduce techniques to mitigate them. We also explore the level of coverage needed to accurately estimate a user’s location, showing that only a small subset of the database is needed to achieve high accuracy.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

Keywords
wifi; wardriving; mobility; location

1. INTRODUCTION
Localization is an increasingly important trend on mobile devices today. Mobile applications use localization to provide users with accurate driving directions, recommendations for local points of interest (e.g., restaurants), and Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

BMC ’15, October 26–30, 2015, Tokyo, Japan.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3848-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2815675.2815711.
even as a form of authentication [10]. Determining a mobile device’s location is typically accomplished in one of two ways: First, mobile devices can use various satellite-based systems (GPS, Galileo, or GLONASS). While most mobile devices today ship with dedicated GPS hardware, relying on GPS alone for determining location has a number of downsides: obtaining an initial GPS fix introduces non-negligible delay, and causes significant power consumption.

Second, mobile devices can use WiFi localization. In brief, WiFi localization works by having the mobile device listen for advertised WiFi networks (each WiFi access point periodically announces its unique identifier or BSSID, as well as the name of the network, referred to as SSID), and report that list to a central server. The server then computes the most likely location of the mobile device and returns the result. Thus, for WiFi localization to be effective, the server must have a pre-computed database of WiFi access points (APs) and their locations. Unfortunately, building such a database is time-consuming and expensive: the database must be comprehensive (covering many locations) and up-to-date (as new APs are deployed and existing ones move).

Originally, the aim of such databases was to enable indoor positioning through finger-printing [3, 9, 20] and later through RF-modeling [15, 5]. Most recent work on indoor localization achieves sub-meter accuracy by rotating the sensing device to simulate directional antennas [14]. As the APs became more wide spread it became possible to use them for outdoor localization as well. The databases were then created by manually going to different locations and recording the observed APs (often termed wardriving) [4, 18, 7, 11]. Today, however, these databases are often built by having dedicated software on the mobile devices collect and report data back both in indoor [21, 27] and outdoor [2, 19, 26] contexts. Therefore, creating such a database at scale is typically only the domain of mobile OS providers (e.g., Apple, Google) or dedicated companies (e.g., Skyhook Wireless).

As a result, the research community currently has a relatively poor understanding of large-scale WiFi localization databases. In this paper, we address this situation by providing insights into the challenges underlying the creation of such a database, and the trade-offs in using them. We first collect a data set based on a deployment of over 800 mobile phones to students at a university in Copenhagen, Denmark for over 1.5 years. These phones run a stock Android OS with custom collection software instrumented to gather GPS location and observed WiFi APs.

Overall, we collect over 1.8M simultaneous measurements of WiFi APs and GPS location, and observe more than 1.3M
unique WiFi APs. Many of the APs are only seen a small number of times, so we focus on the 376K APs that we observe at least five times. To the best of our knowledge, this represents the most comprehensive data set of this kind that has been examined in the research literature. Using this data set, we build a WiFi localization database for Copenhagen. We discuss and identify a number of key challenges and issues in doing so:

The scale of the dataset. Most existing studies were performed either in controlled environments or over a short time. Here, we show that the WiFi landscape is constantly changing, new access points are added and old ones are moved to new locations or retired.

Mobility. With increasing trend of mobile WiFi APs, such as MiFi devices, routers on buses and trains, and mobile phones which also serve as hotspots, we observe that discovering and filtering mobile APs presents a significant challenge. Failing to properly filter these can lead to gross errors when estimating a device’s location.

Noisy data. Unsurprisingly, relying on commodity hardware introduces noise into the measurements of location, signal strength, and detectability of APs, which must be handled when inferring the location and mobility of APs.

We also explore using the database we build to estimate the locations of devices given a set of overheard APs. Specifically, we examine the trade off between the number of APs in the database and the estimation accuracy. We show that knowing the location of only a small fraction of all the APs (3.7%) is actually needed to locate users to within 15 meters 75% of the time.

2. METHODS

We now describe the data we use to build our WiFi localization database.

Phone deployment. We use data collected by the Copenhagen Networks Study experiment [23]. In this experiment, students opt-in to receive a smartphone in exchange for agreeing to let us use to collect data (e.g., Bluetooth and WiFi scan results, location estimations, call and SMS metadata, etc). The students agree to use the device as their primary phone. The experiment has been reviewed and approved by the Danish Data Protection Agency, and participants are provided with a web interface where they can access and remove any of their collected data.

The data analyzed in this work covers a period from September 2013 through March 2015 and involves more than 800 students, with 300–600 participants active on any given day. Because of software failures and physical destruction some phones had to be replaced, and thus 1,000 devices were used in total. The primary focus of the Copenhagen Networks Study experiment is the study of human interactions, hence the setup was not explicitly optimized towards discovering the locations of APs. Nevertheless, we show in this paper that the WiFi scans and GPS data allow us to do so.

Data collection app. On the phones, we install an app based on the Funf framework [1]. It starts automatically when the phones boot, so the users do not need to take action to begin collecting and uploading data.

The app collects data both actively (it requests location and WiFi updates every 5 minutes) and opportunistically (whenever another app requests updates). In order to save the battery, most of the location data is obtained using the network and/or fused provider (i.e., an existing WiFi localization database). Since we intend to use the GPS measurements as ground truth, we focus only on the 10.5% of location readings that are provided by the GPS hardware. As a consequence, while the median sampling period between GPS readings is 1 second, only 29% of per-user hourly bins have at least one GPS sample (i.e., we only know the GPS location of users in 29% of the hours, on average). This distribution is a consequence of apps like Google Maps that either use GPS data constantly or not at all.

Since we are studying WiFi localization databases, in the remainder of the paper we focus on the 1,794,473 GPS samples which happened within the same second as a WiFi access point scan. According to our measurements, a single WiFi scan lasts approximately 500 ms and this time does not depend on the number of saved networks.

It is important to note that the securing the wireless network does not make it impossible to scan it: regardless of the encryption, each router broadcasts its unique identifier and the name of the network in clear text.

Filtering data. In the 567 days of observations, our participants observed 7,203,471 unique APs, out of which 1,320,838 (18.3%) were scanned at least once in the same second as a GPS estimation. However, the majority of these APs were observed with a GPS estimation a very small number of times: 944,984 (71.5%) have less than five observations. Thus, in the remainder of the paper, we focus only on the 375,934 APs that were observed at least five times together with a GPS estimate in the same second to build our WiFi localization database.

3. BUILDING THE DATABASE

We now examine the collected data, with the goal of building a WiFi localization database.

3.1 Estimating the locations of APs

The primary challenge we face is estimating the positions of the APs, given our WiFi scan data. Intuitively, this seems straightforward, but AP mobility presents a number of challenges. In general, we expect APs to fall into one of three categories:

• Static. We expect that many APs are static and have a fixed location that does not change over the course of the experiment.

• Moved. Given that our data covers 1.5 years, some APs may remain static for long periods of time, but may be moved a small number times. For example, businesses may redeploy APs, and residents of Copenhagen may change apartments, taking their APs with them.

1 Allowing for even a short time difference would introduce noise into the measurements. For example, a car driving within city speed limits moves at 14 m/s. Because of uneven and sparse sampling, it is not feasible to calculate the speed of the measuring device and discard the scans that were performed by phones in motion.

2 We note that is possible to hide a network by disabling the access point’s SSID broadcasts (though this provides little actual security [17]). Routers configured this way still broadcast their BSSID and are present in our dataset.
• Mobile. We also expect to see some APs that show no static behavior; these could include APs located on buses and trains, as well as MiFi devices and mobile phone hotspots.

We categorize APs into these three classes by clustering the observed WiFi scan data. Specifically, every time a GPS estimation happens in the same second as a WiFi scan, we add the latitude/longitude to the list of observations of each AP visible in the scan. We then categorize the APs as follows:

**Static access points.** We first compute the geometric median [16] of all locations associated with each AP; if “most” of the observations are “close” together, we then declare the AP to be static, and declare the geometric median to be the AP’s location.³

However, selecting the right thresholds for “most” and “close” to use is more complex than it may seem, as it is difficult to determine the operating range of an AP. First, devices compliant with popular standards can be expected to have a range from 20 meters indoors (the 802.11 standard) to 250 meters outdoors (802.11n) [25]. We therefore set the radius for a static AP to be no more than 300 meters. Second, due to the complex nature of signal propagation, the range can be shortened or enlarged due to characteristics of the local environment (e.g. buildings, narrow corridors). Third, GPS devices are known to sometimes return erroneous readings [6]; to deal with these, we allow for up to 5% of locations associated with an AP to be in a bigger distance than 300 meters from the median position.

We classify the APs that satisfy this condition (95% of readings within 300 meters) as static, and find that 263,281 (70%) of the APs fall into this category.

**Moved and mobile access points.** We assume that the rest of the APs are either moved or mobile. To disambiguate the two cases, we repeat the clustering above but allow for multiple such clusters.

Specifically, we group any two locations within 600 meters (twice the radius) into the same cluster, and discard any clusters that have fewer than 5 measurements. If at least 95% of the points can be associated with one of the clusters, and the clusters can be cleanly separated in time, we categorize the AP as moved. We observe that 1,087 (0.3%) APs fall into this category. Otherwise, we categorize the AP as mobile. We observe that 111,566 (29.7%) APs fall into this category.

### 3.2 Classification evaluation

We now briefly evaluate our classification. As a sanity check, in Figure 1 we show the locations of all APs with the SSID of *dtu*, which is the SSID of APs installed at our university. The left panel shows the APs on a metro area scale; each group of APs is correctly placed at one of the university campuses and out-of-campus buildings. The right panel shows the APs around the main campus of the university. While this is not a definite confirmation of the accuracy of our approach, this example of 1,100 APs shows that we should not expect too many gross errors.

We evaluate our method of identifying the mobile APs by verifying the classification of APs that are nearly certainly static and those that are nearly certainly mobile. First, we choose APs with *eduroam* SSID as examples of APs which we expect to be stationary, since these are the names of APs at universities. Out of 3,654 such APs with at least 5 observations, 3,117 (85.3%) were identified as static and 9 (0.2%) as moved. Universities are known to relocate APs, which may partially explain why our accuracy is not 100%. Next, we choose APs with *Bedrebustur* or *Commutenet* SSIDs as examples of APs we expect to be mobile, since these are the official names of networks on buses and trains in Copenhagen. Out of 650 such APs with at least 5 observations, 642 (98.8%) were identified as mobile, and 8 (1.2%) as static.

It is important to note that access points with more observations are less likely to be classified as mobile (e.g., 29.7% of access points seen at least 5 times are classified as mobile, while only 10.0% of access points seen at least 200 times are classified as mobile). This effect is likely due to the biased sampling of access points by users (i.e., static access points are more likely to be sampled many times, due to their static nature).

Overall, our results suggest that our AP classification methodology is likely to have high accuracy.

### 3.3 Accuracy of database

Next, we explore two aspects of the accuracy of the WiFi localization database: (1) how the number of measurements of a given AP affect our estimate of its location, and (2) how the number of measurements of a given AP affects our ability to classify it as mobile or fixed location.

**Number of measurements needed.** While we cannot measure the error of location estimation without knowing the ground truth location, we can analyze how the location estimation changes with the number of observations. We select 46,000 APs classified as static and with more than 50 measurements. For each of these APs we select *N* random observations, calculate the distance between the location of the AP estimated from all the observations and the estimation based on *N* random observations. We vary *N* from 1 to 50 and repeat the process 10 times.

In Figure 2 we show that even in case of APs with fixed location, using too few measurements leads to significant deviations in the estimated position. For example, calculating the position of the AP based on only two observations leads

---

³Following the definition of accuracy from the Android Location API, we calculate the radius around the median within which 68% of points are enclosed [8].
to a 50 meter error, on average. 15 observations are necessary to ensure that the error is not larger than 50 meters in 99% of cases.

**Mobility and sample size.** Because of the prevalence of APs that are mobile, too few observations might lead to their incorrect classification as stationary. To evaluate this, we select 20,000 APs classified as mobile and more than 50 observations. For each of them we select \( N \) random observations and re-run our classification procedure. We vary \( N \) from 2 to 50 and repeat the process 10 times.

Because we only allow 5% of observations to be outside of the 300 meter radius around the median, with too few observations we might classify a fixed AP as mobile. We repeat the described experiment but with fixed APs and calculate the fraction of misclassified fixed location APs as a function of \( N \).

As we show in Figure 3, the more observations we base our estimations on, the more accurate the results are. The “spikes” at 20 and 40 APs are caused by the fact that the 5% noise threshold translates to 0 noisy samples with less than 20 observations, 1 noisy sample with 20-39 observations, etc.

Taken together, these results suggest that building an accurate WiFi localization database requires large amounts of data collected continuously over time. To better visualize the importance of longitudinal observation, we provide several examples of APs with different patterns of observation. The top left panel of Figure 4 shows a clear example of a mobile AP; in this case, it is installed in a bus. In such cases, a few observations should be sufficient to correctly classify the AP as mobile. In other cases however—as shown in top right and bottom left panels of Figure 4—a long observation period is beneficial. While in the top right example not knowing the new location of the AP would lead to errors at the range of hundreds of meters, the bottom left example shows an AP whose location changes hundreds of kilometers during the observation period. Still in some cases, even a long observation period might not be enough to determine the nature of the AP, as shown in the bottom right panel: the AP seems to have two major placements, but they overlap in time, so we classify this AP as mobile.

## 4. USING THE DATABASE

With our WiFi localization AP database built, we now turn to using the database to estimate the location of a user. In brief, when a user requests their location to be calculated, they present the database with (a) a list of the AP SSIDs and BSSIDs that it current observes, and (b) the received signal strength (RSSI) of each of these APs. We first explore
Figure 5: Using a population of students from one university results in uneven sampling. Each red point on the maps represents a single AP. The inferred locations of APs in the city center indicate that sampling is not uniform across space: the routers seem to be located along the streets, not inside the buildings.

how the signal strength relates to the distance to AP before examining our ability to estimate the user’s location.

4.1 Estimating distance from APs

RSSI. As radio waves propagate through space they become attenuated; the amount of attenuation can be calculated to estimate the distance $d$. There are a number of models describing the attenuation of WiFi signals and one of the simplest is the log-distance path loss (LDPL) model [12], from which the distance can be calculated using Equation 1:

$$d_{ij} = 10^\left(\frac{P_i - P_{ij}}{\gamma}\right).$$

In Equation 1 mobile user $j$ is at distance $d_{ij}$ (m) from access point $i$ and sees the signal strength of $P_{ij}$ (dBm). $P_i$ is the power transmitted by the AP. The path loss exponent $\gamma_i$ captures the rate of fall of RSSI around the AP $i$ which depends on the environment the router is in [13]. If the transmitted power and path loss exponent are known, three non-collinear measurements of the AP should theoretically be enough to determine its position using trilateration.

However, accurately estimating the distance given RSSI has been shown to be a challenging problem. First, because the transmitted power and the propagation loss exponent are different for every router and need to be calculated, two more measurements are necessary to solve the system of LDPL equations. Second, since the receiver characteristics vary greatly even among devices of the same make and model [24, 9, 5], more measurements are necessary to compensate for individual characteristics [5]. Third, due to the inherent noise in the measurements and a dynamically changing environment (e.g., people walking by) the RSSI reading can be very noisy in practice. For example, our previous work observed that the RSSI reading can deviate as much as 10 dB from the mean even when the source and destination are static [22]. We note that while there are methods that take advantage of the variable attenuation introduced by a human body [28], they require accelerometer data to be collected as well (which we were unable to collect in our experiment).

Nevertheless, RSSI has been reported in other studies of war-driving as a useful, if somewhat noisy, proxy for distance [4]. To verify this finding, we randomly select 5.6M observations of 30,000 APs classified as static and present

Figure 6: RSSI (left) and response rate (right) as functions of distance from the AP. The shaded bands represent percentiles 1-99, 5-95, and 25-75, the bold line represents the median value. There is a weak correlation between RSSI and distance with Spearman’s correlation of $\rho = -0.23$ for distances from 0 to 100 meters, and no correlation for larger distances. There is a strong correlation between Response rate and distance ($\rho = -0.64$) for distances from 0 to 100 meters, and a weaker ($\rho = -0.30$) correlation for larger distances. Using non-specialized hardware raises a number of challenges, including noisy measurements of RSSI and location. As a result, RSSI is not a reliable proxy for distance.

RSSI as a function of distance from the inferred location in the left panel of Figure 6. There is only a weak correlation between the measured signal strength ($\rho = -0.23$) and the inferred location, and that correlation disappears for distances larger than 100 meters. The figure also reveals that a strong RSSI can be used as an indicator of close distance, but a weak RSSI does not indicate that the APs is far away. We use Spearman’s rank correlation coefficient, instead of Pearson’s product-moment correlation because we cannot expect a linear relationships between RSSI and distance. Pearson’s $\rho$ values are lower in the analyzed relationships.

The low correlation could still be caused by the differences between routers (the emitted power and the influence of obstacles). We therefore calculate the correlation between distance and RSSI for each router separately. We find that about 35% of the routers with at least 50 observations have statistically significant, negative relation between distance and RSSI with mean $\rho = -0.36$. On the other hand, 16% of such routers have a positive relation between the RSSI and distance, with mean $\rho = 0.32$. All reported correlations are statistically significant with $p_{val} < 0.01$.

Response rate. Here, we reevaluate the response rate as a proxy of distance from the AP, first suggested in [4]. Response rate at distance $d$ is defined as the fraction of WiFi scans at distance $d$ from the position of the AP which report finding the AP. We select a random subsample of 11,700 static APs with at least 50 observations. Then, for each AP, we find all scans recorded at distance $d$ from its inferred location, varying the $d$ from 0 to 1,000 meters. We define the response rate of a AP at distance $d$ as a fraction of scans in which the AP was found. In the right panel of Figure 6 we show the correlation of distance and response rate. As expected, the response rate drops as the distance from the inferred location increases, with a much stronger correlation than RSSI ($\rho = -0.55$ for distances up to 100 meters). However, to measure the response rate, one must perform multiple scans at the same distance from the router.
4.2 Estimating the location of users

We now turn to estimating the locations of users using our database of the location of APs. Unfortunately, while response rate provides a much better correlation with AP distance, it is not ideal for estimating users' locations: when estimating locations, doing so quickly is of paramount importance, and estimating response rate requires a number of scans. Thus, in the approach below, we simply use RSSI, and leave leveraging of response rates to future work.

Knowing the list of APs recently observed, along with their RSSI, we explore estimating the user's location using four different approaches:

Mean coordinates. We ignore the RSSI and calculate the mean latitude and mean longitude among all the APs for which we know the location.

Geometric median. We ignore the RSSI and calculate the geometric median of the APs for which we know the location.

Mean weighted by RSSI. Each AP is assigned a weight based on the RSSI, with the weight defined by RSSI+100.\(^4\)

We examine instances where different numbers of APs are observed in the scans, selecting 100 random instances between 0 and 30 observed APs. In the left panel of Figure 7, we show the cumulative error distributions for estimating the user's location using these three methods. The approach with geometric median location works best, followed closely by mean weighted by RSSI. While there are some differences in the performance of the three selected methods, they are negligible: all methods locate more than 50% of scans within 13 meters from the ground truth, 90% of scans within 70 meters, and 95% of scans within 120 meters.

In the right panel of Figure 7 we compare our best method (based on the geometric median) to the estimations which we acquired from the Google Geolocation API. We show the median error as a function of the number of APs used for the estimation. While our approach performs slightly better than Google's location API, the performance is similar. Google's crowd sourced data is collected using a wide variety of uncalibrated hardware (all of our phones are exactly the same model), which might lead to more measurement noise for Google's database. Since the number of APs in each scan is highly correlated with the population density [22], and the estimation errors are lower with more routers available, we expect that the location estimations will be best in densely populated areas.

4.3 Applicability of the localization database

In total, we identified 263,281 APs as static, constituting only 3.7% of the total of 7.2M unique APs observed. We revisit the original dataset with all the scans collected to verify whether this small set of APs can be used for localization in the broader context. We randomly select 51M of those scans and find that at least two of our static APs are visible in 73% of all scans, meaning we would provide an average error of 15 meters for 73% of all WiFi scans we observed.

The median error of 15 meters means that certain problems—such as car navigation—cannot be solved using WiFi signals alone. There are, however, a number of applications where the advantages outweigh the problems related to a relatively high positioning error. First, using geographic WiFi routers enables tracking the location of mobile devices with sub-minute time resolution at low costs in terms of battery or data consumption. As a consequence, it becomes possible to accurately measure for example time spent at each location, or detect whether the user changed their location in between two location scans pointing to the same place. Second, we show it is feasible to store a lookup database on the mobile devices themselves, thus enabling positioning without access to the Internet. Our database for the Greater Copenhagen area is only 9 MB, it could be a part of a mobile application targeted at tourists.

5. SUMMARY

Being able to quickly and efficiently determine the location of a mobile device is becoming increasingly important. While mobile devices often contain dedicated GPS hardware to do so, they often opt to instead rely on WiFi localization databases as they are much quicker and more power-efficient. However, building such a database requires access to large-scale WiFi scan data over time, and is typically only available to the large mobile OS vendors.

In this work, we explored the opportunities and challenges in building such a database using a deployment of over 800 mobile devices. We found that mobility of access points was a key challenge in ensuring that the database is accurate; a significant fraction (30%) of APs are actually non-static. However, we found that using just the APs that we are confident are static, we can provide a location estimate for 73% of all scans with a median accuracy of 15 meters. Overall, our results provide the largest-scale look at WiFi localization databases that we know of in the research community.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by NSF grants CNS-1054233, CNS-1319019, and CNS-1421444. All map tiles used in the article are by Stamen under CC BY 3.0. Map data by OpenStreetMap under CC BY SA.

\(^4\)The range of RSSI given by Android is -99dBm to 0dBm.
6. REFERENCES


Appendix C

Tracking Human Mobility using WiFi signals
Tracking Human Mobility Using WiFi Signals

Piotr Sapiezynski1*, Arkadiusz Stopczynski1,2, Radu Gatej3, Sune Lehmann1,4

1 Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark, 2 Media Lab, Massachusetts Institute of Technology, Cambridge, MA, United States of America, 3 Department of Economics, University of Copenhagen, Copenhagen, Denmark, 4 Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

* pisa@dtu.dk

Abstract

We study six months of human mobility data, including WiFi and GPS traces recorded with high temporal resolution, and find that time series of WiFi scans contain a strong latent location signal. In fact, due to inherent stability and low entropy of human mobility, it is possible to assign location to WiFi access points based on a very small number of GPS samples and then use these access points as location beacons. Using just one GPS observation per day per person allows us to estimate the location of, and subsequently use, WiFi access points to account for 80% of mobility across a population. These results reveal a great opportunity for using ubiquitous WiFi routers for high-resolution outdoor positioning, but also significant privacy implications of such side-channel location tracking.

Introduction

Due to the ubiquity of mobile devices, the collection of large-scale, longitudinal data about human mobility is now commonplace [1]. High-resolution mobility of individuals and entire social systems can be captured through a multitude of sensors available on modern smartphones, including GPS and sensing of nearby WiFi APs (access points or routers) and cell towers. Similarly, mobility data may be collected from systems designed to enable communication and connectivity, such as mobile phone networks or WiFi systems (e.g. at airports or on company campuses) [2, 3]. Additionally, large companies such as Google, Apple, Microsoft, or Skyhook, combine WiFi access points with GPS data to improve positioning [4], a practice known as ‘wardriving’. While widely used, the exact utility and mechanics of wardriving are largely unknown, with only narrow and non-systematic studies reported in the literature [5, 6]. As a consequence, it is generally not known how WiFi networks can be used for sensing mobility on a societal scale; this knowledge is proprietary to large companies.

In the scientific realm, the mobility patterns of entire social systems are important for modeling spreading of epidemics on multiple scales: metropolitan networks [7–9] and global air traffic networks [10, 11]; traffic forecasting [12]; understanding fundamental laws governing our lives, such as regularity [13], stability [14], and predictability [15]. Predictability and stability of human mobility are also exploited by commercial applications such as intelligent...
assistants; for example Google Now [16] is a mobile application, which learns users’ habits to, among other services, conveniently provide directions to the next inferred location.

Mobility traces are highly unique and identify individuals with high accuracy [17]. Sensitive features can be extracted from mobility data, including home and work locations, visited places, or personality traits [18]. Moreover, location data are considered the most sensitive of all the commonly discussed personal data collected from or via mobile phones [19].

Here, we show that a time sequence of WiFi access points is effectively equal to location data. Specifically, having collected both GPS and WiFi data with high temporal resolution (median of 5 minutes for GPS and 16 seconds for WiFi) in a large study [20], we use six months of data for 63 participants to model how lowering the rate of location sampling influences our ability to infer mobility. The study participants are students with heterogeneous mobility patterns. They all attend lectures on campus located outside of the city center, but live in dormitories and apartments scattered across the metro area at various distances from the university.

By mapping the WiFi data, we are able to quantify details of WiFi-based location tracking, which are usually not available to the general public. We find that the geo-positioning inferred from WiFi access points (APs or routers) could boost efficacy in other data collection contexts, such as research studies. In addition, our findings have significant privacy implications, indicating that for practical purposes WiFi data should be considered location data. As we argue in the following sections, this finding is not recognized in current practices of data collection and handling.

Methods

The dataset

Out of the 130+ participants of the study [20], we selected 63 for which at least 50% of the expected data points are available. The methods of collection, anonymization, and storage of data were approved by the Danish Data Protection Agency, and complies both with local and EU regulations. Written informed consent was obtained via electronic means, where all invited participants read and digitally signed the form with their university credentials. The median period of WiFi scans for these users was 16 seconds, and the median period of GPS sampling was 10 minutes. The data spans a period of 200 days from October 1st, 2012 to April 27th, 2013.

Known routers and coverage

In the article we use a simple model of locating the WiFi routers. We consider an access point as known if it occurred in a WiFi scan within one second of a GPS location estimation. The shortcomings of this approach and possible remedies are described in more detail in S1 File.

We define time coverage as a fraction of ten-minute bins containing WiFi data in which at least one known router was scanned. For example, let us assume that the user has data in 100 out of 144 timebins during a day, and in 80 of these timebins there is a known router visible. Therefore, that user’s coverage for that day is 80%. The average time coverage for a day is the mean coverage of all users who had any WiFi information in that day. This way our results are independent from missing data caused by imperfections in data collection system deployed in the study.

In Fig 1 we present three different approaches to sampling, which we describe here in detail. **Initial-period sampling.** As presented in Fig 1a, we learn the location of the routers sequentially. With each GPS location estimation accompanied with a WiFi scan, we add the visible access points to the list of known routers. The learning curve can be observed for the first seven days (Fig 1a, left panel) or the first 28 days (Fig 1a, right panel). **Random subsampling.** In the random
subsampling scenario we select a set fraction of available GPS location estimations, each paired with a WiFi scan. Each GPS estimation provides information on the position of all routers seen in the paired scan. This scenario can be realized after the data collection is finished, as the location estimations are used to locate the WiFi scans which happened both before and after said estimations. The results are presented in Fig 1b. **Top routers.** We select the top routers in a greedy fashion after the data collection is finished. We sort the routers in descending order by the

---

doi:10.1371/journal.pone.0130824.g001

**Fig 1.** The time coverage provided by the routers with known position depends on who collects the corresponding location data and when it happens. In each subplot the orange line describes the scenario where each individual collects data about themselves and does not share it with others; the blue line corresponds to a system in which the location of routers discovered by one person is made known to other users; the green line presents a situation where each individual can use the common pool of known routers but does not discover access points herself. **a. Stability of location.** Learning the location of APs seen during the first seven (left panel) or 28 (right panel) days, leads to performance gradually decreasing with time in the personal case (orange line). The histograms of time coverage distribution for day 190 show that this decline is driven by a growing number of people who spend only ~ 10% of time in the locations they visited in the beginning of the observation. The global approach (blue line) does not show this tendency, which indicates that people rotate between common locations rather than moving to entirely new places. **b, c. Representativeness of randomly selected locations.** Random subsampling with an average period of 24 hours (less than 1% of all available location estimations) is sufficient to find the most important locations in which people spend more than 80% of their time; using an average period of 4 hours (2.5% percent of all available location data) results in ~ 85% coverage. The personal database does not expire since the location is sampled throughout the experiment, not only in the beginning. **d. Limited number of important locations.** Although querying commercial services for WiFi geolocation is costly, knowing the location of only the 20 most prevalent routers per person in the dataset results in an average coverage of ~ 90%. Since people’s mobility overlaps, there is a benefit of using a global database rather than treating all mobility disjointly.
number of user timebins they occur in. We choose the top one router, and then we select the routers which provide the biggest increase in the number of user timebins covered. Due to high density of access points, each semantic place is described by presence of several routers, but location of only one of them has to be established to find the geographic position of the place. In this sampling method we do not rely on our own GPS data—top routers are found purely based on their occurrence in the WiFi scans, regardless of availability of GPS scans within the one second time delta. The results of such sampling are presented in Fig 1e.

Data collection scenarios

Each subplot in Fig 1 contains series coming from three different simulated collection scenarios. In the global scenario, there is a pool of WiFi routers locations estimations coming from all users, and a router is considered known if at least one person has found its location. This scenario simulates the function of such services as for example mobile Google Maps. In the personal scenario each user can only use their own data, a router can be known to them only if they found its location themselves. It simulates collecting data in a disjoint society, where each person frequents different locations. Finally, in the global with no personal data scenario, each user can exploit estimations created by everybody else, but without contributing their own data.

Results

Ubiquitously available WiFi access points can be used as location beacons, identifying locations based on BSSID (basic service set identifier, uniquely identifying every router) broadcast by APs. These locations are not intrinsically geographical, as the APs do not have geographical coordinates attached. However, since the placement of APs tends to remain fixed, mapping an AP to a location where it was seen once is sufficient to associate all the subsequent scans from the user device with geographical coordinates. See S1 File for details on inferring the geographical locations of routers, as well as identifying (and discarding data from) mobile access points.

WiFi networks are ubiquitous. In our population, 92% of all WiFi scans detect at least one access point, and 33% detect more than 10 APs, as shown in Fig 2c. In densely-populated areas, an average of 25 APs are visible in every scan, with population density explaining 50% of the variance of the number of APs, as shown in Fig 2b. WiFi scans containing at least one visible AP can be used for discovering the location of the user, with a typical spatial resolution on the order of tens of meters.

We investigate three approaches to using access points as location beacons, all of which enable WiFi-based location tracking even with limited resources: (1) recovering APs’ locations from mobility traces collected during an initial training period (exploiting the long-term stability of human mobility), (2) recovering APs’ locations from randomly sampled GPS updates (exploiting low entropy of human mobility, see S1 File for distinction between stability and low entropy), and (3) using only the most frequently observed APs for which location can be feasibly obtained from external databases. The task is to efficiently assign geographical coordinates (latitude and longitude) to particular APs, so they can be used as beacons for tracking user’s location. In the following sections, we refer to time coverage as the fraction of ten-minute timebins, in which at least one router with a known location is observed.

Stability of human mobility allows for efficient WiFi-based positioning

Human mobility has been shown to remain stable over long periods of time [13]. We find that participants in our study have stable routines, with locations visited in the first one, two, three, and four weeks of the study still visited frequently six months later. Learning the locations of
routers seen during the first seven days (corresponding to \(3.5\%\) of the observations, shown in Fig 1a, left panel) provides APs' locations throughout the rest of the experiment sufficient for recovering \(55\%\) of users mobility until the Christmas break around days 75–90. When the location of routers seen by each person is inferred using only this person's data (the personal-only WiFi database case, shown using an orange line in Fig 1), the information expires with time: there is a stable decrease in time coverage after Christmas break. This decline is evident both when a week (Fig 1a, left panel) and four weeks (right panel) are used for training, with the time coverage dropping \(18\%\) between days 60 and 160. The histograms above each plot show the distribution of time coverage in selected points in time (at 7, 80, 190 days respectively). The distribution for day 190 reveals that the expiry of the personal database validity is driven by individuals who significantly altered routines, with \(40\%\) of participants spending only around \(10\%\) of time in locations they have visited in the first week. In contrast, when the inferred locations of routers are shared among people (the global database case, represented by a blue line) the information does not expire and shows no decreasing trend during the observation period. This implies that rather than moving to entirely new
locations, people begin to visit places that are new to them, but familiar to other participants. The histograms of time coverage distribution in both panels of Fig 1a reveal that the individuals are heterogeneous in their mobility. The coverage in most cases is highly affected in the non-personal case (where the person does not collect their own location information, but data from others is used, marked using green in the figures), but 20% of participants retain a coverage of above 80% throughout the observation period, see Fig 1a, left panel. People living and working close to each other (like students in a dormitory) share a major part of their mobility and thus location of the APs they encounter can be estimated using data collected by others.

The demonstrated stability of human mobility patterns over long periods has real-life privacy implications. Denying a mobile application access to location data, even after a short period, may not be enough to prevent it from tracking user’s mobility, as long as its access to WiFi scans is retained.

Human mobility can be efficiently captured using infrequent location updates

Sampling location randomly across time (Fig 1b), rather than through the initial period (Fig 1a) provides a higher time coverage, which is retained throughout the observation. With around one sample per day per person on average, the location can be inferred 80% of the time in case of global lookup base and 70% in personal case (see Fig 1c, at training fraction of 0.007).

The histograms in Fig 1b confirm that distribution of coverage in the non-personal case is bimodal within our population: mobility of some individuals can effectively be modeled using data from people around them, while patterns of others are so distinct they require using self-collected information. The single-mode distribution of coverage in the personal case and the fact that the distribution is unchanged between day 7 and day 190 show the lack of temporal decline when sampling happens throughout the observation period.

The GPS sensor on a mobile device constitutes a major battery drain when active [21], whereas the WiFi frequently scans for networks by default. Our results show that GPS-based location sampling rate can be significantly reduced in order to save battery, while retaining high resolution location information through WiFi scanning. Our analyses also point to another scenario where WiFi time series can result in leaks of personal information. Infrequent location data can be obtained from a person’s (often public) tweets, Facebook updates, or other social networking check-ins and then matched with their WiFi records to track their mobility.

Overall human mobility can be effectively captured by top WiFi access points

As previously suggested [15], people’s mobility has low entropy and thus a few most prevalent routers can work effectively as proxies for their location. Fig 1d shows that inferring the location of just 20 top routers per person on average (which, given the median count of 22 000 routers observed per person, corresponds to 0.1% of all routers seen) translates to knowing the location of individuals 90% of the time. Since our population consists of students, who attend classes in different lecture halls in various buildings across the campus, we expect that the number of access points necessary to describe mobility of persons with a fixed work location can be even lower. There are persons in our study, for whom just four access points correspond to 90% of time coverage (see Fig D in S1 File for details).

That the mobility of individuals in our sample overlaps is apparent in Fig 1d as the time coverage of three top routers in the personal case is the same as in the global coverage using the total of 80 routers (instead of 189 disjoint routers).
As a consequence, a third party with access to records of WiFi scans and no access to location data, can effectively determine the location of each individual 90% of time by sending less than 20 queries to commercial services such as Google Geolocation API or Skyhook.

**Single-user analysis**

To illustrate the ubiquity of WiFi access points and how effectively they can be used to infer mobility patterns, we present a small example dataset containing measured and inferred location information of one of the authors, collected over two days. During the 48 hours of observation, the researcher’s phone was scanning for WiFi with a median period of 44 seconds, measuring on average 19.8 unique devices per scan, recording 3,822 unique access points. Only one scan during the 48 hours was empty, and one scan yielded 113 unique results. Fig 3a shows the corresponding GPS trace collected with a median sampling period of 5 minutes. When dividing the 48 hours of the test period into 10 minute bins, a raw GPS trace provides location estimation in 89% of these bins. Four stop locations are marked with blue circles and include home, two offices, and a food market visited by the researcher. Fig 3b shows the estimation of this trace based on the inferred locations of WiFi routers, see S1 File for detailed information on the location inference. The four stop locations are clearly visible, but the transitions have lower temporal resolution and errors in location estimations. This method provides location information in 97% of temporal bins. Using WiFi increases overall coverage, but might introduce errors in location estimation of routers which were only observed shortly, for example during transition periods. Fig 3c shows the estimation of this trace based on the locations of top 8 (0.2%) WiFi routers. The four important locations have been correctly identified, but information on transitions is lost. Information in 95% of temporal bins is available. Finally, Fig 3d shows a graphical representation of how much time the researcher spends in any one of the top eight locations during the observation time. Note that the first four locations account for an overwhelming fraction of the 48 hours.

Knowing the physical position of the top routers and having access to WiFi information reveals the location of the user for the majority of the time bins. The details of trajectories become lost as we decrease the number of routers we use to estimate locations. With too few routers might not be possible to determine which of possible routes the subject chose or how long she took to travel through each segment of the trip. On the other hand, the high temporal resolution of the scans allows for very precise discovery of arrival and departure times and of time spent in transit. Such information has important implications for security and privacy, as it can be used to discover night-watch schedules, find times when the occupants are not home, or efficiently check work time of the employees.

**Discussion**

Our world is becoming progressively more enclosed in infrastructures supporting communication, mobility, payments, or advertising. Logs from mobile phone networks have originally been considered only for billing purposes and internal network optimization; today they constitute a global database of human mobility and communication networks [13]. Credit card records form high-resolution traces of our spending behaviors [22]. The omnipresent WiFi networks, intended primarily for communication, has now became a location tracking infrastructure, as described here. The pattern is clear: every new cell tower, merchant with credit card terminal, every new private or municipal WiFi network offer benefits to the connected society, but, at the same time, create opportunities and perils of unexpected tracking. Cities entirely covered by WiFi signal provide unprecedented connectivity to citizens and visitors alike; at the same time multiple parties have to incorporate this fact in their policies to limit...
privacy abuse of such infrastructure. Understanding and quantifying the dynamics of privacy and utility of infrastructures is crucial for building connected and free society.

Since the creation of comprehensive databases containing geolocation for APs is primarily carried out by large companies [4], one might assume that WiFi based location tracking by ‘small players’, such as research studies or mobile applications, is not feasible. As we have shown above, however, APs can be very efficiently geolocated in a way that covers a large majority of individuals’ mobility patterns.

Fig 3. 48 hours of location data of one of the authors, with the four visited locations visited marked in blue: home, two offices, and a food market. Even though the author’s phone has sensed 3 822 unique routers in this period, only a few are enough to describe the location more than 90% of time. a. traces recorded with GPS; b. traces reconstructed using all available data on WiFi routers locations—the transition traces are distorted, but all stop locations are visible and the location is known 97% of the time. c. with 8 top routers it is still possible to discover stop locations in which the author spent 95% of the time. In this scenario transitions are lost. d. timeseries showing when during 48 hours each of the top routers were seen. It can be assumed that AP 1 is home, as it’s seen every night, while AP 2 and AP 3 are offices, as they are seen during working hours. The last row shows the combined 95% of time coverage provided by the top 8 routers.

doi:10.1371/journal.pone.0130824.g003
In the results, we focused on outdoor positioning with spatial resolution corresponding to WiFi AP coverage: we assume that if at least one AP is discovered in a scan, we can assign the location of this AP to the user. This is a deliberately simple model, described in detail in S1 File, but we consider the resulting spatial resolution sufficient for many aspects of research, such as studying human mobility patterns. The spatial resolution of dozens of meters is higher than for example CDR data [13], which describes the location with the accuracy of hundreds of meters to a few kilometers. Incorporating WiFi routers as location beacons can aid research by drastically increasing temporal resolution without additional cost in battery drain.

Students live in multiple dormitories on and outside of campus, take multiple routes commuting to the university, frequent different places in the city, travel across the country and beyond. While the students spend most of their time within a few dozens of kilometers from their homes, they also make international and intercontinental trips (see Figs B and C in S1 File for details). Such long distance trips are not normally captured in studies based on telecom operator data. Our population is densely-connected and in this respect it is biased, in the same sense as any population of people working in the same location. We do simulate a scenario in which the individuals do not form a connected group by analyzing the results for personal-only database. We expect the obtained results to generalize outside of our study.

Our findings connect to an ongoing debate about the privacy of personal data [23]. Location data has been shown to be among the most sensitive categories of personal information [19]. Still, a record of WiFi scans is, in most contexts, not considered a location channel. In the Android ecosystem, which constitutes 85% of global smartphone market in Q2 2014 [24], the permission for applications to passively collect the results of WiFi scans is separate from the location permission; moreover, the Wi-Fi connection information (ACCESS_WIFI_STATE) permission is not considered ‘dangerous’ in the Android framework, whereas both high-accuracy and coarse location permissions are tagged as such [25]. While it has been pointed out that Android WiFi permissions may allow for inference of sensitive personal information [26], the effect has not been quantified through real-world data. Here we have shown that inferring location with high temporal resolution can be efficiently achieved using only a small percentage of the WiFi APs seen by a device. This makes it possible for any application to collect scanned access points, report them back, and inexpensively convert these access points into users’ locations. The impact is amplified by the fact that apps may passively obtain results of scans routinely performed by Android system every 15–60 seconds. Such routine scans are even run when the user disables WiFi. See S1 File for additional analysis on data privacy in the Android ecosystem.

Developers whose applications declare both location and WiFi permissions are able to use WiFi information to boost the temporal resolution of any collected location information. We have shown that even if the location permission is revoked by the user, or removed by the app developers, an initial collection of both GPS and WiFi data is sufficient to continue high-resolution tracking of the user mobility for subsequent months. Many top applications in the Play Store request Wi-Fi connection information but not explicit location permission. Examples from the top charts include prominent apps with more than 100 million users each, such as Candy Crush Saga, Pandora, and Angry Birds, among others. We are not suggesting that these or other applications collect WiFi data for location tracking. These apps, however, do have a de facto capability to track location, effectively circumventing Android permission model and general public understanding.

Due to uniqueness of location traces, users can be easily identified across multiple datasets [17]. Our results indicate that any application can use WiFi permission to link users to other public and private identities, using data from Twitter or Facebook (based on geo-tagged tweets and posts), CDR data, geo-tagged payment transactions; in fact any geo-tagged data set. Such
cross-linking is another argument why WiFi scans should be considered a highly sensitive type of data.

In our dataset, 92% of WiFi scans have at least one visible AP. Even in the most challenging scenario, when there are no globally shared locations and each individual frequents different places, top 20 WiFi access points per person can be efficiently converted into geolocations (using Google APIs or crowd-sourced data) and used as a stable location channel. These results should inform future thinking regarding the collection, use, and data security of WiFi scans. We recommend that WiFi records be treated as strictly as location data.

Supporting Information

S1 File. Additional details on the properties of the data and the employed analysis methods.
In this Supporting File we present an example method of inferring the locations of WiFi routers, explain the interplay between the long term stability and low entropy of human mobility, provide a detailed description of the mobility properties of the participants (Figs B and C), show the distributions of time coverage of top routers (Fig D), and explain how Android permission model allows apps to access the WiFi information of the user.

(PDF)

Acknowledgments

We thank Yves-Alexandre de Montjoye and Andrea Cuttone for useful discussions.

Author Contributions

Conceived and designed the experiments: SL AS PS. Performed the experiments: PS AS RG. Analyzed the data: PS RG. Wrote the paper: PS AS RG SL.

References

Inferring location of routers.

In the article we use a deliberately simplistic model of locating the WiFi routers. We assume that if we find a WiFi scan and a GPS location estimation which happened within a one second time difference we can assume that all routers visible in the scan are at the geographical location indicated by the GPS reading. Due to effective outdoor range of WiFi routers of approximately 100 meters, this assumption introduces an obvious limitation of accuracy of location inference. Moreover, there are a number of mobile access points such as routers installed in public transportation or smartphones with hotspot capabilities. Such devices cannot effectively be used as location beacons and will introduce noise into location estimations unless identified and discarded. We propose and test the following method. For each GPS location estimation with timestamp $t_{\text{GPS}}$ we find WiFi scans performed by the same device at $t_{WIFI}$ so that $t_{\text{GPS}} - 1s \leq t_{WIFI} \leq t_{\text{GPS}} + 1s$ and select the one, for which $|t_{\text{GPS}} - t_{WIFI}|$ is the smallest. We then add the location estimation and its timestamp to the list of locations where each of the available WiFi access points was seen. For each device, we fit a density-based spatial clustering of applications with noise (DBSCAN) model [1] specifying 100 meters as the maximum distance parameter $\varepsilon$. If there are no clusters found, or the found clusters contain less than 95% of all locations associated with the said router we assume the router is mobile and to be discarded from further analysis. If only one cluster is detected and it contains at least 95% of all points, we assume the geometric median of these points is the physical location of the router. If there are more clusters found and they contain at least 95% of all points, we verify if these clusters are disjoint in time: if the timestamps of sightings do not overlap between those clusters, we assume the device is a static access point which has been moved to a different place during the experiment. Otherwise, we classify the access point as a mobile device and do not use it as a location proxy.
In the proposed method we assume accuracy of tens of meters is satisfactory, and hence do not find a need to exploit the received signal strength information [2]. Arguably, with the sparse data that we operate on, employing received signal strength could lead to more confusion, as it can vary greatly for one location, depending on the position of the measuring smartphone, and presence of humans and other objects obstructing the signal. Fig A shows timeseries of signal strengths received by a non-moving smartphone, which vary as much as 10 dB, which corresponds to drastic differences in estimated distance to the source, as in free-space propagation model extending the distance $\sqrt{10} \approx 3.16$ times corresponds to 10 dB loss in received signal strength.

In the 200 days of observations, the participants have scanned 487 216 unique routers, out of which 64 983 were scanned within a second of a GPS estimation. As many as 57 912 were only seen less than five times which we assumed to be the minimum number of sightings to be considered a cluster, which left only 7 071 routers for further investigation. In 1 760 cases there were no clusters found, or there was more than 5% noise. In 5 267 cases there was only one cluster and less than 5% of noise. Out of 21 cases there were multiple clusters and less than 5% of noise, 9 revealed no time overlap between clusters. We verified our heuristic of determining which routers are mobile by classifying routers which are very likely mobile, as their networks are called AndroidAP (default SSID for a hotspot on Android smartphones), iPhone (default SSID for iPhones), Bedrebustur or Commutenet (names of networks on buses and trains in Copenhagen). Out of 340 such devices 323, or 95%, were identified as mobile, and 17 as fixed-location devices.

All in all, out of 487 216 unique APs we believe we managed to estimate the location of 5 276, we identified 1 771 as mobile, and did not have enough data to investigate 480 169. Even though we only know the location of approximately 1% of all sensed routers, this knowledge is enough to estimate the location of users in 87% ten-minute timebins in the dataset.

**Long term stability and low entropy of human mobility.**

Long-term stability in the context of human mobility means that individuals keep returning to the same locations over long time periods. Arguably, most people do not often move, change the work place, or find
We use entropy in Shannon’s definition, as presented in equation (1):

\[ H(X) = -\sum_i P(x_i) \log P(x_i), \]  

where \( X \) is the set of all possible locations, and \( P(x_i) \) is the probability of a person being at location \( i \). Therefore, the bigger the fraction of time a person spends in their top few places, the lower the entropy value of that person’s mobility. In this sense, long-term stability is necessary for the low entropy, and both contribute to the predictability of human mobility.

![Fig A: Received signal strength can vary greatly even if the smartphone and the access points do not move.](image)

**Mobility of the studied population.**

This article focuses on a population of students at a university. To show that their mobility is not constrained to the campus only, we present summary statistics about their mobility. Displacements in our dataset can be as big as 10,000 km. Given such extreme statistics, the radius of gyration, while commonly used in literature to describe mobility on smaller scales [3], is not a suitable measure here. Instead, in Fig B we show a qualitative overview in form of a heatmap of observed locations, as well as a distribution of time spent as a function of distance from home. For simplicity, we define the home location for each student as the location of the most prevalent access point in their data. We then calculate the median distance from home for each hour of the observation using their location data. For a more detailed view, we present the distribution for 48 randomly chosen students in Fig C.
Time coverage of top routers.

In this section we present a more detailed view on time coverage of top routers selected separately for each person. Fig DA shows the fraction of time which participants spent near to one of their top 20 routers. It is worth noting, that while home location is immediately apparent, there seems to be no definite "work" location in our population. This can be attributed to the fact that the participants of the observation are students who attend classes in different buildings and lecture halls and do not have an equivalent of an office. Fig DB is an enriched version of Fig 2d from the main text of the article. It shows that even though 20 routers are needed on average to capture 90% of mobility, there are participants for whom just four routers suffice.

Android Permissions.

The scope of Android permission \texttt{ACCESS\_WIFI\_STATE} is described in the developer documentation as “allows applications to access information about Wi-Fi networks” [4]. This permission provides the requesting application with a list of all visible access points along with their MAC identifiers after each scan ordered by any application on the phone (via broadcast mechanism). Moreover, with this permission the applications can start in the background when the first WiFi scan results appear after the phone boots: the app's BroadcastReceiver is called and the data can be collected without explicit \texttt{RECEIVE\_BOOT\_COMPLETED} permission. Requesting a WiFi scan requires the \texttt{CHANGE\_WIFI\_STATE} permission, marked as dangerous, but in most cases it is not necessary to request it: the Android OS by default performs WiFi scans in the intervals of tens of seconds, even when the WiFi is turned off; the setting to disable background scanning when WiFi is off is buried in the advanced settings.

Application developers often use \texttt{ACCESS\_WIFI\_STATE} to obtain information whether the device is connected to the Internet via mobile or WiFi network. This information is useful, for example, to perform larger downloads only when the user is connected to a WiFi network and thus avoid using mobile data. This is an unnecessarily broad permission to use for this purpose, as the same information can be obtained
with `ACCESS_NETWORK_STATE`, which provides all the necessary information without giving access to personal data of WiFi scans:

```java
ConnectivityManager cManager = 
    (ConnectivityManager) getSystemService(Context.CONNECTIVITY_SERVICE);
NetworkInfo mWifi = cManager.getNetworkInfo(ConnectivityManager.TYPE_WIFI);
if (mWifi.isConnected()) { } // wifi is connected
```

Since the `ACCESS_WIFI_STATE` together with `INTERNET` permission (for uploading the results) are effectively sufficient for high-resolution location tracking, we suggest the developers transition to using the correct permissions and APIs for determining connectivity and that accessing the result of WiFi scan requires at least the `ACCESS_COARSE_LOCATION` permission.

References

Fig B: The article focuses on a population of students at a single university, but they are not constrained to the campus only. Our data captures human mobility at different scales: the participants spend most of their time at home (1), but they travel around the neighborhood (2), the city (3), to different cities in Denmark (4), different cities in Europe (5), and finally, other continents (6).
Fig C: Distribution of time spent at different distances from the inferred home location, presented for randomly selected 48 participants. In most cases, we see the home location as the most prevalent, and probably a "work" location as the next peak in the distribution.
Fig D: A more detailed view of time coverage provided by top routers found through the greedy algorithm. A: there is a clear main location for a majority of participants, we therefore assume this to be the home location. B: even though 20 routers are needed on average to capture 90% of mobility, there are participants for whom just four routers suffice.
Appendix D

Inferring Stop Locations from WiFi
RESEARCH ARTICLE

Inferring Stop-Locations from WiFi

David Kofoed Wind*, Piotr Sapiezynski, Magdalena Anna Furman, Sune Lehmann

DTU Compute, Technical University of Denmark, Copenhagen, Denmark

*dawi@dtu.dk

Abstract

Human mobility patterns are inherently complex. In terms of understanding these patterns, the process of converting raw data into series of stop-locations and transitions is an important first step which greatly reduces the volume of data, thus simplifying the subsequent analyses. Previous research into the mobility of individuals has focused on inferring ‘stop locations’ (places of stationarity) from GPS or CDR data, or on detection of state (static/active). In this paper we bridge the gap between the two approaches: we introduce methods for detecting both mobility state and stop-locations. In addition, our methods are based exclusively on WiFi data. We study two months of WiFi data collected every two minutes by a smartphone, and infer stop-locations in the form of labelled time-intervals. For this purpose, we investigate two algorithms, both of which scale to large datasets: a greedy approach to select the most important routers and one which uses a density-based clustering algorithm to detect router fingerprints. We validate our results using participants’ GPS data as well as ground truth data collected during a two month period.

Introduction

With the growing availability of datasets describing human behavior, it has become increasingly feasible to study mobility of individuals and entire social systems [1]. Large-scale records of human mobility can be used to, for example, model spreading of epidemics [2, 3], infer and analyze social networks [4, 5], or to quantify and understand fundamental properties of our behavior, such as predictability [6, 7].

Early mobility research focused primarily on call detail records (CDR) data made available by telecom operators [1]. Such datasets cover large populations—the operators’ entire customer bases—but contain biases in terms of sampling and spatial resolution. These biases might result in an underestimation of individuals’ mobility [8]. On the other hand, the use of GPS data enables a high spatial resolution that allows for accurate estimation of mobility, especially with respect to discovery of stay points and places of interest [9–11]. GPS information is, however, rarely available for populations of comparable size to mobile phone datasets due to, for example, high battery impact [12] and the perceived impact on privacy of such data [13].

Using WiFi as a data source for detecting and classifying mobility is a well-studied research problem. It is possible to calculate the position of a device with accuracy of under 1.5 meters using trilateration [14], but this strategy has only been shown to work indoors and requires an
habits of individuals at a high spatio-temporal resolution. We understand and appreciate the need for transparency in research and are ready to make the rest of the data available to researchers who meet the criteria for access to confidential data, sign a confidentiality agreement, and agree to work under our supervision in Copenhagen. Please direct your queries to Sune Lehmann, the Principal Investigator of the study, at slio@dtu.dk.

Funding: This work was supported by Villum Foundation, http://villumfoundation.dk/C12576AB0041F1180/C12576AB0041F1180/4F7615B6F43A6E51C1257AEF003D9930/; Young Investigator programme 2012; High Resolution Networks (SL) and University of Copenhagen, http://dsi.ku.dk/news/ucph_funds/ through the UCPH 2016 Social Fabric grant (SL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

expensive training phase. One can also classify the mobility state by investigating variance of Received Signal Strength Indication (RSSI), but such approaches require temporal resolution of the data as high as one sample per two seconds [15, 16] and robustness to lower- or variable sampling rates has not yet been demonstrated.

Here we show how to identify stop-locations using WiFi data exclusively. There are multiple motivations for using WiFi data in place of GPS data: First of all, WiFi information is potentially available for large populations. For example, at the time of writing (Q1 2015), 17 out of 20 top free games on Android Play Store required access to WiFi information, while none of them required access to GPS data. Moreover, because of frequent WiFi scans scheduled by the Android operating system (by default even when the user disables WiFi), the WiFi information can be obtained by applications without additional cost to the battery [17].

Secondly, related to the study of human behavior, sequences of latitude and longitude coordinates are not how human beings process location. We argue that a sequence of stop locations is a more natural representation of a day’s activities. An example of a set of stop-location is given below.

17:33 – 07:32: Home
07:40 – 08:07: Coffee shop
08:18 – 16:10: Work

With data represented as labelled intervals, we are able phrase research questions more directly, for example ‘How does the time spent at work relate to x’, where the time spent at work can now be found by adding up the lengths of the intervals labelled ‘work’. Thirdly, in contrast to the GPS representation where mobility is represented as a sequence of pairs of rational numbers (coordinates on a sphere), an individual’s stop-locations constitute a finite alphabet, which we can analyze using, for example, the tools of information theory. Thus, the stop-location representation greatly reduces the dimensionality and sheer volume of data.

In the literature different methods have been developed to extract such personal diaries from data sources such as GPS [10]. Here, we define a stop-location as a location in which a subject is stationary—defined by a start time, an end time and optionally a label for the location. The intervals between stop-locations are denoted trips.

When considering human mobility and especially when inferring stop-locations of people, there is an inherent problem of scale [18–21]. When sitting at your office desk, there are multiple correct stop-locations to report: your chair, your office, your building, your city, your country. Which of these scales to report, depends on the application. Since WiFi data is very local (a typical router has a range of up to around 100 meters), the stop-locations that we can infer based on WiFi are on a scale corresponding to buildings.

Data

The ground truth data was collected using a smartphone (LG Nexus 4 running Android 4.4.3) with software that periodically scans and records scans for WiFi (visible access points), Bluetooth (visible Bluetooth devices) and GPS (location coordinates) [22, 23]. The dataset was collected by a single individual and runs over a period of 60 days between September 9th, 2014 and November 8th, 2014, and contains 41441 WiFi scans (approximately one every second minute), 5982 unique WiFi devices. In total 25161 GPS samples were collected (about one every 3–4 minutes). Over the data collection period 137 stops were recorded. In addition to the automatic recording of WiFi and GPS, the subject manually recorded which state she was in (bike, bus, car, run, stand, train or walk) at all times. It should be noted that the stationary
('stand') entries were not labelled to indicate specific location. A part of this diary is shown below:

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>09-09-2014</td>
<td>16:00</td>
<td>stand</td>
</tr>
<tr>
<td>09-09-2014</td>
<td>17:22</td>
<td>walk</td>
</tr>
<tr>
<td>09-09-2014</td>
<td>17:23</td>
<td>bike</td>
</tr>
<tr>
<td>09-09-2014</td>
<td>17:35</td>
<td>stand</td>
</tr>
<tr>
<td>09-09-2014</td>
<td>17:36</td>
<td>walk</td>
</tr>
<tr>
<td>09-09-2014</td>
<td>17:37</td>
<td>stand</td>
</tr>
<tr>
<td>09-09-2014</td>
<td>17:38</td>
<td>train</td>
</tr>
</tbody>
</table>

One day of the collected WiFi data is visualized in Fig 1. We use this diary of mobility as ground truth to evaluate the accuracy of the algorithms for inferring stop-locations based on the automatically collected WiFi data. Data collection, anonymization, and storage were approved by the Danish Data Protection Agency, and comply with both local and EU regulations.

Structure of this paper

The remainder of the paper is organized as follows. In Section 1 we describe methods for inferring stop-locations based on mobile sensing data. We start by discussing a recent algorithm based on GPS data [10], which we use as a baseline for our two novel approaches. We then discuss WiFi-algorithm 1 (Greedy Router Selection), which uses the most prevalent single routers and treats them as locations. WiFi-algorithm 2 (Density Based Clustering of Time Samples) uses clustered routers as locations. In Section 2 we use two different evaluation schemes to
compare the stop-locations found by the different methods. Finally, in Section 3 we discuss the advantages and shortcomings of the different methods, address potential issues of our analysis and propose future work.

1 Methods

Distance Grouping and Density Based Clustering of GPS Samples

In order to evaluate the usefulness of employing WiFi in order to infer stop-locations, we compare our results to stop-locations obtained using GPS, using a state-of-the-art method [10], which employs a combination of distance grouping and Density Based SCAN (DBSCAN) [24]. The distance grouping algorithm is based on the idea that a stop corresponds to a temporal sequence of locations within a maximal distance \( d_{\text{max}} \) from each other. Locations are examined sequentially by non-decreasing timestamp. Each stop initially contains only a single location \( loc_i \), and each subsequent location \( loc_{i+k} \) is added to the stop while distance(\( loc_{i+k}, loc_i \)) < \( d_{\text{max}} \). Then the process is restarted from \( loc_{i+k+1} \). After the distance grouping is complete, we are left with a number of groups of locations, each corresponding to a stop. Within each group the geometric median (the point minimizing the sum of distances to the points in the group) is identified and finally DBSCAN is run on the set of medians, yielding a number of clusters—each corresponding to a place of interest. The DBSCAN algorithm requires specification of two parameters \( \varepsilon \) and \( M \). The \( \varepsilon \)-parameter dictates that if two points are within distance \( \varepsilon \) from each other, they belong to the same cluster. The \( M \)-parameter specifies the minimum number of points in a cluster. In Ref. [10], \( d_{\text{max}} = 60 \text{m} \) and DBSCAN has parameters \( \varepsilon = 60 \text{m} \) and \( M = 1 \). The distance metric is the haversine metric.

Greedy Router Selection

The greedy approach was to router selection was originally proposed as a method for reducing the WiFi scan data volume in order to describe the mobility using as few routers as possible [17]. Here, we show that routers selected using this method correspond to stop-locations.

Method. We quantize the timestamps of WiFi samples’ into 5 minute time bins, corresponding to the sampling rate of WiFi in the data collector app (more samples may be available due to passive scanning in Android). Next, we sort the list of all routers by the number of unique time bins in which they appear. We then select the most most frequently occurring router and define its set of time bins as covered time bins. The next step is to descend through the sorted list of routers and find the router for which the union of covered time bins with its respective time bins is has the most elements, while discarding the routers with majority of time bins already covered. This step identifies the router, for which the increase in covered time bins is the largest. The new union is now defined as covered time bins and the search is restarted, from top of the list. The algorithm stops where no routers can be found to extend the set of covered time bins by at least \( \Delta N \) (we use \( \Delta N = 1 \) for simplicity). This results in a list of important routers which is much smaller than the set of all routers (typically, 20 routers are enough to describe the location of a person 90% of time [17]).

Post-processing. Upon extracting the important routers, we label each scan in which they appear a ‘stop location: routerid’. Scan results which do not contain any of the important routers are labelled as ‘moving’ state. In order to achieve results comparable with the method presented in [10], we discard all stop locations with duration lower than 15 minutes. We also discard all moving states of duration lower than 15 minutes if their adjacent stop locations correspond to the same important router.
Density Based Clustering of Time Samples

As an alternative to the—potentially non-optimal— greedy method of using single routers as stop-locations, we propose a method which uses multiple routers as a ‘fingerprint’ of a stop-locations below.

Data. From the WiFi samples, we construct a data matrix \( X \) with each row corresponding to an observed router, and each column corresponding to time stamp for which we have a WiFi-sample. The element \( X_{ij} \) is set equal to 1 if we observe the router \( r \) in the sample at time \( t \) and 0 otherwise. Since each WiFi-sample only contains a small portion of the total set of routers in the data set, the columns of this matrix are very sparse (see Fig 1). The rows are not necessarily sparse, since some routers are observed a large percentage of the time.

Pre-processing. Before inferring the stop-locations for the user, we pre-process the matrix. First we bin the data by introducing a time-grid with 5-minute intervals—once again corresponding to wifi sampling rate—and merging WiFi-samples occurring within the same 5-minute interval. In this column merge-step, pairs of subsequent WiFi-observations are combined using a union of the corresponding binary columns (corresponding to observing all routers from both samples at the same time).

Second we merge routers (rows) which are a subset of another router to remove a number of routers which insignificant. As part of the row merge-step the same time we introduce a weighting of the importance of the routers, where each router \( r \) starts off with an initial weight \( w(r) \) of 1. Now, given \( r_w \) and \( r_s \), where observations of \( r_w \) are a strict subset of \( r_s \) observations, then we remove the row corresponding to \( r_w \) and update the weight of \( r_s \) to \( w(r_s) \rightarrow w(r_s) + \frac{|r_s|}{|r_w|} \), where \( |r| \) is the number of observations of router \( r \) in the data set. In the cases where a router \( r_w \) is a subset of multiple routers \( R = r_1, \ldots, r_m \), we choose a random router \( r_s \in R \) and merge \( r_w \) into \( r_s \).

These two merge-steps result in a sparse matrix \( X' \), where no rows are subsets of each other, and a vector of weights \( W \). In Fig 2 a part of the data matrix \( X \) is shown before the merging of routers and a part of \( X' \) after the merging of routers.

Clustering. To identify stop-locations, we assign the columns of \( X' \) clusters using the DBSCAN (Density Based SCAN) algorithm [24]. As above, we must determine the value of DBSCAN’s two parameters: \( \varepsilon \) and \( M \) which are dependent on the problem. Further, we need to select a suitable distance measure for comparing pairs of WiFi-samples.

The Jaccard-distance of two binary vectors \( x \) and \( y \) is defined as:

\[
J(x, y) = 1 - \frac{\sum_{i=1}^{n} I(x_i)I(y_i)}{I(x) + I(y) - I(x)I(y)}
\]  

(1)

where \( I_v \) is an indicator function taking the value 1 if and only if the \( i \)-th element of the vector \( v \) is 1. We use a weighted version of the Jaccard-distance defined in Eq (2):

\[
J_w(x, y) = 1 - \frac{\sum_{i=1}^{n} w_i I(x_i)I(y_i)}{I(x) + I(y) - w_i I(x)I(y)}
\]  

(2)

In order to avoid cases when sporadic noise result in swil clusters, we choose \( M \) to be larger than 1, but keep the value as low as possible (in this case \( M = 2 \)); this allows for stop-locations which were visited only once in the data set. The parameter \( \varepsilon = 0.325 \) was chosen as to match stop-locations on the building-scale.

If we want to run this method live on incoming data (in an online fashion), we can easily update the stop location and regularly recalculate which routers should be merged. When we observe a new time-sample \( x_c \), we it to a cluster by letting \( x_c \) belong to a cluster \( C \) when the
Fig 2. Visualization of merge step for density based clustering. By merging two routers when one of them is a complete subset of the other, we reduce the number of routers in the data set. Here, merging is illustrated for a single day of data. The resulting reduction is from 357 to 29 routers. Note that the first stop-location has been reduced to a single router.

doi:10.1371/journal.pone.0149105.g002
Jaccard-distance between $x_t$ and some point in $C$ is less than $\varepsilon$. Due to the sparsity of the samples (columns of $X$) and the nature of the data (that most pairs of routers never appear together and some almost always do), we can efficiently compute which cluster a new sample belongs to by maintaining a data-structure for finding close points to a new point.

Using this method, each inferred cluster can be viewed as a 'fingerprint' specifying the routers that are typically present at the corresponding stop-location. In Fig 3 we have visualized the distribution of router-presence at a few representative stop-locations. Most clusters contain more than a single router, indicating that the method achieves robustness to a single router disappearing—and many clusters have 1–10 routers appearing 100% of the time.

**Post-processing.** After clustering the time-samples, we perform the following post-processing step: A sequence of clusters $A, B, A$, is merged to a single occurrence of cluster $A$ if the stop in cluster $B$ is shorter than 15 minutes. We also merge two consecutive occurrences of the same cluster if the gap between them is smaller than 15 minutes. These post-processing steps are performed in order to achieve results comparable with the baseline method presented in [10].

### 2 Evaluation and Results

Below we compare the stop-locations inferred by each of the three different methods presented above to the ground truth stop-locations. The problem of inferring stop-locations introduces two challenges. One challenge is to detect when a subject is stationary (which is equivalent to detecting when a subject is transitioning between stop-locations) and another is to infer in which stop-location the subject is stationary. Therefore, we perform two different tests, one

![Figure 3. Six examples of the distribution of routers in a cluster. Each plot corresponds to a single-cluster obtained from DBSCAN. In a plot, each bar (a maximum of 100 bars is shown) corresponds to an access-point, and its height corresponds to the proportion (0 to 1) of the samples in the cluster where the router was present. In most of the clusters, 1–10 routers are all present 100% of the time. doi:10.1371/journal.pone.0149105.g003](image-url)
evaluating at how well each method can predict the start and stop-times of each stop recorded in the ground truth, and one investigating how well the different methods are able to infer stop-locations, which match the true stops in regards to their geographical location.

Overlap of stop-locations

To quantify the estimation of start- and stop times for the different algorithms, we measure the overlap between stop-locations found by each method and the ones given in the ground truth. A visualization of the stop-locations found by the different methods is displayed in Fig 4. Because the ground truth data does not contain labels for the stop-locations, we consider the problem to be a binary classification problem, where the task is to predict whether or not the subject is stationary in a given time bin. We split the time-axis into bins with length 1 minute, and count in how many bins each method agrees with the ground truth, and in how many it disagrees. If the start and stop times for the inferred stop-locations are different than the ground truth, this will result in misclassifications. We compare the stop-locations found using GPS-traces, the ones found using greedy router selection, the ones found using DBSCAN on the WiFi-data and a baseline metric always predicting that the subject is in a stop-location (since approximately 96% of the time is spent in a stop-location).

We use 5 different metrics to compare the methods:

- **Classification error**: \( \frac{FP + FN}{P + N} \)
- **Precision**: \( \frac{TP}{TP + FP} \)
- **Recall**: \( \frac{TP}{TP + FN} \)
- **F1-score**: \( \frac{2TP}{2TP + FP + FN} \)
- **MCC**: \( \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \)

where \( P \) is the number of times the subject was in a stop location, \( N \) is the number of times the subject was not in a stop location, \( TP \) is the number of times the model correctly predicts that the subject is in a stop-location, \( TN \) is the number of times the model correctly predicts that the subject is not in a stop-location, \( FP \) is the number of times the model falsely predicts that the subject is in a stop-location, and \( FN \) is the number of times the model falsely predicts that the subject is not in a stop-location.
the subject is not in a stop-location. Matthews Correlation Coefficient (MCC) is a measure of the quality of a binary classification; it is generally regarded as a balanced measure which can be used for problems with large class imbalance (which is the case here, since people are mostly stationary). Even with a very high fraction of time-bins where the subject is stationary, a simple model always predicting stationarity will receive a MCC of 0.

The results are summarized in Table 1. The greedy router selection achieves the highest classification rate of 98.1%, where the GPS-based method achieves a rate of 94% and the always-one baseline gets an accuracy of 96%. In the $F_1$-metric, the two WiFi-based methods achieve a score of 0.990, the GPS-based a (lower) score of 0.969 and the always-one baseline a score of 0.979. The WiFi-based DBSCAN gets a Matthew’s Correlation Coefficient of 0.737, the greedy router selection scores 0.723, the GPS-based method gets a 0.497 and the always-one baseline scores a MCC of 0.

### Median distance between stop-locations

We now study how well each method is able to infer in which stop-location the subject is stationary. Because our ground truth data does not include labels of the recorded stops, we are not able to easily quantify whether the stops found by the methods correspond to physical locations of interest. Using the GPS-samples collected along with the WiFi data, we therefore evaluate if the clusters found by the different methods are geographically close to the stops recorded in the ground truth. In order to quantify how well the stop-locations inferred from the data correspond to the true stop-locations coordinates, we compare the geographical median of each inferred stop-location to the geographical median of GPS-samples in the ground truth.

For each recorded stop $(g_{\text{start}}, g_{\text{end}})$ in the ground truth data, we determine if the method predicts a stop in cluster $c$ which is at least 70% overlapping with $(g_{\text{start}}, g_{\text{end}})$. We have to select some threshold for how big an overlap two stops need to have before we compare them due to the inherent problem of scale in detecting stop-locations. The threshold of 70% can be chosen anywhere between 55% and 85% giving similar results.

If this is the case, then we compare the geographical median of the GPS-samples collected within $(g_{\text{start}}, g_{\text{end}})$ to all GPS-samples happening while the method predicts cluster $c$ except for those occurring in $(g_{\text{start}}, g_{\text{end}})$ (to avoid using the same GPS-samples data for computing the two medians). See Fig 5 for a visualization of this.

We perform this comparison for all reported stop-locations in the ground truth where every method (GPS, DBSCAN on WiFi and Greedy Router Selection) reports a stop-location with 70% overlap to the ground truth stop (see Fig 6 for a visualization of this). The distribution of

<table>
<thead>
<tr>
<th>Metric</th>
<th>GPS</th>
<th>DBSCAN</th>
<th>Top router</th>
<th>Always 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification error</td>
<td>0.060</td>
<td>0.020</td>
<td>0.019</td>
<td>0.040</td>
</tr>
<tr>
<td>Precision</td>
<td>0.989</td>
<td>0.988</td>
<td>0.985</td>
<td>0.960</td>
</tr>
<tr>
<td>Recall</td>
<td>0.950</td>
<td>0.992</td>
<td>0.995</td>
<td>1.000</td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.969</td>
<td>0.990</td>
<td>0.990</td>
<td>0.979</td>
</tr>
<tr>
<td>MCC</td>
<td>0.497</td>
<td>0.737</td>
<td>0.723</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The results are summarized in Table 1. The greedy router selection achieves the highest classification rate of 98.1%, where the GPS-based method achieves a rate of 94% and the always-one baseline gets an accuracy of 96%. In the $F_1$-metric, the two WiFi-based methods achieve a score of 0.990, the GPS-based a (lower) score of 0.969 and the always-one baseline a score of 0.979. The WiFi-based DBSCAN gets a Matthew’s Correlation Coefficient of 0.737, the greedy router selection scores 0.723, the GPS-based method gets a 0.497 and the always-one baseline scores a MCC of 0.

Median distance between stop-locations

We now study how well each method is able to infer in which stop-location the subject is stationary. Because our ground truth data does not include labels of the recorded stops, we are not able to easily quantify whether the stops found by the methods correspond to physical locations of interest. Using the GPS-samples collected along with the WiFi data, we therefore evaluate if the clusters found by the different methods are geographically close to the stops recorded in the ground truth. In order to quantify how well the stop-locations inferred from the data correspond to the true stop-locations coordinates, we compare the geographical median of each inferred stop-location to the geographical median of GPS-samples in the ground truth.

For each recorded stop $(g_{\text{start}}, g_{\text{end}})$ in the ground truth data, we determine if the method predicts a stop in cluster $c$ which is at least 70% overlapping with $(g_{\text{start}}, g_{\text{end}})$. We have to select some threshold for how big an overlap two stops need to have before we compare them due to the inherent problem of scale in detecting stop-locations. The threshold of 70% can be chosen anywhere between 55% and 85% giving similar results.

If this is the case, then we compare the geographical median of the GPS-samples collected within $(g_{\text{start}}, g_{\text{end}})$ to all GPS-samples happening while the method predicts cluster $c$ except for those occurring in $(g_{\text{start}}, g_{\text{end}})$ (to avoid using the same GPS-samples data for computing the two medians). See Fig 5 for a visualization of this.

We perform this comparison for all reported stop-locations in the ground truth where every method (GPS, DBSCAN on WiFi and Greedy Router Selection) reports a stop-location with 70% overlap to the ground truth stop (see Fig 6 for a visualization of this). The distribution of
the distance between the true stop median position and the median position reported by the three methods is shown in Fig 7. For the three methods, the median of the distance between the median position of the stops found using GPS-traces and the true stop position is 28.86 meters. For DBSCAN on WiFi, the median error is 29.17 meters and for the Greedy router selection, the median error is 29.26 meters. This metric penalizes methods which end up with clusters corresponding to two or more different geographical stop-locations. The reason is that in this case, the geographical coordinates for the center of the cluster (which is the geographical median) will be far off from at least one of the ground truth stops.

Fig 5. During the ground truth stop between time \( g_{\text{start}} \) and \( g_{\text{end}} \) (labeled \( S_1 \)), the GPS-method reports cluster \( G_1 \), the Top-router method reports cluster \( T_1 \) and the DBSCAN-method reports cluster \( D_2 \). Now we want to compare the geographical median of \( S_1 \) to clusters \( G_1 \), \( T_1 \) and \( D_2 \). We do this by —for each method— computing the distance between the geographical median of the gps-samples collected during \( S_1 \) and the geographical median of the gps-samples collected during for example \( G_1 \), excluding the ones collected during \( S_1 \) (to avoid overfitting). In the figure, this is depicted by comparing samples from \( S_1 \) to samples from the non-grayed-out \( G_1 \).

doi:10.1371/journal.pone.0149105.g005

Fig 6. We only make the comparison of medians for the ground truth stops where all methods report stops with at least 70% overlap. In this figure the first example (on the left) is used for comparison whereas the second (on the right) is not since the GPS method does not report a sufficiently overlapping stop. \( g_{\text{start}} \) and \( g_{\text{end}} \) refers to the starting and stopping times of the ground truth stop-location.

doi:10.1371/journal.pone.0149105.g006
3 Discussion

Above, we have analyzed the feasibility of inferring human mobility in the form of stop-locations using WiFi data. The analysis is based on two months of smartphone based WiFi data. We proposed two different approaches to inferring stop-locations from WiFi data, one based on greedily selecting routers as stop-locations and one using router signature fingerprinting with DBSCAN. Each method was evaluated using two evaluation schemes and compared to a baseline method utilizing GPS-data for stop-location inference. The evaluation schemes measured a) how well the start and stop-time of the stop-locations match the ground truth, and b) how well the geographical medians of the inferred stops correspond to the ground truth data.

In the evaluation of start and stop-times, the WiFi based methods outperform the GPS-based method, primarily because of the higher sampling rate for WiFi. In the evaluation of the geographical precision of the stops, all the methods report similar errors. In general, our results demonstrate that it is feasible to infer stop-locations using WiFi. That two different approaches to inferring stop-locations with WiFi (greedy router selection and DBSCAN) both work, indicates that WiFi is a robust data source for this application.

Fig 7. The distribution of distances between the true stop median position and the median position reported by the three methods. The histograms in the right column are log-log versions of the figures in the left column. As seen, most error-distances are less than 100 meters, but a few large errors of around 2000 meters are reported by all methods.

doi:10.1371/journal.pone.0149105.g007
The greedy router selection approach is straightforward to implement, computationally efficient and produces results which can be easily interpreted. However, due to the lack of knowledge of other routers surrounding the selected access points, the results are not robust. Whenever one of the important routers is replaced by another device in its location, it is not possible to recognize and merge the new label with the previous one. Similarly, when one of the important routers is moved to a new physical location, it is not possible to merge the two places.

None of the methods described in this paper require a specification of the number of stop-locations to find. This is an advantage because the problem of scale makes it impossible to give an objectively correct estimation of this. The three different methods find very different number of clusters (see Fig 8 for an example). The GPS-based method infers 16 distinct clusters, the greedy single-router based method infers 35 distinct clusters, and the DBSCAN-based WiFi method infers 69 distinct clusters. Adding to the complexity of the problem, the number of clusters found by the different methods is strongly dependent on the parameters of each method. For the GPS-based method, the parameters are $d_{\text{max}}$ and the two parameters $\varepsilon$ and $M$ for DBSCAN. For the greedy router selection the parameter is $\Delta N$. For the DBSCAN-based WiFi method, the parameters are $\varepsilon$ and $M$ for DBSCAN. Additionally all methods have variability in their pre- and post processing steps, for example the bin-size when time-binning and removal of short stop-locations.

![Fig 8. The three approaches produce a different number of points of interest. Density based clustering of GPS data (left) produces the lowest number of stop locations, followed by greedy selection of routers (middle), and DBSCAN (right). All the stops from GPS are reflected using WiFi data, but WiFi based methods identify locations with a higher spatial resolution.](doi:10.1371/journal.pone.0149105.g008)
Finally, there is the matter of non-stationary stop-locations in WiFi data. When using WiFi to detect stop-locations, it is possible to observe stop-locations which are not spatially stationary—this is for example due to personal MiFi devices and access points located in for example busses and trains. Examples of such non-stationary stop-locations are shown in Fig 9. When evaluating the start and stop-times of stop-locations, such non-stationary stop-locations will affect the results of the WiFi-based methods negatively.

We realize that using the data from a single subject for our study is a limitation to the generalizability of the findings. Nevertheless, the particular individual reveals mobility pattern at least as complex as we would expect from a typical adult: she works at two separate venues, appears to have two home locations (places visited on weeknights), and visits different areas of the city.

**Future work.** To achieve better results in the evaluations, one could filter out mobile routers—either by manually picking out SSID’s or by detecting routers which appear in different geographical locations. The former requires location-specific knowledge as each city/country has a different naming scheme for the routers on public transportation. The latter involves coupling the WiFi information with GPS data; in this work we intended to show that detecting stop-locations is possible with just the WiFi data.

Further, in the proposed methods, we are not explicitly modeling the temporal dimension of the problem. If two routers are often observed close in time, the physical distance between them...
them is likely to be low. Using this temporal closeness might also enable the construction of hierarchical clusters based on WiFi, consequently ameliorating the problem of scale.

Acknowledgments

All map tiles used in the article are provided by Stamen Design under CC BY 3.0 license. All map data acquired from OpenStreetMap under CC BY SA license.

Author Contributions

Conceived and designed the experiments: DKW PS MAF SL. Performed the experiments: MAF. Analyzed the data: DKW PS MAF. Contributed reagents/materials/analysis tools: DKW PS SL. Wrote the paper: DKW PS SL.

References


Appendix E

Conservation laws in human mobility
Conservation laws in human mobility

L. Alessandretti¹, P. Sapiezynski², S. Lehmann²,³*, A. Baronchelli¹*

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

* E-mail: sune.lehmann@gmail.com, andrea.baronchelli.1@city.ac.uk

Faced with effectively unlimited choices of how to spend their time, humans are constantly balancing a trade-off between exploitation of familiar places and exploration of new locations. Previous analyses have shown that at the daily and weekly timescales individuals are well characterized by a geo-spatial signature of repeatedly visited locations. How this geo-spatial signature evolves in time, however, remains unexplored. Here we analyse high-resolution spatio-temporal traces from 850 individuals participating in a 24-month experiment. We find that, although geo-spatial signatures undergo dramatic changes, the number of familiar locations an individual visits at any point in time is a conserved quantity. We show that this number is similar for different individuals, revealing a substantial homogeneity of the observed population. We point out that the observed fixed size of the geo-spatial signature cannot be explained in terms of time constraints, and is therefore a distinctive property of human behavior.
There is a disagreement between the current scientific understanding of human mobility as highly predictable and stable over time (1–3) and the fact that individual lives are constantly evolving due to changing needs and circumstances (4). Recent studies based on the analysis of human digital traces including mobile phone records (5, 6), online location-based social networks (7–11) and GPS location data of vehicles (12–17) have shown that individuals universally exhibit a markedly regular pattern characterized by few locations where they return regularly (18, 19) and predictably (20). However, the observed regularity mainly concerns human activities taking place at the daily (21, 22) or weekly (5, 6, 8) time-scales, such as commuting between home and office (5, 6, 23, 24), pursuing habitual leisure activities, and socializing with established friends and acquaintances (7). The role played by inherent life changes is not well understood and their effects are not included in the available models of human mobility behavior (2, 25–29).

Here, we quantify the development of individuals routines across months and years, characterising how individuals balance the trade off between the exploitation of familiar places and the exploration of new opportunities. Our study is based on the mobility traces of ~850 university students involved in the Copenhagen Networks Study experiment (30) over a period of 24 months (Table S1, Fig. S1). The physical location of individual mobile devices across time is inferred using a combination of an individual’s WiFi scan time-series and GPS coordinate scans (see (31), Figs. S2 - S4). While the location data based on mobile phone calls and location-based social networks typically used for mobility analyses is unevenly sampled (32, 33), the temporal sampling of WiFi traces in this dataset is even, with median time between scans $\Delta t = 16\,\text{sec}$ (Figs. S5 - S7). Fixed rate sampling is important in order to capture mobility patterns beyond highly regular ones such as home-work commuting (34), as individuals’ voice-call/SMS/data activity may be higher in certain preferred locations, leading to possible biases in the analyses based on telecommunication data (23). Also, because of their relatively short range signal, ubiq-
uitous WiFi access points (AP) can be localized with a typical spatial resolution of the order of tens of meters (35, 36). Notwithstanding significant differences in data collection, and the homogeneity of the population sample (university students), we have verified that the Copenhagen Networks Study dataset displays statistical properties across a wide range of measures that are consistent with previously analysed data on human mobility (1, 2) (Figs. S8- S10). Yet its temporal duration and spatial resolution makes it ideal for investigating the long-term evolution of individual geo-spatial behaviors.

When initiating a transition from a place to an other, individuals may either choose to return to a previously visited place, or explore a novel location. To characterize this exploration-exploitation trade-off, we represent individual geo-spatial trajectories as sequences of locations, where ‘locations’ are defined as places where participants in the experiment stopped for more than 10 minutes (31), Fig. 1A. A first question concerning the exploration behavior of the individuals is whether an individual’s set of known locations is continuously expanding, or instead its size saturates over time. We find that the total number of unique locations $L_i(t)$ an individual $i$ has discovered up to time $t$ grows as $L_i \propto t^{\alpha_i}$ (Fig. 1B), and that individuals’ exploration is homogeneous across the population studied, with $\alpha_i$ peaked around $\bar{\alpha} = 0.66$ (Fig. 1C). This sub-linear growth occurs regardless of how locations are defined or when in time the measurement starts (Figs. S11, S12). It is a characteristic signature of Heaps’ law (37) and implies that the rate of discovery of new locations decreases continuously during the entire duration of the experiment. We also find that, while continuously exploring new places, individuals allocate most of their time among a small subset of all visited locations (Fig. S13), in agreement with previous research on human mobility behavior (18–20). Hence, at any point in time, each individual is characterized by an individual geo-spatial signature (GS) of familiar locations (Fig. 1D). We define the geo-spatial signature as the set $GS_i(t) = \{\ell_1, \ell_2, ..., \ell_k, ..., \ell_C\}$ of locations $\ell_k$ that individual $i$ visited at least twice and where she spent on average more than
10 minutes/week during a time-window of 10 consecutive weeks preceding time \( t \). The results described in the following of this paper are robust with respect to variations of this definition, such as changes of the time-window size (Figs. S14, S15, and Table S2) or the definition of a location (Figs. S16, S17).

Thus, individuals continually explore new places yet they are loyal to a limited number of familiar locations forming their GS. But how does discovery of new places modify an individual’s GS? We find that the average probability \( P \) that a newly discovered location will become part of the GS stabilizes at \( P = 20\% \) on the long term (Fig. S18), indicating that individual GS are inherently unstable and new locations are continuously added. However, over time individuals may also cease to visit locations that are part of the GS. The balance between newly added and dismissed familiar locations is captured by the temporal evolution of the GS spatial capacity and net gain. We define spatial capacity \( C_i \) as the number of an individual’s familiar locations, i.e. the GS size, at any given moment. The net gain \( G_i \) is defined as the difference between the number of locations that are respectively added \( (A_i) \) and removed \( (D_i) \) at a specific time, hence \( G_i = A_i - D_i \). Fig. 2A shows the evolution of the average population capacity \( \bar{C} \). We find that the average capacity \( \bar{C} \) is constant in time, with a linear fit of the form \( \bar{C} = a + m \cdot t \) yielding \( m = -0.05 \pm 0.1 \). Thus, despite individual GS evolving over time, the average capacity is a conserved quantity.

The conservation of the average spatial capacity may result from either (i) each individual maintaining a stable number of familiar locations over time or (ii) a substantial heterogeneity of the population with certain individuals shrinking their set of familiar locations and other expanding theirs. We test the two hypotheses by measuring the individual average net gain across time \( \langle G_i \rangle \) and its standard deviation \( \sigma_{G,i} \). If a participant’s average gain is closer than one standard deviation from 0, hence \( |\langle G_i \rangle|/\sigma_{G,i} < 1 \), then the net gain is consistent with \( \langle G_i \rangle = 0 \). If this is true for the majority of participants, the spatial capacity is conserved.
at the individual level and hypothesis (i) holds. If, on the other hand, \( |\langle G_i \rangle|/\sigma_{G,i} \geq 1 \), the individual capacity must either increase or decrease in time, supporting hypothesis (ii). We find that hypothesis (i) holds for 96.35% of individuals (Fig. 2B). For the large majority of the population, the average net gain of familiar locations added or removed to the GS at any instant of time is not significantly different from 0, hence their individual capacity is conserved (Fig. S19, Table S3). Also, we find that the ratio between the average individual capacity and its standard deviation across time is smaller than 23% for 75% of the population (Fig. S20), demonstrating that fluctuations of the capacity are relatively small.

These results indicate that each individual is characterized by a fixed-size but evolving set of familiar locations. While the size depends on the spatial resolution chosen to define locations (Fig. S21), the population is homogeneous within this constraint (Fig. 3B). To clarify the information contained in the precise value of the spatial capacity, we randomize the temporal sequences of locations in two ways, preserving routines of individuals only up to the daily level. After breaking individual time series into modules of 1 day length, (i) we randomize individual timeseries preserving the module/day units (local randomizations) or (ii) we create new sequences by assembling together modules extracted randomly by the whole set of individual traces (global randomization). Due to the absence of temporal correlations, the capacity is constant in time also for the randomized datasets (see Fig. 2A). However, the capacity of the random sets is significantly higher than in the real time series for both randomizations (Fig. S22, Table S4), implying that the observed value in real data is not a simple consequence of time constraints. Instead, the fixed capacity is an inherent property of human behavior.

The time evolution of the GS supports this finding. We measure the turnover of familiar locations using the Jaccard similarity \( J_i(t, \gamma) \) between the GS at \( t \) and at \( t + \gamma \) (Fig. 2C). Despite seasonality effects which imply fluctuations around a typical behavior, \( J_i \) does not depend on the initial point but only on the waiting time \( \gamma \), and we can consider \( J_i(\gamma) \) independently of \( t \).
Do not distribute (Fig. S23). We find that the average similarity decreases as a power law $J \propto \gamma^\lambda$ with coefficient $\lambda = 0.29$. On the other hand, for the randomized sequences, the Jaccard similarity is constant in time as familiar locations are never abandoned ($J \propto \gamma^0$). Also, individuals keep visiting only few locations for long periods of time, in contrast to the randomized cases (Fig. 2D). This confirms that individual sets of familiar locations change continually and individual routines evolve in time.

In order to characterize the sub-structure of the geo-spatial signatures, we investigate how individuals allocate time among different classes of locations defined on the basis of their average visit duration. We consider intervals $\Delta T$, with $\Delta T$ ranging from 10 to 30 minutes per week (the time it takes to visit a bus stop or grocery shop) up to 48 to 168 hours per week (such as for home locations) (Fig. S24). For each of these locations classes, we compute the evolution of the capacity $c_i^{\Delta T}$ and the gain $G_i^{\Delta T}$, and test the hypothesis $G_i^{\Delta T} = 0$, as above. We find that, although the GS subsets are continuously evolving (Fig. S25), $c_i^{\Delta T}$ is conserved for each $\Delta T$ (Fig. 3C, Table S5, Fig. S26), indicating that the number of places where individuals spend a range of time $\Delta T$ does not change over time. This result holds independently of the choice of specific $\Delta T$ (Fig. S27, Table S5) and implies that the individual capacity $C_i = \sum c_i^{\Delta T}$, where both $C_i$ and each $c_i^{\Delta T}$ are conserved across time. Thus, both spatial capacity and time allocation are conserved quantities.

In summary, we have shown that the number of locations an individual visits regularly is conserved in time, even while individual routines are unstable in the long term because of the continual exploration of new locations. This individual spatial capacity is peaked around a typical value across the population, which is significantly lower than expected if only time-constraints were at play. Finally, this spatial capacity is hierarchically structured, indicating that individual time allocation is also conserved.

Taken together, these findings shed new light on the long-term dynamics of human mobility,
with potential impact for a better understanding of phenomena such as urban development and epidemic spreading. They will also help test and improve existing models of human mobility (2, 25–29) which were not designed to account for long-term instabilities and fixed-capacity effects. Extending our scope beyond mobility, it is interesting to note that similar fixed-size effects in the social domain (38–41) have been put in direct relation with human cognitive abilities (38). We anticipate that our results will stimulate new research exploring this connection.
Figure 1: Geo-spatial signatures and exploration of new locations. (A) An example of individual mobility trace. The visiting temporal pattern of the six most visited locations are shown (Loc1, ..., Loc6) along with the black trace including all visits to these 6 locations (Loc1-6). (B) Total number of discovered locations, $L$. Time is measured in $t$ days since an individual received the phone. The figure shows the 50% (dark grey area) and the 90% (light gray area) of the population, the average across users $\bar{L}$ (black line) and a power-law fitting function (red dashed line) with exponent $\alpha$. (C) The distribution of individuals power-law fit coefficients $\alpha_i$ is peaked around its average value $\overline{\alpha} = 0.66$. (D) Example of an individual’s geo-spatial signature. Locations are represented as pins on a map. The six most visited locations are displayed as larger pins using the same color scheme of panel A. Yellow areas show Copenhagen city and DTU University.
Figure 2: Conserved size of evolving geo-spatial signatures. (A) Evolution of individual capacity. The light blue area represents 50% of the population, the blue line its average $\mathcal{C}$, and the blue dashed line is the linear fit. The error on the angular coefficient of the fitting line, reported in the legend, shows that the fit is compatible with a constant line. The capacities resulting from the local (orange line) and global (green line) randomizations are also reported. (B) Gain standard deviation $\sigma_{G,i}$ vs the average gain $\langle G_i \rangle$ (lines obtained through a kernel density estimation from the data). The grey area corresponds to individuals for which $|\langle G_i \rangle| < \sigma_{G,i}$, i.e. whose average gain is compatible with zero (red bar in the illustrative inset). It contains 97.84% of the population. (C) The average Jaccard similarity $\bar{J}$ between geo-spatial signatures measured at $t$ and $t + \gamma$ as a function of $\gamma$ for data (blue line), and the randomized series (orange and green lines). Dashed lines correspond to power-law fits $\bar{J} \sim \gamma^\lambda$. (D) Probability distribution of the time interval between first and last occurrences of a location for data (blue line) and the randomized cases (orange and green lines).
Figure 3: Conservation of time allocation. (A) The average capacities (full lines) $c^{\Delta T}$ and the corresponding fits (dashed lines) as a function of time computed for several categories of locations $\Delta T$. (B) Fit coefficients are consistent with 0 within errors (Table S5). (C) Frequency histogram of individuals according to their average individual capacity $C_i$. 
References and Notes


31. Materials and Methods are available as supporting material on Science Online.


Acknowledgements

Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA. This work was supported by Villum Foundation, http://villumfoundation.dk/C12576AB0041F11B/0/4F7615B6F43A8EA5C1257AEF003D9930?OpenDocument, Young Investigator programme 2012, High Resolution Networks (SL) and University of Copenhagen, http://dsin.ku.dk/news/ucph_funds/, through the UCPH2016 Social Fabric grant (SL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary Material

Materials and Methods
Supplementary Text
Figs. S1 to S27
Tables S1 to S5
References 42 to 45
Materials and Methods

Data description and pre-processing

In this section we provide a description of the raw data used for this study and of the pre-processing operations applied. The data was collected by the Copenhagen Network Study (CNS) experiment [42]. We made use of WiFi scans data and GPS scans data.

WiFi data description

The WiFi dataset provides the time-series of wireless network scans performed by participants’ mobile devices. Each record \((i, \text{timestamp}, \text{SSID}, \text{BSSID}, \text{RSSI})\) indicates the participant ID \(i\), the name of the wireless network scanned \(\text{SSID}\), the MAC address \(\text{BSSID}\) uniquely identifying the AP providing access to the wireless network, the time of scan in seconds \(\text{timestamp}\), and the signal strength \(\text{RSSI}\) in dBm.

APs do not have geographical coordinates attached. However, their time sequence is effectively equal to location data, because APs positions tend to be fixed. In the following we will estimate participants’ position over time using the sequence of WiFi APs, that are evenly scanned, in combination with GPS scans, that are not evenly sampled but provide precise geographical coordinates. We will not take into account the signal strength, as it was proven that there is only a weak correlation between the \(\text{RSSI}\) and the distance between the mobile device and the AP [43]. Table S1 summarises the main characteristics of the WiFi dataset.

- Time resolution

  Throughout the experiment, participants’ devices scanned for WiFi every \(\Delta t\) seconds,
where $1 < \Delta t < 10^7$ (Figure S5-c). In figures (Figure S5-a and b) we show respectively the distribution of participants with respect to the median time and the mean time between consecutive scans. For half of the population the median time is lower than $\Delta t_m = 16 \text{ sec}$. Instead, the average is higher, with median value across the population $\overline{\Delta t}_m = 55 \text{ sec}$. This can be explained considering that participants may switch off Wi-Fi detection on their phone for longer periods of time. As detailed in the next sections, we restrict the analysis to periods with high enough temporal resolution.

- **Number of participants**
  The CNS data collection took place between September 2013 and September 2015. In total, 851 students were involved in the experiment, but not all of them engaged for its entire duration. In figure S1-a, we show the number of participants $N_{\text{active}}$ as a function of time. In figure S1-b we show the histogram of individual total participation time in the CNS experiment.

- **APs**
  A total number of $N = 7596742$ different APs were detected throughout the experiment. To estimate APs locations, we followed the procedure described in [43] making use of participants’ sequences of GPS scans. First, we discarded mobile APs, that are located on buses or trains, and moved APs that were displaced during the experiment (for example by residents of Copenhagen changing apartment, taking their APs with them). Then, we considered all WiFi scans happening within the same second as a GPS scan to estimate APs location. The APs location estimation error is below 50 meters in 99% cases. Using this procedure, $N_{\text{geo}} = 303482$ APs were geolocalised. Most of the APs are located in the Copenhagen area (see figure S2). APs are heterogeneous in terms of number of total visits, visitation frequency and number
of individuals presence. The median length of the longest observation for each APs is 4min. This implies that most of the APs can be discarded for our analysis since they are scanned only for short observational periods.

**Data pre-processing: temporal aspect**

Detailed temporal information is not meaningful for the study of individual long-term behaviour. Data was aggregated in bins of length 1min, and reshaped by defining APs observation intervals. Starting from the series of records \((i, BSSID, t)\), we derived the sequence of observation intervals \(I = (user\_id, BSSID, \{t_0,t_f\})\), such that the AP identified by \(BSSID\) was scanned in all bins between \(t_0\) and \(t_f\) where at least one scan was performed by an individual \(i\).

Individuals long-term mobility behaviour can be estimated for individuals whose position is known for a considerable fraction of time. For each individual \(i\), we measure the **time coverage** \(TC_i\) as the fraction of time an individual location is known. The time coverage is higher than 0.8 for 50% of the population (see Figure S6) when considering only geolocalised APs. However, its value is not constant in time, but displays fluctuations due to seasonality effects (Figure S7).

**Data pre-processing: spatial aspect**

Quantifying individuals mobility behaviour requires to define spatial “locations” identifying places that are significant in human daily experience (i.e. homes, offices, cafés, shops...). There is not a univoque definition of locations: here, we propose a method to cluster APs into “locations” based on their reciprocal distance and simultaneous detection.
The information on simultaneous detection of two APs is included in the indirect graph \( G = (V,E) \). \( V \) is the set of geo-localised APs, links \( e(j,k) \) exist between pairs of access points that have ever been scanned in the same 1min bin by at least one user. The physical distances \( \text{dist}(j,k) \) for all pairs of \( (j,k) \in E \) can be easily computed since their position has been estimated.

We consider the set of links \( E_D \subset E \) such that \( \text{dist}(j,k) < d \), where \( d \) is a threshold value, and we define a new graph \( G_d = (V,E_d) \). Each connected component in the graph \( G_d \) include all APs that are closer than \( d \) to at least one other AP in the same component.

With our definition, a stop-location is a connected component in the graph \( G_d \). For \( d = 5m \) the maximal distance between two APs in the same location is smaller than 10m for most locations and at most 200m (see figure S4-a). The number of APs in the same location is lower that 10 for most locations, but reaches 700 for dense areas such as the University Campuses (one should also consider that APs are replaced throughout the experiment) (see figure S4-b). An example of APs clustering for \( d = 5m \) and \( d = 10m \) is shown in figure S3.

In the main text, we presented results obtained defining locations with threshold \( d = 5m \). However, the findings do not depend on the choice of the threshold (see the following sections). For \( d = 5m \) we find in total \( N_L = 123415 \) locations. Considering simply APs as a proxy for location introduces noise in the analysis.

Robustness Tests

The results presented in the MS do not depend on how locations are defined (many results hold simply considering Wi-Fi access points (APs) as proxy for locations), nor on the time-window used to investigate the long-term behaviour. In this section, we show how the results are derived and we demonstrate their statistical robustness. To avoid confusion, we will indicate with \( \bar{x} \) the average value of a quantity \( x \) across the population, and \( \langle x \rangle \) the
average across time.

*Exploration grows sub-linearly regardless of the definition of location and the measure starting point*

Individual exploration behaviour is quantified measuring the number of locations $L_i(t)$ discovered up to day $t$. In the MS, we show that for $d = 5m$, individual exploration does not saturate, with $L_i(t)$ growing sub-linearly in time. Here, we show that this holds also considering APs and locations with threshold $d = 10m$ (figure S11). Exploration behaviour is not affected by aging, as we verify by repeating the same measures starting $M$ months after the participant received the phone, with $M$ in $\{1,2,5,7,10,12\}$ (See figure S12).

*Conservation of capacity*

The number of locations an individual $i$ visits regularly is equivalent to the geo-spatial signature size $C_i(t) = |GS_i(t)|$. We call this quantity *spatial capacity*. The average individual capacity across the population $\overline{C(t)}$ is constant in time regardless of the definition of location (see Figure S16 and section S3 for the definition of locations) or the choice of the window size $W$ (Table S2 and Figure S14). This is tested by first performing a linear fit of the form $\overline{C(t)} = a + b \cdot t$ and a power-law fit of the form $\overline{C(t)} \propto t^b$, and then testing the hypotheses $H_0 : b = 0$, and $H_1 : \beta = 0$. In table S2 we show the fit coefficients computed with the least squares method, where the errors are estimated taking into account the standard deviation $\sigma_{C(t)}$ quantifying the population dispersion. Hypotheses $H_0$ and $H_1$ hold for all the choices of $W$. Also, the correlation hypothesis test, testing $\rho_{t,C} = 0$, where $\rho_{t,C}$ is the Pearson correlation coefficient between capacity and time, yields that there is not signifi-
cant correlation between time and capacity at $\alpha = 0.05$ with p-value $> \alpha$ for any choice of $W$ (table S2). The coefficient of determination $R^2$ is smaller than 0 in most cases meaning that a horizontal line fits better than the best fit computed with the least squared method.

Individual capacities $C_i(t)$ are also conserved across time. We apply a linear fit of the form $C_i(t) = a_i + m_i \cdot t$ and we find that the sample mean of the linear fit coefficients is $m = -0.054 \pm 0.048$ (Figure S19 and Table S3). The hypothesis $m = 0$ is not rejected at $\alpha = 0.05$ under the t-student hypothesis test, for locations with $d = 5m$, and APs (Table S3). We quantify the fluctuations of the individual capacity measuring the coefficient of variation $\sigma_{C_i}/\langle C_i \rangle$, where $\langle C_i \rangle$ is an individual’s average capacity and $\sigma_{C_i}$ the corresponding standard deviation. For $W = 10$, these fluctuations are smaller than 17% for 50% of individuals (see Figure S20), considering only weeks with time coverage higher than 80% (see Section S3 for the definition of time coverage). The individual average capacities $\langle C_i \rangle$ are homogeneously distributed around a typical value for the population that is determined by the choice made to define location (Figure S21).

**Individual gain is equal to zero**

Changes of the individual capacity are better quantified measuring the net gain, defined as $G_i(t) = A_i(t) - D_i(t)$, where $A_i(t) = |GS_i(t) \setminus GS_i(t - dt)|$ is the number of location added and $D_i(t) = |GS_i(t - dt) \setminus GS_i(t)|$ is the number of location removed from the geo-spatial signature during $dt$, where $dt = 1$ week. Figure S15 shows the individual average capacity $\langle C_i \rangle$, activation $\langle A_i \rangle$, deactivation $\langle D_i \rangle$ and weekly gain $\langle G_i \rangle$, as a function of the windows size $W$. The individual average gain $\langle G_i \rangle$ is consistent with 0 for all choices of $W$, because of the conservation of individual capacities across time (Figure S15, right column).
The individual average gain $\langle G_i \rangle$ is consistent with 0 for more than 95% of individuals, independently of how locations are defined. This is verified testing whether the ratio $\sigma_{G,i}/\langle G_i \rangle > 1$, where $\sigma_{G,i}$ is the standard deviation of the average individual net gain across time (see main text). In figure S17, we show the correlation between $\sigma_{G,i}$ and $\langle G_i \rangle$, for different definitions of locations.

**Absence of aging effects**

Individuals are continuously modifying their geo-spatial signatures: this evolution is quantified measuring the Jaccard similarity $J_i(t, \gamma) = |GS_i(t) \cap GS_i(t + \gamma)|/|GS_i(t) \cup GS_i(t + \gamma)|$ (see MS). The average similarity $\bar{J}(t, \gamma)$ decreases in time: power-law fits of the form $\bar{J}(t, \gamma) = p(t) \cdot \gamma^\lambda(t)$ yield $\lambda < 0$ for all $t$. The fit coefficient $\lambda(t)$ fluctuates around a typical value, because of seasonality effects, but does not substantially increases or decreases as a function of the waiting time $t$ (figure S23, bottom). Hence, we consider $\bar{J}(\gamma)$, regardless of the waiting time $t$ (figure S23, above, full line).

**Conservation of time allocation**

In this section we show how the results presented in the main text hold independently of how these classes are defined.

The individual sub-capacity are defined as $C_i(t)^{\Delta T} = |GS_i(t)^{\Delta T}|$. The average sub-capacities $\overline{C}^{\Delta T}(t)$ are constant in time for several choices of $\Delta T$ and different definitions of location. This is verified with the linear fit test as detailed in section S1.2 (see table S3 and figure...
We consider both arbitrary time intervals and logarithmic spaced intervals.

Individuals net gain, with respect to changes of the GS subsets is zero. The study of the ratios $\sigma_{G,i}^{\Delta T}/\langle G_i^{\Delta T} \rangle$, where $\langle G_i^{\Delta T} \rangle$ is the average individual net gain for a class $\Delta T$, and $\sigma_{G,i}^{\Delta T}$ is the corresponding standard deviation, yields that the ratio is higher than 1 for more than 98% of individuals, for all $\Delta T$ (Figure S26).

**Discrepancy with the randomized cases**

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

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

The individual randomized capacity $\langle C_{rand,i} \rangle$ averaged across time, (see figure S22), is higher than in the real case both for the global and the local randomization cases. We compute the Kolmogorov-Smirnov test-statistics (Table S4 to compare the real sample with the randomized samples. We reject the hypothesis that the two samples are extracted from the same distribution since p-value < $\alpha$ with $\alpha = 0.01$. 

8
do not distribute
Supplementary text

Establishment of individual geo-spatial signatures

At any point in time individuals allocate most of their time among few locations. For each user $i$, we consider the set of locations $GS_i(t) = \{\ell_1, \ell_2, ..., \ell_C\}$ seen in the $W$ weeks preceding $t$ at least twice and such that $T_{i,\ell}(t) > W \cdot 10 \text{min}$, where $T_{i,\ell}(t)$ is the total time of observation of location $\ell$ during the $W$ weeks. We call this subset the geo-spatial signature (GS). In figure S13 (left), we show that for $W = 10$ weeks, the GS contains on average $\sim 4\%$ of all locations seen during the same 10 weeks. Yet the time spent in these locations is on average $\sim 94\%$ of the total time (figure S13, right).

Evolution and composition of the geo-spatial signature subsets

Individuals are continually discovering locations, but only some among them become part of the GS. The probability that a newly discovered locations at time $t$ will be introduced in the geo-spatial signature is $P_i(t) = L_{i,GS}(t)/L_i$, where $L_i$ are all the locations discovered by an individual $i$ at $t$ and $L_{i,GS}$ are the one that will be part of the individual’s signature. We show in figure S18, the average value of $P$ across time. $P$ stabilizes at $\mathcal{P} = 20\%$.

Individuals allocate time heterogeneously among locations, due to their different functions (homes, work-places, shops, universities, leisure places...). We study time allocation between different classes of locations considering subsets of the geo-spatial signature defined on the basis of the total visitation time. The subsets of the geo-spatial signature $GS_i(t)^{\Delta T} \in GS_i(t)$ include all locations seen in the $W$ weeks preceding $t$ at least twice and such that $W \cdot \Delta t(0) < T_{i,\ell}(t) < W \cdot \Delta t(1)$ where $T_{i,\ell}(t)$ is the time of observation of location
We test several choices of $\Delta T$. We find that when $\Delta T$ increases, the subsets are empty for many individuals, since no locations satisfy the above-mentioned criteria. In figure S24, we show the distribution of average individual sub-capacities $\langle C^{\Delta T}_i \rangle$. Only subsets with small enough $\Delta T$ are significant for more than 50% of the population, and typically each individual has 1 location where he/she spend more than 48 hours per week.

The evolution of these subsets in time is quantified measuring the Jaccard similarity, as detailed for the entire GS in the MS. The average similarity $\bar{J}(\gamma)$ decreases in time for all $\Delta T$. However, a power-law fit of the form $\bar{J}(\gamma) \sim \gamma^\lambda$, yields that the coefficients $\lambda$ decreases as $\Delta T$ increases, since places where individuals spend more time are changed less frequently (Figure S25).

**Comparison with previous research**

The CNS dataset provides novel results on the long term stability of human mobility patterns. Nevertheless, this dataset displays statistical properties that are consistent with previously analysed data on human mobility. The average radius of gyration (see [44], SI for definition) displays high fluctuations in time due to seasonality effects (during summer many students travel, for example). However, as highlighted by other studies [44], this quantity grows initially fast, and then tend to saturate (figure S10-A). Individuals are distributed heterogeneously with respect to their radius of gyration measured at the end of the experiment, with the probability distribution $P(r_g)$ (figure S10-B) decaying as a power-law with coefficient $\beta = -1.47$. This is comparable
with the results found in [44], $\beta = -1.65$ and [45] $\beta = -1.55$, where both studies relied on CDRs.

The distribution between consecutive jumps $P(\Delta r)$ is also power law distributed (Figure S10-C), with exponent $\beta = -1.75$. Gonzales et. al [44] found the same exponent for the truncated power-law distribution, Song et. al [45] found a power-law distribution of jumps with exponent $\beta = -1.55$.

The visitation frequency of a location is defined as $f_l = m_l / \sum_{k=0}^{NL} (m_k)$ where $m_l$ is the total number of visits to the $l$th location, and $NL$ is the number of locations. We find that the visitation frequency of a location with rank $r$, where the rank is attributed based on the visitation frequency, follows a Zipf’s law $f(r) \propto r^{-\zeta}$, with $\zeta = 1.03$ (Figure S8). Our result is not far with the one obtained by Gonzales et. al, who found that $f(r) \propto 1/r$ [44], and Song et al who found $f(r) \propto r^{-1.2\pm0.1}$ [45].
<table>
<thead>
<tr>
<th>Participants</th>
<th>APs</th>
<th>Scans</th>
<th>Dataset size</th>
<th>Period of measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>851</td>
<td>$7.6 \cdot 10^6$</td>
<td>$6 \cdot 10^9$</td>
<td>91Gb</td>
<td>1st September 2013 12AM (UTC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30th September 2015 10AM (UTC)</td>
</tr>
</tbody>
</table>

**Table S1:** Main characteristics of the CNS WiFi dataset

*do not distribute*
### Table S2: Conservation of the average capacity

For different windows sizes $W$, we show: the Pearson correlation coefficient $\rho_{t,C}$ with the corresponding p-value, testing the hypothesis $\rho_{t,C} = 0$ (there is no significant correlation between time and capacity at $\alpha = 0.05$ when $p$-value $> \alpha$). The linear fit coefficient $m$, with the corresponding coefficient of determination $R^2$ (the coefficient of determination is negative because a horizontal line fits better than the best fit), and the power-law fit coefficient $\beta$.

<table>
<thead>
<tr>
<th>$W$</th>
<th>$\rho_{t,C}$</th>
<th>p-value</th>
<th>Linear fit coeff $m$</th>
<th>$R^2$</th>
<th>PL fit coefficient $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.057</td>
<td>0.58</td>
<td>-0.006$\pm$0.042</td>
<td>-0.049</td>
<td>-0.023$\pm$0.042</td>
</tr>
<tr>
<td>2</td>
<td>-0.035</td>
<td>0.735</td>
<td>-0.005$\pm$0.061</td>
<td>-0.026</td>
<td>-0.011$\pm$0.061</td>
</tr>
<tr>
<td>6</td>
<td>0.057</td>
<td>0.589</td>
<td>0.032$\pm$0.088</td>
<td>-0.015</td>
<td>0.024$\pm$0.088</td>
</tr>
<tr>
<td>10</td>
<td>0.107</td>
<td>0.325</td>
<td>0.046$\pm$0.103</td>
<td>-0.006</td>
<td>0.033$\pm$0.103</td>
</tr>
<tr>
<td>12</td>
<td>0.097</td>
<td>0.377</td>
<td>0.044$\pm$0.109</td>
<td>-0.011</td>
<td>0.033$\pm$0.109</td>
</tr>
<tr>
<td>16</td>
<td>0.078</td>
<td>0.49</td>
<td>0.038$\pm$0.121</td>
<td>-0.018</td>
<td>0.031$\pm$0.121</td>
</tr>
<tr>
<td>24</td>
<td>-0.172</td>
<td>0.145</td>
<td>-0.022$\pm$0.147</td>
<td>0.016</td>
<td>-0.015$\pm$0.147</td>
</tr>
<tr>
<td>32</td>
<td>-0.149</td>
<td>0.235</td>
<td>-0.023$\pm$0.176</td>
<td>0.014</td>
<td>-0.018$\pm$0.176</td>
</tr>
<tr>
<td></td>
<td>$\bar{m}$</td>
<td>$t$ - statistics</td>
<td>$p$ - value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
<td>------------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>locations, d=5m</td>
<td>$-0.054 \pm 0.048$</td>
<td>-1.128</td>
<td>0.259</td>
<td></td>
<td></td>
</tr>
<tr>
<td>locations, d=10m</td>
<td>$-0.073 \pm 0.027$</td>
<td>-2.677</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APs</td>
<td>$0.134 \pm 0.370$</td>
<td>0.360</td>
<td>0.718</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table S3: Conservation of individual capacities** For locations and APs, we show the results of a t-statistics hypothesis test, testing the hypothesis $\bar{m} = 0$, where $\bar{m}$ is the sample average of individual linear fit coefficients. The hypothesis $\bar{m} = 0$ is not rejected when $p$-value > $\alpha = 0.05$.

<table>
<thead>
<tr>
<th></th>
<th>KS statistics</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>local randomization</td>
<td>0.35</td>
<td>$8 \cdot 10^{-43}$</td>
</tr>
<tr>
<td>global randomization</td>
<td>0.96</td>
<td>$2 \cdot 10^{-321}$</td>
</tr>
</tbody>
</table>

**Table S4:** Results of the Kolmogorov-Smirnov comparing the individual average capacity sample, with both the randomized individual average capacity sample
<table>
<thead>
<tr>
<th>$\Delta t$</th>
<th>Linear fit coeff</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Arbitrary time intervals</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.17-1.0 hw</td>
<td>0.055 ± 0.076</td>
<td>0.006</td>
</tr>
<tr>
<td>1.0-6.0 hw</td>
<td>-0.023 ± 0.037</td>
<td>-0.034</td>
</tr>
<tr>
<td>6.0-12.0 hw</td>
<td>0.005 ± 0.009</td>
<td>0.142</td>
</tr>
<tr>
<td>12.0-24.0 hw</td>
<td>0.005 ± 0.006</td>
<td>0.376</td>
</tr>
<tr>
<td>24.0-48.0 hw</td>
<td>0.001 ± 0.005</td>
<td>-0.01</td>
</tr>
<tr>
<td>48.0-168 hw</td>
<td>0.003 ± 0.005</td>
<td>0.163</td>
</tr>
<tr>
<td><strong>Logarithmic spaced intervals</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.17-0.69 h/w</td>
<td>0.049 ± 0.067</td>
<td>0.009</td>
</tr>
<tr>
<td>0.69-2.85 h/w</td>
<td>-0.015 ± 0.039</td>
<td>-0.056</td>
</tr>
<tr>
<td>2.85-11.70 h/w</td>
<td>-0.002 ± 0.019</td>
<td>-0.018</td>
</tr>
<tr>
<td>11.70-48.0 h/w</td>
<td>0.006 ± 0.008</td>
<td>0.214</td>
</tr>
<tr>
<td>48.0-168 h/w</td>
<td>0.003 ± 0.005</td>
<td>0.163</td>
</tr>
</tbody>
</table>

**Table S5: Conservation of time allocation** For different $\Delta T$, we show the linear fit coefficient of the average capacity, and the coefficient of determination of the linear fit $R^2$. 
Figure S1: Participation of individuals in the CNS experiment

a) Number of active users in the experiment as a function of time

b) Frequency histogram of participants according to the number of days they took part in the experiment.
Figure S2: Distribution of geo-localised APs at different geographical scales Heat maps of the number (in log 10) of geo-localised APs in the world (a), in Europe (b) and in the Copenhagen area (c)
Figure S3: Comparison between locations with threshold $d = 5$ and $d = 10$ An example of the clustering of APs located within Copenhagen city for thresholds $d = 5m$ (left) and $d = 10m$ (right). Dots corresponds to geo-localised APs, coloured according to the cluster they belong to. The coloured regions are the convex hulls of the set of APs in a same cluster. Grey lines are streets.
Figure S4: Changes in locations size and intra-distance for different choices of threshold

a The boxplots of the average (dark orange) and maximal (light orange) distance among APs located in the same location, as a function of the threshold $d$ used to merge APs. Boxes are set at the 1st and 3rd quantile, while whiskers at 2.5% and 97.5%.

b The boxplots of the locations size (# of APs) taking (dark blue) or not (light blue) into account isolated APs (components with size 1) as a function of the threshold $d$. 
Figure S5: High temporal resolution of the CNS Wi-Fi dataset

Frequency histogram of individuals according to the median (a) and mean (b) between consecutive scans. The black line is the median value across the population. 

Figure S6: For half of the population, the location is known for 80% of the time

The frequency histogram of individuals based on their time coverage, defined as $TC_i = T_i / T_{tot}$, where $T_i$ is the number of 1min bins where user $i$ detected at least one router, and $T_{tot}$ is the total time $i$ took part in the CNS experiment, considering all APs (a), and only geo-localised APs (b).
Figure S7: Fluctuations of the time coverage in time Average time coverage (fraction of time a user location is known) computed for APs (left) and Locations (right) considering a time-window of 10 weeks (black line) and without window, every 10 weeks (blue line)
Figure S8: Power-law behaviour of locations frequency-rank plot The visitation frequency $f_k$ as a function of a location rank $k$ for each individual (thin lines), the average rank (black thick line) and the power law fit $f_k \propto k^{-\zeta}$, with $\zeta = 1.03$ (dashed line).
Figure S9: Power-law distribution of individual radius of gyration

The probability density distribution of individuals final radius of gyration $r_{gi}(t_{max})$, where $t_{max}$ is the moment when an individual dropped the experiment, considering the entire individual history (red) and the last 30 days in the experiment (yellow line), the corresponding power-law fit (dashed line).
Figure S10: Power-law distribution of jump lengths

The probability density distribution of jump lengths (in Km) between consecutive stop-locations (where individuals stopped for more than 10 minutes), and the corresponding PL fit (dashed line).
Figure S11: Growth of individuals sets locations. The number of locations discovered up to day $t$ (a) by an individual $i$ (coloured dots), the average across users (black line) and its power-law fit (red dotted line) for locations with thresholds $d = 10m$ and APs (top to bottom). In the insets, the distribution of individuals fit coefficients.
Figure S12: Stability of exploration behaviour in time The average number of locations individually discovered in time, measured after waiting $M$ months, and the corresponding power-law function fit (dotted lines) for different values of $M$.

Figure S13: Establishment of geo-spatial signatures The frequency histogram of individuals according to (left) the average fraction of locations constituting the GS signature, (right) the average amount of time spent in the GS signature (left). The GS is computed for $W = 10$ weeks.
Figure S14: Conservation of capacity for different sliding window sizes. The evolution in time of the average capacity (full lines), and the corresponding linear fit (dashed lines) for several values of sliding window size $W$. The measure is computed for locations with threshold $d = 5m$. 

![Figure S14: Conservation of capacity for different sliding window sizes.](image)
Figure S15: Dependence on the window size for Capacity, Activation, Deactivation and Gain

The boxplots of the individual average capacity $\langle C_i \rangle$ (left, in black), individual activation $\langle A_i \rangle$ (center, in blue) and deactivation $\langle D_i \rangle$ (center, in red) and gain $\langle G_i \rangle$ (right, in green), as a function of the sliding window size for locations with $d = 5m$, $d = 10m$ and APs (from top to bottom). Boxes contains the population interquartile (25 to 75 percentiles) and whiskers contain the 95% of the population (2.5 to 97.5 percentiles).
Figure S16: Conservation of capacity under different definition of locations

The evolution of the capacity in time for locations with thresholds $d = 5m$, $d = 10m$ and APs (from top to bottom). Boxplots contain the interquartile range (25 to 75 percentiles), the full line is the average value, the dashed line is the average capacity linear fit.
**Figure S17: Individual average gain is zero** The 2d frequency histogram of the average individual gain $\langle G_i \rangle$ vs its standard deviation $\sigma_{G,i}$ for locations with thresholds $d = 5\text{m}$ (A), $d = 10\text{m}$ (B) and APs (C). Colors are attributed according to the number of individuals in logarithmic scale. The dark grey area corresponds to $|\langle G_i \rangle| < \sigma_{G,i}^i$, the light gray area to $|\langle G_i \rangle| > \sigma_{G,i}^i$.

**Figure S18: Probability of discovering a new familiar location** The fraction of new locations introduced in the geo-spatial signature as a function of the time elapsed since the beginning of the experiment.
Figure S19: Conservation of individual capacities under different definition of locations

The frequency histogram of individual fit coefficients $m_i$, where individual capacities $C_i(t)$ are modelled as $C_i(t) = a_i + m_i \cdot t$ for locations with thresholds $d = 5m$ (A), $d = 10m$ (B) and APs (C). The sample average coefficient $\bar{m}$ is consistent with 0 for locations and APs under t-statistics hypothesis test.
Figure S20: Limited fluctuations of individual capacity

The frequency histogram of the coefficient of variation $\sigma_{C,i}/\langle C_i \rangle$, where $\langle C_i \rangle$ is the individual capacity averaged across time, and $\sigma_{C,i}$ the corresponding standard deviation for locations with thresholds $d = 5m$ (A), $d = 10m$ (B) and APs (C). We consider only weeks with time coverage higher than 80%.
**Figure S21: Homogeneous distribution of individual capacity** The frequency histogram of the average individual capacity $\langle C_i \rangle$ for locations with thresholds $d = 5m$ (A), $d = 10m$ (B) and APs (C). We show the average value $\overline{C}$ (black line) with standard error SE.

**Figure S22: Difference between the randomized and the real individual capacity** The frequency histogram of the average individual capacity $\langle C_{rand,i} \rangle$ for local (A), global (B) randomizations, and data (C). We show the average value $\overline{C}$ (black line) with standard error SE.
Figure S23: Evolution of individual geo-spatial signatures. (A,B,C) The average overlap (Jaccard similarity) between the geo spatial signature at week $t$ and week $t+\gamma$ regardless of the waiting time $t$ (thick line), with the corresponding PL fit (dashed line), and for different $t$ (thin lines). (D,E,F) The PL fit coefficients $\lambda(t)$ vs $t$. Measures are performed for locations with thresholds $d = 5m$ (A,D), $d = 10m$ (B,E) and APs (C,F).

Figure S24: Composition of the geo spatial signature. The distribution of the average individual sub-capacities $\langle C_i \rangle^{\Delta T}$, for increasing intervals $\Delta T$ (left to right).
The average overlap (Jaccard similarity) between subsets of the geo-spatial signature at time $t$ and $t + \gamma$ vs $\gamma$ (thick lines), and the corresponding power-law fits $J(\gamma) \sim \gamma^\lambda$ (dashed lines) with exponent $\lambda$, for different values of $\Delta T$. Measures are performed for locations with thresholds $d = 5m$ (A), $d = 10m$ (B) and APs (C). Each line corresponds to a different choice of $\Delta T$. It is worth noting that decay of the overlap for the $> 48$ class is slower than for the other classes, and that this overlap is, in general, higher. This implies that locations where individuals spend a great amount of time every week are more stable, in agreement with previous results on the stability of human mobility [44, 45, 46].
Figure S26: Net gain consistency with zero for different choices of $\Delta T$ The 2d frequency histogram of the average individual gain $\langle G^{\Delta T_i} \rangle$ vs its standard deviation $\sigma_i^{G,\Delta T}$ for locations with threshold $d = 5m$ and sliding window $W = 10$. Colours are attributed according to the number of individuals in logarithmic scale. The dark grey area corresponds to $|\langle G_i^{\Delta T} \rangle| < \sigma_i^{G,\Delta T}$, the light gray area to $|\langle G_i^{\Delta T} \rangle| > \sigma_i^{G,\Delta T}$. 

...
Figure S27: Conservation of time-allocation, for different type of locations, and choices of \( \Delta T \) intervals. The evolution of the average sub-capacities in time for locations with thresholds \( d = 5m \) (arbitrary \( \Delta T \)), \( d = 10m \) (arbitrary \( \Delta T \)), APs (arbitrary \( \Delta T \)), and locations with threshold \( d = 5m \) (logarithmic \( \Delta T \)) (top to bottom).
References


Appendix F

Temporal Fidelity in Dynamic Social Networks
Temporal fidelity in dynamic social networks

Arkadiusz Stopczynski1,2, a, Piotr Sapiezynski1, Alex ‘Sandy’ Pentland2, and Sune Lehmann1,3

1 Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
2 Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA
3 Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark

Received 6 July 2015 / Received in final form 17 August 2015
Published online 7 October 2015 – © EDP Sciences, Societ`a Italiana di Fisica, Springer-Verlag 2015

Abstract. It has recently become possible to record detailed social interactions in large social systems with high resolution. As we study these datasets, human social interactions display patterns that emerge at multiple time scales, from minutes to months. On a fundamental level, understanding of the network dynamics can be used to inform the process of measuring social networks. The details of measurement are of particular importance when considering dynamic processes where minute-to-minute details are important, because collection of physical proximity interactions with high temporal resolution is difficult and expensive. Here, we consider the dynamic network of proximity-interactions between approximately 500 individuals participating in the Copenhagen Networks Study. We show that in order to accurately model spreading processes in the network, the dynamic processes that occur on the order of minutes are essential and must be included in the analysis.

1 Introduction

Temporal networks provide an important framework for modeling a variety of real systems [1]. Examples of complex systems where dynamics can play a central role include social networks, energy grids, networks of sexual contacts, and transportation systems [2–8].

Only recently, thanks to technical developments in data collection, it has become possible to collect high-resolution data about physical and virtual interactions in complex social systems. Using sociometric badges or smartphones, it is now possible to record interactions happening on multiple channels and at multiple timescales, measuring events with minute-by-minute resolution [6,9–12]. With access to such data, we can begin to describe the complexity, structure, and dynamics of such social systems [13]. Accurate measurements and models of social systems are necessary in order to understand how diseases spread [6,7], what makes teams productive [14,15], or how friendships form and disappear [11,13].

A fully-formed framework for incorporating network dynamics has yet to be established [16–20]. We know, however, that for many practical applications, it is important to get the details right, because variations in how the time dimension is incorporated can lead to significant differences in the modeling of dynamical processes unfolding on the network. Understanding spreading in dynamic networks is of particular interest, as these may represent a wide variety of processes in the system, including spreading of biological pathogens, information, knowledge, or behaviors.

Recently, there has been a growing interest in how to correctly and efficiently incorporate time dimension in the modeling of disease spread. Over the last few years, studies have focused on the mixing matrices capturing important epidemiological features [20], efficient representation of the spreading networks with coarser temporal representation [16,19], and the fundamental impact that temporal features have on the spreading process [17,18]. Here we study how the fidelity of representation of network behavior at short timescales – on the order of minutes – influences simulated spreading in the network. These minute-to-minute dynamics are particularly interesting because data collection with high temporal resolution tends to be challenging and costly. We consider the ramifications of reducing temporal resolution and which biases such a reduction introduces in terms of understanding spreading process in the temporal network of close proximity interactions.

The dataset

Here we analyze close proximity interactions network of participants of the Copenhagen Networks Study (CNS) [9]. This proximity dataset is based on Bluetooth scans collected using state-of-the-art smartphones. We define an interaction between users $i,j$ in 5-min time-bin $t$. 

* e-mail: arka@mit.edu
2 Results

2.1 Dynamics of a complex social system

The network of close proximity interactions in the CNS dataset displays dynamics at multiple time scales. We can observe distinct weekly and diurnal patterns (Fig. 1a). Concurrently, at the minute-by-minute resolution, we observe significant fluctuations in the number of links active within 5-min windows (Fig. 1a inset).

We can quantify the magnitude of network changes by considering the overlap between active links in consecutive time slices as a function of the duration of the aggregation time window. A link \((i,j)\) is considered active if at least one interaction happened on it within the aggregated time window. We define the overlap as

\[
J = \frac{|L_{t_1}^{\Delta T} \cap L_{t_2}^{\Delta T}|}{|L_{t_1}^{\Delta T} \cup L_{t_2}^{\Delta T}|}
\]

where \(L_{t}^{\Delta T}\) is set of links present in time \(t\), in a time window of size \(\Delta T\). The overlap, averaged over all time-bins in the network, is large at shortest timescales \((J(\Delta T = 5 \text{ min}) = 0.71)\) but drops rapidly as the size of the window increases (Fig. 1b). For example, \(J(\Delta T = 1 \text{ h}) = 0.51\), indicating a substantial turnover even at short timescales.

Within the 5-min time-bins, the network is comprised of disjoint cliques, each one corresponding to a gathering of individuals. The changes of the network during short time intervals can be attributed to people moving between gatherings, as proposed by Sekara et al. [13]. As a system, these are constantly evolving, with members changing associations, and gatherings dissolving and forming (Fig. 2a). These changes lead to the network connectivity that can be observed when aggregating the interactions into time-bins of longer duration (Fig. 2b), even though every single time slice still consists of disjoint cliques.

2.2 Temporal subsampling

To study the effect of temporal subsampling – reduction of dynamical information in the network – on the spreading processes, we consider two sampling schemes. These schemes are motivated by data collection strategies employed in the real studies, and therefore not necessarily an effort to devise the best possible temporal compression strategy.

In the first approach, which we call snapshot sampling, we choose a random 5-min bin from every N bins and consider this to represent the state of the entire network for these N bins (Fig. 3a). This way, we reduce the temporal resolution by a factor of N, but in a coherent fashion, because the network slices we use contain actual observed network states. The ‘snapshot sampling’ is typically the result of data collection methods which use static snapshots of the full population measured simultaneously. This is the case when physical proximity networks are inferred based on photographs or synchronized sensors, for example reports from a WiFi system (i.e. a list of devices connected to a router at the time of taking the snapshot, as in Ref. [15]), or results of Bluetooth scans performed by a fixed location device (as in Refs. [21,22]). We use random 5-min bin from every N bins rather than first bin (or last or middle) to remove the possible bias of choosing always the same part of the large time-bin (for example, always choosing beginning of the hour).
Fig. 2. Dynamics of physical proximity networks in a complex social system. (a) Nodes are colored based on the component they belong at (randomly chosen) $t_0$ (5-min time-bin). While preserving the colors we plot the network in $t_2, t_4, t_6, t_8$, corresponding to 10, 20, 30, 40 minutes later. Nodes that are not present at $t_0$ are marked in black. We can see how nodes move between gatherings. (b) The constant mixing of the nodes presented in (a) connects the initially separate components in to a well connected network when aggregating the interactions – even at relatively short timescales. Largest connected component in the network is highlighted.

In the second approach, which we refer to as link sampling, we sample the state (interacting/not interacting) of every link $(i,j)$ in the network from a random 5-min bin within the sampling interval (Fig. 3b). Thus, for every dyad, we choose a random 5-min bin (out the $N$ eligible bins) use the dyad’s state (on/off) as representative for this link in the subsampled network. This sampling strategy results in a network state which may have never existed at any given point in time, but which also contains a factor of $N$ less temporal information compared to the original network. The link sampling corresponds to sampling occurring in multiple places in the population in an asynchronous way, a situation which occurs when collecting data using mobile phones or sociometric badges.

The temporal subsampling in both modes reduces the information about dynamics and results in a lossy compressed version of the temporal network. The information about the exact dynamics is lost (Fig. 4a), replaced by static representations, with a width corresponding to the subsampling parameter $N$. In both subsampling scenarios the probability that a link $(i,j)$ is active is directly related to the number $n$ of 5-min bins in the $N$-bin interval in which the link is present and this probability is equal to $n/N$. This implies that the average number of interactions after subsampling is the same in both snapshot and link sampling. However, due to the high temporal variability presented in the inset to Figure 1a, we expect that the number of links in the snapshot sampling will have a higher variation. This is, in fact, the case, as we show in Figure 4b. As expected, snapshot sampling results in a larger variability in the total number of interactions, because snapshots with very high or very low number of links may end up being chosen; still, the variability is within ±20% from the number of interactions for the full network. Performing linear regression on the mean values of total number of interactions we test for slope different from 0 ($H_0 : b = 0$). The test statistic is $t = b/s_b$ on $(N - 2)$ degrees of freedom, and for both snapshot and link sampling we do not discover any significant trend in the average number of total interactions ($p = 0.60$ and $p = 0.63$, respectively). On average, when the number of interactions is considered, the subsampled networks are equal. In spite of the fact that the total amount of temporal information and average number of total interactions are equal, the structure of the snapshot and link networks is quite different. Keeping track of the size of the largest connected component (LCC) we notice that the coherent network sampling results in disconnected neighborhoods dominating the network (Fig. 4c). As expected, in link sampling, the network is more connected, with LCC containing up to 50% of the nodes in the network.

2.3 Spreading results

To quantify the effect of temporal subsampling on the modeling of a dynamic process unfolding on the network, we simulate spreading using a Susceptible-Infected-Recovered (SIR) model. In the spreading, we explore a variety of values for the transmission parameter $\beta$, including very slow and very fast transmissions (ranging between $\beta = 0.002$ and $\beta = 0.05$), and maintain a fixed recovery parameter $\mu = 4$ days. We randomly subsample the network 10 times for every value of subsampling parameter $N$ and run 100 simulations per condition, with a random starting time-bin and index patient. We apply
Fig. 3. Models for temporal subsampling. (a) Snapshot sampling: when reducing the sampling resolution $N$ times, for each set of $N$ consecutive 5-min bins we randomly select one and use it to represent these $N$ bins. (b) Link sampling: when reducing the sampling resolution $N$ times, for each dyad in the network during $N$ consecutive 5-min bins we choose its state (interacting or not interacting) in a random bin among the $N$ and use the state to represent the dyad’s status across the $N$ bins.

circular boundary condition to extend the network beyond one month. For our purposes, the spreading simulation is used to understand the impact of sampling on a dynamic processes in the network. We do not attempt to model any particular disease.

Temporal subsampling, both snapshot and link-based, results in decreased spreading. The spreading process is slower, with a smaller peak value, and reduced total outbreak size (Fig. 5a). This effect is more pronounced for rapid spreading (large $\beta$) and the effect is markedly stronger for snapshot sampling.

In Figure 5b, we quantify the effect of temporal subsampling on outbreak size. The drop of the outbreak size with the subsampling parameter $N$ is well explained by linear model (ordinary least squares regression), with a sub-linear effect for low values of $\beta$. Again, the effect is dramatically more pronounced for the snapshot subsampling. Similarly, probability of small outbreaks (reaching only a small fraction of the network) grows as a function of subsampling (Fig. 5d), with effects much more pronounced for the snapshot sampling. Finally, a reduction of temporal fidelity drastically increases the time it takes for the spreading process to reach 50% of the network (Fig. 5e).

2.4 Link-subsampling vs. full resolution

It is interesting to consider why spreading in the link-subsampled network is not faster than spreading in the full network: the number of connections in the link-subsampled network is typically higher than in single time slices of the full-resolution network. To understand why, we consider the structure of the link-subsampled network compared to the full-resolution network aggregated over the same time window. The difference arises from the fact that although the subsampling is performed so that the subsampled and full network have the same number of interactions $(i, j, t)$ in any given time window (Fig. 6a), the way these interactions are distributed on links $(i, j)$ is very different.

To help guide our thinking about the differences, consider the full-resolution network aggregated over a certain time window. Here, the distribution of the link weights is broad, with many weak links and a few very strong connections. By contrast, link-subsampling creates a network where all links have the same weight – because all links are active through the entire window (Fig. 6b). The full-resolution network has many more – but weaker – links...
Fig. 5. Results of SIR spreading process. (a) The shape of the spreading curve is less affected when each dyad is sampled independently, but network sampling leads to a significant underestimation of the outbreak size. The effect is less pronounced for slow spread (see the inset). Solid lines in (b)–(e) represent snapshot sampling, dashed lines represent link sampling. (b) With fast spreading epidemics, snapshot sampling results in a smaller expected size of the outbreak, but slow spreading epidemics are less affected. (c) Subsampling from 5-min bins down to 120-min bins (increasing the sampling interval $T_s$) does not significantly change the expected results of spreading simulations. (d) The probability of a non-outbreak (outbreak smaller than 20% of the network) grows with the temporal subsampling. (e) For fast spreading epidemics, the time needed to infect half of the population grows linearly with the increasing subsampling rate in the network sampling scenario, but stays relatively stable when we sample dyads. Difference for low $\beta$ is not statistically significant between snapshot and link sampling.

This has strong implications for the connectedness of the network. In the link-sampled network, the network is split into a number of separate components, and an infection is rarely able to infect the entire network within a single frame. This is not the case in the full network, which has an effectively much larger connected component active. This has strong implications for the connectedness of the network. In the link-sampled network, the network is split into a number of separate components, and an infection is rarely able to infect the entire network within a single frame. This is not the case in the full network, which has an effectively much larger connected component

Fig. 6. Link weight heterogeneity in the full-resolution network. (a) Links-subsampled and full-resolution network aggregated over the same time window have the same number of interactions ($i, j, t$). Here shown for $N = 72$ i.e. 6 h sampling. (b) The link weights in these views are distributed very differently, the full-resolution network features a long-tailed distribution with many weak links, whereas all links in the subsampled network have the same weight equal to sampling window size (72). (c) The higher number of links in the full-resolution network leads to a greater connectivity, here illustrated by the size of LCC. (d) The difference in the structure impacts the spreading. For fixed starting conditions (time-bin and seed node), it is possible to find regime of $\beta$ where the spreading on link-subsampled network is in fact faster (values above 1 on the plot). This is however not guaranteed for every starting condition and on average the spreading is slower in the link-subsampled network due to lower number of links.

2.5 Slow versus fast spreading

These dynamics are sensitive to the speed of the disease spread. When the disease spreading is slower (low $\beta$) than the changes in the network, the gradual building of the connectivity in the full-resolution network does not slow down the spreading: from the slow disease perspective the network looks well connected. In the case when the disease spread is high (high $\beta$), the lower number of links ($i, j$) in the link-subsampled network becomes the limiting factor. When the transmission parameter is large, the number of links, not the link weights, is important. The disease is unable to reach the full network, for example getting ‘stuck’ in a disconnected component. In both of these cases the full-resolution network facilities spreading, due to higher number of links.

These findings imply there may exist a third regime of $\beta$, where the transmission is faster than changes in the full network, ‘waiting’ for connectivity in the full-resolution network to build up, but slow enough that it
does not run out of links in the subsampled network (never fully filling up its network components). Such regimes can be found in the network for fixed starting conditions (start time and node). But these cases are rare, because each instance depends on an interplay between the structure of the network, size of the sampling window, and starting conditions. Thus, when averaged over many different starting conditions, the spreading is slower in the link-subsampled network due to the lack of the high number of weak links (Fig. 6d). In the following, we discuss findings for the averaged case.

As expected, the impact of losing the temporal fidelity is strongest for fast spreading processes. With the lack of information about the detailed network dynamics, the disjoint gatherings produced by snapshot sampling lead to containment of the disease, resulting in smaller and slower spreading. When the transmission parameter $\beta$ is high (fast spreading), the disease is more likely to infect all nodes in the available neighborhoods, with no possibility to propagate to new places. For slow processes, the loss of temporal fidelity is less significant, as the spreading takes more time to fill up the isolated gatherings. The containment effect is much smaller in the link subsampling, as the network is more connected due to different (non-coherent) configuration of the links.

3 Discussion

Above we have investigated how modeling of spreading processes is impacted by reducing the temporal fidelity of close proximity interaction networks. We found that the network are highly dynamic, even at short timescales. Within short time-bins, nodes gather in disjoint cliques, but with changing affiliation across time. These dynamics create significant interconnectedness when considering network at longer timescales (hours). When these short term dynamics are disregarded, either due to data collection process or data compression, spreading processes are strongly affected—as the temporal fidelity decreases, outbreaks become less frequent and smaller.

Interestingly, subsampling the network in a synchronized way (when the state of the entire network is sampled at once and repeated) has a much greater impact on the spreading results than when sampling is performed independently across links. This is because the disjoint gatherings that appear at shortest timescales inhibit the spreading process, when the minute-to-minute dynamics of nodes switching membership are lost. When we sample every link from an independently chosen time-slice the impact is much smaller, effectively approximating these short timescale dynamics.

The results presented here highlight a fundamental property of close proximity networks in social systems. We show how the dynamics contained within hourly time-bins can be instrumental for spreading process in the society. Simultaneously, from a methodological perspective, we illustrate how inclusion of these dynamics is crucial for understanding of the network of close proximity interactions and dynamical processes unfolding on them.

References

5. A. Vespignani, Nat. Phys. 8, 32 (2012)
7. J. Stehlé et al., BMC Medicine 9, 87 (2011)
Appendix G

Inferring Person-to-person Proximity Using WiFi Signals
ABSTRACT
Today's societies are enveloped in an ever-growing telecommunication infrastructure. This infrastructure offers important opportunities for sensing and recording a multitude of human behaviors. Human mobility patterns are a prominent example of such a behavior which has been studied based on cell phone towers, Bluetooth beacons, and WiFi networks as proxies for location. However, while mobility is an important aspect of human behavior, understanding complex social systems requires studying not only the movement of individuals, but also their interactions. Sensing social interactions on a large scale is a technical challenge and many commonly used approaches—including RFID badges or Bluetooth scanning—offer only limited scalability. Here we show that it is possible, in a scalable and robust way, to accurately infer person-to-person physical proximity from the lists of WiFi access points measured by smartphones carried by the two individuals. Based on a longitudinal dataset of approximately 800 participants with ground-truth interactions collected over a year, we show that our model performs better than the current state-of-the-art. Our results demonstrate both the value of WiFi signals in social sensing as well as potential threats to privacy that they imply.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

Keywords
mobility; wifi; face-to-face; proximity; interactions; social networks

1. INTRODUCTION
We are surrounded by an ever-increasing number of telecommunication infrastructures, such as mobile phone networks, WiFi access points, or Bluetooth beacons. In addition to their intended function of providing connectivity, these infrastructures offer an unprecedented opportunity for sensing, modeling, and subsequent analyzing of a wide range of human behaviors [23]. Here we show how our interactions with other people can be inferred in a reliable and scalable way, using signals from WiFi access points.

Being able to infer person-to-person proximity events with high spatio-temporal resolution enables modeling of phenomena such as spreading of diseases and information [18], formation of social ties [10], as well as group dynamics [39]. Commercial applications vary from distributed ad hoc networking [24] to romantic matchmaking [8].

Despite the importance of understanding networks of close proximity interactions, there is a scarcity of scalable and efficient ways to obtain data for large populations. This is due to the fact that technology has only recently developed to the point, where collection of such high resolution data has become technologically feasible. The data sources used for investigating mobility of individuals, such as call detail records (CDRs) from mobile operators [14], are too coarse in terms of temporal and spatial resolution to allow inference of person-to-person proximity. On the other hand, the current state-of-the-art methods for measurement of physical proximity require using specialized hardware (e.g. sociometric badges) [29, 34] or smartphones sensing each other through Bluetooth [9, 3, 45]. Specialized hardware adds cost and complexity to experimental deployments, effectively limiting their scale. Bluetooth scanning realized on participants' mobile phones increases power consumption [12]—limiting temporal resolution that can be achieved—and requires the devices to be in Bluetooth discoverable mode. This requirement raises privacy [48] and security concerns [37]. When a phone is in discoverable mode the location of its owner can be tracked by third parties, a fact commonly used both by researchers [22, 31], and advertisers [7]. Moreover, whenever a phone is discoverable, a malicious actor can attempt to pair to it in order to steal contact lists, content of messages, etc. For these reasons phone manufacturers make it difficult (or impossible) for a handset to remain discoverable indefinitely. iOS and Android 6.0+ devices disable discoverability whenever the user exits the Bluetooth settings screen. Older Android devices let the user set the discoverability timeout to, at maximum, five minutes. In our study we relied on the fact that in Android versions 4.1 - 6.0 it is still possible to set unlimited discoverability timeout programmatically, but this might change at any point in the future. Apart from the privacy and security issues of using Bluetooth for sensing, another shortcoming is that Bluetooth data lacks location context. When co-presence of individuals is inferred through devices sensing each other, an additional step is usually required to estimate the location of the meeting, for example by comparing Bluetooth scans with GPS measurements [39], by using fixed infrastructure of RFID transmitters [41], or Bluetooth beacons [22]. In the light of these problems, it is clear that alternative methods for tracking person-to-person interactions are needed. We explore the possibility of using WiFi for this purpose. Previous attempts at exploiting the
properties of WiFi signals (eg. [27, 21, 26, 20] further described in the related work section) for social sensing have been limited to controlled environments and lacked verification on longer timescales. We base our research on the intuitions developed in these previous experiments, we extend them, and test in a real-world dataset.

Present work. Here we study the problem of inferring physical proximity between pairs of individuals from a list of WiFi signals sensed by their phones. We use a longitudinal dataset containing WiFi and Bluetooth scan results from hundreds of participants, collected over a year as part of the Copenhagen Network Study [45]. Using Bluetooth as ground-truth for physical proximity, we train a model for comparing the results of WiFi scans from two devices to determine whether two individuals were in close physical proximity. We employ a number of interpretable metrics to compare the lists of visible WiFi access points, such as Jaccard similarity or correlation of received signal strengths. Apart from comparing the lists directly, we can derive context from just the number of routers seen in the lists: more populated areas tend to have more routers available. Furthermore, we exploit the characteristics of interaction dynamics, for example that people are more likely to meet during work hours, or on a Friday afternoon than on a Sunday night. Importantly, our algorithm for using WiFi signals to infer proximity does not rely on positioning the routers in physical space. Co-location is not inferred by thresholding the distance between the estimated location of two individuals. Instead, their WiFi environments are compared and then we estimate the similarity directly. As a final step, we are able to combine these insights using machine learning models to achieve the area under receiver operator curve (AUC ROC) scores of up to 0.89 in the proximity inference task. We show that our model works in a range of environments, does not depend on particular access points, and its performance does not deteriorate over time. Our experiments demonstrate that we are able to track close-proximity interactions over time and in different social and spatio-temporal contexts. Overall, our approach performs better than previously suggested solutions.

Contribution. We present a novel approach for tracking close-proximity person-to-person interactions based on existing infrastructure of WiFi networks and off-the-shelf consumer smartphones and compare its performance against existing methods.

2. EXPERIMENTAL DESIGN

2.1 The Copenhagen Networks Study

The dataset used in this work was collected as part of the Copenhagen Networks Study [45]. It covers mobility and interaction records of approximately 1000 students at Technical University of Denmark, over a two year period. Each student was equipped with a LGE Nexus 4 Android smartphone as a data collecting device. On each phone, an application based on the Funf Open Sensing framework [3] gathered readings from multiple sensors including:

- Bluetooth scans (every 5 minutes): each scan contains a list of discoverable devices, their unique identifiers, user defined names, and received signal strength (RSSI). Because we know which anonymized participant identifier corresponds to which Bluetooth unique identifier, we can monitor proximity between the participants.

- WiFi scans (every 5 minutes): each scan contains a list of WiFi access points (both traditional routers and mobile hotspots), their unique identifiers (BSSID or MAC addresses), network names they transmit (SSIDs), and RSSI.

The collector app additionally collected the data requested by other applications on the phone. Therefore, the temporal resolution of the data for some of the users can be even higher than one sample every 5 minutes [36].

All data in the Copenhagen Networks Study was collected with the participants’ informed consent, with an emphasis on ensuring awareness of the complexity and sensitivity of the collected data [43]. The study setup, including security, privacy, and informed consent has been approved by Danish Data Protection Agency. Further details of the study can be found in Ref. [45].

3. INFERRINGCLOSE-PROXIMITY INTER-AC TIONS

3.1 Problem statement

In brief, our task is to compare the lists of WiFi routers seen by users A and B approximately at the same time (with at most $\Delta t = 300$ seconds difference) and determine whether the two users were in close physical proximity. We use Bluetooth data as ground truth for physical proximity to train and verify our models.

3.2 Data preparation

WiFi. We found that in our dataset there are multiple WiFi routers that share the same MAC address, a phenomenon which might confound our task. We use a simple heuristic to remove these “ambiguous” routers since finding the optimal way of identifying them would warrant a publication on its own. Here we rely on the network name they broadcast. Because the routers at the DTU campus broadcast up to four network names (SSIDs) per MAC address, we remove the scans of routers which broadcast five or more network names throughout the observation. We found 3950 offending MAC addresses, which corresponds to only 0.04% of all unique MAC addresses in the data. However, scans of these routers constitute 1.4% of all scan results.

<table>
<thead>
<tr>
<th></th>
<th>training</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>total observations</td>
<td>5M</td>
<td>115.5M</td>
</tr>
<tr>
<td>% positive</td>
<td>31%</td>
<td>31%</td>
</tr>
<tr>
<td>unique users</td>
<td>812</td>
<td>820</td>
</tr>
<tr>
<td>median number of access points per observation</td>
<td>7.0</td>
<td>7.0</td>
</tr>
<tr>
<td>mean number of access points per observation</td>
<td>11.3</td>
<td>11.3</td>
</tr>
</tbody>
</table>

Table 1: Summary statistics of the dataset used to infer proximity events.
Next, we identify one home router for each participant per month. We employ the following heuristic for each participant:

1. Bin the time information of WiFi scan history. The size of the bin does not influence the results significantly, here we use 10 minutes.
2. Sort the list of routers by the number of timebins they appear in.
3. The router that appears in the biggest number of timebins is assumed to be the home router.

The details of the procedure are described in Ref. [36].

Bluetooth. Due to the imperfect firmware and software running on the phones, Bluetooth data is not always available—not all users are scanning and discoverable at all times. This can introduce a situation in which two persons are proximate, but Bluetooth does not capture that event. We divide the dataset into one hour subsets and select on only the WiFi and Bluetooth data from people who were seen and who saw at least one other person through Bluetooth. This strict approach makes the task more difficult, as it removes long periods where individuals are alone, for example night-time samples of students who do not live with other participants.

To train our model we also need to provide negative examples. For dyads in this category we choose potential interactions between two people who did not see each other on Bluetooth, but whose lists of scan results share at least one overlapping router. Compared to selecting negative samples by randomly sampling dyads this definition brings the task closer to a real-life scenario of discovering very close physical proximity (up to approximately 10 meters).

3.3 Dataset statistics

Table 1 shows the details about the dataset. Through a year of data we found 116M potential interactions. We randomly select .5M of them to train the models.

We also observe that in 99% of cases of Bluetooth sightings the corresponding WiFi scans overlap by at least one access point. This indicates that there is a potential in using WiFi scan results to infer the co-presence with high recall. Conversely, in more than 31% of cases where there is at least one overlapping access point, the two devices are also close according to Bluetooth. This indicates that WiFi signals can be applied to the task resulting in a high precision solution. The majority (53%) of meetings happen during working hours (from 8am to 7pm) on campus.

3.4 Methods of comparison

We use a number of metrics to compare two lists of WiFi scan results and use these metrics as features in a supervised machine learning approach. We divide the features into the following categories: availability of access points, received signal strength, presence + RSSI, timing, and popularity.

Table 2 lists the features we apply, and Figure 1 shows how the probability of an interaction changes as a function of each feature’s value. In this section we describe each feature in detail. Citations refer to the first articles using the features for the purpose of face to face contact detection.

### Availability of access points (AP presence)

First, we compare the list of routers seen by the two phones, regardless of their received signal strength. We introduce the following measures:

- **overlap**: the raw count of overlapping routers [21];
- **union**: size of the union of the two lists;
- **jaccard**: ratio between the size of the intersection and the size of the union of the two lists [20];
- **non-overlap**: the raw count of non-overlapping routers (size of union minus size of overlap) [21];
- **spearman**, **pearson**, **manhattan**, **euclidean**.

### Received Signal Strength Indicator (RSSI)

Next, we focus on comparing the received signal strength of the overlapping routers. While received signal strength (RSSI) is not generally a reliable proxy for distance [35], two colocated people can be expected to have similar RSSI readings for the overlapping routers. We investigate the spearman and pearson correlation coefficients of received signal strengths of the overlapping routers.

<table>
<thead>
<tr>
<th>category</th>
<th>features</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP presence</td>
<td>overlap, union, jaccard</td>
</tr>
<tr>
<td>RSSI</td>
<td>spearman, pearson, manhattan, euclidean</td>
</tr>
<tr>
<td>AP presence +</td>
<td>top AP, top AP +/- 6dB</td>
</tr>
<tr>
<td>RSSI</td>
<td></td>
</tr>
<tr>
<td>timing</td>
<td>hour of week</td>
</tr>
<tr>
<td>popularity</td>
<td>min popularity, max popularity, Adamic-Adar</td>
</tr>
</tbody>
</table>

Table 2: Features used to infer close-proximity interactions.
Figure 1: The larger the number of common routers two phones see, the higher the probability of close proximity. At the same time, the more routers they see in total, the lower the probability of an interaction - densely populated areas have more routers and more people who are not necessarily interacting. Jaccard similarity allows us to recognize interactions regardless of the number of visible access points.

Furthermore, we also calculate the difference between RSSI of overlapping routers by measuring the $\ell_1$ and $\ell_2$ distances and dividing the results by the number of overlapping routers. For simplicity we call these features manhattan and euclidean and define them in Equations 1 [27] and 2 [20] respectively.

$$m = \frac{1}{N} \sum_{i} |RSSI_{A,i} - RSSI_{B,i}|$$

$$e = \sqrt{\frac{1}{N} \sum_{i} (RSSI_{A,i} - RSSI_{B,i})^2}$$

where $RSSI_{A,i}$ is the received signal strength or access point $i$ as measured by user A and $N$ is the total number of overlapping routers. Figure 1c shows that with growing distance, the probability of an interaction falls.

AP presence + RSSI. It has been previously shown that a good heuristic for determining whether a user is in the same location during two measurements is to verify whether they measure a common strongest router [13]. Here, we verify whether this approach can be used for inferring co-location: if two users measure the same router as the strongest one, we assume they are in close proximity. We investigate the strict case, top AP. Additionally, we allow for some variability in the measured strength: feature $\text{top AP} + / - 6 \text{dB}$ assumes a positive value if there is at least one overlapping access point in the lists of routers of A and B within 6 dB from the top router.

Popularity. Additionally, we inspect how many different participants of the study scanned the overlapping routers within five minutes of the meeting - intuitively if only a few persons were in a given location they were more likely to be there together, rather than by chance. We find the least and the most popular among the overlapping routers and report $\min\text{popularity}$ and $\max\text{popularity}$. As we show in Figure 1f., this intuition is not entirely confirmed by the data. The correlation between the number of individuals present and the probability that any two of them are interacting is low (Spearman’s $\rho = 0.15$, $p_{val} < 0.001$). Note that popularity and the size of union are correlated (Spearman’s $\rho = 0.48$, $p_{val} < 0.001$) - more routers are located in popular places, so the more routers there are around, the more people see each of them. However, to achieve a good estimation of popularity, we need data from the entire population, while the number of routers around can be obtained just from data of just the two individuals. Additionally, we use a score inspired by a measure introduced by Adamic and Adar [2], defined as:

$$score(u_1, u_2) = \sum_{i} \frac{1}{\log(\text{popularity}(\text{AP}_i))}$$

Here, each overlapping router is weighted more the fewer people scanned it. In this case, the higher the value, the higher the probability of a meeting between two people.

Timing. In contrast to the other features we described, timing does not rely on comparing the list of scan results. Instead, we use the timestamp of each potential face to face meeting to exploit the temporal characteristics of human interactions. As a reminder, we only consider a potential interaction if both parties have WiFi scans within 300 seconds from one another. For simplicity, we assume here that the timestamp of the potential interaction is the lower of the two scan timestamps. We notice that the prior probability of two people being proximate depends on the time of day and the day of week, as shown in Figure 1i-k. While there is only a small variability between the days of the week (Figure 1j.), the probability of the interaction during a day (Figure 1i.) appears to be driven both by the class.
schedule—the probability is the highest during classes, and drops during lunchtime—and by after-school social activities. Only by combining the two factors (Figure 1k.), we get the full picture: the probability of interactions from Monday to Tuesday is driven by the school schedule; Friday is a mixture of scheduled and social interactions, with the probability remaining high far into the night hours; Saturday is characterized by interactions starting in the late afternoon and into the night; and on Sunday our participants interact mostly during daytime, with no visible lunch breaks. We add a feature to capture these patterns: **hour of week**: from 0 to 167.

### 3.5 Imputing missing values

Two of our features are Pearson and Spearman correlations. There are two cases in which it is not possible to calculate the correlation: (1) if there are fewer than three routers available for comparison, (2) if at least one person reads all the signal strengths at the same level. In such cases we assume a NaN (not-a-number) value of $\rho$ to be imputed later on. Additionally, we assume a NaN value of $\rho$ if the correlation is not significant with the $\rho_{min} < 0.05$. This results in multiple missing values for the two features. The simplest approach is to skip such observations, but that would imply not training the model in cases with few routers available. We therefore impute the values by assigning the mean value of $\rho$ to be imputed to training, which means that the models and thresholds are not overfitted to the training data. We verify in our data that other approaches, such as using the median value of the feature or using $k$ nearest neighbors to impute the missing value [40], do not improve the consecutive predictive performance.

### 4. RESULTS

#### 4.1 Performance of single features

We first show how well one can infer close-proximity interactions by simply thresholding a single feature. By manipulating the threshold we achieve the Receiver Operating Curve. We report the area under this curve (AU ROC) as the first metric of performance in Table 3. Then, we select the threshold at which the $F_1$ score (the harmonic mean between precision and recall) is maximized in the training set. We also report the AU ROC for the test data (111.5 million previously unseen samples) along with the $F_1$ score at the threshold optimal for the training set.

The results are presented in Table 3. We find that the single best performing feature is Jaccard similarity between the two lists of routers. As expected, thresholding on time information is not meaningful (it is equivalent to assuming that all interactions after a certain hour of a certain day of week are close proximity interactions). It is important to note that the performance in test does not drop compared to training, which means that the models and thresholds are not overfitted to the training data.

#### 4.2 Performance of feature sets

We train a Gradient Boosting Classifier for each category of features and present the results in Table 3. The parameters of the classifier are tuned each time through a grid search or parameter space with 5-fold cross validation. Furthermore, we compare the model based on the features proposed by Krumm et al. [21] to models based on richer sets of features, see Table 4. In the original work, Krum et al. did not find any performance improvements of using a combined model over using single features. Here, we show that combining the features they proposed does improve the performance. Our Simple model is based on features that do not require long term data collection and are not specific to our deployment. It performs better than any single feature or group of features, and it outperforms the model based on the features introduced by Krumm. Enhancing the model with the information on popularity (the General model) further improves the performance. Finally, using all features, including timing and location (which might be specific to this experiment as they depend on our campus as location and the time schedule typical for students), does not improve the performance of the classifier.

#### 4.3 WiFi similarity and physical proximity

Here, we verify whether there is a correlation between how close people are in physical space (approximated by the received Bluetooth signal strength measured on their phones) and the probability that our models misclassify the sample as “non-interaction”. As we show in Figure 2, the shorter the distance over which an interaction happens (high Bluetooth RSSI), the lower the probability of missing that interaction. This shows that the similarity measure between WiFi lists introduced by our models has a physical interpretation: a more similar WiFi environment indicates proximity in a more granular way than just the Bluetooth 10 meter

<table>
<thead>
<tr>
<th>Category</th>
<th>AU ROC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP presence</td>
<td>overlap</td>
<td>0.77</td>
</tr>
<tr>
<td>association</td>
<td>jaccard</td>
<td>0.84</td>
</tr>
<tr>
<td>union</td>
<td>0.53</td>
<td>0.48</td>
</tr>
<tr>
<td>non-overlap</td>
<td>0.74</td>
<td>0.58</td>
</tr>
<tr>
<td>combined</td>
<td>0.85</td>
<td>0.69</td>
</tr>
<tr>
<td>RSSI</td>
<td>spearman</td>
<td>0.70</td>
</tr>
<tr>
<td>pearson</td>
<td>0.71</td>
<td>0.59</td>
</tr>
<tr>
<td>manhattan</td>
<td>0.60</td>
<td>0.51</td>
</tr>
<tr>
<td>euclidean</td>
<td>0.59</td>
<td>0.51</td>
</tr>
<tr>
<td>combined</td>
<td>0.78</td>
<td>0.62</td>
</tr>
<tr>
<td>Location</td>
<td>at DTU</td>
<td>0.61</td>
</tr>
<tr>
<td>at home</td>
<td>0.64</td>
<td>0.55</td>
</tr>
<tr>
<td>combined</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>Timing</td>
<td>time of week</td>
<td>0.51</td>
</tr>
<tr>
<td>Popularity</td>
<td>min</td>
<td>0.54</td>
</tr>
<tr>
<td>max</td>
<td>0.59</td>
<td>0.49</td>
</tr>
<tr>
<td>adamic adar</td>
<td>0.77</td>
<td>0.62</td>
</tr>
<tr>
<td>combined</td>
<td>0.79</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 3: Performance of single features and feature categories in the task of inferring close proximity interactions. Jaccard similarity between lists of routers seen by the two devices is the best performing single feature. The performance of the feature categories is given for a GradientBoostingClassifier trained on the corresponding category. $F_1$ are given for a threshold that maximizes $F_1$ in the training set.
NearMe: overlap, non-overlap, spearman, euclidean

Simple: AP presence, RSSI, Presence + RSSI

General: all but at DTU and Timing

Full: all features

<table>
<thead>
<tr>
<th>featureset</th>
<th>AU ROC train</th>
<th>train test</th>
<th>F1 train</th>
<th>F1 test</th>
</tr>
</thead>
<tbody>
<tr>
<td>NearMe</td>
<td>0.87</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Simple</td>
<td>0.88</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>General</td>
<td>0.89</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>Full</td>
<td>0.89</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 4: Performance of feature sets in the task of inferring close proximity interactions. We train a Gradient Boosted Classifier on selected subsets of features: NearMe [21], Simple (no features that are specific to this experiment or require longer term data collection), General (without features that could be specific to this experiment), and Full (all listed features). All models based on the feature combinations we propose perform better than NearMe. Using features which could be specific to the experiment does not improve performance further.

4.4 Training period and performance in test

Figure 3 shows how the number of samples used for training influences the performance of the full model in test. We compare the performance of a random forest classifier and a gradient boosted classifier and find that the latter has a slightly higher performance for training sets larger than 1000 samples. On the other hand training of the random forest classifier can paralleled thus making the process faster, which works to the advantage of this classifier.

4.5 Importance of features

In addition to reporting the performance of single features, we also show how important each feature is for the machine learning model. In the implementation we use [22] the feature importance is defined as the total decrease in node impurity weighted by the probability of reaching that node, averaged over all trees of the ensemble [1]. Figure 4 shows the accumulated results from 30 training rounds of the gradient boosted classifier on randomly selected subsets of the training data, each with 100 000 samples. We find that Jaccard similarity is the most important, followed by the overlap among the strongest routers, Pearson’s correlation of signal strenghts, and Adamic-Adar (which exploits the overlap and the popularity of routers).

4.6 Validity of the model in different scenarios

Figure 5 shows the performance of the gradient boosting classifier in different contexts and across time.

Number of routers. As described before, the number of routers in an environment is positively correlated with the population density. We divide the test data in three equally-sized subsets, depending on the size of the union of routers seen by two people. Figure 5a. shows that the performance of the model is best in the low and mid sets (AUC > 0.9) and observably lower (AUC ≈ 0.85) for environments with the highest number of routers. Thus, we show that the model performs well in typical environments.

Location. Because our the data was collected by students of one university, with the majority of interactions happening on campus, there is a risk that the model would overfit towards such situation. This is, in fact not the case. Figure 5b. shows that while the performance of the model is high on campus, it becomes even better for the meetings outside.

Timing. As shown in Figure 5c. the performance of the model does not drop significantly during special periods, such as Christmas of summer vacation (grey areas in the plot correspond to periods with no university classes). Instead, it remains stable throughout the experiment.
Figure 4: Gradient Boosted Classifier reports the relative importance of each feature (the decrease in node impurity it provides). After 30 training rounds we see that Jaccard is the most important feature, followed by overlap among the strongest routers (top AP±6), Adamic-Adar, and Pearson correlation between the signal strengths.

The performance does vary with the time of week, as shown in Figure 5d-f. When we compare it to Figure 1k, we see that the model performs better in situations where the prior probability of meeting is lower (for example during week nights). Nevertheless, it retains high performance of $AUC \geq 0.8$ throughout the week.

5. RELATED WORK

In this section we discuss related work that explores the application of mobile data to deepen our understanding of aspects relevant to this paper.

Location and mobility. CDR data has been used as a proxy for human mobility at large, societal scale. It has been shown that our movements are regular [14], stable [25], and predictable [40]. It yet remains to be verified whether these findings hold fully if the analysis were to be performed on data with higher spatial and temporal resolution (such as WiFi data). At smaller scales, the scientific community investigated the potential of WiFi routers in applications of indoor [4, 15, 33] and outdoor [6, 28, 11, 17] localization. Our recent work investigates how large companies can crowd source the creation of databases with router locations [35, 28, 11] and how people’s mobility on societal scale can be described using only a small subset of available routers [36]. WiFi signals can also be analyzed to discover places of interest and stop locations in an unsupervised manner, i.e. without explicit location information as reference [19, 47].

It is important to stress that the work presented in this article does not rely on location estimation (in terms of geographical coordinates) but instead on relative comparison between the environments sensed by two parties.

Figure 5: Our model for detecting person-to-person proximity events performs well regardless of the number of available routers (a) and location (b). Its performance does not drop during holidays (marked with grey areas in (c)). The situation in which the performance is the worst is the Friday evenings and nights (f), but even then, the AUC ROC is high.

Interactions. Complementary to mobility, the question of social interactions has been recently considered in various contexts, with the results indicating that collection of high-resolution behavioral traces is instrumental for understanding of complex processes in society [9, 39, 44, 42]. However, from a technical point of view, collection of such data remains a challenge.

The most popular methods for quantitative and scalable collection of close-proximity interactions include using specialized hardware (e.g. sociometric badges) [29, 34] or Bluetooth enabled smartphones [9, 3, 45]. In case of badges, interactions are usually inferred using radio-frequency identification (RFID) transmissions or infrared. This way, badges worn around participants’ necks can usually sense not just proximity but also whether individuals are facing each other, resulting in recordings of face-to-face interactions. Sens-
ing performed using Bluetooth-enabled mobile phones is less granular. The proximity can be detected in a binary fashion or further refined using the received signal strength as a proxy for distance [38]. However, the orientation of the individuals cannot be sensed. The subjects’ devices must remain in Bluetooth-discoverable state, which raises a number of security and privacy concerns, as described in the Introduction. There has been some developments in substituting Bluetooth with WiFi, an approach in which one of the phones acts as a hotspot and is sensed by others [5]. In controlled test environments this approach appears to offer a distance estimation resolution of 0.5m [30], providing a better understanding of the nature of the contacts [16]. However, the claim has not been tested in the wild and the method potentially introduces even more privacy and security problems than Bluetooth.

An alternative way of sensing interactions between two persons with smartphones relies on comparing the two devices’ radio frequency perceptions of the environment. If a similarity is above a certain threshold, the two devices are assumed to be in physical proximity. The idea of comparing WiFi signals to measure proximity was initially explored more than a decade ago. Initially, researchers relied on single-feature measures of similarity, such as Manhattan distance [27] or overlap [26]. NearMe project [21] introduced more features, such as rank correlation between the lists of overlapping routers sorted by signal strength, Euclidean distance, and the number of non-overlapping APs. The authors explored combining the features into a regression model, but this approach did not outperform single features. Moreover, their model would overfit for the rooms where it was trained and thus underperform in previously unseen environments. Kjærgaard and Nurmi also the differences in environments where the sensing takes place among the most important obstacles in using WiFi for social sensing [20].

We note that the differences in environments can actually be used to increase the performance of the model. We can exploit the characteristics of human interactions: from a technical standpoint, environments with a smaller number of routers offer lower accuracy of distance estimation; however, two people in an environment with fewer access points are more likely to be actually interacting (see Figure 1).

6. DISCUSSION

6.1 Privacy implications

There are two main privacy implications of this work. Firstly, the ability to track face-to-face interactions using WiFi can help us move away from relying on Bluetooth. By not requiring the participants’ phones to remain Bluetooth discoverable we protect the privacy and security of the subjects. While currently most phones advertise their presence and identity by scanning for WiFi, this problem is being addressed. Both Android and iOS randomize the MAC address of the device every time it sends WiFi probe requests making it more difficult to identify the user.1

Secondly, our results indicate a potential erosion of privacy of Android users. As we have previously shown, WiFi can be efficiently used for high-resolution mobility tracking of entire populations [35, 36, 47]. Here we go a step further and infer who people interact with, not only where they are. Thus, results of WiFi scans—collected by major manufacturers of mobile devices and available to majority of mobile application developers—constitute very sensitive datasets. For example, a vast majority of the applications available in Google Play Store has access to WiFi information, including all the scan results requested by the system as often as every 15 seconds [36]. This problem is addressed since Android 6.0—in the latest versions of the system an application has to hold a location permission to listen to WiFi scan results. However, the vast majority of handsets currently in use will not receive these crucial updates. Thus, WiFi signals remain a major privacy risk for years to come.

6.2 Limitations of the WiFi-based social inference

While our approach to inference of social interactions and ties using WiFi scans offers an important new method in computational social science, we want to recognize its limitations. WiFi-based inference of physical proximity requires data from both individuals for matching. As a consequence, datasets containing all the sensed access points for the individuals can be significantly larger when compared to physical interactions data collected using Bluetooth- or RFID-based approaches.

The inference in the approach presented here depends on the WiFi routers being present in the environment. While today WiFi networks are omnipresent, especially in densely-populated areas [36], we find that in our longitudinal and diverse dataset approximately 5% of the WiFi scans did not report any nearby networks, preventing inference of physical proximity.

In this study, all phones collecting data were of the same make and model. When considering a broader application of the method, differences in WiFi hardware transmitters and firmware and software of the phones may result in less consistent scan data, making it more difficult to devise a robust model as the one presented here.

Furthermore, due to the lack of ground truth data, we cannot prove that our model accurately estimates the distance between users. We show, that our model is more likely to recognize interactions with a higher Bluetooth RSSI, but this property does not trivially translate to distance estimation.

Finally, we should note that it is not our argument that the values of all model features for discovering particular interactions and reconstructing the overall social network are generally applicable to different populations. Depending on the specific population and social context under consideration, the weights in the model might be different or even entirely new features might be useful. Our results indicate however that physical proximity can be inferred in a feasible fashion using WiFi signals collected by smartphones, even in very densely-connected populations.

6.3 Conclusion

In this work we showed how WiFi scan results can reveal a great deal about our daily interactions with others and our social ties. By using behavioral traces, placed in context through meta information and our basic understanding of the inner working of social systems, we can transform a noisy data source to a strong social signal. Our findings have
important privacy implications, especially given our previous work which shows that it is possible to use WiFi signals for tracking human mobility. On the other hand, WiFi scans also constitute a great opportunity for companies with access to such data on a global scale, to contribute e.g. better epidemic models built on proximity data of billions of people. Finally, we hope that this method of social sensing will substitute Bluetooth sensing in future Computational Social Science deployments.

Acknowledgements

The authors would like to thank Andrea Cuttone for useful discussions as well as Urvashi Khandelwal and Jana Huisman for the important feedback. In this work we used the implementations of machine learning models from the scikit-learn [32] Python package.

7. REFERENCES


[28] B. Meyerson. AOL introduces location plug-in for instant messaging so users can see where buddies are. http://goo.gl/2W1uYh.


Appendix H

Offline Behaviors of Online Friends
Offline Behaviors of Online Friends

Abstract
In this work we analyze traces of mobility and co-location among a group of nearly 1,000 closely interacting individuals. We reconstruct the Facebook friendship graph, Facebook interaction network, as well as call and SMS networks from longitudinal records of face-to-face offline proximity. We find subtle, yet observable behavioral differences between pairs of people who communicate using the different channels and show that the signal of friendship is strong enough to stand out from the noise of serendipitous offline interactions between familiar strangers. Our study also offers an overview of methods for link inference based on offline behavior and provides new features to improve the performance in the prediction task.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

1. INTRODUCTION
The social bonds we form have a lasting impact on our lives and constitute a fundamental building block of our societies. They facilitate access to resources [16], dissemination of opinions and innovation [5], as well as the spread of habits [7, 15]. As described by Krackhardt et al. [18], one of the key factors in human bond formation is propinquity, physical and psychological proximity between people. However, until recently, scientists lacked tools to monitor person-to-person proximity over extended periods and at large scales. Recent technological advancements, such as mobile networks, the Internet, and online social networks have created a new opportunity for studying human behavior. Massive behavioral datasets are collected as part of operations of service providers and in research projects. It is thus increasingly feasible to study mobility, interactions, and habits in large populations, over long periods of time, and across multiple channels [20]. Data is now recorded in a more seamless fashion and with a significantly reduced involvement on behalf of the researchers. This approach potentially lowers the impact of the observation on the participants' behavior, compared to traditional methods. Rather than depending on the self-reported proximity among participants [4], we can now sense the interactions using sociometric badges [36], software running on mobile phones utilizing Bluetooth [11, 13, 31], fixed-location sensors [19], mobile network towers [3] and Call Detail Records (CDRs) [10], mining check-ins [9, 28], or geo-tagged online content [8].

The data set used for this study, is one year of person-to-person proximity records among nearly 1,000 students. We use a novel approach, which leverages the ubiquitous infrastructure of WiFi routers [26]. Then, we infer friendships among this population. We use a number of communication channels as ground-truth proxies for social ties, in lieu of self-reported relationships. Given a history of a dyad's person-to-person interactions, we are able to determine whether they are connected on Facebook or call each other, or whether they are just a pair of interacting familiar strangers [22]. We find that the size of the group people meet in is an important factor in accurately inferring social ties, with friends spending more time in smaller groups. We also show that interactions between friends are less likely to follow a particular schedule, compared to interactions between non-friends. The number of people our subjects interact with in the real world is orders of magnitude larger than the number of their online contacts. The participants of our study attend lectures, perform group work, and socialize during events organized by the university. Nevertheless, the signal of friendship is strong enough not to be lost in the noise of proximity interactions with strangers.

There are three central findings in this work. First, we verify that the proxies of friendship used in computational social science (Facebook interactions, calls, and short messages) are reflected in the offline behaviors and are positively correlated with the intensity of person-to-person contact. Second, we show that there is a surprisingly low overlap between dyads who communicate on Facebook and those who call each other. Offline behaviors can indicate which means of communication are adopted by each dyad. Finally, we compare the performance of a number of behavioral traits in discerning the actual social ties from those imposed by the structure of the studies. The insights can be applied to aid research and empower social applications, but also raise important questions regarding privacy of millions of smartphone users.

2. EXPERIMENTAL DESIGN

2.1 The Copenhagen Networks Study
The dataset used in this work was collected as part of the Copenhagen Networks Study [31]. In this study, we tracked the lives of approximately 1,000 student volunteers at the Technical University of Denmark. We used a custom app running on smartphones which distributed among the participants. The population is densely connected, and includes the majority of freshmen. The data was collected at all times over two full years (2013–2015) and is thus not limited to the university context. It contains a variety of interactions, both work-related and recreational [29].

Each participant of the study received an Android smartphone (LG Nexus 4) and installed the data collection app. The collector software was based on Funf Open Sensing framework [1] and ran in the background, collecting a variety of sensor readings. The data, compressed and encrypted, was periodically uploaded to a server located at the university. The data was collected with high temporal resolution and includes for example location estimations, records of phone calls and short messages, and person-to-person proximity events. In addition to data collected directly from the participants’ phones, we also collected snapshots of participants’ Facebook data every 24 hours (server-side collection), including lists of their friends, as well as likes, tags, and posts.

All data in the Copenhagen Networks Study was collected with the participants’ informed consent, with an emphasis on ensuring awareness of the complexity and sensitivity of the collected data [30]. The study setup, including security, privacy, and informed consent has been approved by the Danish Data Protection Agency. Further details of the study can be found in Ref. [31].

The high-resolution data collected in the Copenhagen Networks study offers an opportunity for an unprecedented insight into the dynamics of a complex social system seen across multiple channels. The location data allows for high-resolution fixed rate tracking of the participants’ mobility, both inside the buildings and outdoors. Calls, texts, and Facebook interactions reveal the social structure of the population. Communication on these channels is rarely incidental, thus considering the number of interactions on these channels provides a proxy for quantifying the strength of social ties.

We note that discovering the internal social structure of a densely-connected population—such as the population considered here—based on physical proximity is a non-trivial task. Non-acquainted participants in the study meet on a regular basis during lectures, in cafeterias, or at the gym. Here, we attempt to extract social signal above and beyond the already-dense hairball of interactions [29] to reconstruct the entire social network of the participants.

2.2 Data availability

Figure 1 summarizes the properties of the dataset by showing the number of users (a), dyads (b), and the density of all five networks (c): face to face, Facebook friendship graph, Facebook interaction graph, SMS, and call. Note in Figure 1a that behavioral (WiFi) data is available for a significantly larger number of users than the ground truth data. Facebook authorisations expire after three months and not all users renewed them. Therefore, the number of users with Facebook data declines over time. The relatively low number of users with calls and short messages is not caused by the users failing to collect this data: it shows that these forms of communication are less popular among the experiment participants. Figure 1b emphasizes that the sets of interactions in different communication channels are highly imbalanced: there are an order of magnitude more links in the Facebook friendship network compared to call/sms and the Facebook interaction network, and an order of magnitude more interactions face to face compared to Facebook friend links. As shown in Figure 1c, contacts through communication networks are not driven by the academic schedule as much as face to face interactions: the fraction of active dyads does not decrease during the summer time, as opposed to the face to face network. Finally, Figure 2 shows that there is a surprisingly small overlap between the dyads who interact on Facebook and those who call or message each other. Both telecommunication exchanges and Facebook comments have been used as proxies for friendship in scientific literature. This low overlap indicates that relying on only one of the channels might give a distorted picture of the actual social network.

2.3 Proximity inference

In this work, we rely on proximity events inferred from WiFi data. The details of the inference procedure can be found in Ref. [26]. The findings presented here do not rely on using WiFi as a proxy for person-to-person proximity. However, when using alternative methods for social sensing (such as Bluetooth or RFID badges), an additional step is necessary to determine the location of each interaction.

3. METHODS

The task is to infer social ties in a group of people, given their history of interactions in physical space. In lieu of self-reported relationships, we infer four proxies of social ties: Facebook friendships, Facebook interactions (comments on each other’s content), phone calls, and text messages. For simplicity we treat all networks as non-directional and assume that the links are reciprocated (although we acknowledge that in real-world networks the perceived friendships are not always reciprocal [2]). To accomplish the task, we take the following steps:

1. We contextualize each meeting by describing its social makeup, timing, and location.
2. We create a set of features that describe and summarize the properties of the person-to-person interactions of each dyad among participants.
3. We train supervised machine learning models on a subset of dyads and infer the links among the remainder of population. We do so separately for each of the four types of links.

We infer the links for each month from the interactions during the same month. Additionally, we answer the following questions: (1) does knowing the history of offline interactions for longer than a month increase performance of the inference? (2) can we infer (using the history of offline interactions) which communication tool a dyad uses, knowing that they do communicate either through Facebook comments or calls (but not both)?
3.1 Features

For each dyad \( A, B \) we define 16 features grouped into the following categories (see Table 1 for an overview):

Network similarity. It is commonly assumed that social
**category** | **features**
--- | ---
**Time spent together** | total time together, on campus, outside of campus, at home, weighted by the number of people, weighted by the number of APs
**Regularity** | entropy of hour of the day, entropy of the day of the week, entropy of the hour of the week, mean time between meetings, median time between meetings, entropy of locations
**Network similarity** | overlap among top 5 contacts, overlap among top 15 contacts, overlap among top 25 contacts, overlap among top 50 contacts

Table 1: Features used to infer social ties. Each feature has three variants: total, in-role (considering only interactions during working hours on weekdays), and extra-role (considering only interactions outside of working hours and on weekends).

relations often are transitive: if A is friends with C and C is friends with B, then A is likely to be friends with B. We measure the Jaccard similarity of top contacts between A and B, assuming that the more similar their top contacts are, the more likely A and B are to be friends. We use the values of overlap among top 5 contacts, overlap among top 15 contacts, overlap among top 25 contacts, and overlap among top 50 contacts (we note that extending the search beyond 50 top contacts does not increase the performance of the models). The neighborhood similarity has been previously exploited in the problem of link prediction for example in [9].

Each of the 16 features has three variants: total, in-role (considering only interactions during working hours on weekdays), and extra-role (considering only interactions outside of working hours and on weekends). Previous research showed that the distinction is crucial and that extra-role interactions are more indicative of friendship [12]. We train a random forest classifier for each month of the data using the four networks (Facebook friendship, Facebook interactions, call, and text messages) as ground truth.

### 3.2 Inference performance

We quantify the performance of our classifiers using two popular measures: Area Under Receiver Operator Curve (AUROC) and Matthew’s Correlation Coefficient (MCC). AUROC can be interpreted as follows: given a dyad of friends and a dyad of non-friends, what fraction of times does the classifier rank the friends higher? This metric provides an idea of how different a typical dyad of friends is from a typical dyad of non-friends. However, because of the severe class imbalance problem (see Figure 1c) a large number of non-friends can still be misclassified as friends. Therefore, we also provide the MCC measure, defined as follows:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP (true positives) is the number of friends correctly identified by the classifier, TN (true negatives) is the number of correctly identified non-friends, FP (false positives) is the number of non-friends incorrectly classified as friends, and FN (false negatives) is the number of friends that the classifier missed. The MCC does not have a straightforward interpretation similar to AUROC. However, it does reflect the problem of class imbalance—a value of MCC close to 0 indicates that a classifier should not be used, while a value closer to 1 shows that the classifier overcomes the imbalance problem.

### 4. RESULTS

We find that all four networks—Facebook friendships, Facebook interactions, text messages, and calls—can be reliably inferred from the close proximity events collected over a year. Our data suggests that there are different levels of psychological propinquity necessary for an edge to exist in each of the four networks. The population we study is better connected on Facebook than via text messages and calls, which indicates a lower threshold for becoming friends online than maintaining telephone communication exchanges.

Furthermore, we find that dyads who actively communicate (rather than just being friends on Facebook) also interact with higher intensity. For example in the month of March 2014:

- **Facebook friends.** 73% of Facebook friends met at least once during the month, including 43% who met outside of campus; 35% spent at least one hour together.
- **Facebook friends, who actively interact online.** 89% of Facebook friends who interact on Facebook met at least once during the month, including 64% who met outside of campus; 58% spent at least one hour together.
- **SMS friends.** 94% of text message contacts met at least once, including 80% who met outside of campus; 68% spent at least one hour together.
- **Call friends.** 93% of call contacts met at least once, including 80% who met outside of campus; 71% spent at least one hour together.

The results for each month are presented in Figure 3, revealing that the tendency holds during other months as well, with higher fraction of telecommunication contacts meeting on and outside of campus, and spending more time together.

Given the 16 features, we infer the four kinds of friendships among the study participants. We perform a five-fold cross validation procedure with a random forests classifier. In Figure 5a we report the mean AUROC and MCC scores (see Methods for interpretation) for each of the models in each prediction task. Additionally, we perform a similar procedure, this time only investigating links among people studying the same majors. The results reported in Figure 5b and 5d indicate that there is a strong signal of friendship even among people who are “forced” to spend multiple hours per
Figure 4: Relative importance of features in predicting the three kinds of links. The most important feature for predicting call/sms networks is the extra-role time weighted by the number of people present. In-role interactions are more important for inferring Facebook links than extra-role. Time at home is consistently the least important feature for inferring Facebook ties. The presented values are median importances of 10 runs of five-fold cross validation training of a Random Forest Classifier.

Figure 3: A vast majority of people who contact each other via phone/sms also meet in the real world, often outside of campus and for longer periods. People who interact on Facebook do as well, but only a smaller subset of them.

day together. AU ROC scores indicate that the difference between a pair Facebook typical friends and a pair of typical non-friends is less pronounced in the behavioral data that it is the case with interaction (facebook comments, call, sms) networks, see Figure 5. This finding is consistent with previous research on strength of Facebook ties [35, 17]. Finally, we present the importance of each feature as estimated by the Random Forest Classifier in Figure 4.

Figure 3 shows that they tend to spend most time together (e), both at campus (f) and outside (g), and at each other’s homes (h). We also notice that they tend to meet with fewer people present (i) and in places with a lower population density (j). Their interevent time is the lowest (k, l), which means that they not only spend more time together, but also meet more often. Higher entropy values (m-o) indicate that the timing of their meetings has less of a scheduled character than it is the case with other relationships. We also observe that call friends also have most similar friends. Behavioral signatures of Facebook friends (blue line) are somewhere between those of calling friends and people who are not friends on Facebook.

4.1 Long-term relationships

Finally, we investigate whether extending the observation time helps in the inference task. We build a matrix in which each row corresponds to a dyad and each column represents a month. For each month we train a random forest classifier to infer links in the call network. The classifier estimates and reports the probability of the existence of each link. We store this probability in the corresponding cells of our matrix. Then, for each month, we build a Logistic Regression model which uses these probabilities for the month and $N$ months before (Figure 6a) or after (Figure 6b). Figure 6 reports the increase in area under receiver operator curve introduced by exploiting more than one month of data. Using past or future data increases the performance of the inference for all
In this section we ask the following question: knowing the history of offline interactions of two people who communicate, can we determine which channel of communication they choose? In order to answer this question, in each month of data we find dyads which communicate using one channel but not another. We then build a model from the same 48 features we used for friendship inference and classify the type of friendship. We find that it is possible, to a certain degree, to discern people who interact on Facebook from people who interact via SMS or call.

4.2 Determining the type of friendship

We have already indicated that importance of features in the inference task depends on the friendship definition. In this section we ask the following question: knowing the history of offline interactions of two people who communicate, can we determine which channel of communication they choose? In order to answer this question, in each month of data we find dyads which communicate using one channel but not another. We then build a model from the same 48 features we used for friendship inference and classify the type of friendship. We find that it is possible, to a certain degree, to discern people who interact on Facebook from people who call and send messages. However, possibly due to the very high overlap between dyads who call and send messages (see Figure 2) our model fails to tell these two groups apart. We present the results in Table 2.

5. RELATED WORK

Infering social ties from co-presence events. Eagle et. al in their seminal work [12] were the first to explore the relationship between self-reported social ties and behavioral data collected through smartphones. Their analysis revealed a stronger correlation between the reported friendships and extra-role (off campus, off hours) than in-role meetings (ρ = 0.35 and ρ = 0.08 respectively). Crandall et al. extended the approach to include context beyond the simple on/off campus indication [9]. Their model included popularity of interaction locations, temporal entropy of the meetings, and neighborhood similarity between the nodes, and it outperformed the approach of Eagle et al.

Wang et al. [33] have shown that even co-locations inferred from comparatively low-resolution CDR data can be used to infer social ties.

In parallel to these development, researchers have also worked on the link prediction problem in settings where the continuous behavioral data is unavailable. Crandall et al. investigated the relationship between the number of unique locations visited by two people and the probability of them being friends in a photo-sharing service [8]. Scellato et al. extended this approach by introducing additional, inferred properties of locations shared among two people, such as the social entropy [28]. Other works showed that probability of friendship decreases with growing geographical distance [21], that clusters of friends tend to live nearby [27], and that friends meet in diverse locations [24]. There have also been developments into coupling the social and the mobility data beyond the task of link prediction. Intuitively, since maintaining a bond requires physical proximity, some of people’s mobility is driven by social factors. Several works argue that many non-routine travels observed in real data can be attributed to individuals seeking interaction with their social contacts [14, 32, 6].

Communication networks as proxies for real-world relationships. In this article we rely on networks of Facebook friendship and interactions as well as telecommunication networks as proxies for the existence of social ties. It is therefore important to cover research describing the applicability of such data in this context.

Wiese et al. [34] compared phone networks and self-reported friendships of 40 subjects. They found that while frequent communication indicates strong ties, lack of communication does not necessarily indicate a weak tie. Among other contributing factors they list the realization that people use multiple channels of communication (including face to face) and their phone networks do not fully describe their social network.

Table 2: The observable behaviors allow us to partially recognize friendship types: the behaviors of fb-only dyads are measurably different from behaviors of calling dyads.
networks. This finding might imply that many dyads who our models misclassify as call friends, are friends indeed, but use different communication means.

While studying users of Facebook, Golder et al. [13] and Wilson et al. [35] found that only a fraction of links present in the social graph represent dyads which interact actively on Facebook. Wilson recommends using the interaction graph instead of the declared friendships to better model the underlying social network. These insights were further confirmed by Jones’ research on inferring self-reported friendship ties from online interaction data [17]. They found that the strength of tie is correlated with the intensity of contact on Facebook, especially with commenting each other’s wall content. Furthermore, they found that private messages, to which we do not have access in this study, do not constitute a better indicator of real word friendship than wall posts.

Figure 7: Compared to Facebook friends (blue line) and Facebook non-friends (black line) people who call (yellow line) or message (green line) each other spend more time together (e-h) especially with only few others around (i, j), meet more often (k, l), and irregularly (m-o). They meet similar people (a-d). Behavioral signatures of Facebook friends lie between those of calling friends and people who are not friends on Facebook. The ties of people who call each other appear slightly more pronounced in face to face meetings than those who message each other.
Given the research described in this section, we believe that the existence of phone communication links can be treated as friendship signal among our population. We further confirm the findings from Wilson and Jones showing that Facebook interaction networks are more predictable from offline behavioral data than the Facebook friendship graph.

6. LIMITATIONS

Arguably, our study suffers from the lack of self-reported relationships and relying on communication channels for ground-truth. It would be an oversimplification to claim that two people calling one another or commenting on each other’s content online are necessarily friends. At the same time, there some friends might choose to communicate over means we have no access to (email, instant messaging, etc.). Our “ground-truth” unfortunately misses such cases, and our models perform worse as a result. There is, however, a body of research indicating that friendship does, in fact, manifest itself through the communication channels which we considered [12, 17, 34]. We believe that going towards large-scale experiments, scientist will increasingly need to rely on observable behaviors as proxies for cognitive relationships.

7. CONCLUSION

In this work we explored the interplay between offline and online interactions of a tight-knit group of nearly 1 000 students. We found that there are two orders of magnitude more links in the personal proximity network than those expressed through Facebook comments or phone calls. Despite this imbalance, it is possible to identify the signal of friendship in this pool of interactions. It has previously been shown that both Facebook interactions and phone calls reflect the existence of social ties. In this work we show that these reflections are not equivalent: dyads in our dataset tend to choose one of the two means of communication. Furthermore, we found that this choice is not random and can be predicted from offline interaction data. The relationships we form are instrumental to our well-being, give us access to resources, and influence our chances of success in life. This work shows that these ties can be inferred at scale from observable data on offline interactions.

Acknowledgements

The authors would like to thank Aniko Hannak and Jana Husman for their feedback on the manuscript. In this work we used the machine learning models implementations from scikit-learn Python package [23].

8. REFERENCES


[2] Google ceo on privacy (video): ‘if you have something you don’t want anyone to know, maybe you shouldn’t be doing it’. The Huffington Post.


[62] C. Grothoff and J. Porup. The nsa’s skynet program may be killing thousands of innocent people. *arstechnica.co.uk*.


